THE EVOLUTION OF DNA REPAIR
GENES, PROTEINS, AND PROCESSES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL SCIENCES
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jonathan Andrew Eisen

November 1998

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Dr. Philip C. Hanawalt (Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Dr. Marcus W. Feldman

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Dr. Allan M. Campbell

Approved for the University Committee on Graduate Studies

THE EVOLUTION OF DNA REPAIR
GENES, PROTEINS, AND PROCESSES

Jonathan A. Eisen
Stanford University, 1998

ABSTRACT

Genomic integrity is under nearly constant threat in all species. The primary mechanism by which organisms maintain their genomic integrity in the face of such threats is through DNA repair. In this thesis I discuss the interface between evolution and DNA repair. First, I discuss the use of comparative studies of repair genes and processes to study evolution. Specifically I discuss the development of the RecA gene as a model molecule for molecular systematic studies of bacteria. Then I discuss how differences in repair can drive evolution by discussing how differences in mismatch repair lead to variation in mutation rates and patterns at microsatellite loci. In the third section, I discuss how evolutionary studies can benefit our understanding of repair both in regard to structure-function studies (of the RecA protein) and in regard to studying diverse multigene families (in this case, the SNF2 family). Finally, in the last main section, I discuss my development of what I refer to as phylogenomics which combines evolutionary reconstructions and genome sequence studies into one composite analysis. The main reason I have developed the phylogenomic approach is that evolutionary studies can improve our understanding of genome sequences and genome sequences can improve inferences of evolutionary history so there is a feedback loop between the two types of study. In addition, I also present some additional results in Appendices regarding DNA turnover in *E. coli*, DNA repair in the extremely halophilic Archaea, and additional studies of the evolution of RecA.

ACKNOWLEDGEMENTS

My thesis has represented a relatively long and twisting road. In acknowledging the people who have helped make this possible I think it is useful to provide some of the history of each of the projects and my scientific development along the way. I have tried to make these brief and have put them in somewhat chronological order.

I owe my general interest in science and science research to my parents, Howard and Laura Eisen and to my grandfather Benjamin Post. They did not force me to become a scientist, but they did help me learn how to think critically and to appreciate some of the wonders of science.

As an undergraduate at Harvard, I was very fortunate to interact with many great biology researchers and teachers. During that time, I become interested in evolutionary biology and in particular in molecular evolution. The people I am particularly grateful to include: *Stephen J. Gould* (for his excellent class on evolution which was my first introduction to evolutionary biology as a science); *Wayne* and *David Maddison* (who, as Teaching Assistants for Gould's class, introduced me to computational evolutionary biology); *Stephen Austad* (for getting me interested in field biology research through the field "laboratories" for his Ecology class); *Eric Fajer* and *Scott Melvin* (for giving me my first experience designing a semi-independent research project for their Conservation Biology); *Alan Launer* (for lending me the equipment to collect fish from the pond, which, according to him, I never returned); *William A. Calder III* (for giving me my first experience as doing real science research at the Rocky Mountain Biological Laboratory); *Fakhri A. Bazzaz* (for being an excellent advisor throughout my time at Harvard and afterward, and for giving me a chance to do my first truly independent research project); *Peter Wayne, David Ackerly,* and *Susan Morse* for helping me with the jet-lag experiment; *P. Wayne,* again, for hiring me as a research assistant and teaching me about science research and culture; *Jennifer Doudna* (for introducing me to molecular evolution and for teaching me how to critically read a scientific paper); *Dennis Powers* (for introducing me to molecular ecology), *Colleen M. Cavanaugh* (for too many things to list here including teaching me to keep a good notebook and to do controls for every experiment, for introducing me to microbiology, and introducing me to the powers of a phylogenetic perspective in biological research); *Rob Dorit* and *Hiroshi Akashi* who helped me learn how to do some molecular biology experiments; and all my other teachers and colleagues at Harvard including *Woody Hastings, Karl Liem, Peter Ashton,* and *Otto Solbrig*.

And so, with a great debt to all of these people, I moved on to Stanford. Although I have officially worked on DNA repair in Phil Hanawalt's lab, I have benefited a great deal from many people at Stanford including: *Ward Watt* (for teaching me about biochemical evolution), *Sharon Long* (who somehow taught me many valuable lessons in a short rotation project but in particular, for infusing in me the benefits of working on an organism that has good genetic tools available); *Shi-Kau Liu* (for help in initial projects and for initiating all of my work on RecA); *Charlie Yanofsky* (for many things but in particular for helping me realize just how powerful it is to have a crystal structure of the protein one is interested in); *Kurt Gish,* for help with sequencing and cloning; *Allan Campbell* and *Richard Lenski* (for helping me discover some of the flaws of adaptive mutation experiments and thus leading me to look for a new project); *Patrick Keeling* and *W. Ford Doolittle* for getting me started working on *Haloferax volcanii; Mitch Sogin* and the Woods Hole Molecular Evolution course (for teaching me how molecular evolutionary methods worked and how to think about evolutionary questions at the molecular level); *David Botstein* (for convincing me that teaching and research are not incompatible, despite what many Stanford professors try to claim); all the people involved in the SME core, in particular *D. Botstein, Rick Myers, David Cox, Bob Simoni,* and *Brad Osgood; Rick Myers* (for encouraging me to develop methods and ideas behind phylogenomics); my brother *Mike Eisen* for help with virtually everything; *Sam Karlin* and *Volker Brendel* for teaching me about mathematical methods in molecular biology, and for making me be more critical of some of the methods used in molecular evolutionary studies; *David Ackerly* for teaching me about the uses of evolutionary approaches in comparative biology; *Russ Fernald* for general advice on life and science as well as for valuable discussions on the uses of evolutionary analysis in molecular biology; *Marc Feldman*, for many helpful discussions about evolution; *Steve Smith, David Swofford,* and *Joe Felsenstein* for making the GDE, PAUP, and PHYLIP computer programs freely available; *Steve Henikoff* and *Amos Bairoch* for inspiring me to put information about the gene families I work on onto the World Wide Web; and many of the people I have collaborated with over the years including *Bob Shafer* and *Michael Lerman*.

I am particularly grateful to my advisor *Philip C. Hanawalt* for allowing me the freedom to explore the areas of science that interested me and for being not only a great advisor, but a great human being too, showing me that one can do good scientific research without losing touch with ones humanity. I am also grateful to many of the members of the *Hanawalt Lab* including *David Crowley, Justin Courcelle,* and *Jennifer Halliday*.

Many people provided basic resources that helped me get my work done and get through my time at Stanford more easily including the library staff, in particular *Jill Otto; Steve* and *Pat* at the copy center; all the people in the Biostores especially *Darnell, Joe* and *Manual;* and all the people in the Biology Department main office.

On a personal level, I want to thank all of my many friends who have helped me get through graduate school whether it was by playing hockey and softball, going for bike rides, going hiking and camping doing many other things. These people include *Bob Fisher, Chester Washington, Jochen Kumm, David Pollock, David Goldstein, Joanna Mountain, Mo Badi, Grant Hoyt, Becky Taylor, James Keddie, Healy Hamilton, Sinan Suzen,* and *Aviv Bergman*. Many of these people are both good friends and colleagues. I am also grateful to my family, especially my brother *Mike Eisen,* my sister *Lisa Eisen,* and my mother *Laura Eisen,* my mothers parents *Annie* and *Ben Post,* and my little brother *Matthew Glenn,* my uncle *David Post,* and my pseudo-brother *Saul Jacobson* for support and encouragement. And in particular, I want to thank *Maria-Ines Benito* for being there through all the good and bad times and for listening to me read sections of papers she had no interest in, and for proofreading various papers and for just about everything.

I dedicate this thesis to the memory of my father, *Howard J. Eisen*.

AUTHORSHIP

Chapter 1 is entirely the work of myself (JAE). **Chapter 2a** is entirely the work of JAE. **Chapter 2b** is the result of a collaboration between JAE, Don A. Bryant, Tanja M. Gruber and Kurt Gish. It was written predominantly by TMG and JAE. The PCR experiments were done by JAE. The mapping experiments and the identification of full length clones were done by TMG. The sequencing was done by JAE, KG and TMG. The sequence and phylogenetic analysis was done by JAE. **Chapter 3** is entirely the work of JAE. **Chapter 4** is the result of a collaboration between JAE, Shi-Kau Liu, Irwin Tessman and Philip C. Hanawalt. It was written predominantly by JAE and SKL. The experimental work was done by SKL in IT's laboratory. The sequence and structural analysis was done by JAE. **Chapter 5** is the result of a collaboration between JAE, Kevin S. Sweder and Philip C. Hanawalt. It was written predominantly by JAE with some help from KSS. All analysis was done by JAE. **Chapter 6a** is the result of collaboration between JAE, Richard M. Myers and Dale Kaiser. It was written predominantly by JAE and RMM. The idea for the phylogenomic analysis and the analysis of MutS was the work of JAE. **Chapter 6b** and **Chapter 6c** are entirely the work of JAE. **Chapter 7** is the result of a collaboration between JAE and Philip C. Hanawalt. It was written predominantly by JAE. All analysis was done by JAE. All **Appendices** are entirely the work of JAE.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1 TABLES AND FIGURES

CHAPTER 2A TABLES AND FIGURES

CHAPTER 2B TABLES AND FIGURES

CHAPTER 3 TABLES AND FIGURES

CHAPTER 4 TABLES AND FIGURES

CHAPTER 5 TABLES AND FIGURES

CHAPTER 6A TABLES AND FIGURES

CHAPTER 6B TABLES AND FIGURES

CHAPTER 6C TABLES AND FIGURES

CHAPTER 7 TABLES AND FIGURES

APPENDIX A TABLES AND FIGURES

APPENDIX B TABLES AND FIGURES

APPENDIX C TABLES AND FIGURES

APPENDIX D TABLES AND FIGURES

APPENDIX E TABLES AND FIGURES

APPENDIX F TABLES AND FIGURES

APPENDIX G TABLES AND FIGURES

CHAPTER 1

Introduction
and
Summary of Thesis Work

"Nothing in biology makes sense except in the light of evolution."

Theodosius Dobzhansky[1]

"The theory of evolution is quite rightly called the greatest unifying theory in biology. The diversity of organisms, similarities and differences between kinds of organisms, patterns of distribution and behavior, adaptation and interaction, all this was merely a bewildering chaos of facts until given meaning by the evolutionary theory."

Ernst Mayr[2]

[1] in Dobzhansky, T.H. 1973. *Nothing in biology makes sense except in the light of evolution.* American Biology Teacher, 35, 125-129.

[2] in Mayr, E. 1970. *Populations, Species, and Evolution*. Harvard University Press, Cambridge, MA.

SUMMARY

Studies of evolution and studies of DNA repair have a great deal of overlap. This interface between evolution and DNA repair has been the focus of my thesis research. In this chapter, I give a brief introduction to the field of molecular evolution and to DNA repair processes. I discuss some of the different aspects of the interface between repair and evolution including how evolutionary studies can benefit our understanding of repair and how comparative studies of repair can help better understand evolution. In addition, I provide a summary of the different sections of this thesis and how they relate to the interface between evolution and DNA repair.

THE INTERFACE BETWEEN EVOLUTION AND DNA REPAIR

Molecular Evolution

The field of molecular evolution has sought to combine molecular biology and evolutionary biology into one area of study. While there are many different facets to molecular evolutionary studies, I believe it is possible to divide the field into two major subfields: (a) the application of molecular techniques and data to evolutionary questions (which I refer to as molecular evolutionary biology) and (b) the application of evolutionary techniques and data to molecular questions (which I refer to as evolutionary molecular biology). While molecular evolution can trace its roots back to studies of population genetics from the early 1900s, the impetus for much of molecular evolution came in 1962 with the publication of a classic paper by Emil Zuckerkandl and Linus Pauling, entitled "Molecules as documents of evolutionary history" (1). In this article, Zuckerkandl and Pauling argued that comparison of gene sequences between species could be used to infer the evolutionary history of species. Since the publication of this paper, there has been a revolution in molecular biology. This revolution is best seen in regard to gene sequencing techniques that continue to get faster, easier, and cheaper, allowing sequence data to accumulate at an amazing pace. While Zuckerkandl and Pauling discussed the uses of comparisons of single genes of different species, it is now possible to compare the sequences of entire genomes of different species.

The revolution in molecular biology has been seized upon by evolutionary biologists and the molecular evolutionary biology side of molecular evolution has developed spectacular momentum. Thus, the primary way to determine the evolutionary history of species is now through gene sequence comparisons. Sequence comparisons have also been used to infer selective constraints on different genes, to study the process of selection at the molecular level, and for many other evolutionary studies. While molecular evolutionary biology has flourished, the same cannot be said for evolutionary molecular biology. This is somewhat surprising because it is well established that an evolutionary perspective can benefit any aspect of comparative biology and comparative molecular biology is no exception. The reason that an evolutionary perspective is beneficial in comparative studies is that all organisms have an evolutionary history, and thus, to understand what the differences and similarities among species mean, it is helpful to understand how and why these differences arose. An evolutionary perspective has been used extensively in many fields of comparative biology including physiology, developmental biology, and ecology. So why has an evolutionary perspective not seen much use in comparative molecular biology? It is certainly not because there are no examples of the uses of evolution in molecular biology. Many areas of molecular biology have benefited a great deal from evolutionary analysis (see Table 1 for examples). I believe there are two major reasons for the limited use of evolutionary methods in comparative molecular biology. First, with the accumulation of so much molecular data, the focus of most comparative molecular biology research has been simply on quantifying the similarities and differences among species rather than studying the origins of these similarities and differences. In addition, most evolutionary biologists who work in the field of molecular evolution have focused on the first side of molecular evolution - the use of molecular data to study evolutionary questions. For molecular evolution to be truly a field that works at the interface of evolutionary and molecular biology there needs to be much more focus on the second side of molecular evolution - evolutionary molecular biology.

DNA Repair

Genomic integrity is under constant threat in all species. Threats come in the form of endogenous and exogenous agents that damage DNA and/or interfere with DNA metabolic processes, as well as spontaneous base loss or deamination and errors in DNA metabolism such as nucleotide misincorporation during replication. These threats lead to a variety of alterations in the normal DNA structure including single- and double-strand breaks, chemically modified bases, abasic sites, bulky adducts, inter- and intra-strand cross-links, and base-pairing mismatches. The direct effects of these abnormalities include mutations at or near the site of the abnormality, genetic recombination, and the inhibition or alteration of cellular processes such as replication and transcription. These direct effects can lead in turn to many indirect effects including chromosomal aberrations, tumorigenesis, developmental abnormalities, apoptosis, and/or necrosis.

The primary mechanism by which organisms maintain their genomic functions in the face of these threats is by removing the abnormalities from the DNA and restoring the genomic integrity, a process known as DNA repair. Experimental studies in a variety of species have documented an incredible diversity of repair pathways. Fortunately, the comparison of repair pathways is simplified by the fact that all repair pathways can be placed into one of three classes based on its general mechanism of action: direct repair, recombinational repair, and excision repair. In direct repair, alterations in the structure of DNA are simply reversed. Examples include photoreactivation, alkyltransfer, and DNA ligation. In recombinational repair, sections of altered or damaged DNA are corrected by homologous recombination with undamaged templates (see (2) for review). Thus, there is a great deal of overlap between the pathways involved in general recombination and those involved in recombinational repair. Finally, in excision repair, first a section of one strand of the DNA double-helix containing the abnormality is excised, then the other strand is used as a template to correctly resynthesize the removed section, and finally the patch is ligated into place (see (3) for review). There are three major forms of excision repair that are distinguished by the type of abnormality removed and by the mechanism of its recognition and removal. In base excision repair (BER), inappropriate, damaged, or modified bases are removed and the resulting abasic site is repaired by a process that replaces only one or a few nucleotides; in nucleotide excision repair (NER) abnormal

DNA structures are removed as part of an oligonucleotide and longer patches are introduced; and in mismatch repair (MMR) base mismatches or unpaired loops are removed as part of a very long stretch of nucleotides. More detail on the different repair pathways is provided in Chapter 7.

Comparisons on a species by species basis reveal that some aspects of repair are similar between species and some are different. All species examined in detail have been found to exhibit multiple repair pathways, usually including many of the different classes and types of repair. Although the use of multiple repair pathways is likely universal, the repertoire of types of repair frequently differs between species. In addition, although each particular class of repair is similar in all species in which it is found, close examination of the details of the processes in different species reveals a great deal of diversity in how particular species carry out the respective classes of repair. For example, although all PHR processes are similar in different species, the specificity varies between and even within species. In some species PHR reverses only pyrimidine dimers, in others it reverses only 6-4 photoproducts, and some species have multiple PHR processes that are able to repair both CPDs and 6-4s. Differences in specificity, some subtle, some large, are found in almost all classes of repair. Thus, the finding that two species exhibit the same repertoire of repair types does not mean that they have identical repair processes.

Molecular Evolution and DNA Repair

The study of DNA repair and evolution have a great deal of overlap and these areas represent the interface between evolution and DNA repair (see Table 2). First, comparative studies of repair genes and processes can be used to study evolution. These are aspects of this interface that represent molecular evolutionary biology. For example, comparative studies of repair genes can be used to infer evolutionary history of species. In addition, comparative studies can be used to understand the evolutionary history of DNA repair processes. Also, since differences in repair processes can have profound biological effects (see Table 2 for a listing of some), to understand the evolution of these phenotypes it is necessary to understand the differences in repair.

Second, evolutionary analysis can improve our understanding of repair genes and processes. These are the aspects of the interface that represent evolutionary molecular biology. Evolutionary studies of repair processes help understand the differences between species. For example, evolutionary studies of photoreactivation show that all photoreactivation processes are homologous and that the differences between species (6-4 vs. CPD photoreactivation) are due to functional changes in photolyase enzymes (4). Evolutionary studies have many other potential uses in the study of repair including the characterization of genes that are part of multigene families (5-7), the prediction of functions for uncharacterized genes (8); the identification of motifs conserved among particular homologs (7); and the study of structure-function relationships of repair genes.

The interface between repair and evolution is of particular interest because of the role that repair processes have in influencing evolutionary patterns. Since repair processes influence mutation rates and patterns, differences in repair can lead to different mutational and evolutionary patterns within and between species. For examples, the high mutation rate in animal mitochondria relative to the animal nucleus could be explained by deficiencies in certain repair processes in the mitochondria. Similarly, the high rate of nucleotide substitution in mycoplasmas may be due in part to deficiencies in DNA repair (9,10). Other cases of mutation rate differences being due to DNA repair differences include the higher rate of mutation in rodents than primates (11), the low microsatellite mutation rate in flies (12), the strand bias in C->T changes in *E. coli* (13) and mutation hotspots within the p53 gene in humans (14).

Finally, the fact that repair processes play a part in controlling the mutation rates and patterns of different species means that all analysis that make use of sequence comparisons between species can benefit from a better understanding of repair processes in different species. For example, many aspects of sequence analysis such as database searches, phylogenetic analysis, sequence alignment generation and population analysis are improved when they include information on mutation rates and patterns such as transition-transversion patterns, microsatellite mutation mechanisms and insertion-deletion rates. The reverse of this is also true - since repair processes influence mutation rates and patterns, evolutionary analysis of sequences can be used to identify mutation rates and patterns.

The many areas of overlap between DNA repair and evolution mean that it is of interest to study the interface between repair and evolution. Below I summarize the different areas of my thesis work and give a brief description of how each fits in to studies of evolution and DNA repair.

SUMMARY OF CHAPTERS AND APPENDICES

Using comparative analysis of DNA repair genes to study evolutionary history

In **Chapter 2,** the use of comparative analysis of sequences of *recA* genes from different species to infer the evolutionary relationships among species is discussed. **Chapter 2a** reports the results of a comparison of evolutionary trees of RecAs and 16s rRNAs from the same set of species (15). In this study, I found that, when the same species sets and methods were used to generate trees of RecAs and rRNAs the trees were highly congruent and had similar powers to resolve phylogenetic relationships. The main conclusions of this analysis are 1) that molecular phylogenetic analysis is reliable 2) if lateral transfers of genes have occurred between bacterial species they likely did not involve *recA* or *rRNA* genes and 3) that RecA comparisons are a useful tool for systematic studies of bacteria. **Chapter 2b** reports the cloning and phylogenetic analysis of *recA* genes from species in the genera *Chlorobium* (green sulfur bacteria) and *Chloroflexus* (green non-sulfur bacteria) (16). The phylogenetic analysis in this study is consistent with the results in Chapter 2a. In addition, this analysis helps confirm the phylogenetic position of *Chlorobium* and *Chloroflexus.* Finally, **Appendix E** presents two figures relating to the cloning of the *recA* gene of *Caulobacter crescentus* which was analyzed in Chapter 2b.

Effect of differences in DNA repair on evolution

In **Chapter 3,** I review the literature concerning how differences in the mechanism of mismatch repair (MMR) can lead to differences in mutation patterns at microsatellites (loci that contain small 1-10 bp tandem repeats).

Using evolutionary analysis to better understand DNA repair processes

In **Chapter 4,** evolutionary analysis of RecA sequences is used to help understand second-site mutations in the *E. coli recA* gene that suppress the phenotype of the *recA1202* mutation (17). Some additional figures relating to this are reported in **Appendix B.** This analysis of these mutations was followed up by a more detailed analysis of RecA structural evolution which is reported in **Appendix C.** In **Chapter 5,** evolutionary analysis is used to help better understand the SNF2 family of proteins (5). In this chapter I introduce the concept of using evolutionary analysis to make functional predictions for uncharacterized genes. In addition, I also discuss some additional uses of evolutionary analysis in studies of genes in multigene families.

Development of phylogenomic analysis

In **Chapters 6-7,** I present work on the development of a new approach that combines the analysis of complete genome sequences with evolutionary reconstructions into one phylogenomic analysis. In **Chapter 6a,** I introduce the concept of "phylogenomic" analysis for the prediction of gene functions (18). In **Chapter 6b,** I discuss the use of phylogenomic analysis for functional predictions in more detail and discuss some of the advantages of evolutionary functional predictions over other methods such as blast searches (8). In **Chapter 6c,** I present a more complete phylogenomic methodology in which evolutionary and genomic analyses are used for all aspects of the study of a gene family (and not just functional predictions) (6). I focus this analysis on the MutS family of proteins. Finally, in **Chapter 7,** I present a phylogenomic analysis of all repair genes. This analysis is used to reconstruct the evolutionary history of DNA repair proteins and DNA repair processes; to infer repair genes likely to have been present in the ancestor of all living organisms are identified and to predict the likely repair capabilities of these species.

DNA repair in Haloferax volcanii

In **Appendix F** I present the results of studies that document the repair of UV irradiation induced cyclobutane pyrimidine dimers in extremely halophilic Archaea *Haloferax volcanii* and discuss some of the reasons why studies of repair in Archaea, and in this species in particular are of interest. In **Appendix G** I present results on the cloning of a homolog of the mismatch repair gene MutL from *H. volcanii.*

DNA turnover, thymineless death and stationary phase mutations

In **Appendix A,** I present some results of preliminary experiments on DNA turnover, thymineless death and stationary phase mutagenesis in *E. coli.* Certain features of these phenomena have suggested that they may be related.

REFERENCES

1. Zuckerkandl, E. and Pauling, L. (1965) *J Theor Biol,* **8,** 357-366.
2. Camerini-Otero, R. D. and Hsieh, P. (1995) *Annu Rev Genet,* **29,** 509-552.
3. Sancar, A. (1996) *Annu Rev Biochem,* **65,** 43-81.
4. Kanai, S., Kikuno, R., Toh, H., Ryo, H. and Todo, T. (1997) *J Mol Evol,* **45,** 535-548.
5. Eisen, J. A., Sweder, K. S. and Hanawalt, P. C. (1995) *Nucleic Acids Res,* **23,** 2715-2723.
6. Eisen, J. A. (1998) *Nucleic Acids Res,* **26,** 4291-4300.
7. Henikoff, S., Greene, E. A., Pietrovsky, S., Bork, P., Attwood, T. K. and Hood, L. (1997) *Science,* **278,** 609-614.
8. Eisen, J. A. (1998) *Genome Res,* **8,** 163-167.
9. Dybvig, K. and Voelker, L. L. (1996) *Annu Rev Microbiol,* **50,** 25-57.
10. Labarére, J. (1992) In Maniloff, J. (ed.), Mycoplasmas: Molecular Biology And Pathogenesis. American Society For Microbiology, Washington, D. C., pp. 309-323.
11. Li, W. H., Gouy, M., Sharp, P. M., O'hUigin, C. and Yang, Y. W. (1990) *Proc Natl Acad Sci U S A,* **87,** 6703-6707.
12. Schug, M. D., Mackay, T. F. and Aquadro, C. F. (1997) *Nature Genetics,* **15,** 99-102.
13. Francino, M. P., Chao, L., Riley, M. A. and Ochman, H. (1996) *Science,* **272,** 107-109.
14. Tornaletti, S. and Pfeifer, G. P. (1994) *Science,* **263,** 1436-1438.
15. Eisen, J. A. (1995) *J Mol Evol,* **41,** 1105-1123.
16. Gruber, T. M., Eisen, J. A., Gish, K. and Bryant, D. A. (1998) *FEMS Microbiol Lett,* **162,** 53-60.
17. Liu, S. K., Eisen, J. A., Hanawalt, P. C. and Tessman, I. (1993) *J Bacteriol,* **175,** 6518-6529.
18. Eisen, J. A., Kaiser, D. and Myers, R. M. (1997) *Nature (Medicine),* **3,** 1076-1078.

Table 1. Examples of uses of evolutionary analysis in molecular biology

Use of Evolution	Comments
Characterization of mutation patterns and mutational events	Phylogenetic analysis of sequences in different species has allowed the identification of mutation biases and patterns. Transition-transversion bias Mutation rates depend on neighboring sequences Mutation patterns different in nucleus vs. mitochondria Phylogenetic analysis helps identify gene duplications, inversions, deletions, and lateral transfers
Prediction or confirmation of secondary and tertiary structures	Amino-acid/nucleotide substitution patterns useful in structural studies. Allowed solving of tRNA and rRNA structures
Sequence searching matrices	Evolutionary substitution patterns improve sequence searching matrices (e.g., PAM, Blosum)
Motif analysis	Motif analysis is improved when evolutionary distances between sequences are incorporated in analysis (e.g., BLOCKS method).
Prediction of gene function	Phylogenetic relationships among genes can be used to improve functional predictions for uncharacterized genes.
Classification of multigene families	Phylogenetic relationships among genes are the best way of classifying genes in multigene families.
Identification of correlated evolutionary events	Correlated gene loss may reveal interactions among genes.

Table 2. The interface between evolution and DNA repair

DNA repair differences influence many biological phenotypes

- Lifespan
- Pathogenesis
- Cancer rates
- Codon usage and GC content
- Evolutionary rates
- Speciation
- Survival in extreme environments
- Diurnal/nocturnal patterns

DNA repair differences can lead to differences in mutation and evolutionary rates and patterns

- High relative rate in mitochondria, rodents, mycoplasmas
- Hotspots within genes (e.g., p53)
- Log vs. stationary phase
- Strand bias

DNA substitution patterns between species help identify mutation and repair biases

- Transitions >> transversions in many species
- C -> T transitions very high in mitochondria
- Patterns vary with neighboring bases

Evolutionary analysis helps characterized proteins and pathways

- Division of multigene families into subfamilies/orthologous groups
- Identification of conserved motifs (e.g., BLOCKS)
- Predict or confirm protein structures
- Predict functions of uncharacterized genes
- Understanding of functional changes (e.g., Phr)
- Correlated gain/loss of genes may help understand pathways
- Loss of genes may be correlated with other biological changes

Information about mutation mechanisms improves sequence comparisons/searches

- Transition-transversion bias in calculating distances between sequences
- PAM and BLOSUM matrices in database searches
- Microsatellite step-wise mutations for population analysis
- INDEL rates useful for searches and alignments

Evolutionary history of repair helps put other information in perspective

- Which processes lost in reduced genomes?
- Lateral transfers of repair genes.
- Which processes are ancient?
- Organismal utility of repair processes.

CHAPTER 2

Using Comparative Analysis of DNA Repair Genes to Study the Evolutionary History of Species

PART A

The RecA Protein as a Model Molecule for Molecular Systematic Studies of Bacteria: Comparison of Trees of RecAs and 16S rRNAs from the Same Species[3]

[3] Previously published as *Jonathan A. Eisen. 1995. Journal of Molecular Evolution 41(12): 1105-1123.* Reprinted with permission of Springer-Verlag Incorportated.

ABSTRACT

The evolution of the RecA protein was analyzed using molecular phylogenetic techniques. Phylogenetic trees of all currently available complete RecA proteins were inferred using multiple maximum parsimony and distance matrix methods. Comparison and analysis of the trees reveal that the inferred relationships among these proteins are highly robust. The RecA trees show consistent subdivisions corresponding to many of the major bacterial groups found in trees of other molecules including the α, β, γ, δ, and ε Proteobacteria, cyanobacteria, high-GC gram-positives, and the *Deinococcus*-*Thermus* group. However, there are interesting differences between the RecA trees and these other trees. For example, in all the RecA trees the proteins from gram-positives species are not monophyletic. In addition, the RecAs of the cyanobacteria consistently group with the RecAs of the high-GC gram-positives. To evaluate possible causes and implications of these and other differences, phylogenetic trees were generated for small-subunit rRNA sequences from the same (or closely related) species as represented in the RecA analysis. The trees of the two molecules using these equivalent species-sets are highly congruent and have similar resolving power for close, medium, and deep branches in the history of bacteria. The implications of the particular similarities and differences between the trees are discussed. Some of the features that make RecA useful for molecular systematics and for studies of protein evolution are also discussed.

INTRODUCTION

Molecular systematics has become the primary way to determine evolutionary relationships among microorganisms because morphological and other phenotypic characters are either absent or change too rapidly to be useful for phylogenetic inference (Woese 1987). Not all molecules are equally useful for molecular systematic studies and the molecule of choice for most such studies of microorganisms has been the small-subunit of the rRNA (SS-rRNA). Comparisons of SS-rRNA sequences have revolutionized the understanding of the diversity and phylogenetic relationships of all organisms, and in particular those of microorganisms (Fox et al. 1980, Olsen 1988, Olsen et al. 1994, Pace et al. 1986, Sogin 1989, Woese 1991, Woese 1987). Some of the reasons that SS-rRNA sequence comparisons have been so useful include: SS-rRNAs are present in, and have conserved sequence, structure, and function among, all known species of free-living organisms as well as mitochondria and chloroplasts (Pace et al. 1986, Woese 1987); genes encoding SS-rRNAs are relatively easy to clone and sequence even from uncharacterized or unculturable species (Eisen et al. 1992, Lane et al. 1985, Medlin et al. 1988, Olsen et al. 1986, Weisburg et al. 1991); the conservation of some regions of primary structure and large sections of secondary structure aids alignment of SS-rRNA sequences between species (Woese 1987); the evolutionary substitution rate between species varies greatly within the molecule allowing for this one molecule to be used to infer relationships among both close and distant relatives (Pace et al. 1986, Woese 1987); and it is generally considered unlikely that SS-rRNA genes have undergone lateral transfers between species (Pace et al. 1986), thus the history of SS-rRNA genes should correspond to the history of the species from which they come. The accumulating database of SS-rRNA sequences, which now includes over 3000 complete or nearly complete sequences (Maidak et al. 1994), provides an extra incentive to focus on this molecule.

Despite the advantages and successes of using SS-rRNA sequences to determine microbial phylogenetic relationships, there are potential problems with relying on only SS-rRNA-based phylogenetic trees (e.g., Hasegawa and Hashimoto 1993, Rothschild et al. 1986). First, there are some characteristics of SS-rRNA genes that may lead to trees based on them being inaccurate including: over-estimation of the relatedness of species with similar nucleotide frequencies (such as could occur in unrelated thermophiles) (Embley et al. 1993, Vawter and Brown 1993, Viale et al. 1994, Weisburg et al. 1989b, Woese et al. 1991), non-independence of substitution patterns at different sites (Gutell et al. 1994, Schoeniger and Von Haeseler 1994), variation in substitution rates between lineages (e.g., Wolfe et al. 1992, Bruns and Szaro 1992, Nickrent and Starr 1994), and ambiguities in alignments between distantly related taxa. Even if the trees inferred from SS-rRNA genes accurately reflect the evolutionary history of these genes, they might not accurately reflect the history of the species as a whole. For example, lateral transfers between species might cause the genomes of some species to have mosaic evolutionary histories. Although it is unlikely that SS-rRNAs have been stably transferred between species (see above), other genes may have been. Therefore, to understand the history of entire genomes, and to better understand the extent of mosaicism within species, it is important to compare and contrast the histories of different genes from the same species. Finally, since SS-rRNA genes are present in multiple copies in many bacteria (Jinks-Robertson and Nomura 1987, Nomura et al. 1977), it is possible that the genes being compared between species are paralogous not orthologous. This could cause the gene trees to be different from the species trees. For these and other reasons, researchers interested in microbial systematics have begun to compare and contrast the relationships of other molecules with those of the SS-rRNA. The choice of which additional molecule to use is a difficult one. Many potential candidates have arisen and each has its advantages. Examples include HSP70 (Boorstein et al. 1994, Gupta et al. 1994, Rensing and Maier 1994), GroEL (Viale et al. 1994), EF-TU (Ludwig et al. 1994; Delwiche et al. 1995), ATPase-β-subunit (Ludwig et al. 1994), 23S rRNA (Ludwig et al. 1992), and RNA polymerases (Klenk and Zillig 1994). Another potential choice is RecA.

The RecA protein of *Escherichia coli* is a small (352 aa) yet versatile protein with roles in at least three distinct cellular processes: homologous DNA recombination, SOS induction, and DNA damage induced mutagenesis (Kowalczykowski et al. 1994). This diversity of genetic functions is paralleled by multiple biochemical activities including DNA binding (double and single-stranded), pairing and exchange of homologous DNA, ATP hydrolysis, and coproteolytic cleavage of the LexA, λcI, and UmuD proteins (Kowalczykowski et al. 1994). It has been 30 years since the isolation of the first *recA* mutants in *E. coli* (Clark and Margulies 1965) and 15 years since the sequencing of the corresponding *recA* gene (Sancar et al. 1980; Horii et al. 1980). In that time, studies of the wild type and mutant RecA proteins and genes have yielded a great deal of information about the structure-function relationships of the protein, as well as about the general mechanisms of homologous recombination (Clark and Sandler 1994, Kowalczykowski 1991, Roca and Cox 1990). Such studies have been facilitated greatly by the publication of the crystal structure of the *E. coli* RecA protein alone, and bound to ADP (Story and Steitz 1992, Story et al. 1992).

Genes encoding proteins with extensive amino-acid sequence similarity to the *E. coli* RecA have been cloned and sequenced from many other bacterial species. Included among these are complete open reading frames from many of the major bacterial phyla as well as an open reading frame from the nucleus of *Arabidopsis thaliana* that encodes a protein that functions in the chloroplast (Table 1). Partial open reading frames are available from many additional bacterial species. The high levels of sequence similarity, even between proteins from distantly related taxa, and the demonstration that many of the functions and activities of the *E. coli* RecA are conserved in many of these other proteins (Angov and Camerini-Otero 1994, Gutman et al. 1994, Roca and Cox 1990, Wetmur et al. 1994), suggest that these proteins are homologs of the *E. coli* RecA.

The diversity and number of species from which sequences are available makes RecA a potentially useful tool for molecular systematic studies of bacteria. Previously, Lloyd and Sharp (1993) tested the utility of RecA comparisons for phylogenetic studies. They concluded that RecA comparisons were probably only useful for determining relationships among closely related bacterial species. However, they were limited by the number and diversity of RecA sequences that were available at the time. I have re-analyzed the evolution of RecA using 40 additional sequences. In this paper, analysis is presented that shows that the RecA protein is a good alternative or supplement to SS-rRNA for molecular systematic studies of all bacteria, not just of closely related species. Phylogenetic trees of the 65 complete RecA protein sequences were inferred using a variety of phylogenetic methods. Statistical analysis and comparisons of trees generated by the different phylogenetic methods suggests that the RecA phylogeny is highly consistent and robust. The RecA trees are compared to trees of SS-rRNA sequences from the same or very closely related species as represented in the RecA trees. Overall, the trees of the two molecules are highly congruent. The implications of the particular similarities and differences between the RecA-based and SS-rRNA-based trees are discussed. Some of the features of RecA that make it a potentially useful molecular chronometer are also discussed.

METHODS

Sequences and alignment

All RecA sequences used in this paper were obtained from the National Center for Biotechnology Information (NCBI) databases by electronic mail (Henikoff 1993) except for those from *Methylophilius methylotrophus* (Emmerson 1995), *Xanthomonas oryzae* (Mongkolsuk 1995), *Synechococcus* sp. PCC7942 (Coleman 1995), and *Borrelia burgdorferi* (Huang 1995) which were kindly provided prior to submission. Accession numbers for those in databases are given in Table 1. The amino-acid sequences of the RecA proteins were aligned both manually and with the *clustalw* multiple sequence alignment program (Thompson et al. 1994). The RecA alignment was used as a block and aligned with the sequences of the RadA protein from an Archaea (Clark and Sandler 1994, Clark 1995) and RecA-like proteins from eukaryotes (Ogawa et al. 1993), also using *clustalw*.

For the comparison of RecA and SS-rRNA trees, a complete or nearly complete SS-rRNA sequence was chosen to represent each species for which a complete RecA protein was available. For most of the RecA proteins, a complete SS-rRNA sequence was available from the same species. The remaining species (those for which a RecA sequence was available but a complete or nearly complete SS-rRNA was not) were represented by a "replacement" SS-rRNA from a different species. The choice of which replacement sequence to use was determined in one of two ways. For those RecAs for which a partial SS-rRNA was available from the same species, the complete or nearly complete SS-rRNA that was most similar to the partial sequence was used. Similarity was determined by comparisons using the Ribosomal Database Project (RDP) computer server (Maidak et al. 1994) and blastn searches (Altschul et al. 1990) of the NCBI databases by electronic mail (Henikoff 1993). For those RecAs for which even a partial SS-rRNA sequence was not available from the same species, a replacement SS-rRNA was chosen from a species considered to be a close relative. A SS-rRNA was not used to represent the *Shigella flexneri* RecA because this protein was identical to the *E. coli* RecA. For the majority of the SS-rRNA phylogenetic analysis, only one SS-rRNA sequence was used to represent the two RecAs from *Myxococcus xanthus.* For some of the analysis an additional SS-RNA from a close relative of *M. xanthus* was also included. The SS-rRNA sequences used and the species from which they come are listed in Table 1. The SS-rRNA sequences were obtained already aligned from the RDP (Maidak et al. 1994), with the exception of those from *Corynebacterium glutamicum* and *Anabaena* sp. PCC7120, which were obtained from the NCBI and were aligned to the other sequences manually. Entry names and numbers are listed in Table 1.

Phylogenetic trees

Phylogenetic trees were generated from the sequence alignments using computer algorithms implemented in the PHYLIP (Felsenstein 1993), PAUP (Swofford 1991), and GDE (Smith 1994, Smith et al. 1994) computer software packages. Trees of the RecA sequences were generated using two parsimony methods (the *protpars* program in PHYLIP and the *heuristic search* algorithm of PAUP) and three distance methods (the least-squares method of De Soete (De Soete 1983) as implemented in GDE, and the Fitch-Margoliash (Fitch and Margoliash 1967) and neighbor-joining methods (Saitou and Nei 1987) as implemented in PHYLIP). Trees of the SS-rRNA sequences were generated using one parsimony method (the *dnapars* algorithm of PHYLIP) and the same three distance methods as used for the RecA trees. For the trees generated by the *protpars*, *dnapars*, Fitch-Margoliash, and neighbor-joining methods, 100 bootstrap replicates were conducted by the method of Felsenstein (1985) as implemented in PHYLIP.

For the distance-based phylogenetic methods listed above, estimated evolutionary distances between each pair of sequences were calculated for input into the tree-reconstruction algorithms. Pairwise distances between RecA proteins were calculated using the *protdist* program of PHYLIP and the PAM matrix-based distance correction (Felsenstein 1993). Pairwise distances between SS-rRNA sequences were calculated in two ways: the method of Olsen (1988) (as implemented by the *count* program of GDE) was used for the trees generated by the De Soete method; and the two-parameter model of Kimura (1980) (as implemented by the *dnadist* program of PHYLIP) was used for the Fitch-Margoliash and neighbor-joining trees.

Regions of the alignments for which homology of residues could not be reasonably assumed were excluded from the phylogenetic analysis. For the SS-rRNA trees, the alignment of SS-rRNA sequences was extracted from an alignment of thousands of sequences in the RDP database (Maidak et al. 1994). This RDP alignment was generated using both primary and secondary structures as a guide to assist in the assignment of homology (Maidak et al. 1994). Therefore it was assumed that the aligned regions were likely homologous. Nevertheless, regions of high sequence variation (as determined by a 50% consensus mask using the *consensus* program of GDE) were excluded from the phylogenetic analysis since these regions are perhaps most likely to contain non-homologous residues. The SS-rRNA alignment and a list of the 1061 alignment positions used for phylogenetic analysis are available on request. For the RecA analysis, the assignment of homology in the alignment was based only on similarity of primary structure (as determined by the *clustalw* program). Regions of ambiguity in the alignment were considered to potentially include non-homologous residues and thus were excluded from the phylogenetic analysis. Such regions were identified by comparing alignments generated by the *clustalw* program using a variety of alignment parameters. Parameters varied included scoring matrices (PAM, BLOSUM, and identity matrices were used) and gap opening and extension penalties. Alignments were compared by eye to detect differences and those regions that contained different residues in the different alignments were considered ambiguous.

Character states and changes

Analysis of character states and changes over evolutionary history was done using the MacClade 3.04 program (Maddison and Maddison 1992). For each alignment position, all unambiguous substitutions as well as all unambiguous non-conservative substitutions were counted. Non-conservative substitutions were defined as amino-acid changes that were not within the following groups: (V-I-L-M), (F-W-Y), (D-E-N-Q), (K-R), (G-A), and (S-T).

Computer programs

GDE, PHYLIP, and *clustalw* were obtained by anonymous FTP from the archive of the Biology Department at the University of Indiana (ftp.bio.indiana.edu). PAUP was obtained from David Swofford and is now available from Sinauer Associates, Inc., Sunderland, MA. GDE, PHYLIP, and *clustalw* were run on a Sparc10 workstation and MacClade and PAUP on a Power Macintosh 7100/66. Unless otherwise mentioned, all programs were run with default settings.

RESULTS AND DISCUSSION

The potential of using RecA for phylogenetic studies of bacteria was first addressed by Lloyd and Sharp (1993). In a detailed analysis of the evolution of *recA* genes from 25 species of bacteria, they showed that phylogenetic trees of RecA proteins appeared to be reliable for determining relationships among closely related bacterial species. Specifically, for the Proteobacteria, the branching patterns of RecA proteins were highly congruent to branching patterns of SS-rRNA genes from the same or similar species. However, the RecA and SS-rRNA trees were not highly congruent for relationships between sequences from more distantly related species. Lloyd and Sharp concluded that this was due to a low resolution of the deep branches in the RecA tree. However, this low resolution of deep branches could have been due to poor representation of certain taxa in their sample set. Of the *recA* sequences available at the time, only six were from species outside the Proteobacteria. The diversity as well as the number of *recA* sequences available has increased greatly since Lloyd and Sharp's study (see Table 1). Therefore, I have re-analyzed the evolution of *recA* including these additional sequences with a specific focus on determining whether *recA* comparisons can provide reasonable resolution of moderate to deep branches in the phylogeny of bacteria. The analysis presented here focuses on amino-acid comparisons for two reasons. First, for highly conserved proteins such as RecA, it is likely that amino-acid trees will be less biased by multiple substitutions at particular sites and base-composition variation between species than trees of the corresponding nucleotide sequences (Hasegawa and Hashimoto 1993; Viale et al. 1994, Lloyd and Sharp 1993). In addition, Lloyd and Sharp (1993) presented specific evidence suggesting that DNA-level comparisons of the *recA* genes between distantly related taxa might be misleading.

Alignment of RecA sequences

An alignment of the sequences of the complete RecA proteins is shown in Figure 1. Aligning sequences is an integral part of any molecular systematic study because each aligned position is assumed to include only homologous residues from the different molecules. Assignment of homology, as represented by the sequence alignments, can be highly controversial, and differences in alignments can cause significant differences in phylogenetic conclusions (Gatesy et al. 1993, Lake 1991). To limit such problems, regions for which homology of residues cannot be unambiguously assigned should be excluded from phylogenetic analysis. Thus for a molecule to be useful for molecular systematic studies, alignments between species should be relatively free of ambiguities. This is one of the main advantages of using SS-rRNA genes over other genes for phylogenetic analysis. Assignment of homology for SS-rRNA sequences can be aided by alignment of both primary and secondary structures (Woese 1987). In addition, regions of high primary structural conservation that are interspersed throughout the molecule help align less conserved regions. Since RecA is a highly conserved protein, it has the potential to be useful for phylogenetics because the assignment of homology should be relatively unambiguous (Lloyd and Sharp 1993). Regions of ambiguity in the RecA alignment shown in Fig. 1 were determined by comparing this alignment to those generated using different alignment parameters (see Methods). Regions of the alignment were considered to be ambiguous if they contained different residues in the different alignments, as suggested by Gatesy et al. (1993). Overall, the majority of the alignment was determined to be free of ambiguities and thus can be used with confidence for the phylogenetic analysis. The four regions of ambiguity (the C- and N-termini (corresponding to *E. coli* amino-acids 1-7 and 320-352) and two short regions corresponding to *E. coli* amino-acids 36-37 and 231-236)) were excluded from the phylogenetic analysis. The 313 alignment positions used are indicated by the sequence mask shown in Fig. 1.

Another potential source of variation and error in phylogenetic reconstruction from sequences lies in assigning a weight to give insertion or deletion differences (indels) between species. Other than in the C- and N-terminal regions, there are few indels in the RecA alignment (see Fig. 1). Most of the indels are in regions of ambiguous alignment as identified above, and thus were not included in the phylogenetic analysis. The phylogenetic results were not affected whether the few remaining indels were included or not (data not shown). Of the indels in regions of unambiguous alignment most are isolated (in only one species) and only one amino acid in length. There are two very large indels - one in each of the Mycobacterium RecAs. These are protein introns that are removed by post-translational processes (Davis et al. 1991, Davis et al. 1994). There is a 4 aa indel in the *Thermotoga maritima* RecA (see Fig. 1). There only indels that have obvious phylogenetic relevance are the single amino acid gaps found in the cyanobacterial and the *A. thaliana* RecAs all at the same position --*E. coli* position 53 (see below for discussion of this).

Another aspect of the RecA alignment that is relevant to molecular systematics is the degree of conservation of different alignment positions. I have used the RecA phylogeny and parsimony character-state analysis to characterize the patterns of amino-acid substitutions at different sites of the molecule (see Methods). The number of inferred substitutions varies a great deal across the molecule. The number of total substitutions ranges from 0 (at 58 positions) to 38 (at one position) with a mean of 9.4. The number of non-conservative substitutions varies from 0 (at 111 positions) to 27 (at one position) with a mean of 4.8. The variation in the substitution patterns across the molecule suggests that RecA comparisons may have phylogenetically useful information at multiple evolutionary distances.

Generation of phylogenetic trees

To examine the utility of the RecA comparisons for molecular systematics, the RecA trees were compared to trees of the same species based on studies of other molecules. Such a comparison is useful for a few reasons. First, congruence among trees of different molecules indicates both that the genomes of the species are not completely mosaic and that the molecular systematic techniques being used are reliable (Miyamoto and Fitch 1995). Differences in the branching patterns between trees of different molecules can help identify genetic mosaicism, unusual evolutionary processes, or inaccuracies in one or both of the trees. Differences in resolution and significance of particular branches can help identify which molecules are useful for specific types of phylogenetic comparisons. Since differences in species sampled have profound effects on tree generation (e.g., (Lecointre et al. 1993)), to best compare the phylogenetic resolution of trees of different molecules the analysis should include sequences from the same species. Fortunately, SS-rRNA sequences were available for most of the species represented in the RecA data set. Therefore it was possible to generate SS-rRNA trees for essentially the same species-set as represented in the RecA trees. For those species for which RecA sequences were available but SS-rRNA sequences were not, SS-rRNA sequences were used from close relatives (see Methods). A list of the sequences used is in Table 1.

Phylogenetic trees of the RecAs and SS-rRNAs were generated from the sequence alignments using multiple phylogenetic techniques (see Methods). The trees were generated without an outgroup and thus can be considered unrooted. However, since rooting of trees is helpful for a variety of reasons, a root was determined for both the RecA and SS-rRNA trees. In both cases, the root was determined to be the sequence from *Aquifex pyrophilus*. For the SS-rRNA trees, this rooting was chosen because analyses of sequences from all three kingdoms of organisms indicate that the deepest branching bacterial SS-rRNA is that of *A. pyrophilus* (Burggraf et al. 1992; Pitulle et al. 1994). Although it seems reasonable to assume that the deepest branching bacterial RecA would also be that of *A. pyrophilus*, if there have been lateral transfers or other unusual evolutionary processes, the RecA trees could be rooted differently than the SS-rRNA trees. Therefore the rooting of the RecA sequences was tested by constructing trees using likely RecA homologs from Archaea and eukaryotes as outgroups (see Methods). In both neighbor-joining and *protpars* trees, the deepest branching bacterial protein was that of *A. pyrophilus* (not shown). However, the alignments of the RecAs with the Archaeal and eukaryotic RecA-like proteins include many regions of ambiguity. Therefore, only 140 alignment positions were used in this analysis and the trees showed little resolution within the bacteria. In addition, the bootstrap values for the deep branching of the *A. pyrophilus* RecA were low (<30 in all cases). Thus although the rooting of the RecA trees to the *A. pyrophilus* protein is reasonable it should be considered tentative. The rooting will likely be better resolved as more sequences become available from eukaryotes and Archaea.

The analysis and comparison of the phylogenetic trees focused on a few specific areas. First, bootstrap values were used to get an estimate of the degree that the inferred branching patterns reflect the characteristics of the entire molecule. In addition, since phylogenetic methods differ in the range of evolutionary scenarios in which they accurately reconstruct phylogenetic relationships (Hillis 1995), comparison of the trees generated by the different methods was used to identify the phylogenetic patterns that were most robust for that particular molecule. To summarize the differences and similarities among the trees inferred by the different methods, strict-consensus trees of all the trees of each molecule were generated (Figure 2). Since consensus trees lose some of the information of single trees and since they only show the areas of agreement among trees (and not the phylogenetic patterns in the areas of difference), it is also useful to examine individual trees. A comparison of the Fitch-Margoliash trees for the two molecules is shown in Figure 3. The other trees are available from the author on request. Finally, the SS-rRNA trees determined here were compared to those determined with more sequences to help identify patterns that might be due to poor sampling of the species here.

A quick glance at the trees in Fig. 2 and 3 shows that the patterns for each molecule are highly robust (there is high resolution in the consensus trees) and that the patterns are similar between the two molecules. To aid comparison of the trees of the two molecules, sequences have been grouped into consensus clades based on the patterns found in the consensus trees (Fig. 2, Table 2). Clades of RecA sequences were chosen to represent previously characterized bacterial groups as well as possible. Comparable clades were determined for the SS-rRNA sequences (Table 2). The clades are named after the rRNA-based classification of most of the members of the clade (Maidak et al. 1994). These clades are highlighted in the trees in Fig. 2 and 3. Sequences from the same or similar species are aligned in the middle in Fig. 2 to ease comparison of the two consensus trees. Besides being found in trees generated by all the phylogenetic methods used, the consensus clades have high bootstrap values for the methods in which bootstrapping was performed (Table 2). Thus we believe that the clades are consistent and reliable groupings of the RecA and SS-rRNA sequences. In the following sections, some of the implications of the similarities and differences within and between the RecA and SS-rRNA trees are discussed. The discussion has been organized by phylogenetic groups.

Proteobacteria

The Proteobacteria phylum includes most but not all the traditional gram-negative bacterial species (Stackebrandt et al. 1988). This phylum has been divided into five phylogenetically distinct groups (α, β, γ, δ, and ε) mostly based on SS-rRNA comparisons (Olsen et al. 1994, Rainey et al. 1993, Stackebrandt et al. 1988, Woese 1987). The available RecA sequences are heavily biased towards the Proteobacteria (Table 1) and thus much of the discussion will focus on this phylum. With the species represented in this analysis, the Proteobacterial RecA sequences form a monophyletic clade in all phylogenetic methods (Fig. 2). In contrast, with essentially the same species-set, the Proteobacterial SS-rRNA sequences do not consistently form a clade (Fig. 2, positions of *Campylobacter jejuni*, *Helicobacter pylori*, and *Myxococcus xanthus*), although they do in some of the phylogenetic methods (e.g., Fig. 3). This was surprising since the Proteobacterial group was defined based on SS-rRNA comparisons (Stackebrandt et al. 1988). When additional SS-rRNA sequences are included in phylogenetic analysis, *M. xanthus*, *C. jejuni*, and *H. pylori* consistently branch with the other Proteobacteria (Maidak et al. 1994; Olsen et al. 1994). The lack of resolution of the position of these species in the SS-rRNA versus RecA trees was not due to using only one SS-rRNA sequence to represent the two *M. xanthus* RecAs -- the same pattern was seen when the SS-rRNA sequence from another δ species was also included. Thus in this case the RecA trees can be considered to have higher resolution than the SS-rRNA trees since the RecA trees show a relationship between species that is only consistently detected in SS-rRNA trees with more sequences.

Subdivisions corresponding to the α, β, γ, δ, and ε groups are detected in both the RecA and SS-rRNA trees and the placement of species into these subdivisions is nearly the same for the two molecules (Fig. 2, Table 2). Thus the RecA comparisons support the division of the Proteobacteria into these groups as well as the classification of particular species into the groups here. There are other phylogenetic patterns that are the same in the RecA and SS-rRNA trees here. Examples include the separation of the

Pseudomonas-Azotobacter γs (γ2 here) from the *Haemophilus*, *Proteus,* and enteric γs (γ1 here); the monophyly of the enteric bacteria (represented here by *E. coli*, *S. flexneri*, *Erwinia carotovara*, *Enterobacter agglomerans* and *Yersinia pestis*); the relatedness of the *Rhizobium* species, *Agrobacterium tumefaciens* and *Brucella abortus;* the placement of *Acinetobacter calcoaceticus* into the γ supergroup; an affiliation between the γ's and the β's into what can be called a β-γ supergroup; and the grouping of *Legionella pneumophilia*, *Neisseria gonorrhoeae*, *Xanthomonas oryzae*, and the *Thiobacillus* species somewhere in the γ-β supergroup. In all these cases, the relationships have been suggested by other studies of SS-rRNA sequences (see (Maidak et al. 1994; Olsen et al. 1994; Woese 1987)). The finding of the same patterns in the RecA trees serves to confirm the previous suggestions of the phylogenetic associations indicated between these species. Thus even though the RecA trees are based on analysis of highly conserved protein sequences, they do appear to have resolution for even close relatives as suggested by Lloyd and Sharp (1993).

Most of the differences between the RecA and SS-rRNA trees for the Proteobacteria are in areas of low resolution (differences among the trees generated by the different methods) or low bootstrap values for one or both of the molecules and thus are probably not biologically significant. For example, the differences in the grouping of the δ and ε clades within the Proteobacteria discussed above appears to be due to a lack of resolution of the SS-rRNA trees with the species represented here. In addition, the branching order between *Haemophilus influenzae*, the *Proteus* species, the *Vibrios*, and the enteric species is ambiguous in the SS-rRNA trees yet it is consistent in the RecA trees. In other cases, the SS-rRNA trees appear to have more resolution than the RecA trees. For example, the specific position of the RecA from *L. pneumophilia* is ambiguous (Fig. 2a) yet the SS-rRNA of this species consistently groups with the γ1 and γ2 groups, and thus can be considered part of the γ clade (Fig. 2b, Table 2). Analysis of other SS-rRNA sequences suggests that the position of the Legionellaceae in the γ subgroup is robust (Fry et al. 1991; Weisburg et al. 1989a). Similarly, the exact position of the *N. gonorrhoeae* RecA is ambiguous, yet the *N. gonorrhoeae* SS-rRNA groups consistently with the β clade.

There are branching patterns within the Proteobacteria that have high resolution and robustness for each molecule but are different between the two. The most striking example of this is the phylogenetic position of the sequences from *Acidiphilium facilis.* The *A. facilis* RecA branches with the *Thiobacillus ferrooxidans* RecA in the β–γ supergroup in all trees (Fig. 2) and the node joining these two species has very high bootstrap values (Table 2). However, the corresponding *A. facilis* SS-rRNA consistently branches with species in the α clade also with high bootstrap values. Thus either the SS-rRNA and RecA genes of *A. facilis* have different phylogenetic histories, or one of the trees is inaccurate. The grouping of *Acidiphilium* species within the α subgroup appears to be a reliable representation of the SS-rRNA relationships (Lane et al. 1992; Sievers et al. 1994), so it is unlikely that the SS-rRNA tree here is biased by species sampling. It has been suggested that the *A. facilis* RecA sequence contains many sequencing errors and it is currently being resequenced (Roca 1995). Errors in the sequence would explain the unusual amino acids found in the *A. facilis* RecA in otherwise highly conserved regions (Fig. 1) and the extremely long branch length for this sequence in all phylogenetic methods (Fig. 3). Thus the position of the *A. facilis* RecA in the trees may not represent the actual evolutionary history of this gene.

M. xanthus, the only δ Proteobacteria represented in this analysis, is the only species known to encode two RecA proteins. There are at least two plausible explanations for this: lateral transfer from another species or gene duplication. The phylogenetic analysis of the two proteins helps limit the possibilities for when and how a duplication or lateral transfer could have occurred. In all the RecA trees, the two *M. xanthus* proteins branch together, showing that they are more related to each other than to any other known RecAs. However, the node joining them is quite deep indicating that the degree of evolutionary separation between them is quite high. Thus if a duplication event was what led to these two genes in the same species, it apparently happened reasonably early in the history of the δ clade. If one of these sequences was obtained by a lateral transfer from another species, most likely, the donor was another δ species. It is interesting that the bootstrap values for the node joining the two *M. xanthus* RecAs are relatively low in all methods (Table 2). This indicates that the branching together is not very stable and is affected by the choice of alignment positions used in the phylogenetic analysis. Perhaps there was a gene conversion event after a lateral transfer or duplication and only certain regions of the *recA* genes underwent the conversion. Alternatively, the low bootstraps could also be explained if a duplication occurred right at or near the time of separation of the δ clade from the other Proteobacterial groups. The specific history of these two genes will probably be best resolved by studies of RecAs in other δ species.

Gram-positive bacteria

Previous studies have shown that gram-positive species are divided into multiple phylogenetically distinct groups (Woese 1987). Whether these distinct groups are monophyletic has been the subject of a great deal of research and debate (e.g., (Gupta et al. 1994; Van De Peer et al. 1994; Weisburg et al. 1989c; Woese 1987)). For example, studies of HSP70 genes (Viale et al. 1994) and some studies of rRNA genes (Woese 1987) suggest the gram-positives are monophyletic while studies of EF-TUs (Ludwig et al. 1994), ATPaseβ (Ludwig et al. 1994) and different studies of rRNA genes (Van De Peer et al. 1994) suggest they are not.

Species from two of the gram-positive groups, the low-GCs and the high-GCs, are represented in the analysis here (Table 1). In all the RecA and SS-rRNA trees inferred in this study, the sequences from the high-GC species cluster together (Fig. 2). In addition these species have the same branching patterns within this group in all trees of both molecules. Thus the RecA data support the phylogenetic coherence of as well as the branching topology within the high-GC clade. In contrast, the RecA and SS-rRNA trees are not congruent for the relationships among sequences from low-GC gram-positive species. In all the SS-rRNA trees, the sequences from species considered to be low-GC gram-positives are monophyletic, as might be expected, since the classification of these species was based on SS-rRNA comparisons. However in all the RecA trees the sequences from the low-GCs are not monophyletic (e.g., Fig. 3). This may be due to a combination of poor species sampling and unusual evolutionary patterns. In four of the five RecA trees only one RecA, that of the spirochaete *Borrelia burgdorferi*, prevents the low-GCs as a whole from being monophyletic (e.g., Fig. 3). The bootstrap values for the position of the *B. burgdorferi* RecA are relatively low in all of these trees, and since this is the only sample from the spirochaetes, its position may be unreliable. In addition, in three out of four of the SS-rRNA trees, the *B. burgdorferi* sequence is an outgroup to the low-GCs. Thus with the species sampled here the *B. burgdorferi* sequences tend to group with the sequences from low-GCs. Yet another factor that could contribute to a biased placement of the *B. burgdorferi* RecA is the apparent high rate of sequence change in the mycoplasmal RecAs, which can be seen by their long branch lengths (Fig. 3a). A rapid rate of mycoplasmal protein evolution has been thought to complicate trees of other proteins (e.g., (Ludwig et al. 1994)). The inclusion of additional sequences from the spirochetes and other low-GC gram-positives may help resolve whether this difference between the RecA and SS-rRNA trees is biologically significant.

With the species represented here, the branching between the high and low-GCs is unresolved in both the RecA and SS-rRNA trees. Interestingly, in all the RecA trees, the proteins from the high-GCs form a group with the cyanobacterial proteins. Thus the gram-positives are non-monophyletic for RecA proteins. Analysis of other genes has suggested that the cyanobacteria and gram-positives are sister groups (e.g., (Van De Peer et al. 1994; Viale et al. 1994; Woese 1987)). However this is one of the few if not the only case in which the cyanobacterial genes consistently group with genes from high-GCs to the exclusion of those from the low-GCs. Since this relationship is found in all the RecA trees it appears to be robust. However, the bootstrap values for the node linking these two groups are moderate (31-40) indicating that this association is a good, but not great, representation of the relationships of RecA sequences.

Cyanobacteria

The RecA and SS-rRNA trees both show the cyanobacteria forming a coherent clade. The nuclear encoded chloroplast RecA from *A. thaliana* groups consistently with the cyanobacterial RecAs. This suggests that the *A. thaliana recA* gene is derived from the *recA* gene of a cyanobacterial-like ancestor to the *A. thaliana* chloroplast and that, as has been demonstrated for many other genes, it was transferred to the nucleus after endosymbiosis. Given the high degree of sequence conservation in RecAs, it is possible that studies of chloroplast evolution might be aided by sequencing of additional nuclear encoded chloroplast RecAs. In addition, all the RecAs from this group (including the *A. thaliana* RecA) contain an alignment gap not found in any other RecAs (see Fig. 1). This could serve as a sequence signature for cyanobacterial and chloroplast RecAs and further serves to demonstrate the relatedness among chloroplasts and cyanobacteria. As discussed above, the cyanobacterial RecAs group with those of the high-GC gram-positives in all trees.

Deinococcus/Thermus group

The RecAs of *Deinococcus radiodurans* and the two *Thermus* species form a clade with high bootstrap values in all the trees (see Table 2, Fig. 2). Analysis of other data suggests that these species are part of a clade (Ludwig et al. 1994; Weisburg et al. 1989b). However, these sequences do not consistently form a clade in the SS-rRNA trees here (they form a clade only in the *dnapars* tree (not shown)). Inclusion of additional SS-rRNA sequences allows for better resolution of this clade, probably because of GC content variation among the species (Embley et al. 1993). Thus with the species used here, the RecA trees show resolution of the *Deinococcus-Thermus* group while the SS-rRNA trees do not. This may be due to less of a GC bias in the RecA sequences this in the SS-rRNA sequences, as suggested by Lloyd and Sharp (1993). The RecA analysis also supports previous assertions that this group is one of the deeper branching bacterial phyla (Weisburg et al. 1989b), and shows that RecA has resolution even for deep branches. However, this conclusion relies on the rooting of the RecA tree to the *A. pyrophilus* sequence which has low support (see above).

Other taxa

There is little resolution in the RecA trees regarding the position of the *Thermotoga maritima*, *Chlamydia trachomatis*, and *Bacteroides fragilis* proteins. These RecA proteins do not show consistent affiliations with any individual sequences or groups (Fig. 2, Fig. 3) and the bootstrap values for their positions in the individual trees are low (Fig. 3). I believe that this is due to these sequences being the only representatives from large phylogenetic groups (Thermotogales, Chlamydia, and Bacteroides, respectively). Using the same sets of sequences as in the RecA trees, the SS-rRNA trees show a similar lack of resolution for sequences that are individual representatives of large groups (in this case, *C. trachomatis, B. fragilis,* and *Borrelia burgdorferi)*. It would be useful to have more RecA genes from these phylogenetic groups to better determine if the RecA and SS-rRNA based trees are congruent for these bacterial groups. It is interesting that although the specific positions of the *T. maritima* RecA is ambiguous, it never branches below the *Deinococcus-Thermus* sequences as the *T. maritima* SS-rRNA does in all the SS-rRNA trees. Thus even if the rooting of the RecA tree with *A. pyrophilus* is incorrect, the *A. pyrophilus* and *T. maritima* RecAs never branch immediately near each other as they do in the SS-rRNA trees. Since the RecA tree appears to be less biased by GC content variation (as suggested by Lloyd and Sharp (1993)) than SS-rRNA analysis, it seems plausible that the close branching of the *T. maritima* and *A. pyrophilus* SS-rRNAs may be caused by GC content convergence.

Conclusions

Comparison of phylogenetic results for particular taxa using different genes can help determine what genes are useful for evolutionary studies as well as whether different genes have different histories (as could be caused by lateral transfers). However, in order to make direct comparisons it is important to remove as many variables in the studies of the different genes. For example, many researchers studying bacterial systematics compare phylogenetic trees of particular genes to standard trees of SS-rRNA sequences. Yet when these trees have differences with the SS-rRNA trees it is not always clear whether the differences are due to use of different techniques (SS-rRNA trees tend to be constructed with maximum likelihood methods while such methods are still difficult to apply to large numbers of protein sequences), the inclusion of different sets of species (there are some 3000 SS-rRNA sequences that can be used), or true differences in branching or resolution power of different molecules. In the analysis presented here I have compared phylogenetic trees of RecA and SS-rRNA sequences using similar techniques from essentially the same sets of species. Overall, the branching patterns and powers of resolution of the two molecules are highly similar. The similar branching patterns lend support to the general pattern of bacterial systematics inferred from SS-rRNA sequences. This indicates either that the potential problems with SS-rRNA trees have little effect on phylogenetic results or that the RecA trees are biased in the same ways by these problems. In some cases, the RecA trees have resolution where the SS-rRNA trees do not (e.g., for the monophyly of the Proteobacteria and the grouping of *D.*

radiodurans and the *Thermus* species) and in other cases the reverse is true -- the SS-rRNA trees have resolution (e.g., the position of *T. maritima*; the placement of *L. pneumophilia* within the γ-Proteobacteria and the monophyly of the low-GC gram-positives). The lack of resolution of some of the deep branches in the RecA trees is likely related to the species sampled -- a similar lack of resolution is seen in SS-rRNA trees when using the same species set. Therefore RecA appears to be as good a model for studies of molecular systematics of bacteria as SS-rRNA. It remains to be seem whether some of the unusual patterns in the RecA trees (such as the grouping of the cyanobacteria with the high-GC gram-positives and the branching of *T. maritima* above the Deinococci-Thermus group) are supported by future studies.

In conclusion I would like to emphasize some of the features of RecA that make it a good choice for molecular systematic studies. Among protein encoding genes RecA is relatively easy to clone from new species -- either by degenerate PCR *(*e.g., (Duwat et al. 1992a, Duwat et al. 1992b, Dybvig et al. 1992, Dybvig and Woodard 1992, Quivey and Faustoferri 1992)) or functional complementation of the radiation sensitivity of *recA* mutants from other species (Calero et al. 1994, De Mot et al. 1993, Favre et al. 1991, Gomelsky et al. 1990, Tatum et al. 1993). RecA protein function appears to be conserved in all bacteria and there are similar proteins in eukaryotes and Archaea (Clark and Sandler 1994), although whether these can be used reliably for phylogenetic analysis of all three kingdoms remains to be seen. Like with SS-rRNAs, some regions of RecA are virtually completely conserved between species and other regions are variable even between close relatives. This allows for resolution of relationships among both close and distant relatives. The high conservation of size and sequence among RecAs makes alignments virtually unambiguous, limiting complications due to incorrect assignment of homology. In addition since RecA sequences can be compared at the protein and the DNA level it may be possible to limit problems due to nucleotide composition convergence between species. However, perhaps most importantly, I have shown here that phylogenetic trees of RecA sequences have similar topologies and similar resolution to trees of SS-rRNA sequences from the same species. This not only demonstrates that the genomes of these species are not completely mosaic (these two genes have similar phylogenies) but also that molecular systematics of bacteria is reliable and that RecA comparisons are useful for such molecular systematic studies.

Finally, I would like to suggest two additional reasons why researchers might want to choose RecA for molecular systematic studies. First, the cloning and sequencing of *recA* genes from new species facilitates the creation of *recA* mutants which are useful to have for laboratory studies of bacterial species. Also, with the availability of the crystal structure of the *E. coli* protein and with information about the phenotypes of 100s of *recA* mutants, I believe RecA can become a model for studies of protein evolution.

ACKNOWLEDGEMENTS

I would like to thank P. C. Hanawalt for support and encouragement; M. B. Eisen for help with computer analysis; S. Smith and J. Felsenstein for making their computer programs freely available; M. Huang, J. Coleman, A. J. Clark, S. Sandler, W. Finch, and S. Mongkolsuk for making sequences available prior to publication; M. Feldman for the use of computer equipment; D. Pollock, M. B. Eisen, M-I. Benito, H. Hamilton, D. Distel, and J. D. Palmer for helpful discussions and suggestions. I also want to thank A. I. Roca for help in all aspects of this project including and especially the identification of errors in *recA* sequences and aid in getting access to unpublished sequences. During the course of this research I have received support from a pre-doctoral fellowship from the N. S. F., an N. I. H. grant to P. C. Hanawalt, and a tuition grant from the Woods Hole Institution to attend the Summer Workshop on Molecular Evolution, 1994.

REFERENCES

Aigle B, Schneider D and Decaris B (1994) Genbank entry Z30324

Akaboshi E, Yip ML and Howard-Flanders P (1989) Nucleotide sequence of the *recA* gene of *Proteus mirabilis*. Nucleic Acids Res 17: 4390

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410

Angov E and Camerini-Otero RD (1994) The *recA* gene from the thermophile *Thermus* *aquaticus* YT-1: cloning, expression, and characterization. J Bacteriol 176: 1405-1412

Ball TK, Wasmuth CR, Braunagel SC and Benedik MJ (1990) Expression of *Serratia marcescens* extracellular proteins requires *recA*. J Bacteriol 172: 342-349

Bayles KW, Brunskill EW, Iandolo JJ, Hruska LL, Huang S, Pattee PA, Smiley BK and Yasbin RE (1994) A genetic and molecular characterization of the *recA* gene from *Staphylococcus aureus*. Gene 147: 13-20

Berson AE, Peters MR and Waleh NS (1990) Nucleotide sequence of *recA* gene of *Aquaspirillum magnetotacticum*. Nucleic Acids Res 18: 675

Billman-Jacobe H (1994) Genbank entry X77384

Binet M-N, Osman M and Jagendorf AT (1993) Genomic nucleotide sequence of a gene from *Arabidopsis thaliana* encoding a protein homolog of *Escherichia coli* RecA. Plant Physiol 103: 673-674

Boorstein WR, Ziegelhoffer T and Craig EA (1994) Molecular evolution of the HSP70 multigene family. J Mol Evol 38: 1-17

Bruns TD and Szaro TM (1992) Rate and mode differences between nuclear and mitochondrial small-subunit rRNA genes in mushrooms. Mol Biol Evol 9: 836-855

Burggraf S, Olsen GJ, Stetter KO and Woese CR (1992) A phylogenetic analysis of *Aquifex pyrophilus*. Syst Appl Microbiol 15: 352-356

Calero S, Fernandez de Henestrosa AR and Barbe J (1994) Molecular cloning, sequence and regulation of expression of the *recA* gene of the phototrophic bacterium *Rhodobacter sphaeroides*. Mol Gen Genet 242: 116-120

Cerutti H, Osman M, Grandoni P and Jagendorf AT (1992) A homolog of *Escherichia coli* RecA protein in plastids of higher plants. Proc Natl Acad Sci USA 89: 8068-8072

Clark AJ, and Margulies AD (1965) Isolation and characterization of recombinant-deficient mutants of *Escherichia coli*. Proc Natl Acad Sci USA 53: 451-459

Clark AJ and Sandler SJ (1994) Homologous genetic recombination: the pieces begin to fall into place. Crit Rev Microbiol 20: 125-142

Clark AJ (1995) Personal communication

Coleman J (1995) Personal communication

Davis EO, Sedgwick SG and Colston MJ (1991) Novel structure of the *recA* locus of *Mycobacterium tuberculosis* implies processing of the gene product. J Bacteriol 173: 5653-62

Davis EO, Thangaraj HS, Brooks PC and Colston MJ (1994) Evidence of selection for protein introns in the *recA*s of pathogenic mycobacteria. EMBO J 13: 699-703

De Mot R, Laeremans T, Schoofs G and Vanderleyden J (1993) Characterization of the *recA* gene from *Pseudomonas fluorescens* OE 28.3 and construction of a *recA* mutant. J Gen Microbiol 139: 49-57

De Soete G (1983) A least squares algorithm for fitting additive trees to proximity data. Psychometrika 48: 621-626

Delwiche, CF, Kuhsel, M and Palmer, JD (1995) Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. Mol Phylogen Evol 4: 110-128.

Dunkin SM and Wood DO (1994) Isolation and characterization of the *Rickettsia prowazekii recA* gene. J Bacteriol 176: 1777-1781

Duwat P, Ehrlich SD and Gruss A (1992a) A general method for cloning *recA* genes of gram-positive bacteria by polymerase chain reaction. J Bacteriol 174: 5171-5175

Duwat P, Ehrlich SD and Gruss A (1992b) Use of degenerate primers for polymerase chain reaction cloning and sequencing of the *Lactococcus lactis* subsp. lactis *recA* gene. Appl Environ Microbiol 58: 2674-2678

Dybvig K, Hollingshead SK, Heath DG, Clewell DB, Sun F and Woodard A (1992) Degenerate oligonucleotide primers for enzymatic amplification of *recA* sequences from gram-positive bacteria and mycoplasmas. J Bacteriol 174: 2729-2732

Dybvig K and Woodard A (1992) Cloning and DNA sequence of a mycoplasmal *recA* gene. J Bacteriol 174: 778-784

Eisen JA, Smith SW and Cavanaugh CM (1992) Phylogenetic relationships of chemoautotrophic bacterial symbionts of *Solemya velum* Say (Mollusca: Bivalvia) determined by 16S rRNA sequence analysis. J Bacteriol 174: 3416-3421

Embley TM, Thomas RH and Williams RAD (1993) Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. Syst Appl Microbiol 16: 25-29

Emmerson PT (1995) Personal communication

Favre D, Cryz SJ Jr. and Viret JF (1991) Cloning of the *recA* gene of *Bordetella pertussis* and characterization of its product. Biochimie 73: 235-44

Favre D and Viret JF (1990) Nucleotide sequence of the *recA* gene of *Bordetella pertussis*. Nucleic Acids Res 18: 4243

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evol 39: 783-791.

Felsenstein J (1993) PHYLIP version 3.5c. University of Washington, Seattle, WA

Fernandez de Henestrosa AR (1994) Genbank entry X82183

Finch WM (1995) Personal communication

Fitch WM and Margoliash E (1967) Construction of phylogenetic trees. Science 155: 279-284

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Leuhrsen KH, Chen KN and Woese CR (1980) The phylogeny of prokaryotes. Science 209: 457-463

Fry NK, Warwick S, Saunders NA and Embley TM (1991) The use of 16S ribosomal RNA analyses to investigate the phylogeny of the family Legionellaceae. J Gen Microbiol 137: 1215-1222

Fyfe JA and Davies JK (1990) Nucleotide sequence and expression in *Escherichia coli* of the *recA* gene of *Neisseria gonorrhoeae*. Gene 93: 151-156

Gammie AE and Crosa JH (1991) Co-operative autoregulation of a replication protein gene. Mol Microbiol 5: 3015-3023

Gatesy J, Desalle R and Wheeler W (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol Phylog Evol 2: 152-157

Gomelsky M, Gak E, Chistoserdov A, Bolotin A and Tsygankov YD (1990) Cloning, sequence and expression in *Escherichia coli* of the *Methylobacillus flagellatum recA* gene. Gene 94: 69-75

Goodman HJ and Woods DR (1990) Molecular analysis of the *Bacteroides fragilis recA* gene. Gene 94: 77-82

Gregg-Jolly LA and Ornston LN (1994) Genbank entry L26100

Guerry P, Pope PM, Burr DH, Leifer J, Joseph SW and Bourgeois AL (1994) Development and characterization of *recA* mutants of *Campylobacter jejuni* for inclusion in attenuated vaccines. Infect Immun 62: 426-432

Gupta RS, Golding GB and Singh B (1994) Hsp70 phylogeny and the relationship between archaebacteria, eubacteria, and eukaryotes. J Mol Evol 39: 537-540

Gutell RR, Larsen N and Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol Rev 58: 10-26

Gutman PD, Carroll JD, Masters CI and Minton KW (1994) Sequencing, targeted mutagenesis and expression of a *recA* gene required for the extreme radioresistance of *Deinococcus radiodurans*. Gene 141: 31-37

Haas R (1994) Genbank entry Z35478

Hasegawa M and Hashimoto T (1993) Ribosomal RNA tree misleading? Nature 361: 23

Henikoff S (1993) Sequence analysis by electronic mail server. Trends Biochem Sci 18: 267-268

Hillis DM (1995) Approaches for assessing phylogenetic accuracy. Syst Biol 44: 3-16

Horii T, Ogawa T and Ogawa H (1980) Organization of the *recA* gene of *Escherichia coli*. Proc Natl Acad Sci USA 77: 313-317

Huang WM (1995) Personal communication

Inagaki K, Tomono J, Kishimoto N, Tano T and Tanaka H (1993) Cloning and sequence of the *recA* gene of *Acidiphilium facilis*. Nucleic Acids Res 21: 4149

Inouye M (1995) Personal communication

Jinks-Robertson S and Nomura M (1987) Ribosomes and tRNA, In: F.C. Neidhardt (Ed.), *Escherichia coli* and *Salmonella typhimurium* cellular and molecular biology. American Society for Microbiology, Washington, D.C., pp. 1358-1385

Kato R and Kuramitsu S (1993) RecA protein from an extremely thermophilic bacterium, *Thermus thermophilus* HB8. J Biochem 114: 926-929

Kerins SM, Fitzpatrick R, O'Donohue M and Dunican L (1994) Genbank entry X75085.

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120

King KW, Woodard A and Dybvig K (1994) Cloning and characterization of the *recA* genes from *Mycoplasma pulmonis* and *M. mycoides* subsp. mycoides. Gene 139: 111-115

Klenk H-P and Zillig W (1994) DNA-dependent RNA polymerase subunit b as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. J Mol Evol

38: 420-432

Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SS and Rehrauer WM (1994) Biochemistry of homologous recombination in *Escherichia coli*. Microbiol Rev 58: 401-465

Kowalczykowski SC (1991) Biochemical and biological function of *Escherichia coli* RecA protein: behavior of mutant RecA proteins. Biochimie 73: 289-304

Kryukov VM, Suchkov IY, Sazykin IS and Mishankin BN (1993) Genbank entry X75336.

Lake JA (1991) The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol 8: 378-385

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML and Pace NR (1985) Rapid determination of 16S rRNA sequences for phylogenetic analysis. Proc Natl Acad Sci USA 82: 6955-6959

Lane DJ, Harrison AP, Stahl DA, Pace B, Giovannoni, SJ, Olsen GJ and Pace NR (1992) Evolutionary relationships among sulfur- and iron-oxidizing eubacteria. J Bacteriol 174: 269-278

Larsen SH (1994) Genbank entry U16739

Lecointre G, Philippe H, Van Le HL and Le Guyader H (1993) Species sampling has a major impact on phylogenetic inference. Mol Phylog Evol 2: 205-224

Lloyd AT and Sharp PM (1993) Evolution of the *recA* gene and the molecular phylogeny of bacteria. J Mol Evol 37: 399-407

Ludwig W, Kirchhof G, Klugbauer N, Weizenegger M, Betzl D, Ehrmann M, Hertel C, Jilg S, Tatzel R, Zitzelsberger H, Liebl S, Hochberger M, Shah J, Lane D, Wallnöfer PR and Shleifer KH (1992) Complete 23S ribosomal RNA sequences of gram-positive bacteria with a low DNA G plus C content. Syst Appl Microbiol 15: 487-501

Ludwig W, Neumaier J, Klugbauer N, Brockmann E, Roller C, Jilg S, Reetz K, Schachtner I, Ludvigsen A, Bachleitner M, Fischer U and Schleifer KH (1994) Phylogenetic relationships of bacteria based on comparative sequence analysis of elongation factor TU and ATP-synthase beta-subunit genes. Antonie Van Leeuwenhoek 64: 285-305

Luo J, Burns G and Sokatch JR (1993) Construction of chromosomal *recA* mutants of *Pseudomonas putida* PpG2. Gene 136: 263-266

Maddison WP and Maddison DR (1992) MacClade Version 3. Sinauer Associates, Inc., Sunderland, MA

Maidak BL, Larsen N, McCaughey MJ, Overbeek R, Olsen GJ, Fogel K, Blandy J and Woese CR (1994) The ribosomal database project. Nucleic Acids Res 22: 3485-3487

Margraf RL, Roca AI and Cox MM (1995) The deduced *Vibrio cholerae* RecA amino acid sequence. Gene 152: 135-136

Martin B, Ruellan JM, Angulo JF, Devoret R and Claverys JP (1992) Identification of the *recA* gene of *Streptococcus pneumoniae*. Nucleic Acids Res 20: 6412

Medlin L, Elwood HJ, Stickel S and Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like ribosomal RNA-coding regions. Gene 71: 491-500

Michiels J, Vande Broek A and Vanderleyden J (1991) Molecular cloning and nucleotide sequence of the *Rhizobium phaseoli recA* gene. Mol Gen Genet 228: 486-490

Miyamoto MM and Fitch WM (1995) Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 44: 64-76

Mongkolsuk S (1995) Personal communication

Murphy RC, Bryant DA, Porter RD and de Marsac NT (1987) Molecular cloning and characterization of the *recA* gene from the cyanobacterium *Synechococcus sp.* strain PCC 7002. J Bacteriol 169: 2739-2747

Murphy RC, Gasparich GE, Bryant DA and Porter RD (1990) Nucleotide sequence and further characterization of the *Synechococcus sp.* strain PCC 7002 *recA* gene: complementation of a cyanobacterial *recA* mutation by the *Escherichia coli recA* gene. J Bacteriol 172: 967-976

Nakazawa T, Kimoto M and Abe M (1990) Cloning, sequencing, and transcriptional analysis of the *recA* gene of *Pseudomonas cepacia*. Gene 94: 83-88

Nickrent DL and Starr EM (1994) High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. J Mol Evol 39: 62-70

Nomura M, Morgan EA and Jaskunas S (1977) Genetics of bacterial ribosomes. Ann Rev Genet 11: 297-347

Nussbaumer B and Wohlleben W (1994) Identification, isolation and sequencing of the *recA* gene of *Streptomyces lividans* TK24. FEMS Microbiol Lett 118: 57-63

Ogawa T, Yu X, Shinohara A and Egelman EH (1993) Similarity of the yeast RAD51 filament to the bacterial RecA filament. Science 259: 1896-1899

Olsen GJ (1988) Phylogenetic analysis using ribosomal RNA. Meth Enzymol 164: 793-812

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR and Stahl DA (1986) Microbial ecology and evolution: a rRNA approach. Ann Rev Microbiol 40: 337-365

Olsen GJ, Woese CR and Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176: 1-6

Owttrim GW and Coleman JR (1989) Regulation of expression and nucleotide sequence of the *Anabaena variabilis recA* gene. J Bacteriol 171: 5713-5719

Pace NR, Olsen GJ and Woese CR (1986) Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell 45: 325-326

Pitulle C, Yang Y, Marchiani M, Moore ERB, Siefert JL, Aragno M, Jurtshuk PJ and Fox GE (1994) Phylogenetic position of the genus *Hydrogenobacter*. Int J Syst Bacteriol 44: 620-626

Quivey RG, Jr. and Faustoferri RC (1992) In vivo inactivation of the *Streptococcus mutans recA* gene mediated by PCR amplification and cloning of a *recA* DNA fragment. Gene 116: 35-42

Rainey FA, Toalster, R and Stackebrandt E (1993) *Desulfurella acetivorans,* a thermophilic, acetate-oxidizing and sulfur-reducing organism, represents a distinct lineage within the Proteobacteria. Syst Appl Microbiol 16: 373-379.

Ramesar RS, Abratt V, Woods DR and Rawlings DE (1989) Nucleotide sequence and expression of a cloned *Thiobacillus ferrooxidans recA* gene in *Escherichia coli*. Gene 78: 1-8

Rappold CSJ and Klingmueller W (1993) Genbank entry P33037

Rensing SA and Maier UG (1994) Phylogenetic analysis of the stress-70 protein family. J Mol Evol 39: 80-86

Ridder R, Marquardt R and Esser K (1991) Molecular cloning and characterization of the *recA* gene of *Methylomonas clara* and construction of *recA* deficient mutant. Appl Microbiol Biotechnol 35: 23-31

Roca AI (1995) Personal communication.

Roca AI and Cox MM (1990) The RecA protein: structure and function. Crit Rev Biochem Mol Biol 25: 415-456

Rothschild LJ, Ragan MA, Coleman AW, Heywood P and Gerbi SA (1986) Are rRNA sequences the Rosetta stone of phylogenetics. Cell 47: 640

Saitou N and Nei M (1987) The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Sancar A, Stachelek C, Konigsberg W and Rupp WD (1980) Sequences of the *recA* gene and protein. Proc Natl Acad Sci USA 77: 2611-2615

Sano Y and Kageyama M (1987) The sequence and function of the *recA* gene and its protein in *Pseudomonas aeruginosa* PAO. Mol Gen Genet 208: 412-419

Schoeniger M and Von Haeseler A (1994) A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylog Evol 3: 240-247

Selbitschka W, Arnold W, Priefer UB, Rottschafer T, Schmidt M, Simon R and Puhler A (1991) Characterization of *recA* genes and *recA* mutants of *Rhizobium meliloti* and *Rhizobium leguminosarum* biovar viciae. Mol Gen Genet 229: 86-95

Sievers M, Ludwig W and Teuber M (1994) Phylogenetic positioning of *Acetobacter*, *Gluconobacter*, *Rhodopila* and *Acidiphilium* species as a branch of acidophilic bacteria in the alpha-subclass of proteobacteria based on 16S ribosomal DNA sequences. Syst Appl Microbiol 17: 189-196

Smith SW (1994) Genetic Data Environment. Version 2.2a. Harvard Genome Laboratory, Cambridge, MA

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment: an expandable GUI for multiple sequence analysis. CABIOS 10: 671-675

Sogin ML (1989) Evolution of eukaryotic microorganisms and their small subunit ribosomal RNA. Amer Zool 29: 487-500

Stackebrandt E, Murray RGE and Trüper HG (1988) *Proteobacteria* classis nov., a name for the phylogenetic taxon that includes 'purple bacteria and their relatives'. Int J Syst Bacteriol 38: 321-325

Story RM, Bishop DK, Kleckner N and Steitz TA (1993) Structural relationship of bacterial RecA proteins to recombination proteins from bacteriophage T4 and yeast. Science 259: 1892-1896

Story RM and Steitz TA (1992) Structure of the RecA protein-ADP complex. Nature

355: 374-376

Story RM, Weber IT and Steitz TA (1992) The structure of the *E. coli* RecA protein monomer and polymer. Nature 355: 318-325

Stranathan MC, Bayles KW and Yasbin RE (1990) The nucleotide sequence of the *recE*+ gene of *Bacillus subtilis*. Nucleic Acids Res 18: 4249

Stroeher UH, Lech AJ and Manning PA (1994) Gene sequence of *recA*+ and construction of *recA* mutants of *Vibrio cholerae*. Mol Gen Genet 244: 295-302

Swofford D (1991) Phylogenetic Analysis Using Parsimony (PAUP) Version 3.0d. Illinois Natural History Survey, Champaign, Ill.

Tatum FM, Morfitt DC and Halling SM (1993) Construction of a *Brucella abortus recA* mutant and its survival in mice. Microb Pathog 14: 177-185

Tayama K, Fukaya M, Takemura H, Okumura H, Kawamura Y, Horinouchi S and Beppu T (1993) Cloning and sequencing the *recA*+ genes of *Acetobacter polyoxogenes* and *Acetobacter aceti*: construction of *recA*- mutants of by transformation- mediated gene replacement. Gene 127: 47-52

Thompson JD, Higgins DG and Gibson TJ (1994) Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680

Tolmasky ME, Gammie AE and Crosa JH (1992) Characterization of the *recA* gene of *Vibrio anguillarum*. Gene 110: 41-48

Van De Peer Y, Neefs JM, De Rijk P, De Vos P and De Wachter R (1994) About the order of divergence of the major bacterial taxa during evolution. Syst Appl Microbiol 17: 32-38

Vawter L and Brown WM (1993) Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics 134: 597-608

Venkatesh TV and Das HK (1992) The *Azotobacter vinelandii recA* gene: sequence analysis and regulation of expression. Gene 113: 47-53

Viale AM, Arakaki AK, Soncini FC and Ferreyra RG (1994) Evolutionary relationships among eubacterial groups as inferred from GroEL (chaperonin) sequence comparisons. Int J Syst Bacteriol 44: 527-533

Wardhan H, McPherson MJ, Harris CA, Sharma E and Sastry GR (1992) Molecular analysis of the *recA* gene of *Agrobacterium tumefaciens* C58. Gene 121: 133-6

Weisburg WG, Barns SM, Pelletier DA and Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173: 697-703

Weisburg WG, Dobson ME, Samuel JE, Dasch GA, Mallavia LP, Baca O, Mendelco L, Sechrest JE, Weiss E and Woese CR (1989a) Phylogenetic diversity of the Rickettsiae. J Bacteriol 171: 4202-4206

Weisburg WG, Giovannoni SG and Woese CR (1989b) The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. Syst Appl Microbiol: 128-134

Weisburg WG, Tully JG, Rose DL, Petzel JP, Oyaizu H, Yang D, Mandelco L, Sechrest J, Lawrence TG (1989c) A phylogenetic analysis of the mycoplasmas: basis for their classification. J Bacteriol 171: 6455-6467

Wetmur JG, Wong DM, Ortiz B, Tong J, Reichert F and Gelfand DH (1994) Cloning, sequencing, and expression of RecA proteins from three distantly related thermophilic eubacteria. J Biol Chem 269: 25928-25935

Woese CR (1991) The use of ribosomal RNA in reconstructing relationships among bacteria, In: Selander RK, Clark AG and Whittam TS (eds.), Evolution at the molecular level. Sinauer Associates, Inc., Sunderland, MA, pp. 1-24

Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271

Woese CR, Achenbach L, Rouviere P and Mandelco L (1991) Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in the light of certain coposition-induced artifacts. Syst Appl Microbiol 14: 364-371

Wolfe KH, Katz-Downie DS, Morden CW and Palmer JD (1992) Evolution of the plastid ribosomal RNA operon in a nongreen parasitic plant: accelerated sequence evolution, altered promoter structure, and tRNA pseudogenes. Plant Mol Biol 18: 1037-1048

Yao W and Vining LC (1994) Cloning and sequence analysis of a *recA*-like gene from *Streptomyces venezuelae* ISP5230. FEMS Microbiol Lett 118: 51-56

Zhang D, Fan H, McClarty G and Brunham RC (1994) Genbank entry U15281

Zhao X and Dreyfus LA (1990) Expression and nucleotide sequence analysis of the *Legionella pneumophila recA* gene. FEMS Microbial Lett. 70: 227-232

Zhao XJ and McEntee K (1990) DNA sequence analysis of the *recA* genes from *Proteus vulgaris, Erwinia carotovora, Shigella flexneri* and *Escherichia coli* B/r. Mol Gen Genet 222: 369-376

Zulty JJ and Barcak GJ (1993) Structural organization, nucleotide sequence, and regulation of the *Haemophilus influenzae rec-1+* gene. J Bacteriol 175: 7269-7281

Table 1. RecA and SS-rRNA sequences.

Species (by Phylum)	Abbr.	RecA.	#aa	SS-rRNA[1], [2]	RecA Refs.
Proteobacteria					
Acetobacter polyoxogenes	*Act.po*	D13183	348	ABA.PASTER*	(Tayama et al. 1993)
Acidiphilium facilis	*Acd.f*	D16538	354	ACDP.FACI2	(Inagaki et al. 1993)
Acinetobacter calcoaceticus	*Acn.c*	L26100	349	ACN.CALCOA	(Gregg-Jolly and Ornston 1994)
Agrobacterium tumefaciens	*Ag.t*	L07902	363	AG.TUMEFAC	(Wardhan et al. 1992)
Azotobacter vinelandii	*Az.v*	S96898	349	F.LUTESCEN*	(Venkatesh and Das 1992)
Bordetella pertussis	*Bd.p*	X53457	352	BRD.PERTUS	(Favre et al. 1991, Favre and Viret 1990)
Brucella abortus	*Br.a*	L00679	360	BRU.ABORTS	(Tatum et al. 1993)
Burkholderia cepacia[3]	*Bu.c*	D90120	347	BUR.CEPACI	(Nakazawa et al. 1990)
Campylobacter jejuni	*Ca.j*	U03121	343	CAM.JEJUNI	(Guerry et al. 1994)
Enterobacter agglomerans[4]	*En.a*	P33037	354	ER.HERBICO	(Rappold and Klingmueller 1993)
Erwinia carotovara	*Er.c*	X55554	342	ER.CAROTOV	(Zhao and McEntee 1990)
Escherichia coli	*Es.c*	V00328	353	E.COLI	(Horii et al. 1980, Sancar et al. 1980)
Haemophilus influenzae	*Ha.i*	L07529	354	H.INFLUENZ	(Zulty and Barcak 1993)
Helicobacter pylori	*He.p*	Z35478	347	HLB.PYLOR3	(Haas 1994)
Legionella pneumophila	*Le.p*	X55453	348	LEG.PNEUMO	(Zhao and Dreyfus 1990)
Magnetospirillum magnetotacticum[5]	*Ma.m*	X17371	344	MAG.MAGNE2	(Berson et al. 1990)
Methylobacillus flagellatum	*Mb.f*	M35325	344	MBS.FLAGEL	(Gomelsky et al. 1990)
Methylomonas clara	*Mm.c*	X59514	342	MLM.METHYL*	(Ridder et al. 1991)
Methylophilus methylotrophus	*Mp.m*	unpub.	342	MLP.METHY1	(Emmerson 1995, pers. commun)
Myxococcus xanthus 1	*Mx.x1*	L40367	342	MYX.XANTHU	(Inouye 1995, pers. commun.)
Myxococcus xanthus 2	*Mx.x2*	L40368	358	n/a[6]	(Inouye 1995, pers. commun.)
Neisseria gonorrhoeae	*Ne.g*	X17374	348	NIS.GONORR	(Fyfe and Davies 1990)
Proteus mirabilis	*Pr.m*	X14870	355	ARS.NASONI*	(Akaboshi et al. 1989)
Proteus vulgaris	*Pr.v*	X55555	325	P.VULGARIS	(Zhao and McEntee 1990)
Pseudomonas aeruginosa	*Ps.a*	X52261	346	PS.AERUGIN	(Sano and Kageyama 1987)
Pseudomonas fluorescens	*Ps.f*	M96558	352	PS.FLAVESC*	(De Mot et al. 1993)
Pseudomonas putida	*Ps.p*	L12684	355	PS.PUTIDA	(Luo et al. 1993)
Rhizobium leg. phaseoli	*Rz.p*	X62479	360	RHB.LEGUM6*	(Michiels et al. 1991)
Rhizobium leg. viciae	*Rz.l*	X59956	351	RHB.LEGUM8	(Selbitschka et al. 1991)
Rhizobium meliloti	*Rz.m*	X59957	348	RHB.MELIL2	(Selbitschka et al. 1991)
Rhodobacter capsulatus	*Rh.c*	X82183	355	RB.CAPSUL2	(Fernandez de Henestrosa 1994)
Rhodobacter sphaeroides	*Rh.s*	X72705	343	RB.SPHAER2	(Calero et al. 1994)
Rickettsia prowazekii	*Ri.p*	U01959	340	RIC.PROWAZ	(Dunkin and Wood 1994)
Serratia marcescens	*Se.m*	M22935	354	SER.MARCES	(Ball et al. 1990)
Shigella flexneri	*Sh.f*	X55553	353	n/a	(Zhao and McEntee 1990)
Thiobacillus ferrooxidans	*Tb.f*	M26933	346	THB.CALDUS*	(Ramesar et al. 1989)
Vibrio anguillarum	*Vi.a*	M80525	348	V.ANGUILLA	(Gammie and Crosa 1991, Tolmasky et al. 1992)
Vibrio cholerae	*Vi.c*	U10162	354	V.CHOLERAE	(Margraf et al. 1995, Stroeher et al. 1994)
Xanthomonas oryzae	*Xa.o*	unpub.	355	XAN.ORYZAE	(Mongkolsuk 1995, pers. commun.)
Yersinia pestis	*Ye.p*	X75336	356	YER.PESTIS	(Kryukov et al. 1993)
Gram Positives					
Acholeplasma laidlawii	*Acp.l*	M81465	331	ACP.LAIDLA	(Dybvig and Woodard 1992)
Bacillus subtilis	*Ba.s*	X52132	347	B.SUBTILIS	(Stranathan et al. 1990)
Corynebacterium glutamicum	*Co.g*	X77384	376	Z46753	(Billman-Jacobe 1994, Kerins et al. 1994)
Lactococcus lactis	*La.l*	M88106	365	LCC.LACTIS	(Duwat et al. 1992a)
Mycobacterium leprae	*Myb.l*	X73822	711	MYB.LEPRAE	(Davis et al. 1994)
Mycobacterium tuberculosis	*Myb.t*	X58485	790	MYB.TUBER2	(Davis et al. 1991)
Mycoplasma mycoides	*Myp.m*	L22073	345	M.MYCOIDES	(King et al. 1994)
Mycoplasma pulmonis	*Myp.p*	L22074	339	M.PULMONIS	(King et al. 1994)
Staphylococcus aureus	*Sta.a*	L25893	347	STP.AUREUS	(Bayles et al. 1994)
Streptococcus pneumoniae	*Stc.p*	Z17307	388	STC.SALIVA*	(Martin et al. 1992)
Streptomyces ambofaciens	*Stm.a*	Z30324	372	STM.AMBOFA	(Aigle et al. 1994)
Streptomyces lividans	*Stm.l*	X76076	374	STM.LIVIDA	(Nussbaumer and Wohlleben 1994)
Streptomyces violaceus[7]	*Stm.v*	U04837	377	STM.COELI3*	(Yao and Vining 1994)
Cyanobacteria/Chloroplasts					
Arabidopsis thaliana	*Ar.t*	M98039	439	NICO.TAB_C*	(Binet et al. 1993, Cerutti et al. 1992)
Anabaena variabilis	*An.v*	M29680	358	X59559*	(Owttrim and Coleman 1989)
Synechococcus sp. PCC7942	*Sy.79*	unpub.	361	PHRM.MINUT*	(Coleman 1995)
Synechococcus sp. PCC7002	*Sy.70*	M29495	348	SYN.6301*	(Murphy et al. 1987, Murphy et al. 1990)
Deinococcus-Thermus Group					
Deinococcus radiodurans[8]	*De.r*	U01876	363	D.RADIODUR	(Gutman et al. 1994)
Thermus aquaticus	*Th.a*	L20095	340	T.AQUATICU	(Angov and Camerini-Otero 1994, Wetmur et al. 1994)
Thermus thermophilus	*Th.t*	D13792	340	T.THMOPHL	(Kato and Kuramitsu 1993, Wetmur et al. 1994)
Chlamydia/Planctomyces					
Chlamydia trachomatis	*Ch.t*	U16739	352	CLM.TRACHO	(Larsen 1994, Zhang et al. 1994)
Spirochaetes					
Borrelia burgdorferi	*Bo.b*	unpub.	365	BOR.BURGDO	(Huang 1995, pers. commun.)
Bacteroides					
Bacteroides fragilis	*Bct.f*	M63029	318	BAC.FRAGIL	(Goodman and Woods 1990)
Thermophilic O_2 Reducers					
Aquifex pyrophilus	*Aq.p*	L23135	348	AQU.PYROPH	(Wetmur et al. 1994)
Thermotogales					
Thermotoga maritima	*Tg.m*	L23425	356	TT.MARITIM	(Wetmur et al. 1994)

[1] Names refer to Ribosomal Database Project entries (Maidak et al. 1994). Numbers are Genbank entries.

[2] The SS-rRNA sequences that come from a different species than the RecA sequences are indicated by an asterix *. The species are ABA.PASTER (*Acetobacter pasteurianus*), F.LUTESCEN("*Flavobacterium*" *lutescens*, MLM.METHYL (*Methylomonas methylovora*), ARS.NASONI (*Arsenophonus nasoniae*), PS.FLAVESC (*Pseudomonas flavescens*), STM.COELI3 (*Streptomyces coelicolor*), STC.SALIVA (*Streptococcus salivarius*) NICO.TAB_C (*Nicotiana tabacum*), X59559 (*Anabaena* sp. PCC7120), PHRM.MINUT (*Phormidium minutum*), and SYN.6301 (*Synechococcus* sp. PCC 6301).

[3] also known as *Pseudomonas cepacia*

[4] also known as *Erwinia herbicola*

[5] also known as *Aquaspirillum magnetotacticum*

[6] For most of the analyses only one SS-rRNA was used for the two *M. xanthu*s RecAs. For some analyses the SS-rRNA of *Cystobacter fuscus* (CYS.FUSCUS) was also used.

[7] Also known as *Streptomyces venezuelae*

[8] Also known as *Micrococcus radiodurans*

Table 2: Consensus Phylogenetic Groups

		Comprable SS-RNA Consensus ?[1,2,3]	RecA Bootstrap[4]			sRNA Bootstrap[5]		
Clade	**Species in RecA Consensus Clade[6]**		**PP**	**NJ**	**FM**	**DP**	**NJ**	**FM**
Proteobacteria - γ1[7]	*Escherichia coli, Shigella flexneri, Yersinia pestis, Erwinia carotovara, Serratia marcescens, Enterobacter agglomerans, Proteus vulgaris, Pr. mirabilis, Vibrio cholerae, V. anguillarum, Haemophilus influenzae*	YES	78	91	100	100	100	100
Proteobacteria - γ2	*Azotobacter vinelandii, Pseudomonas aeruginosa, Ps. putida, Ps. fluorescens*	YES	100	100	100	100	100	100
Proteobacteria - γ	γ1, γ2, *Acinetobacter calcoaceticus*	YES (+ *Legpn*)	33	63	75	48	85	92
Proteobacteria - β1	*Methylobacillus flagellatum, Methylomonas clara, Methylophilus methylotrophus, Burkholderia cepacia, Bordetella pertussis*	YES (+ *Neigo*)	74	84	88	100	100	100
Proteobacteria - β2	*Thiobacillus ferrooxidans, Acidiphilium facilis*	No	100	100	100	*	*	*
Proteobacteria - βγ	γ, β1, β2, *Xanthomonas oryzae, Neisseria gonorrhoeae, Legionella pneumophila*	YES (-*Acifa*)	53	86	95	90	94	95
Proteobacteria - α	*Rhodobacter capsulatus, Rho. sphaeroides, Rhizobium meliloti, Rhi. viciae, Rhi. phaseoli, Acetobacter polyoxogenes, Magnetospirillum magnetotacticum, Brucella abortus, Agrobacterium tumefaciens, Rickettsia prowazekii*	YES (+*Acifa*)	14	68	72	100	100	100
Proteobacteria - αβγ	α, β, γ	YES	10	57	58	93	96	96
Proteobacteria - δ	*Myxococcus xanthus* 1, *M. xanthus* 2	YES	43	71	42	*[8]	*	*
Proteobacteria - ε	*Campylobacter jejuni, Helicobacter pylori*	YES	100	100	100	100	100	100
Proteobacteria	γ, β, α, δ, ε	NO	14	38	49	*	*	36
Gram "+" High GC	*Corynebacterium glutamicum, Streptomyces ambofaciens, S. violaceus, S. lividans, Mycobacterium tuberculosis, Myb. leprae*	YES	97	100	100	100	100	100
Gram "+" Low GC	*Bacillus subtilis, Lactococcus lactis, Streptococcus pneumoniae, Staphylococcus aureus, Acholeplasma laidlawii*	YES (+ *Mycpn, Mycge*)	27	59	63	50	56	80
Mycoplasmas	*Mycoplasma mycoides, Myp. pulmonis*	YES (+ *Achla)*	88	100	98	71	88	84
Cyanobacteria	*Arabidopsis thaliana, Anabaena variabilis, Synechococcus* sp. PCC7942, *Syn.* sp. PCC7002	YES	100	96	91	100	100	100
Deinococcus-Thermus	*Deinococcus radiodurans, Thermus aquaticus, T. thermophilus*	NO	95	96	95	*	*	*

[1] For those groups which have 1 or 2 additional species in the SS_rRNA tree, the extra species are listed
[2] Groups found in trees generated by neighbor-joining, Fitch-Margoliash, De Soete and *dnapars.*
[3] Abbreviations are for *Legionella pneuomnphila, Neisseria gonorrhoeae, Acidiphilium facilis, Mycosplasma pneumonia, M. genitalium, and Acholeplasma laidlawii*
[4] PP = protein parsimony, NJ = neighbor-joining, FM = Fitch-Margoliash, DP = DNA parsimony
[5] Bootstrap values are shown for comprable clade
[6] Groups found in trees generated by neighbor-joining, Fitch-Margoliash, De Soete, *protpars* and PAUP
[7] Not applicable.
[8] Bootstraps were only calculated for trees with the one δ sequence (see Methods)

Nterminus

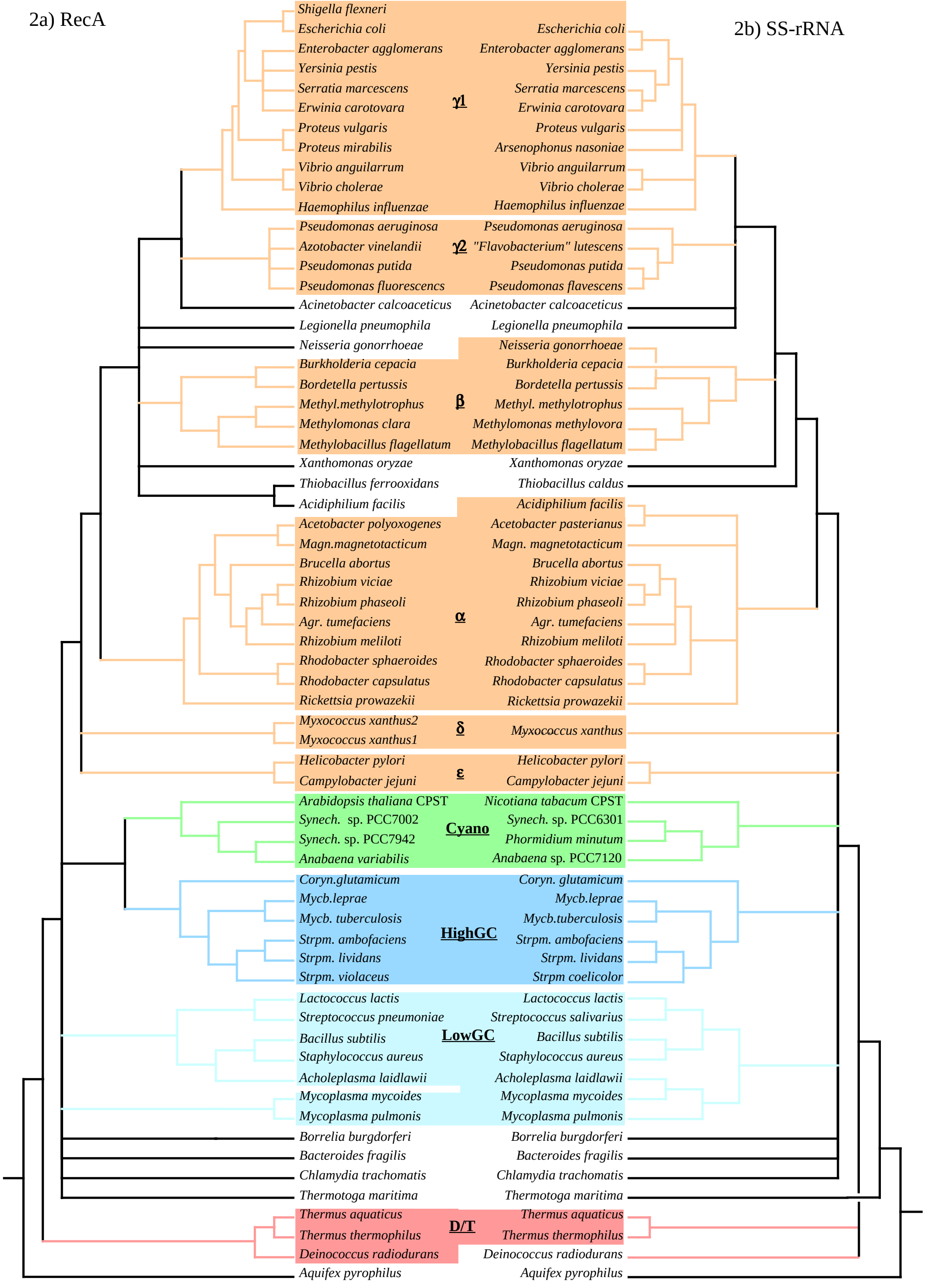
```
#                                 10        20        30          40        50        60        70        80        90       100       110       120       130       140       150       160       170       180
                                   |         |         |           |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
Es.c     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTGEQALEICDALARSGAVDVIVVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKQS
Sh.f     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTGEQALEICDALARSGAVDVIVVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKQS
En.a     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYAKKLGVDIDNLLCSQPDTGEQALEICDALTRSGAVDVIIVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKNA
Se.m     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYAKKLGVDIDNLLCSQPDTGEQALEICDALTRSGAVDVIIVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKNA
Pr.v     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMN--VETISTGSLYLDVALGAGGLPRGRIVEIYGPESSGKTTLTLQVIAAAQREGKICAFIDAEHALDPIYAQKLGVDIDNLLCSQPDTGEQALEICDALSRSGAVDVIVVDSVAALTPKAEIEGEIGDSHVGLAARMMSQAMRKLAGNLKNS
Ye.p     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYAKKLGVDIDNLLCSQPDTGEQALEICDALTRSGAVDVIIVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKNA
Pr.m     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMN--VETISTGSLSLDVALGAGGLPRGRIVEIYGPESSGKTTLTLQVIASAQREGKICAFIDAEHALDPIYAQKLGVDIDNLLCSQPDTGEQALEICDALSRSGAVDVIVVDSVAALTPKAEIEGEIGDSHVGLAARMMSQAMRKLAGNLKNS
Er.c     ----------------MAIDENKQKALAAALGQIEKQFGKGSIMRLGEDRSMD--VETISTGSLSLDIALGAGGLPMGRIVEIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYAKKLGVDIDNLLCSQPDTGEQALEICDALTRSGAVDVIIVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNLKQA
Vi.a     ----------------KVMDENKQKALAAALGQIEKQFGKGSIMRLGDNRTMD--VETISTGSLSLDIALGAGGLPMGRIVEVYGPESSGKTTLTLELIAAAQRVGKTCAFIDAEHALDPIYAKKLGVNIDELLVSQPDTGEQALEICDALARSGAIDVIVIDSVAALTPKAEIEGEMGDSHMGLQARMLSQAMRKLTGNLKQS
Vi.c     ------------------MDENKQKALAAALGQIEKQFGKGSIMRLGDNRAMD--VETISTGSLSLDIALGAGGLPMGRIVEIFGPESSGKTTLTLELIAAAQREGKTCAFIDAEHALDPVYAKKLGVNIDELLVSQPDTGEQALEICDALARSGAVDVIVVDSVAALTPKAEIEGEMGDSHMGLQARMLSQAMRKLTGNLKQS
He.i     ----------------MATQEEKQKALAAALGQIEKQFGKGSIMKLGDTKTLD--VESISTGSLGLDVALGIGGLPMGRIVEIFGPESSGKTTLTLSVIAQAQKAGKTCAFIDAEHALDPIYAAKLGVDVKELFVSQPDNGEQALEICDALVRSGAIDVIIVDSVAALTPKAEIEGDMGDSHMGLQARLMSQALRKLTGQIKNA
Ps.f     ------------------MDDNKKKALAAALGQIERQFGKGAVMRMGDHDRQA--IPAISTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLSVIAQAQKMGATCAFVDAEHALDPEYAGKLGVNVDDLLVSQPDTGEQALEITDMLVRSNAIDVIVVDSVAALVPKAEIEGEMGDMHVGLQARLMSQALRKITGNIKNA
Ps.p     ------------------MDDNKKRALAAALGQIERQFGKGAVMRMGDHERQA--IPAISTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLSVIAEAQKNGATCAFVDAEHALDPEYAGKLGVNVDDLLVSQPDTGEQALEITDMLVRSNAVDVIIVDSVAALVPKAEIEGEMGDMHVGLQARLMSQALRKITGNIKNA
Ps.a     ------------------MDENKKRALAAALGQIERQFGKGAVMRMGDHERQA--IPAISTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLSVIAEAQKQGATCAFVDAEHALDPDYAGKLGVNVDDLLVSQPDTGEQALEITDMLVRSNAVDVIIVDSVAALVPKAEIEGEMGDAHVGLQARLMSQALRKITGNIKNA
Az.v     ------------------MDENKKRALAAALGQIEKQFGKGAVMRMGDHERQA--IPAISTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLSTIAEAQKQGATCAFVDAEHALDPDYAAKLGVNVDDLLVSQPDTGEQALEITDMLVRSNAVDVIIVDSVAALVPKAEIEGEMGDQHVGLQARLMSQALRKITGNIQNA
Le.p     ------------------MEENKQKALSAAVSQIERQFGKGSVMRMGDSTVSRD-IEAISTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLQVIAECQKMGGTAAFIDAEHALDPSYAQKLGVKVDELLVSQPDTGEQALEITDMLVRSAAVDVVIIDSVAALTPKAEIEGEMGDSHVGLQARLMSQALRKLTANIKRS
Acn.c    ------------------MDDNKSKALQAALSQIEKQFGKNTVMRLGDNTVQA--VEAVSTGSLTLDIALGIGGLPKGRIVEIYGPESSGKTTMTLQAIAQCQKAGGTCAFIDAEHALDPQYARKLGVDIDNLLVSQPDHGEQALEIADMLVRSGAIDLIVVDSVAALTPKAEIEGEMGDSHMGLQARLMSQALRKITGNAKRS
Bd.p     ----------MDDKTSKAAAAEKAKALAAALSQIEKQFGKGSIMRYGDNEVEHD-IQVVSTGSLGLDIALGVGGLP-AAVIEVYGPESSGKTTLTLQVIAEMQKLGGTCAFVDAEHALDVQYASKLGVNLTDLLISQPDTGEQALEITDALVRSGSVDLIVIDSVAALVPKAEIEGEMGDSLPGLQARLMSQALRKLTATIKRT
Bu.p     ------------------MTAEKSKALAAALAQIEKQFGKGSIMRMGDGEAAED-IQVVSTGSLGLDIALGVGGLPRGRVVEIYGPESSGKTTLTLQVIAELQKLGGTAAFIDAEHALDVQYAAKLGVNVPELLISQPDTGEQALEITDALVRSGSIDMIVIDSVAALVPKAEIEGEMGDSLPGLQARLMSQALRKLTGTIKRT
Mm.c     ------------------MDDNKSKALAAALSQIEKQFGKGSIMRMGDADIGED-LQVVSTGSLGLDIALGVGGLPRGRIIEIYGPESSGKTTLTLSAIAEMQKLGGVAAFIDADDALDPQYASKLGVNVPELLISQPDTGEQALEIADMLVRSGSVDIVVVDSVAALTPRAEIEGEMGDSHMGLQARLMSQALRKLTGNIKRT
Mb.f     ------------------MDENRSKALAAALSQIEKQFGKGSIMRMGDTDVAAD-IQAVSTGSLGLDIALGIGGLPRGRIVEIYGPESSGKTTLTLSVIAQMQKLGGTAAFIDAEHALDPVYAQKLGVNVSDLLISQPDTGEQALEIADMLVRSGSVDVVVVDSVAALTPKAEIEGEMGDSHMGLQARLMSQALRKLTANIKRT
Tb.f     ------------------MDEQRSKGLSAALSQIDKQFGKGAVMRLGDHNAIKD-IEVYSTGSLGLDLALGVGGLPRGRVVEIYGPESSGKTTLTLHAIASCQAAGGTAAFIDAEHALDPGYAHKLGVDLENLLISQPDTGEQALEIADMLVRSGAVDLIVIDSVAALTPKAEIEGEMGDSHVGLQARLMSQALRNLTANISRS
Acd.f    -----------------MDEPGGKIEFSPAFMQIEGQFGKGAVMRAGDKPGIND-PDVKSTGSLGLDGALGQGGLPRGRVVEIYGPESSGKTTLTLKAIASAQAEGATPAFTDAEHALDPGFASKLGVNVKRLLISQPDTGEQALEIADMLFRSGAVDVIVKDSVAALTPKAEIEGEMGDSHQGLHARLMSQALRNKTANISRW
Ne.g     ------------------MSDDKSKALAAALAQIEKSFGKGAIMKMDGSQQEEN-LEVISTGSLGLDLALGVGGLPRGRIVEIFGPESSGKTTLCLEAVAQCQKNGGVCAFVDAEHAFDPVYARKLGVKVEELYLSQPDTGEQALEICDTLVRSGGIDMVVVDSVAALVPKAEIEGDMGDSHVGLQARLMSQALRKLTGHIKKT
Rz.p     ------MSQNSLRLVEDKSV-DKSKALEAALSQIERSFGKGSIMKLGSNENVIE-IETISTGSLGLDIALGVGGLPKGRIIEIYGPESSGKTTLALQTIAESQKKGGICAFVDAEHALDPVYARKLGVDLQNLLISQPDTGEQALEITDTLVRSGAVDVLVVDSVAALTPRAEIEGEMGDSLPGLQARLMSQALRKLTASISKS
Rz.v     --------------------VDKSKALEAALSQIERSFGKGSIMKLGSNENVVE-IETISTGSLGLDIALGVGGLPRGRIIEIYGPESSGKTTLALQTIAEAQKKGGICAFVDAEHALDPVYARKLGVDLQNLLISQPDTGEQALEITDTLVRSGRVDVLVVDSVAALTPRAEIEGEMGDTVPGLQARLMSQALRKLTASISKS
Rz.m     ------MAQNSLRLVEDKSVMDKSKALEAALSQIERSFGKGSIMKLGAKDSVVE-IETVSTGSLGLDIALGIGGLPKGRIIEIYGPESSGKTTLALQTIAEAQKKGGICGFVDAEHALDPVYARKLGVDLENLLISQPDTGEQALEITDTLVRSGAIDILVIDSVAALVPRAEIEGEMGDSLPGMQARLMSQALRKLTASISKS
Ag.t     ---------------EDKSV-DKSKALEAALSQIERSFGKGSIMKLGSNENVVE-VETVSTGSLSLDIALGIGGLPKGRIIEIYGPESSGKTTLALQTIAEAQKKGGICAFVDAEHALDPVYARKLGVDLQSLLISQPDTGEQALEITDTLVRSGAVDVLVIDSVAALTPRAEIEGEMGDSLPGLQARLMSQALRKLTASISKS
Br.a     ------MSQNSLRLVEDNSV-DKTKALDAALSQIERAFGKGSIMRLGQNDQVVE-IETVSTGSLSLDIALGVGGLPKGRIVEIYGPESSGKTTLALHTIAEAQKKGGICAFVDAEHALDPVYARKLGVHLENLLISQPITGEQALEITDTLVRSGAIDVLVVDSVAALTPRAEIEGEMGDSH-GLQARLMSQAVRKLTGSISRS
Ma.m     --------------------MDRQKALEAAVSQIERAFGKGSIMKLGGKDQVVE-TEVVSTRILGLDVALGIGGVPRGRIIEVYGPESSGKTTLALHIIAEAQKKGGTCAFVDAEHALDPSYARKLG-ALDELLISEPDAGEQALEIADTLVRPGAVDVLVVDSVAALVPRGELEGEMGDNHMGLHARLMSQALRKLTGSVSKS
Act.p    --------------MAQAPGIDKSKALEGALSQIERAFGKGSIMRMGERPTEQ--VDVISTGSLGLDIALGIGGLPRGRIIEIYGPESSGKTTMALHAIAEAQRKGGTCAFIDAEHALDPGYARKLGVDVDNLLISQPDAGEQALEIADTLVRSGAVDVLVVDSVAALVPRAELEGDMGDSHVGLHARLMSQALRKLTGSVSRS
Rh.s     --------------------MDKAKALESALAQIERQFGKGSIMKLGANSPVME-IEATSTGSLGLDIALGIGGLPKGRIIEIYGPESSGKTTLTLHVVAEEQKKGGVCAFVDAEHALDPQYAKKLGVNLDELLISQPDTGEQALEIVDTLVRSGAVNLIVVDSVAALTPKSEIEGDMGDMQMGSQARLMSQAMRKLTASIGRS
Rh.c     ------MATTGLFEMNDKGKADKQKALECALAQIERQFGKGSIMKLGGDTPPPE-IEATSTGSLGLDIALGIGGLPKGRIVEIYGPESSGKTTLTLHCIAEEQKKGGVCAFVDAEHALDPYYAKKLGVSLEDLLISQPDTGEQALEIVDTLVRSGAVSLVVVDSVAALTPKAEIEGDMGDATVGAQARLMSQAMRKLTASIGRS
Ri.p     -----------------MSNIDKERAIAAALAQIEKSYGKGSVMKLGQRPNVD--IEAISTGSLGLDIALGIGGVPKGRIIEIFGPESSGKTTLTLHLIAESQKKGGTCAFIDAEHALDPAYAKKLGVNIDELIISQPDTGEQALEIADTLIRSGGIDMIIIDSVAALVPKSEIEGEMGDAQMASQARLMSQALRKLTASINRT
Mx.x1    ----------------MSKLAEKLKAVAAAVASIEKQFGRGSVMTLGGEAREQK-VAVIPSGSVGVDRALGVGGYPRGRVVEVFGNESSGKTTLTLHAIAQVQAAGGVAAFIDAEHALDVSYARKLGVRVEELLVSQPDTGEQALEITEHLVRSGAVDLIVVDSVAALVPRAEIEGEMGDAHMGVQARLMSQALRKLTGAVSRS
Mx.x2    ----------------MAVNQEKEKAIELAMSAVERQFGKGSIMRLGNDEPMMRDVQAIPTGSISLDIALGVGGVPKGRIIEIFGPESSGKTTLCLHIVAEAQKRGGICGYVDAEHALDVGYARKLGVRTDDLLLSQPDTGEQALEIAEMLVRSGAIDVLVVDSVAALVPKAELEGEMGDAHMGVQARLMSQALRKLTGTIAKS
He.p     ---------------MAIDEDKQ-KAISLAIKQIDKVFGKGALVRLGDKQVEK--IDAISTGSLGLDLALGIGGVPKGRIIEIYGPESSGKTTLSLHIIAECQKNGGVCAFIDAEHALDVHYAKRLGVDTQNLLVSQPDTGEQALEILETITRSGGIDLVVVDSVAALTPKAEIDGDMGDQHVGLQARLMSHALRKITGVLHKM
Ca.j     -----------------MDDNKR-KSLDAALKSLDKTFGKGTILRLGDKEVEQ--IDSIGTGSVGLDLALGIGGVPKGRIIEIYGPESSGKTTLTLHIIAECQKAGGVCAFIDAEHALDVKYAKNLGVNTDDLYVSQPDFGEQALEIVETIARSGAVDLIVVDSVAALTPKAEIEGDMGDQHVGLQARLMSQALRKLTGIVHKM
An.v     -------------MAINTDTSGKQKALTMVLNQIERSFGKGAIMRLGDATRMR--VETISTGALTLDLALG-GGLPRGRVIEIYGPESSGKTTVALHAIAEVQKEGGIAAFVDAEQALDPTYASALGVDIQNLLVSQPDTGESALEIVDQLVPSAAVDIVVIDSVAALVPRAEIEGDMGDAHVGLQARLMSQALRKITGNIGKS
Sy.70    -------------MSAISNNPDKEKALNLVLNQIERNFGKGAIMRLGDAAQMK--VATIPSGALTLDQAMG-GGFPRGRIVEIYGPESSGKTTVALHAIAEVQKAGGVAAFIDAEHALDPTYSAALGVDIENLLVAQPDNGESALEIADQLVRSAAVDLIVIDSVAALVPRAEIEGEMGDVQVGLQARLMSKALRKIAGNMGRS
Ar.t     SSEFDDRINGALSPDADSRFLDRQKALEAAMNDINSSFGKGSVTRLGSAGGAL--VETFSSGILTLDLALG-GGLPKGRVVEIYGPESSGKTTLALHAIAEVQKLGGNAMLVDAEHAFDPAYSKALGVDVENLIVCQPDNGEMALETADRMCRSGAVDLICVDSVSALTPRAEIEGEIGMQQMGLQARLMSQALRKMSGNASKA
Myp.m    --MSTELQNTIENNDIRESQMWNSKELKEAIKEIEKMFGKGSIMVLGQSDNLN--IETFSSGSLLLDNALGIGGYPKGRIIEIYGPESSGKTTLSLHAICEVQKLGGIAAFIDAEHSLEPKYCQTLGIDTNKLLVSQPDNGEQALDILEMLINSNSIDLIVVDSVAALVPKTELDGEMSDQSIGLQARMMSKALRKLNWLIAKS
Myp.p    ----------MSENNQSNQNNQINKIIKSTIEEIEKKFGNESIMLLGQKEKCD--VDVFSSGSYAINSALGIGGFPKGRIIEIFGPESSGKTTIALHTIAEIQKKNGFAAFIDVEHSIDPVYAKNLGIDIDNLLISQPDSGEQALEIVDILAKSGSIDLIVVDSVAALVPEAELNGEMKDQSIGLQARLMSKALRKITGSLSKN
Ba.s     -------------------MSDRQAALDMALKQIEKQFGKGSIMKLGEKTDTR--ISTVPSGSLALDTALGIGGYPRGRIIEVYGPESSGKTTVALHAIAEVQQQR-TSAFIDAEHALDPVYAQKLGVNIEELLLSQPDTGEQALEIAEALVRSGAVDIVVVDSVAALVPKAEIEGDMGDSHVGLQARLMSQALRKLSGAINKS
Sta.a    ------------------MDNDRQKALDTVIKNMEKSFGKGAVMKLGDNIGRR--VSTTSTGSVTLDNALGVGGYPKGRIIEIYGPESSGKTTVALHAIAEVQSNGGVAAFIDAEHALDPEYAQALGVDIDNLYLSQPDHGEQGLEIAEAFVRSGAVDIVVVDSVAALTPKAEIEGEMGDTHVGLQARLMSQALRKLSGAISKS
Stc.p    ----MAKKPKKLEEISKKFGAEREKALNDALKLIEKDFGKGSIMRLGERAEQK--VQVMSSGSLALDIALGSGGYPKGRIIEIYGPESSGKTTVALHAVAQAQKEGGIAAFIDAEHALDPAYAAALGVNIDELLLSQPDSGEQGLEIAGKLIDSGAVDLVVVDSVAALVPRAEIDGDIGDSHVGLQARMMSQAMRKLGASINKT
La.la    ---MATKKKTNFDDITKKYGAERDKALADALALIEKDFGKGSLMRLGEAANQK--VSVVSSGSLALDIALGAGGYPKGRIVEIYGPESSGKTTVALHAVAAVQKEGGIAAFIDAENALDPEYAKALGVNIDELLLSQPDYGEQGLQIAEKLITSGAVDLVVIDSVAALVPKAEIDGEIGDSSVGLQARMMSQAMRKLAGHINKT
Ac.l     -----------------MSDNKKQQALELALKQIEKQFGKGSIMKLGDGADHS--IEAIPSGSIALDIALGIGGYPRGRIIEVYGPESSGKTTLTLHAMASAQKQGGTVAFIDAEHALDPNYAKALGVDLDNLVLSQPDTGEQALDIAEALIKSGSIDMIVIDSVAALVPEAEIAGDMSANHVGLQARMMSQAMRKMSGVISKS
Stm.l    -----------------MAGTDREKALDAALAQIERQFGKGAVMRMGDRTNEP--IEVIPTGSTALDVALGVGGIPRGRVVEVYGPESSGKTTLTLHAVANAQKAGGQVAFVDAEHALDPEYAKKLGVDIDNLILSQPDNGEQALEIVDMLVRSGALDLIVIDSVAALVPRAEIEGEMGDSHVGLQARLMSQALRKITSALNQS
Stm.a    -----------------MAGTDREKALDAALAQIERQFGKGAVMRMGDRSKEP--IEVIPTGSTALDVALGVGGLPRGRVIEVYGPESSGKTTLTLHAVANAQKAGGQVAFVDAEHALDPEYAQKLGVDIDNLILSQPDNGEQALEIVDMLVRSGALDLIVIDSVAALVPRAEIEGEMGDSHVGLQARLMSQALRKITSALNQS
Stm.v    -----------------MAGTDREKALDAALAQIERQFGKGAVMRMGDRTQEP--IEVISTGSTALDIALGVGGLPRGRVVEIYGPESSGKTTLTLHAVANAQKAGGQVAFVDAEHALDPEYAKKLGVDIDNLILSQPDNGEQALEIVDMLVRSGALDLIVIDSVAALVPRAEIEGEMGDSHVGLQARLMSQALRKITSALNQS
Co.g     -----MHPKKTATKATAAKGNDRQKALDAALALIEKDFGKGAVMRLGDENRPP--IQTISSGNTAIDIALGIGGFPRGRIVEVYGPESSGKTTVALHAIAQAQKAGGIAAFIDAEHALDPDYARKLGVDTDALLVSQPDTGEQALEIADMLVRSGAIDIIVIDSVAALTPKAEIEGEMGDSHVGLQARLMSQALRKMTGALYNS
Myb.t    ----------------MTQTPDREKALELAVAQIEKSYGKGSVMRLGDEARQP--ISVIPTGSIALDVALGIGGLPRGRVIEIYGPESSGKTTVALHAVANAQAAGGVAAFIDAEHALDPDYAKKLGVDTDSLLVSQPDTGEQALEIADMLIRSGALDIVVIDSVAALVPRAELEGEMGDSHVGLQARLMSQALRKMTGALNNS
Myb.l    ----------------MAQVPDREKALELAMAQIEKNYGKGSVMRLGDEMCQP--ISVIPTGSIALDVALGIGGLPRGRIVEIYGPESSGKTTVALHAVANAQAVGGVAAFIDAEHALEPEYAKKLGVDTDSLLVSQPDTGEQALEIADMLIRSGALDIVVIDSVAALVPRAELEGEMGDSYVGLQARLMSQALRKMTGALSNS
Ch.t     -----------------MSVPDRKRALEAAIAYIEKQFGAGSIMSLGKHSSAHE-ISTIKTGALSLDLALGIGGVPKGRIVEIFGPESSGKTTLATHIVANAQKMGGVAAYIDAEHALDPNYAALIGANINDLMISQPDCGEDALSIAELLARSGAVDVIVIDSVAALVPKSELEGEIGDVHVGLQARMMSQALRKLTATLART
Bct.f    -----------------MASSEKLKALQAAMDKIEKSFGKGSIMKMGEEVVEQ--VEVIPTGSIALNAALGVGGYPRGRIIEIYGPESSGKTTLAIHAIAEAQKAGGIAAFIDAEHAFDRFYAAKLGVDVDNLFISQPDNGEQALEIAEQLIRSSAIDIIVVDSVAALTPKAEIEGDMGDNKVGLQARLMSQALRKLTSAVSKT
Tg.m     ----------------MPEEKQKKSVLEKALKRIEENFGKGSIMILGDETQVQP-VEVIPTGSLAIDIATGVGGYPRGRIVEIFGQESSGKTTLALHAIAEAQKMGGVAAFIDAEHALDPVYAKNLGVDLKSLLISQPDHGEQALEIVDELVRSGVVDLIVVDSVAALVPRAEIEGAMGDMQVGLQARLMSQALRKIAGSVNKS
De.r     -----MSKDATKEISAPTDAKERSKAIETAMSQIEKAFGKGSIMKLGAESKLD--VQVVSTGSLSLDLALGVGGIPGGRITEIYGPESGGKTTLALAIVAQAQKAGGTCAFIDAEHALDPVYARALGVNADELLVSQPDNGEQALEIMELLVRSGAIDVVVVDSVAALTPRAEIEGDMGDSLPGLQARLMSQALRKLTAILSKT
Th.a     -----------------MEENKR-KSLENALKTIEKEFGKGAVMRLGEMPKLQ--VDVIPTGSLGLDLALGIGGIPRGRVTEIFGPESGGKTTLALTIIAQAQKGGGVAAFVDAEHALDPLYAKKLGVDVQELLVSQPDTGEQALEIVELLARSGAVDVIVVDSVAALVPKAEIEGEMGDQHVGLQARLMSQALRKLTAVLSKS
Th.t     -----------------MDESKR-KALENALKAIEKEFGKGAVMRLGEMPKQQ--VDVIPTGSLALDLALGIGGIPRGRIVEIYGPESGGKTTLALTIIAQAQRRGGVAAFVDAEHALDPLYAQRLGVQVEDLLVSQPDTGEQALEIVELLARSGAVDVIVVDSVAALVPRAEIEGEMGDQHVGLQARLMSQALRKLTAVLAKS
Aq.p     ------------MARVSENLSEKMKALEVALSSIEKRFGKGAVMPLKAVETVE--VETIPTGSISLDIATGVGGIPKGRITEIFGVESSGKTTLALHVIAEAQKRGGVAVFIDAEHALDPKYAKKLGVDVDNLYISQPDYGEQALEIAESLINSGAVDVIVVDSVAALVPKDELEGEMGEAQVGKQARLMSQALRKLKGAVHRS
                                          .   .*  ..                 .    .  * * ** *   . *..* ** ****..   ..  *         * . . .  ..  .*     *  ..*  ** *           . ..  ***.** * *  *. * .     .  **..*.*.*

Mask     00000000000000000000000011111111111111111111111111110001111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
```

Cterminus

```
#         190       200          210       220       230            240       250          260       270       280       290         300       310       320       330             340       350
              |         |            |         |         |              |         |            |         |         |         |           |         |         |         |               |         |
Es.c   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGAV-KEGE----NVVGSETRVKVVKNK--IAAPFKQAEFQILYGEGINFYGELVDLGVKEKLIEKAGAWYSYKGEK--IGQGKANATAWLKDNPETAKEIEKKVRELLLSNPN-------STPDFSVDDSEGVAETNEDF------------------
Sh.f   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGAV-KEGE----NVVGSETRVKVVKNK--IAAPFKQAEFQILYGEGINFYGELVDLGVKEKLIEKAGAWYSYKGEK--IGQGKANATAWLKDNPETAKEIEKKVRELLLSNPN-------STPDFSVDDSEGVAETNEDF------------------
En.a   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGAI-KEGD----VVVGSETRVKVVKNK--IAAPFKQAEFQILYGEGININGELIDLGVKHKLIEKAGAWYSYNGEK--IGQGKANSCNYLKENPKVAAELDKKLRDMLLSGTG---E---LSVATTAEDADDNMETSEEF------------------
Se.m   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGAI-KEGD----EVVGSETRVKVVKNK--IAAPFKQAEFQIMYGEGINSRGELVDLGVKHKMIEKAGAWYSYNGEK--IGQGKANACNFLKENPAIAAELDKKLRDLLLHSGG---E---LVAASGDDFEDDEAETSEQF------------------
Pr.v   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGSV-KNGD----EVVGSETRVKVVKNK--IAAPFKQAEFQIMYGEGINTFGELIDLGVKHKLVEKAGAWYSYNGEK--IGQGKANATTYLKENPEMYNELNTKLREMLLNHAGEF-T---SARDFANDSDDAADIEETEK------------------
Ye.p   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGAV-KDGD----VVVGSETRVKVVKNK--IAAPFKQAEFQILYGEGININGELVDLGVKLKLIEKAGAWYSYYGDK--IGQGKANASNYLKENPANAAELDKKLREMLLNGGNGE-Q---PVAAATAEFADGADETNEEF------------------
Pr.m   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRIGSV-KNGD----EVIGSETRVKVVKNK--VAAPFKQAEFQIMYGEGINTYGELIDLGVKHKLVEKAGAWYSYNGEK--IGQGKANATNYLKEHPEMYNELNTKLREMLLNHAGEF-T---SAADFAGEESDSDADDTKE-------------------
Er.c   NTLLIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGAI-KDGE----EVVGSETRVKVVKNK--VAAPFKQAEFQILYGEGINIHGELVDLGVKHKLIEKAGAWYSYNGDK--IGQGKANACNFLKENSLVKATKNFNGC-----------------------------------------------------
Vi.a   NCMCIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGSI-KEGD----EAVGNETRIKVVKNK--IAAPFKQADTQILYGQGFNREGELVDLGVKHKLVEKAGAWYSYNGDK--IGQGKANACKFLRENPAAAMALDTKLREMLLNP--AE-L---IVEEPILSEMPQE----EEL------------------
Vi.c   NCMCIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGAI-KEGE----EVVGNETRIKVVKNK--IAAPFKEANTQIMYGQGFNREGELIDLGVKHKMVEKSGAWYSYNGDK--IGQGKANACKYLKENPEIAKTLDKKLREMLLNPENMQ-L---IAETSSAADDVEFGAVPEEF------------------
He.i   NCLVVFINQIRMKIGVM-FG--NPETTTGGNALKFYSSVRLDIRRTGSV-KDGE----NIIGNETRVKVVKNK--LAAPFRQVDFQILYGEGISKAGELLELGVKHKLVEKSGAWYSYNGEK--IGQGKANSMKWLNENIEKSDELEARLRAELVANPE-------QALMADIEQSENNTESESDFE-----------------
Ps.f   NCLVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRTGAV-KEGD----EVVGSETRVKVVKNK--VAPPFRQAEFQILYGKGIYLNGEMIDLGVLHGFVEKSGAWYAYNGSK--IGQGKANSAKFLADNPDIVATLEKQIRDKLLTPAP---D---VKAAANREPVEEVEEADTDI------------------
Ps.p   NCLVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRTGAV-KEGD----EVVGSETRVKIVKNK--VSPPFRQAEFQILYGKGIYRNGEIIDLGVSQGLVEKSGAWYAYQGNK--IGQGKANAAKYLAENPAIGAEIEKQIRDKLLTSGAVAAA---GKAAAVEADADDMADADAGY------------------
Ps.a   NCLVIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGAV-KEGD----EVVGSETRVKVVKNK--VSPPFRQAEFQILYGKGIYRTGEIIDLGVQLGLVEKSGAWYSYQGSK--IGQGKANAAKYLEDNPEIGSVLEKTIRDQLLAK-----------SGPVKADAEEVADAEAD-------------------
Az.v   NCLVIFINQIRMKIGVM-FG--NPETTTGANALKFYASVRLDIRRTGAV-KESD----EVIGSETRVKVVKNK--VAPPFRQAEFQILYGKGIYRNGEIIDLGVQLGLLEKSGAWYSYQGSK--IGQGKANAAKFLEDNPEVAAAVEKSIRDQLLAAP---------ASARPAALADEPADADLDY------------------
Le.p   NTLVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGSI-KKGE----EILGSETRVKVVKNK--VAPPFKMTEFDILYNEGISRESEIINLGVQLNLIEKSGAWYSYKQEK--IGQGKENVRLYLKENPQVAAELEQQIRTELLEKKLSVLA---SSSE----DLFETIDD----------------------
Acn.c  NCMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGQV-KEGD----EIIGSETKVKVVKNK--MAPPFREAIFQILYGKGVNQLGELVDLAVQQNIVQKAGAWYSYQGNK--IGQGKNNVIRYLEENPQISTEIEAVIREQLLTKAS-------DQTAAHDETEEEPDLLES--------------------
Bd.p   NCMVIFINQIRMKIGVM-FG--NPETTTGGNALKFYSSVRLDIRRIGAI-KKGD----EVVGNETRVKVVKNK--VAPPFKQAEFDIMYGSGISREGEIIDLGVQANVVDKSGAWYSYSGNR--IGQGKDNVREYLKEHKEMAIEIENKVRENQGIVSRAATF---PASE-------AEDGE----------------------
Bu.p   NCLVIFINQIRMKIGVM-FG--NPETTTGGNALKFYSSVRLDIRRIGSI-KKND----EVIGNETRVKVVKNK--VSPPFREAIFDILYGEGISRQGEIIDLGVQAKIVDKAGAWYSYNGEK--IGQGKDNAREFLRENPEIAREIENRIRESLGVVAMPDGA---GNEA-------EAMDEEE--------------------
Mm.c   NTLVIFINQIRMKIGVM-FG--NPETTTGGNALKFYSSVRLDIRRTGAI-KKGD----EVIGSETKVKVIKNK--VAPPFKQAEFDIMYGEGISRLGEIIELGTNLKLVEKSGAWYSYNGEK--IGQGKENAKEFLRENPAIAAEIEAKIRDNSNVLADSMTA---ARSE-------DD-------------------------
Mb.f   NTLVIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGAI-KKGD----EVTGSETRVKVVKNK--VAPPFKLAEFDILYGEGISREGEIIELGVNLKLIEKAGAWYSYKGEK--IGQGKDNAREFLREHPEIANEIDAKIREHSNLANAAMTT---APDE-------ESDE-----------------------
Tb.f   NTLVIFINQIRMKIGVM-YG--SPETTTGGNALKFYASVRLDIRRIGAI-KKSD----EVVGNDTRVKVVKNK--VAPPFREAEFAIYYGEGISRLSELVDLGVKFDIVEKSGAWYSYQGHR--IGQGKDNARQYLKVHPELAANIEQRIR-AAAAGHPLAFA---EEVESPQRSAS---------------------------
Acd.f  NKLVIFKKQIRMKMGV--YG--RPETTTGGNALKFYASVRLDIRRMGAM-KKSAT---KSYDWSTRVKVVKNK--VAPPFRQAELAIYYGEGIYRGSEPVDLGVKLENVEKSGGWYSYPGRR--IGQGKANARQYLRVKPEFPGIFEQGIRGAMAAPHPLGFG---ERRDVQQESGEPYGNNG---------------------
Ne.g   NTLVVFINQIRMKIGVM-FG--SPETTTGGNALKFYSSVRLDIRRTGSI-KKGE----EVLGNETRVKVIKNK--VAPPFRQAEFDILYGEGISWEGELIDIGVKNDIINKSGAWYSYNGAK--IGQGKDNVRVWLKENPEISDEIDAKIRALNGVEMHITEG---TQDETDGERPEE--------------------------
Rz.p   NTMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGSV-KERE----EVIGNQTRVKVVKNK--MAPPFKQVEFDIMYGEGVSKTGELVDLGVKAGIVEKSGAWFSYNSQR--LGQGGENAKTFLRDNPDLAREIELALRENAGLIADRFLQ---NGGPDADDGDG---ADM---------------------
Rz.v   NTMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGAV-KERE----EVIGNQTRVKVVKNK--MAPPFKQVEFDIMYGEGVSKTGELVDLGVKAGIVEKSGAWFSYNSQR--LGQGRENAKTFLRDNPDLAREIELSLRQNAGLIADRFLQ---NGGPDPDDGDGDATAEM---------------------
Rz.m   NCMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGSV-KERE----EVVGNQTRVKVVKNK--MAPPFKQVEFDIMYGEGVSKTGELIDLGVKAGIVEKSGAWFSYNSQR--LGQGRENAKLFLRENPELLREIETALRQNAGLIADRFLE---NGGPES-DGDE--AADM---------------------
Ag.t   KCMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGAV-KERE----EVVGNQTRVKVVKNK--MAPPFKQVEFDIMYGEGVSKTGELVDLGVKAGIVEKSGAWFSYNSQR--LGQGRENAKTFLRDNPDTANEIELALRQNAGLIADRFLQ---NGGPDAGEGDDG--SDEG--------------------
Br.a   NCMVIFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRLDIRRIGSI-KERD----EVVGNQTRVKVVKNK--LAPPFKQVEFDIMYGAGVSKVGELVDLGVKAGVVEKSGAWFSYNSQR--LGQGRENAKQYLKDNPEVAREIETTLRQNAGLIAEQFLD---DGGPEE---DAAGAAEM---------------------
Ma.m   KTIVIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRMEIRRVGAI-KDRD----EVVGNQTRVKVVKNK--LAPPFKVVDFDIMYGEGISKMGELIDLGVKANVVKKSGAWFSYNSTR--IGQGRENAKQFLRDNPAMAAEIEGAIRQNAGLISEALAA---VPDLDGTPVAE---------------------------
Act.p  NTMLIFLNQIRLKIGVM-FG--SPESTTGGNALKFYASVRMDIRRIGSI-KDKD----EVTGNQTRVKVVKNK--MAPPFRQVEFDIMYGEGISKVGELIDLGVKAGIVEKSGAWFSCDSQR--IGQGRENAKQFLRDHPEMAADIERRVREQAGVVAEAMLV---GPDEDGAEH-----------------------------
Rh.s   NCMVIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRLDIRRTGAI-KDRD----NVIGNTTKVKVVKNK--VAPPFREVEFDIMYGEGISKTGELVDLGVKAGVVEKSGSWYSYGDER--IGQGRENAKAFLRANPTVAGDIEDRIRASHGLDFSTGED---GKGDDLVDM-----------------------------
Rh.c   NCMVIFINQIRMKIGVM-FG--SPETTSGGNALKFYASVRLDIRRTGAI-KDRD----EVIGNQTRVKVVKNK--VAPPFREVEFDILYGEGISKVGELVDLGVKAGVVAKSGAWYSYGDER--IGQGRENAKQFLRDNPDIAYEIEDKIRASHGLEFGV--D---PTAEDLTEED----------------------------
Ri.p   NCITVFINQIRMKIGVM-FG--SPETTTGGNALKFYASVRIDIRRIGSI-KDKE----EVIGSQTKVKVVKNK--VSPPFKTADFDIMYGSGISKEGEIIDLGVKLEIIEKSGSWFSYNKIR--IGQGRENVKQYLKEHPQISNEIEKIIREKSSAITNINLD---QTEE----------------------------------
Mx.x1  GTCIIFINQIRMKIGVM-FG--NPETTTGGNALKFYASVRMEIRRTGNI-KDGD----AVVGSKARVKVVKNK--VAPPFQEAEFDLMYGSGIHRVGEVLDLGVATGLIEKSGSYFSLRGER--IGQGRERAAEWLREHPDVLEALGKEITG-----TSALPS------SPAPVEVAA--------------------------
Mx.x2  QTCVIFINQIRMKIGVM-FG--NPETTTGGNALKFYASQRLDIRRIGAI-KNGD----NVVGSRTRVKVVKNK--VAPPFKEVEFDIMYGTGISREGDLIDLASNENIVEKSGSWFSFNGER--IGQGRENVKEYLREHPEIAKDIEGRVLEKYGIGKSGAPSP-RRRTSPRRPKVAARSAAV---------------------
He.p   NTTLIFINQIRMKIGMMGYG--SPETTTGGNALKFYASVRIDIRRIATL-KQNE----QHIGNRAKAKVVKNK--VAPPFREAEFDIMFGEGISKEGEIIDYGVKLDIVDKSGAWLSYQDKK--LGQGRENAKALLKEDKALADEITLKIKESIGSNEEIM-----PLPDEPLEEME---------------------------
Ca.j   NTTVIFINQIRMKIGAMGYG--TPETTTGGNALKFYASVRLDVRKVATL-KQNE----EPIGNRVKVKVVKNK--VAPPFRQAEFDVMFGEGLSREGELIDYGVKLDIVDKSGAWFSYKDKK--LGQGRENSKAFLKENPEIADEITKAIQNSMGIEGMIS-----GSEDDEGEE-----------------------------
An.v   GCTVIFINQLRQKIGVT-YG--SPETTTGGNALKFYASVRLDIRRIQTL-KKGT----DEFGNRVKVKVAKNK--VAPPFRIAEFDIIFGKGVSTLGCLVDLAEETGILLRKGAWYSYNGDN--ISQGRDNAIKYLEEKPEFAEQIKQQVREKLDKGAVVSAN---SVAKANEEDEEDVDLDEEE-------------------
Sy.70  GCTVIFLNQLRQKIGIS-YG--NPEVTTGGTALKFYASVRLDIRRIQTL-KKGSE---GEFGIRAKVKVAKNK--VAPPFRIAEFDIIFGKGISRVGCMLDLAEQTGVITRKGAWYSYEGDN--IAQGRDNAVKYLEENPDVAAIVTQKVRENLDMSSMGFGD---EHHTTEEE------------------------------
Ar.t   GCTLIFLNQIRYKIGVY-YG--NPEVTSGGIALKFFASVRLEIRSAGKI-KSSKGD--EDIGLRARVRVQKSK--VSRPYKQAEFEIMFGEGVSKLGCVLDCAEIMEVVVKKGSWYSYEDQR--LGQGREKALQHLRENPALQDEIEKKVRLLMLDGEVHRST---PLMSSSSSSASHREEEEEDSLDDFQ-------------
Myp.m  NTTVIFINQLREKIGVI-FG--NPETTTGGKALKFFSSIRLEVRKAENI-LNNY----EIIGNKIKIKVVKNK--TAIPFKTTTISLLYNKGIDKLGELVDLLVSYEIIEKSGVWYSYQNEK--IGQGRTSVIQWLNADENRINELTEQVKKLIKQD-----------------------------------------------
Myp.p  KTSVIFINQVREKIGVV-FG--NPETTPGGRALKFYASIRLDVRKSTWI-MLNN----DISGNQIRVKVVKNK--LAPPFKIAETEIIFSKGINKFGEVADLALVHDVLQKKGAWFSYNGNN--IAQGRQKLIAQLESNNELFEEIFQKIVEKENQKLS---------------------------------------------
Ba.s   KTIAIFINQIREKVGVM-FG--NPETTPGGRALKFYSSVRLEVRRAEQL-KQGN----DVMGNKTKIKVVKNK--VAPPFRTAEVDIMYGEGISKEGEIIDLGTELDIVQKSGSWYSYEEER--LGQGRENAKQFLKENKDIMLMIQEQIREHYGLDNNGV------VQQQAEETQEELEFEE---------------------
Sta.a  NTTAIFINQIREKVGVM-FG--NPETTPGGRALKFYSSVRLEVRRAEQL-KQGQ----EIVGNRTKIKVVKNK--VAPPFRVAEVDIMYGQGISKEGELIDLGVENDIVDKSGAWYSYNGER--MGQGKENVKMYLKENPQIKEEIDRKLREKLGISDG--------DVEETEDAPKSLFDEE---------------------
Stc.p  KTIAIFINQLREKVGVM-FG--NPETTPGGRALKFYASVRLDVRGNTQI-KGTGDQKETNVGKETKIKVVKNK--VAPPFKEAVVEIMYGEGISKTGELLKIASDLDIIKKAGAWYSYKDEK--IGQGSENAKKYLAEHPEIFDEIDKQVRSKFGLIDGEE------VSEQDTENKKDEPKKEEAVNEEVPLDLGDELEIEIEE
La.la  KTTAIFINQLREKVGVM-FG--SPETTPGGRALKFYASVRLDVRGSTKI-EEGSGDNKTQIGKITKIKVVKNK--VAPPFKVALVDIMFGEGISSTGELLNIAVEEGIIKKSGAWFAYNDEK--IGQGAEKAKNYLKEHQDVFDEIDHKVRAAHGLLDDSE------VAETEEETTAFKN------------------------
Ac.l   NVVAIFINQIREKVGVM-FG--NPETTPGGRALKFFSSVRLEIRRAEAI-KQGS----EMIGIKSNVKVVKSK--VAPPLKTASIDIMYGTGISRSGEVLDLSVELNLVNKSGAWYNIGEEK--LGQGRDNAKQYLEDKPELLNELEKKVRTHFKLTK----------------------------------------------
Stm.l  KTTAIFINQLREKIGVM-FG--SPETTTGGRALKFYASVRLDIRRIETL-KDGT----DAVGNRTRVKVVKNK--VAPPFKQAEFDILYGQGISREGGLIDMGVENGFVRKAGAWYTYEGDQ--LGQGKENARNFLKDNPDLANEIEKKIKQKLGVGVHPEE---SATEPGADAASAAPADAAPAVPAPTTAKATKSKAAAAKS
Stm.a  KTTAIFINQLREKIGVM-FG--SPETTTGGRALKFYASVRLDIRRIETL-KDGT----DAVGNRTRVKVVKNK--VAPPFKQAEFDILYGQGISREGGLIDMGVEHGFVRKAGAWYTYEGDQ--LGQGKENARNFLKDNPDLANEIEKKIKEKLGVGVRPEEP--TATESGPD---AATAESAPAVPAPATAKVTKAKAAAAKS
Stm.v  KTTAIFINQLREKIGVM-FG--SPETTTGGRALKFYASVRLDIRRIETL-KDGT----DAVGNRTRVKVVKNK--VAPPFKQAEFDILYGQGISREGGLIDMGVEHGFVRKAGAWYTYEGDQ--LGQGKENARNFLKDNPDLADEIERKIKEKLGVGVRPDAAKAEAATDAAAADTAGTDDAAKSVPAPASKTAKATKATAVKS
Co.g   GTTAIFINQLREKIGVM-FG--SPETTTGGKALKFYASVRCDIRRIQTL-KDGQ----DAIGNRTRLKVVKNK--VSPPFKIAEFDIMYGEGISRESSVIDLAVDNGIVKKSGSWFTYEGEQ--LGQGKEKVRLSLKENPELTDELEDKIFKKLGVGKYAAAS-DELTDDPVELVPNVDFDDEADTEADAED------------
Myb.t  GTTAIFINQLRDKIGVM-FG--SPETTTGGKALKFYASVRMDVRRVETL-KDGT----NAVGNRTRVKVVKNK••CSPPFKQAEFDILYGKGISREGSLIDMGVDQGLIRKSGAWFTYEGEQ--LGQGKENARNFLVENADVADEIEKKIKEKLGIGAVVTD--DPSND-GVLPAPVDF-------------------------
Myb.l  GTTAIFINQLREKIGVM-FG••SPETTTGGKALKFYASVRMDVRRIETL-KDGV----DAVGNRTRVKIVKNK--VSPPFKQAEFDILYGKGISREGSLIDMGVEQGFVRKSGSWFTYEGEQ--LGQGKENARNFLLENADVANEIEKKIKEKLGIG-------AVVTDDDILPTP-VDF------------------------
Ch.t   NTCAIFINQIREKIGVS-FG--NPETTTGGRALKFYSSIRIDIRRIGSI-KGGEN---FDIGNRIKVKVAKNK--LAPPFRTAEFDILFNEGISSAGCIIDLAVEKNIIDKKGSWFNYQDRK--LGQGREAVREELKRNKELFHELERRIYESVQASQAPA------AACVDSESREVAEAAK---------------------
Bct.f  RTTCIFINQLREKIGVM-FG--NPETTTGGNALKFYASVRLDIRGSQQI-KDGE----EVIGKQTKVKVVKNK--VAPPFRKAEFDIMFGEGISHSGEIIDLGADLGIIKKSGSWYSYNDTK--LGQGRDAAKQCIADNPELAEELEGLIFEKLREHK----------------------------------------------
Tg.m   KAVVIFTNQIRMKIGVM-FG--SPETTTGGLALKFYATMRMEVRR-GEPIKEGK----DVIGNVISVKIVKNK--VAPPFKTAQTYIIYGKGIDREYELFNIAVNEGIVDRKGSWYYYTTLK••LGQGSSNAVQFLKDNPEIAGEIERRIREKYGLLSVEKEE----QRKEKKSSGEEAS------------------------
De.r   GTAAIFINQVREKIGVM-YG--NPETTTGGRALKFYASVRLDVRKIGQPTKVGN----EAVANAVKIKTVKNK--VAAPFKEVELALVYGKGFDQLSDLVGLAADMDIIKKAGSFYSYGDER--IGQGKEKTIAYIAERPEMEQEIRDRVMAAIRAGNAGEAPA-LAPAPAAPEAAEA--------------------------
Th.a   NTAAIFINQVREKVGVM-YG--NPETTPGGRALKFYSSVRLDVRKSGQPIKVGN----EAVGIKVKVKVVKNK--LAPPFREAELEIYFGRGLDPVMDLVNVAVAAGVIEKAGSWFSYGEHR--LGQGKEKAAEYLRERPELLEEIRAKVLERA------DKVV-LAAGEEEGE------------------------------
Th.t   NTAAIFINQVREKVGVT-YG--NPETTPGGRALKFYASVRLDVRKSGQPIKVGN----EAVGVKVRVKVVKNK--LAPPFREAELEIYFGRGLDPVADLVNVAVAAGVIEKAGSWFSYGELR--LGQGKEKAAEALRERPELLEEIRAKVLERS------DQVV-LAAGEDEGE------------------------------
Aq.p   NTALIFINQIREKIGVM-FG--NPETTPGGRALKFFSDMRLEVRRLGDV-KEGG----EKKGYRVKVRVVKNK--LAPPFQEAEFDVIYGEGICRICDIIDTAANLGVITKSGSWYSYGEKR--LGQGREQAKKYLLEHPEMLEEIERKVREVS------GLVR-PDTENSVGEKSE----------------------------
           .*  *.* ..*    *   ** * *. ****..  * ..*                       .  *       * .     . .  *                .   *           . **
Mask   111111111111111110110011111111111111111111111111100000000011111111111111100111111111111111111111111111111111111111111111100111111111111111111111111111110000000000000000000000000000000000000000000000000
```

Figure 2. Comparison of consensus trees for RecA and SS-rRNA.

Strict-rule consensus trees representing the phylogenetic patterns found in all trees generated by multiple methods for each molecule are shown. The RecA consensus (A) was generated from the PAUP, *protpars*, Fitch-Margoliash, De Soete and neighbor-joining trees (see Methods). The SS-rRNA consensus (B) was generated from the *dnapars*, Fitch-Margoliash, De Soete and neighbor-joining trees. Comparable species are aligned in the middle and species are ordered to minimize branch crossing (note two crossed branches in SS-rRNA tree). Consensus clades are shaded for each molecule.

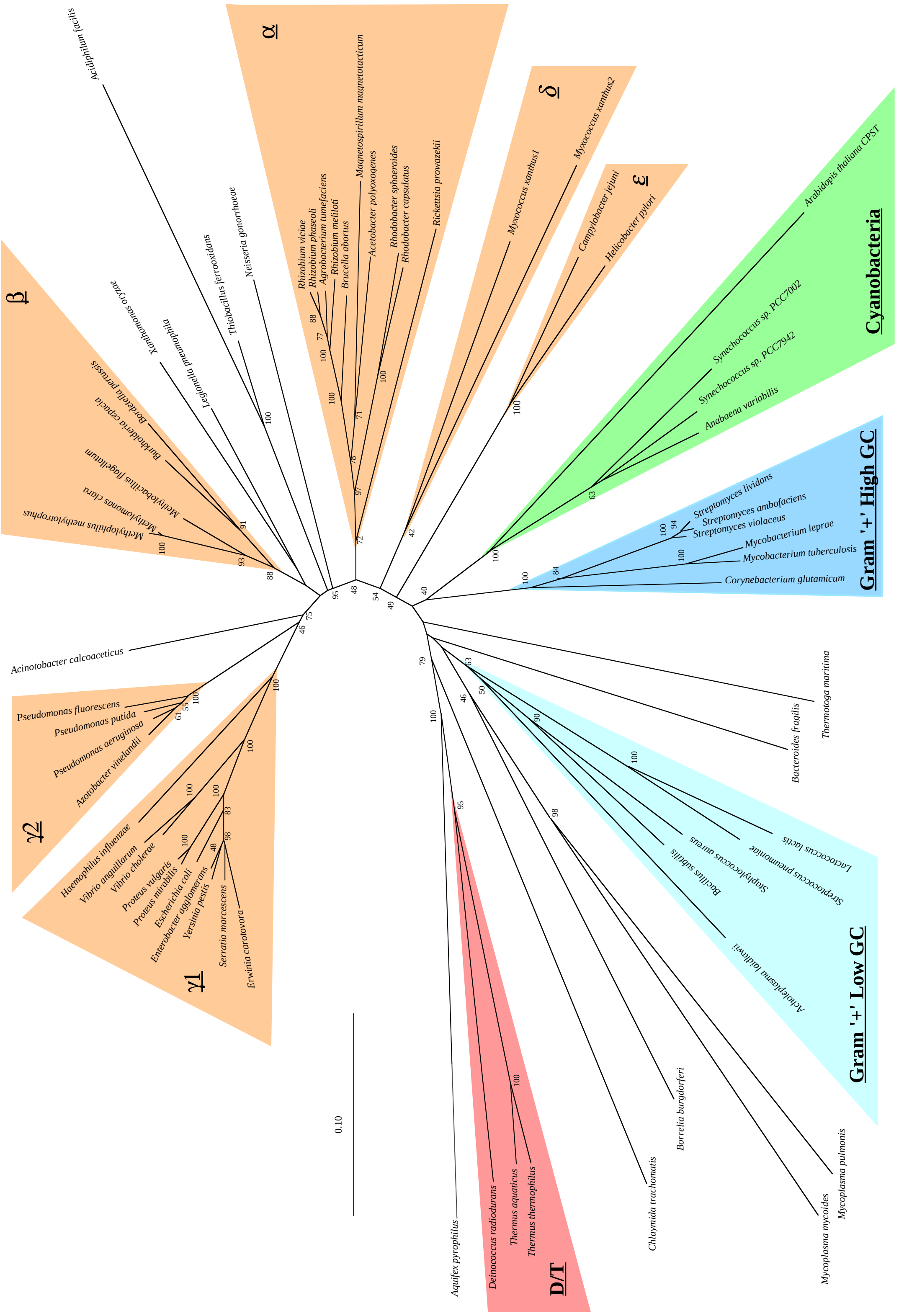


2a) RecA

2b) SS-rRNA

Figure 3. Fitch-Margoliash trees for RecA (A) and SS-rRNA (B).

Trees were generated from the multiple sequence alignments by the method of Fitch and Margoliash. Regions of ambiguous alignment and indels were excluded from the analysis (see Methods). For the RecA tree, distances were calculated using the *protdist* program of PHYLIP with a PAM-matrix based distance correction. For the SS-rRNA tree, distances were calculated using the *dnadist* program of PHYLIP and the Kimura-2-parameter distance correction. Consensus clades representing groups found in all phylogenetic methods are highlighted. Branch lengths and scale bars correspond to estimated evolutionary distance. Bootstrap values when over 40 are indicated.


α
Acidiphilium facilis
Thiobacillus ferrooxidans
Neisseria gonorrhoeae
Rhizobium viciae
Rhizobium phaseoli
Agrobacterium tumefaciens
Rhizobium meliloti
Brucella abortus
Magnetospirillum magnetotacticum
Acetobacter polyoxogenes
Rhodobacter sphaeroides
Rhodobacter capsulatus
Rickettsia prowazekii
δ
Myxococcus xanthus1
Myxococcus xanthus2
ε
Campylobacter jejuni
Helicobacter pylori
Cyanobacteria
Arabidopsis thaliana CPST
Synechococcus sp. PCC7002
Synechococcus sp. PCC7942
Anabaena variabilis
Gram '+' High GC
Streptomyces lividans
Streptomyces ambofaciens
Streptomyces violaceus
Mycobacterium leprae
Mycobacterium tuberculosis
Corynebacterium glutamicum
Thermotoga maritima
Bacteroides fragilis
Gram '+' Low GC
Lactococcus lactis
Streptococcus pneumoniae
Staphylococcus aureus
Bacillus subtilis
Acholeplasma laidlawii
Mycoplasma pulmonis
Mycoplasma mycoides
Borrelia burgdorferi
Chlaymida trachomatis
D/T
Thermus thermophilus
Thermus aquaticus
Deinococcus radiodurans
Aquifex pyrophilus
0.10
γ1
Erwinia carotovora
Serratia marcescens
Yersinia pestis
Enterobacter agglomerans
Escherichia coli
Proteus mirabilis
Proteus vulgaris
Vibrio cholerae
Vibrio anguillarum
Haemophilus influenzae
γ2
Azotobacter vinelandii
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas fluorescens
Acinotobacter calcoaceticus
β
Methylophilus methylotrophus
Methylomonas clara
Methylobacillus flagellatum
Burkholderia cepacia
Bordetella pertussis
Xanthomonas oryzae
Legionella pneumophila

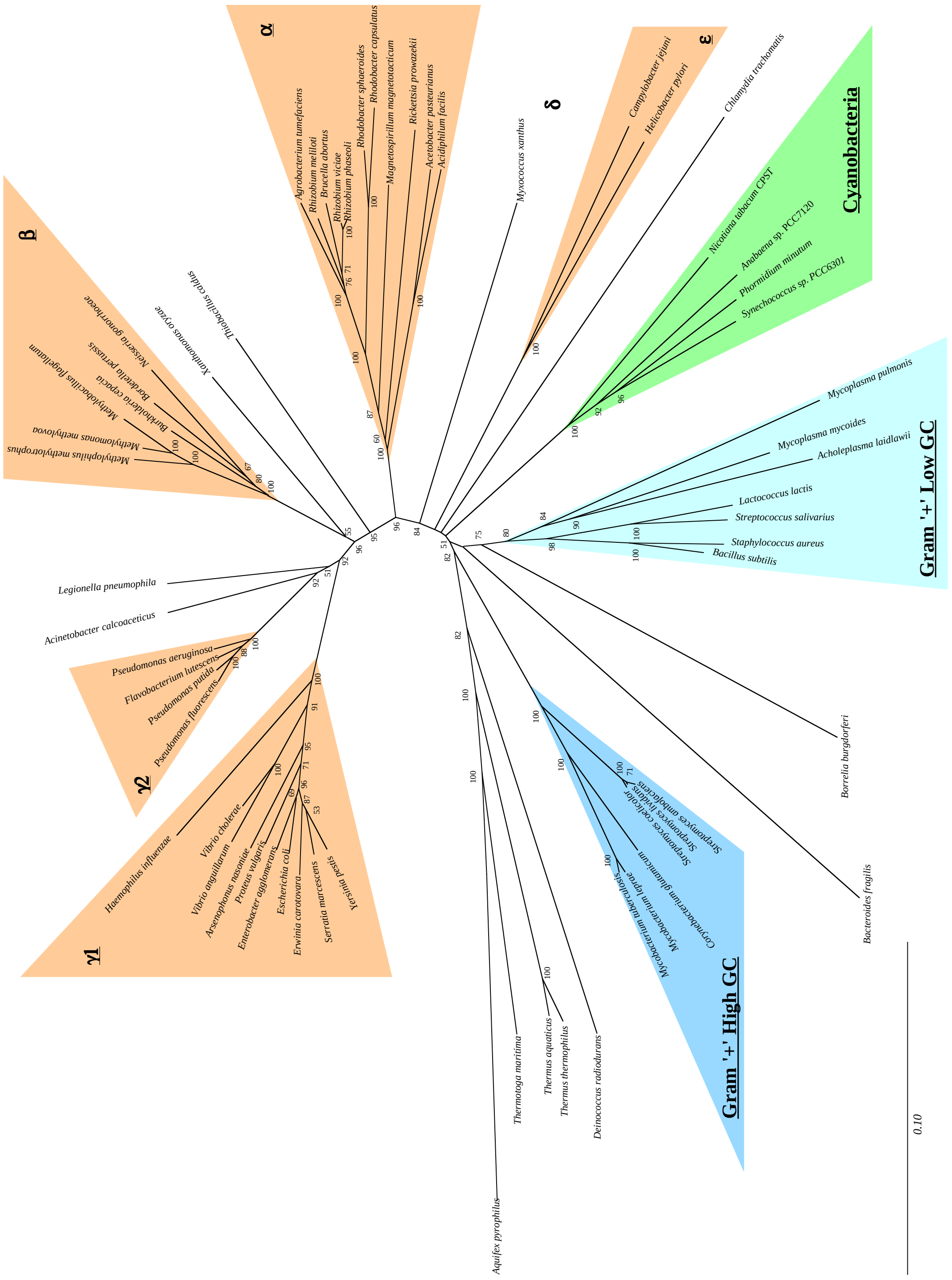
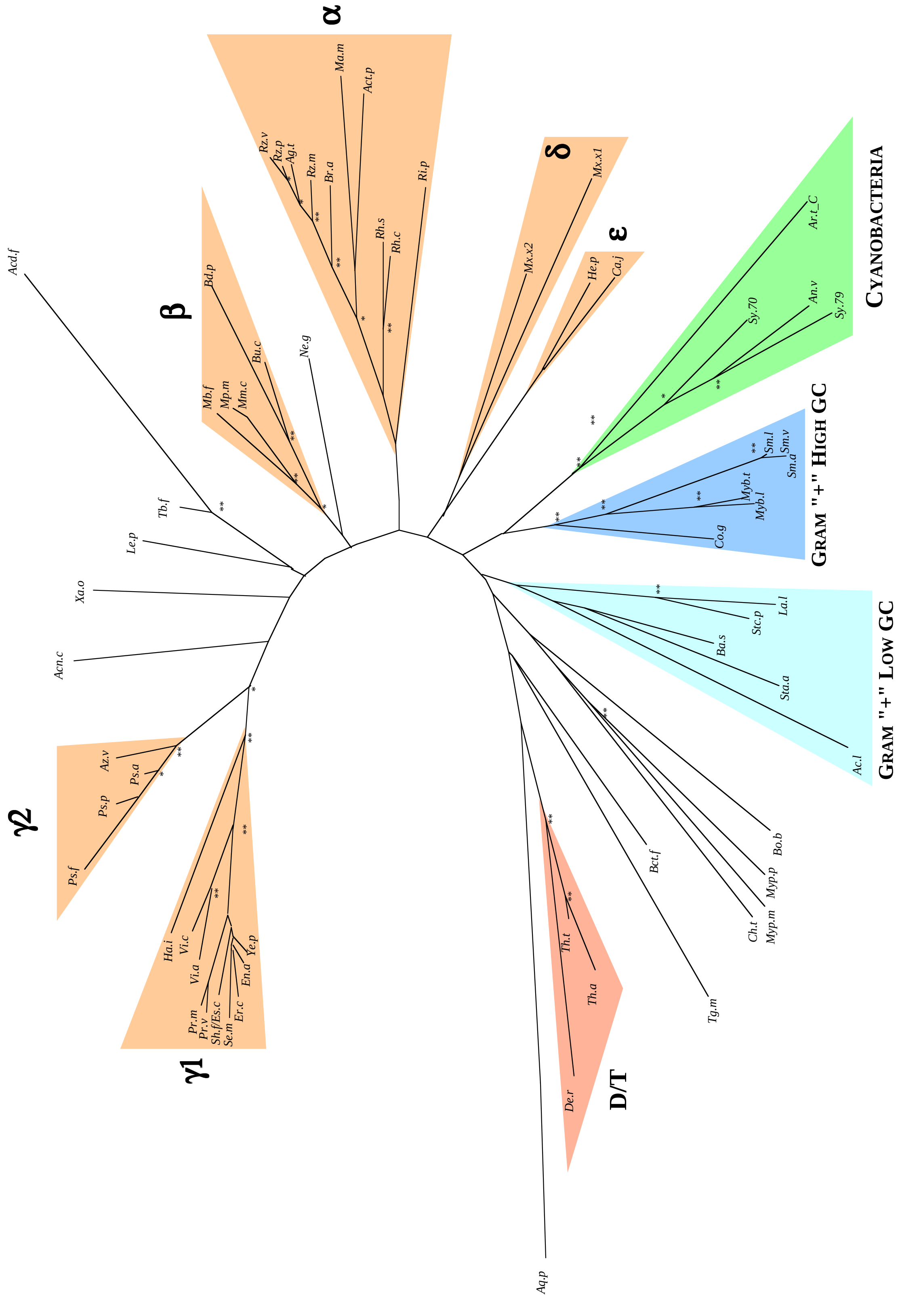

α
Agrobacterium tumefaciens
Rhizobium meliloti
Brucella abortus
Rhizobium viciae
Rhizobium phaseoli
Rhodobacter sphaeroides
Rhodobacter capsulatus
Magnetospirillum magnetotacticum
Rickettsia prowazekii
Acetobacter pasteurianus
Acidiphilium facilis
β
Neisseria gonorrhoeae
Bordetella pertussis
Burkholderia cepacia
Methylobacillus flagellatum
Methylomonas methylovora
Methylophilus methylotrophus
Thiobacillus caldus
Xanthomonas oryzae
δ
Myxococcus xanthus
ε
Campylobacter jejuni
Helicobacter pylori
Chlamydia trachomatis
Cyanobacteria
Nicotiana tabacum CPST
Anabaena sp. PCC7120
Phormidium minutum
Synechococcus sp. PCC6301
Gram '+' Low GC
Mycoplasma pulmonis
Mycoplasma mycoides
Acholeplasma laidlawii
Lactococcus lactis
Streptococcus salivarius
Staphylococcus aureus
Bacillus subtilis
Borrelia burgdorferi
Bacteroides fragilis
Gram '+' High GC
Streptomyces ambofaciens
Streptomyces lividans
Streptomyces coelicolor
Corynebacterium glutamicum
Mycobacterium leprae
Mycobacterium tuberculosis
Deinococcus radiodurans
Thermus thermophilus
Thermus aquaticus
Thermotoga maritima
Aquifex pyrophilus
Legionella pneumophila
Acinetobacter calcoaceticus
γ2
Pseudomonas aeruginosa
Flavobacterium lutescens
Pseudomonas putida
Pseudomonas fluorescens
γ1
Haemophilus influenzae
Vibrio cholerae
Vibrio anguillarum
Arsenophonus nasoniae
Proteus vulgaris
Enterobacter agglomerans
Escherichia coli
Erwinia carotovora
Serratia marcescens
Yersinia pestis
0.10

Figure 4. RecA parsimony tree.

Tree was generated from the multiple sequence alignments using the protpars method of the program Phylip according to the methods described in the text. Regions of ambiguous alignment and indels were excluded from the analysis (see Methods). Consensus clades representing groups found in all phylogenetic methods are highlighted. Branch lengths and scale bars correspond to estimated number of amino-acid substitutions. Bootstrap values when over 70 are indicated by ** and when between 40 and 70 by *. Not published in the Journal of Molecular Evolution article.


α
Ma.m
Act.p
Rz.v
Rz.p
Ag.t
Rz.m
Br.a
Rh.s
Rh.c
Ri.p
β
Bd.p
Bu.c
Mb.f
Mp.m
Mm.c
Ne.g
Acd.f
Tb.f
Le.p
Xa.o
Acn.c
δ
Mx.x1
Mx.x2
ε
He.p
Ca.j
CYANOBACTERIA
Ar.t_C
Sy.70
An.v
Sy.79
GRAM "+" HIGH GC
Sm.l
Sm.v
Sm.a
Myb.t
Myb.l
Co.g
GRAM "+" LOW GC
La.l
Stc.p
Ba.s
Sta.a
Ac.l
Bo.b
Myp.p
Myp.m
Ch.t
Bct.f
Tg.m
D/T
Th.t
Th.a
De.r
Aq.p
γ2
Az.v
Ps.a
Ps.p
Ps.f
γ1
Ha.i
Vi.c
Vi.a
Pr.m
Pr.v
Sh.f/Es.c
Se.m
Er.c
En.a
Ye.p

PART B

The Phylogenetic Relationships of *Chlorobium tepidum* and *Chloroflexus aurantiacus* Based upon their RecA Sequences[4]

[4] Originally appeared as *Tanja M. Gruber, Jonathan A. Eisen, Kurt Gish, Donald A. Bryant. 1998. FEMS Microbiology Letters 162: 53-60.* Reprinted with permission of Elsevier Science.

ABSTRACT

Using RecA as the phylogenetic marker, the relationships of the green sulfur bacterium *Chlorobium tepidum* and the green gliding bacterium *Chloroflexus aurantiacus* to other eubacteria were investigated. The *recA* genes of the two organisms were cloned, and the resulting protein sequences aligned with 86 other eubacterial RecA sequences. *Cb. tepidum* was placed as the nearest relative to the *Cytophaga/ Flexibacter/Bacteroides* group, a relationship supported by results obtained with several phylogenetic markers. *Cf. aurantiacus* was placed near *Chlamydia trachomatis* and the high-GC gram-positives; however, this placement was not strongly supported statistically. Possible reasons for this ambiguity are discussed.

INTRODUCTION

The green sulfur bacteria (also called Chlorobiaciae) and the green gliding bacteria comprise a relatively small number of identified genera which have not been exhaustively characterized phylogenetically. Using sulfide or sulfur as electron donors, the green sulfur bacteria are obligately anaerobic and photolithoautotrophic. Recently, SS-rRNA sequences from 18 strains belonging to the genus *Chlorobium* were analyzed to study their phylogenetic relationships [1]. The *Chlorobium* sp. analyzed were all very closely related to each other, and *Cb. tepidum*, the only known thermophile of the genus *Chlorobium*, was found to be placed near the *Chlorobium limicola* cluster. The thermophilic nature of *Cb. tepidum* is believed to result from rapid, divergent evolution rather than from inherited growth characteristics derived from an ancestral thermophilic relative [1].

The green gliding bacteria are composed of both photosynthetic and non-photosynthetic members. The photosynthetic thermophile *Chloroflexus aurantiacus* is the best characterized member of the green gliding bacteria. This organism is very interesting from an evolutionary perspective due its combination of characteristics that are found in very different and diverse groups of phototrophic prokaryotes [2]. The photosynthetic green gliding bacteria have chlorosomes as their light-harvesting antenna system, like the green sulfur bacteria [3]. However, their reaction centers are similar to those of the photosynthetic Proteobacteria and to photosystem II of the cyanobacteria [4], and differ significantly from those of the green sulfur bacteria, the heliobacteria, and photosystem I of cyanobacteria. The overall cell morphology, carotenoid composition, and mat-forming behavior resemble certain cyanobacteria [5]. *Cf. aurantiacus* also displays some features which are unique among autotrophs, such as its autotrophic CO_2 fixation mechanism by the 3-hydroxypropionate pathway [6].

The RecA protein in *E. coli* takes part in a number of cellular processes, among them homologous DNA recombination, SOS induction, and DNA-damage-induced mutagenesis [7]. Although the RecA protein sequence and function is highly conserved within bacteria, it is not absolutely essential for cell survival in most organisms. Related proteins have also been found in Archaea and eukaryotes [8]. Eisen [9] has shown that RecA comparisons are informative in studies of molecular systematics of bacteria. The molecule fulfills a number of criteria that make it a useful marker for phylogenetic analyses. Some of these are: the molecule is of reasonable size, thus allowing statistical analyses to be performed; some regions of RecA are conserved between species and other regions are highly variable, thus allowing comparisons between both close and distant relatives; and the gene is relatively easily cloned.

In this work the *recA* genes of *Cb. tepidum* and *Cf. aurantiacus* have been cloned and analyzed phylogenetically. It was our goal to use RecA as a marker to examine specifically the placement of these two phylogenetically ambiguous phyla within the eubacterial kingdom. Furthermore, we updated the previous phylogenetic tree derived from RecA sequences [9] by including 26 additional eubacterial RecA sequences.

MATERIALS AND METHODS

Recombinant DNA procedures

Chlorobium tepidum was kindly provided by Dr. Michael Madigan (Southern Illinois University, Carbondale, IL), and *Chloroflexus aurantiacus* J-10-fl kindly provided by Dr. Beverly Pierson (University of Puget Sound, Tacoma, WA). Total chromosomal DNA from *Cb. tepidum* and *Cf. aurantiacus* was isolated as described [10] with the inclusion of a CTAB (hexadecyl-trimethylammonium bromide) extraction. Clones containing the *recA* genes were isolated by using size-directed plasmid libraries as described [11]. DNA sequences were determined by the dideoxy chain termination method [12], with the Sequenase Version 2.0 DNA sequencing kit from U. S. Biochemical (Cleveland, OH) or were determined by an automated sequencer (Perkin-Elmer/Applied Biosystems, Foster City, CA). Oligonucleotides for sequencing were synthesized on a Model 392 Applied Biosystems (Foster City, CA) automated DNA/RNA synthesizer. Oligonucleotides for PCR were obtained from Genset Corporation. Sequence data were analyzed with MacVector Sequence Analysis Programs Version 6.0 (Eastman-Kodak, Rochester, NY).

Cloning of recA genes

Degenerate PCR primers were synthesized that span conserved regions of the RecA protein corresponding to amino acids 91-101 (primer sequence is 5' GCITTYRTIGAYGCIGARCAYGCIYTIGAYCC 3') and amino acids 206-212 (primer sequence 5' CCICCIGKIGTIGTRTCIGG 3') of *E. coli* [9]. The resulting PCR products were used as hybridization probes to obtain genomic clones containing the complete *recA* genes. Southern blots with digests of *Cb. tepidum* and *Cf. aurantiacus* chromosomal DNAs were hybridized with the respective PCR products. Based on these hybridization experiments, a 3.2 kb *Eco*RI fragment of *Cb. tepidum* was cloned (see Fig. 1) to obtain the entire sequence of the *recA* gene. A portion of the *Cf. aurantiacus recA* gene was initially cloned on a 1.8 kb *Hin*cII fragment; subsequently, a 0.5 kb *Kpn*I-*Hin*cII fragment was cloned to obtain the remaining coding region of the gene (see Fig. 1). The DNA sequences for the *Cb. tepidum recA* gene and the *Cf. aurantiacus recA* gene have been deposited in GenBank under the accession numbers AF037259 and AF037258, respectively.

Phylogenetic analyses

In 1995 Eisen [9] aligned and analyzed phylogenetically 65 RecA sequences.

Since then, 26 new sequences of RecA have been identified and deposited in the databases. These 26 sequences, as well as the sequences of *Cb. tepidum* and *Cf. aurantiacus,* have been added to most of the alignment obtained previously [9] . The alignment is available at http://www-leland.stanford.edu/~jeisen/RecA/ RecA.Alignment.html. The phylogenetic tree was generated using algorithms available from the PHYLIP software package [13] . The pairwise distances between the RecA proteins were calculated with the *protdist* program in PHYLIP, using the PAM matrix-based distance correction [13] . The tree was generated by the neighbor-joining methods [14] as implemented in PHYLIP. Bootstrap replicates were carried out 100 times by the method of Felsenstein [15] .

RESULTS AND DISCUSSION

As shown in Fig. 1, the *recA* gene of *Cb. tepidum* is flanked upstream by a gene with significant sequence similarity to dihydroflavonol-4-reductase (*dfr*) and downstream by genes with significant sequence similarity to the nitrogen regulatory gene *nifR3* and to aspartate semialdehyde dehydrogenase (*asd*). The *Cb. tepidum recA* gene predicts a protein of 346 amino acids with a predicted molecular mass of 37.1 kDa. No sequences with significant similarity to genes in the databases were identified downstream from the *recA* gene of *Cf. aurantiacus*; the *recA* gene of this bacterium predicts a protein of 351 amino acids with a predicted mass of 37.8 kDa. The deduced protein sequences were aligned with 86 other RecA sequences obtained from the databases. Figure 2 shows the phylogenetic tree obtained for the RecA sequences, using the procedures described in the Materials and Methods.

Green sulfur bacteria

Using RecA as the phylogenetic marker (Fig. 2), *Cb. tepidum* is placed as the closest relative to the *Cytophaga/Flexibacter/Bacteroides* group. Based on the close relationship of the identified green sulfur bacteria among themselves [1] , it can be assumed that the entire group will be closely related to the

Cytophaga/Flexibacter/Bacteroides group. The relationship between *Cb. tepidum* and the *Cytophaga/Flexibacter/Bacteroides* group is very highly supported in the RecA tree (bootstrap value of 100%), although the position of this entire clade within the tree lacks statistical significance. The green sulfur bacteria have also been placed as the nearest relatives to the *Cytophaga/Flexibacter/ Bacteroides* group based on SS-rRNA data [16,17] . This association has been confirmed by further studies that included representative members of these two phyla, by using the ATP-synthase β subunit and EF-Tu as markers [18] . In analyses using sigma factors as the phylogenetic marker, the green sulfur bacteria are seen to be most closely related to the green gliding bacteria [Gruber and Bryant, submitted]; however, this study did not include any sequence from the *Cytophaga/Flexibacter/Bacteroides* group. Thus, there appears to be a consensus among a range of phylogenetic markers that the green sulfur bacteria and the *Cytophaga/Flexibacter/Bacteroides* group are close relatives.

All green sulfur bacteria described so far have similar physiological characteristics. All are strictly anaerobic, are obligately phototrophic, and can use carbon dioxide as the sole carbon source [19] . All species can use sulfide, which is oxidized to sulfate with the intermediate accumulation to elemental sulfur globules outside the cells, as the electron donor for growth. The *Cytophaga/Flexibacter/Bacteroides* group is composed of a mixture of physiological types [20] . The *Bacteroides* sp. are obligately anaerobic and primarily fermentative organisms, while the *Cytophaga* and *Flexibacter* sp. are heterotrophic gliding bacteria.

Green gliding bacteria

The green gliding bacteria are phylogenetically confusing organisms, and the best studied member of these organisms, *Cf. aurantiacus*, has even been termed a 'chimeric organism' due to its unique set of phenotypic properties [21] . In the present study using RecA as the marker, the position of *Cf. aurantiacus* is unfortunately not further clarified. Although *Cf. aurantiacus* is placed nearest *Chlamydia trachomatis* and the high-GC gram-positives, it is also fairly closely related to *Cb. tepidum* (Fig. 2). However, the bootstrap values supporting all of these relationships are low. In earlier studies using SS-rRNA as the marker, the green gliding bacteria were placed as the closest relatives to the

Deinococcus/*Thermus* group [17] , whereas later studies showed the group to be positioned between the Thermotogales and the Planctomycetales [16] . In both analyses based upon SS-rRNA, the green gliding bacteria were observed to diverge very early within the eubacterial line of descent. In an analysis using EF-Tu as the marker [18] , the green gliding bacteria were placed between the *Deinococcus*/*Thermus* branch and the branch composed of the green sulfur bacteria and the *Cytophaga/Flexibacter/Bacteroides*. In the study using sigma factors, *Cf. aurantiacus* is the closest relative to *Cb. tepidum*, an association that is fairly well supported statistically [Gruber and Bryant, submitted]. In a study using reaction center proteins as phylogenetic markers [21] , it was unequivocally shown that the type II reaction center of *Cf. aurantiacus* is most closely related to the reaction centers of the photosynthetic Proteobacteria. The closest relative of *Cf. aurantiacus* using reaction center proteins as markers was *Rhodopseudomonas viridis* [21]. The reaction center of green gliding bacteria also shows very distant similarity to Photosystem II of cyanobacteria, whereas green sulfur bacteria and heliobacteria have reaction centers which share a closer evolutionary relationship with Photosystem I of cyanobacteria [22] . There is also evidence that the membrane-bound bacteriochlorophyll *a*-containing antenna complexes and the membrane-bound cytochrome that donates electrons to the reaction center are similar in the photosynthetic Proteobacteria and *Cf. aurantiacus* [21] . Since none of the phylogenetic markers employed to date suggest a close relationship between the Proteobacteria and *Cf. aurantiacus*, the simplest explanation for these observed differences is that a lateral gene transfer event may have been responsible for the transfer of these photosynthetic genes into or out of an ancestor of *Cf. aurantiacus*. In fact, in the photosynthetic Proteobacteria these genes are clustered as a 46 kb region [23] , and a similar clustering of photosynthetic genes has also recently been demonstrated in *Heliobacillus mobilis* (Dr. C. Bauer, personal communication). Such clustering of photosynthetic genes has not been observed in *Cf. aurantiacus* [24] , *Cb. tepidum* [24] , or cyanobacteria [25,26] . Sequence comparisons of some of the proteins found in the chlorosomes of green sulfur bacteria and green gliding bacteria indicate an evolutionary relatedness, albeit limited, between some proteins of the antenna complexes of the two groups [27] .

In the present study, *Cf. aurantiacus* is the only member of the green gliding bacteria included. Before any consensus regarding the placement of the green gliding bacteria in phylogenetic trees can be established, it will be necessary to include more examples of the green gliding bacteria, both photosynthetic and non-photosynthetic, in the phylogenetic analyses using various markers. Since the photosynthetic apparatus appears to be chimeric (or alternatively gave rise to different lineages through divergence), it seems as if the components of the photosynthetic apparatus are not appropriate markers to represent the organism as a whole in comparative phylogenetic studies. Different components would give very different results; and although this is very intriguing, it would not be particularly informative. It would be very interesting to include non-photosynthetic green gliding bacteria in studies using the sigma factor marker. *Cf. aurantiacus* is now the sole representative [Gruber and Bryant, submitted] in such analyses, and it should be possible to determine if the group would still be most closely related to the green sulfur bacteria and the cyanobacteria. The placement of this group would probably show much better statistical support if more sequences were included. Due to the apparent chimeric nature of *Cf. aurantiacus*, it would be very useful to obtain the sequence of the entire genome to examine completely the origins of this sequence diversity. Such an analysis could help to define specifically the phylogeny of green gliding bacteria, but might also provide more clues to the origins and evolution of this puzzling genome.

Other phyla

A detailed description and discussion of the use of RecA as a phylogenetic marker can be found in Eisen [9] . More than half of the species displayed in the tree (Fig. 2) belong to Proteobacteria phylum. Of the 26 newly added species to the alignment, 14 represent Proteobacterial species. The general relationships in the phylum have not changed significantly due to these additions, and the five distinctive subgroups (α, β, γ, δ, and ε) are maintained. Four gram-positive organisms have been added, and it is notable that the gram-positives still do not form a monophyletic clade, as was observed in the previous study using RecA [9] . Although the high-GC gram-positives cluster together, the low-GC gram-positives do not cluster with the high-GC gram positives and are not themselves monophyletic either. *Clostridium perfringens* is not placed within the gram-positive bacteria in the RecA tree, whereas this organism is placed within the low-GC gram positive organisms based on SS-rRNA data [17] . Although *C. perfringens* was not included in studies using sigma factors as the marker, based on this marker the gram-positive bacteria clearly constitute a monophyletic clade that is statistically highly supported [Gruber and Bryant, submitted]. The cyanobacteria also form a coherent group, with the nuclear encoded chloroplast RecA from *Arabidopsis thaliana* falling within this group. The closest group to the cyanobacteria is the spirochetes, although this association is statistically very weakly supported. The RecAs of *Deinococcus radiodurans* and the two *Thermus* species form a well supported group, which is placed near the Aquificales (as seen previously [9]). A number of phyla are presently represented by only one or two sequences (e.g., the thermotogales, chlamydia, green-gliding bacteria, and green sulfur bacteria) which contributes to the inability to produce a meaningful and precise placement of some of these organisms within the tree. This is reflected by relatively low bootstrap values for most of these species. The placement of these groups should increase in statistical significance once more RecA sequences have been analyzed from such groups.

Conclusions

The position of *Cb. tepidum*, the representative of the green sulfur bacteria used in these studies, as the closest relatives to the *Cytophaga/Flexibacter/ Bacteroides* group, is very well supported in the phylogenetic tree based on RecA sequences. This result is consistent with a number of other analyses using several phylogenetic markers. On the other hand, the position of the green gliding bacterium *Cf. aurantiacus* remains highly ambiguous, and at the present time no definite conclusions regarding the relationship of this organism to other eubacteria can be drawn from the analyses using RecA or other phylogenetic markers.

ACKNOWLEDGMENTS

This work was supported by USPHS grant GM-31625 and DOE grant DE-FG02-94ER20137 to D.A.B. In addition, J.A.E. was supported in part by Outstanding Investigator Grant CA44349 from the National Cancer Institute to Philip C. Hanawalt. We would also like to thank Alberto I. Roca for help in tracking down unpublished RecA sequences and Rob Wheeler for sharing the *Caulobacter crescentus* RecA sequence prior to publication.

REFERENCES

[1] Figueras, J.B., Garcia-Gil, L.J. and Abella, C.A. (1997) Phylogeny of the genus *Chlorobium* based on 16S rDNA sequence. FEMS Microbiol. Lett. 152, 31-36.

[2] Pierson, B.K. and Castenholz, R.W. (1995) Taxonomy and physiology of filamentous anoxygenic phototrophs. In: Anoxygenic photosynthetic bacteria, Vol. 2, pp. 32-47 (Blankenship, R.E., Madigan, M.T. and Bauer, C.E., Eds.) Kluwer Academic Publishers, Dordrecht.

[3] Gibson, J., Ludwig, W., Stackebrandt, E. and Woese, C.R. (1985) The phylogeny of green photosynthetic bacteria: absence of a close relationship between *Chlorobium* and *Chloroflexus*. Syst. Appl. Microbiol. 6, 152-156.

[4] Pierson, B.K. and Olson, J.M. (1987) Photosynthetic bacteria. In: Photosynthesis, pp. 21-42 (Amesz, J., Ed.) Elsevier Science Publishers B. V.

[5] Pierson, B.K. and Castenholz, R.W. (1974) A phototrophic gliding filamentous bacterium of hot springs, *Chloroflexus aurantiacus*, gen. and sp. nov. Arch. Microbiol. 100, 5-24.

[6] Holo, H. (1989) *Chloroflexus aurantiacus* secretes 3-hydroxy-propionate, a possible intermediate in the assimilation of CO_2 and acetate. Arch. Microbiol. 151, 252-256.

[7] Kowalczykowski, S.C., Dixon, D.A., Eggleston, A.K., Lauder, S.S. and Rehrauer, W.M. (1994) Biochemistry of homologous recombination in *Escherichia coli*. Microbiol. Rev. 58, 401-465.

[8] Clark, A.J. and Sandler, S.J. (1994) Homologous genetic recombination: the pieces begin to fall into place. Crit. Rev. Microbiol. 20, 125-142.

[9] Eisen, J.A. (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. 41, 1105-1123.

[10] de Lorimier, R., Bryant, D.A., Porter, R.D., Liu, W.-Y., Jay, E. and Stevens, S.E. (1984) Genes for the α and β subunits of phycocyanin. Proc. Natl. Acad. Sci. USA

81, 7946-7950.

[11] Gruber, T.M. and Bryant, D.A. (1997) Molecular systematic studies of eubacteria, using σ^{70}-type sigma factors of group 1 and group 2. J. Bacteriol. 179, 1734-1747.

[12] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

[13] Felsenstein, J. (1989) PHYLIP-Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-166.

[14] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425.

[15] Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783-791.

[16] Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. Science 276, 734-740.

[17] Olsen, G.J., Woese, C.R. and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. 176, 1-6.

[18] Ludwig, W. et al. (1994) Phylogenetic relationships of bacteria based on comparative sequence analysis of elongation factor TU and ATP-synthase beta-subunit genes. Antonie Van Leeuwenhoek 64, 285-305.

[19] Imhoff, J.F. (1995) Taxonomy and physiology of phototrophic purple bacteria and green sulfur bacteria. In: Anoxygenic photosynthetic bacteria, pp. 1-15 (Blankenship, R.E., Madigan, M.T. and Bauer, C.E., Eds.) Kluwer Academic Publishers, Dordrecht.

[20] Brock, T.D. and Madigan, M.T. (1984) Prentice Hall, Englewood Cliffs.

[21] Blankenship, R.E. (1992) Origin and early evolution of photosynthesis. Photosynth. Res. 33, 91-111.

[22] Golbeck, J.H. and Bryant, D.A. (1991) Photosystem I. In: Current Topics in Bioenergetics: Light Driven Reactions in Bioenergetics, Vol. 16, pp. 83-177 (Lee, C.P., Ed.) Academic Press, New York.

[23] Young, D.A., Bauer, C.E., Williams, J.C. and Marrs, B.L. (1989) Genetic evidence for superoperonal organization of genes for photosynthetic pigments and pigment binding proteins in *Rhodobacter capsulatus*. Mol. Gen. Genet. 218, 1-12.

[24] Shiozawa, J.A. (1995) A foundation for the genetic analysis of green sulfur, green filamentous and Heliobacteria. In: Anoxygenic Photosynthetic Bacteria, pp. 1159-1173 (Blankenship, R.E., Madigan, M.T. and Bauer, C.E., Eds.) Kluwer Academic Publishers, Dordrecht.

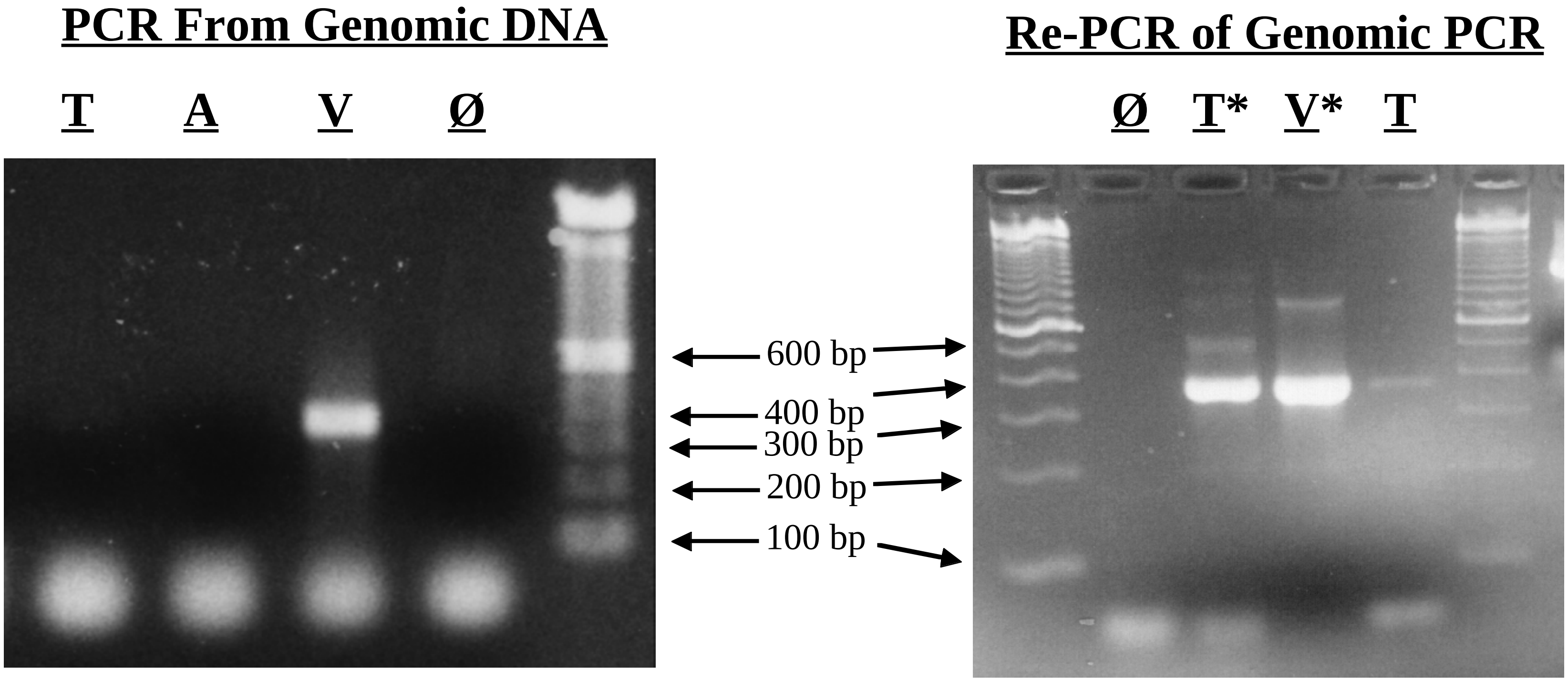
[25] Barry, B.A., Boerner, R.J. and de Paula, J.C. (1994) The use of cyanobacteria in the study of the structure and function of Photosystem II. In: The Molecular Biology of Cyanobacteria (Bryant, D.A., Ed.) Kluwer Academic Publishers, Dordrecht.

[26] Golbeck, J.H. (1994) Photosystem I in cyanobacteria. In: The Molecular Biology of Cyanobacteria (Bryant, D.A., Ed.) Kluwer Academic Publishers, Dordrecht.

[27] Wagner-Huber, R., Brunisholz, R., Frank, G. and Zuber, H. (1988) The BChl *c/e*-binding polypeptides from chlorosomes of green photosynthetic bacteria. FEBS Lett. 239, 8-12.

Figure 1. PCR of *recA* from *Chloroflexus* and *Chlorobium.*

Agarose gels showing results of PCR amplification of *recA* genes using primers and conditions described in the methods section. 5 ul of each PCR reaction was mixed with loading dye and run on 2% low melting point agarose gels. Bands of the expected size were excised and used for cycle-sequencing or for Southern hybridization. Not shown in FEMS Microbiology publication.



PCR From Genomic DNA

Re-PCR of Genomic PCR

T = *Chlorobium tepidum*
V = *Chlorobium vibrioforme*
A = *Chloroflexus aurantiacus*
Ø = Negative control (no DNA)
* = Size purified before re-PCR

Figure 2. Physical maps of RecA clones.

Physical maps of 3.2-kb *Eco*RI fragment encoding the *recA* gene of *Chlorobium tepidum* and of the 2.2-kb *Hin*cII fragment encoding the *recA* gene of *Chloroflexus aurantiacus*. The ORF found upstream of *recA* is homologous to dihydroflavonol-4-reductase (*dfr*) sequences. The ORFs upstream of *recA* are homologous to *nifR3* sequences and aspartate semialdehyde dehydrogenase (*asd*) sequences. Abbreviations for restriction enzymes are: C = *Hin*cII; E = *Eco*RI; H = *Hin*dIII; K = *Kpn*I; R = *Eco*RV.

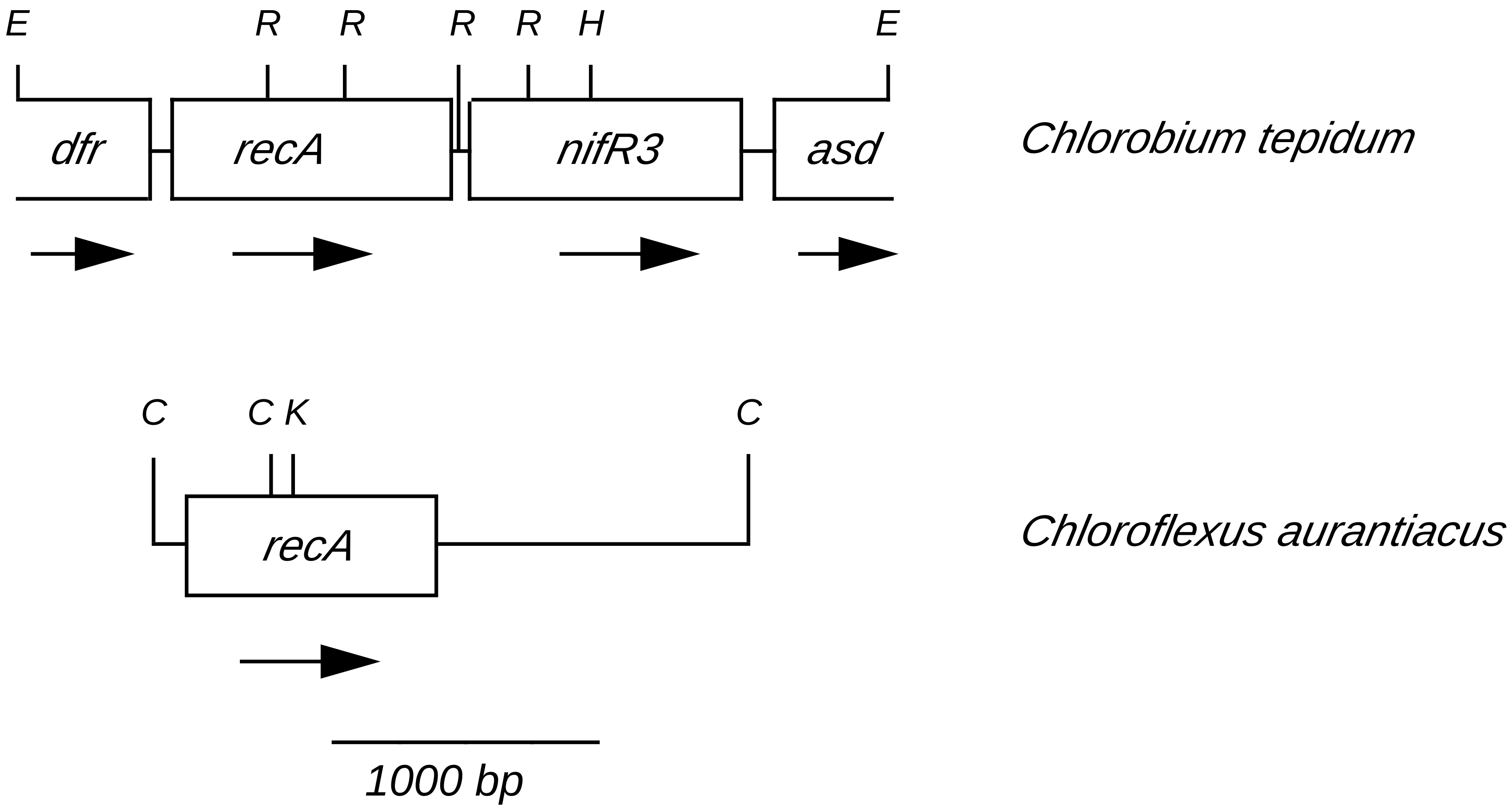
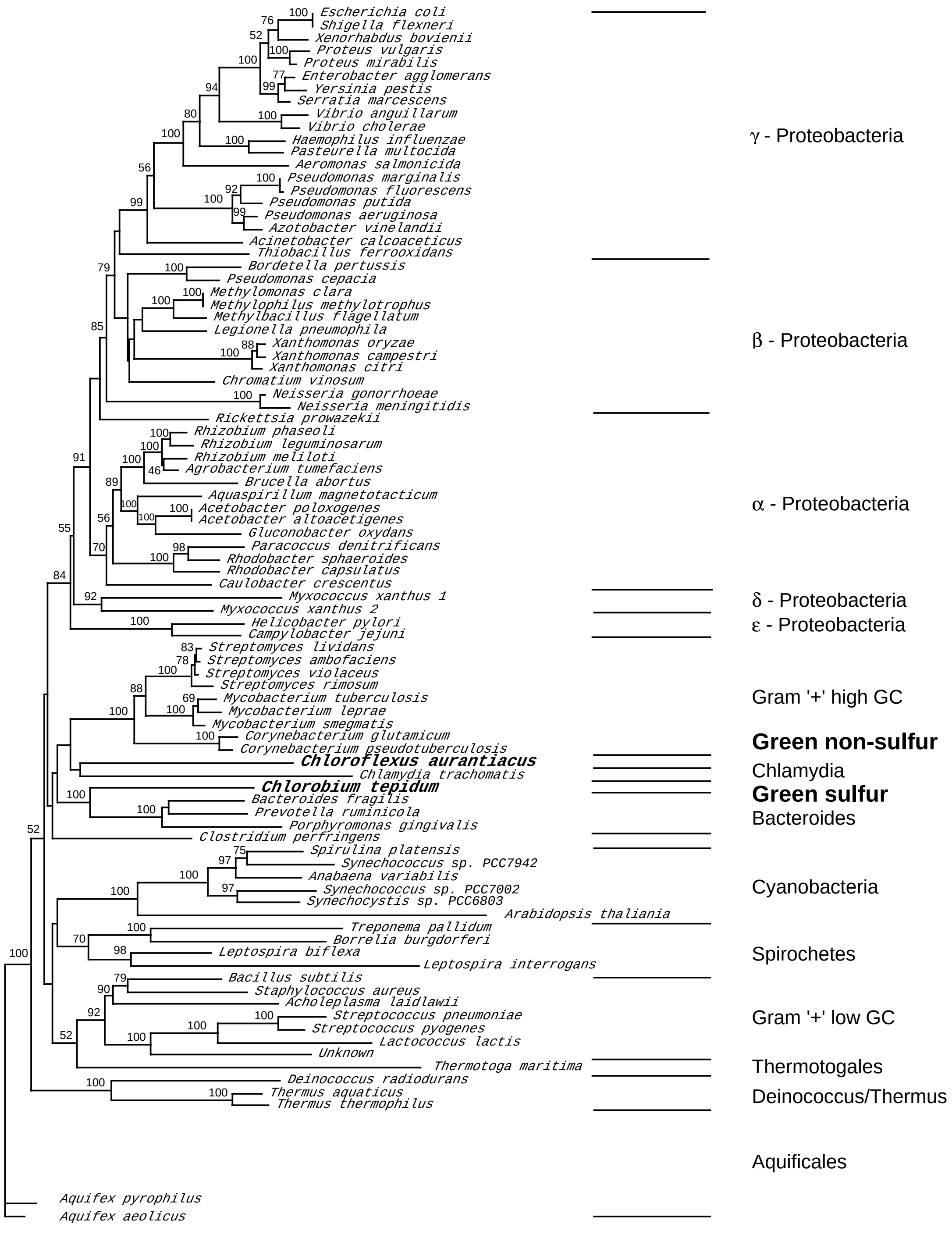

E
R
R
R
R
H
E
dfr
recA
nifR3
asd
Chlorobium tepidum
C
C K
C
recA
Chloroflexus aurantiacus
1000 bp

Figure 3. Neighbor joining tree for RecA

The distances were calculated using the *protdist* program of PHYLIP with a PAM-matrix based distance correction. Bootstrap values were obtained after 100 replications and are indicated when over 40. (References for the sequences used can be obtained in [9], in Genbank, in appropriate webpages of completed genome sequences, and at http://www-leland.stanford.edu/~jeisen/RecA/RecA.html).


Escherichia coli
Shigella flexneri
Xenorhabdus bovienii
Proteus vulgaris
Proteus mirabilis
Enterobacter agglomerans
Yersinia pestis
Serratia marcescens
Vibrio anguillarum
Vibrio cholerae
Haemophilus influenzae
Pasteurella multocida
Aeromonas salmonicida
Pseudomonas marginalis
Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas aeruginosa
Azotobacter vinelandii
Acinetobacter calcoaceticus
Thiobacillus ferrooxidans
Bordetella pertussis
Pseudomonas cepacia
Methylomonas clara
Methylophilus methylotrophus
Methylbacillus flagellatum
Legionella pneumophila
Xanthomonas oryzae
Xanthomonas campestri
Xanthomonas citri
Chromatium vinosum
Neisseria gonorrhoeae
Neisseria meningitidis
Rickettsia prowazekii
Rhizobium phaseoli
Rhizobium leguminosarum
Rhizobium meliloti
Agrobacterium tumefaciens
Brucella abortus
Aquaspirillum magnetotacticum
Acetobacter poloxogenes
Acetobacter altoacetigenes
Gluconobacter oxydans
Paracoccus denitrificans
Rhodobacter sphaeroides
Rhodobacter capsulatus
Caulobacter crescentus
Myxococcus xanthus 1
Myxococcus xanthus 2
Helicobacter pylori
Campylobacter jejuni
Streptomyces lividans
Streptomyces ambofaciens
Streptomyces violaceus
Streptomyces rimosum
Mycobacterium tuberculosis
Mycobacterium leprae
Mycobacterium smegmatis
Corynebacterium glutamicum
Corynebacterium pseudotuberculosis
Chloroflexus aurantiacus
Chlamydia trachomatis
Chlorobium tepidum
Bacteroides fragilis
Prevotella ruminicola
Porphyromonas gingivalis
Clostridium perfringens
Spirulina platensis
Synechococcus sp. PCC7942
Anabaena variabilis
Synechococcus sp. PCC7002
Synechocystis sp. PCC6803
Arabidopsis thaliania
Treponema pallidum
Borrelia burgdorferi
Leptospira biflexa
Leptospira interrogans
Bacillus subtilis
Staphylococcus aureus
Acholeplasma laidlawii
Streptococcus pneumoniae
Streptococcus pyogenes
Lactococcus lactis
Unknown
Thermotoga maritima
Deinococcus radiodurans
Thermus aquaticus
Thermus thermophilus
Aquifex pyrophilus
Aquifex aeolicus
γ - Proteobacteria
β - Proteobacteria
α - Proteobacteria
δ - Proteobacteria
ε - Proteobacteria
Gram '+' high GC
Green non-sulfur
Chlamydia
Green sulfur
Bacteroides
Cyanobacteria
Spirochetes
Gram '+' low GC
Thermotogales
Deinococcus/Thermus
Aquificales
100
76
52
100
100
77
99
94
100
80
100
100
56
100
92
100
99
99
79
100
100
100
85
88
100
100
100
100
91
100
46
100
89
100
100
100
55
56
98
70
100
84
92
100
83
78
100
88
69
100
100
100
100
100
52
75
97
100
97
100
100
70
98
100
79
90
100
92
100
100
52
100
100

CHAPTER 3

Effects of Differences in DNA Repair on Evolution

Mechanistic Basis for Variation in Microsatellite Mutation Rates[5]

[5] In press, to be published as *Jonathan A. Eisen. Mechanistic basis of microsatellite instability. In "Microsatellites: Evolution and Applications" (DB Goldstein and C Schlotterer, eds). Oxford University Press, Oxford.* Reprinted with permission of Oxford University Press.

ABSTRACT

The inherent instability of microsatellite loci makes them exceptionally useful for evolutionary and genetic studies. This instability is predominantly due to changes in the number of copies of the microsatellite repeat. Most copy number changes at microsatellites are caused by slip-strand mispairing errors during DNA replication. Some of these errors are corrected by exonucleolytic proofreading and mismatch repair, but many escape repair and become mutations. Thus microsatellite instability can be considered to be a balance between the generation of replication errors by slip-strand mispairing and the correction of some of these errors by exonucleolytic proofreading and mismatch repair. The factors that cause this process to occur much more frequently in microsatellites that in non-repeat containing DNA are discussed. However, not all microsatellites are equally unstable because not all are equally prone to this mutation process. The mechanisms by which a variety of factors cause this variation in stability among microsatellites are discussed.

INTRODUCTION

The characteristic that makes loci that contain microsatellite repeats particularly useful for evolutionary and genetic studies is their inherent instability. The mutation rates at most microsatellite loci are usually orders of magnitude higher than mutation rates at other loci within the same genome. Although many types of mutations occur at microsatellite loci, the elevated mutation rate is primarily caused by an elevated rate of one particular class of mutations -- changes in the length of the repeat tract. Thus the term "microsatellite instability" is frequently used to specifically refer to these tract length changes. Since most of these tract length changes result from changes in integral number of copies of the repeat, they are also frequently referred to as copy number changes

Ever since it was recognized that microsatellites are so prone to changes in tract length, researchers have been trying to determine why. A variety of approaches have been useful this purpose. Evolutionary and population genetic comparisons have been used to document the patterns of tract length variation at microsatellites and to test the robustness of different mutation models when averaged over long time scales. Biochemical experiments with purified proteins or cell extracts have been used to characterize each step in the mutation process and to determine the factors that control that step. Genetic studies have given insight into the genes that control microsatellite stability, and have allowed the accurate quantification of the stability of different microsatellites in controlled genetic backgrounds. Only by combining the results of these different types of studies has the mechanism of the mutation process become well characterized. Since evolutionary studies of the mutation mechanism are described in detail elsewhere in this book, I focus here on the biochemical and genetic studies.

To have a complete understanding of the mechanism of microsatellite instability one must also explain why stability varies both within and between species. Clues to the cause of this variation have come from the identification of factors that correlate with the level of microsatellite stability. Such factors include size of the repeat unit, number of copies of the repeat, presence of variant repeats, and amount of transcription in the region of DNA containing the repeat. Many studies that use data on microsatellite variation use models of the mutation process to enhance the analysis being done. Such studies should be improved by a better understanding of the mechanism underlying microsatellite instability as well as the causes of differences in stability among microsatellites. In this chapter, I summarize what is known about the mechanism underlying microsatellite instability and discuss some of the factors that cause variation in stability within and between species.

DISCUSSION

Microsatellite Mutation Models

The central debate about the mechanism of microsatellite instability has focused on two competing but not necessarily mutually exclusive models. One model proposes that microsatellite instability is caused by an elevated rate of unequal crossing-over

(UCO) within microsatellite repeats. Unequal crossing-over is the result of recombination between homologous chromosomes that are imperfectly aligned. The UCO microsatellite instability model suggests that UCO occurs at an elevated rate in microsatellites because the presence of repeats increases the likelihood of misalignment between homologs. A similar proposal has been made to explain the high rates of copy number changes observed in tandemly repeated genes (37). The alternative model proposes that microsatellite instability is caused by an elevated rate of slip-strand mispairing (SSM) errors during DNA replication. The SSM process, which was first proposed to explain frameshift mutations in any type of DNA (9), begins with the DNA polymerase "slipping" during replication, causing the template and newly replicated strands to become temporarily unaligned. For replication to continue, the strands must realign. Mutations will be generated if this realignment is imperfect. The SSM microsatellite instability model proposes that SSM occurs at an elevated rate in microsatellites because the presence of repeats increases the likelihood of misalignment after slippage (since repeats can easily be looped out of the DNA double-helix) (40).

The results of many studies indicate that an elevated rate of SSM is the main cause of microsatellite instability. The key evidence that supports the SSM model against the UCO model is summarized below (see (35) for review):

• Microsatellite stability is unaffected by defects in genes with major roles in recombination (*recA* in *Escherichia coli* (25), *rad52* in *Saccharomyces cerevisiae* (13)). This suggests against the UCO model since mutations are dependent on recombination in this model.

• In humans, copy number changes at microsatellites can be generated without exchange of flanking genetic markers (and thus probably without recombination) (30).

• In *S. cerevisiae*, microsatellite stability is similar in mitotic and meiotic cells (38). Since recombination occurs more frequently in meiosis than mitosis, if the UCO model were correct, microsatellites should be more unstable during meiosis.

• Microsatellite stability is reduced by defects in genes involved in DNA replication error correction pathways. This is consistent with the SSM model since this model requires DNA replication to occur. In addition, genetic and biochemical experiments show that these error correction pathways can recognize and repair the types of DNA loops that would be created by SSM (1, 31).

• The orientation of a microsatellite relative to the leading and lagging strands of replication influences its stability (10). This is not expected by UCO model but is consistent with the SSM model since the leading and lagging strands have somewhat different mechanisms of replication.

These and other results show that SSM is an integral component of the mutation process leading to microsatellite instability. However, SSM alone does not provide a full picture of this mutation process. As suggested above, not all SSM errors become mutations -- some are "repaired" by error correction mechanisms. The two error correction pathways that have been shown to be important in repairing SSM errors are exonucleolytic proofreading and post-replication mismatch repair. Thus a complete description of the mutation process must include both the generation of replication errors by SSM and the correction of some of these errors by mismatch repair and proofreading (see Figure 1). In the following sections I discuss each of the steps in the microsatellite instability mutation process, providing some details about the mechanism of each step and the methods used to study those mechanisms. In addition, I discuss how variability in each step contributes to variation in microsatellite stability within and between species (see Table 1).

Mutation Mechanism I: Slip-strand Mispairing

To study the mechanism of the SSM process, one must functionally isolate SSM from the downstream error correction steps. One approach to achieve such functional separation is to study the replication of DNA *in-vitro* (20, 21, 34). *In-vitro* studies allow straightforward comparisons of replication errors by different polymerases as well as comparisons of errors by the same polymerase using different templates. However, *in-*

vitro studies are limited because they may not accurately reflect what occurs during intracellular replication conditions. To study SSM errors *in-vivo*, researchers have used strains with defects in either exonucleolytic proofreading or mismatch repair or both. In such strains, since SSM errors are not corrected, SSM error rates and patterns can be inferred directly from observed mutations (e.g., (49)). Results from many such *in-vitro* and *in-vivo* studies show that the SSM process can be subdivided into three distinct steps: slippage of the DNA polymerase during replication, mis-realignment of the template and newly replicated DNA strands, and continuation of replication from a misaligned template (see Fig. 1).

These studies confirm the prediction of the SSM model that SSM errors are more likely to occur in microsatellite repeats than in "normal" DNA. However it has not been determined which step of SSM is most affected by the presence of repeats: slippage, misalignment or extension. It is almost certain that misalignment is more common in repeat regions than in “normal” DNA. Loops generated by misalignment will be more stable in microsatellites than in non-repeat regions since base-pairing is not significantly changed when one or more copies of a repeat are in a loop (see Fig. 1). However, there is also reason to believe that slippage occurs more frequently in microsatellite repeats than in normal DNA. *In-vitro* studies show that DNA containing microsatellite repeats is particularly prone to the formation of unusual DNA structures. Such structures likely interfere with the replication process, which could lead to slippage by the polymerase (16, 33). Thus, the elevated SSM rates at microsatellites relative to normal DNA may be caused by an increased likelihood of both slippage and misalignment.

SSM variation: effects of the nature of the microsatellite

Although in general SSM errors are more frequent in microsatellite containing regions than other regions of the genome, the rate and type of such errors are not equal for all microsatellites. The nature of the microsatellite itself has a large impact on SSM. For example, the likelihood of SSM for a particular microsatellite is correlated with the number of copies of the repeat. The most detailed study of this copy number effect is that of Wierdl et al. (49) in which the stability of five microsatellites with different numbers of copies of a GT repeat was analyzed. The mutation *rate* was found to increase with more repeats (as is expected since there are more places to slip and misalign) but the increase was greater than expected (more than two orders of magnitude between loci with 7.5 and 52.5 repeats). The *types* of mutations also differed between the microsatellites with different numbers of the repeat. The long tracts (those with more repeats) were more likely to have large, multi-repeat deletions than short tracts. In addition, the mutations that resulted in single repeat changes (plus or minus one repeat) were different between long and short tracts. The single copy changes in long tracts were mostly additions while those in short tracts included roughly equal numbers of additions and deletions. Wierdl et al. showed that these copy number effects were not due to biases in mismatch repair since the effects were seen in mismatch repair mutants. Therefore, they concluded that the copy number effects were probably caused by differences in SSM between microsatellites with different numbers of repeats. However, they were not able to determine the step of SSM that was influenced by copy number. One possibility is that the unusual DNA structures discussed above as a potential cause of increased slippage in microsatellite repeats may be even more likely to occur as the number of repeats increases. Regardless of the exact mechanism, the details of the effects of copy number on SSM (and thus on microsatellite stability) help to explain why the number of repeats at a particular microsatellite is somewhat stable over evolutionary time. Long tracts may be biased towards getting shorter (due to the large deletions) and short tracts may be biased towards getting longer (because of a slight bias in additions over deletions). An effect of copy number may also explain why certain microsatellites (e.g., those associated with some human diseases) become particularly unstable after they cross a threshold number of copies of the repeat (see Chapter by Rubinsztein).

Another aspect of the microsatellite that influences the likelihood of SSM is the presence of variant repeats. Evolutionary and genetic studies have shown that the presence of variant repeats is correlated with the stability of a microsatellite (e.g., (11)). Petes et al. have studied this effect in controlled laboratory conditions in *S. cerevisiae* to try to determine the underlying mechanism (32). This study showed that the presence of variant repeats leads to an approximately five-fold stabilization of GT repeats. Since this stabilizing effect was also seen in mismatch repair mutants, the authors suggested that the variant repeats exerted their effect by reducing the likelihood of SSM errors. However, as with the copy number effect described above, it has not been possible to determine what step of SSM was most “stabilized” by variant repeats.

SSM variation: effects of external factors

There are many reasons to believe that external factors (i.e., factors other than characteristics of the microsatellite) can influence SSM error rates and patterns. For example, base misincorporation error rates and patterns are influenced by many external factors. Since base misincorporation and SSM are both forms of polymerase error, it is likely that these factors will also influence the SSM process. External factors that influence misincorporation errors include local DNA sequence (e.g., the GC content or the ability to form secondary structures), genome position (e.g., proximity to replication origins or chromosome ends), and even the chromosome in which a sequence is found (e.g., nuclear, organellar, plasmid) (14, 19, 51, 52). In addition, misincorporation error rates are dependent on many conditional factors including methylation state, amount of chromosome packaging, temperature, phase of the cell cycle during which a particular section of DNA is replicated, and amount of DNA damage and repair prior to replication. Future studies of microsatellite mutation mechanisms would benefit by examining whether some of these factors influence SSM errors.

SSM variation: differences between individuals or species

Although the SSM mechanism and its role in causing microsatellite instability are conserved between species, it is likely that the specific rates and patterns of SSM differ greatly between species. For example, polymerases from different species have significantly different base misincorporation error rates (22) and thus likely also have different SSM rates at microsatellites. In addition, many of the factors described above as influencing SSM errors within a species differ greatly between species (e.g., GC content, temperature, methylation). Thus it remains to be seen whether all species are affected by copy number and variant repeats in the same ways as described above.

Mutation Mechanism II: Exonucleolytic Proofreading

Exonucleolytic proofreading is a process in which DNA that has been recently synthesized is examined for errors made by the DNA polymerase. If errors are found, the exonuclease will degrade the newly replicated strand, the DNA polymerase will back up, and the strand will be recopied. Thus many errors made by the DNA polymerase will not become mutations because they will be “erased” by proofreading. Proofreading was originally characterized for its role in limiting mutations due to base misincorporation errors. The role of proofreading in regulating microsatellite stability has been determined by methods that are similar to those used to study SSM. *In-vitro* studies have been used to compare the error rates and patterns of polymerases with and without associated exonucleases and to determine the types of substrates that the proofreading exonucleases will degrade. *In-vivo* studies have allowed the determination of errors with and without exonucleases under realistic cellular conditions. In such *in-vivo* studies, it has been helpful to use strains with defects in mismatch repair so that the role of the proofreading step is clear.

Studies such as the ones described above have shown that proofreading is involved in regulating the stability of microsatellites, but the extent of this role is limited in two ways. First, proofreading only significantly influences the stability of a subset of microsatellites: those with both small unit size (mostly mono- and di-nucleotide repeats) (18, 35, 39) and few copies of the repeat (18, 41, 47). In addition, even for this subset of microsatellites, the impact of proofreading is limited -- the stability of such microsatellites only decreases by about five to ten fold in exonuclease mutants.

The details of the mechanism of proofreading help to explain why this process has only a limited role in regulating microsatellite stability (for review see (5, 22)). Proofreading exonucleases detect errors by monitoring the DNA that has just been replicated to determine whether it forms normal double-helical DNA structures with the template strand. Abnormal DNA structures trigger the exonuclease activity. This is how proofreading prevents many base misincorporation errors from becoming mutations. A base misincorporation error will lead to a base:base mismatch between the newly replicated and template DNA strands and many such mismatches will be recognized by proofreading exonucleases. However, proofreading exonucleases are only able to monitor the DNA within a few bases of the active site of the polymerase. This proximity effect explains why proofreading has at most a small impact on microsatellite stability.

Most loops generated by SSM will be too far from the replication fork to be recognized by proofreading exonucleases. The lack of a role of exonucleases in repairing most SSM errors at microsatellites helps to explain the high rate of microsatellite copy number changes relative to point mutation rates.

Proofreading variation

The impact of proofreading on microsatellite stability is limited, variation in proofreading can account for some of the variation in stability of microsatellites. As with SSM, the nature of the microsatellite has a profound impact on proofreading. The best example of this was described above -- proofreading only works on microsatellites that are short and in which the repeat unit size is small. The mechanism of both of these biases is directly related to the proximity effect described above. As the number of copies of a repeat increases, the impact of proofreading decreases because those loops that are generated by SSM will be even more likely to be far from the replication fork. In addition, in microsatellites with repeats of large unit size (e.g., 5 bp repeats), a loop just one repeat away from the replication fork may be too far away to be proofread (the base-pairing of one repeat may be enough to stabilize the DNA structure at the fork). Proofreading is also likely to be affected by many external factors. For example, the efficiency of some exonucleases is affected by both GC content and sequence context (19). Thus the sequence around a mononucleotide repeat may influence its mutation rate by altering the efficiency of proofreading. Finally, the impact of proofreading on microsatellite stability is also likely to vary greatly between species. For example, some species do not even have proofreading exonucleases associated with their DNA polymerases. Microsatellites with short mono- and di-nucleotide tracts should be more unstable in species without proofreading than in species with proofreading.

Mutation Mechanism III: Mismatch Repair

Mismatch repair was named based on its role in recognizing and repairing base:base mismatches that arise due to base misincorporation errors. It is now clear that the same process can repair DNA containing loops such as those generated by SSM at a microsatellite (see Fig. 1). Mismatch repair has a much more significant impact on microsatellite stability than proofreading. Defects in mismatch repair can cause microsatellite instability to increase by many orders of magnitude (see below for more details). Since mismatch repair plays such a key role in regulating microsatellite stability, differences in the repair of loops by mismatch repair could account for a great deal of the variation in microsatellite stability within and between species.

Before discussing the specifics of loop repair and how it varies within and between species, it is useful to review some details about the general mechanism of mismatch repair. Mismatch repair has been found in a variety of species from bacteria to humans. It has been characterized in the most detail in *E. coli*. In the other species in which it has been characterized, the overall scheme of mismatch repair works in much the same way as in *E. coli*. Thus the *E. coli* system has served as a useful model for mismatch repair of all species. The first critical step in mismatch repair in *E. coli* is the recognition of mismatched DNA by the MutS protein (see (29) for review). Specifically, a dimer of MutS (two MutS proteins bound together) binds to the site of a mismatch in double-stranded DNA. Subsequently, through an interaction between the MutS dimer, a dimer of the MutL protein, and a single MutH protein, a section of one of the DNA strands at that location is targeted for removal. Other proteins complete the repair process: the section of DNA that has been targeted is removed and degraded, a patch is synthesized using the complementary strand as a template, and the patch is ligated into place resulting in a repaired section of double-stranded DNA without mismatches.

The evidence that mismatch repair is involved in repairing SSM errors at microsatellites comes from three types of studies. First, defects in mismatch repair cause decreases in microsatellite stability (anywhere from 10 to 5000 fold depending on the species and the microsatellite). In addition, when DNA containing loops is transformed into cells, the loops can be repaired, but only if the cells have functional mismatch repair (1, 4, 31). Finally, *in-vitro* studies have shown that repair of loops can be carried out by purified mismatch repair proteins (23, 31). Each of these results has been found in a variety of species, showing that the role of mismatch repair in repairing loops at microsatellites is highly conserved. Incidentally, this is what led to the discovery that mismatch repair genes are defective in hereditary non-polyposis colon cancer in humans -- cells from patients with this disease showed high levels of microsatellite instability. In summary, these studies show that the repair of loops is very similar to the repair of mismatches.

MMR variation: effects of the nature of the microsatellites

Perhaps the most important cause of variation in mismatch repair is the nature of the microsatellite. Loops are not all recognized equally by mismatch repair system and this specificity varies between species. One factor that is very important to the recognition step is the size of loop. For example, in *E. coli*, transformation studies have shown that loops of 1-3 bases are repaired well, those of 4 bases are repaired poorly, and those greater than 4 bases are not repaired at all. *In-vitro* studies of purified mismatch repair proteins show that this is due to inability of MutS to recognize loops larger than 4 bases in size (23, 31). Thus in *E. coli*, microsatellites in which the repeat unit size is 4 bp or greater have especially high rates of instability since SSM errors in such regions are not repaired well. Mismatch recognition is also biased by loop size in many other species, although the specific size preferences are not completely conserved. For example, the yeast mismatch repair system appears to be able to recognize and repair loops up to 6 bp well (and possibly even up to 14 bp, although this has not been confirmed). More details about the mechanism causing the different size preference are given in the section on variation in mismatch repair between species. For the purposes of the discussion here, all that is important is that in many species the size limits of loop recognition help to explain why microsatellites with different repeat unit sizes have different mutation *rates*.

The size specificity of loop recognition also helps to explain variation in mutation *patterns* between microsatellites with different sized repeats. For example, in *S. cerevisiae*, the majority of mutations in mononucleotide repeats are additions or deletions of one repeat (i.e., plus or minus 1 bp). However, the majority of mutations at microsatellites with 5 bp repeats are additions or deletions of two or more repeats (36). To understand this phenomenon, it is important to recognize that the mutation rate and pattern for a microsatellite is determined by a combination of the rate and type of SSM errors and how well these errors are repaired. Thus a particular mutation may occur at a high rate either because it is a common SSM error or because it is repaired poorly. For the mononucleotide repeat described above, most SSM errors are repaired about equally well (errors involving even five repeats at a time can be repaired by mismatch repair). Thus the most common mutations are those that are the most common SSM errors. In contrast, for the microsatellite with the 5 bp repeat, mismatch repair will only repair single repeat changes. Thus, although SSM errors involving two or more repeats are not very frequent, most of the mutations are changes in two or more repeats because many of the single repeat changes are repaired. The size dependence of mismatch repair also explains why 20 bp repeats are so unstable in *S. cerevisiae* (36); mismatch repair will not recognize any SSM error involving such a large repeat. Since both the number of repeats and the size of the repeat influence microsatellite stability, it is important to compare repeats of the same unit size when studying copy number effects and repeats with the same number of copies when studying unit size effects.

One aspect of loop repair that has been poorly studied is the role of the *type* of microsatellite (e.g., GT vs. GA repeats). Since base:base mismatch repair is not uniform for all mismatches (e.g., C:C mismatches are not repaired well in many species), it is likely that loop repair will also not be uniform. Since most of the studies of microsatellite mutation mechanisms have been done on limited types of microsatellites, it will be important to determine if the results of these studies are universal to all types of repeats.

MMR variation: effects of external factors

As with SSM and proofreading, many factors in addition to the nature of the microsatellite itself can influence the effectiveness of mismatch repair. For example, the location of the mismatch within the genome is important. In *S. cerevisiae*, loop recognition appears to be biased between loops on the template versus nascent strand of replication. For loops including a single repeat, mismatch repair appears to preferentially repair those that are on the template strand, resulting in a bias towards single repeat additions. The exact mechanism of this strand bias is not known although some of the genes involved have been identified (44, 45). Another effect of location is whether the mismatch is in nuclear or organellar DNA. Although organellar mismatch repair has not been characterized in detail, it is likely quite different from nuclear mismatch repair. The surrounding DNA also influences mismatch repair. For example, studies of base:base mismatches have shown that mismatch recognition is affected by sequence context (2), GC content (15). It is likely that the recognition of loops will also be affected by these factors. Finally, mismatch repair can also be influenced by conditional factors including the presence of strand recognition signals, methylation state, and level of transcription.

MMR variation: differences within a species

Differences in mismatch repair among individuals of a particular species have been well documented. For example, many strains of *E. coli* in the "wild" are defective in mismatch repair (24, 27). Since there are adaptive benefits to having modest increases in mutation rates in certain circumstances (42, 43), and since one way to alter mutation rates is by altering mismatch repair, many strains may be found to have defects in mismatch repair. Also, since mismatch recognition is involved in other cellular processes such as the regulation of interspecies recombination, there may be other selective pressures that lead to variation in mismatch repair capabilities within a species. Finally, since organisms appear to be able to turn mismatch repair on and off in certain situations (12, 26, 46), environmental conditions may play a major role in determining mismatch repair capabilities.

MMR variation: differences between species

Although mismatch repair is a highly conserved process, there are many ways in which it varies between species. For example, the mismatch recognition process is not completely conserved between bacteria and eukaryotes. The best characterized eukaryotic mismatch repair system is that of *S. cerevisiae*. As suggested above, the general mechanism of *S. cerevisiae* mismatch repair in very similar to that of *E. coli* (see (17) for review). In particular, the role of the MutS and MutL proteins is highly conserved - *S. cerevisiae* uses homologs of these proteins in essentially the same way that they are used in *E. coli*. Even the use of the proteins as dimers is conserved. However, unlike *E. coli*, *S. cerevisiae* uses multiple homologs of both MutS and MutL for mismatch repair. These multiple homologs are used to make separate mismatch repair complexes with unique and distinct functions. The specificity of each of these complexes is determined almost entirely by its particular combination of MutS homologs (which are referred to as MSH proteins for MutS Homolog). For mismatch repair of nuclear DNA there are two recognition complexes: an MSH2-MSH6 heterodimer for recognizing and repairing base:base mismatches and loops of 1-2 bases, and an MSH2-MSH3 heterodimer for recognizing and repairing loops of 1-6 bases (and possibly even up to 14 bases -- see (36)). Genetic studies suggest that there may also be a mitochondrial specific mismatch repair complex. Defects in another MutS homolog, MSH1, cause increases in the mutation rates in mitochondrial DNA. However, the details of mitochondrial mismatch repair are not well understood. In particular, it is not known what the role mismatch repair plays in microsatellite stability in mitochondrial DNA. Interestingly, *S. cerevisiae* encodes two additional MutS homologs (MSH4 and MSH5) that do not function in mismatch repair, but instead appear to use mismatch recognition to regulate meiotic crossing-over and chromosome segregation. The mismatch recognition process of other eukaryotes is highly similar to that of *S. cerevisiae* (8). One of the results of the differences in mismatch repair between eukaryotes and *E. coli* is that eukaryotes can repair loops of larger sizes than *E. coli*. This explains why microsatellites with these larger sized repeats are more stable in eukaryotes than in *E. coli*.

Another major difference in mismatch repair between species is in the mechanism used to determine which strand is the recently replicated strand (and thus is the strand that contains the error). In *E. coli* the “incorrect” strand is determined by its methylation state -- the newly replicated strand is unmethylated and thus can be distinguished from the template strand. In some other species, strand recognition is thought to be based on the presence of nicks, which are more likely to occur on the newly replicated strand. In such species, there may be differences in mismatch repair efficiency between the leading and lagging strands, since nicks are more common on lagging strand.

Although the process of mismatch repair is highly conserved, some species may not have the process at all. For example, analysis of complete genome sequences shows that some bacterial and Archaeal species do not encode any likely MutS or MutL homologs (6, 7). It is likely that these species do not have any mismatch repair, since functional MutS and MutL homologs are absolutely essential to the mismatch repair process. Any species without mismatch repair should have significantly elevated levels of microsatellite instability. In addition, differences between species could arise from the number and types of MutS and MutL homologs that are present.

Mutation Mechanism IV: Additional Factors that Affect Microsatellite Stability

Although the studies of microsatellite mutation mechanisms have been extensive, there are still many factors that have been found to influence microsatellite stability but for which the mechanism of the effect is unknown. For example, Wierdl et al. (50), following up previous studies (3), showed that transcription leads to a 4-9 fold destabilization of polyGT repeat. One explanation for this is that transcription will increase the likelihood of repair by the process of transcription-coupled repair and this process is mutagenic in some conditions (48). Alternatively, transcription could interfere with either mismatch repair or replication. Another unexplained observation is that microsatellites in the chromosome are usually more stable than identical microsatellites on a plasmid (13). Finally, many studies have shown that microsatellite stability is dependent on the orientation of the microsatellite within the DNA (10, 16, 28). For example, Freudenreich et al showed that a microsatellite with 130 CTG repeats was more unstable when the CTG was on the lagging strand (10). They suggested that this could be due to differences in the likelihood of slippage on the leading vs. lagging strand of replication. However, they could not rule out differences in mismatch repair, transcription, proofreading or other factors between the strands as the explanation. An alternative explanation for the orientation effect is that loops may be better recognized on the nascent strand than on the template strand (36). More detailed studies of microsatellite mutation mechanisms will likely sort out how these and factors influence microsatellite stability.

Conclusions and Summary

The mutation process at microsatellites can be considered to be a balance between the generation of replication errors by slip-strand mispairing and the correction of some of these errors by exonucleolytic proofreading and mismatch repair. The mutation rate and pattern for a particular microsatellite will be determined by the rate and type of SSM errors as well as how well these errors are recognized and repaired by exonucleases and mismatch repair. The details of the mutation mechanism explain why microsatellites are so unstable. First, SSM occurs much more frequently in microsatellites than in normal DNA. In addition, exonucleolytic proofreading, which prevents a large proportion of base misincorporation errors from becoming mutations, has only a limited role in preventing SSM errors from becoming mutations. The details of the mutation mechanism also help to understand why microsatellite stability varies within and between species. For example, the high mutation rate of microsatellites with repeats of large unit size can be explained by the inability of mismatch repair to recognize SSM errors in such large repeats. In addition, the positive correlation between number of copies of a repeat and stability can be explained by an increased likelihood of SSM errors in microsatellites with more repeats. The details not only help to understand the mutation process causing microsatellite instability, but they can be used to improve models of microsatellite evolution. Just as better models of nucleotide substitution processes have improved the analysis of DNA sequence variation, better models of microsatellite instability should improve the analysis of copy number variation at microsatellite loci.

REFERENCES

1. **Bishop, D. K., J. Andersen, and R. D. Kolodner.** 1989. Specificity of mismatch repair following transformation of *Saccharomyces cerevisiae* with heteroduplex plasmid DNA. Proceedings of the National Academy of Sciences U.S.A. **86:**3713-3717.
2. **Cheng, J. W., S. H. Chou, and B. R. Reid.** 1992. Base pairing geometry in GA mismatches depends entirely on the neighboring sequence. Journal of Molecular Biology **228:**1037-1041.
3. **Datta, A., and S. Jinks-Robertson.** 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science **268:**1616-1619.
4. **Dohet, C., R. Wagner, and M. Radman.** 1986. Methyl-directed repair of frameshift mutations in heteroduplex DNA. Proceedings of the National Academy of Science U.S.A. **83:**3395-3397.
5. **Echols, H., and M. F. Goodman.** 1991. Fidelity mechanisms in DNA replication. Annual Review of Biochemistry **60:**477-511.
6. **Eisen, J. A.** 1998. Evolution of the MutS family of proteins: gene duplication, loss, and functional divergence. Submitted

7. **Eisen, J. A., D. Kaiser, and R. M. Myers.** 1997. Gastrogenomic delights: a movable feast. Nature (Medicine) **3:**1076-1078.
8. **Fishel, R., and T. Wilson.** 1997. MutS homologs in mammalian cells. Current Opinion in genetics and Development **7:**105-113.
9. **Fresco, J. R., and B. M. Alberts.** 1960. The accommodation of noncomplemetary bases in helical polyribonucleotides and deoxyribonucleic acids. Proceedings of the National Academy of Sciences U.S.A. **85:**311-321.
10. **Freudenreich, C. H., J. B. Stavenhagen, and V. A. Zakian.** 1997. Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Molecular and Cellular Biology **17:**2090-2098.
11. **Goldstein, D. B., and A. G. Clark.** 1995. Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucleic Acids Research **23:**3882-3886.
12. Harris, R. S., G. Feng, K. J. Ross, R. Sidhu, C. Thulin, S. Longerich, S. K. Szigety, M. E. Winkler, and S. M. Rosenberg. 1997. Mismatch repair protein MutL becomes limiting during stationary-phase mutation. Genes and Development 11:2426-2437.
13. **Henderson, S. T., and T. D. Petes.** 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae*. Molecular and Cellular Biology **12:**2749-2757.
14. **Hess, S. T., J. D. Blake, and R. D. Blake.** 1994. Wide variation in neighbor-dependent substitution rates. Journal of Molecular Biology **236:**1022-1033.
15. **Jones, M., R. Wagner, and M. Radman.** 1987. Mismatch repair and recombination in *Escherichia coli*. Cell **50:**621-626.
16. **Kang, S., A. Jaworski, K. Ohshima, and R. D. Wells.** 1995. Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. Nature Genetics **10:**213-218.
17. **Kolodner, R.** 1996. Biochemistry and genetics of eukaryotic mismatch repair. Genes and Development **10:**1433-1442.
18. **Kroutil, L. C., K. Register, K. Bebenek, and T. A. Kunkel.** 1996. Exonucleolytic proofreading during replication of repetitive DNA. Biochemistry **35:**1046-1053.
19. **Kunkel, T.** 1992. Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays **14:**303-308.
20. **Kunkel, T. A.** 1986. Frameshift mutagenesis by eucaryotic DNA polymerases in vitro. Journal of Biological Chemistry **261:**13581-13587.
21. **Kunkel, T. A.** 1990. Misalignment-mediated DNA synthesis errors. Biochemistry **29:**8004-8011.
22. **Kunkel, T. A.** 1992. DNA replication fidelity. Journal of Biological Chemistry **267:**18251-18254.
23. **Learn, B. A., and R. H. Grafstrom.** 1989. Methyl-directed repair of frameshift heteroduplexes in cell extracts from *Escherichia coli*. Journal of Bacteriology **171:**6473-6481.
24. **LeClerc, J. E., B. Li, W. L. Payne, and T. A. Cebula.** 1996. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science **274:**1208-1211.

25. **Levinson, G., and G. A. Gutman.** 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Research **15:**5323-5338.

26. **Macintyre, G., K. M. Doiron, and C. G. Cupples.** 1997. The Vsr endonuclease of *Escherichia coli*: an efficient DNA repair enzyme and a potent mutagen. Journal of Bacteriology **179:**6048-6052.

27. **Matic, I., M. Radman, F. Taddei, B. Picard, C. Doit, E. Bingen, E. Denamur, and J. Elion.** 1997. Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. Science **277:**1833-1834.

28. **Maurer, D. J., B. L. O'Callaghan, and D. M. Livingston.** 1996. Orientation dependence of trinucleotide CAG repeat instability in *Saccharomyces cerevisiae.* Molecular and Cellular Biology **16:**6617-6622.

29. **Modrich, P.** 1991. Mechanisms and biological effects of mismatch repair. Annual Review of Genetics **25:**229-253.

30. **Morral, N., V. Nunes, T. Casals, and X. Estivill.** 1991. CA/GT microsatellite alleles within the cystic fibrosis transmembrane conductance regulator (CFTR) gene are not generated by unequal crossingover. Genomics **10:**692-698.

31. **Parker, B. O., and M. G. Marinus.** 1992. Repair of DNA heteroduplexes containing small heterologous sequences in *Escherichia coli*. Proceedings of the National Academy of Sciences U.S.A. **89:**1730-1734.

32. **Petes, T. D., P. W. Greenwell, and M. Dominska.** 1997. Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. Genetics **146:**491-498.

33. **Samadashwily, G. M., G. Raca, and S. M. Mirkin.** 1997. Trinucleotide repeats affect DNA replication in vivo. Nature Genetics **17:**298-304.

34. **Schlötterer, C., and D. Tautz.** 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Research **20:**211-215.

35. **Sia, E. A., S. Jinks-Robertson, and T. D. Petes.** 1997. Genetic control of microsatellite stability. Mutation Research **383:**61-70.

36. **Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell, and T. D. Petes.** 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology **17:**2851-2858.

37. **Smith, G. P.** 1973. Unequal crossover and the evolution of multigene families. Cold Spring Harbor Symposium of Quantitative Biology **38:**507-513.

38. **Strand, M., T. A. Prolla, R. M. Liskay, and T. D. Petes.** 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature **365:**274-276.

39. **Strauss, B. S., D. Sagher, and S. Acharya.** 1997. Role of proofreading and mismatch repair in maintaining the stability of nucleotide repeats in DNA. Nucleic Acids Research **25:**806-813.

40. **Streisinger, G., Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, and M. Inouye.** 1966. Frameshift mutations and the genetic code. Cold Spring Harbor Symposium of Quantitative Biology **31:**77-84.

41. **Streisinger, G., and J. Owen.** 1985. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics **109:**633-659.
42. **Taddei, F., I. Matic, B. Godelle, and M. Radman.** 1997. To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. Trends in Microbiology **5:**427-428.
43. **Taddei, F., M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, and B. Godelle.** 1997. Role of mutator alleles in adaptive evolution. Nature 387:700-702.
44. **Tishkoff, D. X., A. L. Boerger, P. Bertrand, N. Filosi, G. M. Gaida, M. F. Kane, and R. D. Kolodner.** 1997. Identification and characterization of *Saccharomyces cerevisiae* EXO1, a gene encoding an exonuclease that interacts with MSH2. Proceedings of the National Academy of Sciences U.S.A. **94:**7487-7492.
45. **Tishkoff, D. X., N. Filosi, G. M. Gaida, and R. D. Kolodner.** 1997. A novel mutation avoidance mechanism dependent on *S. cerevisiae* RAD27 is distinct from DNA mismatch repair. Cell **88:**253-263.
46. **Torkelson, J., R. S. Jarris, M.-J. Lombardo, J. Nagendran, C. Thulin, and S. M. Rosenberg.** 1997. Genome-wide hypermutation in a subpopulation of stationary-phase cells underlies recombination-dependent adaptive mutation. EMBO Journal **16:**3303-3311.
47. **Tran, H. T., J. D. Keen, M. Kricker, M. A. Resnick, and D. A. Gordenin.** 1997. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. Molecular and Cellular Biology **17:**2859-2865.
48. **Wang, G., M. M. Seidman, and P. M. Glazer.** 1996. Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. Science **271:**802-805.
49. **Wierdl, M., M. Dominska, and T. D. Petes.** 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:**769-779.
50. **Wierdl, M., C. N. Greene, A. Datta, S. Jinks-Robertson, and T. D. Petes.** 1996. Destabilization of simple repetitive DNA sequences by transcription in yeast. Genetics **143:**713-721.
51. **Wolfe, K., W.-H. Li, and P. Sharp.** 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast and nuclear DNAs. Proceedings of the National Academy of Sciences U.S.A. **84:**9054-9058.
52. **Wolfe, K. H., P. M. Sharp, and W. H. Li.** 1989. Mutation rates differ among regions of the mammalian genome. Nature **337:**283-285.

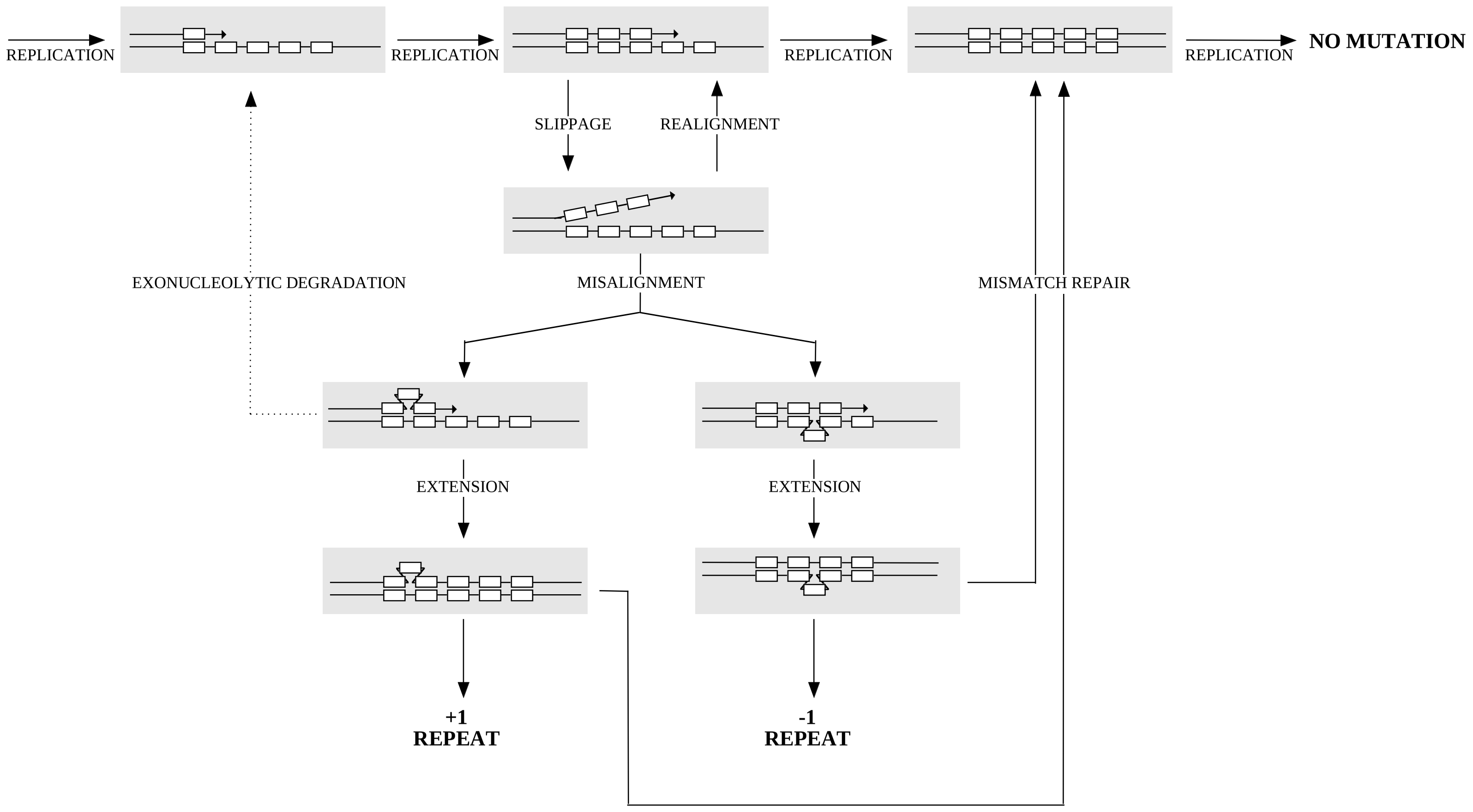
Table 1. Factors that lead to variation in mutation rates and patterns at microsatellite loci[1]

	Nature of the Microsatellite					Local DNA[2]		Cellular Conditions				Species Level	
Step in Mutation Process Affected by Factor	Repeat Unit Size	Number of Repeats	Type of Repeat	Variant Repeats	Replication Orientation	GC Content	Sequence Context	Transcription	Methylation State	Cell Cycle Stage	Pathway Used[3]	Pathway Presence[4]	Pathway Biases[5]
SSM (any step)	±	+	±	+	±	+	+	±	±	+	+	?	±
Replication Slippage	±	±	±	±	±	±	±	±	±	±	+	?	±
Misalignment	±	±	±	±	?	±	±	?	?	±	+	?	±
Extension[6]	?	?	±	?	±	±	±	?	?	±	+	?	±
Exonuclease	+	+	±	?	±	+	+	?	?	±	+	+	+
Mismatch Repair	+[7]	?	+	?	+	+	+	±	+	+	+	+	+

[1] +, documented experimentally; ±, suggested or likely but not yet well documented; ?, effect not known.
[2] Some of these effects have only been shown for base-misincorporation errors.
[3] For example, different polymerases are used for chromosome replication and DNA repair replication.
[4] Mismatch repair is absent in many strains and species and not all polymerases have associated exonucleases.
[5] For example, the ability to recognize loops for mismatch repair varies greatly between mismatch repair systems in different species.
[6] The extension and exonuclease steps are related in that they both work with the same substrate (see Fig. 1) but they can be functionally separated.
[7] Mismatch repair is affected by the total size of the loop, thus both number and size of repeats are important.

Figure 1. Model of the mutation process at microsatellite loci.

Cartoons of double-stranded DNA containing a microsatellite repeat are shown at different stages of the replication and mutation process. In the cartoons, DNA strands are represented by thin lines, microsatellite repeats by small boxes, and ongoing replication by small arrows. Flow arrows point down for steps that lead to mutations, up for steps that prevent mutations from occurring, and to the right for steps in the DNA replication process. The exonuclease step is shown with a dashed line since it has only a limited role in regulating microsatellite mutations. Details about each step are provided in the main text.


REPLICATION
REPLICATION
REPLICATION
REPLICATION
NO MUTATION
SLIPPAGE
REALIGNMENT
EXONUCLEOLYTIC DEGRADATION
MISALIGNMENT
MISMATCH REPAIR
EXTENSION
EXTENSION
+1
REPEAT
-1
REPEAT

CHAPTER 4

Using Evolutionary Analysis to Characterize DNA Repair Processes I:

Structure-Function Analysis of DNA Repair Proteins

recA mutations That Reduce the Constitutive Coprotease Activity of the RecA1202(Prt^{c}) Protein: Possible Involvement of Interfilament Association in Proteolytic and Recombination Activities[6]

[6] Previously published as *Shi-Kau Liu, Jonathan A. Eisen, Philip C. Hanawalt, and Irwin Tessman. 1993. Journal of Bacteriology 175(20): 6518-6529.* Reprinted with permission from the American Society of Microbiology.

ABSTRACT

Twenty-eight *recA* mutants, isolated after spontaneous mutagenesis generated by the combined action of RecA1202(Prt^{c}) and UmuDC proteins, were characterized and sequenced. The mutations are intragenic suppressors of the *recA1202* allele and were detected by the reduced coprotease activity of the gene product. Twenty distinct mutation sites were found, among which two mutations, *recA1620* (V-275→D) and *recA1631* (I-284→N), were mapped in the C-terminal portion of the interfilament contact region (IFCR) in the RecA crystal. An interaction of this region with the part of the IFCR in which the *recA1202* mutation (Q-184→K) is mapped could occur only intermolecularly. Thus, altered IFCR and the likely resulting change in interfilament association appear to be important aspects of the formation of a constitutively active RecA coprotease. This observation is consistent with the filament-bundle theory (R. M. Story, I. T. Weber, and T. A. Steitz, Nature (London) 335:318-325, 1992). Furthermore, we found that among the 20 suppressor mutations, 3 missense mutations that lead to recombination-defective (Rec^{-}) phenotypes also mapped in the IFCR, suggesting that the IFCR, with its putative function in interfilament association, is required for the recombinase activity of RecA. We propose that RecA-DNA complexes may form bundles analogous to the RecA bundles (lacking DNA) described by Story et al. and that these RecA-DNA bundles play a role in homologous recombination.

INTRODUCTION

The *recA* gene product of *Escherichia coli* is a small yet versatile protein composed of 353 amino acids (41). Two major and well-studied roles of RecA are to promote homologous recombination (7) and to induce the SOS response (35, 51). In homologous recombination, RecA is required for both strand pairing and an ATP-dependent strand exchange reaction (18, 37, 40, 56). In the SOS response, RecA is activated to a coprotease state by cofactors such as single-stranded DNA (ssDNA) and ATP or dATP (8, 38, 39). This activated RecA then mediates the cleavage of the LexA repressor (22, 24) and allows the expression of SOS genes, which are those under the repression of LexA and include *lexA, umuDC, and recA* itself (1, 4, 25). The activated RecA also mediates the cleavage of UmuD into two fragments, the larger of which, the C-terminal UmuD' is essential for the function of UmuD in SOS mutagenesis (5, 33, 43).

Although RecA is required for the cleavage of LexA, UmuD, and phage repressors (35, 51) in vivo and in vitro (under physiological conditions) and may act as a protease, the term coprotease (23) has been adopted to describe its proteolytic activity because the protein substrates of RecA can undergo autodigestion at alkaline pH in vitro (21, 44). We find it convenient, however, to retain the designation Prt (47) to describe the protease phenotype of RecA.

In wild-type *E. coli* cells, the RecA protein is not proteolytically active without inhibition of DNA replication or exogenous DNA-damaging treatments (35, 51). Some mutations in either *recA* or other genes result in the activation of the RecA protein in the absence of DNA-damaging agents. Mutations such as *dnaB(Ts)* and *dnaE(Ts)* can lead to changes in DNA metabolism and indirectly activate RecA and induce the SOS response constitutively (31, 42). This activation is likely due to an increase in the availability of ssDNA regions in the cell as a result of abnormal DNA replication (42). In addition, mutations in the *recA* gene, designated *recA*(Prt^c), can confer constitutive coprotease activity to RecA and turn on the SOS response at all times.

By using a method that involves plating mutagenized *λrecA* $^+$ (a λ phage carrying the *recA*$^+$ gene) on indicator strains with *recA* deleted and containing Mu d(Ap *lac)* fusions in SOS genes *(dinD* and *sulA*), Tessman and Peterson isolated several classes of novel *recA*(Prt^c) mutants (47), some of which are recombinase negative and are designated *recA*(Prt^c Rec$^-$) (48). Unlike mutants carrying the classical *recA441 (tif-1)* allele, which confers the Prt^c phenotype only at high temperature (6, 14), these newly isolated *recA*(Prt^c) mutants display constitutive coprotease activity at any growth temperature, with some having considerably greater coprotease activity than *recA441* strains (47). Among these, *recA1202* cells showed the strongest coprotease activity (47).

Further studies on two of these *recA*(Prt^c) mutants with strong RecA coprotease activity, the *recA1202 and recA1211* mutants, demonstrated that the strong RecA(Prt^c)

phenotype for these strains is likely due to two factors: (i) the RecA1202 and RecA1211 proteins can use any one of the other natural nucleoside triphosphates besides ATP or dATP as a cofactor in activating the cleavage of LexA (52), and (ii) they can use tRNA or rRNA besides ssDNA as a cofactor in the cleavage reaction (55). These novel biochemical properties of RecA(Prt^{c}) proteins provide an explanation for a mutagenic phenomenon observed with *recA1202*(Prt^{c}) cells; this phenomenon is termed proximal mutagenesis because the *recA1202* gene and nearby regions are preferentially mutated (26).

The proximal mutagenic activity was used in the present study to isolate mutants that reduce the constitutive coprotease activity of the *recA1202* allele. We characterized 28 such mutants. Each carried an additional *recA* mutation that can be viewed as an intragenic suppressor of the *recA1202* constitutive coprotease activity. These new double recA mutants, carried by λ phages, have been characterized for both recombination and coprotease phenotypes in a strain with its chromosomal *recA* gene deleted.

Story et al. (46) have solved the molecular structure of the RecA protein by X-ray crystallography to a 2.3-Å (0.23-nm) resolution. The crystallized RecA protein can be divided into three domains: a large central domain, and two smaller flanking domains at the amino and carboxyl termini (from residue 1 to about residue 30 and from about residue 270 to residue 328, respectively), both of which protrude from the central domain (Fig. 1a). There are two types of interactions among RecA molecules in the crystal. First, monomers pack together to form a filament coil, with six monomers per turn of the coil. The coil of the filament is relatively open, with intermolecular associations only between adjacent monomers in the filament polymer. In addition, there are interfilament associations between monomers which allow filaments to form a filament bundle in the crystal. Thus, each monomer interacts with four other monomers, two within the filament and two from another filament. Each intrafilament (intermolecular) contact is extensive, involving at least 54 amino acids; in contrast, the interfilament contact is less extensive, involving only about 20 amino acids (46).

While Story et al. (46) pointed out that these interfilament contacts may be an artifact of crystallization, they suggested that they may be biologically relevant because mutations in and around these residues have major effects upon RecA functions. In particular they noted that many of the mutations that lead to constitutive coprotease activity are located in the interfilament contact region (IFCR). On the basis of this observation, they proposed that the formation of a RecA filament, from the dissociation of bundles of RecA polymer (a storage form lacking DNA), is an important step in forming an active coprotease. Thus, mutations mapped in the IFCR, including *recA1202* (Q-184→K), could reduce the interfilament contact and shift the equilibrium toward the formation of active RecA filaments, which in turn results in the generation of a constitutively active coprotease (46). The existence of RecA bundles (with or without DNA) has been documented by in vitro studies (3, 10, 11, 57).

To further understand the structure-function relationship of RecA, we integrated into the crystal structure of RecA the phenotypic and sequencing data for 28 suppressor mutants that we isolated. Although the crystal structure may not reflect the exact RecA conformation in vivo, the 2.3-Å structure (46) can serve as a model upon which the analysis of newly obtained data can be based. Our analysis of some of the 28 suppressor mutations provides additional evidence supporting the theory proposed by Story et al. (46). Furthermore, our extended analysis of the locations of Rec$^-$ mutations indicates that interfilament association may also play an important role in recombination. We propose a theory involving RecA-DNA multifilament bundles to explain why the IFCR is involved in recombination and how single *recA* mutations might result in Prt^c Rec$^-$ phenotypes.

MATERIALS AND METHODS

Bacterial strains and media

The host bacterial strain for *λrecA* mutant phages used in this work was *E. coli* K-12 strain EST2411 (*ΔrecA306 sulA11 dinD1*::Mu d(Ap *lac*) *supE44* S13^s) (27), which is a derivative of AB1157. The *recA1202* control strain was IT1993 (*EST24111,λrecA1202* *c*I *ind*). *The recA*$^+$ strain was EST2422 (*λrecA*$^+$ *c*I *ind*) (27). The media, M9-CAA (a Casamino-Acids-supplemented M9), LB broth, and SFLB (a salt-free plating agar based on LB broth), have been previously described (26, 27). The antibiotics used were kanamycin (30 µg/ml) and rifampin (25 µg/ml). 5-Bromo-4-chloro3-indolyl-B-D-galactoside (X-Gal) was used at 60 µg/ml, and mitomycin (MitC) was used at 0.5 µg/ml.

Isolation and characterization of the suppressor mutants of the recA1202 allele

$CaCl_2$-treated IT3111 (EST2411/ *λrecA1202 cI857 ind*) cells were mixed with DNA from the high-copy-number plasmid pSE117 ($umuD^+C^+$ Kan^r) (12), heat shocked at 42°C for 2 min, diluted sixfold with LB broth and incubated at 30°C for 1 h before being spread on plates containing M9-CAA plus kanamycin and X-Gal. Strains with unmutated *recA1202* alleles produce dark blue (DB) colonies on X-Gal plates because the high coprotease activity derepresses the *dinD* gene and the fused *lacZ* gene (47). However, in combination with the pSE117 plasmid, the *recA1202* gene exhibits a very high frequency of mutation in the *recA* gene, which results in decreased RecA coprotease activity and pale blue (PB) colonies (26). Thus, after incubation at 32°C for 24 h, there were many transformants with stable PB or blue (B) colors that could be picked and purified. From these PB or B mutants, λrecA mutant phages were heat induced and then used to lysogenize EST2411. All phenotypic characterizations refer to these lysogenized strains. The temperature for all phenotypic characterizations was 32°C.

Sensitivity to UV was determined by spotting 10 µl of cells grown overnight in M9-CAA medium onto the surface of M9-CAA plates, which were then UV irradiated with a 15-W germicidal lamp with fluxes of 0, 20, 47, 75, and 103 J/m^2. Strains that were completely inactivated by 20 J/m^2 were designated S for sensitive, strains resistant to 103 J/m^2 were designated R for resistant, and strains inactivated by 47 or 75 J/m^2 were designated R/S. Determinations of the fraction of lethal lesions repaired (repair sector, *W*) by Weigle reactivation of UV-irradiated S13 and of the Rif^r frequencies were as described previously (26, 27).

Phage λ DNA purification

To induce the phage, lysogens of λ were grown in LB plus 0.01 M $MgCl_2$ to mid-log phase, heat induced at 45°C for 15 min, and incubated for another 2 h at 38°C.

Chloroform was added to complete the lysis, and bacterial debris were removed by centrifugation. The λ DNA was extracted and purified from the lysate by a λ DNA minikit with the protocol provided by the manufacturer (Qiagen).

DNA sequence determination

The λ DNA containing the mutated *recA* gene was digested with *Eco*RI. A 1.8- and a 1.3-kb *Eco*RI fragment, which contained three-fourths (N terminal) and one-fourth (C terminal) of the *recA* gene, were purified by using low-melting-point agarose gel electrophoresis (29). The DNA fragments that contained parts of the *recA* gene were sequenced by inserting the purified fragments into M13 mpl9. The orientation of the cloned *Eco*Rl fragments in M13 mpl9 was determined by a DNA hybridization test (13). Sequencing was performed by using a Sequenase kit (United States Biochemical). In order to sequence the entire *recA* gene without subcloning parts of it, three 19-mer synthetic DNA primers with the following sequences were used in addition to the universal primer: 5'-GCGGTGCGTCGTCAGGCTA-3', 5'-GCCGCAGCGCAGCGTGAAG-3', and 5'-CTCCTGTCATGCCGGGTAA-3'. The 5' nucleotides of these primers correspond to nucleotides -98 and 293 in the nontranscribed strand and to nucleotide 44 in the transcribed strand of the gene.

Structural analysis

We downloaded the spatial coordinates of the *E. coli* RecA crystal as solved by Story et al. (46) from the Brookhaven protein data base. All subsequent structural analysis was performed with the Midas computer program on an Iris workstation. Possible effects of changes in the amino acid sequences of *E. coli* RecA on the tertiary structure were predicted using the computer programs of Lee and Levitt (20).

Evolutionary comparisons

The amino acid sequences of the RecA proteins from 32 species of bacteria, including *E. coli,* were downloaded from the National Center for Biotechnology Information database via the Internet. These species covered a wide evolutionary range within the eubacterial kingdom, including enterobacteria, gram-positive bacteria, Bacteroides, spirochetes, mycoplasmas, cyanobacteria, and species from the α, β, and γ subgroups of the *Proteobacteria (58).* Sequences were aligned by the computer program CLUSTAL V (16) with the aid of the Genetic Data Environment computer program (kindly provided by Steve Smith, Millipore Corp.). Alignment ambiguities were limited to the C terminus (residue 315 or greater in the *E. coli* protein); the species showed a high degree of homology through the rest of the protein. Positions were scored for degree of conservation among the sequences. Completely conserved positions were those that are identical in all, or in all but one, species. Highly conserved positions were those with only conservative alterations among all species (e.g., valine, isoleucine, or leucine in all species) or among all but one. Moderately conserved positions were those that were identical or conservatively different in most species (>80%) but in which some nonconservative alterations were also present.

RESULTS

Isolation and characterization of recA mutants

The *recA1202* allele, in the absence of the *umuD*$^{+}$*C*$^{+}$ plasmid, produces a DB colony on X-Gal plates because it completely derepresses the SOS regulon, which includes the *dinD*::Mu d*(lac)* gene (17) in our strain. The high-copy-number *umuD*$^{+}$*C*$^{+}$ plasmid pSE117, in combination with the *recA 1202* allele, causes an extraordinarily high frequency of proximal mutations in the *recA1202* gene itself (26). The basis of our isolation procedure was the fact that many of the mutations are easy to detect because they weaken the Prt^{c} phenotype, resulting in B or PB colonies that are easily distinguished from the DB parent colonies.

Immediately after transformation with pSE117 *(umuD*$^{+}$*C*$^{+}$ Kan^{r}), the cells were incubated for 1 h to allow expression of the Kan^{r} phenotype before being spread on M9-CAA-XGal-kanamycin plates. After 24 h of incubation at 32°C, Kan^{r} colonies with various degrees of blue color were observed. Among 660 transformed colonies, 66% were DB and 34% were B or PB. The DB colonies were similar to the plasmidless parent

IT3111 in colony color except that many also contained B or PB sectors. The other B or PB colonies were presumed to have a change in color due to reduced constitutive coprotease activity. The 34% B or PB colonies looked homogeneous in colony color, suggesting that the mutations had occurred during the 1-h incubation period before plating. From these, 37 mutants were chosen to give a wide distribution of colony sizes and colors. These mutants were tested for their RecA functions by Weigle reactivation (an indication of SOS repair) of UV-irradiated phage S13, UV sensitivity, and crystal violet sensitivity (47, 48) as described in Materials and Methods. By these tests, 35 of the mutants could be distinguished from the *recA1202* control strain IT1993; it was inferred that these mutants were further mutated in their *recA1202* genes.

From the 35 potential *recA* mutants, 28 were selected by the ease with which the *λrecA* DNA could be isolated, and they were then analyzed for their DNA sequence and phenotype. The *λ* lysates from the other seven *recA* mutants gave consistently low titers, which may be an indication that the proximal mutagenesis phenomenon resulted in mutations in some important λ genes located near *recA1202* in the prophage. These mutants were not further characterized. The phage lysates from the 28 mutants were also used to lysogenize EST2411, in which characterization of the mutant RecA phenotype could be carried out free of the multicopy *umuDC* plasmid that might have complicated the studies. Sequencing of the DNA revealed that all 28 mutations represented 20 distinct sites within the *recA* gene.

The distinctive properties of the 20 different *recA* mutants allowed us to classify them into six groups, each containing a unique combination of Prt and Rec phenotypes (Table 1). Three tests, color on M9-CAA-X-Gal, repair sector (*W*) for the Weigle reactivation of UV-damaged S13 in unirradiated cells, and spontaneous mutation frequency to Rif, were used to measure the constitutive coprotease strength of RecA mutants in vivo. The correlation between the RecA coprotease activity and these three phenotypes has been established (47, 49). All 20 *recA* mutants appeared to have weaker-than-normal constitutive coprotease activity as indicated by the reduced values of W and the Rif^r frequencies, properties we expected from our mutant isolation strategy. Sensitivity to MitC and UV were used to estimate the recombinase activity of the mutants, since recombinase activity of RecA is a major determinant of resistance to UV and other DNA-damaging agents such as MitC (47, 54). The Rec^- phenotype of all the UV- and MitC-sensitive mutants was further confirmed (data not shown) by the lack of recombination with DNA from an Hfr donor strain, as previously described (48).

In cases when constitutive coprotease activity of RecA was low, the ability of MitC to induce activity of RecA coprotease was also tested. Those mutants which were PB on M9-CAAX-Gal and became B on plates with added MitC were then classified as Prt^+ if they also showed low-level constitutive Weigle reactivation (W ≤0.01) and a low Rif frequency, both of which are characteristic of the *recA*$^+$ reference strain EST2422 (Table 1). This method could not be used to determine the inducible coprotease phenotype (Prt^+ or Prt^-) of Rec^- strains because they are sensitive to MitC (designated S in Table 1). Therefore, the $\mathrm{Prt}^{\pm}$ Rec^- and Prt^- Rec phenotypes were determined by measuring Weigle reactivation of UV-inactivated phage S13 in cells induced with UV light at 16 J/m^2. The single mutant IT3200, classified as $\mathrm{Prt}^{\pm}$, had the intermediate value of W = 0.08 when UV irradiated, while all the mutants classified as Prt^- Rec^- showed negligible values of W (≤0.01). Because W was 0.04 for IT3170 in the absence of UV induction (Table 1), further study was unnecessary inasmuch as the Prt^c phenotype was apparent.

Distribution of the mutations in the recA gene

The DNA sequence changes and inferred amino acid substitutions were determined for all 28 mutants (Table 2). All mutants retained the original *recA1202* mutation, which is a Gln→Lys (GAG→AAG) change at amino acid residue 184. Thus, all alleles have double mutations in the *recA* gene, but for convenience only the second-site change is indicated. Among, the 20 distinct second-site mutations, one was in the promoter region of *recA*, one was a 10-bp deletion, and two were nonsense mutations. The remaining 16 were single base pair missense mutations, which were distributed between residues 111 and 284. Of these mutations, 14 were in the region corresponding to the central domain of the RecA protein crystal, two were in the C-terminal domain, and none were in the N-terminal domain (Fig. 1). For each of the missense mutations, we analyzed the degree of evolutionary conservation of the residues shown in Table 2. The results indicate that 88% (14 of 16) of the suppressor mutations resulted in changes at either completely conserved (6 of 16) or highly conserved (8 or 16) residues of RecA (Table 2), implying that most of the mutated residues play important roles in aspects of RecA structure and function.

Mutations mapped in the three-dimensional vicinity of Gln-184

Three suppressor mutations, *recA1630* (K-177 →Q), *recA1623* (L-182→Q), and *recA1625* (T-187→A), alter residues that are very close to residue 184, the site of the *recA1202* mutation (Fig. 2), and they therefore may directly interact with it. In addition to compensating for the structural effect caused by *recA1202*, all three could also affect another area of the IFCR (see below). The importance of this region is also suggested by analysis of the *recA1623* allele. The suppressor mutation in this allele is a change from a polar to a nonpolar residue at a highly conserved residue that is in a nonexposed packing region immediately next to the original mutation (Fig. 2). Such a change would seem likely to destabilize the whole region and destroy its associated functions, a prediction that nicely fits with the Prt$^-$ Rec$^-$ phenotype of this allele. Of the remaining 13 suppressor mutations, only two appear to be reasonable candidates for changes that could have some direct influence on residue 184*: recA1636* (L-132→Q) and *recA1626* (D-139→G). These residues are relatively close to the original change at residue 184 in the three-dimensional structure. In particular, the mutation of *recA1626* is at a residue in the same hydrophobic packing region as *recA1623 and recA1625.*

Mutations mapped in the putative binding site for the LexA repressor

Story et al. (46) proposed that a pocket (Fig. 3) formed by two adjacent RecA monomers in the crystal may be the binding site for the UmuD, LexA, and phage repressors. This proposal was based on mutation information and physical considerations (46). Residues 229 and 243, cited by Story et al. (46), have also been indicated as a contact region between RecA and LexA in a recent study on the structure of the LexA-RecA filament complex (60). Three of our suppressor mutations, *recA1627, recA1628,* and *recA1642,* were mapped to this region of the crystal (Table 2; Fig. 3). These three mutations resulted in a Prt$^+$ phenotype (Table 2). Thus, the resultant RecA mutant proteins are no longer constitutively active and have a coprotease activity like that of wild-type RecA: they become active only after DNA-damaging treatments. It is possible that the reduction of the RecA Prt^{c} activity in these mutants is due to defects in binding to the LexA repressor.

Mutations mapped in the putative intermolecular packing regions in a RecA filament

Four of the suppressor mutations were mapped to residues that may be involved in the contact regions between RecA monomers within a polymer filament (Table 2) (46). It is not surprising that mutations in these regions would reduce the coprotease activity, since it has been indicated that the coprotease activity depends upon the formation of a filament in the presence of ATP and ssDNA (46, 60). None of the four mutations produced a Prt^{-} phenotype; two of them produced a Prt^{c} phenotype with reduced coprotease activity, and the other two mutants exhibited a Prt^{+} phenotype (Tables 1 and 2).

Mutations in possible DNA- or ATP-binding sites of RecA

Story et al. suggest that three regions are particularly likely to be involved in DNA binding: regions in or near loop 1 (residues 157 to 164), loop 2 (residues 195 to 209), and helix G (residues 213 to 218) (Fig. la) (46). This suggestion was based on structural comparisons with other DNA-binding proteins and on DNA-binding properties of mutant proteins with known mutation sites (46). It should be emphasized that the structure of a RecA-DNA complex has not been solved, and these assignments should be considered tentative. Four mutations were mapped to or near these three regions (Table 2). Story and Steitz solved the crystal structure of a RecA-ADP complex (45). One suppressor mutation was found in sequences corresponding to the proposed ATP-binding site (Table 2) (45). If these regions are indeed involved in DNA and ATP binding, these mutations may reduce the Prt^{c} activity by affecting the binding activity of RecA1202 for DNA (or similarly RNA) or ATP (or similarly other nucleoside triphosphates).

Mutations in the IFCR

The remaining five suppressor mutations were mapped to the IFCR of the crystal. Of these, three were discussed above as likely having effects partly due to direct interaction with the mutant residue 184. The two others *(recA1620* and *recA1631)* are missense changes in the C-terminal domain of the RecA monomer, and the altered sites evidently could not interact intramolecularly with the IFCR near residue 184 (Fig. la) (46). Interestingly, these sites of *recA1620* (V-275→D) and *recA1631* (I-284→N) are in the C-terminal portion of the IFCR that is close to the original *recA1202* (Q-184→K) site in the adjacent filament (Fig. 2). Both of these mutations are nonpolar-to-polar changes in a hydrophobic core of the C-terminal domain. These drastic changes could lead to an altered spatial position of the IFCR and thus compensate for the alteration caused by the mutation at residue 184 in the IFCR. Such a situation is similar to that suggested by Story et al. (46) for the temperature-sensitive suppression of the Prt^{c} phenotype of the *recA1211* (E-38→K) allele by an I-298→V change (53). Thus, our finding is consistent with the filament-bundle theory, which suggests that the interference with the interfilament association is the cause of the constitutive coprotease activity of RecA(Prt^{c}) mutant proteins (46).

Effect of the suppressor mutations on recombination

The *recA1202* allele has a wild-type recombination phenotype (Rec^{+}). Of the 16 missense suppressors, 8 led to defective Rec phenotypes (Rec^{-} and Rec$^{\pm}$), and they were mapped in four different functional regions of RecA (Table 2). All eight Rec and Rec$^{\pm}$ mutations were mapped at residues that are either completely conserved or highly conserved among bacterial species; seven of the eight mutations produced a change from nonpolar to polar (Table 2). Thus, the relatively dramatic change in phenotype caused by these mutations can be explained by the fact that almost all cause drastic changes at critical residues. The fact that changes at DNA-binding and intermolecular packing regions can lead to a Rec$^{\pm}$ phenotype is further testimony that DNA binding and filament formation are required for RecA to promote recombination activities (3, 40). Of the eight recombination-defective mutations, four were mapped in the IFCR, strongly suggesting that this region is involved in recombination (see below).

Involvement of IFCR in the recombination activities

While Story et al. (46) emphasized the involvement of the IFCR in the Prt^{C} phenotype, we explored the possibility of IFCR involvement in recombination. Among the 16 distinct missense mutations, 3 changes resulted in a Rec^{-} phenotype (Table 1) Strikingly, all three mutations, *recA1623* (L-182→Q), *recA1631* (I-284→N), and *recA1620* (V-275→N), were mapped in the IFCR. In fact, among the five suppressor mutations mapped in the IFCR, four led to defective Rec phenotypes (Table 2) These observations imply that this region and its putative function in interfilament association play a role in promoting recombination activities (see Discussion).

DISCUSSION

Usefulness of the proximal mutagenic system

The proximal mutagenic system provides a simple way to produce and isolate mutations in the *recA1202* gene and in nearby genes as well. In the presence of a high-copy-number *umuDC* plasmid, the mutation frequency in the *recA1202* gene can be as high as about 5% per generation (26). The proximal mutagenic system is self-controlled because the agent of mutagenesis (the *recA1202* allele) is preferentially targeted for mutation, and frequently such mutations suppress mutagenic activity (26), thereby precluding multiple mutations. In none of the 28 mutants we sequenced was there a multiple mutation in the *recA* gene.

General considerations for a structure-function analysis of the suppressor mutations

Our analysis of the *recA1202* suppressors was organized as an attempt to understand how the structure of RecA might contribute to the coprotease and recombination functions, particularly in the context of the filament-bundle theory proposed by Story et al. (46). First, the possible effects of the original *recA1202* mutation (Q-184→K) on the structure of RecA were examined by using a computer program designed by Lee and Levitt (20). The only predicted effects are in the immediate vicinity of position 184. This appears to be a reasonable prediction inasmuch as Q-184 is on the surface of the protein and the hydrophilic side chain is in solution and not buried; the *recA1202* mutation replaces a polar amino acid with one containing a basic side chain, which would not be likely to have much effect on the hydrophobicity of this region. Thus, the structural change caused by *recA1202* is probably restricted to a small region around residue 184 of the IFCR. This is consistent with a basic part of the theory proposed by Story et al. (46), which explains the phenotypic effects of the *recA1202* mutation: this change at residue 184 alters only the local structure involved in contact between filaments.

We can imagine that the new mutations may suppress the Prt^{c} phenotype of the *recA1202* mutant either (i) directly by a change complementary to Q-184 that essentially restores the original structure or (ii) indirectly by altering a completely separate functional site that reduces the overall coprotease activity. We attempted to distinguish between these two possibilities by determining whether the secondary mutations would be likely to alter the structure in the same region as the original mutation (direct effect) or in some other regions (indirect effect) or both. It should be emphasized, however, that although a mutation may map in what appears to be a distinct functional region of the protein, it could also have effects on other functional regions as well, depending on how drastically it affects the structure and how close it is to other functional regions. Thus, its effect on the phenotype could be due to changes in other functional regions. Nevertheless, when several different mutations causing a similar phenotype all map in the same region, it is likely that the function of that region is directly responsible for the altered phenotype.

Interfilament association and the Prt^{c} activity

It is possible, as indicated in the filament-bundle theory (46), that without an inducing signal(s), RecA^{+} monomers tend to form protein filaments which in turn have the tendency to form multifilament bundles. These protein bundles are a storage from of RecA and are not active in promoting proteolytic reactions (Fig. 4) (46). When a Prt^{c}

mutation occurs in the IFCR sequences (Fig. 2), the following two events leading from a RecA(Prt$^+$) to a RecA(Prt^c) phenotype may both occur: (i) the mutation shifts the equilibrium from bundle formation to favor single filament formation, and (ii) the mutation causes a conformation change that allows the RecA filament to bind to unusual cofactors and thus promote the constitutive cleavage of repressors. Therefore, a second-site mutation at the IFCR in a RecA(Prt^c) protein that shifts the equilibrium back toward bundle formation can suppress the Prt^c phenotype (Fig. 4). The mutagenic effect could produce a phenotype that is either Prt^c (with reduced coprotease strength), Prt$^+$, or Prt$^-$, depending on how much it shifts the equilibrium.

IFCR mutations that also affect the recombinase activity

In addition to the four IFCR suppressor mutations described in Results, there are several previously known *recA* mutations that affect the recombinase activity and also map in the IFCR. *recA803* is a V→M change at residue 37 (32), which is in the N-terminal portion of the IFCR (Fig. 2) (46). The *recA803* mutant is Prt$^+$, but the mutant protein shows a higher-than normal rate of strand pairing and strand exchange activity (32). Thus, not only can mutations in or around the IFCR decrease the recombinase activity, but some can also enhance the RecA recombination activities, an observation in agreement with our suggestion that the IFCR is involved in the recombinase activity of RecA.

A total of three single-site Prt^c Rec$^-$ mutations were previously sequenced by Wang and Tessman: *recA1206* (G-301→D), *recA1601* (G-301→S), and *recA1203* (R-169→C) (54). All three single-site mutations, like the *recA1620*(Prt^{c}Rec$^-$) allele, mapped in or near the IFCR, where changes could cause defective association between filaments (Fig. 2) (46). Therefore, it appears that the IFCR and its possible function of interfilament association are important in regulating proteolytic activity and promoting recombination activity. It should be noted that the mutation site of *recA1203*, residue 169, is also close to a proposed DNA-binding site (46), and its change could also affect the DNA-binding activity.

Roles of interfilament association: bundles and bundle-like structure

Why is the IFCR important for the recombinase activity? How can single point mutations at the IFCR give rise to a Prt^{c} Rec^{-} split-phenotype mutant protein? These observations can be explained if we assume the following: (i) Prt^{c} mutations affect the IFCR structure to favor the formation of the active filament (containing DNA) rather than the formation of the pure protein bundles; (ii) the pairing of homologous DNA strands and the subsequent strand exchange, which are required in part for the Rec^{+} phenotype, are enhanced when the RecA-DNA combination forms a bundle-like structure; and (iii) the bundle-like structure is similar, but not identical, to the inactive form of bundle proposed by Story et al. (Fig. 4) (46) that is inactive for both coprotease and recombinase activities. Thus, if RecA bundle formation reduces coprotease activity but a similar bundle-like RecA-DNA structure promotes homologous recombination, mutations that hinder the formation of both bundle structures can produce a Prt^{c} Rec phenotype. These mutations could favor the formation of individual active filaments, which presumably results in enhanced coprotease activity; in our model these mutations would also remove the RecA-DNA bundle-like structures required for recombination (Fig. 4), thereby producing the Prt^{c} Rec^{-} split-phenotype effect. Mutations in the IFCR that block the formation of only one of the two bundle structures will give rise to a phenotype that is either Prt^{c} Rec^{-} (defective only in bundle formation) or Prt^{+} Rec^{-} (defective in the formation of the RecA-DNA bundle-like structure).

The bundle-like structure proposed here may be a transient and dynamic aggregate of RecA-DNA filaments. In order to complete the pairing of homologous DNA strands, hydrogen bonds are likely to be formed between complementary DNA strands, which could require dynamic exchanges between two or more multifilament structures. The relatively weak contact between RecA filaments, as compared with that between monomers within a filament, may be an important factor in acquiring these dynamics. Specifically, we postulate a sandwich form of the bundle-like structure, consisting of RecA-DNA filament-DNA-RecA-DNA filament, which may continuously exchange with the other form of bundle-like structure composed of only RecA-DNA filaments (Fig. 4). The DNA wrapped in the RecA-DNA filament could be either ssDNA or double-stranded DNA (dsDNA).

Several observations are consistent with the notion that the RecA-DNA filament can associate not only with another RecA-DNA filament but also with a DNA (ssDNA or dsDNA) molecule at the IFCR. First, the C-terminal portion of the IFCR between residues 280 and 310 is rich in basic and aromatic residues and fits the sequence of the DNA-binding domain of some DNA-binding proteins (35, 37). In addition, the *recA441* allele, which consists of two missense mutations (53) that are both in the IFCR (46), codes for a mutant RecA protein that has altered DNA-binding kinetics (28, 30). The RecA1202 and RecA1211 Prt^{c} proteins, whose alterations are at different parts of the IFCR (Fig. 2) (46), also have enhanced DNA-binding activity (52, 55). Furthermore, in the three-dimensional crystal structure of the IFCR, most of the C-terminal residues described above, and the region around residue 184**,** are not buried in the intersurface between contacting filaments (Fig. 2). In fact, the two regions defined by residues around 184 and nonintersurface residues between 280 and 310 look like two sides of a small pocket in the three-dimensional structure, and the small pocket is big enough to provide a binding site for an ss- or dsDNA molecule (Fig. 2). Thus, the binding of a RecA-DNA filament to DNA may be enhanced, rather than excluded, by the association between two such filaments. This could explain, in part, the importance of the postulated RecA-DNA bundle-like structure in recombination.

The bundle-like structure may be formed after the pairing of complementary DNA strands, and its involvement in the strand exchange reactions could still require a dynamic structure. Mutations that affect the dynamics and flexibility of this structure could result in a Rec^{-} phenotype. Such mutations could actually stabilize the transient association between RecA filaments and reduce the coprotease activity as well. All three Rec^{-} suppressors isolated in this study resulted from nonpolar-to-polar changes at residues involved in interfilament association (Table 2). These mutations could significantly disrupt the integrity of the IFCR and form additional hydrogen bonds, leading to the stabilization effect described above.

The RecA-DNA bundle-like structure that we postulate and its involvement in strand pairing and/or exchange activities is supported by several lines of evidence. In a study designed to understand how RecA promotes the alignment of homologous DNA strands, it has been observed that under strand-pairing conditions, RecA-ssDNA filaments and heterologous dsDNA formed coaggregates (50). This aggregation may provide a concentration effect that facilitates the search for homologous DNA sequences (15). This coaggregate structure is conceivably a form of the proposed bundle-like structure,. Furthermore, in various electron microscopy studies of the structure of the RecA-DNA filament, bundle formation from RecA-DNA filaments has been repeatedly observed (10, 11). In the presence of Mg^{2+} and ATP-γS, RecA-DNA filaments regularly aggregate into bundles composed typically of three, or six RecA-DNA filaments (10).

If formation of a bundle-like structure is important for promoting homologous recombination, how can one reconcile this with the data showing that RecA protein monomers truncated at the C terminus can still have recombinase activity, both in vivo (19) and in vitro (2)? The answer may lie in the electron microscopic study of Yu and Egelman (59). In their study of the conformational change of a truncated RecA protein, it was found that removal of 18 residues from the C terminus of the RecA protein results in a significant change in the structure of the RecA-DNA filament; a 15-Å (1.5-nm) outward (from the DNA axis) movement of an inner domain is observed, but RecA monomers still form a stable filament complex with DNA (59). Their result suggests that the RecA structure is flexible in forming a RecA-DNA filament. Our inability to isolate Prt^- suppressor mutations in regions involved in DNA binding and filament formation (see below) is also consistent with this implied flexibility. It is conceivable, therefore, that a 50-amino-acid deletion at the C terminus (2) may result in a mutant RecA protein whose conformation is altered even more dramatically than that of the truncated RecA with 18 amino acids removed from its C terminus. The drastic conformation change caused by the 50-amino-acid deletion (2) could allow the formation of RecA-DNA filaments which in turn are capable of forming the bundle-like structure and promoting homologous recombination. Without such a notable change in conformation, most point mutations affecting the interfilament association will not allow efficient formation of the bundle-like structure, and, as a result, they will lead to defective recombinase activity. Thus, the evidence from the studies on the deletion mutants is not sufficient to rule out our theory that parts of the C-terminal domain containing the IFCR are important for recombination. In any case, it appears that conclusions from structure-function analysis based on deletions, especially multiresidue deletions, could be misleading.

Flexibility of RecA1202 structure in promoting repressor cleavage

It is intriguing that 11 mutations, consisting of all 5 mutations mapped in the DNA- or ATP-binding sites and all 7 mutations in repressor-binding and intermolecular packing regions, failed to reduce the coprotease activity to a level less than that of RecA^{+} (Tables 1 and 2). This is significant because most of the amino acid substitutions caused by the 11 mutations are in themselves rather severe: 7 of the 11 substitutions are either nonpolar-to-polar or polar-to-nonpolar changes, and I of the other 4 substitutions is from a moderately polar to an extremely polar residue (Asn→Lys) (Table 2). This suggests flexibility in the three-dimensional structures required for DNA and repressor binding and subsequent repressor cleavage. If this is the case, most single-residue alterations in these regions are not likely to cause a dramatic change in the protein structure and function and resultant coprotease activity. An alternative explanation for this specific lack of Prt^{-} mutations is that such mutations confer a selective disadvantage under our experimental procedures. We do not favor this alternative hypothesis because we were able to isolate many Prt^{-} mutants; among 20 distinct mutants isolated, 6 showed a coprotease activity much weaker than that of the wild type (5 Prt^{-} and 1 Prt$^{\pm}$ [Table 1]). This number is equal to the number of Prt^{c} mutants isolated and only slightly lower than the number of Prt^{+} mutants isolated (Tables 1 and 2). Thus, while Prt^{-} mutants can be easily isolated by our procedure, 11 mutations in the three regions described above did not produce a Prt^{-} phenotype.

Furthermore, one of the nonsense mutations, *recA1640* (E-259→Oc), mapped at residue 259, and the resulting RecA is missing 94 C-terminal residues (Fig. 1b and Table 2). Surprisingly, this mutant protein could still be partially activated to a coprotease by UV irradiation (16 J/m^{2}) of the cell; it induced W to rise from 0.01 to 0.08, which can be compared with W = 0. 18 for UV-activated RecA^{+} protein (27). It seems, therefore, that the coprotease activity of RecA1202 is "buffered" by structural flexibility in maintaining the appropriate conformation to promote the cleavage of LexA repressor. Amino acid substitutions at a number of sites, including all the missense mutations at intermolecular packing, repressor-binding, and DNA-binding sites, and the 94-amino-acid truncation (caused by the *recA1640* mutation) at the C terminus do not eliminate the coprotease activity completely. It is unclear whether such a structure-function flexibility also exists in the RecA$^+$ protein. If so, it would be consistent with the evidence that RecA is not a true protease, but rather a coprotease (20, 23, 44), and can play a role in proteolytic reactions that is relatively easy to fulfill.

Inducibility of mutant proteins by MitC treatment

It is intriguing that many mutations mapped in the putative sites for repressor binding, intermolecular packing, and DNA binding lead to a coprotease that is inducible by MitC but not by constitutive cofactors such as nucleoside triphosphates and RNAs (Tables 1 and 2). One possibility is that cofactors for constitutive activation of RecA1202 (nucleoside triphosphates and RNAs) may activate RecA1202 to a conformation different from that activated by damaged DNA; mutations in different regions of RecA1202 may exaggerate this difference.

Conclusions

Analyses of our newly isolated suppressors of the *recA1202*(Prt^c) mutant provide additional evidence for the molecular mechanism proposed by Story et al. (46), by which *recA* mutations can lead to a Prt^c phenotype. Of 16 distinct missense suppressor mutations, 5 were mapped in the proposed IFCR, and 4 of them markedly changed the parental Prt^c Rec$^+$ phenotype of *recA1202* (Q-184→K) (Table 2). Two of these five suppressor mutations were mapped in the RecA C-terminal domain. This small domain, according to the 2.3-Å X-ray crystal structure, does not interact intramolecularly with the large central domain where the *recA1202* mutation resides. The two mutations were located in a part of the IFCR that apparently can interact intermolecularly with residue 184 in an adjacent, contacting filament (Fig. 2). Thus, alterations in interfilament association could lead to a mutant RecA with a Prt^c phenotype and also to a reduction of the Prt^c activity of an existing Prt^c protein. A major part of the model proposed by Story et al. was based on the sequencing studies of the *recA441* double-mutant allele (E-38→K and I-298→V) (53), a temperature-dependent Prt^c mutant. The first mutation confers a Prt^c phenotype, and the second one is a temperature-sensitive suppressor (53); though far apart in the linear sequence, both mutations map in the proposed IFCR of the crystal (46). Our finding of more suppressor mutations mapped in the same regions strengthens the idea that alterations such as the suppressor mutations may favor the formation of inactive bundles and reduce the formation of individual filaments that produce the Prt^c phenotype.

We also found that three suppressor mutations give rise to the Rec$^-$ phenotype; surprisingly, all three map in the IFCR. In addition, all the sequenced Prt^c Rec$^-$ mutations, including one reported in this study, map in or near the IFCR. Therefore, it appears that the IFCR is involved in both proteolytic and recombination activities. To account for the observations, we propose that while shifting from inactive RecA bundles to the active RecA-DNA single filaments is essential for coprotease activity, formation of a RecA-DNA bundle-like structure is required for RecA-promoted recombination activities such as strand pairing and/or strand exchange. For our analysis of the possible composition of the bundle-like structure, we suggest that a form of the structure contains a DNA molecule between two contacting RecA-DNA filaments (Fig. 4). Determination of the exact conformation and functions of this hypothetical bundle-like structure awaits further studies, particularly on the Prt^c Rec$^-$ mutants.

In addition to mutations mapped in the IFCR, including three located in the vicinity of *recA1202,* there are suppressor mutations mapped in sites possibly involved in repressor binding, intermolecular packing, and binding to DNA and ATP. Most of these mutation sites do not seem to interact with residue 184 either intra- or intermolecularly. Thus, the conformations of these putative binding or packing sites have to be appropriately maintained to form a constitutively active coprotease. The fact that there were no Prt$^-$ mutations among 11 distinct mutations mapped in the above-mentioned four sites suggests that the RecA protein is structurally flexible in its ability to form a RecA-DNA filament upon activation and to bind to repressors to promote the subsequent cleavage.

ACKNOWLEDGMENTS

We are grateful to several people for their kind assistance: Michael Eisen for help with downloading the structural information and converting it to a useful format, Christopher Lee for help with details of the Midas program and for discussions on the interpretations of structural changes, Matthew A. Kennedy for measuring the repair sector of some of the suppressor mutants, Ann Ganesan for critical reading of the manuscript, and Michael Levitt and Marc Feldman for use of computer equipment. The reviewers were particularly helpful in their criticisms. This work was supported, in part, by a grant (GM35850) from the National Institute of General Medical Services (to I.T.), an Outstanding Investigator grant (CA44349) from the National Cancer Institute (to P.C.H.), and an NSF predoctoral fellowship (to J.A.E.).

REFERENCES

1. Bagg, A., C. J. Kenyon, and G. C. Walker. 1981. Inducibility of a gene product required for UV and chemical mutagenesis in *Escherichia coli.* Proc. Natl. Acad. Sci. USA 78:5749-5735.
2. Benedict, R. C., and S. C. Kowalczykowski. 1988. Increase of the DNA strand assimilation activity of *recA* protein by removal of the C terminus and structure-function studies of the resulting protein fragment. J. Biol. Chem. 263:15513-15520.
3. Brenner, S. L., A. Zlotnick, and J. D. Griffith. 1988. RecA protein self-assembly multiple discrete aggregation states. J. Mol. Biol. 204:959-972.
4. Brent, R., and M. Ptashne. 1980. The *lexA* gene product represses its own promoter. Proc. Natl. Acad. Sci. USA 77:1932-1936.
5. Burckhardt, S. E., R. Woodgate, R. H. Scheuermann, and H. Echols. 1988. UmuD mutagenesis of *Escherichia coli* and cleavage by RecA. Proc. Natl. Acad. Sci. USA 85:1811-1815.
6. Castellazzi, M., J. George, and G. Buttin. 1972. Prophage induction and cell division in *E. coli.* I. Further characterization of the thermosensitive mutation *tif-1* whose expression mimics the effects of UV irradiation. Mol. Gen. Genet. 119:139-152.
7. Clark, A. J., and A. D. Margulies. 1965. Isolation and characterization of recombination-deficient mutants of *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA 53:451-459.
8. Craig, N. L., and J. W. Roberts. 1980. *E. coli recA* protein-directed cleavage of phage lambda repressor requires polynucleotide. Nature (London) 283:26-30.

9. Dutreix, M., P. L. Moreau, A. Bailone, F. Gailbert, J. R. Battista, G. C. Walker, and R. Devoret. 1989. New RecA mutations that dissociate the various RecA protein activities in *Escherichia coli* provide evidence for an additional role for RecA protein in UV mutagenesis. J. Bacteriol. 171:2415-2423.

10. Egelman, E. H., and A. Stasiak. 1986. Structure of helical RecADNA complexes: complexes found in the presence of ATPγS filaments. J. Mol. Biol. 191:677-697.

11. Egelman, E. H., and A. Stasiak. 1988. Structure of helical RecADNA complexes. II. Local conformation changes visualized in bundles of RecA-ATP-γS filaments. J. Mol. Biol. 200:329-349.

12. Elledge, S. J., and G. C. Walker. 1983. Proteins required for ultraviolet light and chemical mutageiiesis: identification of the products of the *umuC* locus of *Escherichia coli*. J. Mol. Biol. 164:175-192.

13. Gardener, R. C., A. J. Howarth, P. Hahn, M. Brown-Luedi, R. J. Shepherd, and J. Messing. 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13 mp7 shotgun sequencing. Nucleic Acid Res. 9:2871-2880.

14. Goldthwait, D., and F. Jacob. 1964. Sur le méchanisme de l'induction du développement du prophage chez les bacteries lysogenes. C.R. Acad. Sci. 259:661-664.

15. Gonda, D. K., and C. M. Radding. 1986. The mechanism of the search for homology promoted by *recA* protein. J. Biol. Chem. 261:13087-13096.

16. Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Applic. Biosci. 8:189-191.

17. Kenyon, C. J., and G. C. Walker. 1980. DNA-damaging agents stimulate gene expression at specific loci in *Escherichia coli.* Proc. Natl. Acad. Sci. USA 77:2819-2823.

18. Kowalczvkowski, S. C. 1987. Mechanistic aspects of the DNA strand exchange activity of *E. coli* RecA protein. Trends Biochem. 12:141-145.

19. Larminat, F., and M. Defais. 1989. Modulation of SOS response by truncated RecA protein. Mol. Gen. Genet. 216:106-112.

20. Lee, C., and M. Levitt. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. Nature (London) 352:448-451.

21. Little, J. W. 1984. Autodigestion of LexA and phage λ repressor. Proc. Natl. Acad.Sci. USA 81:1375-1379.

22. Little, J. W. 1982. Control of the SOS regulatory system by the level of RecA protease. Biochimie 64:585-589.

23. Little, J. W. 1991. Mechanism of specific LexA cleavage: autodigestion and the role of RecA coprotease. Biochimie 73:411-422.

24. Little, J. W., S. H. Edmiston, L. Z. Pacelli, and D. E. Mount. 1980. Cleavage of the *Escherichia coli* LexA protein by the RecA protease. Proc. Natl. Acad. Sci. USA 77:3225-3229.

25. Little, J. W., D. W. Mount, and C. R. Yanisch-Perron. 1981. Purified LexA protein is a repressor of the *recA* and *lexA* genes. Proc. Natl. Acad. Sci. USA 78:4199-4203.

26. Liu, S.-K., and I. Tessman. 1990. Mutagenesis by proximity to the *recA* gene of *Escherichia coli*. J. Mol. Biol. 211:351-358.

27. Liu, S.-K., and I. Tessman. 1990. *groE* genes affect SOS repair in *Escherichia coli*. J. Bacteriol. 172:6135-6138.

28. Lu, C., and H. Echols. 1987. RecA protein and SOS. Correlation of mutagenesis phenotype with binding of mutant RecA proteins to duplex DNA and LexA cleavage. J. Mol. Biol. 196:497-504.

29. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

30. McEntee, K., and G. M. Weinstock. 1981. *tif-1* mutation alters polynucleotide recognition by the RecA protein of *Escherichia coli*. Proc. Natl. Acad. Sci. USA 78:6061-6065.

31. Monk, M., and J. Kinross. 1972. Conditional lethality of *recA* and *recB* derivatives of a strain of *Escherichia coli* K-12 with a temperature-sensitive deoxyribonucleic acid polymerase I. J. Bacteriol. 109:971-978.

32. Murty, V. V., S. Madiraju, A. Templin, and A. J. Clark. 1988. Properties of a mutant *recA*-encoded protein reveal a possible role for Escherichia coli *recA*-encoded protein in genetic recombination. Proc. Natl. Acad. Sci. USA 85:6592-6596.

33. Nohmi, T., J. R. Battista, L. A. Dodson, and G. C. Walker. 1988. RecA-mediated cleavage activates UmuD for mutagenesis: mechanistic relationship between transcriptional depression and posttranslational activation. Proc. Natl. Acad. Sci. USA 85:1816-1820.

34. Ogawa, H., and T. Ogawa. 1986. General recombination: functions and structure of RecA protein. Adv. Blophys. 21:135-148.

35. Peterson, K. R., N. Ossanna, A. R. Thliveris, D. G. Ennis, and D. W. Mount. 1988. Derepression of specific genes promotes DNA repair and mutagenesis in *Escheichia coli*. J. Bacteriol. 170:1-4.

36. Priestle, J. P. 1988. Ribbon, a stereo cartoon drawing program for proteins. J. Appl. Crystallogr. 21:572-576.

37. Radding, C. M. 1982. Homologous pairing and strand echange in genetic recombination. Annu. Rev. Genet. 16:405-437.

38. Roberts, J. W., C. W. Roberts, and N. L. Craig. 1978. *Escherichia coli recA* gene product inactivates phage lambda repressor. Proc. Natl. Acad. Sci. USA 75:4714-4718.

39. Roberts, J. W., C. W. Roberts, N. L. Craig, and E. M. Phizicky. 1979. Activities of the *Escherichia coli recA* gene product. Cold Spring Harbor Symp. Quant. Biol. 43:917-920.

40. Roca, A. I., and M. M. Cox. 1990. The RecA protein: structure and function. Crit. Rev. Biochem. Mol. Biol. 25:415-456.

41. Sancar, A., A. Stachelek, W. Konigsberg, and W. D. Rupp. 1980. Sequences of the *recA* gene and protein. Proc. Natl. Acad Sci. USA 77:2611-2615.

42.Schuster, H., D. Beyersmann, M. Mikolajczyk, and M. Schlicht. 1973. Prophage induction by high temperature in thermosensitive *dna* mutants lysogenic for bacteriophage lambda. J. Virol. 11:879-885.

43.Shinagawa, H., H. Iwasaki, T. Kato, and A. Nakata. 1988. RecA protein-dependent cleavage of UmuD protein and SOS mutagenesis. Proc. Natl. Acad. Scl. USA 85:1806-1810.

44. Slilaty, S. N., and J. W. Little. 1987. Lysine-156 and serine-119 are required for LexA repressor cleavage: a possible mechanism. Proc. Natl. Acad. Sci. USA 84:3987-3991.

45.Story, R. M., and T. A. Steitz. 1992. Structure of the RecA protein-ADP complex. Nature (London) 355:374-376.

46.Story, R. M., I. T. Weber, and T. A. Steitz. 1992. The structure of the *E. coli* RecA protein monomer and polymer. Nature (London) 355:318-325.

47.Tessman, E. S., and P. Peterson. 1985. Plaque color method for rapid isolation of novel *recA* mutants of *Escherichia coli* K-12: new classes of protease-constitutive *recA* mutants. J. Bacteriol. 163: 677-687.

48.Tessman, E. S., and P. K. Peterson. 1985. Isolation of proteaseproficient, recombinase-deficient *recA* mutants of *Escherichia coli* K-12. J. Bacteriol. 163:688-695.

49.Tessman, E. S., I. Tessman, P. K. Peterson, and J. D. Forestal. 1986. Roles of RecA protease and recombinase activities of *Escherichia coli* in spontaneous and UV-Induced mutagenesis and in Weigle repair. J. Bacteriol. 168:1159-1164.

50.Tsang, S. S., S. A. Chow, and C. M. Radding. 1985. Networks of DNA and RecA protein are intermediates in homologous pairing. Biochemistry 24:3226-3232.

51.Walker, G. C. 1984. Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli.* Microbiol. Rev. 48:60-93.

52.Wang, W.-B., M. Sassanfar, I. Tessman, J. W. Roberts, and E. S. Tessman. 1988. Activation of protease-constitutive RecA proteins of *Escherichia coli* by all of the common nucleoside triphosphates. 170:4816-4822.

53.Wang, W.-B., and E. S. Tessman. 1985. Evidence that the *recA441 (tif-1)* mutant of *Escherichia coli* K-12 contains a thermosensitive intragenic suppressor of RecA constitutive protease activity. J. Bacteriol. 163:407-409.

54.Wang, W.-B., and E. S. Tessman. 1986. Location of functional regions of the *Escherichia coli* RecA protein by DNA sequence analysis of RecA protease-constitutive mutants. J. Bacteriol. 168: 901-910.

55.Wang, W.-B., E. S. Tessman, and I. Tessman. 1988. Activation of protease-constitutive RecA proteins of *Escherichia coli* by rRNA and tRNA. J. Bacteriol. 170:4823-4827.

56.West, S. C. 1988. Protein-DNA interactions in genetic recombination. Trends Genet. 4:8-13.

57.Wilson, D. H., and A. S. Benight. 1990. Kinetic analysis of the preequilibrium steps in the self-assembly of RecA protein from *Escherichia coli.* J. Biol. Chem. 265:7351-7359.

58.Woese, C. 1987. Bacterial evolution. Microbiol. Rev. 51:221-271.

59.Yu, X., and E. H. Egelman. 1991. Removal of the RecA C-terminus results in a conformational change in the RecA-DNA filament. J. Struct. Biol. 106:243-254,

60. Yu, X., and E. H. Egelman. 1993. The LexA repressor binds within the deep helical groove of the activated RecA filament. J. Mol. Biol. 231:29-40.

Table 1. Properties of the *recA* mutants

		Color on M9-CAA-X-Gal[1]:				
Strain	Allele[2]	w/o MitC	w/ MitC[3]	Sensitive to UV[4]	W[5] (±0.01)	Mutant frequency[6] (Rif^{S}->Rif^{R})
Prt^{C} Rec^{+}						
IT3174	*recA1624*	PB+	B	R	0.06	(7.4 ± 0.5) x 10^{-8}
IT3175	*recA1625*	B	DB	R	0.17	(3.2 ± 1.3) x 10^{-7}
IT3186	*recA1626*(2)	B	B	R	0.20	(4.1 ± 1.1) x 10^{-7}
IT3194	*recA1634*	PB+	B	R	0.02	(1.9 ± 0.7) x 10^{-7}
IT3167	*recA1647*(2)	PB+	B	R	0.06	(8.2 ± 1.1) x 10^{-8}
Prt^{C} Rec^{-}						
IT3170	*recA1620*(2)	B	S	S	0.04	(2.5 ± 1.2) x 10^{-8}
Prt^{+} Rec^{+}						
IT3172	*recA1622*	PB	B	R	<0.01	(3.6 ± 0.7) x 10^{-8}
IT3190	*recA1641*	PB+	DB	R	<0.01	(4.4 ± 1.9) x 10^{-8}
IT3162	*recA1642*(3)	PB	B	R	0.01	(5.1 ± 0.8) x 10^{-8}
Prt^{+} Rec$^{\pm}$						
IT3177	*recA1627*(2)	PB	B	R/S	<0.01	< 3.0 x 10^{-8}
IT3178	*recA1628*	PB	B	R/S	<0.01	(1.6 ± 0.4) x 10^{-8}
IT3180	*recA1630*	PB	B	R/S	<0.01	(1.2 ± 0.1) x 10^{-8}
IT3196	*recA1636*	PB	B	R/S	<0.01	(2.0 ± 0.6) x 10^{-8}
IT3169	*recA1649*(2)	PB+	B	R/S	<0.01	(1.5 ± 0.2) x 10^{-8}
Prt$^{\pm}$ Rec^{-}						
IT3200	*recA1640*	PB+	S	S	0.01	< 1.6 x 10^{-8}
Prt^{-} Rec^{-}						
IT3173	*recA1623*(2)	PB+	S	S	<0.01	< 1.5 x 10^{-8}
IT3181	*recA1631*	PB+	S	S	<0.01	< 1.7 x 10^{-8}
IT3193	*recA1633*	PB	S	S	0.01	(3.4 ± 1.3) x 10^{-8}
IT3195	*recA1635*	PB	S	S	<0.01	(1.7 ± 0.9) x 10^{-8}
IT3163	*recA1643*	PB+	S	S	0.01	< 3.5 x 10-8
Reference						
EST2411	Δ*recA306*	PB-	S	S	<0.01	NT[7]
EST2422	*recA*$^{+}$	PB	B	R	<0.01	(1.1 ± 0.4) x 10^{-8}
IT3111	*recA1202*	DB	DB	R	0.26	(2.6 ± 0.3) x 10^{-6}

[1] PB+, between PB and B; PB-, between PB and white

[2] Multiple occurrences are indicated in parentheses. All mutations are missense except one ochre mutation (*recA1640*), one amber mutation (*recA1635*), and one deletion (*recA1643*).

[3] MitC was used at 0.5 ug/ml. S, sensitive to MitC and failed to grow.

[4] R, resistant; S, sensitive; R/S intermediate sensitivity as described in Materials and Methods

[5] The value of *W* is the repair sector for UV-irradiated phage S13. $W = 1 - \log S_a/\log S_b$, where S_a and S_b are the fraction of viruses surviving after and before reactivation, respectively. The viral survival after UV irradiation was between 1.5 x 10^{-6} and 3.9 x 10^{-7}. The *recA* mutant cells were not irradiated.

[6] Rifr mutant frequencies were determined by growing cells in M9-CAA from small inocula (1,000-2,000 cells) to mid-log phase at 32°C. The cells were then spread on LB and LB-rifampin (25 ug/ml) plates and incubated at 32°C for 24h. Each value is the average from three cultures ± standard error of the mean.

[7] NT, not tested

Table 2. Amino acid substitutions of the *recA* mutants classified by the putative functions affected by the mutations.

Allele[1]	Phenotype	Amino acid no.	Amino acid change	Evolutionary conservation[2]
IFCRs				
recA1620	Prt^{c} Rec^{-}	275	Val -> Asp	HC
recA1623	Prt^{-} Rec^{-}	182	Leu -> Gln	HC
recA1625	Prt^{c} Rec^{+}	187	Thr -> Ala	HC
recA1630	Prt^{+} Rec$^{\pm}$	177	Lys -> Gln	CC
recA1631	Prt^{-} Rec^{-}	284	Ile -> Asn	HC
Repressor-binding sites				
recA1627	Prt^{+} Rec$^{\pm}$	244	Val -> Glu	HC
recA1628	Prt^{+} Rec$^{\pm}$	231	Val -> Glu	HC
recA1642	Prt^{+} Rec^{+}	238	Val -> Glu	MC
Intermolecular packing regions				
recA1622	Prt^{+} Rec^{+}	111	Ile -> Met	HC
recA1626	Prt^{c} Rec^{+}	139	Asp -> Gly	CC
recA1634	Prt^{c} Rec^{+}	213	Asn -> Lys	MC
recA1636	Prt^{+} Rec$^{\pm}$	132	Leu -> Gln	CC
DNA-binding sites				
recA1624	Prt^{c} Rec^{+}	208	Thr -> Asn	CC
recA1634[3]	Prt^{c} Rec^{+}	213	Asn -> Lys	MC
recA1641	Prt^{+} Rec^{+}	152	Lys -> Ile	HC
recA1649	Prt^{+} Rec$^{\pm}$	214	Ala -> Ser	CC
ATP-binding site				
recA1647	Prt^{c} Rec^{+}	145	Ser -> thr	CC
Others				
recA1633	Prt^{-} Rec^{-}	-12[4]	Gln -> Am	
recA1635	Prt^{-} Rec^{-}	84	Glu -> Oc	
recA1640	Prt$^{\pm}$ Rec^{-}	259		

[1] Unlisted is the *recA1643* mutations, which is a 10-base deletion between T-446 and T-456 in the sequences 5'CGCCCT-446GGCGCGTTCT-456GGCG-3'; this sequence contains a 4-base direct repeat and a 4-base inverted repear, either of which might coneivably encourage formation of the deletion.

[2] CC, completely conserved; HC, highly conserved; MC, moderately conserved.

[3] The mutation site of *recA1634*, residue 213, is likely to be involved in both intermolecular packing and DNA binding.

[4] Refers to the base at -12 in the *recA* promoter.

Table 3: Phenotype of double mutants, and site and possible function of mutated residues.

Allele	Phenotype	Residue Number	Amino Acid Change	Conservation[1]	Putative Function of Residue / Region[2]
recA1626	Prt^{c} Rec^{+}	139	Asp→Gly	C	Filament Formation
recA1647	Prt^{c} Rec^{+}	145	Ser→Thr	C	ADP Binding
recA1625	Prt^{c} Rec^{+}	187	Thr→Ala	H	Interfilament Contact
recA1624	Prt^{c} Rec^{+}	208	Thr→Asn	C	DNA Binding
recA1634	Prt^{c} Rec^{+}	213	Asn→Lys	M	DNA Binding
recA1622	Prt^{+} Rec^{+}	111	Ile→Met	H	Filament Formation
recA1641	Prt^{+} Rec^{+}	152	Lys→Ile	H	DNA Binding
recA1642	Prt^{+} Rec^{+}	238	Val→Glu	M	Repressor Binding
recA1636	Prt^{+} $\mathrm{Rec}^{\pm}$	132	Leu→Gln	C	Filament Formation
recA1630	Prt^{+} $\mathrm{Rec}^{\pm}$	177	Lys→Gln	C	Interfilament Contact
recA1649	Prt^{+} $\mathrm{Rec}^{\pm}$	214	Ala→Ser	C	DNA Binding
recA1628	Prt^{+} $\mathrm{Rec}^{\pm}$	231	Val→Glu	H	Repressor Binding
recA1627	Prt^{+} $\mathrm{Rec}^{\pm}$	244	Val→Glu	H	Repressor Binding
recA1620	Prt^{c} Rec^{-}	275	Val→Asp	H	Interfilament Contact
recA1623	Prt^{-} Rec^{-}	182	Leu→Gln	H	Interfilament Contact
recA1631	Prt^{-} Rec^{-}	284	Ile→Asn	H	Interfilament Contact

[1] C=Complete conservation. H=high (conservative alterations only). M=moderate (some non-conservative alterations)

[2] Putative functions of regions as suggested by Story et al (1992).

Figure 1. RecA structure and sites of the suppressor mutations.

A. RecA monomer structure. The original coordinates (46) were obtained from the Brookhaven protein data base. The three-dimensional structure was drawn using the computer program RIBBON (36). The labeling of secondary structure elements is according to Story et al. (46). The 10 α-helices are lettered A to J; the 11 b-strands are numbered 0 to 10. The numbers of the approximate first and last amino acid residues of each helix and strand are indicated, and the numbering is identical to that in panel b, which shows the mutations sites. The two disordered loops (L1 and L2) suggested as DNA-binding sites are labeled with dashed lines. The disordered N and C termini are not shown.

B. Map of the sites of the suppressor mutations in the RecA protein. Not shown is the *recA1643* mutation, which is a 10-bp deletion (Table 2) that alters the reading frame following codon 69 for proline.

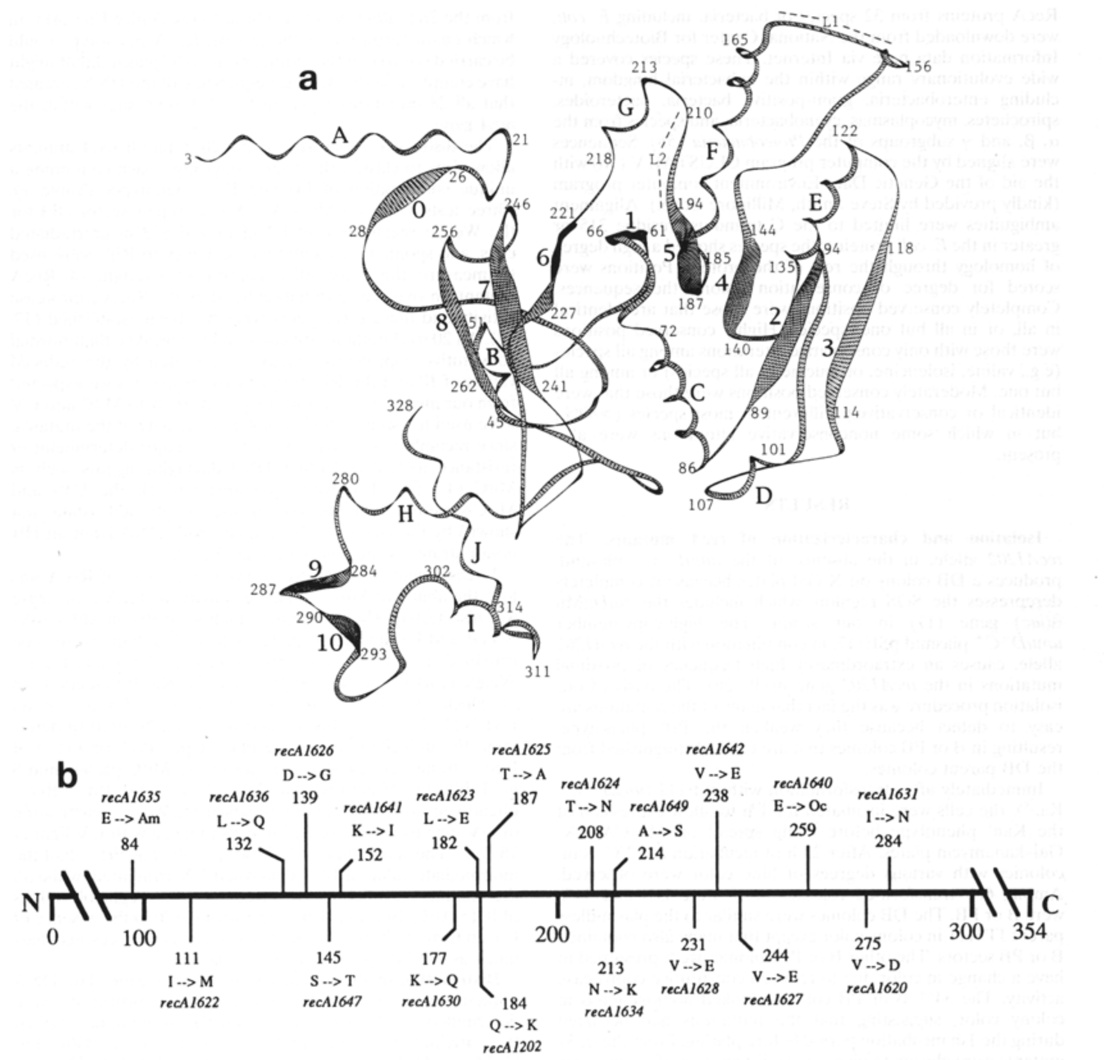

a
3
A
21
26
0
28
246
256
7
8
51
B
262
45
241
227
6
221
66
1
61
72
213
G
218
210
L2
F
194
5
185
187
4
144
135
2
140
L1
165
156
122
E
94
118
3
114
C
89
86
101
107
D
328
280
H
J
302
9
284
287
290
10
293
I
314
311
b
N
0
100
200
300
354
C
recA1635
E --> Am
84
111
I --> M
recA1622
recA1636
L --> Q
132
recA1626
D --> G
139
145
S --> T
recA1647
recA1641
K --> I
152
177
K --> Q
recA1630
recA1623
L --> E
182
184
Q --> K
recA1202
recA1625
T --> A
187
recA1624
T --> N
208
213
N --> K
recA1634
recA1649
A --> S
214
231
V --> E
recA1628
recA1642
V --> E
238
244
V --> E
recA1627
recA1640
E --> Oc
259
275
V --> D
recA1620
recA1631
I --> N
284

Figure 2. Sites of mutations at the IFCR in the RecA crystal.

The side chains are shown for the wild-type amino acids. The Prt^c Rec$^+$ mutation sites are *recA1202* (Q-184→K), *recA1211* (E-38→K), *recA1235* (T-39→I) (54), *recA1625* (Q-184→K, T-187→A) (Table 2), and the temperature-dependent Prt^C allele *recA441* (E-38→K, I-298→V) (53). Prt^c Rec$^+$ mutations also mapped in the IFCR but not shown are *recA1222* (S-25→F) and *recA1213* (A-179→V) (54). Prt^c Rec$^-$ mutation sites are *recA1620* (Q-184→K, V-275→D) (Table 2), *recA1601* (G-301→S), *recA1206* (G-301→D), and *recA1203* (R-169→C) (54). Prt$^-$ Rec$^-$ mutations sites are *recA1623* (Q-184→K, L-182→K) and *recA1631* (Q-184→K, I-284→N) (Table 2). The *recA1630* (Q-184→K, K-177→Q) allele showed a Prt^c Rec$^\pm$ phenotype (Table 2). The *recA803* mutation at residue 37 (V-37→M) enhances the recombinase activity (32). The arrow indicates the location of the immediate contact points (intersurface) between two contacting filaments (A and B). The actual contacting amino acids, determined from our three-dimensional analysis with the computer program Midas, are residues 12, 15, 16, 19, 23, 33, 35, 36, 38, and 60 in the N-terminal domain, residue 183 in the central domain, and residues 290, 294, 296, 297, 298, 308, 311, 312, and 314 in the C-terminal domain.

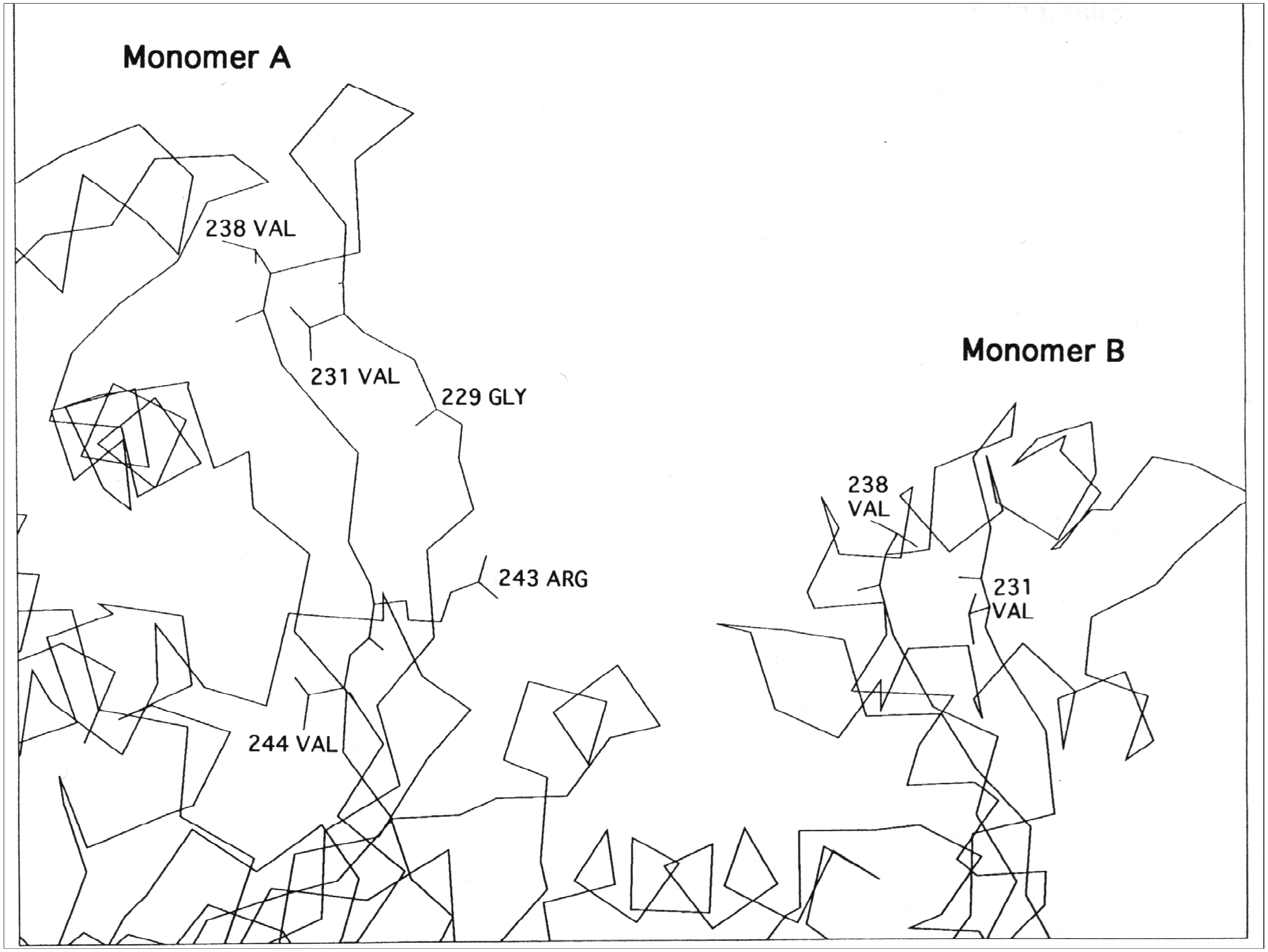
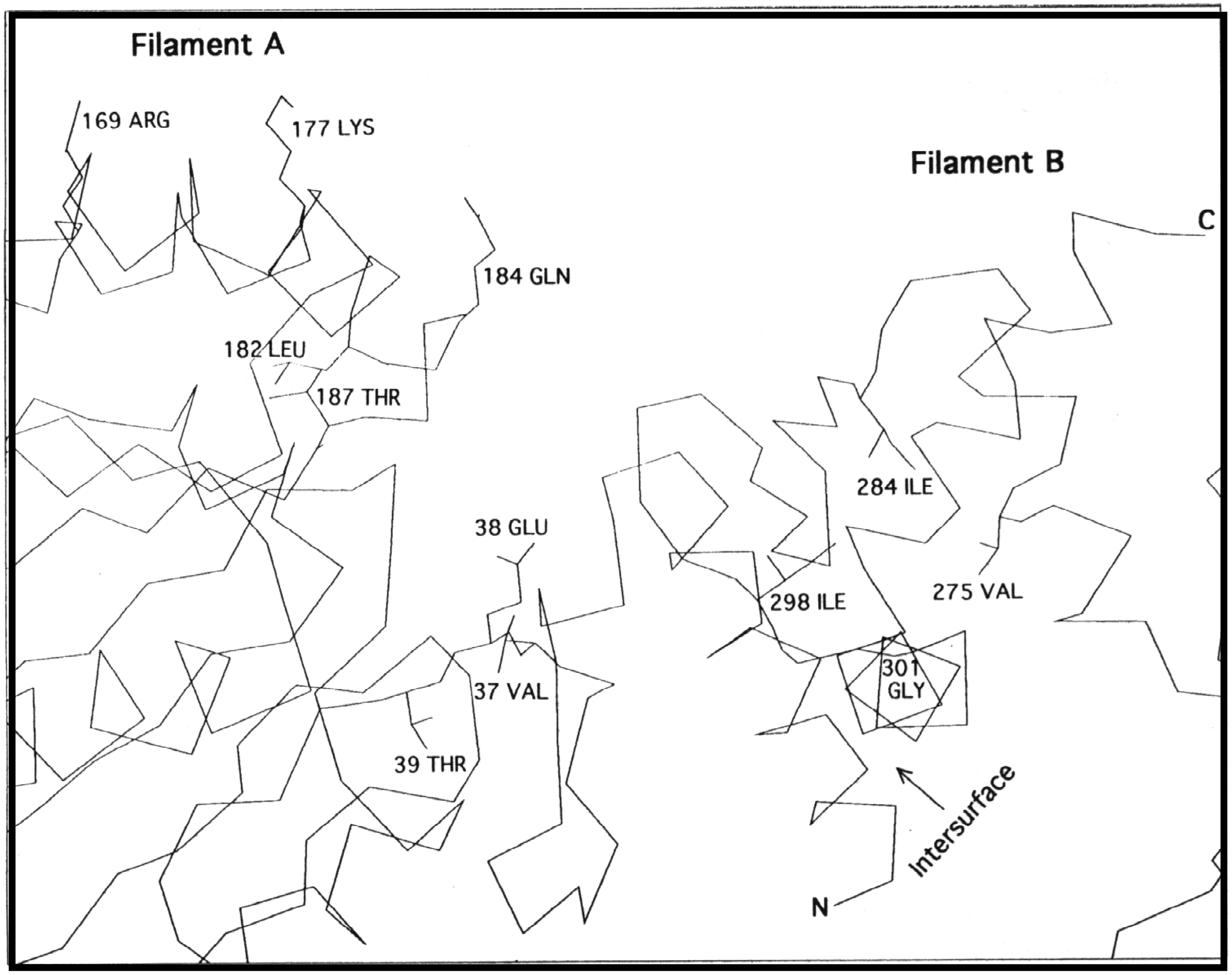

Monomer A
238 VAL
231 VAL
229 GLY
243 ARG
244 VAL
Monomer B
238
VAL
231
VAL

Figure 3. Locations of mutations at the repressor-binding sites of RecA.

The repressor-binding sites of RecA consist of regions of two adjacent monomers in a RecA filament (46), indicated as Monomer A and Monomer B. The mutation sites are at three residues: 231 (*recA1628*), 238 (*recA1642*), and 244 (*recA1627*). Two additional mutation sites, residue 229 (*recA91*)(34) and residue 243 (*recA1734*) (9), are also indicated; these mutations differentially affect cleavage of repressors, and the sites are close to those of the three suppressor mutations notes above. The side chains shown belong to the wild-type forms of the mutated amino acids.



Filament A
169 ARG
177 LYS
184 GLN
182 LEU
187 THR
38 GLU
37 VAL
39 THR
Filament B
C
284 ILE
298 ILE
275 VAL
301 GLY
Intersurface
N

Figure 4. RecA structural models to explain how single *recA* mutations can give rise to the Prt^{c} Rec^{+}, Prt^{+} Rec^{-}, or Prt^{c} Rec^{-} phenotype.

Mutations that affect step A, B or D can shift the equilibrium toward filament formation and result in Prt^{c} coprotease activity (46), mutations at step C, E, or G can shift the equilibrium toward monomer or filament formation and lead to a Rec^{-} phenotype, a mutation that acts at both steps A and G can cause a Prt^{c} Rec^{-} phenotype, and a mutation can act at step F or H to cause more efficient formation of the bundle-like structure and give rise to a mutant protein with enhanced recombinase activity. The DNA shown can be ssDNA or dsDNA.

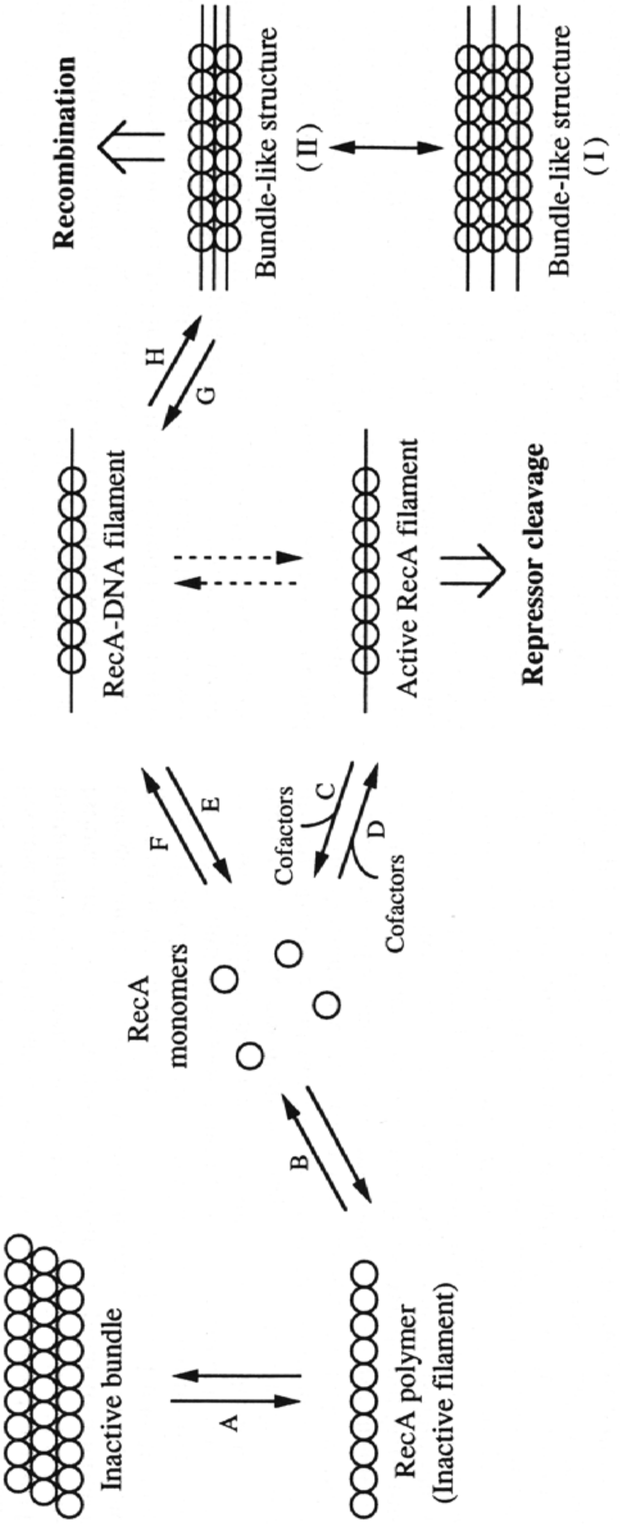

Inactive bundle
A
RecA polymer
(Inactive filament)
B
RecA
monomers
F
E
Cofactors
C
D
Cofactors
RecA-DNA filament
Active RecA filament
Repressor cleavage
H
G
Recombination
Bundle-like structure
(II)
Bundle-like structure
(I)

Figure 5. Outline of experimental methodology.

Strain Infomration

EST2411 has the chromosomal *recA* gene deleted and a *lacZ* gene under the regulation of and SOS promoter. Without functional RecA protein, the SOS-*lacZ* fusion is always repressed. IT3111 is EST2411 lysogenized with λ*recA1202* (a λ phage carrying the *recA1202* allele). This strain shows constitutive SOS induction (dark blue color on X-gal even without induction).

Spontaneous Proximal Mutagenesis

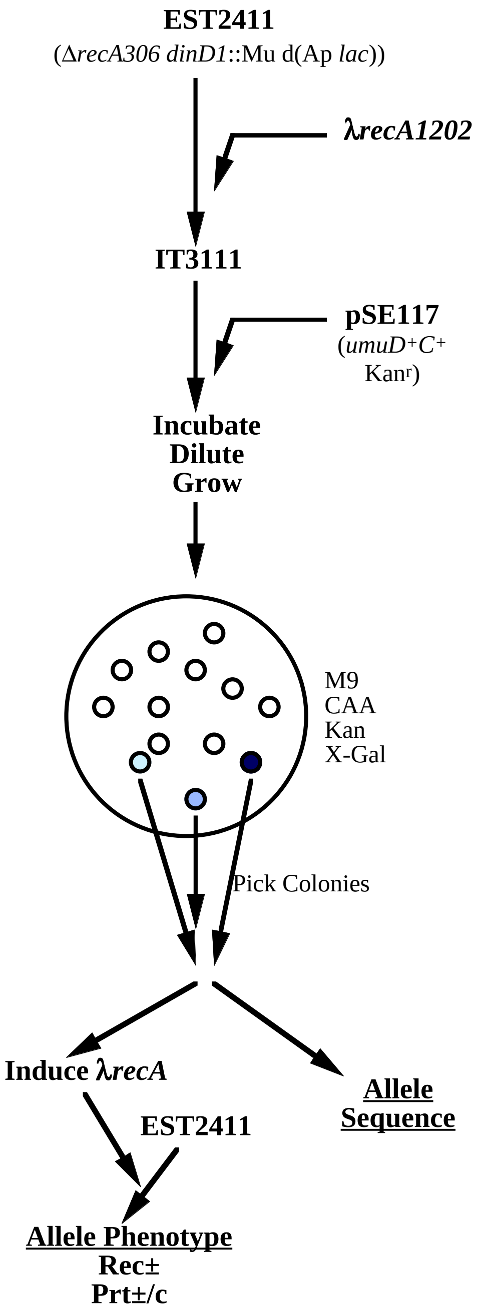
The *recA1202* allele was targeted for mutation by proximal mutagenesis (Liu and Tessman, 1992). In this method the spontaneous mutation rate in an around *recA* (Prt^{c}) alleles is increased relative to the rest of the genome. This process is self-limiting because mutations eventually decrease the Prt^{c} activity and thus decrease the mutation rate.

Mutant Screen

Mutants with decreased protease activity were selected by a decrease in the expression of the SOS-*lacZ* fusion. LacZ expression was quantified by color on M9 X-Gal plates. 28 mutants, chosen to include a wide range of expression, were used for further analysis.

Mutant Characterization

Mutant alleles were sequenced to determine the nucleotide and/or amino acid changes relative to the *recA1202* allele. In addition, alleles were characterized for their effects on protease and recombination activities in reference to *recA*+ (Prt^{+} Rec^{+}), *recA*- (Prt^{-} Rec^{-}) and *recA1202* (Prt^{c} Rec^{+}). First, the alleles were transferred to EST2411, in order to remove possible background effects. Protease activity was determined by color on X-Gal, Weigle reactivation, and spontaneous mutation frequency. Recombination activity was determined by UV and MitC sensitivity as well as ability to recombine with an Hfr donor strain.



CHAPTER 5

Using Evolutionary Analysis to Characterize DNA Repair Processes II:

Evolution of Multigene Families that Include DNA Repair Genes

Evolution of the SNF2 Family of Proteins: Subfamilies with Distinct Sequences and Functions[7]

[7] Previously published as *Jonathan A. Eisen; Kevin S. Sweder; Philip C. Hanawalt. 1995. Nucleic Acids Research 23(14): 2715-2723.* Reprinted with permission from Oxford University Press.

ABSTRACT

The SNF2 family of proteins includes representatives from a variety of species with roles in cellular processes such as transcriptional regulation (e.g., MOT1, SNF2, BRM), maintenance of chromosome stability during mitosis (e.g., lodestar), and various aspects of processing of DNA damage including nucleotide excision repair (e.g., RAD16, ERCC6), recombinational pathways (e.g., RAD54), and post-replication daughter strand gap repair (e.g., RAD5). This family also includes many proteins with no known function. To better characterize this family of proteins we have used molecular phylogenetic techniques to infer evolutionary relationships among the family members. We have divided the SNF2 family into multiple subfamilies, each of which represents what we propose to be a functionally and evolutionarily distinct group. We have then used the subfamily structure to predict the functions of some of the uncharacterized proteins in the SNF2 family. We discuss possible implications of this evolutionary analysis on the general properties and evolution of the SNF2 family.

INTRODUCTION

Proteins with extensive amino acid sequence similarity to the yeast transcriptional activator protein SNF2 have been grouped into a protein family. This family includes proteins from a variety of species with roles in cellular processes such as transcriptional regulation, recombination, and various types of DNA repair (see Table 1; for reviews see (1, 2)). In addition to the sequence similarity with other family members, all proteins in the SNF2 family contain sequence motifs similar to those found in many DNA and RNA helicase protein families (1). Proteins with these "helicase" motifs have been divided into multiple superfamilies based upon amino acid sequence patterns found within the motifs (3). By this method, the SNF2 family has been assigned to helicase superfamily 2 that also includes the ERCC3, RAD3, PRIA, EIF4A, and PRP16 protein families (3).

The number of proteins assigned to the SNF2 family has increased rapidly over the last few years and continues to expand. Many new family members have been cloned by methods that do not provide any information about their function such as in genome sequencing projects or by homology-based cloning. Considering the number of proteins in the family, the diversity of their genetic roles, and the large number of proteins with unknown function, we thought some insights could be provided by deducing the evolutionary relationships among the family members. Our phylogenetic analysis leads us to propose that the SNF2 family is composed of evolutionarily distinct subfamilies of proteins. We suggest that these subfamilies represent groups of homologous proteins with similar functions and activities and that the functions of some of the uncharacterized members of the SNF2 family can be predicted by their assignment to particular subfamilies. The evolutionary relationships determined here provide insight into the diversity of genetic functions within the family as well as the likely common biochemical activities of all family members. Finally, we discuss the implications of this analysis on studies of the function of RAD26 and ERCC6 and their role in transcription-coupled repair (TCR) in eukaryotes.

METHODS

Sequence alignment

Sequences used in this paper were downloaded from the National Center for Biotechnology Information databases using an electronic mail server (retrieve@ncbi.nlm.nih.gov). Accession numbers are given in Table 1. Similarity searches were conducted using the blastp and tblastn (4), MPsrch (5), and fasta (6, 7) computer programs via electronic mail servers (8). Motif searches were conducted using the blocks electronic mail server (9). Alignment of protein sequences was conducted using the clustalv (10) and clustalw (11) multiple sequence alignment programs. The computer generated alignments were optimized by some minor manual adjustment.

Phylogenetic trees

Phylogenetic trees were generated from the sequence alignments using programs available in the PHYLIP (12), PAUP (13), and GDE (14) computer software packages. Parsimony analysis was conducted using the protpars program in PHYLIP and the heuristic search algorithm of PAUP. Multiple runs searching for the shortest tree were conducted using a variety of starting parameters and branch swapping options. For the distance based methods, we first generated matrices representing the estimated evolutionary distances between all pairs of sequences using the protdist program of PHYLIP, with default settings. Phylogenetic trees were then generated from these matrices using the least-squares method of De Soete (15) as implemented in GDE and the Fitch-Margoliash (16) and neighbor-joining methods (17) as implemented in PHYLIP. Since in both parsimony and distance methods each alignment position (the column containing one amino acid from each species) is assumed to include residues that share a common ancestry among species, regions of ambiguous alignment were excluded from the phylogenetic analysis. For similar reasons, regions in which some sequences had alignment gaps were also excluded. Bootstrap resampling was conducted by the method of Felsenstein (18) as implemented in PHYLIP. In bootstrapping, new data sets are made by resampling the alignment positions used in the original data set by random removal and replacement. The result of a single bootstrap is a data set with the same total number of alignment positions as in the original but in which some original alignment positions may not be represented while others may be represented multiple times. Phylogenetic trees are generated based upon each of these modified data sets. Comparison of the trees generated with multiple bootstraps can thus give a measure of the consistency of the original tree. We conducted 100 bootstrap replicates for the protpars, neighbor-joining and Fitch-Margoliash methods.

Computer programs

GDE, PHYLIP, clustalv and clustalw were obtained by anonymous FTP from the archive of the Biology Department at the University of Indiana (ftp.bio.indiana.edu).

RESULTS

Alignment of protein sequences

The presence of a highly conserved domain averaging approximately 400 amino acids in length has been used to define the SNF2 family (1). We will refer to this conserved region as the SNF2 domain. We first aligned the amino acid sequences of all previously characterized members of the SNF2 family. We then used the SNF2 domains from each of these proteins as query sequences in searches of sequence databases to identify potential additional members of the SNF2 family. A list of all the sequences containing a complete SNF2 domain and some relevant information about these sequences is given in Table 1. In addition to these sequences, we have detected some incompletely sequenced open reading frames that encode peptides that are highly similar to portions of the SNF2 domain. These include a partial open reading frame from chicken (19), two from *Mycoplasma genitalium* (U01723 and U02179 in (20)), and many expressed sequence tags from *Caenorhabditis elegans*. The high similarity of the proteins encoded by these sequences to segments of the SNF2 domain suggests that these are also members of the SNF2 family. A new alignment was generated to include all likely members of the SNF2 family. We used this alignment as a block and aligned this block to other proteins with the helicase motifs using the profile alignment method of the clustalv program. A schematic diagram of the alignment of the sequences containing the entire SNF2 domain is shown in Figure 1. A peptide encoded by an incompletely sequenced open reading frame from *Bacillus cereus* is shown in the alignment because it has been previously grouped into the SNF2 family (21). The labeling of particular helicase domains is based on the relative alignment to the suggested helicase domains of these other proteins, as well as previously published assignment of helicase domains to the proteins in the SNF2 family.

The SNF2 domain and the position of the helicase motifs in our final alignment are essentially identical to that presented by others (e.g., (1, 22, 23)). The degree of amino acid conservation varies greatly within the SNF2 domain. We define conserved regions as those regions in which the alignment is unambiguous, the number of amino acids is the same among all the proteins, and the percentage of amino acid similarity between proteins is high. Alignments were considered ambiguous if slight alterations in the alignment parameters, such as changing the scoring matrix used by the clustalv and clustalw programs, greatly altered the relative position of amino acids from the different sequences. Using these definitions, we find that the SNF2 domain is composed of many small conserved regions separated by less-conserved spacers that vary in length among the family members (see Fig. 1). The only notable difference between our alignment and other published alignments of the proteins in the SNF2 family is the relative position of part of the *Escherichia coli* HepA protein. We could not obtain an unambiguous alignment for the region of HepA between helicase domains III and V. There is also no consensus among other researchers in the alignment of these regions of HepA (e.g., see (1, 21)). One possible explanation for the difficulty in aligning this region of HepA to the other members of the SNF2 family is that the amino acid sequence of this region of HepA is somewhat ambiguous. It is necessary to postulate a frameshift in translation or a sequencing error in this region to align the downstream portion of the protein (1) and the exact position of the postulated change may not be correct. Alternatively this region may be poorly conserved between bacteria and eukaryotes which would also make unambiguous alignment difficult. The alignment is available on request.

Phylogenetic trees of SNF2 domain

We generated phylogenetic trees of the proteins in the SNF2 family using multiple distance and parsimony based methods. These trees were generated by comparisons of the regions conserved among all family members (i.e., the conserved regions within the SNF2 domain). Less conserved regions (such as the regions flanking the SNF2 domain and the variable spacer regions) were not used because of problems in obtaining unambiguous alignments in these regions (see methods) and because there is no established method of scoring alignment gaps in phylogenetic reconstruction. Since the phylogenetic methods are more accurate with more alignment positions, we excluded those proteins, like the *B. cereus* partial sequence, that do not have an entire SNF2 domain. The trees generated using the different distance based methods were identical in topology. Similarly, the most parsimonious trees found by the two parsimony methods were identical. In Figure 2 we present a comparison of the trees generated by the parsimony versus distance methods. As can be seen, there are only slight differences between the parsimony and distance based trees. Bootstrap values for each node are shown on the trees. The root of each tree was determined using proteins that contain the helicase motifs but are not members of the SNF2 family as outgroups. In particular we used the vaccinia virus cI proteins since these proteins are considered to be the closest relatives of proteins in the SNF2 family (24). In trees generated by all the methods using these proteins as outgroups, HepA was determined to be the deepest branching member of the SNF2 family. Thus the trees are shown with HepA as an outgroup.

Sequence motifs and similarities in less conserved regions

We were also interested in sequence patterns and relationships in the regions of each of the proteins that were not conserved among all family members (i.e., in the variable spacers within the SNF2 domain and in the regions on the C- and N-terminal sides of the SNF2 domain). We conducted two types of analysis on these less conserved regions: motif searches and sequence similarity searches. Some interesting amino acid motifs have previously been identified in these less conserved regions of some of the members of the SNF2 family. For example, SNF2, STH1, BRM, hBRM, mBRG1 and hBRG1 proteins have all been shown to contain a bromodomain motif on the C-terminal side of the SNF2 domain (e.g., (25)). We did not find bromodomain-like motifs in any of the remaining members of the SNF2 family. RAD5, RAD16, and spRAD8 have all been shown to contain a RING-finger like motif between helicase motifs III and IV (26-30). We find a similar motif in HIP116A (aa 766-836), also between helicase motifs III and IV. Finally, CHD1 has been shown to have a chromodomain motif on the N-terminal side of the SNF2 domain (31). We have found a similar motif in the same relative positions in the yeast sequence SYGP4 (aa 203-235). No other significant matches to any motif profiles in the blocks database were found. The motifs described above are highlighted in Figure 1.

We used a variety of sequence comparison programs (see Methods) to search sequence databases for proteins or possible open reading frames with similarity to the less conserved regions of each the SNF2 family members. We defined significant similarities as those with p values less than $1x10^{-4}$ for at least one of the search methods.

Other than in the regions of the motifs described above, the only significant sequence similarities in the less-conserved regions of any of the proteins were with other SNF2 family members. In all cases, the significant similarities detected were between proteins that branch close to each other in the phylogenetic trees. All similarities detected between two proteins were in comparable regions of the proteins. For example, the regions on the C-terminal sides of the SNF2 domain only showed similarity to other C-terminal regions. Overall, the similarities we found allowed us to divide the SNF2 family into six distinct groups of proteins. All proteins within a group have significant similarity outside the SNF2 domain to all other members of the same group but not to any other proteins in the SNF2 family. These groups are 1: (SNF2L, ISWI, F37A4, YB95); 2: (CHD1, SYGP4); 3: (ERCC6, RAD26); 4: (hNUCP, mNUCP); 5: (RAD54, DNRPPX); 6: (SNF2, STH1, BRM, hBRM, mBRG1, hBRG1) and 7: (RAD16, HIP116A, RAD5, spRAD8). MOT1, ETL-1, FUN30, YB53, lodestar, HepA, and NPH42 showed no significant similarity outside the SNF2 domain to any other SNF2 family members. A few of the proteins not included in the groups do show small regions of less-significant similarity to some other members of the SNF2 family. For example, YB53 has a small region of similarity to RAD54 and DNRPPX.

The regions of significant sequence similarity between group members vary within and among the groups. For example, mBRG1 and hBRG1 are significantly similar throughout their entire lengths, including the regions on the C- and N-terminal side of the SNF2 domain as well as the variable spacers. In contrast, mBRG1 and SNF2 show little similarity in the variable spacers, some similarity in the regions on the N-terminal side of the SNF2 domain, and extensive similarity in the region on the C-terminal side. To summarize we have characterized the groups by the regions that are significantly similar among all group members: groups 1, 2, 3, and 4 (both the C- and N-terminal sides of the SNF2 domain); group 5 (N-terminal side); group 6 (C-terminal side with a small region on the N-terminal side); and group 7 (the spacer between helicase domains III and IV -- which is the location of the RING finger motif in all of these sequences). A summary of the regions of significant sequence similarity is given in Table 2. The regions of similarity among all group members are highlighted in Figure 1.

DISCUSSION

Molecular phylogenetic analysis can be used to infer the evolutionary history of genes. Such phylogenetic information can provide insight into the function of particular sequences as well as into the forces that have affected their evolution. We have applied molecular phylogenetic techniques to infer the evolutionary history of the SNF2 family of proteins. Based upon this analysis we propose that the SNF2 family is composed of evolutionarily distinct subfamilies. The subfamilies we propose are listed in Table 1 and outlined in Fig. 1 and Fig. 2. We have named each subfamily after one of the proteins in that subfamily. To avoid confusion, we use *ITALICS* when referring to the subfamily.

We based our selection of subfamilies upon the following criteria. First, each subfamily had to be monophyletic. Monophyly for a group occurs when all the members of the group share a common ancestor that no other sequences share. Thus one subfamily could not have evolved from within another subfamily. Second, each subfamily had to be inferred by each of the phylogenetic reconstruction methods used. All phylogenetic reconstruction methods rely on assumptions about the evolutionary process. Each class of methods has a distinct range of evolutionary scenarios over which it reliably reconstructs true evolutionary relationships (18). Congruence among trees inferred by different methods therefore indicates robustness of the phylogenetic conclusions. We used two different classes of methods (parsimony-based and distance-based) and multiple types of each method. All proposed subfamilies were found by all methods. Third, the node defining each subfamily had to have high bootstrap values. Bootstrap values for the node defining a subfamily indicate the percentage of times that the sequences in the subfamily grouped together to the exclusion of other sequences in trees generated using different subsamples of a particular alignment. Bootstrapping is thus a method for assessing whether a particular branching pattern has been biased by the sampling of alignment positions. The bootstrap values were very high (between 90-100%) for the nodes that define most of the subfamilies (see Table 2). The only proposed subfamily with consistently moderate to low bootstrap values is the *RAD16* subfamily. It is possible that this subfamily would be divided into multiple subfamilies with the availability of sequences from more species.

Our phylogenetic analysis shows that the sequences within the proposed subfamilies are historically more related to each other than to any other characterized proteins, including other members of the SNF2 family. We propose that these evolutionary subdivisions are paralleled by functional subdivisions, and therefore that function is conserved within but not between subfamilies. In the cases for which the information is available, protein function does appear to be conserved within subfamilies (see Table 1). For example, both members of the *ERCC6* subfamily, RAD26 and ERCC6 are involved in the process of transcription coupled DNA repair (32, 33). In addition, all the proteins in the *SNF2* subfamily for which functional information is available are known to function in transcriptional activation (see Table 1). The *RAD16* subfamily is the only subfamily that includes proteins with known dissimilar genetic functions. This subfamily includes RAD16 which is involved in nucleotide excision repair of nontranscribed regions of the genome and RAD5 which is involved in post-replication repair and mutagenesis. As discussed above we believe it is possible that the proposed *RAD16* subfamily may include sequences from multiple subfamilies. However, we note that recent experiments suggest that RAD5 and RAD16 may functionally interact (34).

Other genetic evidence supports our proposal that function is conserved within but not between the proposed subfamilies. For example, expression of BRG1 (35) and BRM genes can restore growth and transcription activity to yeast SNF2 mutants but expression of the hSNF2L gene (which is from another subfamily) cannot (22). In addition, expression of genetic chimeras in which the sequence coding for the SNF2 domain of the SNF2 protein is replaced with the corresponding region of BRG1 (35), BRM (23) or STH1 (36) can restore growth and transcription to SNF2 mutants. However if the SNF2 domain of ISWI (a member of a different subfamily) is used as a donor, function is not restored (23).

We believe that our sequence comparisons of the regions outside the SNF2 domain also support our proposal of functional distinctness of the subfamilies. By definition the less conserved regions were not found in all the proteins in the SNF2 family and were not used in the phylogenetic analysis. Because of the possibility of processes such as domain swapping, exon shuffling, and recombination, it is theoretically possible that the phylogenetic relationships of the SNF2 domain would not correspond to the relationships of the less conserved regions. We therefore examined patterns of sequence similarity outside the conserved regions of the SNF2 domain (see Results, Table 2). Of the similarities we detected, some have been noted previously (e.g., 32, 35). Most relevant to this study, among SNF2 family members, the only significant sequence similarity outside the SNF2 domain is within our proposed subfamilies. In most cases, significant similarity outside the SNF2 domain was detected among all members of our proposed subfamilies. This is true for the *SNF2*, *SNF2L*, *ERCC6*, *CHD1* and *RAD16* subfamilies (see Table 2). Thus these regions are conserved within but not between subfamilies.

We believe that the sequence conservation within but not between subfamilies is due to conservation of function within the subfamilies. The regions conserved within subfamilies may be important in providing specific functions to each of the subfamilies. We believe that analysis of these regions will help identify the function conserved within each subfamily. Some of the proteins in the SNF2 family contain sequence motifs also found in proteins outside the SNF2 family. Other researchers have used the nature of these motifs to help predict the functions of the proteins that have the motifs. We have found that these motifs are conserved within subfamilies and propose that the nature of these motifs may help identify the function conserved within the subfamily. For example, all members of the *SNF2* subfamily contain a bromodomain motif (see Results, Fig 1). This motif is found in a variety of proteins involved in transcription regulation (25) and it has been suggested that it may be involved in protein-protein interactions (37). It is not known what function the bromodomain provides to the members of the *SNF2* subfamily -- it can be deleted from SNF2 (38) and hBRM (39) with no discernible phenotypic effect. Recent studies of BRG1 suggest that the region containing the bromodomain may be involved in binding the retinoblastoma protein (40). Both proteins in the *CHD1* subfamily contain a chromodomain motif. This motif is found in a few other proteins and is proposed to play a role in chromatin compaction (41), but it is not known what role it plays in the function CHD1 or SYGP4 (31). Finally, a RING finger motif is found in all the proteins in the *RAD16* subfamily. This motif is related at the sequence and structural levels to the zinc finger motif (42, 43). It is found in many proteins that interact with DNA (including the DNA repair protein RAD18, the p53 associated protein MDM2, and the protooncogene mel-18) and it is thought that it is involved in DNA binding (42). We believe that the presence of this motif in all the members of the proposed *RAD16* subfamily, but not in any other proteins in the SNF2 family, lends support to the idea that these sequences form a distinct group.

If, as we suggest above, function is conserved within subfamilies, then the functions of some of the uncharacterized proteins in the SNF2 family can be predicted by comparison to other members of the same subfamily. For example, we predict that STH1, the only member of the *SNF2* subfamily for which a genetic role is unknown, is involved in transcription activation as are all the other members of this subfamily. STH1 is in a monophyletic evolutionary group with the other proteins in the *SNF2* subfamily in every phylogenetic method. In addition, it contains the same sequence motifs, including the bromodomain, found in all the other members of the *SNF2* subfamily. Since STH1 mutants do not have the same phenotype as SNF2 mutants (36), STH1 may have a slightly different function from SNF2. For example, STH1 may be involved in transcription activation only in certain environmental conditions or in certain stages of the cell cycle. We also predict that HIP116A may have some function in DNA repair. HIP116A branches consistently within the *RAD16* subfamily and contains a sequence motif (the RING finger) found in all members of this subfamily but not in any other members of the SNF2 family. Two of the other members of the *RAD16* subfamily are involved in DNA repair (RAD16 and RAD5) and the third is likely involved in repair (spRAD8) (26). The subfamily structure also allows us to identify likely homologs of uncharacterized mammalian proteins in species in which function may be easier to ascertain. For example, the human SNF2L has no known function (22). We suggest that it will be informative to study likely SNF2L homologs ISWI, YB95, or F37A4 in the more tractable systems of *Drosophila melanogaster*, *S. cerevisiae*, and *C. elegans*, respectively. Similarly, we believe the elucidation of the function of CHD1 and ETL1 may be facilitated by studying their likely homologs in *S. cerevisiae*, SYGP4 and FUN30 respectively.

The evolutionary relationships among subfamilies are less strongly resolved than those that define the subfamilies. For example the evolutionary position of some of the subfamilies is different in the parsimony versus distance based trees (see Fig. 1). In addition, bootstrap values for the nodes that define the branching patterns between subfamilies are low, indicating that changes in the choice of alignment positions used to generate the trees affect the inferred relationships among subfamilies. More accurate determination of the evolutionary relationships among subfamilies should be possible once more sequences are available in each subfamily. However, we believe that most of the overall topology of the relationships among subfamilies will not change significantly from that presented here. For example, the *SNF2*, *CHD1*, and *SNF2L* subfamilies form a coherent supergroup -- the bootstrap values for this supergroup are 100 in all trees and the estimated distances (branch lengths) between these subfamilies are low. In addition, we find it intriguing that the proteins known to be involved in DNA repair have deeper branches than those known to be involved in transcription. It is possible that the transcription functions evolved later in the history of this family. However, until more is known about the genetic and biochemical activities of many of the proteins in the SNF2 family, the implications of the inter-subfamily relationships are unclear.

Regardless of the specific phylogenetic relationships among the subfamilies, it is apparent from the number of proteins in the SNF2 family from single species that there have been many duplications in the history of the SNF2 family (see Table 1). We believe the phylogenetic analysis reveals a great deal about the timing of these duplications. Since *S. cerevisiae* has a representative in each subfamily and mammals have a representative in all but the *MOT1* subfamily, we believe that many of the duplications occurred before the separation of fungal and animal ancestors. The rooting of the tree with HepA and the absence of bacterial representatives from the rest of the tree suggests that the majority of the duplications occurred after the separation of bacterial and eukaryotic ancestors. Until complete bacterial genomes are available it is impossible to know for certain if any bacterial species encodes multiple members of the family. Unfortunately the only likely members of this family from bacterial species other than *E. coli* have not been sequenced completely and are currently too short to use reliably in phylogenetic methods. Complete sequences of these will help better determine the history of these proteins in bacteria. Since in most cases, all the proteins within a subfamily contain sequence motifs that are not found in any other members of the SNF2 family we propose that many of the duplications of the SNF2 domain were accompanied by the addition of these subfamily specific motifs.

The high conservation of amino acid sequence in the SNF2 domain has led to much speculation about whether any particular biochemical activity is shared by all members of the SNF2 family. The presence of the helicase motifs in the SNF2 domain has been used to suggest that the conserved activity is helicase activity. While helicase activity is needed for the processes (i.e., transcription, recombination and DNA repair) in which these proteins are known to be involved, helicase activity has never been detected in any protein in the SNF2 family. This is despite extensive efforts to detect such activity, especially for SNF2 (44) and MOT1 (45). Despite the presence of the motifs, Henikoff proposed that the SNF2 proteins are not helicases (24) and that the "helicase" motifs are indicative of a broader DNA-dependent ATPase activity of which helicase activity is a subset. Consistent with this proposal, SNF2, MOT1, and HIP116A have all been shown to be DNA-dependent ATPases. Thus, the SNF2 family members may share another activity that requires a DNA-dependent ATPase function.

We believe the phylogenetic analysis presented here may help understand the common function of the proteins in the SNF2 family. For example, the apparent massive duplication in eukaryotes suggests that either there is something specific about eukaryotes that required or allowed for the diversification of this protein family or there is something in bacteria that prevented the diversification. Understanding what influenced this diversification in eukaryotes might provide a clue about the common function of these proteins. We believe that recent work on MOT1 helps identify what that eukaryotic specific factor is. Auble et al. have shown that MOT1 functions to remove TATA-binding protein (TBP) from DNA. They suggest that the common function of the SNF2 family members is the ability to remove proteins from DNA utilizing the energy of ATP hydrolysis (45). We believe that this activity may have been particularly important during the early evolution of eukaryotes because of the higher complexity of DNA packaging with proteins and other protein-DNA interactions in eukaryotes versus bacteria. Auble et al. suggest that the particular details of protein removal from DNA varies among SNF2 family members. We suggest that these specific details will be conserved within our proposed subfamilies. For example, if the suggestion that SNF2 functions to remove histones from DNA (e.g., (46)) is confirmed, we would suggest that hBRM, BRM, BRG1, and STH1 will have similar activities.

Of the proteins in the SNF2 family, we are particularly interested in the human ERCC6 protein. ERCC6 protein is defective in individuals with Cockayne's syndrome-complementation group B (CS-B) (33). Cockayne's syndrome is an autosomal recessive disorder characterized by growth retardation, severe photosensitivity, developmental abnormalities, and neural degeneration. Cells from patients with CS-B lack transcription-coupled repair (TCR), the preferential repair of DNA damage on the transcribed strand of an actively transcribing gene relative to the non-transcribed strand of the same gene (47, 48). It is not known whether the symptoms associated with CS-B are due to their lack of TCR or to another activity of ERCC6 in transcriptional regulation, as has been suggested (49).

Since its discovery in the DHFR gene in hamster cells (50), TCR has been shown to be widespread (48). Mellon and Hanawalt suggested that the mechanism of TCR might involve the blockage of transcription by DNA damage and that the recognition of this blockage serves as a signal to the nucleotide excision repair proteins (51). Selby and Sancar subsequently showed that, in an *in-vitro E. coli* system, TCR is an active process requiring a transcription-repair coupling factor (TRCF) and that this TRCF is encoded by the *mfd* gene. They have also shown that the product of the *mfd* gene can remove an *E. coli* RNA polymerase stalled at a DNA lesion (52-54). Selby and Sancar propose that the Mfd protein also serves to recruit the nucleotide excision repair system to that lesion. The Mfd protein, like ERCC6 and RAD26, contains motifs like those found in helicases. As with the proteins in the SNF2 family, despite the presence of the helicase motifs, helicase activity has not been detected in Mfd (55). Although Mfd and ERCC6 both contain the helicase motifs, they are not true homologs . Each is more similar to many other proteins than to the other (for example, ERCC6 and RAD26 are more closely related to all the other members of the SNF2 family than they are to Mfd) (55). This suggests that perhaps ERCC6/RAD26 and Mfd do not function in a similar way. Despite this complication, there are many similarities between the eukaryotic and prokaryotic processes of TCR. In an *in-vitro* eukaryotic system, DNA damage in the transcribed strand of an expressed gene is an absolute block to transcription elongation (56). Like in *E. coli*, this RNA polymerase complex stalled at the site of DNA damage must then be moved to allow access to repair proteins (56). The moving of a stalled RNA polymerase is similar to the predicted general function of the SNF2 family of proteins -- removing proteins from DNA. Thus, we predict that ERCC6 and RAD26 function in the moving of stalled RNA polymerase away from the site of DNA damage. If this is true, the lack of homology of Mfd and ERCC6 suggests that eukaryotes and prokaryotes have separately evolved the ability to move a stalled RNA polymerase. It has been suggested that it would be beneficial to eukaryotes for TCR to allow for continued RNA synthesis after DNA repair (because of the amount of energy invested in synthesizing some large RNAs (57)). Thus, unlike in bacteria, eukaryotes may somehow translocate the RNA polymerase but not remove it.

In conclusion, we believe that molecular phylogenetics is a useful tool in studies of protein families. In the present case we believe molecular phylogenetics has helped to 1) understand the common properties of the SNF2 family members; 2) make reasonable predictions of the functions of uncharacterized members of the family; 3) divide the family into functionally distinct subfamilies; and 4) identify amino acid sequences conserved within but not between subfamilies. These regions conserved within subfamilies are likely important in providing specific functions to the proteins; therefore the characteristics of these regions (e.g., charge, presence of known motifs) may help identify the activity(s) conserved within the subfamilies. The subfamily specific activities are also determined in part by the characteristics of the highly conserved SNF2 domain -- swapping the SNF2 domain leads to functional proteins only when the donor and recipient are from the same subfamily (see above). Related to this, we have identified proteins that do not share any particular motifs outside the SNF2 domain but which consistently group together in the phylogenetic analysis. Examples of this include the *ETL1* subfamily in which FUN30 and ETL1 branch together in every analysis but have no significant sequence similarity outside the SNF2 domain and the *RAD54* subfamily which includes two sub-groups which show no similarity between the groups. The phylogenetic analysis is particularly helpful is these cases.

ACKNOWLEDGMENTS

We would like to thank J.H.J. Hoeijmakers for helpful comments on earlier versions of this work and Marc Feldman for the use of computer equipment. KSS was supported by a Postdoctoral Training Grant (T32AR07422) from the National Institute of Arthritis and Musculoskeletetal and Skin Diseases to the Department of Dermatology in the Stanford University School of Medicine. JAE was supported by a National Science Foundation Pre-Doctoral Fellowship and a tuition grant from the Marine Biological Laboratory to attend the 1994 Workshop on Molecular Evolution at which much of this research was initiated. This work was also supported by an Outstanding Investigator Grant (CA44349) from the National Cancer Institute to PCH.

REFERENCES

1. Bork,P. and Koonin,E.V. (1993), *Nucleic Acids Res.*, **21**, 751-752.
2. Carlson,M. and Laurent,B.C. (1994), *Curr. Opin. Cell. Biol.*, **6**, 396-402.
3. Gorbalenya,A.E. and Koonin,E.V. (1993), *Curr. Opin. Struct. Biol.*, **3**, 419-429.
4. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990), *J. Mol. Biol.*, **215**, 403-410.
5. Sturrock,S.S. and Collins,J.F. (1993) MPsrch, Version 1.3. Biocomputing Research Unit, Edinburgh, UK.
6. Pearson,W.R. and Lipman,D.J. (1988), *Proc. Natl. Acad. Sci.*, **85**, 2444-2448.
7. Pearson,W.R. (1990), *Meth. Enzymol.*, **183**, 63-98.
8. Henikoff,S. (1993), *Trends Biochem. Sci.*, **18**, 267-268.
9. Henikoff,S. and Henikoff,J. (1994), *Genomics*, **19**, 97-107.
10. Higgins,D., Bleasby,A. and Fuchs,R. (1992), *Comput. Appl. Biosci.*, **8**, 189-191.
11. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994), *Nucleic Acids Res.*, **22**, 4673-4680.
12. Felsenstein,J. (1989), *Cladistics*, **5**, 164-166.
13. Swafford,D. (1991) Phylogenetic Analysis Using Parsimony (PAUP), Version 3.0d. Illinois Natural History Survey, Champaign, Ill.
14. Smith,S.W. (1991) Genetic Data Environment., Version 2.2a. Harvard Genome Laboratory, Cambridge, MA.
15. De Soete,G. (1983), *Psychometrika,* **48**, 621-626.

16. Fitch,W.M. and Margoliash, E. (1967), *Science*, **155**, 279-284.
17. Saitou,N. and Nei, M. (1987), *Mol. Biol. Evol.*, **4**, 406-425.
18. Felsenstein,J. (1985), *Evol.*, **39**, 783-791.
19. Funahashi,J., Sekido,R., Murai,K., Kamachi,Y. and Kondoh,H. (1993), *Dvlpmt.*, **119**, 433-446.
20. Peterson,S.N., Hu,P.-C., Bott,K.F. and Hutchison, C.A.I. (1993), *J. Bacteriol.*, **175**, 7918-7930.
21. Kolsto,A.B., Bork,P., Kvaloy,K., Lindback,T., Gronstadt,A., Kristensen,T. and Sander,C. (1993), *J. Mol. Biol.*, **230**, 684-688.
22. Okabe,I., Bailey,L.C., Attree,O., Srinivasan,S., Perkel,J.M., Laurent,B.C., Carlson,M., Nelson,D.L. and Nussbaum,R.L. (1992), *Nucleic Acids Res.*, **20**, 4649-4655.
23. Elfring,L.K., Deuring,R., McCallum,C.M., Peterson,C.L. and Tamkun,J.W. (1994), *Mol. Cell. Biol.*, **14**, 2225-2234.
24. Henikoff,S. (1993), *Trends Biochem. Sci.*, **18**, 291-292.
25. Tamkun,J.W., Deuring,R., Scott,M.P., Kissinger,M., Pattatucci,A.M., Kaufman,T.C. and Kennison,J.A. (1992), *Mol. Cell. Biol.*, **12**, 1893-1902.
26. Doe, C.L., Murray, J.M., Shayeghi, M., Hoskins, M., Lehmann, A.R., Carr, A.M. and Watts, F.Z. (1993), *Nucleic Acids Res.*, **21**, 5964-5971.
27. Ahne,F., Baur,M. and Eckardt-Schupp,F. (1992), *Curr. Genet.*, **22**, 277-282.
28. Bang,D.D., Verhage,R., Goosen,N., Brouwer,J. and van de Putte,P. (1992), *Nucleic Acids Res.*, **20**, 3925-3931.
29. Mannhaupt,G., Stucka,R., Ehnle,S., Vetter,I. and Feldmann,H. (1992), *Yeast*, **8**, 385-395.
30. Johnson,R.E., Henderson,S.T., Petes,T.D., Prakash,S., Bankmann,M. and Prakash,L. (1992), *Mol. Cell. Biol.*, **12**, 3807-3818.
31. Delmas,V., Stokes,D.G. and Perry,R.P. (1993), *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 2414-2418.
32. Van Gool,A.J., Verhage,R., Swagemakers,S.M.A., van de Putte,P., Brouwer,J., Troelstra,C., Bootsma,D. and Hoeijmakers,J.H.J. (1994), *Embo J.*, **13**, 5361-5369.
33. Troelstra, C., van Gool,A., de Wit,.J., Vermeulen,W., Bootsma,D. and Hoeijmakers,J.H.J. (1992), *Cell*, **71**, 939-953.
34. Glassner,B.J. and Mortimer,R.K. (1994), *Radiat. Res.*, **139**, 24-33.
35. Khavari,P.A., Peterson,C.L., Tamkun,J.W., Mendel,D.B. and Crabtree,G.R. (1993), *Nature*, **366**, 170-174.
36. Laurent,B.C., Yang,X. and Carlson,M. (1992), *Mol. Cell. Biol.*, **12**, 1893-1902.
37. Haynes,S.R., Dollard,C., Winston,F., Beck,S., Trowsdale,J. and Dawid,I.B. (1992), *Nucl. Acids Res.*, **20**, 2603.
38. Laurent,B.C., Treich,I. and Carlson,M. (1993), *Genes Dev.*, **7**, 583-591.
39. Muchardt,C. and Yaniv,M. (1993), *Embo J.*, **12**, 4279-4290.
40. Dunaief,J.L., Strober,B.E., Guha,S., Khavari,P.A., Ålin,K., Luban,J., Begemann,M.,

Crabtree,G. and Goff,S.P. (1994), *Cell*, **79**, 119-130.

41. Paro,R. and Hogness,D.S. (1991), *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 263-267.
42. Lovering,R., Hanson,I.M., Borden,K.L.B., Martin,S., O'Reilly,N.J., Evan,G.I., Rahman,D., Pappin,D.J.C., Trowsdale,J. and Freemont,P.S. (1993), *Proc. Natl. Acad. Sci. U. S. A.*, **90**, 2112-2116.
43. Barlow,P.N., Luisi,B., Milner,A., Elliott,M. and Everett,R. (1994), *J. Mol. Biol.*, **237**, 201-211.
44. Cote,J., Quinn,J., Workman,J.L. and Peterson,C.L. (1994), *Science*, **265**, 53-60.
45. Auble,D.T., Hansen,K.E., Mueller,C.G.F., Lane,W.S., Thorner,J. and Hahn,S. (1994), *Genes & Dvlpmt.*, **8**, 1920-1934.
46. Wolffe,A.P. (1994), *Curr. Biol.*, **4**, 525-528.
47. Venema,J., van Hoffen,A., Karcagi,V., Natarajan,A.T., van Zeeland,A.A. and Mullenders,L.H. (1991), *Mutat. Res.*, **255**, 123-141.
48. Hanawalt,P.C. and Mellon,I. (1993), *Curr. Biol.*, **3**, 67-69.
49. Bootsma,D. and Hoeijmakers,J.H.J. (1993), *Nature*, **363**, 114-115.
50. Mellon,I., Spivak,G. and Hanawalt,P.C. (1987), *Cell*, **51**, 241-249.
51. Mellon,I. and Hanawalt,P.C. (1989), *Nature*, **342**, 95-98.
52. Selby,C.P., Witkin,E.M. and Sancar,A. (1991), *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 11574-11578.
53. Selby,C.P. and Sancar,A. (1991), *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 8232-8236.
54. Selby,C.P. and Sancar,A. (1993), *Science*, **260**, 53-58.
55. Selby,C.P. and Sancar,A. (1994), *Microbiol. Rev.*, **58**, 317-329.
56. Donahue,B.A., Yin,S., Taylor,J.-S., Reines,D. and Hanawalt,P.C. (1994), *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 8502-8506.
57. Hanawalt,P.C., Donahue,B.A. and Sweder,K.S. (1994), *Curr. Biol.*, **4**, 518-521.
58. Yoshimoto,H. and Yamashita,I. (1991), *Mol. Gen. Genet.*, **228**, 270-280.
59. Laurent,B.C., Treitel,M.A. and Carlson,M. (1991), *Proc. Natl. Acad. Sci. U. S. A.*, **88**, 2687-2691.
60. Tsuchiya,E., Uno,M., Kiguchi,A., Masuoka,K., Kanemori,Y., Okabe,S. and Miyakawa,T. (1992), *Embo J.*, **11**, 4017-4026.
61. Randazzo,F.M., Khavari,P., Crabtree,G., Tamkun,J. and Rossant,J. (1994), *Dev. Biol.*, **161**, 229-242.
62. Chiba,H., Muramatsu,M., Nomoto,A. and Kato,H. (1994), *Nucleic Acids Res.*, **22**, 1815-1820.
63. Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Becks,M. and Bonfield,J. (unpublished) gi458966.
64. Aljinovic,G., Pohl,F.M. and Pohl,T.M. (unpublished) Z36114.
65. Mulligan,J.T., Dietrich,F.S., Hennessey,K.M., Sehl,P., Komp,C., Wei,Y., Taylor,P., Nakahara,K., Roberts,D. and Davis,R.W. (unpublished) gi172808.
66. Soininen,R., Schoor,M., Henseling,U., Tepe,C., Kisters-Woike,B., Rossant,J. and Gossler, A. (1992), *Mech. Dev.*, **39**, 111-123.

67. Kaback,D.B. and Busey,H. (1992), *Yeast*, **8**, 133-145.

68. Barton,A.B. and Kaback,D.B. (1994), *J Bacteriol.*, **176**, 1872-1880.

69. Davis,J.L., Kunisawa,R. and Thorner,J. (1992), *Yeast*, **8**, 397-408.

70. Huang,M.E., Chuat,J.C. and Galibert,F. (1994), *Biochem. Biophys. Res. Commun.*, **201**, 310-317.

71. Emery,H.S., Schild,D., Kellogg,D.E. and Mortimer,R.K. (1991), *Genes Dev.*, **5**, 1786-1799.

72. Muris,D.F.R., Vreeken,K., Smit,C., Carr,A.M., Broughton,B.C., Lehman,A.R., Lohman,P.H.M. and Pastink,A. (unpublished) Z29640.

73. Steensma,H.Y. and Van der Aart,Q.J.M. (unpublished) Z35942.

74. Van Der Aart,Q.J.M., Barthe,C., Doignon,F., Aigle,M., Crouzet,M. and Steensma,H.Y. (1994), *Yeast*, **10**, 959-964.

75. Gecz,J., Pollard,H., Consalez,G., Villard,L., Stayton,C., Millasseau,P., Khrestchatisky,M. and Fontes,M. (1994), *Hum. Mol. Genet.*, **3**, 39-44.

76. Schild,D., Glassner,B.J., Mortimer,R.K., Carlson,M. and Laurent,B.C. (1992), *Yeast*, **8**, 385-395.

77. Sheridan,P.L., Schorpp,M., Voz,M.L. and Jones,K.A. (unpublished) L34673.

78. Ayers,M., Howard,S., Kuzio,J., Lopez-Ferber,M. and Possee,R. (1994), *Virology*, **202**, 586-605.

79. Girdham,C.H. and Glover,D.M. (1992), *Gene Expr.*, **2**, 81-91.

80. Lewis,L.K., Jenkins,M.E. and Mount,D.W. (1992), *J. Bacteriol.*, **174**, 3377-3385.

81. Iwasaki,H., Ishino,Y., Toh,H., Nakata,A. and Shinagawa,H. (1991), *Mol. Gen. Genet.*, **226**, 24-33.

Table 1. Proteins in the SNF2 Family.

Protein	*#aa*	Species	Sub-family	Function/ Comments	Genbank	Refs.
SNF2[1]	*1703*	*S.cerevisiae*	*SNF2*	Transcription activation. DNA-dependent ATPase. Alters chromatin structure?	M61703	(58, 59)
STH1[2]	*1359*	*S.cerevisiae*	"	Cell cycle control. Required for normal growth.	M83755	(59, 60)
BRM	*1638*	*D.melanogaster*	"	Transcription activation of homeotic genes.	M85049	(25, 40)
BRG1	*1022*	Mouse	"	Binds retinoblastoma protein.	S68108	(61)
BRG1[3]	*1613*	Human	"	Transcription coactivation w/ hormone receptors.	S66910	(35, 62)
hBRM[4]	*1586*	Human	"	Transcription coactivation w/ hormone receptors.	X72889	(39, 62)
SNF2L	*976*	Human	*SNF2L*	?	M89907	(22)
ISWI	*1027*	*D.melanogaster*	"	?	L27127	(23)
F37A4.8	*971*	*C.elegans*	"	?	gi458966	(63)
YB95[5]	*1143*	*S.cerevisiae*	"	?	Z36114	(64)
CHD-1[6]	*940*	Mouse	*CHD1*	Binds DNA.	L10410	(31)
SYGP4	*1468*	*S.cerevisiae*	"	?	gi172808	(65)
ETL-1	*1136*	Mouse	*ETL1*	Expressed very early in development. Concentrated in CNS & epithelium.	X69942	(66)
FUN30[7]	*1131*	*S.cerevisiae*	"	Mutants have increased UV resistance.	gi171856	(67, 68)
MOT1	*1867*	*S.cerevisiae*	*MOT1*	Transcription repression. Removes TBP from DNA. DNA-dependent ATPase.	M83224	(69)
RAD26[8]	*1085*	*S.cerevisiae*	*ERCC6*	Transcription-coupled repair.	X81635	(32, 70)
ERCC6	*1493*	Human	"	Transcription-coupled repair. Defective in Cockayne's syndrome group B.	L04791	(33)
RAD54	*898*	*S.cerevisiae*	*RAD54*	Recombination repair.	M63232	(71)
DNRPPX	*852*	*S.pombe*	"	?	Z29640	(72)
YB53[9]	*958*	*S.cerevisiae*	"	?	Z35942	(73, 74)
NUCPRO	*1298*	Human	"	?	L34363	(75)
NUCPRO	*996*	Mouse	"	?	L34362	(75)
RAD16	*790*	*S.cerevisiae*	*RAD16*	Nucleotide excision repair of silent genes.	M86929	(28,29,76)
RAD5[10]	*1169*	*S.cerevisiae*	"	Post-replication repair. GT repeats more stable in mutants.	M96644	(27, 30)
RAD8	*1133*	*S.pombe*	"	Mutants have increased sensitivity to UV & gamma irradiation.	X74615	(26)
HIP116A	*1009*	Human	"	DNA-dependent ATPase. Binds HIV & SPH motifs of SV40 enhancer.	L34673	(77)
NPHCG42	*506*	*A. californica*	none	Viral encoded protein.	L22858	(78)
lodestar	*1061*	*D.melanogaster*	none	Mutants have excessive chromosome breakage & tangling in mitosis.	X62629	(79)
HepA	*968*	*E.coli*	none	Induced by DNA damage.	M81963	(80, 81)

[1]=GAM1, SWI2, TYE3

[2]=NPS1

[3]=SNF2B

[4]=SNF2A

[5]=YBR245C, YBR1633

[6]=MMKYBP

[7]=YAL019 , YAL001, YAB9

[8]=GTA1085

[9]=SCTRAAA_3 , YBRO73W , YBR0715

[10]=REV2

Table 2. Characteristics of proposed subfamilies

	Members w/ sequence similarity (In region relative to the SNF2 domain)				Bootstrap Values			
Sub-family	N-terminal	C-terminal	Variable Spacers	Conserved Motifs	Pars.	Fitch	NJ	Conserved Function
SNF2	ALL*	ALL	-	bromodomain	100	100	100	Transcription activation; remove histones from DNA?
SNF2L	ALL	ALL	-	-	100	100	100	?
CHD1	ALL	ALL	-	chromodomain	100	100	100	?
ETL1	-	-	-	-	97	100	100	?
ERCC6	ALL	ALL	-	-	100	100	100	Transcription-coupled repair; moves stalled RNA polymerase?
RAD16	some	some	ALL	ring finger	47	83	62	?
RAD54	some	some	-	-	81	92	94	Recombination repair?
MOT1	n/a	n/a	n/a	-	n/a	n/a	n/a	Removes TATA-binding protein from DNA

* Similarity among all members only over a small stretch of amino acids

Figure 1. Schematic alignment of the proteins in the SNF2 family.

The alignment was generated using the clustalv and clustalw programs and some manual modification. Continuous stretches of amino-acids in the alignment are boxed. Alignment gaps are indicated by lines joining boxes. Conserved regions of the SNF2 domain are in black. Colors were chosen to highlight proposed subfamilies. Regions flanking the SNF2 domain are colored for those that show significant similarity to other flanking regions. Blank regions show no significant similarity to other proteins in the family. The presence of motifs is indicated: C=chromodomain, BR=bromodomain, R=RING finger. Scale bar corresponds to numbers of amino acid residues in boxed regions.

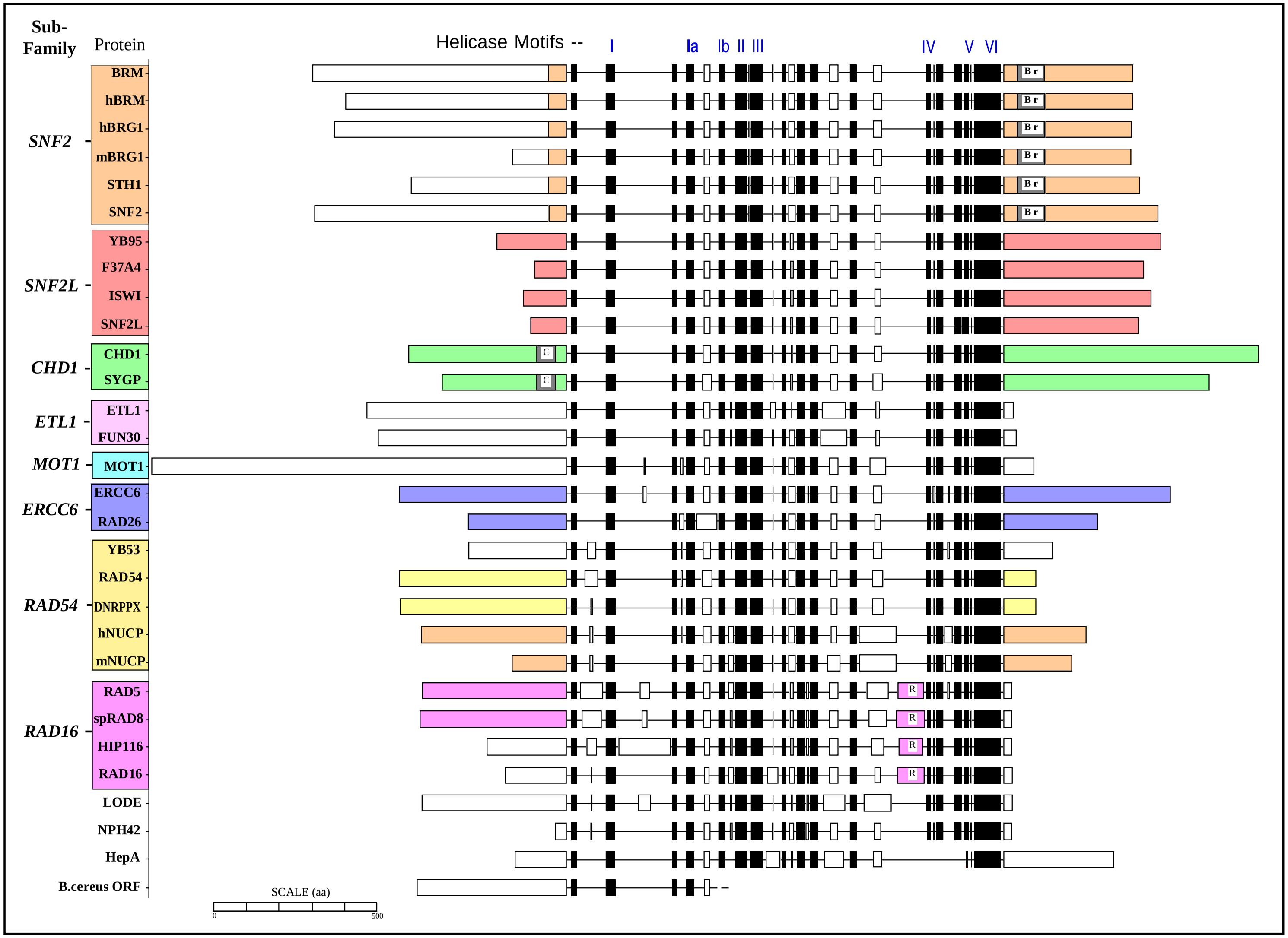
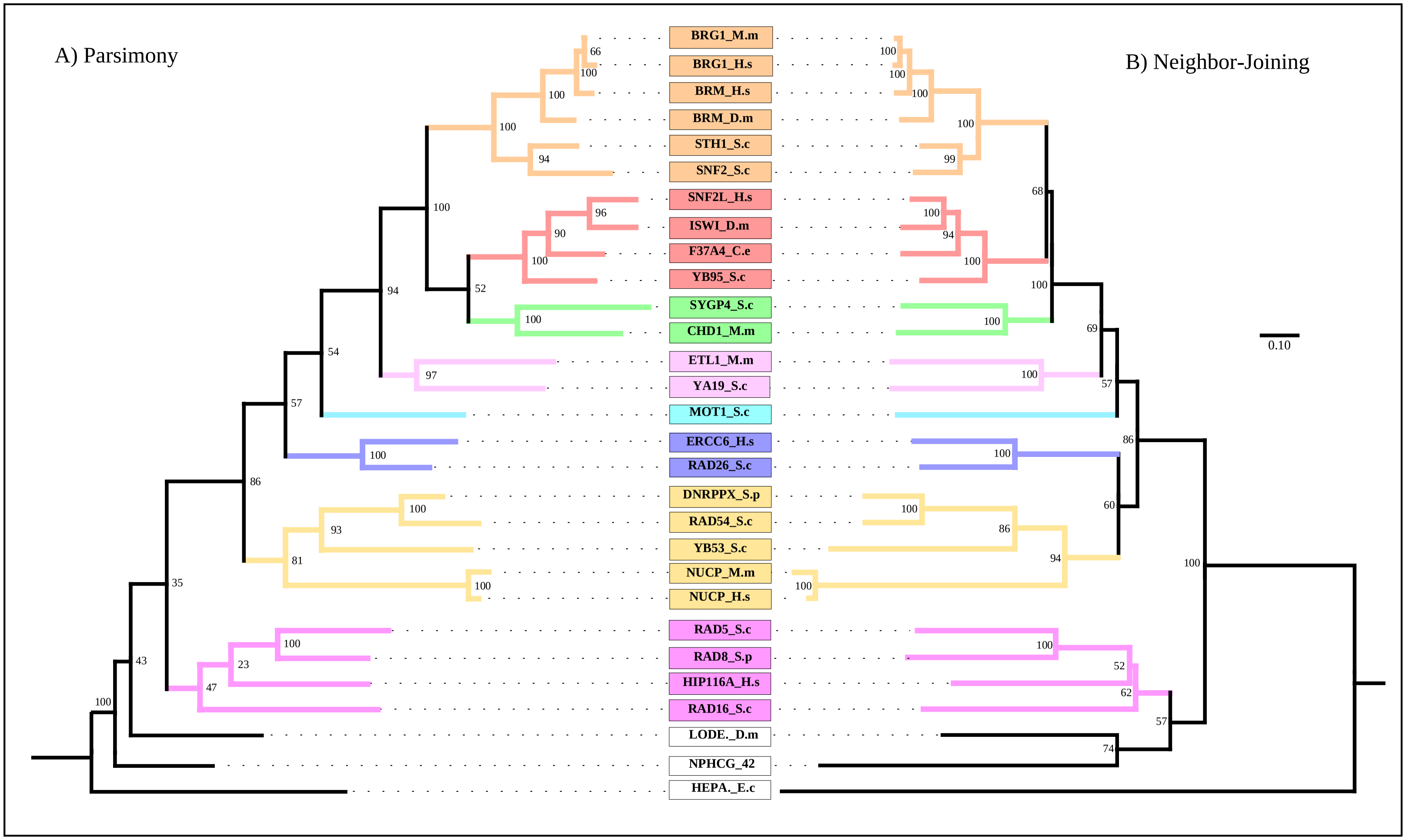

Sub-
Family
Protein
Helicase Motifs --
I
Ia
Ib
II
III
IV
V
VI
SNF2
BRM
hBRM
hBRG1
mBRG1
STH1
SNF2
SNF2L
YB95
F37A4
ISWI
SNF2L
CHD1
CHD1
SYGP
ETL1
ETL1
FUN30
MOT1
MOT1
ERCC6
ERCC6
RAD26
RAD54
YB53
RAD54
DNRPPX
hNUCP
mNUCP
RAD16
RAD5
spRAD8
HIP116
RAD16
LODE
NPH42
HepA
B.cereus ORF
Br
C
R
SCALE (aa)
0
500

Figure 2. Phylogenetic trees of the SNF2 family of proteins.

A) Parsimony tree. B) Neighbor-joining tree. Trees were generated from an alignment generated by the clustalv and clustalw programs. Regions of ambiguous alignment were excluded from the analysis. Bootstrap values, indicating the number of times a particular node was found in trees generated from 100 boostrap replicates of the alignment, are shown on the trees. The roots of the trees were determined by comparisons with other helicase domain containing proteins. Branch lengths correspond to minimum number of inferred amino acid substitutions (in A) or estimated evolutionary distance (in B). Sequences and branches are colored according to proposed subfamilies. Names are shown in the middle to aid in comparison of the two trees. For more details on tree generation see Methods.


A) Parsimony
B) Neighbor-Joining
BRG1_M.m
BRG1_H.s
BRM_H.s
BRM_D.m
STH1_S.c
SNF2_S.c
SNF2L_H.s
ISWI_D.m
F37A4_C.e
YB95_S.c
SYGP4_S.c
CHD1_M.m
ETL1_M.m
YA19_S.c
MOT1_S.c
ERCC6_H.s
RAD26_S.c
DNRPPX_S.p
RAD54_S.c
YB53_S.c
NUCP_M.m
NUCP_H.s
RAD5_S.c
RAD8_S.p
HIP116A_H.s
RAD16_S.c
LODE._D.m
NPHCG_42
HEPA._E.c
0.10
66
100
100
100
94
100
96
90
100
52
100
94
97
54
57
100
86
100
93
81
100
35
100
23
47
43
100
100
100
100
100
99
100
68
100
94
100
100
100
69
100
57
86
100
60
100
86
94
100
100
100
52
62
57
74

CHAPTER 6

Using Evolutionary Analysis to Characterize DNA Repair Processes III:

The Development of Phylogenomics

PART A

Gastrogenomic Delights: A Movable Feast[8]

[8] Previously published as *Jonathan A. Eisen; Dale Kaiser ; Richard M. Myers. 1997. Nature Medicine 3(10): 1076-1078. 1997.* Reprinted with permission from Nature Press.

ABSTRACT

The complete genome sequences of *Escherichia coli* and *Helicobacter pylori* provide insights into the biology of these species.

INTRODUCTION

Recently, we biologists have been treated to a feast of the complete genome sequences of two gut bacteria: and *Helicobacter pylori* reported by Tomb *et al*. in *Nature* (Tomb et al, 1997) and *Escherichia coli* reported by Blattner *et al*. in *Science*. (Blattner et al. 1997). Complete sequences of eight microbes have now been published (Table 1), and there are over 30 additional projects underway and slated for completion in the next 12-18 months. The finished genome sequence of *E. coli* -- metabolic generalist, workhorse of biochemical genetics, molecular biology and biotechnology, and occasional pathogen -- has special, almost emotional, significance to today's biologists, many of whom have grown up with its cultures in one form or another. By contrast, *H. pylori* -- metabolic specialist, gastric pathogen and causative agent of peptic ulcers -- is a relative newcomer to the scientific scene.

DISCUSSION

Why sequence whole genomes?

There are numerous reasons for going to the trouble of determining complete and accurate genome sequences of micro-organisms. In those microbes with pathogenic properties, the total set of instructions provides a potentially powerful basis for developing vaccines and other therapeutic agents. Genomic sequences offer insights into the range of functions an organism possesses, the relative importance natural selection attaches to each function, and the organism's evolutionary history. In addition, the availability of complete genome sequences has spawned an enormous array of creative approaches for global functional analysis of genes and gene networks. There is particular virtue in having contiguous sequence of an entire genome; not only is it possible to predict all or almost all of the proteins that are present in the organism, but what is absent also becomes meaningful.

Functional predictions

The genome sequence of an organism is like the Rosetta stone: it is impressive to see but it must be translated to have value. The most important initial steps in translating a genome are identifying all of the genes and assigning functions to them. Genes can be identified by genetic and biochemical experiments or predicted by computational analysis of the genome sequence. Functions of genes can be assigned also by experimental and computational methods, but accurate prediction of function based solely on sequence information is not so straightforward. In the case of *E. coli*, computational prediction of gene function is less important because of the vast wealth of genetic and biochemical data collected from this organism over the last fifty years (Riley, 1993). However, for *H. pylori* and for most of the species for which complete genome sequences are published, far less experimentally derived functional information is available. Thus, analysis of these genomes, and most of the ones that will be sequenced in the future, depends heavily on computational methods.

Tomb, et al. use the BLAZE program (Brutlag et al., 1993) to assign function to each predicted *H. pylori* gene based on the function of the previously characterized gene in the sequence database that is most similar in sequence to the predicted gene, but only if the likelihood of the match is much higher than that expected by chance. Blattner's group go one step further. They identify multiple similar sequences in existing databases, and if most of these genes appear to have the same physiological role, this function is assigned to the new gene. If the top scoring sequences have different physiological roles, attempts are made to identify a common denominator, such as transport activity, and this general activity, with unknown specificity, is then assigned to the new gene. Although both approaches are likely to result in correct functional assignments for most genes, there are many cases where either approach will lead to incorrect predictions.

One example where caution seems warranted is in the prediction that *H. pylori* is capable of mismatch repair, based on the assignment of methyl transferase, MutS, and UvrD functions to several of its genes. (Tomb, et al., 1997). However, it is unlikely that this DNA repair process is present in *H. pylori* because its genome sequence does not contain a homolog of MutL, a protein required for mismatch repair in all organisms studied from bacteria to humans (Modrich and Lahue, 1996). Furthermore, phylogenetic analysis suggests that there has been an ancient duplication in the *mutS* gene family, and that the "*mutS*" gene (HP0621) in *H. pylori* is not an orthologue (a gene originating from a speciation event), but is rather a paralogue (a gene originating from a gene duplication event) of the *E. coli mutS* gene (Fig. 1). Genes that are orthologs of the *E. coli mutS* gene, (Fig. 1, blue), are absolutely required for mismatch repair in many bacterial species. By contrast, the *mutS* paralogs, (Fig. 1, red), have no known function. Why was the HP0621 gene called *mutS*, and not identified as a *mutS* paralog? Analysis of the database search used by Tomb, et al. (see their web site, Table 1) indicates that the gene was given this designation because its highest sequence similarity hit was with the gene sll1772 from *Synechocystis* sp. (strain PCC6803), a cyanobacterium. The researchers annotating the *Synechocystis* sp. genome sequence earlier gave the name *mutS* to gene sll1772, again because it scored highly similar to *mutS* in a similar type of analysis. However, gene sll1772 is only one of two *mutS*-like genes in *Synechocystis* sp.; a second gene (gene sll1165) predicted from its genome sequence is much more similar to *mutS* from *E. coli*, and is the likely *mutS* orthologue in *Synechocystis* sp. *H. pylori*, for unknown reasons, does not encode an orthologue of the *mutS* genes known to be involved in mismatch repair. As this example shows, database errors are often self-propagating.

This difficulty in assigning function on the basis of sequence data is likely to be widespread, particularly because so many microbial genome sequences are forthcoming. Some simple precautions may help to alleviate the problem. Perhaps the most obvious rule is to avoid assuming that a function assigned to a sequence is correct just because it already appears in a database. The method used by Blattner et al. of examining many high scoring sequences at once may reduce the likelihood of being misled by a single database misannotation, because it assigns a function only if many of the top scoring genes have the same function. A second simple precaution is to recognize that sequence similarity indicates only the *potential* for a biochemical activity. Close similarity does not readily identify the physiological role for a protein and is not definitive evidence that two proteins have the same biochemical activity. Likewise, the absence of a homologous gene in a whole genome sequence does not necessarily mean that the activity is absent in the organism.

Phylogenomics

The MutS story above and many other examples provide evidence that classifying members of multigene families is one of the most difficult parts of assigning function. Molecular phylogenetics is probably a better method for dividing multigene families into groups of orthologous genes than simply relying on database searches. As orthologues frequently have functions distinct from paralogues, a "phylogenomic" methodology is likely to improve the accuracy of function assignment to members of multigene families identified in complete genome sequences. In addition, assignment of function on the basis of DNA sequence data will likely become more accurate as we learn how to integrate knowledge about biochemical pathways and regulatory networks into the computational methods.

Other information from genome sequences

In addition to stimulating predictions of the functions of individual genes, the complete genome sequences of *H. pylori* and *E. coli* provide clues about their global metabolic capabilities. One striking difference between these two organisms is that *H. pylori* has many fewer genes than *E. coli.* Three of the other bacteria whose complete sequences are published also have reduced genome sizes. How can this phenomenon be explained? One argument is that organisms with broad ecological niches need more genes (Hinegardner, 1976). For example, *E. coli*, with a genome of 4.6 million base pairs, can be thought of as a metabolic generalist because it is capable of growing under a variety of conditions. It is equipped to grow in the lower gut of animals where it meets a variety of sugars that have not been absorbed by its host's digestive tract. Absorption being an efficient process, the residual sugars and amino acids are dilute. The lower gut is also anaerobic; *E. coli* is a facultative anaerobe, capable of fermentative metabolism. *E. coli* survives when it is released to the environment where it can be disseminated to new hosts. It grows faster in air than in the gut, metabolizing carbon to CO_2. Its metabolic generalism shows in its genome; there are many different transport proteins to accumulate dilute substrates from the gut contents. There are 700 known gene products for central intermediary metabolism, degradation of small molecules, and energy metabolism. Helping *E. coli* adjust to a variety of growth conditions are the 400 regulatory genes (some known on the basis of experiments and some attributed for reasons of sequence similarity), or 4.5% of the total genome. By contrast, *H. pylori*, with a genome of only 1.66 million base pairs, is an ecological specialist, apparently living nowhere but in the mucosa of the stomach. To survive in this highly acidic environment, *H. pylori* encodes genes that allow it to develop a positive inside membrane potential and has double the number of basic amino acids in most of its proteins compared to other microbes. Consistent with this restricted ecological niche, the genome sequence of *H. pylori* indicates that it is much more limited in its metabolic capabilities and its regulatory networks (Tomb, et al., 1997). The genome sequence also provides clues as to how *H. pylori* survives in the highly acidic environment of the stomach. The proteins encoded by the *H. pylori* genome have twice the number of basic amino acids compared to proteins of other microbes; this may help in establishing a positive inside membrane potential. These comparisons provide but one of the many valuable insights that can be learned from sequences of complete genomes.

Summary

We have every reason to be delighted by the feast that has just been served to us. These complete genomic sequences have a major impact on the study of these two gut bacteria, and will likely speed up our understanding of the mechanisms by which they cause disease. Because these and the other available bacterial sequences are from widely divergent microbes, we are already getting an idea of which genes are universal and perhaps form the core of a micro-organism (Mushegian and Koonin, 1996). By contrast, as complete genome sequences from closely related pairs of microbes become available, we will learn more about mutation and recombination processes, as well as features such as codon usage, genome structure, and horizontal gene transfer, that change on a shorter evolutionary time scale (for example, Lawrence and Ochman, 1997). Today's feast will likely seem meager in comparison to the lavish smorgasbord expected in the future.

REFERENCES

Blattner, F. R., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453-1462..

Brutlag, D. L., Dautricourt, J. P., Diaz, R., Fier, J., Moxon, B. and Stamm, R. (1993). BLAZE: An implementation of the Smith-Waterman comparison algorithm on a massively parallel computer. *Computers and Chemistry* **17:** 203-207.

Bult, C. J., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:** 1058-1073.

Fleischmann, R. D., et. al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Fraser, C. M., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397-403.

Goffeau, A., et al., (1997). The yeast genome directory. *Nature* **387:** (Suppl.) 5-105.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C., Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24:** 4420-4449.

Hinegardner, R. (1976). Evolution of genome size. In F. J. Ayala, *Molecular Evolution*, Sinauer, Sunderland, MA. pp 179-199.

Kaneko, T., et al. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res*. **3:** 109-136.

Lawrence, J. G. and Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol*. **44:** 383-397.

Modrich, P. and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Ann. Rev. Biochem*. **65:** 101-133.

Mushegian, A. R. and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes *Proc. Natl. Acad. Sci. U. S. A.* **93**: 10268-10273.

Riley, M. (1993). Functions of the gene products of *Escherichia coli*. *Micro. Rev*. **57:** 862-952.

Tomb, J-F., et al. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539-547.

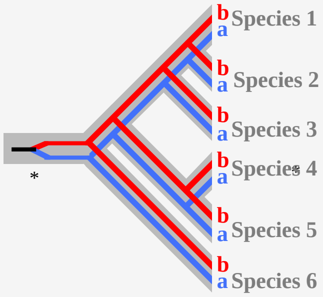
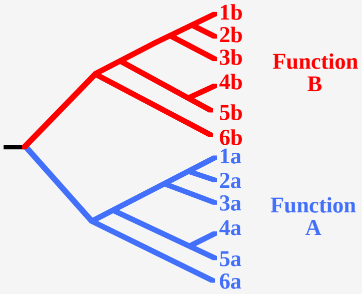
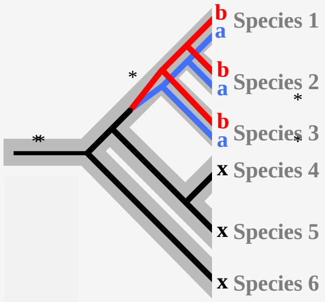
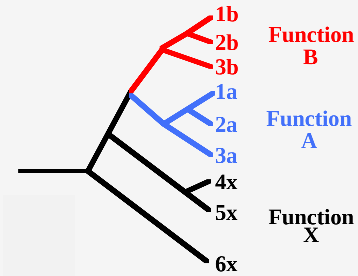
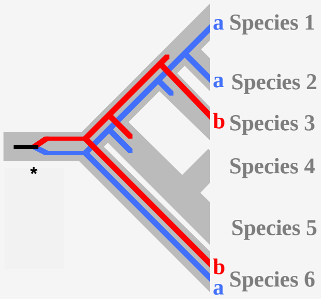
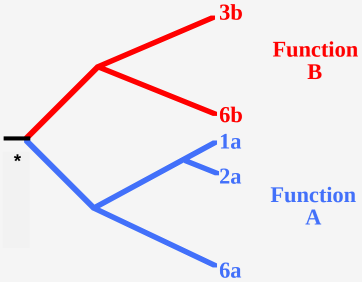
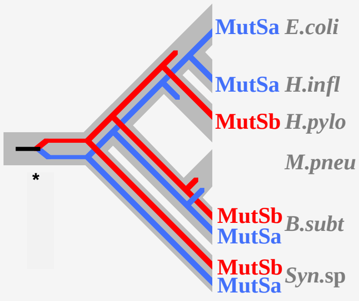
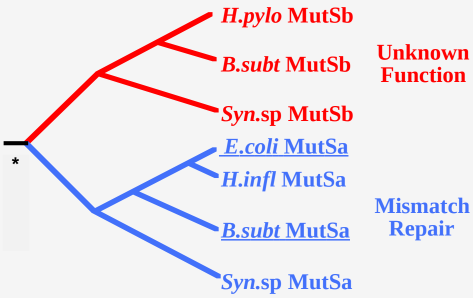
Table 1. Complete genomes.

Species	Classification	Size (mb)	Orfs	Ref.
Bacteria				
Mycoplasma genitalium G-37	LowGC Gram +	0.58	470	Fraser et al. 1995[1]
Mycoplasma pneumoniae M129	LowGC Gram +	0.82	679	Himmelreich et al. 1996[2]
Escherichia coli K-12	Proteobacteria (γ)	4.60	4288	Blattner et al. 1997[3]
Haemophilus influenzae KW20	Proteobacteria (γ)	1.83	1743	Fleischman et al. 1995[4]
Helicobacter pylori 26695	Proteobacteria (ε)	1.67	1590	Tomb et al. 1997[5]
Synechocystis sp. PCC6803	Cyanobacteria	3.57	3168	Kaneko et al. 1996[6]
Archaea				
Methanococcus jannascii	Euryarchaeota	1.66	1738	Bult et al. 1996[7]
Eukaryote				
Saccharomyces cerevisiae	Fungi	13.0	5885	Goffeau et al. 1997[8]

[1] www.tigr.org/tdb/mdb/mgdb/mgdb.html
[2] www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html
[3] www.genetics.wisc.edu:80/index.html
[4] www.tigr.org/tdb/mdb/hidb/hidb.html
[5] www.tigr.org/tdb/mdb/hpdb/hpdb.html
[6] www.kazusa.or.jp/cyano/cyano.html
[7] www.tigr.org/tdb/mdb/mjdb/mjdb.html
[8] genome-www.stanford.edu/Saccharomyces/

Figure 1. Evolution of multigene families.

left, Gene trees (thin lines) is shown embedded within the species tree (thick grey lines). Gene duplication events (marked by an asterisk) result in multiple paralogous genes (distinguished by different colors and gene subscripts a or b) within one species. Gene loss in some lineages is indicated when the gene tree stops within the species tree. *right*, The gene trees are extracted from the species tree and untwisted to better show the relaitonships among the different gene forms. A-C: hypothetical scenarios. D. Reconstruction of the evolution of MutS-like proteins in bacteria using molecular phylogenetics. MutS-like protein sequences were aligned and a tree of these sequences was generated using molecular phylogenetic methods (*details are available from the authors on request*). The gene duplication event occurred prior to the divergence of these bacterial species and led to the presence of two paralogous MutS-like subgroups. Not the loss of genes in some lineages. Only one lineage (in blue) includes genes with established roles in mismatch repair. The genes in the second lineage (in red) have no known function. Because the *H. pylori* gene is a member of this second lineage, it should not be assigned the MutS function.

	Gene history traced within species phylogenetic tree.	Gene history traced independently of species phylogenetic tree
A. Ancient Duplication	 b a Species 1 b a Species 2 b a Species 3 b a Species 4 b a Species 5 b a Species 6 *	 1b 2b 3b 4b 5b 6b Function B 1a 2a 3a 4a 5a 6a Function A
B. Recent Duplication	 b a Species 1 b a Species 2 b a Species 3 x Species 4 x Species 5 x Species 6 * *	 1b 2b 3b Function B 1a 2a 3a Function A 4x 5x 6x Function X
C. Ancient Duplication and Gene Loss	 a Species 1 a Species 2 b Species 3 Species 4 Species 5 b a Species 6 *	
D. Evolution of the MutS Family of Proteins in Bacteria	 MutSa *E.coli* MutSa *H.infl* MutSb *H.pylo* *M.pneu* MutSb MutSa *B.subt* MutSb MutSa *Syn.sp* *	 *H.pylo* MutSb *B.subt* MutSb *Syn.sp* MutSb Unknown Function *E.coli* MutSa *H.infl* MutSa *B.subt* MutSa *Syn.sp* MutSa Mismatch Repair *

PART B

Phylogenomics:
Improving Functional Predictions for Uncharacterized Genes
by Evolutionary Analysis[9]

[9] Previously published as: *Jonathan A. Eisen. 1998. Genome Research 8(3): 163-167.* Reprinted with permission from Cold Spring Harbor Press.

INTRODUCTION

The ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age where numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization, and quantification of sequence-similarity between the gene of interest and genes for which functional information is available; since sequence is the prime determining factor of function, sequence-similarity is taken to imply similarity of function. There is no doubt that this assumption is valid in most cases. However, sequence-similarity does not ensure identical functions, and it is common for groups of genes that are similar in sequence to have diverse (although usually related) functions. Therefore, the identification of sequence-similarity is frequently not enough to assign a predicted function to an uncharacterized gene; one must have a method of choosing among similar genes with different functions. In such cases, most functional prediction methods assign likely functions by quantifying the levels of similarity among genes. I suggest that functional predictions can be greatly improved by focusing on *how* the genes became similar in sequence (i.e., evolution) rather than on the sequence-similarity itself. It is well established that many aspects of comparative biology can benefit from evolutionary studies (Felsenstein 1985) and comparative molecular biology is no exception (e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic* approach to the prediction of gene function, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

DISCUSSION

Sequence-similarity, homology, and functional predictions

To make use of the identification of sequence-similarity between genes, it is helpful to understand how such similarity arises. Genes can become similar in sequence either as a result of *convergence* (similarities that have arisen without a common evolutionary history) or descent with modification from a common ancestor (also known as *homology*). It is imperative to recognize that sequence-similarity and homology are not interchangeable terms. Not all homologs are similar in sequence (i.e., homologous genes can diverge so much that similarities are difficult or impossible to detect) and not all similarities are due to homology (Reeck et al. 1987; Hillis 1994). Similarity due to convergence, which is likely limited to small regions of genes, can be useful for some functional predictions (Henikoff et al. 1997). However, most sequence-based functional predictions are based on the identification (and subsequent analysis) of similarities that are thought to be due to homology. Since homology is a statement about common ancestry, it cannot be proven directly from sequence-similarity. In these cases, the inference of homology is made based on finding levels of sequence-similarity that are thought to be too high to be due to convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions because not all homologs have the same function. The available similarity-based functional prediction methods can be distinguished by how they choose the homolog whose function is most relevant to a particular uncharacterized gene (Table 1). Some methods are relatively simple -- many researchers use the highest scoring homolog (as determined by programs like BLAST or BLAZE) as the basis for assigning function. While highest-hit methods are very fast, can be automated readily, and are likely accurate in many instances, they do not take advantage of any information about how genes and gene functions evolve. For example, gene duplication and subsequent divergence of function of the duplicates can result in homologs with different functions being present within one species. Specific terms have been created to distinguish homologs in these cases: genes of the same duplicate group are called *orthologs* (e.g., beta globin from mouse and humans) and different duplicates are called *paralogs* (e.g., alpha and beta globin) (Fitch 1970). Since gene duplications are frequently accompanied by functional divergence, dividing genes into groups of orthologs and paralogs can improve the accuracy of functional predictions. Recognizing that the one-to-one sequence comparisons used by most methods do not reliably distinguish orthologs from paralogs, Tatusov et al. developed the COG clustering method (Tatusov et al. 1997, see Table 1). While the COG method is clearly a major advance in identifying orthologous groups of genes, it is limited in its power because clustering is a way of classifying levels of similarity and is not an accurate method of inferring evolutionary relationships (Swofford et al. 1996). Thus, since sequence-similarity and clustering are not reliable estimators of evolutionary relatedness, and since the incorporation of such phylogenetic information has been so useful to other areas of biology, evolutionary techniques should be useful for improving the accuracy of predicting function based on sequence-similarity.

Phylogenomics

There are many ways in which evolutionary information can be used to improve functional predictions. In this commentary, I present an outline of one such *phylogenomic* method (see Fig. 1) and I compare this method to non-evolutionary functional prediction methods. This method is based on a relatively simple assumption -- since gene functions change as a result of evolution, reconstructing the evolutionary history of genes should help predict the functions of uncharacterized genes. The first step is the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. Such trees are distinct from clusters and other means of characterizing sequence similarity because they are inferred by special techniques that help convert patterns of similarity into evolutionary relationships (see Swofford et al. 1996). After the gene tree is inferred, biologically determined functions of the various homologs are overlaid onto the tree. Finally, the structure of the tree and the relative phylogenetic positions of genes of different functions are used to trace the history of functional changes, which is then used to predict functions of uncharacterized genes. More detail of this method is provided below:

Step 1. Identification of homologs

The first step in studying the evolution of a particular gene is the identification of homologs. As with similarity-based functional prediction methods, likely homologs of a particular gene are identified through database searches. Since phylogenetic methods benefit greatly from more data, it is useful to augment this initial list by using identified homologs as queries for further database searches or using automatic iterated search methods such as PSI-BLAST (Altschul et al. 1997). If a gene family is very large (e.g., ABC transporters), it may be necessary to only analyze a subset of homologs. However, this must be done with extreme care since one might accidentally leave out proteins that would be important for the analysis.

Step 2. Alignment and masking

Sequence alignment for phylogenetic analysis has a particular purpose -- it is the assignment of *positional homology*. Each column in a multiple sequence alignment is assumed to include amino-acids or nucleotides that have a common evolutionary history and each column is treated separately in the phylogenetic analysis. Therefore, regions in which the assignment of positional homology is ambiguous should be excluded (Gatesy et al. 1993). The exclusion of certain alignment positions, (also known as *masking*) helps to give phylogenetic methods much of their discriminatory power. Phylogenetic trees generated without masking (as is done in many sequence analysis software packages) are less likely to accurately reflect the evolution of the genes than trees with masking.

Step 3. Phylogenetic trees

For extensive information about generating phylogenetic trees from sequence alignments see (Swofford et al. 1996). In summary, there are three methods commonly used: parsimony, distance, and maximum likelihood, and each has its advantages and disadvantages. I prefer distance methods because they are the quickest when using large data sets. Before using any particular tree it is important to estimate the robustness and accuracy of the phylogenetic patterns it shows (through techniques such as the comparison of trees generated by different methods and bootstrapping). Finally, in most cases, it is also useful to determine a root for the tree.

Step 4. Functional predictions

To make functional predictions based on the phylogenetic tree, it is necessary to first overlay any known functions onto the tree. There are many ways this "map" can then be used to make functional predictions but I recommend splitting the task into two steps. First, the tree can be used to identify likely gene duplication events in the past. This allows the division of the genes into groups of orthologs and paralogs (e.g., Eisen et al. 1995). Uncharacterized genes can be assigned a likely function if the function of any ortholog is known (and if all characterized orthologs have the same function). Second, parsimony reconstruction techniques (Maddison and Maddison 1992) can be used to infer the likely functions of uncharacterized genes by identifying the evolutionary scenario that requires the fewest functional changes over time (Fig. 1). The incorporation of more realistic models of functional change (and not just minimizing the total number of changes) may prove to be useful but the parsimony minimization methods are probably sufficient in most cases.

Is the phylogenomic method worth the trouble?

Phylogenomic methods require many more steps and usually much more manual labor than similarity-based functional prediction methods. Is the phylogenomic approach worth the trouble? Many specific examples exist in which gene function has been shown to correlate well with gene phylogeny (Eisen et al. 1995; Atchley and Fitch 1997). While no systematic comparisons of phylogenetic versus similarity-based functional prediction methods have been done, there are a variety of reasons to believe that the phylogenomic method should produce more accurate predictions than similarity-based methods. In particular, there are many conditions in which similarity-based methods are likely to make inaccurate predictions but which can be dealt with well by phylogenetic methods (see Table 2).

A specific example helps illustrate a potential problem with similarity based methods. Molecular phylogenetic methods show conclusively that mycoplasmas share a common ancestor with low-GC Gram-positive bacteria (Weisburg et al. 1989). However, examination of the percent similarity between mycoplasmal genes and their homologs in bacteria does not clearly show this relationship. This is because mycoplasmas have undergone an accelerated rate of molecular evolution relative to other bacteria. Thus a BLAST search with a gene from *B. subtilis* (a low GC Gram-positive species) will result in a list in which the mycoplasma homologs (if they exist) score lower than genes from many species of bacteria less closely related to *B. subtilis*. When amounts or rates of change vary between lineages, phylogenetic methods are better able to infer evolutionary relationships than similarity methods (including clustering) because they allow for evolutionary branches to have different lengths. Thus, in those cases in which gene function correlates with gene phylogeny and in which amounts or rates of change vary between lineages, similarity-based methods will be more likely than phylogenomic methods to make inaccurate functional predictions (see Table 2).

Another major advantage of phylogenetic methods over most similarity methods comes from the process of masking (see above). For example, a deletion of a large section of a gene in one species will greatly affect similarity measures but may not affect the function of that gene. A phylogenetic analysis including these genes could exclude the region of the deletion from the analysis by masking. In addition, regions of genes that are highly variable between species are more likely to undergo convergence and such regions can be excluded from phylogenetic analysis by masking. Masking thus allows the exclusion of regions of genes in which sequence-similarity is likely to be "noisy" or misleading rather than biologically important signal. The pairwise sequence comparisons used by most similarity-based functional prediction methods do not allow such masking. Phylogenetic methods have been criticized because of their dependence (for most methods) on multiple sequence alignments which are not always reliable and unbiased. However, multiple sequence alignments also allow for masking which is probably more valuable than the cost of depending on alignments.

The conditions described above and highlighted in Table 2 are just some examples of conditions in which evolutionary methods are more likely to make accurate functional predictions than similarity-based methods. Phylogenetic methods are particularly useful when the history of a gene family includes many of these conditions (e.g., multiple gene duplications plus rate variation) or when the gene family is very large. The principle is simple -- the more complicated the history of a gene family, the more useful it is to try to infer that history. Thus although the phylogenomic method is slow and labor intensive I believe it is worth using if accuracy is the main objective. In addition, information about the evolutionary relationships among gene homologs is useful for summarizing relationships among genes and for putting functional information into a useful context.

Summary

Despite the evolution of these methods, and likely continued improvements in functional predictions, it must be remembered that the key word is *prediction*. All methods are going to make inaccurate predictions of functions. For example, none of the methods described can perform well when gene functions can change with little sequence change as has been seen in proteins like opsins (Yokoyama 1997). Thus sequence databases and genome researchers should make clear which functions assigned to genes are based on predictions and which are based on experiments. In addition, all prediction methods should use only experimentally determined functions as their grist for predictions. This will hopefully limit error propagation that can happen by using an inaccurate prediction of function to then predict the function of a new gene, which is a particular problem for the highest-hit methods since they rely on the function of only one gene at a time to make predictions (Eisen et al. 1997). Despite these and other potential problems, functional predictions are of great value in guiding research and in sorting through huge amounts of data. I believe that the increased use of phylogenetic methods can only serve to improve the accuracy of such functional predictions.

REFERENCES

Altschul, S.F., R.J. Carroll and D.J. Lipman. 1989. *J. Mol. Biol.* **207**: 647-653.

Altschul, S.F., T.L. Madden, A.A. Schaeffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. *Nucleic Acids Res.* **25**: 3389-3402.

Atchley, W.R. and W.M. Fitch. 1997. *Proc. Natl. Acad. Sci. U.S.A.* **94**: 5172-5176.

Blattner, F.R., G.I. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, et al. 1997. *Science* **277**: 1453-1462.

Bolker, J.A. and R.A. Raff. 1996. *Bioessays* **18**: 489-494.

Eisen, J.A., D. Kaiser and R.M. Myers. 1997. *Nature (Medicine)* **3**: 1076-1078.

Eisen, J.A., K.S. Sweder and P.C. Hanawalt. 1995. *Nucleic Acids Res.* **23**: 2715-2723.

Felsenstein, J. 1985. *Am. Nat.* **125**: 1-15.

Fitch, W.M. 1970. *Syst. Zool.* **19**: 99-113.

Gatesy, J., R. Desalle and W. Wheller. 1993. *Mol. Phylog. Evol.* **2**: 152-157.

Goldman, N., J.L. Thorne and D.T. Jones. 1996. *J. Mol. Biol.* **263**: 196-208.

Henikoff, S., E.A. Greene, S. Pietrovsky, P. Bork, T.K. Attwood and L. Hood. 1997. *Science* **278**: 609-614.

Henikoff, S., S. Pietrokovski and J.G. Henikoff. 1998. *Nucleic Acids Res.* **26**: 311-315.

Hillis, D.M. 1994. In *Homology: the Hierarchical Basis of Comparative Biology* (ed. B. K. Hall), pp. 339-368. Academic Press, Inc., San Diego.

Maddison, W.P. and D.R. Maddison. 1992. *MacClade*. Sinauer Associates, Inc., Sunderland, MA.

Reeck, G.R., C. Haën, D.C. Teller, R.F. Doolittle, W.M. Fitch, R.E. Dickerson, P. Chambon, A.D. McLachlan, E. Margoliash, T.H. Jukes, et al. 1987. *Cell* **50**: 667.

Swofford, D.L., G.J. Olsen, P.J. Waddell and D.M. Hillis. 1996. In *Molecular Systematics* (ed. D. M. Hillis, C. Moritz and B. K. Mable), pp. 407-514. Sinauer Associates, Sunderland, MA.

Tatusov, R.L., E.V. Koonin and D.J. Lipman. 1997. *Science* **278**: 631-637.

Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, et al. 1997. *Nature* **388**: 539-547.

Weisburg, W.G., J.G. Tully, D.L. Rose, J.P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T.G. Lawrence, V.E. J., et al. 1989. *J. Bacteriol.* **171**: 6455-6467.

Yokoyama, S. 1997. *Ann. Rev. Genet.* **31**: 315-336.

Zardoya, R., E. Abouheif and A. Meyer. 1996. *Trends Genet.* **12**: 496-497.

Table 1. Methods of predicting gene function when homologs have multiple functions.

Highest Hits.

The uncharacterized gene is assigned the function (or frequently, the annotated function) of the gene that is identified as the highest hit by a similarity search program (e.g., Tomb et al. 1997).

Top Hits.

Identify top10+ hits for the uncharacterized gene. Depending on the degree of consensus of the functions of the top hits, the query sequence is either assigned a specific function, a general activity with unknown specificity, or no function (e.g., Blattner et al. 1997).

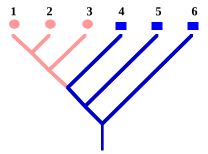
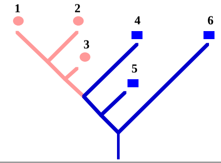
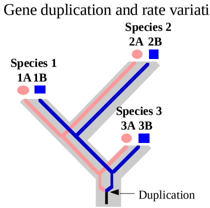
Clusters of Orthologous Groups.

Divides genes in groups of orthologs based on a cluster analysis of pairwise similarity scores between genes from different species. Uncharacterized genes are assigned the function of characterized orthologs (Tatusov et al. 1997).

Phylogenomics.

Known functions are overlaid onto an evolutionary tree of all homologs. Functions of uncharacterized genes are predicted by their phylogenetic position relative to characterized genes (e.g., Eisen et al. 1995; Eisen et al. 1997).

Table 2. Examples of conditions in which similarity methods produce inaccurate predictions of function.

		Highest Hit Method		Phylogenomic Method		
Evolutionary Pattern and Tree of Genes and Functions[1]	Gene With Unknown Function[2]	Predicted Function[3]	Accurate?	Predicted Function[4]	Accurate?	Comments
A. Functional change during evolution.  	1 ●	●	+	●	+	• Phylogenomic method cannot predict functions for all genes, but the predictions that are made are accurate. • Highest hit method is misleading because function changed among homologs but hierarchies of similarity do not correlate with the function (see Bolker and Raff 1996).
	2 ●	●	+	●	+	
	3 ●	●	+	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●/■	±	■	+	
	6 ■	●/■	±	■	+	
B. Functional change & rate variation.  	1 ●	●	+	●	+	• Similarity based methods perform particularly poorly when evolutionary rates vary between taxa. • Molecular phylogenetic methods can allow for rate variation and reconstruct gene history reasonably accurately.
	2 ●	●	+	●	+	
	3 ●	■	-	●/■	±	
	4 ■	●	-	●/■	±	
	5 ■	●	-	■	+	
	6 ■	■	+	■	+	
C. Gene duplication and rate variation.  Species 2 2A 2B Species 1 1A 1B Species 3 3A 3B Duplication	1 A ●	●	+	●	+	• Most-similarity based methods are not ideally set up to deal with cases of gene duplication since orthologous genes do not always have significantly more sequence similarity to each other than to paralogs (Eisen et al. 1995; Zardova et al. 1996; Tatusov et al. 1997). • Similarity-based methods perform particularly poorly when rate variation and gene duplication are combined. This even applies to the COG method (see Table1) since it works by classifying levels of similarity and not by inferring history. Nevertheless, the COG method is a significant improvement over other similarity based methods in classifying orthologs. • Phylogenetic reconstruction is the most reliably way to inferr gene duplication events and thus determine orthology.
	2 A ●	●	+	●	+	
	3 A ●	■	-	●	+	
	1 B ■	■	+	■	+	
	2 B ■	■	+	■	+	
	3 B ■	●	-	■	+	

[1] The true tree is shown but it is assumed that it is not known. Different colors and symbols represent different functions. Numbers correspond to different species.
[2] The function of all other genes is assumed to be known.
[3] The top hit can be determined from the tree by finding the gene is the shortest evolutionary distance away (as determined along the branches of the tree).
[4] It is assumed that the tree of the genes can be reproduced accurately by molecular phylogenetic methods (see Fig. 1).

Table 3. Types of molecular homology.

Type of Homology	**Definition**	**Examples**
Homologs	Genes that are descended from a common ancestor.	All globins.
Orthologs	Homologous genes that have diverged from each other after *speciation* events.	Human beta globin and chimp beta globin.
Paralogs	Homologous genes that have diverged from each other after *gene duplication* events.	Beta and gamma globin.
Xenologs	Homologous genes that have diverged from each other after *lateral gene transfer* events.	Antibiotic resistance genes in bacteria.
Positional Homology	Common ancestry of specific amino-acid or nucleotide positions in different genes.	Conserved oxygen binding histidine in globins.

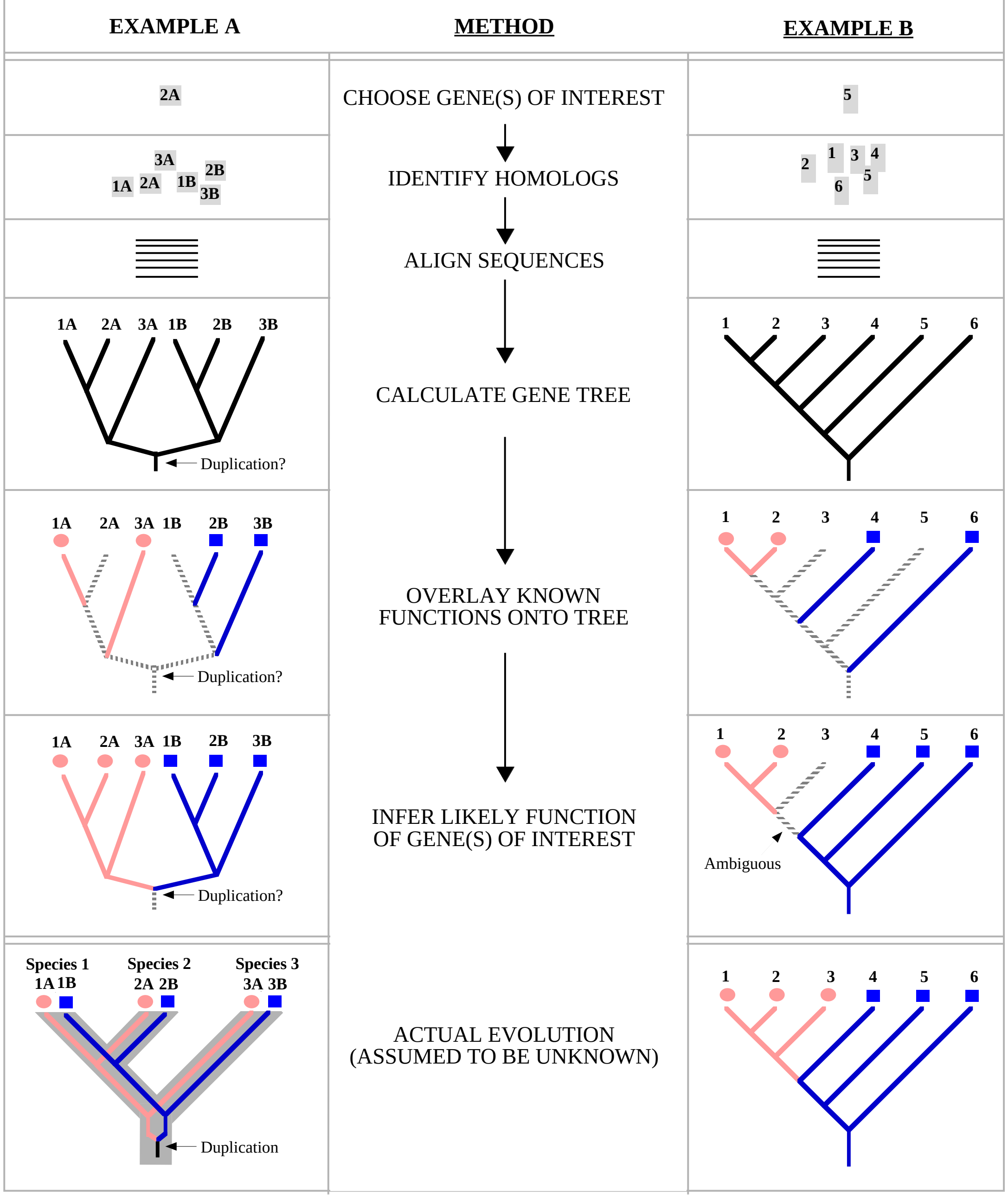
Table 4. Molecular phylogenetic methods.

Method	Description
Parsimony	Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino-acid substitutions) that would be required over evolutionary time to fit the sequences into that tree. The optimal tree is considered to be the one requiring the fewest changes (the most parsimonious tree).
Distance	The optimal tree is generated by first calculating the estimated evolutionary distance between all pairs of sequences. Then these distances are used to generate a tree in which the branch patterns and lengths best represent the distance matrix.
Maximum Likelihood	Similar to parsimony methods in that possible trees are compared and given a score. The score is based on how likely the given sequences are to have evolved in a particular tree given a model of amino-acid or nucleotide substitution probabilities. The optimal tree is considered to be the one that has the highest probability.
Bootstrapping	Alignment positions within the original multiple sequence alignment are resampled and new data sets are made. Each bootstrapped data set is used to generate a separate phylogenetic tree and the trees are compared. Each node of the tree can be given a bootstrap percentage indicating how frequently those species joined by that node group together in different trees. Bootstrap percentage does not correspond directly to a confidence limit.

Figure 1. Outline of a phylogenomic methodology.

In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. In A) a gene family has undergone a gene duplication that was accompanied by functional divergence. In B) gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the bottom. The genes are referred to by numbers (which represent the species from which these genes come) and letters (which in A represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in A) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions and gray (with hatching) represents either unknown or unpredictable functions.

PHYLOGENENETIC PREDICTION OF GENE FUNCTION



EXAMPLE A

METHOD

EXAMPLE B

CHOOSE GENE(S) OF INTEREST

IDENTIFY HOMOLOGS

ALIGN SEQUENCES

CALCULATE GENE TREE

OVERLAY KNOWN FUNCTIONS ONTO TREE

INFER LIKELY FUNCTION OF GENE(S) OF INTEREST

ACTUAL EVOLUTION (ASSUMED TO BE UNKNOWN)

PART C

A Phylogenomic Study of the MutS Family of Proteins[10]

[10] Previously published as *Jonathan A. Eisen. 1998. Nucleic Acids Research 26(18): 4291-4300.* Reprinted with permission from Oxford University Press.

ABSTRACT

The MutS protein of *Escherichia coli* plays a key role in the recognition and repair of errors made during the replication of DNA. Homologs of MutS have been found in many species including eukaryotes, Archaea, and other bacteria, and together these proteins have been grouped into the MutS family. Although many of these proteins have similar activities to the *E. coli* MutS, there is significant diversity of function among the MutS family members. This diversity is even seen within species – many species encode multiple MutS homologs with distinct functions. To better characterize the MutS protein family, I have used a combination of phylogenetic reconstructions and analysis of complete genome sequences. This phylogenomic analysis is used to infer the evolutionary relationships among the MutS family members and to divide the family into subfamilies of orthologs. Analysis of the distribution of these orthologs in particular species and examination of the relationships within and between subfamilies is used to identify likely evolutionary events (e.g., gene duplications, lateral transfer and gene loss) in the history of the MutS family. In particular, evidence is presented that a gene duplication early in the evolution of life resulted in two main MutS lineages, one including proteins known to function in mismatch repair and the other including proteins known to function in chromosome segregation and crossing-over. The inferred evolutionary history of the MutS family is used to make predictions about some of the uncharacterized genes and species included in the analysis. For example, since function is generally conserved within subfamilies and lineages, it is proposed that the function of uncharacterized proteins can be predicted by their position in the MutS family tree. The uses of phylogenomic approaches to the study of genes and genomes are discussed.

INTRODUCTION

The ability to recognize and repair mismatches in DNA after replication has occurred has been well documented in many species. While some such mismatch repair (MMR) is carried out by pathways that repair only specific DNA replication errors, most is performed by broad specificity "general" MMR pathways. The most extensively studied general MMR system is the MutHLS pathway of the bacterium *Escherichia coli* (see (1,2), for review). In the first critical step in this pathway, the MutS protein (in the form of a dimer) binds to the site of a mismatch in double-stranded DNA. Through a complex interaction between MutS, MutL and MutH, a section of the newly replicated DNA strand (and thus the strand with the replication error) at the location of the mismatch bound by MutS is targeted for removal. Other proteins complete the repair process: the section of DNA that has been targeted is removed and degraded, a patch is synthesized using the complementary strand as a template, and the patch is ligated into place resulting in a section of double-stranded DNA without mismatches.

The ability of the MutHLS pathway to repair many types of replication errors is due to the broad specificity of MutS recognition and binding. Since MutS binds to many types of base:base mismatches, the MutHLS pathway can repair many types of base misincorporation errors. Similarly, since MutS binds to heteroduplex loops (in which one strand contains extra-helical bases) the MutHLS pathway can repair frameshift replication errors. This ability to repair loops was somewhat surprising since this pathway was originally characterized as being involved in repairing mismatches. The repair of loops is particularly important in the regulation of the stability of microsatellites (loci that contain small 1-10 bp tandem repeats). Microsatellites are particularly prone to a special class of frameshift replication errors due to a process known as slip-strand mispairing (SSM). This process leads to the generation of loops of one or more copies of repeat unit (3,4). The MutHLS pathway helps keep microsatellite mutation rates in check by repairing many of the loops generated by SSM (5). While the specificity of MutS binding (and thus the MutHLS pathway) is quite broad, it is not uniform. For example, MutS does not bind C:C mismatches well and therefore the misincorporation of a C opposite a C will not be repaired well by the MutHLS pathway (6). Binding of MutS to heteroduplex loops is also not uniform. MutS only binds loops of up to four bases in size and only binds well to those up to three bases in size (7). Thus frameshift errors are only repaired if they produce loops of four bases or smaller. Since loops generated by SSM in microsatellites are usually one repeat unit in size, microsatellites with repeats larger than four base pairs are highly unstable in *E. coli*. The non-uniformity of MutS recognition causes the MutHLS pathway to influence not only the mutation *rate*, but also the mutation *spectrum*.

The overall scheme of the MutHLS pathway (mismatch recognition, strand discrimination and excision, and resynthesis) is conserved in the general MMR systems of other species (1). However, the degree of conservation of specific details varies greatly between the different steps in the process. Some steps (e.g., strand recognition) do not even use the same general mechanism between species. Others (e.g., exonucleolytic degradation) are similar in biochemical mechanism but make use of non-homologous proteins in different species. Nevertheless, some of the specific details of the MMR process are highly conserved. In particular, homologs of MutL and MutS are required for general MMR in all species examined and these proteins function in much the same way as the *E. coli* MutL and MutS (1). The conservation of MutS between species makes the specificity of MMR similar to that of *E. coli.* As with the *E. coli* MutHLS pathway, all characterized general MMR systems can repair both mismatches and loops. Incidentally, this is what led to the discovery that hereditary non-polyposis colon cancer (HNPCC) can be caused by defects in MMR (8). Cells from patients with HNPCC showed exceptionally high levels of microsatellite instability, due to defects in loop repair.

While the ability to repair both loops and mismatches is conserved, the specificity of other species MMR is not identical to that of *E. coli.*. As with *E. coli,* dissecting the specificity of MMR in other species requires dissection of the binding preferences of MutS (or in these cases MutS homologs). However, in many cases the comparison to the *E. coli* MutS is complicated. For example, the best-studied eukaryotic MMR system is that of the yeast *Saccharomyces cerevisiae*. Unlike *E. coli, S. cerevisiae* encodes six MutS homologs, referred to as MSH proteins (for MutS Homolog) (9). The best characterized of these are MSH2, MSH3 and MSH6 which are involved in MMR in the nucleus. These proteins are combined to create two distinct heterodimers, one for recognizing and repairing base:base mismatches and loops of one to two bases (composed of MSH2 and MSH6) and one for recognizing and repairing larger loops (composed of MSH2 and MSH3) (4,10). Thus since MSH2 is in both heterodimers it is required for all MMR in the nucleus, while MSH3 and MSH6 provide the specificity for the type of replication error recognized. The roles of the other MutS homologs in *S. cerevisiae* are not as well understood. MSH1 is involved in the repair of mismatches in mitochondrial DNA, although its exact function is not known (11-13). MSH4 and MSH5 do not even function in MMR, but instead are involved in meiotic crossing-over and chromosome segregation (14-16). The role of MutS homologs in processes other than correction of replication errors is not surprising since mismatches can arise in a variety of cellular circumstances. The proteins in the *E. coli* MutHLS pathway also have alternative cellular roles including the regulation of interspecies recombination and the repair of certain types of DNA damage (1,17). It may be that some of the multiple roles of the *E. coli* MutS have been divided up among the many *S. cerevisiae* MutS homologs.

Mismatch recognition and repair in humans and other animals in quite similar to that of *S. cerevisiae* (18-20). Preliminary studies suggest that this is also true for plants (21). These similarities suggest that the complex MMR system of *S. cerevisiae* was established prior to the divergence of animal and fungal and plant ancestors. While studies of MMR in model species like humans, *S. cerevisiae*, and *E. coli* will likely continue, most new information about the MutS family of proteins is coming in the form of sequence data. Sequences of MutS homologs continue to pour into sequence databases, most without any accompanying functional information. An important new source of these sequences has been genome projects and the results coming out of these projects are somewhat surprising. For example, two MutS homologs have been found in many bacterial species as a result of bacterial genome projects (22,23); but it is not known if their functions are distinct. In addition, some bacteria do not encode any MutS homologs. In addition, some species do not encode any MutS homologs while others encode a MutS homolog but no MutL homolog (24).

How can one make sense out of the ever-expanding MutS family, the diversity of MutS proteins within particular species, and these unusual distribution patterns in complete genome sequences? In this paper, I describe a new type of analysis, which I refer to as phylogenomics, focused specifically on the MutS family of proteins. This analysis provides insight into the evolution of the MutS protein family and the diversity of functions within and between species. In addition, it allows improved predictions of the functions of uncharacterized genes in the MutS family, and the likely phenotypes of species for which complete genomes are available. Such a phylogenomic analysis can be useful to studies of any gene family.

METHODS

The sequences of previously characterized MutS-like proteins were downloaded from the National Center for Biotechnology Information (NCBI) databases (accession numbers are given in Table 1). Additional members of the MutS family were searched for using the blast (25), blast2 and PSI-blast (26) computer programs. Databases searched included the NCBI non-redundant database and unpublished nearly complete genome sequences of *Deinococcus radiodurans* and *Treponema pallidum* from The Institute for Genomic Research (27) and *Streptococcus pyogenes* and *Neisseria gonorrhoeae* from University of Oklahoma (28).

Protein sequences were aligned using the *clustalw* (29) and *clustalx* (30) multiple sequence alignment programs with some manual adjustment using the GDE computer software package (31,32). Regions of ambiguity in this alignment were determined by comparison to alternative alignments generated using modifications of the alignment parameters (such as different gap penalties).

Phylogenetic trees were generated from the sequence alignments using the PAUP* program (33) on a PowerBook 3400/180. Parsimony analysis was conducted using the *heuristic search* algorithm. The total branch lengths of trees was quantified using either an identity matrix, a PAM250 matrix, or a MutS-specific matrix (based on the frequency of amino-acid substitutions in the evolution of the MutS protein family as estimated by the MacClade program (34)). Multiple runs searching for the shortest tree were conducted for each matrix. Distance-based phylogenetic trees were generated by the neighbor-joining (35) and UPGMA algorithms using estimated evolutionary calculated from the matrices described above. Bootstrap resampling was conducted by the method of Felsenstein (36). Character state analysis for the study of gene loss was conducted using the MacClade computer program (34).

RESULTS AND DISCUSSION

The publication in 1995 of the complete genome sequence of the bacterium *Haemophilus influenzae* (37) signaled the beginning of a new era in biological research. Genome sequences provide a wealth of information not only about a single organism but also about all of the genes that they encode. As genome and other sequence data continue to pour into databases at an amazing pace, we need to develop new methods to sort out this information. In developing such methods it is important to recognize that analysis of genomes can benefit from studies of individual gene families and analysis of genome sequences can provide a great deal of information about gene families. For example, many genomes encode dozens or even hundreds of members of some multigene families. Making accurate predictions of the phenotype of these species from the genome sequence requires making accurate predictions of the functions of genes in multigene families. Similarly, a simple analysis of the presence and absence of particular genes in a genome can reveal a great deal about different multigene families. Most methods currently being used to analyze gene and genome data rely on the identification and quantification of similarity between the gene or genome of interest and those of other species. While such methods are useful, they tend to ignore the fact that biological similarities have a historical component (i.e., evolution). It is well documented that the incorporation of an evolutionary perspective can greatly benefit any comparative biological study. The benefits of the evolutionary perspective come from focusing not just on similarities and differences, but on *how* and *why* such similarities and differences arose. Therefore, I believe that studies of genes and genomes can also benefit greatly from an evolutionary focus. I refer to the combined evolutionary study of genes and genomes as phylogenomics (38,39).

I report here a phylogenomic analysis that is focused on the MutS family of proteins. The MutS family is an ideal case study for phylogenomic analysis for a variety of reasons. First, there is a good deal of functional diversity within this gene family. Thus, classifying uncharacterized genes may help improve functional predictions. In addition, this diversity of functions may have major effects on species phenotypes – in particular any phenotype related to mutation rate and pattern. Thus identifying which genes are present in a particular genome may help improve predictions of that species phenotype. Finally, as mentioned in the Introduction, there are many unusual patterns of distribution of MutS homologs in currently available complete genome sequences. I have divided the phylogenomic analysis of the MutS family into multiple sections. In the first few sections, the evolutionary history of the MutS family is inferred by analysis of genes and genomes currently available. In the remaining sections this evolutionary information is used to place some of the studies of the members of this gene family into a useful context and also to make predictions for uncharacterized genes and species.

Identification and alignment of MutS homologs

Multiple sequence searching algorithms were used to identify proteins with extensive amino-acid sequence similarity to the previously characterized members of the MutS family. To increase the likelihood of identifying all available MutS homologs, highly divergent members of the MutS family and a MutS consensus sequence were used as query sequences. In addition, the PSI-blast program was used to identify any proteins with similar motifs to other MutS-like proteins. Proteins were considered to be members of the MutS family if they showed significant sequence similarity to any of the previously identified MutS proteins, and if this similarity extended throughout the protein. All identified complete or nearly complete MutS family members are listed in Table 1.

The sequences of the proteins listed in Table 1 were aligned to each other using the *clustalw* multiple sequence alignment algorithm. This alignment was enhanced both manually and with the *clustalx* program, which allows local *clustalw* alignments to be performed within a larger alignment[11]. The alignment reveals that there are motifs that are highly conserved among all MutS-like proteins. Most of these conserved motifs are confined to one section that is on average about 260 amino-acids in length. This section can be considered the core MutS-family domain. For most of the members of the MutS family, the MutS-family domain is near the C-terminal end of each protein. The alignment of this domain is shown for a representative sample of the proteins in the MutS family in Figure 1. The levels of identity and similarity among the MutS family members ranges from 32% similarity and 18% identity between some distantly related members to

70% similarity and 60% identity between putative orthologs from human and mouse[12]. The level of similarity among all these proteins is much higher than one would expect to occur by convergence, suggesting that all these proteins share a common ancestor and thus should be considered homologs. Although all family members have a MutS-family domain, some sequence patterns were conserved only among subsets of the MutS-like proteins. These motifs may be responsible for providing specific functions to the individual MutS proteins (see below).

Phylogenetic trees of the MutS homologs

Phylogenetic trees of the proteins in the MutS family were determined from the alignment using distance and parsimony methods, each with multiple parameters (see Methods). Since each alignment position is assumed to include residues that share a common ancestry among species, regions of ambiguous alignment were excluded from the phylogenetic analysis. Regions of particularly low sequence conservation were also excluded. In total, 313 amino-acid alignment positions were used[13]. The trees generated with the different methods and parameters were very similar in topology to each other. Therefore only one tree (the neighbor-joining tree) is shown here (Figure 2). Bootstrap analysis revealed that most of the patterns shown in the tree are highly robust (bootstrap values > 70%). Bootstrap values of particular branches are discussed in more detail below and are shown in some of the subsequent Tables and Figures. Overall, the similarity of the trees generated by multiple methods and the high bootstrap values for most branches indicate that most of the patterns shown in Figure 2 are highly robust.

In addition to assessing the internal consistency of the results, it is also useful to compare the results presented here to those of other studies. Unfortunately, many previous studies of the evolution of the MutS family of proteins have not described the methods used to generate the trees and thus are not comparable to this study (e.g., (18)). In addition, some studies have used multiple sequence alignment programs like *clustalw* and *pileup* to generate trees directly and thus cannot be considered reliable phylogenetic studies (e.g., (40,41)). There have been only two studies of the evolution of MutS

[11] This complete alignment is available at http://www-leland.stanford.edu/~jeisen/MutS/MutS.html

[12] A matrix with pairwise similarities and identities is available at the MutS web site described above.

homologs using standard phylogenetic methods (21,42). These studies should be considered limited because they did not include many of the more divergent members of the MutS family. Nevertheless, most of the results of these studies are similar to those reported here. Some specific differences and similarities are discussed below.

Beyond gene trees: identifying evolutionary events in the MutS family's history

As with any gene family, the phylogenetic tree of the MutS proteins simply shows the relationships among homologs. It is almost always useful to go beyond this gene tree to identify specific evolutionary events in a gene family's history. For example, identification of the types of homology (orthology, paralogy, and xenology) in this tree allows the detection of the particular evolutionary event (speciation, gene duplication, and lateral gene transfer, respectively) that led to the divergence of homologs. To identify these and other evolutionary events, it is necessary to integrate the gene tree with other information, such as gene function, species phenotype, or species phylogeny.

Subfamilies of orthologs

As the first step in going beyond the MutS gene tree, I divided the MutS family into subfamilies that I propose represent distinct groups of orthologs (i.e., sets of genes that diverged from each other due to speciation events). Each subfamily has been given a name based on the name of one of the better-studied proteins in that group (*italics* are used to distinguish the subfamilies from individual proteins). The proposed subfamilies are highlighted in Fig. 2b-d and the proteins in each subfamily are listed in Table 1. Some characteristics of each subfamily are given in Table2. The assertion that these subfamilies are distinct evolutionary groups is supported by five lines of evidence: (1) each was found in trees generated by all the phylogenetic methods used; (2) each has reasonably high bootstrap values with different methods (Table 2); (3) the branches leading up to the subfamilies are relatively long indicating that each is evolutionarily distinct from other subfamilies; (4) protein size is somewhat conserved within subfamilies (see Table 1); and (5) there are sequence motifs conserved within but not between subfamilies (not shown). The assertion that these evolutionarily distinct

[13] Available at the MutS web site.

subfamilies are distinct orthologous groups is supported by two factors: (1) the phylogenetic relationships of proteins within each group are roughly congruent to the likely relationships of the species from which they come; and (2) function has been conserved within subfamilies.

Overall, eight orthologous subfamilies were identified – six that include only proteins from eukaryotes (corresponding to the six yeast MutS homologs) and two that include only proteins from bacteria. Most of these subfamilies correspond well to groups that have been suggested previously. For example, the animal and yeast proteins in each eukaryotic subfamily have been identified as likely orthologs of each other by standard sequence similarity searches and other non-phylogenetic methods. The phylogenetic analysis simply confirms that these are indeed orthologs. The identification of two distinct bacterial subfamilies represents a novel finding (although it was suggested in (38)). This finding shows one of the benefits of phylogenetic analysis over standard sequence-similarity searches. In addition to the subfamilies, two proteins (one from *M. thermoautotrophicum* and one from the mitochondrial genome of *S. glaucum*) are closely related to the *MutS2* subfamily but they were not placed into this subfamily. Although these two genes group with the *MutS2* subfamily in every tree, it is possible that they may have been involved in lateral transfer events and therefore may not be orthologs of the *MutS2* proteins. Nevertheless, they are close relatives of the *MutS2* subfamily.

Examination of the species represented in each orthologous group can help determine when that group originated. For example, all the eukaryotic subfamilies except *MSH1* include proteins from yeast and humans suggesting that these subfamilies originated prior to the divergence of the common ancestor of fungi and animals. Similarly, the *MutS1* and *MutS2* subfamilies are composed of proteins from diverse bacterial species including some of the deeper branching bacterial taxa (e.g., *D. radiodurans* and *A. aeolicus*). Therefore the origin of these bacterial subfamilies probably predates the divergence of most of the bacterial phyla. While this type of analysis can help time the origin of the orthologous groups, it does not provide any information about *how* these groups originated. That is, did the orthologous groups originate by gene duplication or lateral transfer? Many other questions also cannot be answered by the simple division into groups of orthologs. Therefore additional analysis is required.

Unusual distributions of MutS orthologs help identify specific evolutionary events

One way to identify particular evolutionary events in the history of a gene family is to analyze unusual distribution patterns of the different orthologs. Such unusual distributions can be explained either by lateral transfer to the species with an "unexpected" presence of a gene, or by gene loss in the lineages with an unexpected absence of certain genes. These two possibilities can be distinguished by comparing the gene tree to the tree of the species from which these genes come. If an unusual distribution is caused by gene loss, then the gene and species trees should be congruent (as though the species which do not encode a particular gene were just cut out of a larger tree of life). If instead lateral transfer caused an unusual distribution, then the gene and species trees should be incongruent.

Analysis of the distribution of proteins used to be relatively haphazard. However, the availability of complete genome sequences allows for the first time the reliable determination (through sequence analysis) of what genes are present or absent in a species. This of course assumes that homologs can be detected by the sequence analysis methods used. Given the level of conservation among a diverse collection of MutS homologs (see Fig. 1), it is likely that most MutS homologs were identified using the search methods described here. A simple identification of homologs in a species does not provide a complete picture of gene presence and absence. It is important to determine presence and absence of specific orthologs. This step is another area in which phylogenetic analysis and genome analysis can be combined. Although other methods have been developed to determine orthology, phylogenetic methods are preferable (39). Thus, using a combination of sequence searches and phylogenetic analysis, the presence and absence of particular orthologs was determined for all species for which complete genomes are available (Table 3).

Since most of the available complete genome sequences are from bacteria, I focused first on distribution patterns in the bacteria. Every possible pattern of presence and absence of the MutS1 and MutS2 proteins is found in the bacteria (Table 3) - some species encode members of both subfamilies, while others encode only one or none.

There are two reasonable explanations for this: either rampant gene loss after gene duplication or multiple lateral transfer events. As discussed above, one way of testing which occurred is to compare the phylogenetic trees of the two subfamilies. If there was an ancient duplication, then the branching patterns within the *MutS1* and *MutS2* subfamilies should be identical. However, it is not valid to simply extract the MutS1 and MutS2 evolutionary relationships from the gene tree shown in Figure 2. This is because the MutS1 and MutS2 genes in this tree do not all come from the same species and species sampling can have a major effect on phylogenetic results (43). To get around this species sampling effect, I generated new trees using only proteins from species that encode both MutS1 and MutS2 (Figure 3a). As can be seen, the branching patterns in the two subfamilies are congruent when these identical species sets are used. It is important to note that this shared topology is not congruent to that of the rRNA tree of life. The reasons for this are not known but it may simply be due to the limited number of MutS sequences that are available. Regardless, the fact that the branching patterns of the two subfamilies are congruent indicates that a gene duplication gave rise to these two subfamilies. Thus the absence of *MutS1* and *MutS2* orthologs from some species is most likely caused by gene loss. I inferred likely gene loss events within the *MutS1* and *MutS2* subfamilies by using standard parsimony character state reconstruction (Fig. 3b). The identification of specific gene loss events relies on the accuracy of the species tree onto which the presence and absence of genes in overlaid. The choice of the particular species tree to use is somewhat difficult, since some results suggest that bacterial “species” do not have a single tree. However, in this case, the choice of the specific tree is not particularly important since all of the inferred gene loss events are in lineages with well-established phylogenies. For example, the inference of gene loss in the mycoplasmas essentially only depends on the well-supported assumption that mycoplasmas are members of the lowGC gram-positive group (since other lowGC gram-positives encode both *MutS1* and *MutS2* orthologs). Thus although the species tree used may not be accurate, the inferred gene loss events are likely correct. The implications of specific gene loss events are discussed in more detail below.

The evidence presented above shows that the *MutS1* and *MutS2* subfamilies are most likely related by a gene duplication event. However, the evidence does not specify when this duplication occurred. Based on a variety of evidence, I propose that the duplication was ancient and that the root of the MutS tree is most accurately placed such that it divides the family into two main lineages which I refer to as *MutS-I* and *MutS-II*. *MutS-I* includes the *MutS1*, *MSH1*, *MSH2*, *MSH3*, and *MSH6* subfamilies and *MutS-II* includes the *MutS2*, *MSH4*, and *MSH5* subfamilies. Three pieces of information support the division into these two main lineages: (1) these two groups were found in all trees regardless of methods or parameters used; (2) function is generally conserved within but not between lineages - the proteins involved in MMR are all in the MutS-I lineage and those involved in meiotic crossing-over are in the MutS-II lineage (Table 1); (3) such an ancient duplication is consistent with the presence of bacterial and eukaryotic subfamilies in each lineage and is also consistent with the evidence for a duplication prior to the emergence of the major bacterial groups. Since these arguments are somewhat circumstantial and, since the bootstrap values defining the two supergroups are relatively low, this hypothesis should be considered highly tentative. A consensus tree, using the proposed rooting but in which those patterns that are not robust are collapsed, is shown in Figure 4. Even assuming the duplication occurred as proposed, since the relationships among the subfamilies within each lineage are not well resolved in the current analysis, it is not possible to determine the exact patterns of duplications or lateral transfers within each lineage. It is likely that as the sequences of additional members of each subfamily become available the relationships between the subfamilies will become better resolved.

The ancient duplication theory proposed above does not describe all of the unusual distribution patterns in the MutS family. One such pattern is the presence of only one MutS homolog among the three Archaea for which complete genomes are available. This is the MutS2-like protein of *M. thermoautotrophicum*. As discussed above, since the MutS proteins are highly conserved (including the one MutS homolog from Archaea) it is unlikely that other MutS homologs are present in these Archaeal species but were not identified. With the data currently available, it is not possible to resolve the origins of this gene. One reason for this is the lack of a consensus concerning the evolutionary history of the major domains of life. If the Archaea are a sister group to the eukaryotes (as suggested by some studies), then the distribution pattern is probably best explained by gene loss in the history of these Archaea. If instead the bacteria and eukaryotes are sister groups (or even just for the parts of the genome encoding the MutS proteins), then the MutS gene family may have evolved after the Archaea formed a separate lineage. Thus the distribution pattern could be explained simply by lateral transfer to *M. thermoautotrophicum*. Another reason for difficulty resolving this unusual distribution pattern is that these three species do not represent much of the Archaeal evolutionary diversity. It is likely that additional Archaeal genomes will help resolve the history of the Archaeal MutS homolog(s).

Another unusual distribution pattern is the presence of a MutS homolog (sgMutS) in the mitochondrial genome of the coral *S. glaucum*. Although this mitochondrial genome is not completely sequenced, many other mitochondrial genomes have been and none of these encodes a MutS homolog. In a detailed phylogenetic study, Pont-Kingdon et al. found that the sgMutS branched most closely to the yeast MSH1 (42). Since MSH1 is encoded by the nucleus but functions in the mitochondria, this seemed like a possible case of lateral transfer from the mitochondria to the nucleus. However, since the sgMutS did not branch within any bacterial group of proteins and since most mitochondria do not encode a MutS homolog, they concluded that the sgMutS represented a case of “reverse” lateral transfer from the nucleus to the mitochondria. Although their analysis was sound, it was not complete because they did not include proteins from all of the MutS subfamilies. With the more complete sample of MutS homologs, the sgMutS branches closely to the *MutS2* subfamily and not with the *MSH1* subfamily (Fig. 2). This branching pattern is robust – it was seen in the trees generated by all methods used and it has high bootstrap values. I further tested the robustness of this branch pattern by determining the parsimony score for trees with a variety of lateral transfer scenarios involving the sgMutS and *MSH1* proteins including (1) a mitochondrial origin of the *MSH1* subfamily (2) a mitochondrial origin of the sgMutS and (3) a *MSH1* origin of the sgMutS (as suggested by Pont-Kingdon et al.). Each of these scenarios requires many more steps than the tree in which sgMutS grouped with the *MutS2* subfamily. Thus the results of Pont-Kingdon et al. were probably biased by not including proteins from all of the MutS subfamilies. There are two reasonable explanations for the close relationship of the sgMutS to the MutS2 family. It is possible that there was a lateral transfer of a MutS2-like gene to the mitochondria of an ancestor of *S. glaucum*. Alternatively, the sgMutS may be a true mitochondrial gene and *S. glaucum* may be one of the few species in which this gene still remains. The ability to resolve the origins of the sgMutS will likely improve with the inclusion of more members of the *MSH1* subfamily and sequences from alpha-Proteobacterial species which are considered to be the closest living relatives to mitochondria.

Using the evolutionary information

The benefits of using evolutionary analysis in molecular biology come from improving both our understanding of observed molecular characteristics and our ability to make biological useful predictions. What are the particular uses of the evolutionary analysis of the MutS family described above? First, I used the phylogenetic information to infer likely functions for uncharacterized members of the MutS family (Figure 1b-d). Such a phylogenomic prediction of function is preferable to similarity-based functional predictions for a variety of reasons (see (39) for review). In summary, since function is conserved within orthologous subfamilies, I have assigned predicted functions to uncharacterized genes based on the subfamily in which they are placed. This ortholog rule cannot be applied to those proteins in the *MutS2* subfamily since none of the proteins in this subfamily have a known function. In addition, it cannot be applied to the two MutS2-like proteins since they may not be orthologs of any of the MutS family members. Interestingly, many of the proteins in the *MutS2* subfamily (as well as the two MutS2-like proteins) have been given the name MutS and assigned a likely role in MMR based predominantly on similarity searches (see (38)). The phylogenetic analysis suggests that these functional assignments are likely to be wrong. First, these proteins are all evolutionarily distant from proteins known to be involved in mismatch repair. In addition, many of these proteins are found in species that do not even encode a MutL homolog (e.g., *H. pylori* (24) and *M. thermoautotrophicum* (44)) and a functional MutL homolog is required for MMR. It is much more reasonable to assign these proteins a possible function in chromosome segregation or crossing-over since they are in the *MutS-II* lineage with proteins in the *MSH4* and *MSH5* subfamilies. Thus the phylogenetic analysis helps suggest what the functions of the genes in the MutS2 subfamily may be and analysis of additional genome data (the presence and absence of MutL homologs) aids in the prediction of function.

The phylogenetic-functional analysis suggest not only that functions have been conserved within orthologous groups but also that the generation of the orthologous groups was accompanied by functional divergence. The evolutionary analysis on its own does not provide a complete explanation of the functions of the MutS genes. There must be some sequence patterns that explain the functional similarities and differences in the family. Since the MutS-family domain is highly conserved among all the MutS-like proteins, this domain likely provides some general activity to all the proteins in the family such as the ability to recognize and bind to unusual double-stranded DNA structures. In addition, there must be some sequence patterns that are conserved within but not between subfamilies (either in these proteins or in regulatory regions) that provide specific functions to each subfamily. The phylogenetic analysis can help identify functionally important motifs because they can be searched for only within subfamilies (45). Thus the phylogenetic analysis can help understand the mechanism of the specificity of each subfamily.

The phylogenetic-functional analysis can be used in combination with gene presence and absence data to predict organismal phenotypes for those species for which complete genomes are available. For example, it is likely that the species that do not encode a protein in the *MutS-I* lineage do not have the MMR process as it has been found in other species. Such an inference is supported by the fact that all species that do not encode a protein in the *MutS-I* lineage also do not encode a MutL homolog (see above and (38)). Such a conclusion is supported by the fact that some of the species that do not encode a MutS1 also have a high mutation rate (e.g., the mycoplasmas) which is consistent with an absence of MMR. However, since it is possible that other enzymatic mechanisms could have evolved to deal with mismatches, without experimental verification it is not possible to know for certain if these species have MMR. Since no function is known for the proteins in the *MutS2* subfamily it is difficult to determine the significance of the absence of orthologs of these genes from species like *E. coli* and *H. influenzae*.

Combining functional predictions for genes with the gene loss analysis allows a better understanding of why the loss of these genes occurred. The gene loss data shows that losses of *MutS1* and *MutS2* occurred in multiple lineages. Many theories have been put forward to explain gene loss during evolution (46,47). Many of these theories involve genome level phenomena such as selection for reduced genome size, or Muller's ratchet destroying some genes. However, the loss of MutS homologs may be a more gene-specific event - there is likely a selective benefit for the loss of MutS genes in some lineages. Defects in MMR have been suggested to be beneficial in certain conditions such as under nutrient stress (48) and selection for pathogenesis (49,50). It is likely that many of these benefits are due to an increased mutation rate, although some may also be due to changes in other functions associated with MMR proteins. While these benefits have been shown by comparing different strains of the same species, it is possible that such benefits may also occur in comparisons between species. For example, it has been suggested that *H. pylori* varies its antigens through a microsatellite mutation process (24). Such mutations would occur at a much higher rate in a MMR deficient strain and could explain the loss of *MutS1* from *H. pylori* sometime in the past.

Conclusions

I have used a combination of phylogenetic reconstruction methods and analysis of complete genome sequences to better understand the MutS family of proteins. Since studies of multigene families and genomes are interdependent it is useful to combine analysis into one study. Phylogenomic methodology similar to that used here can be applied to any multigene family. First, molecular phylogenetic analysis should be used to determine the evolutionary relationships among the genes in the gene family. Then, integration of species information can be used to divide the family into subfamilies of orthologs and to infer evolutionary events such as gene duplications, lateral transfers and gene loss. This evolutionary information can be used in combination with genome information to improve functional predictions for uncharacterized genes. For example, the phylogenetic analysis shows that the proteins in the *MutS2* subfamily are distant and distinct from those involved in mismatch repair and genome analysis shows that many of the species that encode these genes do not encode other proteins required for mismatch repair. Thus these proteins are likely not involved in mismatch repair. The phylogenomic analysis can also be used to characterize functionally important sequence motifs, to predict the phenotypes of species for which complete genomes are available and to better understand why events such as gene loss and gene duplication may have occurred. In summary, since any comparative biological analysis benefits from evolutionary perspective, the use of evolutionary methods can only serve to improve what can be learned from ever increasing amounts of gene and genome data.

ACKNOWLEDGEMENTS

I would like to thank P. Hanawalt for his continued support; J. Halliday, R. Myers, and M-I. Benito for helpful discussions; F. Taddei, D. Berg, K. Culligan, and G. Meyer-Gauen for suggestions on earlier versions of this work; D. Swofford for making PAUP* available prior to release and The Institute for Genomic Research and Oklahoma University for making sequences available prior to publication. The comments of two anonymous reviewers were particularly helpful. I was supported by Outstanding Investigator Grant CA44349 from the National Cancer Institute to P. Hanawalt. I apologize to the authors of primary papers on the various MutS homologs for being unable to cite them all here. References relating to each can be found at the MutS web site: http://www-leland.stanford.edu/~jeisen/MutS/MutS.html.

REFERENCES

1. Modrich, P. and Lahue, R. (1996) *Ann. Rev. Biochem.*, **65,** 101-133.
2. Kolodner, R. D. (1995) *Trends Bioch. Sci.*, **20,** 397-401.
3. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. and Inouye, M. (1966) *Cld. Spr. Hrbr. Symp. Q. Biol.*, **31,** 77-84.
4. Sia, E. A., Jinks-Robertson, S. and Petes, T. D. (1997) *Mutat. Res.*, **383,** 61-70.
5. Levinson, G. and Gutman, G. A. (1987) *Nucl. Acids Res.*, **15,** 5323-5338.
6. Modrich, P. (1991) *Ann. Rev. Genet.*, **25,** 229-253.
7. Parker, B. O. and Marinus, M. G. (1992) *Proc. Natl. Acad. Sci. U. S. A.*, **89,** 1730-1734.
8. Lynch, H. T., Smyrk, T. and Lynch, J. (1997) *Canc. Genet. Cytogenet.*, **93,** 84-99.

9. Kolodner, R. (1996) *Genes Dev.*, **10,** 1433-1442.

10. Marsischky, G. T., Filosi, N., Kane, M. F. and Kolodner, R. (1996) *Genes Dev.*, **10,** 407-420.

11. Chi, N. W. and Kolodner, R. D. (1994) *J. Biol. Chem.*, **269,** 29984-29992.

12. Reenan, R. A. and Kolodner, R. D. (1992) *Genetics*, **132,** 963-973.

13. Reenan, R. A. and Kolodner, R. D. (1992) *Genetics*, **132,** 975-985.

14. Hollingsworth, N. M., Ponte, L. and Halsey, C. (1995) *Genes Dev.*, **9,** 1728-1739.

15. Pochart, P., Woltering, D. and Hollingsworth, N. M. (1997) *J. Biol. Chem.*, **272,** 30345-30349.

16. Ross-Macdonald, P. and Roeder, G. S. (1994) *Cell*, **79,** 1069-1080.

17. Matic, I., Taddei, F. and Radman, M. (1996) *Trends Microbiol.*, **4,** 69-73.

18. Fishel, R. and Wilson, T. (1997) *Curr. Opin. Genet. Dvlp.*, **7,** 105-113.

19. Modrich, P. (1997) *J. Biol. Chem.*, **272,** 24727-24730.

20. Villanuve, A., personal communication.

21. Culligan, K. M. and Hays, J. B. (1997) *Plant Physiol.*, **115,** 833-839.

22. Kunst, A., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., Bertero, M., Bessieres, P., Bolotin, A., Borchert, S., et al. (1997) *Nature*, **390,** 249-256.

23. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. (1996) *DNA Res.*, **3,** 109-136.

24. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., et al. (1997) *Nature*, **388,** 539-547.

25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215,** 403-410.

26. Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) *Nucl.Acids Res.*, **25,** 3389-3402.

27. The Institute for Genomic Research, personal communication.

28. Roe, B. A., Clifton, S., and Dyer, D. W., personal communication.

29. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucl. Acids Res.*, **22,** 4673-4680.

30. Thompson, J. and Jeanmougin, F. (1997). *Clustal X* Strassborg University, Strassborg.

31. Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W. and Gillevet, P. M. (1994) *CABIOS*, **10,** 671-675.

32. Eisen, J. A. (1997) In Swindell, S. R. (ed.), Methods In Molecular Biology, Vol. 70. Sequence Data Analysis Guidebook. Humana Press Inc., Totowa, New Jersey, USA., Vol. 70, pp. 13-38.

33. Swofford, D. (1991). *Phylogenetic Analysis Using Parsimony (PAUP) 3.0d*. Illinois Natural History Survey, Champaign, Ill.

34. Maddison, W. P. and Maddison, D. R. (1992). *MacClade 3*. Sinauer Associates, Inc., Sunderland, MA.

35. Saitou, N. and Nei, M. (1987) *Mol. Biol. Evol.*, **4,** 406-425.

36. Felsenstein, J. (1985) *Evolution*, **39,** 783-791.

37. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., et al. (1995) *Science*, **269,** 496-498, 507-512.

38. Eisen, J. A., Kaiser, D. and Myers, R. M. (1997) *Nature (Med.)*, **3,** 1076-1078.

39. Eisen, J. A. (1998) *Genome Res.*, **8,** 163-167.

40. Pont-Kingdon, G. A., Okada, N. A., Macfarlane, J. L., Beagley, C. T., Wolstenholme, D. R., Cavalier-Smith, T. and Clark-Walker, G. D. (1995) *Nature*, **375,** 109-11.

41. Paquis-Flucklinger, V., Santucci-Darmanin, S., Paul, R., Saunieres, A., Turc-Carel, C. and Desnuelle, C. (1997) *Genomics*, **44,** 188-194.

42. Pont-Kingdon, G. A., Okada, N. A., Macfarlane, J. L., Beagley, C. T., Watkins-Sims, C. D., Cavalier-Smith, T., Clark-Walker, G. D. and Wolstenholme, D. R. (1998) *J. Mol. Evol.*, **46,** 419-431.

43. Eisen, J. A. (1995) *J. Mol. Evol.*, **41,** 1105-1123.

44. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., et al. (1996) *J. Bacteriol.*, **179,** 7135-7155.

45. Eisen, J. A., Sweder, K. S. and Hanawalt, P. C. (1995) *Nucl. Acids Res.*, **23,** 2715-2723.

46. Maniloff, J. (1996) *Proc. Natl. Acad. Sci. U. S. A.*, **93,** 10004-10006.

47. Moran, N. A. (1996) *Proc. Natl. Acad. Sci. U. S. A.*, **93,** 2873-2878.

48. Harris, R. S., Feng, G., Ross, K. J., Sidhu, R., Thulin, C., Longerich, S., Szigety, S. K., Winkler, M. E. and Rosenberg, S. M. (1997) *Genes Dev.*, **11,** 2426-2437.

49. LeClerc, J. E., Li, B., Payne, W. L. and Cebula, T. A. (1996) *Science*, **274,** 1208-1211.

50. Sniegowski, P. D., Gerrish, P. J. and Lenski, R. E. (1997) *Nature*, **387,** 703-705.

Table 1. Proteins in the MutS Family[1]

MutS Lineage				
Subfamily	Gene	Accession	Predicted	Experimentally Determined
Species	Name[2]	(gi)	Size (aa)	Function(s)
***MutS-I* Lineage**				
Bacteria				
*MutS1*Subfamily				
Escherichia coli	MutS	127556	853	Mismatch repair (all)
Salmonella typhimurium	MutS	1171081	861	Mismatch repair (all)
Haemophilus influenzae	MutS	417330	854	*
Azotobacter vinelandii	MutS	127555	855	Mismatch repair (all)[3]
Neisseria gonorrhoeae	MutS	*	*	*
Synechocystis sp.	MutS	1652903	912	*
Treponema pallidum	MutS	*	*	*
Borrelia burgdorferi	MutS	2688751	862	*
Streptococcus pneumoniae	HexA	123080	844	All mismatch repair
Streptococcus pyogenes	MutS	*	*	*
Bacillus subtilis	MutS	1709189	852	Mismatch repair (all)
Thermus thermophilus	MutS	1871501	819	Mismatch recognition in vitro
Thermus aquaticus	MutS	1203807	811	Mismatch recognition in vitro
Deinococcus radiodurans	MutS	*	*	*
Thermotoga maritima	MutS	1619909	793	*
Aquifex aeolicus	MutS	2983001	859	
Aquifex pyrophilus	MutS	1619907	855	*
Chlamydia trachomatis	MutS	*	*	*
Eukaryotes				
MSH2 Subfamily				
Human	MSH2	1171032	934	Mismatch repair (all)
Rat	MSH2	1709122	933	*
Mouse	MSH2	726086	935	Mismatch repair (all)
Xenopus leavis	MSH2	1079288	933	*
Drosophila melanogaster	SPE1	1174416	913	*
Yeast	MSH2	172002	964	Mismatch repair (all)
Neurospora crassa	MSH2	2606088	937	*
Arabidopsis thaliana	atMSH2	2522362	937	*
MSH3 Subfamily				
Human	hMSH3	1490521	1128	Mismatch repair (loops)[4]
Mouse	Rep3	400971	1091	*
Arabidopsis thaliana	MSH3	2980796	1076	
Saccharomyces cerevisiae	MSH3	127089	1047	Mismatch repair (loops)
S.pombe	Swi4	135075	993	Mismatch repair (loops?)[5]
MSH6 Subfmaily				
Human	GTBP	1082386	1292	Mismatch repair (base:base)[6]
Mouse	GTBP	2506881	1358	*
Saccharomyces cerevisiae	MSH6	1588283	1242	Mismatch repair (base:base)
Arabidopsis thaliana	MSH6	2104531	1362	*
MSH1 Subfamily				
Saccharomyces cerevisiae	MSH1	730065	959	Mismatch repair in mtDNA?
S. pombe	MSH1	2330782	780?	*

MutS Lineage Subfamily *Species*	Gene Name[7]	Accession (gi)	Predicted Size (aa)	Experimentally Determined Function(s)
***MutS-II* Lineage**				
Bacteria/Archaea/Mitochondria				
MutS2 Subfamily[8]				
Helicobacter pylori	MutS2[9]	2313742	762	*
Bacillus subtilis	MutS2	2635323	785	*
Streptococcus pyogenes	MutS2	*	*	*
Borrelia burgdorferi	MutS2	2687977	780	*
Synechocystis sp.	MutS2	1652751	822	*
Aquifex aeolicus	MutS2	2983682	762	
Deinococcus radiodurans	MutS2	*	*	*
MutS2-like				
Met. thermoautotrophicum	MutS2	2622891	647	*
Sarcophyton glaucum mt	sgMutS	2147739	982	*
Eukaryotes				
MSH4 Subfamily				
Saccharomyces cerevisiae	MSH4	1078105	878	Meiotic cross-over, segregation
Human	hMSH4	2463653	936	*
C. elegans	MSH4	1330382	688?	Meiotic cross-over
MSH5 Subfamily				
Saccharomyces cerevisiae	MSH5	2497997	901	Meiotic cross-over, segregation
Human	hMSH5	2653649	834	*
C. elegans	MSH5	1340008	1139	Meiotic cross-over

[1] Only complete or nearly complete proteins are included. Additional information about each protein can be found in Genbank and at http://www-leland.stanford.edu/~jeisen/MutS/MutS.html.
[2] Unnamed open reading frames are given a proposed name which is underlined.
[3] Determined by increased mutation rate in lines with defects in this gene.
[4] Genetic and biochemical studies suggest the MSH3 proteins are only involved in repair of large loops.
[5] Mutants show an increased rate of small duplications consistent with a possible role in loop repair.
[6] Genetic and biochemical studies suggest that MSH6 proteins are only involved in the repair of base:base mismatches and small loops.
[7] Unnamed open reading frames are given a proposed name which is underlined.
[8] The last two of these may not be true orthologs of the others (see Discussion).
[9] I suggest changing the names of the sequences in this groups to MutS2 to reflect their distinctness from the proteins in the MutS1 subgroup.
*Information not available.

Table 2. Properties of MutS subfamilies.

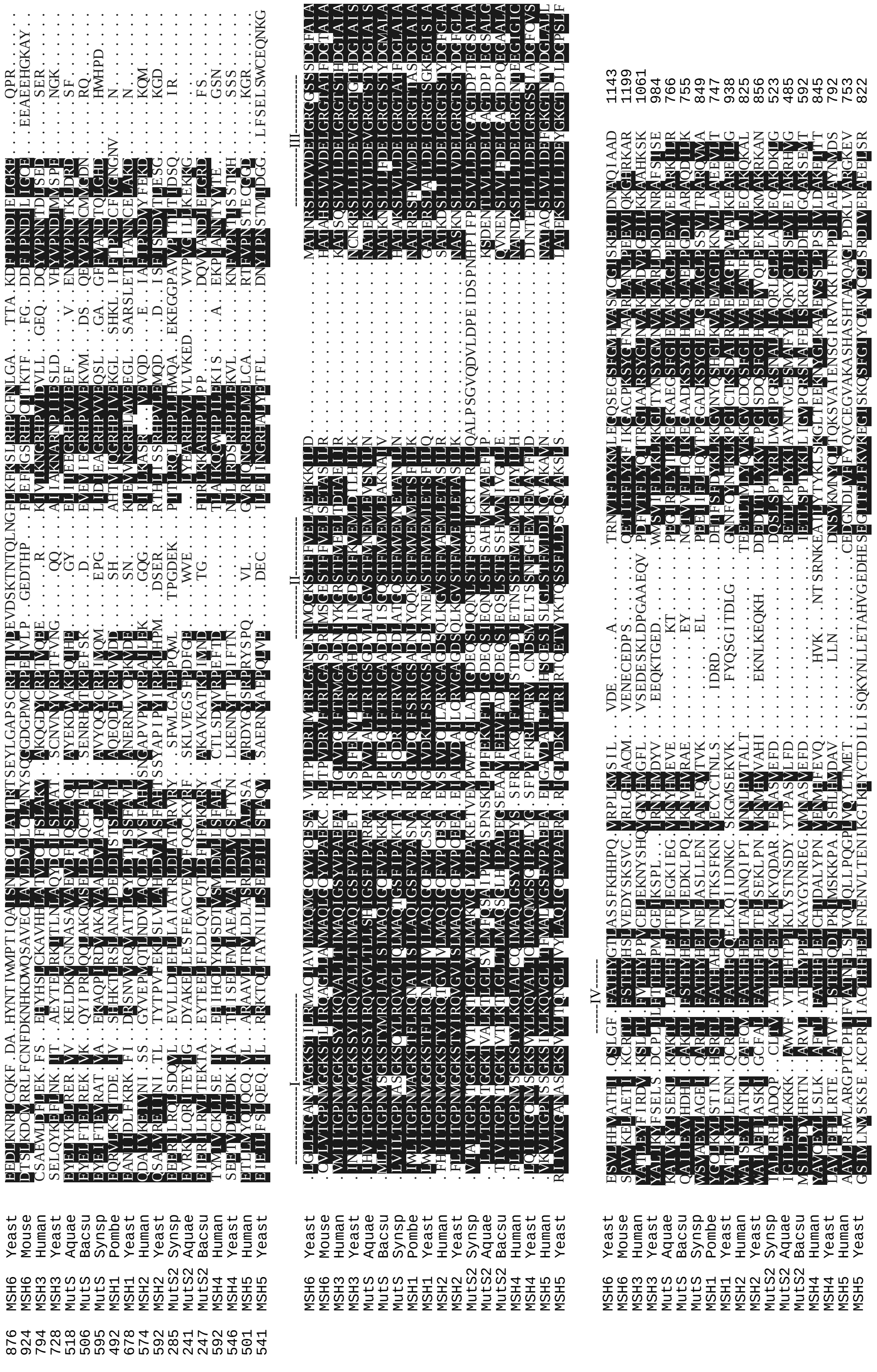
Subfamily	Conserved Function	Comments	Boostrap value		
			NJ	UPG	Pars.
MutS-I	**Mismatch Repair**				
MutS1	All mismatch repair	In most bacteria.	96	100	25
MSH1	Mitochondrial mismatch repair?	Eukaryotic, not yet found in humans.	100	100	95
MSH2	All mismatch repair in nucleus.	Eukaryotic. Defective in some HNPCC.	100	100	100
MSH3	Repair of loops (small & large) in nucleus.	Eukaryotic. Defective in some HNPCC.	79	100	100
MSH6	Repair of mismatched base pairs & small loops in nucleus.	Eukaryotic. Defective in some HNPCC.	95	100	90
MutS-II	**Chromosome Segregation?**				
MutS2	Unknown.	In some bacteria.	74	95	60
MSH4	Facilitate X-over, chromosome segregation.	Eukaryotic. Role in humans unknown.	96	97	85
MSH5	Facilitate X-over, chromosome segregation.	Eukaryotic. Role in humans unknown.	100	100	55

Table 3. Presence of MutS and MutL homologs in complete genomes sequences.

Species	# of MutS Homologs	Which Subfamilies?	# of MutL Homologs
Bacteria			
Escherichia coli K12	1	*MutS1*	1
Haemophilus influenzae Rd KW20	1	*MutS1*	1
Neisseria gonorrhoeae	1	*MutS1*	1
Helicobacter pylori 26695	1	*MutS2*	-
Mycoplasma genitalium G-37	-	-	-
Mycoplasma pneumoniae M129	-	-	-
Bacillus subtilis 169	2	*MutS1,MutS2*	1
Streptococcus pyogenes	2	*MutS1,MutS2*	1
Mycobacterium tuberculosis	-	-	-
Synechocystis sp. PCC6803	2	*MutS1,MutS2*	1
Treponema pallidum Nichols	1	*MutS1*	1
Borrelia burgdorferi B31	2	*MutS1,MutS2*	1
Aquifex aeolicus	2	*MutS1,MutS2*	1
Deinococcus radiodurans R1	2	*MutS1,MutS2*	1
Archaea			
Archaeoglobus fulgidus VC-16, DSM4304	-	-	-
Methanococcus janasscii DSM 2661	-	-	-
Methanobacterium thermoautotrophicum ΔH	1	*MutS2*	-
Eukaryotes			
Saccharomyces cerevisiae	6	*MSH1-6*	3+
Homo sapiens	5	*MSH2-6*	3+

Figure 1. Alignment of a conserved region of the MutS proteins from representative members of the MutS family.

The alignment was generated using the *clustalw* and *clustalx* programs and modified slightly manually. Shading was done based on degree of identity or conservation using the MacBoxshade program. Previously described MutS motifs are referred to by Roman numerals. The beginning and ending amino-acids for each protein are numbered.



MSH6 Yeast
MSH6 Mouse
MSH3 Human
MSH3 Yeast
MutS Aquae
MutS Bacsu
MutS Synsp
MSH1 Pombe
MSH1 Yeast
MSH2 Human
MSH2 Yeast
MutS2 Synsp
MutS2 Aquae
MutS2 Bacsu
MSH4 Human
MSH4 Yeast
MSH5 Human
MSH5 Yeast
I
II
III
IV
1143
1199
1061
984
766
755
849
747
938
825
856
523
485
592
845
792
753
822

Figure 2. Phylogenomic analysis of the MutS family of proteins.

A. Unrooted neighbor-joining tree of the proteins in the MutS family. The tree was generated from a *clustalw* based sequence alignment (with regions of ambiguous alignment excluded) with the PAUP* program. Some of the bacterial MutS1 proteins are left out for clarity. B. Proposed subfamilies of orthologs are highlighted (see Discussion for details). C. Known functions of genes are overlaid onto the tree. For simplicity's sake, only two colors are used, red for mismatch repair and blue for meiotic-crossing over and chromosome segregation. D. Prediction of functions of uncharacterized proteins based on position in the tree.

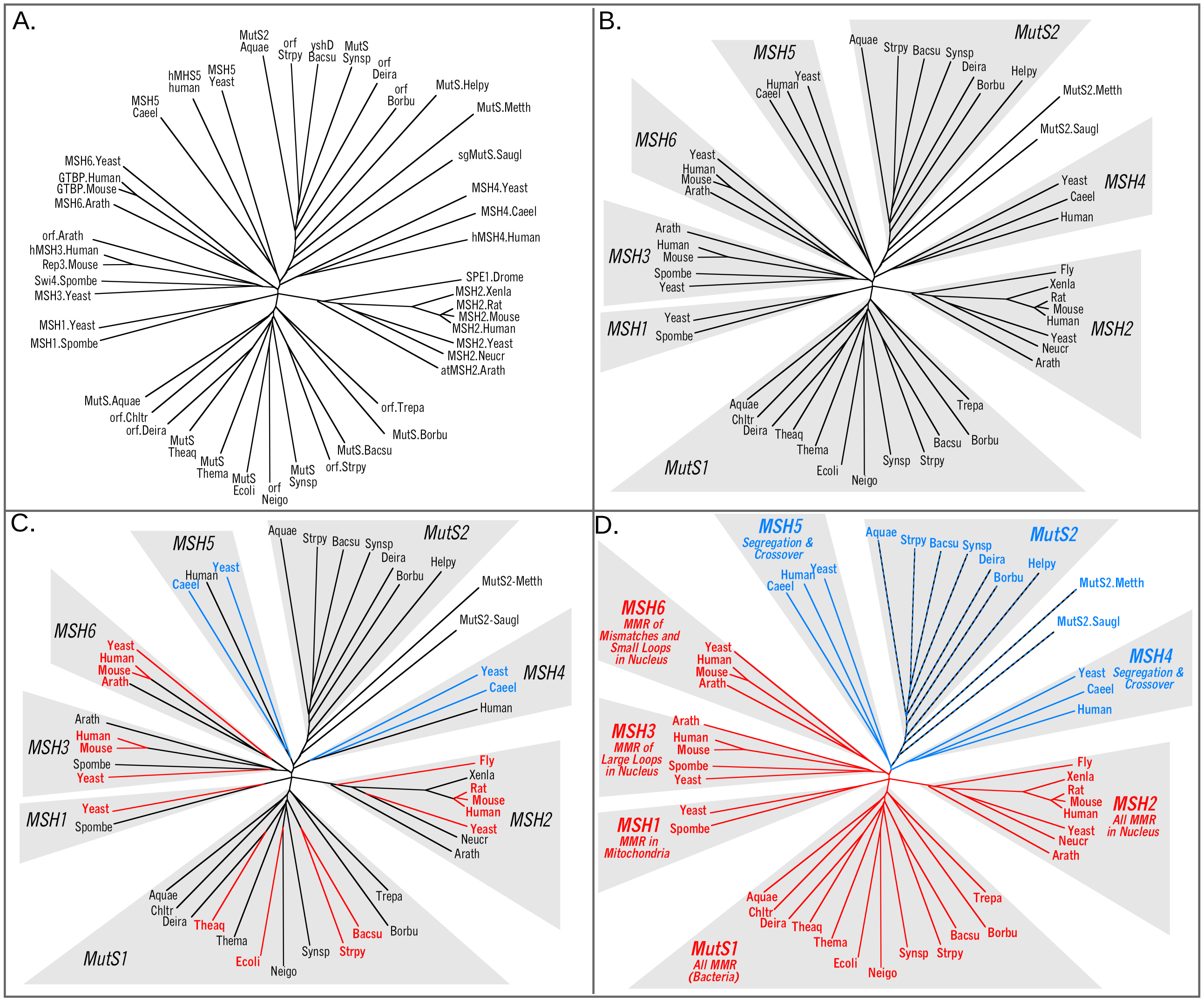

A.
MutS2 Aquae
orf Strpy
yshD Bacsu
MutS Synsp
orf Deira
orf Borbu
MutS.Helpy
MutS.Metth
sgMutS.Saugl
MSH4.Yeast
MSH4.Caeel
hMSH4.Human
SPE1.Drome
MSH2.Xenla
MSH2.Rat
MSH2.Mouse
MSH2.Human
MSH2.Yeast
MSH2.Neucr
atMSH2.Arath
orf.Trepa
MutS.Borbu
MutS.Bacsu
orf.Strpy
MutS Synsp
orf Neigo
MutS Ecoli
MutS Thema
MutS Theaq
orf.Deira
orf.Chltr
MutS.Aquae
MSH1.Yeast
MSH1.Spombe
orf.Arath
hMSH3.Human
Rep3.Mouse
Swi4.Spombe
MSH3.Yeast
MSH6.Yeast
GTBP.Human
GTBP.Mouse
MSH6.Arath
MSH5 Caeel
hMHS5 human
MSH5 Yeast
B.
MSH5
Human
Yeast
Caeel
MutS2
Aquae
Strpy
Bacsu
Synsp
Deira
Borbu
Helpy
MutS2.Metth
MutS2.Saugl
MSH4
Yeast
Caeel
Human
MSH6
Yeast
Human
Mouse
Arath
MSH3
Arath
Human
Mouse
Spombe
Yeast
MSH1
Yeast
Spombe
MSH2
Fly
Xenla
Rat
Mouse
Human
Yeast
Neucr
Arath
MutS1
Aquae
Chltr
Deira
Theaq
Thema
Ecoli
Neigo
Synsp
Strpy
Bacsu
Borbu
Trepa
C.
MSH5
Human
Yeast
Caeel
MutS2
Aquae
Strpy
Bacsu
Synsp
Deira
Borbu
Helpy
MutS2-Metth
MutS2-Saugl
MSH4
Yeast
Caeel
Human
MSH6
Yeast
Human
Mouse
Arath
MSH3
Arath
Human
Mouse
Spombe
Yeast
MSH1
Yeast
Spombe
MSH2
Fly
Xenla
Rat
Mouse
Human
Yeast
Neucr
Arath
MutS1
Aquae
Chltr
Deira
Theaq
Thema
Ecoli
Neigo
Synsp
Strpy
Bacsu
Borbu
Trepa
D.
MSH5
Segregation &
Crossover
Human
Yeast
Caeel
MutS2
Aquae
Strpy
Bacsu
Synsp
Deira
Borbu
Helpy
MutS2.Metth
MutS2.Saugl
MSH4
Segregation &
Crossover
Yeast
Caeel
Human
MSH6
MMR of
Mismatches and
Small Loops
in Nucleus
Yeast
Human
Mouse
Arath
MSH3
MMR of
Large Loops
in Nucleus
Arath
Human
Mouse
Spombe
Yeast
MSH1
MMR in
Mitochondria
Yeast
Spombe
MSH2
All MMR
in Nucleus
Fly
Xenla
Rat
Mouse
Human
Yeast
Neucr
Arath
MutS1
All MMR
(Bacteria)
Aquae
Chltr
Deira
Theaq
Thema
Ecoli
Neigo
Synsp
Strpy
Bacsu
Borbu
Trepa

Figure 3. Gene duplication and gene loss in the history of the bacterial MutS homologs.

A. Neighbor-joining phylogenetic tree of the *MutS1* and *MutS2* subfamilies (using only those proteins from species with both). The identical topology of the tree in the two subfamilies suggests the occurrence of a duplication prior to the divergence of these bacteria. B. Gene loss within the bacteria. Gene loss was determined by overlaying the presence and absence of MutS1 and MutS2 orthologs onto the tree of the species for which complete genomes are available (since only with a complete genome sequence can one be relatively certain that a gene is absent from a species). The thick gray lines represent the evolutionary history of the species based on a combination of the MutS and rRNA trees for these species. The thin colored lines represent the evolutionary history of the two MutS subfamilies (*MutS1* in red and *MutS2* in blue). Branch lengths do not correspond to evolutionary distance. Gene loss is indicated by a dashed line and each loss is labeled by a number: (1) MutS2 loss in enterobacteria; (2) MutS1 loss in *H. pylori*; (3) MutS2 loss in the mycoplasmas; (4) MutS1 loss in the mycoplasmas; and (5) MutS2 loss in *T. pallidum*.

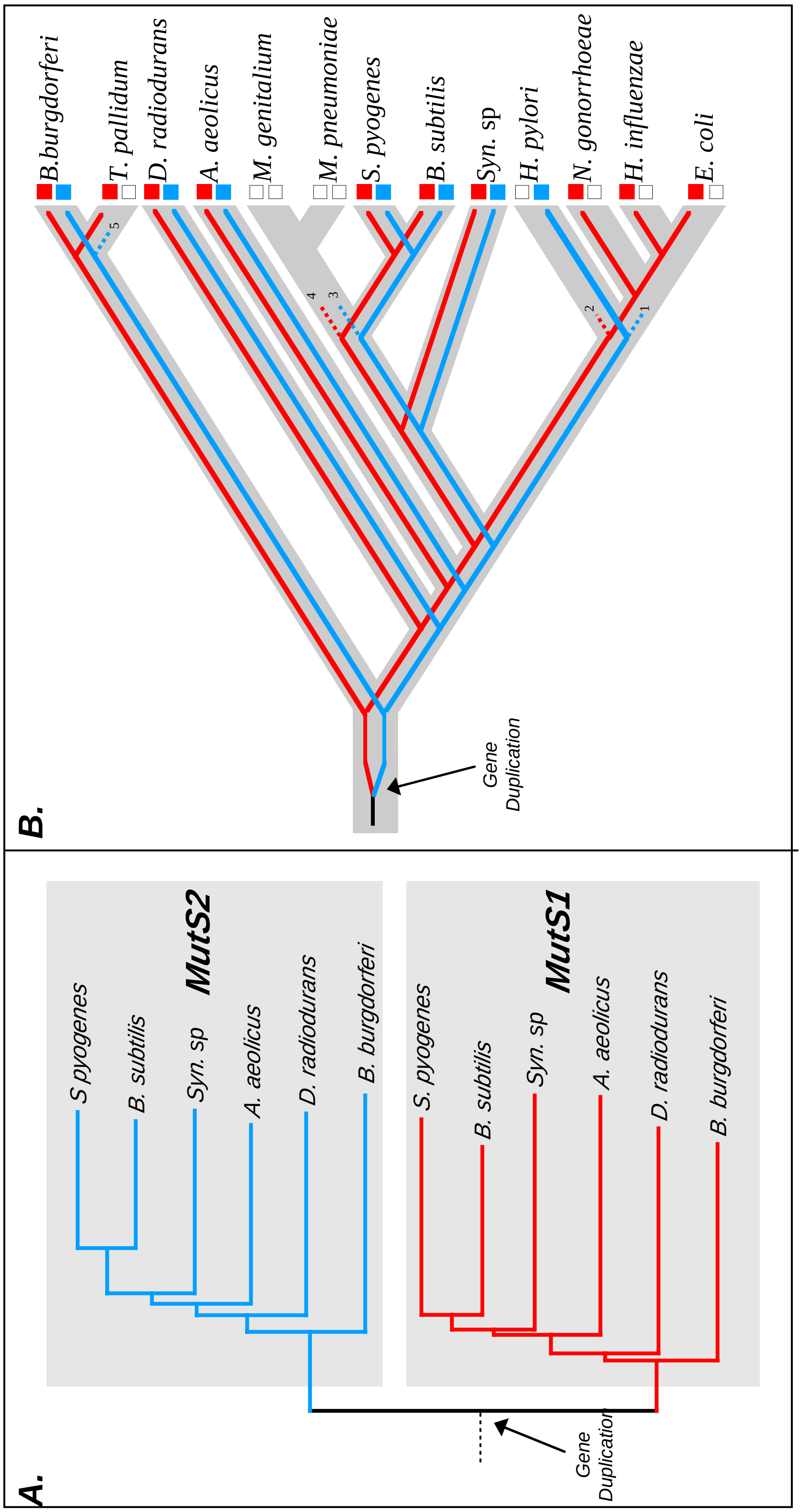
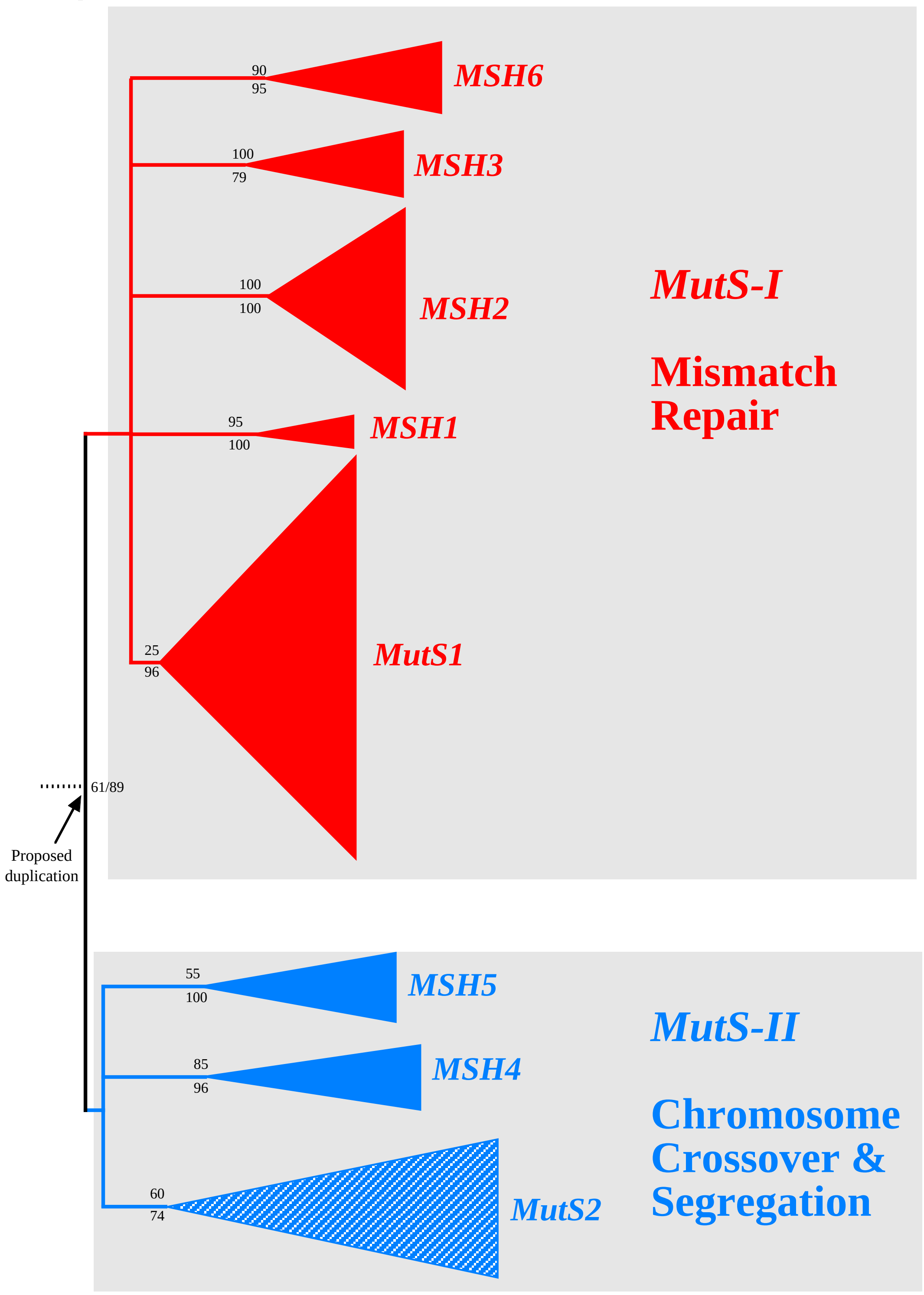

A.
MutS2
S pyogenes
B. subtilis
Syn. sp
A. aeolicus
D. radiodurans
B. burgdorferi
MutS1
S. pyogenes
B. subtilis
Syn. sp
A. aeolicus
D. radiodurans
B. burgdorferi
Gene Duplication
B.
B.burgdorferi
T. pallidum
D. radiodurans
A. aeolicus
M. genitalium
M. pneumoniae
S. pyogenes
B. subtilis
Syn. sp
H. pylori
N. gonorrhoeae
H. influenzae
E. coli
1
2
3
4
5
Gene Duplication

Figure 4. Consensus phylogenetic tree of MutS family of proteins.

Branches with low bootstrap values or that were not-identical in trees generated with different methods were collapsed. Only the proposed subfamilies are shown (sequences in each group are listed in Table 1). In addition, two proteins that are related to the *MutS2* subfamily are grouped with it. The height of each subgroup corresponds to the number of sequences in that group and the width corresponds to the longest branch length within the group. Bootstrap values for specific nodes are listed when over 40% (neighbor-joining on the top, parsimony on the bottom). The root of the tree was assigned as discussed in the text between the groups labeled *MutS-I* and *MutS-II*. Conserved functions for the different groups are listed.


90
95
MSH6
100
79
MSH3
100
100
MSH2
MutS-I
Mismatch Repair
95
100
MSH1
25
96
MutS1
61/89
Proposed duplication
55
100
MSH5
MutS-II
85
96
MSH4
Chromosome Crossover & Segregation
60
74
MutS2

CHAPTER 7

A Phylogenomic Study of DNA Repair Genes, Proteins, and Processes

ABSTRACT

The ability to recognize and repair abnormal DNA structures is common to all forms of life. Studies in a variety of species have identified an incredible diversity of DNA repair pathways. This diversity is seen with regard to the specificity, complexity, and mechanisms of the different pathways as well as the overlap with other cellular functions. Based upon general mechanisms of action, the pathways can be classified as direct repair (e.g., PHR, alkylation reversal, ligation), excision repair (base, mismatch or nucleotide) and recombinational repair. Within any particular species, multiple types of repair are usually found. Comparisons between species reveal that some aspects of repair are similar between species while many features are different. Documenting and characterizing the similarities and differences in repair between species has important value for understanding the origin and evolution of repair pathways as well as for improving our understanding of phenotypes affected by repair (e.g., mutation rates, lifespan, tumorigenesis, survival in extreme environments). Unfortunately, while repair processes have been studied in quite a few species, the ecological and evolutionary diversity of such studies has been limited. Complete genome sequences can provide potential sources of new information about repair in different species. In this paper we present a global comparative analysis of DNA repair proteins and processes based upon the analysis of publicly released complete genome sequences. We use a new form of analysis that combines genome sequence information and phylogenetic recreations into one composite phylogenomic analysis. We use this phylogenomic analysis to study the evolution of repair proteins and processes and to predict the repair phenotypes of those species for which we now know the complete genome sequence.

INTRODUCTION

Genomic integrity is under constant threat in all species. Threats come in the form of endogenous and exogenous agents that damage DNA and/or interfere with DNA metabolic processes, as well as spontaneous base loss or deamination and errors in DNA metabolism such as nucleotide misincorporation during replication. These threats lead to a variety of alterations in the normal DNA structure including single- and double-strand breaks, chemically modified bases, abasic sites, bulky adducts, inter- and intra-strand cross-links, and base-pairing mismatches. The direct effects of these abnormalities include mutations at or near the site of the abnormality, genetic recombination, and the inhibition or alteration of cellular processes such as replication and transcription. These direct effects can lead in turn to many indirect effects including chromosomal aberrations, tumorigenesis, apoptosis, developmental abnormalities, and/or necrosis.

The primary mechanism by which organisms maintain their genomic functions in the face of these threats is by removing the abnormalities from the DNA and restoring the genomic integrity, a process known as DNA repair. Experimental studies in a variety of species have documented an incredible diversity of repair pathways. One aspect of this diversity relates to the types of abnormalities that can be repaired. Overall, pathways have been found that can repair almost any type of lesion, but pathways differ a great deal from each other in their specificity. Some are dedicated to repairing a specific abnormality while others are able to deal with a broad spectrum of abnormalities. Another aspect of the diversity of repair pathways is a diversity of functions. The functions of repair pathways include the correction of replication errors, resistance to killing by DNA damaging agents, chromosome duplication and segregation, cell cycle control, generation of antibody diversity in vertebrates, regulation of interspecies recombination, meiotic and mitotic recombination, transcription or replication elongation and tumor suppression. Since function is determined in a large part by specificity, the functions and specificity of particular pathways are closely interrelated.

Understanding the diversity among DNA repair pathways requires an understanding of the mechanisms of these pathways. Not surprisingly, these mechanisms are also diverse. Some are simple, involving single enzymes and single steps; others are incredibly complex, involving many steps and dozens of enzymes working in concert. Fortunately, the comparison of repair pathways is simplified by the fact that all repair pathways can be placed into one of three classes based on its general mechanism of action: direct repair, recombinational repair, and excision repair. In direct repair, alterations in the structure of DNA are simply reversed. Examples include photoreactivation (the light activated reversal of UV induced cyclobutane pyrimidine dimers and/or 6-4 photoproducts), alkyltransfer (the removal of inappropriate alkyl groups from DNA) and DNA ligation (the restoration of simple phosphodiester bond breaks in the DNA backbone). In recombinational repair, sections of altered or damaged DNA are corrected by homologous recombination with undamaged templates (see (1) for review). Thus, there is a great deal of overlap between the pathways involved in general recombination and those involved in recombinational repair. Finally, in excision repair, first a section of one strand of the DNA double-helix containing the abnormality is excised, then the other strand is used as a template to correctly resynthesize the removed section, and finally the patch is ligated into place (see (2) for review). Thus the excision repair pathways capitalize on the redundancy of the information in the complementary strands of DNA to restore the correct DNA structure. There are three major forms of excision repair that are distinguished by the type of abnormality removed and by the mechanism of its recognition and removal. In base excision repair (BER), inappropriate, damaged, or modified bases are removed and the resulting abasic site is repaired by a process that replaces only one or a few nucleotides; in nucleotide excision repair (NER) abnormal DNA structures are removed as part of an oligonucleotide and longer patches are introduced; and in mismatch repair (MMR) base mismatches or unpaired loops are removed as part of a very long stretch of nucleotides. The diversity of mechanisms, specificity, and functions of repair pathways outlined above includes the diversity of all known repair pathways.

It is also useful to compare repair processes on a species by species basis. Such comparisons reveal that some aspects of repair are similar between species and some are different. All species examined in detail have been found to exhibit multiple repair pathways, usually including many of the different classes and types of repair. For example, *Escherichia coli* and *Saccharomyces cerevisiae* perform photoreactivation (PHR), alkylation reversal, BER, MMR, NER, and recombinational repair. Although the use of multiple repair pathways is likely universal, the repertoire of types of repair frequently differs between species. For example, although PHR is found in a wide range of species, many species, including humans, lack it. In addition, there are some types of repair that are found in only a small range of species (e.g., a process called spore reactivation is found only in *Bacillus subtilis* and some closely related species).

Another way to compare repair processes between species is to characterize the similarities and differences within each class of repair between species (e.g., compare MMR in *S. cerevisiae* and *E. coli*). From a broad perspective, each particular class of repair is similar in all species, and these similarities even go beyond the characteristics that were used to broadly define the class. For example, all known NER processes, from bacteria to humans, follow the same steps (damage recognition, incision (generally on both the 5' and 3' sides of the lesion), excision, repair synthesis and ligation) and all have similar broad recognition capabilities (all can repair many types of lesions). Similarly, although not all species have PHR, all known PHR processes are single enzyme pathways that have very similar mechanisms of action. However, closer examination of the details of the processes in different species reveals a great deal of diversity in how particular species carry out the respective classes of repair. For example, although all PHR processes are similar, the specificity varies between and even within species. In some species PHR reverses only pyrimidine dimers, in others it reverses only 6-4 photoproducts, and some species have multiple PHR processes that are able to repair both CPDs and 6-4s. The specificity, range and spectrum of MMR also frequently differ between species. Each species exhibits preferences for repairing particular mismatches and particular sizes and types of unpaired loops. Differences in specificity, some subtle, some large, are found in almost all classes of repair. Since specificity and function are closely interrelated, the differences in specificity frequently lead to differences in function. Thus, the finding that two species exhibit the same repertoire of repair types does not mean that they have identical repair processes.

We have been interested in documenting and understanding the similarities and differences in repair processes between species. A major rationale for this is that differences in repair between species can have profound biological effects. For example, it has been suggested that the accelerated mutation rate in mycoplamsas may be due in part to deficiencies in DNA repair (3,4). Examples of biological outcomes that may be due to differences in repair include cancer rates both within and between species (5), lifespan (6,7), pathogenesis in bacteria (8-10), codon usage and GC content (11,12), evolutionary rates (13), survival in extreme environments (14), speciation of bacteria

(15,16), and diurnal/nocturnal patterns (17). Thus, to understand differences in any of these phenotypes, it is useful to understand differences in repair. In addition, understanding differences in these phenotypes can have secondary uses. For example, many aspects of sequence analysis such as database searches, phylogenetic analysis, sequence alignment generation and population analysis are optimized when they include information on mutation rates and patterns. Characterization of repair processes and mutation rates and patterns in many species should help optimize these analyses.

We are also interested in using comparative data on DNA repair processes to understand the evolution of repair proteins and processes. Since DNA repair is a major cellular process, it is of interest to understand how different repair pathways originated and how differences between species came to be. In addition, information about the evolution of repair provides a useful perspective for comparative repair studies and thus helps us understand the differences in repair between species as well as the mechanisms and functions of particular repair processes within a species. For example, evolutionary studies of PHR show that all PHR processes are homologous and that the differences between species are due to functional changes in photolyase enzymes (18). Evolutionary studies have many other potential uses in the study of repair including the characterization of genes that are part of multigene families (19-21), the prediction of functions for uncharacterized genes (22) and the identification of motifs conserved among particular homologs (21). In general, an evolutionary perspective is useful in any comparative biological study because it allows one to go beyond identifying what is similar or different between species to understanding how and why such similarities and differences may have arisen.

Unfortunately, evolutionary and comparative studies of DNA repair processes have been limited because of the lack of detailed studies of repair in a wide ecological and evolutionary diversity of species. Although new model systems for repair are being developed, the majority of repair studies have been carried out in only a few bacterial species, yeast, and animals. Recently, a potential new source of comparative biological data has emerged: complete genome sequences. Complete genome sequences provide an unprecedented view into the entire genetic makeup of individual species. In theory, complete genome sequences should enable the prediction of all one could want to know about a particular strain or species, while providing a wealth of data for comparative analysis. In practice, however, obtaining useful information from complete genome sequences is quite difficult. We have been developing a new approach that combines the analysis of complete genome sequences with evolutionary reconstructions into one phylogenomic analysis. We present here a global phylogenomic analysis of DNA repair proteins and processes. We use this phylogenomic analysis to make predictions about the repair phenotypes of species for which genomes have been sequenced and to infer the evolutionary history of repair pathways and the respective proteins that comprise them. In addition, we discuss the value and uses of evolutionary analysis in studies of complete genome sequences and the value and uses of complete genome sequences in studies of evolution as well as the advantages of the combined phylogenomic approach.

METHODS

Database of repair proteins

We created a database of proteins with established roles in DNA repair processes. We focused on proteins from model organisms such as *E. coli*, *B. subtilis*, yeast, and humans. The database is available at http://www-leland.stanford.edu/~jeisen/Repair/Repair.html. In addition, a variety of supplemental data sets related to this analysis are also at this site.

Searching for homologs

Sequences similar to that of each protein in our database were identified using the blast and blast2 search algorithms (23). Databases searched included the nr and EST databases at NCBI, the TIGR genome database (for *B. burgdorferi*, *T. pallidum* and *D. radiodurans*) and the Oklahoma University genome database (for *N. gonorhoeae* and *S. pyogenes*). Iterative search techniques were also performed (either using PSI-blast or by manually selecting lower scoring sequences that were still above the threshold) to be used as new query sequences.

Sequence and evolutionary analysis

Protein sequence alignment was performed using the clustalw program (24). Profiles and blocks were made of some alignments using various world wide web servers. Some of these were then used for additional database searches to identify sequences containing motifs similar to those that were aligned together. Alignments and blocks are available at the above web site.

Phylogenetic trees were generated from the sequence alignments (excluding poorly conserved regions) by the neighbor-joining and parsimony methods of the PAUP* program (25). Evolutionary distribution patterns were analyzed as described in the discussion section using the MacClade program (26). Presence and absence of genes was treated as a binary character state for parsimony analysis. This was used to identify the timing of gene gain and loss events. Gene duplication and lateral transfers were incorporated into this analysis if they were identified by the methods described in the discussion section. Absence was only determined for species for which complete genome sequences were available. Presence was determined from the database searches. Homologs were considered present in a particular species by the criteria described in the Results. Predictions of functions for uncharacterized genes were performed using methods previously described (22).

RESULTS AND DISCUSSION

It is well established that many aspects of comparative biology benefit from an evolutionary perspective. This is because all biological processes and entities have a history, and inferring that history can only serve to benefit comparative studies. The benefits of an evolutionary perspective have been taken for granted in many areas of comparative biology. However, comparative molecular biology has tended to focus on quantifying the levels of similarity among species and not on how and why those similarities arose. This is particularly true for comparative genomics and genome analysis in general. We believe that an evolutionary perspective is just as useful in comparative genomics as it has been in other aspects of comparative biology.

Specifically, we have been developing methods that combine evolutionary reconstructions and genome analysis into a single phylogenomic analysis. The principle behind combining evolutionary reconstructions and genome analysis into a composite phylogenomic approach is that evolutionary reconstructions improve what can be learned from complete genome sequences and conversely that complete genome sequences improve what can be learned about evolution. Since this phylogenomic approach is novel, we first discuss some of its general principles and some details of the methods we used before focusing on the phylogenomic analysis of DNA repair proteins and processes.

Summary of Phylogenomic Analysis

Our phylogenomic approach can be considered to be a feedback loop, since evolutionary information is used to improve genome analysis and genome analysis is used to improve inferences of evolutionary history. An outline of our approach is presented in Figure 1, and some details of the different steps are described in Table 2. Each step is discussed in more detail below. In summary, we first used database searches to identify the presence and absence of homologs of known repair genes in complete genome sequences. Then we used a variety of methods to infer the evolutionary history of each group of homologs. This analysis depended in large part on the presence/absence data and it was used in turn to refine the presence/absence data to render it more accurate and informative. The evolutionary information and the refined presence/absence data was then used to characterize particular repair genes and pathways (e.g., their evolution, functions, specificity, etc.) and to predict species phenotype.

Presence and absence of homologs

The first step in our phylogenomic analysis was the determination of the presence and absence of homologs of known repair genes in species for which complete genomes are available (see Table 1 for a listing of species). Presence and absence of particular genes have many uses in studying individual species as well as in learning about the evolution of genes and pathways. The ability to determine both presence and absence of homologs of known genes in a particular species is one of the great benefits of having complete genome sequences. Prior to the "genomics era", presence of homologs in different species could only be determined if one were able to clone the gene of interest from that species. Absence of homologs could only be surmised from the negative results of experiments like degenerate PCR or complementation analysis.

To determine presence and absence of genes in the complete genomes we used a variety of sequence-searching methods. Since homology is an inference about common ancestry, it was necessary to set limits for the level of sequence similarity we considered to imply homology. We used a conservative operational definition of homology (i.e., high threshold of sequence similarity) to limit the number of false positive results (i.e., identifying genes as homologs that do not share common ancestry). For blastp searches a p value less than 1×10^{-6} was used as a threshold. Since this conservative approach might lead to false negatives, we additionally used iterative search methods (e.g., PSI-blast and manual methods) to increase the likelihood of identifying highly divergent homologs of the reference protein. In some cases, this threshold was lowered if other evidence suggesting homologs was highly divergent. More detail is provided in the respective sections on each pathway. We used these methods to identify presence and absence of homologs in the complete genome sequences as well as presence (but not absence) in other species. Since this presence/absence data was refined by some of the evolutionary analysis, it is elaborated below.

Evolutionary relationships among homologs: gene trees

The second step in our analysis was the determination of gene trees (the evolutionary relationships among all homologs of each gene). Gene trees were generated using standard phylogenetic methods for each group of homologs. All homologs were used to generate the gene trees (not just those from complete genome sequences) since the use of more sequences usually improves the accuracy of the trees. The gene trees were used for many purposes including the prediction of gene functions, the division of gene families into subfamilies, and the identification of evolutionary events such as gene duplications and gene loss in each gene family (each of these is discussed separately below).

Evolutionary distribution patterns and the identification of evolutionary events

One way that the genome analysis was combined with evolutionary analysis was in the determination of evolutionary distribution patterns. This involved overlaying the presence/absence information for particular species onto an evolutionary tree of those species. Evolutionary distribution patterns tell a great deal about the evolutionary history of particular genes (see Table 3). For example, if a gene is present in only one subsection of the species tree (referred to as “uniform presence”) then it probably originated at the base of that section. If a gene is present in all species except for one lineage (referred to as “uniform absence”) then it is likely that the gene was lost at the base of that lineage. In addition, if a gene is found in all species (“universal” distribution) it is likely that it is an ancient gene that was present prior to the divergence of the main lineages of life (and also that it may be universally required in all species). Some distribution patterns do not have a single likely mechanism of generation and thus require further analysis before being used to identify specific evolutionary events. For example, an uneven distribution pattern (that is, scattered presence and absence throughout the species tree) can be explained by either lateral transfer to the species with an unexpected presence of the gene or by gene loss in species with an unexpected absence. In addition, the presence of multiple homologs in some species can be explained either by lateral transfer to the species with multiple copies or a gene duplication event in some lineages. In these cases, ascertaining which event occurred can usually be accomplished by comparing the gene tree to the species tree and testing for congruence. In the uneven distribution, if there had been a lateral transfer, then the species tree and the gene tree should be incongruent (i.e., they should have different branching topology); if there were gene loss, then the gene and species trees should be congruent (except that some species will not be represented in the gene tree). In the multicopy example, if there had been a gene duplication, the gene tree should have two separate lineages that run parallel to the species tree (these two lineages are called paralogous); if there had been a lateral transfer then the species and gene trees should be incongruent and the “transferred” gene lineage should be more related to genes in the donor’s lineage than to those in closely related species. The importance of the presence/absence data in identifying these events is one of the reasons that complete genome sequences are so powerful in studies of evolution.

When we identified particular events, we used parsimony reconstruction methods to determine the timing of these events. In short, one attempts to identify the evolutionary scenario that requires the fewest gene gain and loss events to arrive at the current distribution patterns. Since this type of analysis is not commonly used for molecular data, we provide an example (for tracing gene loss) in Figure 2.

An essential component in the above is the species tree - in order to infer evolutionary events one must have an accurate picture of the evolutionary relationships among the species being compared. Unfortunately, there is no general consensus concerning the relationships among all of the species analyzed here. For presentation purposes, we have used a species tree based upon the Ribosomal Database Project trees (27) in which Archaea, bacteria and eukaryotes are each monophyletic and in which Archaea are a sister group to bacteria. Within the bacterial part of the tree, we divide the species into major phyla but have collapsed the branches joining the different phyla to indicate that the relationships among these phyla are ambiguous. In the sections on specific repair pathways, we discuss whether and how alternative species trees affect our conclusions about the evolutionary events in particular gene families.

Refining homology groups: subfamilies

As mentioned above, one of the ways we used the evolutionary analysis was to refine the presence/absence lists. This was necessary because some genes are members of multigene families and therefore the mere presence of a homolog in a particular species is incomplete information. In such cases, it is much more informative to know whether a homolog from the same subfamily as the query gene is present in the species of interest. We used evolutionary analysis to divide up multigene families in two ways. First, if gene duplication events were identified, then genes were divided into groups of orthologs and paralogs. In addition, even if particular duplication events were not identified, we used the gene trees to subdivide the gene families into evolutionarily distinct subfamilies. The results of our refined analysis are presented in Table 4.

Additional detail can be found in the discussion sections on each specific repair pathway.

Functional predictions and functional evolution

In order to make predictions about the phenotypes of species for which complete genomes were available we needed to make predictions for the functions of each of their genes. Usually, such functional predictions are accomplished by identifying homologs and assigning the uncharacterized gene the function of its homologs. However, identification of homologs is not always adequate – frequently not all homologs have the same function. In such cases, one needs an approach to choose which homolog to use to assign a function to the uncharacterized gene. In these cases, function has usually been assigned based on the highest scoring homolog (with the score based on blast or other searching programs). We have argued that such "highest-hit" methods can frequently be inaccurate and that the best way to predict functions in these cases is through evolutionary analysis (22). The problems with similarity of blast-based functional predictions include that they are prone to database error propagation, they cannot be used to identify orthologous groups reliably, they perform poorly in cases of evolutionary rate variation and non-hierarchical trees, they can be easily misled by modular proteins or large insertion/deletion events, and they are not set up to deal with expanding data sets. Our evolutionary-based functional prediction involves tracing the evolutionary history of the genes of interest and then overlaying onto this history any experimentally determined functions for any of the genes. Predictions for uncharacterized genes are made based on their position in the tree relative to the genes with known functions and based on identifying evolutionary events such as gene duplications that may identify groups of genes with similar functions (22,28). We searched the literature for any functional information on the identified homologs of the gene of interest and used the methods described (22) to make functional predictions. It should be remembered however that all such sequence analysis methods are only predictions and need to be confirmed by experimental approaches. It is imperative that all sequence databases explicitly state which database annotation information is based on experimental studies and which is based on predictions of function.

This analysis was also used to study functional evolution of particular genes. We were interested in how frequently functions changed and whether this could be used to infer anything about how likely such functions were to evolve in other species. If there were many functional changes in the history of a particular gene or gene family, then the identification of the presence of homologs of such genes in a species would not be sufficient information to predict the presence of any activity.

Characterization of Pathways

We used the refined presence/absence results, the gene evolutionary information, and the functional evolution information to characterize the evolution of entire pathways. The first question we asked for a particular pathway was whether that pathway was characterized in other species. Then we asked, if one gene in the pathway is present in a particular species, are the other genes in that pathway also always present? If so, this strongly suggests a conserved association among the genes in the pathway. If the genes are not always present together, there can be multiple possible explanations. In some cases, a species may replace one gene with another (also known as non-orthologous gene replacement). Thus the absence of a particular gene may not necessarily mean the absence of the whole pathway. Alternatively, the pathway may not work the same way in other species. Finally, perhaps most interestingly, a lack of a conserved association across species can also suggest that genes that we think operate together in the model species used to characterize the pathway, may not work together as we thought.

We used the evolutionary analysis to learn more about particular distribution patterns and particular pathways. First, the functional predictions and functional evolutionary information were used to characterize how the genes in a pathway may have changed functions over time. Then, the information on evolutionary events was used to see if any evolutionary events occurred at the same time to the different genes in a pathway. Such correlated events suggest some sort of conserved association among genes. For example, if two genes work together and cannot perform their functions alone, then it would be expected that if one of the genes was lost in a lineage, the other gene might be lost soon after since there would be little selective benefit to its maintenance. A conserved association among genes would be even more strongly supported if the correlated events occurred multiple times in different lineages.

Another aspect of our analysis of pathways was identifying a likely timing of the origin of the pathway, based on combined analysis of the origins of each of the genes in the pathway. This was useful in understanding the origins of the pathway and of differences between species. In addition, it was also useful for predicting species phenotypes, in much the same way that information on functional evolution of individual genes was useful. If pathways with a particular activity evolved only once, then the presence and absence of the genes required for that pathway can be used as a good estimator of the presence and absence of the activity. However, if an activity evolved many times, then one should be careful when making phenotypic predictions. In such cases, the presence of genes that perform that activity may be useful in predicting the presence of an activity. However, the absence of such genes is not very informative since there may be yet another set of genes that have yet to be characterized can perform the same activity and these may be present in the species of interest.

Focusing on why events occurred

In addition to identifying evolutionary events, a lot can be learned about the function and evolution of genes by trying to determine why the events occurred. For example, gene loss events can occur for many reasons and identifying which occurred in particular cases can reveal a lot about the genes of interest. Possible reasons for gene loss include that the gene has low utility in the lineages in which it was lost (and thus there would be little negative selection against the loss of these genes), that the gene is less stable than other genes and thus more prone to loss, or that there is or has been a selective advantage to losing the gene. To determine the likely reasons why such events occurred it is particular helpful to determine if the event has occurred multiple times and if there is anything similar about the lineages in which the event has occurred.

Phylogenomic Analysis of Specific Repair Pathways

We have divided up the discussion of our phylogenomic analysis of repair into two main sections. In the first section, we discuss our results on a pathway by pathway basis. For each pathway, we review what is known about the pathway and the proteins in that pathway in the species in which the pathway is best characterized. Then we discuss what is known about this pathway in other species. Finally we present the results of our phylogenomic analysis as well as results of other comparative or evolutionary studies of this pathway. In the second section, we discuss the results from a broader perspective, looking at all repair pathways together.

Direct Repair I: Photoreactivation (PHR)

Photoreactivation (PHR) is a general term used to refer to the ability of cells to make use of visible light to reverse the toxic effects of UV irradiation. PHR has been found in a wide diversity of species, including bacteria, Archaea, and eukaryotes. Despite the highly general way that PHR was defined, all characterized enzymatic PHR processes involve a similar type of direct repair of UV irradiation induced DNA lesions. Therefore the term photoreactivation is frequently used more narrowly to refer to this type of DNA repair. The first well-characterized PHR process was that of *E. coli.* This process is carried out by a single photolyase enzyme which uses the energy of visible light to reverse UV induced cyclobutane pyrimidine dimers (CPDs) in DNA. PHR processes have now been characterized in detail in many other species. Two different types of PHR have been discovered – the most common one involving the reversal of CPDs and the other involving the reversal of 6-4 pyrimidine-pyrimidone photoproducts (also formed by UV irradiation but with a lower quantum yield than for CPDs). Despite the different substrates, all known PHR processes are actually quite similar to each other. All are single step processes, like that in *E. coli*, and all of the enzymes that perform PHR are homologous. Thus the 6-4 photolyase and CPD photolyase have descended from a common ancestor even though the respective photoproduct substrates are quite different structures. No photolyase can repair both 6-4s and CPDs and some photolyase homologs do not repair any lesions but instead function as blue-light receptors (29). Interestingly, a photolyase homolog has been cloned from humans but has been shown to not exhibit any photolyase activity. Other differences between photolyases include the action spectrum and wavelength of light required for peak activity and in the particular cofactor used to facilitate energy transfer (18).

Evolutionary and comparative analysis provides a great deal of insight into the photolyase gene family (see (18) for a review). Comparative sequence analysis of photolyase genes reveals that the photolyase gene family can be divided into two subfamilies, referred to as classI and classII. In classI are the photolyases of *E. coli*, *H. halobium* and yeast, as well as the blue-light receptors from plants and the human gene with no known function. In classII are the photolyases from *M. xanthus*, *M. thermoautotrophicum*, goldfish and marsupials. Our analyses, as well as previous studies (18), suggest that these two subfamilies are related by an ancient gene duplication. First, the finding of homologs of PhrI and PhrII in at least some species from each of the major domains of life (Table 4) suggests that each of these were present in a common ancestor of all life. In addition, phylogenetic trees of photolyases suggest that the gene duplication was ancient (18). Phylogenetic reconstructions also help understand the functions of the different photolyase homologs and show that functional changes have been quite common in the history of photolyase homologs. For example, the blue-light receptors have likely descended from genes with photolyase activity and thus sometime in their evolution they lost photolyase activity but retained the ability to absorb blue-light.

Despite the conservation of sequence and general function among all photolyases across all domains of life, our genome analysis shows that many species do not encode any photolyase homolog and most that do encode either a PhrI or a PhrII. We believe that this uneven distribution pattern can be explained mostly by gene loss events in some lineages. For example, the absence of PhrI from *H. influenzae*, is likely due to a relatively recent gene loss since many other gamma-Proteobacteria do encode a PhrI (including *E. coli*, *N. gonorrhoeae*, and *S. typhimurium*). The rampant loss of Phr genes is not particularly surprising since most of these genes have very specific functions in repairing UV induced DNA damage. First, many species are not exposed to significant levels of UV irradiation and thus a Phr gene may not be of any use. In addition, other processes such as NER can repair the same lesions that are repaired by Phr enzymes so photolyases are redundant in the presence of NER. It is possible that recent gene loss has occurred in humans as well. Marsupials have been found to encode a classII photolyase, but no such gene has yet been found in humans. Since humans apparently lack any photolyase activity (30), either humans have a classII photolyase gene which encodes a protein that does not function as a photolyase (like the classI photolyase gene) or humans have lost their classII gene sometime since diverging from marsupials. Because the history of photolyases is filled with functional changes and loss of function, we do not believe that the presence of a photolyase homolog in a species can be used to unambiguously predict the presence of photolyase activity or its nature (e.g., CPD vs. 6-4).

The ancient origin of the photolyase genes and the fact that most members of this gene family encode functional photolyases suggests that the ancestral protein was a photolyase and thus that the common ancestor of all life forms could perform PHR. The specific origin of photolyase enzymes is difficult to determine since the Phr gene family does not show any obvious homology to any other proteins. However, it is useful to recognize that limited photolyase activity can be provided by a tripeptide sequence (Lys-Trp-Lys) (31-34), suggesting that a photolyase protein could have evolved relatively easily early in evolution. Photolyase genes could have been more essential in the early evolution of life since there was no ozone layer then to attenuate the intense solar UV flux.

Direct Repair II: Alkylation Reversal

A common form of damage to DNA bases is alkylation, in which alkyl groups (especially methyl and ethyl groups) are covalently linked to different sites of a base. Such alkylation can be caused by many chemical and enzymatic processes. One of the ways that cells repair this damage is by transferring the alkyl-group onto a protein - a form of direct repair called alkyltransfer (35,36). As with PHR, alkyltransfer repair has been characterized in a wide diversity of species including bacteria, Archaea, and eukaryotes, and the process is highly similar in different species. All alkyltransferase processes involve single enzymes which perform the reaction only once - thus the transfer of the alkyl group to the protein is a suicide process. All alkyltransferases characterized to date have been found to repair only O-6-alkyl guanine. As with photolyases, all alkyltransferase are homologous to each other - they all share a common core alkyltransferase domain that is highly conserved (37,38). The comparison of alkyltransferase proteins is somewhat complicated however because some contain additional domains. For example, *E. coli* encodes two different alkyltransferases: Ada and Ogt. Ogt contains only the alkyltransferase domain while Ada contains the alkyltransferase domain and a transcriptional regulatory domain (which itself is related to a large group of transcriptional regulator proteins). Ada uses the second domain as part of an inducible response to alkylation damage, obtaining the signal interestingly from phosphotriesters (i.e., the alkylation of the DNA backbone). Thus, methylation of Ada activates its transcription activation domain, leading to induction of the *alkA* and *ada* genes.

Our comparative analysis shows that, as with photolyases, although alkytransferase homologs are found in a wide diversity of species they are not universal. Based on the finding of alkyltransferase homologs in at least some species from each of the major domains of life, we conclude that alkyltransferases are ancient repair proteins - present in a common ancestor of all species. Since all known members of this gene family function as alkyltransferases, we also conclude that the common ancestor was able to perform this type of repair. The addition of the transcriptional regulatory motif onto the Ada protein appears to have occurred recently - only close relatives of *E. coli* encode proteins like Ada. Interestingly Gram positive bacteria also encode a fused protein with an Ada motif: but in this case the Ada motif is fused to an alkyl glycosylase motif (see Figure 3). Since the alkyltransferase family appears to be ancient, the species (*N. gonorhoeae*, *S. pyogenes*, the two Mycoplasmas, *Synechocystis* and *Borrelia borgdorferi*) that do not encode any homolog of an alkyltransferase evidently have lost the gene in the recent past.

The presence of an alkyltransferase homolog in a species indicates the likely presence of alkyltransferase activity since all members of this gene family that have been characterized have been found to have the activity. The absence of alkyltransferase homologs likely indicates the absence of alkyltransferase activity since no other proteins have been found to have this activity. However, the species without alkyltransferase may still be able to repair alkylation damage since it can also be repaired by forms of BER and by NER. Although some of these species do not encode homologs of alkylation glycosylases, almost all encode likely NER systems (see below).

Direct Repair III: DNA Ligation

DNA ligation is the process of joining together two separate DNA strands. This process is required for replication, recombination, and any form of excision repair. If used to repair single-strand or double-strand breaks, ligation can be considered to be a type of direct repair. However, there are strict constraints on the chemical nature of the DNA strand ends to be joined; a 3’ OH end may be ligated to a 5’ phosphate end to complete the phosphodiester linkage. The ligases that are used for DNA repair can be divided into two families that are apparently not evolutionarily related to each other. Proteins in family I are NAD-dependent DNA ligases. These have been found and characterized in many bacterial species and all have similar functions. Proteins in ligase family II are ATP-dependent DNA ligases. These have been characterized in many viruses, Archaea, and eukaryotes. Multiple members of family II have been found from many eukaroytes and these have similar but not completely overlapping functions.

Our comparative analysis shows that all bacteria encode a member of ligase family I and all Archaea and eukaryotes encode a member of ligase family II. Given the degree of conservation of function among the ligases, and the need for ligation activities in many cellular functions, it is certain that all species have ligation activity. Since all Archaea only encode one protein in ligase family II and many eukaryotes encode multiple members, we conclude that there have been many gene duplication events in this gene family in the history of eukaryotes. Interestingly, some bacteria encode a member of ligase family II (see Table 4) in addition to the universal bacterial ligase I. No function has yet been assigned to any of these bacterial ligase family II genes. Their presence in bacteria suggests either that they were transferred to these species laterally, or that the ligase II genes predate the separation of bacteria, Archaea, and eukaryotes. We believe it is more likely that these were transferred. Thus, we conclude that ligase family I originated early in bacterial evolution and that ligase family II originated in a common ancestor of Archaea and eukaryotes.

Mismatch Excision Repair

The ability to recognize and repair mismatches in DNA has been well documented in many species. Since mismatches can be generated in many ways, MMR processes have many functions including the repair of some types of DNA damage (e.g., deamination of methyl-C leads to a GT mismatch), the regulation of recombination (recombination between non-identical DNA sequences produces mismatches), and perhaps most importantly, the prevention of mutations due to replication errors. MMR processes come in two forms: specific and general. Specific MMR includes dedicated processes that repair special types of mismatches, such as GT mismatches. Some of these are glycosylases and are discussed in the base-excision repair section. The majority of MMR is carried out by broad recognition generalized MMR processes. General MMR was first characterized in *E. coli*. In *E. coli* the process works in the following way. First, the MutS protein binds to a mismatch or a small unpaired loop (small loops are formed by frameshift replication errors) and, with the cooperation of MutL, one of the two strands at the site of the mismatch is targeted for excision repair. The choice of which strand is determined a methylation-endonuclease system. The Dam protein methylates the A’s in GATC sites throughout the genome. However, newly replicated GATCs are transiently unmethlyated leaving a "window" of hemimethylated DNA behind the growing fork. MutH binds to hemimethlyated GATC sites in double-stranded DNA and cuts the unmethylated strand (and thus the newly replicated, error-containing strand) when activated by the MutS-MutL complex. Various exonucleases and the UvrD helicase excise the strand and a very large repair patch is resynthesized using the intact strand as a template.

The overall scheme of general MMR is similar in all species in which it has been characterized. However, not all details have been conserved between species. For example, while all species exhibit strand specificity, the mechanism of strand recognition is different between species. In addition, the post-cleavage steps (exonuclease, resynthesis, and ligation) involve non-homologous proteins between species. However, there is a conserved core of general MMR: homologs of the *E. coli* MutS and MutL proteins are absolutely required for MMR in all species (5,39). MutS (and its homologs) are always responsible for the recognition step and MutL (and its homologs) have an as of yet poorly characterized structural role. The comparison to the *E. coli* MutS and MutL proteins is somewhat complicated, however, because some species have multiple homologs of these proteins, and not all of them have the same functions, and in fact some are not even involved in MMR. In particular, many MutS homologs have no evident role in MMR.

Our comparative analysis reveals that many species encode neither MutS nor MutL homologs. Surprisingly, despite genetic studies that show that MutS and MutL homologs are both required for MMR, some species encode a MutS homolog but not a MutL homolog. Interpreting this distribution pattern is helped by phylogenetic analysis of the MutS family of proteins. In a previous study, we presented evidence that the MutS family underwent an ancient duplication into two lineages – which we refer to as MutS1 and MutS2 (20). The division into two lineages helps us understand the functional diversity within the MutS family. All the MutS homologs known to be involved in MMR are in the MutS1 lineage while those known to be involved in chromosome segregation are in the MutS2 lineage. Thus we believe that with regard to MMR it is only useful to document the presence and absence of genes in the MutS1 lineage. This conclusion is supported by the finding that MutS1 and MutL genes are always either present or absent as a unit. That is, all species with a gene in the MutS1 lineage also encode a MutL homolog, and all species without a gene in the MutS1 lineage also do not encode a MutL homolog (Table 4). In addition, it is also supported by the finding that the absence of these genes from many species is not due to a single gene loss event but to parallel events in which both genes were lost in different lineages. Thus we believe that the presence of MutS1 and MutL can be used as a predictor for the presence of MMR and those species without these genes (*H. pylori*, *M. tuberculosis*, the two mycoplasmas, and the three Archaea) likely do not have MMR. The two species that encode a MutS homolog but do not encode a MutL homolog (*H. pyrlori* and *M. thermoautotrophiucm*) likely do not have MMR both because they do not have a MutL homolog and because the MutS homolog they encode is in the MutS2 lineage.

Although no MutL and MutS1 genes have yet been found in Archaea, we still conclude that MutS and MutL are ancient proteins. Thus the absence of these genes from some species is inferred to be due to gene loss. Why would some species lose MMR? To answer this question it is helpful to recognize that there have been multiple parallel losses of the MutL and MutS1 genes. They have been lost in the mycoplasmal lineage (they are absent from the mycoplasmas but present in other lowGC gram-positive), in the

H pylori lineage (they are absent from this species but present in other Proteobacteria) and in the in the *M. tuberculosis* lineage and Euryarchaeota lineages (see (20) for a more detailed discussion of loss of MMR genes). Such multiple cases of gene loss suggest either that these genes are particularly unstable and are easily lost, or that there is some advantage to the loss of these genes. We believe that the latter explanation is more likely in this case. Although general MMR is thought to be a particularly important process in many species (e.g., defects in MMR lead to higher rates of colon cancer in humans), the absence of MMR can have advantages too. Theory has suggested that elevated mutation rates might be adaptive in unstable changing environments (10,40). Recently it has been shown that strains defective in MMR can outcompete relatives with normal MMR (10,41). In addition, many strains of *E. coli* and *Salmonella* isolated from the environment are defective in MMR (8). Thus the species that have lost their MMR genes may have done so at a time when it was advantageous to have a higher mutation rate. In particular, absence of MMR would result in a very high mutation rate in microsatellite sequences, a process thought to be particularly important in generating diversity in antigen proteins of species like *M. tuberculosis* (42) (which is one of the species without MMR genes). In addition, since MMR plays a role in other processes such as the regulation of interspecies recombination, differences in MMR could also affect these processes (43).

Since the MutS and MutL gene families only show limited similarity to other proteins it is not possible to infer their origin. However, the origins of the MutH-based strand recognition system are quite revealing. The limited distribution of MutH homologs supports evidence that only close relatives of *E. coli* use methyl-directed strand recognition. Interestingly, MutH is closely related to the restriction enzyme Sau3A from *S. aureus* (44). Perhaps the *mutH* methylation based system evolved from a restriction modification system. It is possible that other species have co-opted separate restriction systems for strand recognition, although we do not yet have any candidates. Any protein sensitive to methylation state could potentially serve in recognition of newly replicated strands. This may explain why many species encode a Dam homolog but not a MutH homolog. Possibly related to this, the Vsr mismatch endonuclease, which is involved in mismatch repair of GT mismatches, also has many functional and structural similarities to restriction enzymes (45). The Vsr system also appears to be of recent origin in the Proteobacterial lineage.

Nucleotide Excision Repair

Nucleotide excision repair is a generalized repair process that allows cells to remove many types of bulky DNA lesions (46,47). As with MMR, the overall scheme of NER is highly conserved between species. The general scheme works in the following way: recognition of DNA damage; cleavage of the strand containing the damage (usually on both the 5' and 3' sides of the lesion); the removal of an oligonucleotide containing the damage; the resynthesis of a repair patch to fill the gap; followed by ligation to the contiguous strand at the end of the gap. As with MMR, NER can repair many types of DNA lesions because the recognition step is very broad. However, unlike MMR, the biochemical details of the NER process are quite different between bacteria and eukaryotes. Therefore we have divided the analysis into two sections, one focusing on proteins shown to be involved in NER in bacteria and the other on proteins involved in NER in eukaryotes.

Bacterial NER – UvrABCD pathway

As with many other repair processes, NER in bacteria has been most thoroughly studied in *E. coli.* In *E. coli,* four proteins form the core of the NER process: UvrA, UvrB, UvrC, and UvrD. The details of the functions of each of these proteins have been reviewed elsewhere (2,48). In summary, a homodimer of UvrA initially recognizes the putative lesion and recruits UvrB to aid in the recognition and verification that a lesion exists. UvrA leaves the site and UvrB then recruits UvrC revealing a cryptic endonuclease activity to produce dual incisions 12-13 nucleotides apart bracketing the lesion. The UvrD helicase in concert with DNA polymerase I removes the damaged oligonucleotide and completes a repair patch that is then sealed into place by DNA ligase to restore the intact DNA. At least one additional protein, Mfd, has a particularly important role in NER. It is involved in targeting NER to the transcribed strand of actively transcribing genes – a subpathway known as transcription coupled repair (TCR) (49,50). Studies in a variety of other species have shown that the roles of the UvrABCD proteins are highly conserved among bacteria – homologs of these proteins are required for NER in these species. Homologs of Mfd have been cloned from many species but its function has only been studied in one species other than *E. coli* – *B. subtilis*. In *B. subtilis*, Mfd is also required for TCR, although it may also have some role in recombination (51,52).

Our comparative analysis shows that orthologs of the UvrA, UvrB, UvrC and UvrD proteins are found in all the bacterial species analyzed (Table 4). Since these genes have conserved function, it is likely that all these bacteria can perform NER. The correlated presence/absence of all four proteins suggests that all species with these proteins perform excision repair in much the same way as does *E. coli*. In addition, somewhat surprisingly, orthologs of UvrABCD are also found in the Archaea *M. thermoautotrophicum*. Since UvrABCD are present in all bacteria, we infer that these genes were present in the common ancestor of all bacteria. The presence of these genes in one Archaea can either be explained by lateral transfer to this Archaea or by an origin prior to the divergence of Archaeal and bacterial ancestors, with subsequent gene loss from some Archaeal lineages. At this time we do not have enough evidence to determine which occurred. Interestingly, these genes are present together in the same region of the genome in *M. thermoautotrophicum* but not in most bacteria. Orthologs of Mfd are found in all bacteria except the mycoplasmas and *A. aeolicus.* Therefore Mfd likely also originated near the beginning of bacterial evolution and was then lost from the mycoplasma lineage and the *A. aeolicus* lineage. Since Mfd is absolutely required for TCR in *E. coli* and *B. subtilis* it is likely that the species without Mfd cannot perform TCR.

The specific origins of each of these proteins individually are interesting and help understand the origins of the bacterial NER process. UvrA is a member of the ABC transporter family of proteins (53,54) and thus it likely originated by a gene duplication from an ancestral ABC transporter. The ABC transporter family includes proteins involved in transmembrane transport of many types of molecules including various toxins (e.g., the multi-drug receptor (MDR) proteins). Perhaps the NER system evolved from a system for removing DNA damaging agents from the cell. Although UvrA has not been shown to have transport activity, given that all other characterized ABC transporter family members are transporters, it would be useful to determine whether UvrA is additionally involved in transporting damaged DNA fragments out of the cell. Alternatively, UvrA may have lost its transport activity. Even in this case, the relationship to ABC transporters may help understand the means by which NER is putatively associated with the bacterial membrane (55). In addition, it may also explain why a UvrA homolog in *Streptomyces peucetius* is responsible for resistance to daunorubicin and possibly transport of this antibiotic out of the cell (56). UvrB is a member of the helicase superfamily of proteins. It is particularly closely related to the Mfd and RecG proteins. It appears that early in bacterial evolution, there were multiple duplications of an ancestral gene that gave rise to the UvrB, Mfd and RecG proteins. The relationship of UvrB and Mfd is of interest in the sense that both interact with UvrA. UvrD is also a member of the helicase superfamily, although its origins can be traced to a different helicase family than that of UvrB and Mfd. UvrD is part of a subfamily that includes the RecB, rep, and helicase IV proteins of bacteria and RadH from yeast. UvrC shows some sequence similarity to Ercc1 which has a similar function in eukaryotic NER to that of UvrC (Ercc1 is part of a heterodimer that performs the 5’ incision for NER in eukaryotes). However, the sequence similarity is limited, and many other proteins, including proteins in the ligase (familyI) and RadC families, also share the same motif. Thus UvrC and Ercc1 are probably not homologs. However, our analysis suggests that UvrC may share a common ancestry with homing endonucleases from mitochondrial introns. UvrC shares extensive sequence similarity with these genes in regions of the protein covering more than the motif shared with Ercc1. In addition, *E. coli* and many other bacteria have additional open reading frames (with no known function) similar to UvrC and the homing endonucleases. Thus the genes involved in NER in bacteria have ancient origins in that they are all part of multigene families that are likely very old. However, our analysis shows that the origins of these particular genes, by gene duplications within these multigene families, occurred separately from the origins of the genes involved in eukaryotic NER indicating a separate origin of these similar processes (see below).

Eukaryotic NER – XP pathways

NER in eukaryotes has been most thoroughly studied in yeast and humans. Within eukaryotes, the process is highly conserved (47,57). Interestingly the biochemical reactions are nearly identical to those in bacterial NER but many more proteins are needed to carry them out. Thus one of the obvious differences between bacterial and eukaryotic NER is complexity. In humans, multiple proteins are involved in the initial damage recognition steps, including XPA, RPA, XPE and XPC. The helicase activities are provided by those of XPB and XPD in the basal transcription factor TFIIH, that interestingly serves dual functions in transcription and NER. In the latter, it forms a bubble to enable separate flap endonucleases XPG and XPF-ERCC1 heterodimer to produce incisions 3' and 5' of the lesion, respectively, about 30 nucleotides (or three turns of the helix) apart. Repair replication is then carried out by the same proteins required for genomic replication, namely RPA, RFC, PCNA and DNA polymerase δ/ε. The basic NER system of yeast works in the identical manner as the human system described above. Although NER is generally conserved among eukaryotes, some major differences among eukaryotes exist in targeting NER to particular parts of the genome. For example, in humans the CSA and CSB proteins are involved in targeting NER to the transcribed-strands of transcribed genes (the TCR process mentioned above), but yeast only encodes an ortholog of CSB. Similarly, in yeast, Rad7 and Rad16 are involved in targeting repair to non-transcribed regions but no orthologs of Rad7 or Rad16 have yet been found in humans. Conversely, XPC in humans is required for global genome repair but the homolog of XPC in yeast, Rad4, does not appear to have a similar function. There are also more subtle differences in targeting lesions between humans and rodents. In particular, humans and rodents are nearly identical in the repair of 6-4 photoproducts but rodents do not carry out efficient global repair of CPDs as well as humans, evidently because they lack inducible up-regulation of NER.

Our comparative analysis reveals that most of the proteins involved in NER in eukaryotes are of eukaryotic origin - no homologs are found in bacteria or Archaea. Of the genes that have homologs in bacteria or Archaea, most are part of large multigene families and most of these arose by gene duplications within the eukaryotic lineage and thus are clearly not orthologs of any bacterial or Archaeal proteins. For example, the CSB gene (required for TCR) arose by a gene duplication within the SNF2 family of helicases during eukaryotic evolution (19). The differences between eukaryotic and bacterial NER are most striking when comparing the genes involved in particular steps in the process. As mentioned above, many proteins are involved in damage recognition in eukaryotes (including XPA, the three subunits of RPA Rad23, XPC, and XPE) and not one of these is homologous to the bacterial damage recognition protein UvrA. Thus the damage recognition processes for NER in eukaryotes and bacteria are of distinct origins. Similarly, the early initiation steps in eukaryotes require many proteins, in particular those that make up the basal transcription factor TFIIH, yet none of the proteins in TFIIH share a direct common ancestry with any of the bacterial NER proteins. Interestingly, as noted above, TFIIH includes the 5’-3’ and 3’-5’ helicases encoded by XPB and XPD, respectively, so the functional role of TFIIH can be considered analogous to that of UvrB in bacterial NER. TFIIH might also serve the corresponding role of the UvrD helicase in removing the damaged oligomer but this has not been established. Even the exonucleases used to cut 5’ and 3’ to the site of damage are not homologous with those of bacteria. The TCR processes also have a separate origin between bacteria and eukaryotes. Even though Mfd and CSB are both part of the helicase superfamily of proteins – they are only distantly related within this superfamily. Interestingly, although these proteins are distantly related, they are both examples of proteins with the helicase motifs that do not have helicase activity. Thus the presence of "helicase" motifs cannot be used to predict the presence of helicase activity.

Only three of the eukaryotic NER proteins have likely orthologs in bacteria or Archaea - Rad1, Rad2, and Rad25. Rad1 and Rad2 are also found in all Archaea suggesting that these genes originated in a common ancestor to Archaea and Eukaryotes. Their function in Archaea is not known. However, it is important to note that Rad1 has been found to have roles in recombination so its presence in Archaea does not imply that Archaea have a limited NER system. In addition, Rad1 works in concert with Rad10 (as do the homologs of these in eukaryotes ERCC1 and XPF) but no Rad10 homologs are found in any of the Archaea. Rad2 is a member of the FEN1 family of endonucleases (58) with diverse functions, so it is possible that the use of Rad2 in NER evolved within eukaryotes even if the gene duplication that gave rise to the Rad2 lineage occurred prior to the divergence of Archaea and eukaryotes. The presence of Rad25 in some Archaea and some bacteria is perplexing. It is possible that this represents a case of lateral transfer.

There are only a very few limited links one can make between the proteins involved in NER in eukaryotes and those in bacteria. ERCC1 has been found to share a motif with UvrC but this motif is also in many other proteins and it may not even represent common ancestry. XPD is a distant relative of DinG, which is an SOS induced gene, with unknown function. Overall, we conclude that the eukaryotic NER system evolved within the eukaryotic lineage. Despite the differences between NER in bacteria and eukaryotes, these processes are still similar in general scheme. Thus they appear to be analogous systems, having evolved independently in two separate lineages. NER has not yet been characterized in any Archaea, although Archaea do have some form of dark repair that may be NER (59)(Eisen et al., unpublished). With the separate origin of the bacterial and eukaryotic NER systems it is possible that Archaea have also evolved their own NER system. Thus it would be useful to conduct genetic studies of NER in Archaea to see what genes might be involved. In addition, it would be interesting to characterize NER in *M. thermoautotrophicum* which encodes UvrABCD homologs in its genome.

Base Excision Repair (BER)

In BER, damaged or altered bases are detached from the DNA backbone by DNA glycosylases that cleave the glycosylic bond. Subsequently the backbone of the DNA is incised by an abasic-site endonuclease, the sugar is removed, and a single nucleotide repair patch is synthesized using the base opposite the excised base as a template. Sometimes the patch extends a few more nucleotides. For a review of BER see (60). In this section, we discuss the evolution of different DNA base glycosylases. The evolution of abasic-site endonucleases is discussed in the following section since they are also involved in other repair pathways.

Uracil DNA glycosylase (UDG or UNG)

Uracil can appear in DNA by two routes – incorporation during replication and by spontaneous deamination of cytosine. In the replication pathway, U is sometimes incorporated because of the dUTP pool that coexists with dTTP. This is partially controlled through dUTPase activities. Thus, to limit the amount of U incorporation, cells have pathways that minimize the amount of dUTP present. A likely more harmful source of uracil in DNA is deamination of cytosine. The deamination of cytosine to uracil is potentially mutagenic because replication will lead to an A being incorporated opposite the U, rather than the G that should have been incorporated opposite the C. Removal of U's from DNA is thus a way to reduce mutations due to deamination. Uracil DNA glycosylases have been cloned and characterized from many species. In *E. coli,* one protein Ung is the major uracil DNA glycosylase. Ung homologs have been found in many bacterial and eukaryotic species, as well as in many viruses (mostly herpes related viruses) and these proteins have strikingly similar structures and functions as their *E. coli* counterpart. A variety of other proteins have also been shown to possess uracil-DNA glycosylase activity, including GAPDH (glyceraldehyde-3-phosphate dehydrogenase (61), a cyclin like protein (62), and even the GT mismatch repair protein. However, most studies suggest that the major activity for all species is provided by Ung homologs (63).

Our comparative analysis shows that Ung homologs are found in many, but not all bacteria, but that they are not found in any of the Archaeal complete genome sequences. Within the bacteria, we believe that the absence of Ung homologs (from *T. pallidum, Syn. sp* , and *A. aeolicus*) is due to gene loss and that Ung was present in the common ancestor to all bacteria. However, we cannot resolve whether Ung was present in the common ancestor of all species because it is possible that the eukaryotic Ung is of mitochondrial origin. The mitochondrial origin of the eukaryotic Ung homologs is suggested by a few lines of evidence. First, the human Ung does function in the mitochondria (in an alternatively spliced form) (64). In addition, phylogenetic trees of all Ung proteins suggest that the eukaryotic Ung homologs evolved from within the bacterial lineage. However, since no Ung sequence is yet available from the α-Proteobacteria which are thought to be the closest living relatives of the mitochondria, we cannot conclusively resolve the origin of the eukaryotic Ung genes. If Ung is ancient then the absence from the three Archaea must be due to gene loss sometime in their history. Alternatively, the Archaea may never have encoded an Ung homolog if Ung evolved within bacteria and was laterally transferred to eukaryotes.

Due to the high degree of functional conservation among Ung homologs, it is likely that the species with Ung homologs have uracil glycosylase activity. However, the absence of an Ung homolog does not mean the absence of uracil glycosylase activity, because, as discussed above, many proteins have some uracil glycosylase activity. The absence of uracil glycosylase activity would be particularly surprising in thermophiles since the deamination of cytosine is strongly temperature dependent (it increases with increasing temperature). Perhaps these species have a novel means of preventing or limiting deamination. *In vivo* studies have found uracil glycosylase activity in many thermophiles (although not the species examined here) (65). Thus clearly some thermophiles are able to remove uracil from their DNA.

Nth-MutY family

The MutY and Nth proteins of *E. coli* are related to each other, and homologs of these proteins make up the MutY-Nth family. MutY and Nth (short for endonuclease III) are both DNA-glycosylases and both are involved in BER. Although they share many sequence and structural features (66,67), they have quite different substrate specificity and cellular functions. MutY cleaves the glycosylic bond of A from G:A, C:A, 8-oxo-G:A or 8-oxoA:A. Thus, one of MutY's primary roles is protection against mutations due to oxidative damage of G's. Nth has a very broad specificity and excises a variety of damaged pyrimidines. In addition, it also has β-lyase activity. Members of the MutY-Nth gene family have been cloned from many species. All of these have been shown to have some type of DNA glycosylase activity, however the specificity varies enormously. Examples include the pyrimidine dimer glycosylase of *Micrococcus luteus* (68), the “MutY” of mammals (which has similar activity to the *E. coli* MutY (69,70), “Nth” homologs in mammals (71), the yeast NTG1 and NTG2 (which excise similar substrates to the *E. coli* Nth as well as ring opened purines, the formamidopyrimidines (FAPY)), and the GT mismatch repair enzyme of the Archaea *M. thermoformicum* (72).

Our comparative analysis shows that all species except the two *Mycoplasma* species encode at least one member of the MutY-Nth gene family. We attempted unsuccessfully to use phylogenetic analysis to divide this gene family into subfamilies of orthologs. The evolutionary trees of the MutY-Nth family are ambiguous. Some proteins are clearly more related to MutY or to Nth than others, but there is no obvious, well-supported subdivision. Since there is a great deal of diversity of function in this gene family, and since the trees were ambiguous, we list the MutY-Nth gene family together without attempting to distinguish orthologs of these two genes. Therefore it is not possible to make any specific predictions about activities for any species or to characterize the evolution of this gene family in detail. Since this family is so widespread, we conclude that it is ancient and thus that a common ancestor of all life had a MutY-Nth like protein, or possibly two. However, since the activity is not conserved among these proteins we cannot infer any activity other than a broad "glycosylase" activity for the ancestral protein. Interestingly, the MutY-Nth family is distantly related to the Ogg and AlkA glycosylases (see below). Thus all three of these gene families likely descended from a single ancestral glycosylase gene.

Fpg-Nei family

The Fpg glycosylase (also known as MutM) in *E. coli* excises damaged purines from dsDNA (including 8-oxoG and FAPY). Its primary function is the protection against mutation due to oxidative DNA damage (8-oxoG is mutagenic). Homologs of Fpg have been isolated from a variety of bacterial species and these have functions similar to that of the *E. coli* protein (73,74). Somewhat surprisingly, when the Nei protein was cloned, it was found to be a homolog of Fpg (75,76). Nei is a glycosylase that excises thymine glycol and dihydrothymine. Thus the Nei-Fpg family has a great deal of functional diversity, while exhibiting a common theme of the repair of DNA damage due to reactive oxygen species.

Our comparative analysis shows that although members of the Fpg-Nei family are found in many bacterial species, they are not found in Archaea or eukaryotes. Therefore this family is of bacterial origin. Our phylogenetic analysis of the members of this family has allowed us to divide it into clear Fpg and Nei orthologous groups (therefore they are listed separately in Table 4). Of the proteins in the family, most are orthologs of Fpg. The distribution of Fpg homologs suggests that Fpg was present in the ancestor of most bacteria. Therefore, the species without Fpg (*H. pylori*, the spirochetes and *A. aeolicus*) likely lost this gene in their history. Since Fpg proteins have similar activities between species, the presence of an Fpg homolog likely indicates the presence of FAPY-and 8-oxoG glycosylase activity. The origin of Nei is somewhat less clear. Only one species other than *E. coli* (*M. tuberculosis)* has been found to encode a likely ortholog of Nei. It is not possible to determine if there was a lateral transfer between these two lineages or if there was a gene duplication in the common ancestor and subsequent gene loss of Nei from many species.

Ogg1 and 2

The Ogg1 and Ogg2 proteins of yeast act as 8-oxoG glycosylases (77). Ogg1 excises 8-oxoG if it is opposite C or T and Ogg2 if opposite G or A. Initially, it was reported that Ogg1 and Ogg2 were both homologs of Fpg. Although these proteins have similar substrate specificity, and are both β-lyases like Fpg, in fact they are not homologs of Fpg. They do show some very limited sequence similarities to the MutH-Nth family of proteins and may actually represent very distant evolutionary relatives of the MutY-Nth family. Orthologs of Ogg1 and Ogg2 have been cloned from humans (78-80) and some isoforms of these function in the nucleus and others in the mitochondria (81). Our comparative analysis reveals that a homolog of Ogg1 is present in *M. thermoautotrophicum*, but not in any other Archaea or bacteria. Thus, either Ogg is of eukaryotic origin and *M. thermoautotrophicum* obtained its Ogg protein by lateral transfer or Ogg originated prior to the divergence of Archaeal and eukaryotic ancestors and then was subsequently lost from some Archaeal lineages.

Alkylation glycosylases

Alkylation glycosylases can be divided into three gene families. One includes AlkA of *E. coli* (also known as TagII) and MAG of yeast. Another includes TagI of *E. coli*. the third includes MPG of mammals. All of these proteins have glycosylase activity for some type or types of alkylated base. TagI of *E. coli* is highly specific for 3-meA, although it can also remove 3-meG, but with much lower efficiency. AlkA has a much broader specificity than TagI and it can excise many alkyl-base lesions (e.g., 3-meA, 3meG, 7meG, and 7meA), and a variety of other damaged bases including hypoxanthine. The MAG protein of yeast is a homolog of AlkA and has a similar broad specificity. The mammalian MPG protein is not similar in sequence to either AlkA or TagI but it also has quite a broad specificity.

Our comparative analysis shows some interesting patterns for alkylation glycosylases. TagI is only found in bacteria (and only in a limited number of species). Thus it likely evolved within bacteria. AlkA homologs are found in bacteria, Archaea, and eukaryotes, although many species do not encode homologs. Thus, we believe AlkA is an ancient protein. The origin of MPG is not clear since it is found in many eukaryotes (including many species not listed in Table 4) and *B. subtilis*. Perhaps the *B. subtilis* protein was laterally transferred from eukaryotes. Many species do not encode a homolog of any of the known alkyl-base glycosylases (the mycoplasmas, the spirochetes, *A. aeolicus*, *N. gonorrhoeae* and *M. jannascii*). Given the number of proteins that exhibit this activity, however it is possible that these species have proteins with the activity that have not yet been characterized. Even if they do not have alkyl-base glycosylase activity, since alkylation damage can be repaired by other pathways (e.g., NER and alkyltransferases (although the alkyltransferases only repair O-6-meG)) these species may be adequately protected. In some cases it may turn out that the species that have lost a particular glycosylase do not live in environments in which the particular substrates for those repair enzymes are prevalent.

T4 Endonuclease V

T4 phage encodes its own unique DNA-glycosylase, known as endonuclease V or DENV. This protein acts specifically on UV irradiation induced CPDs as possibly a back-up system for the hosts NER enzymes. Interestingly, DENV can functionally complement any mutants in bacteria or eukaryotes with deficiencies in the early steps in NER. Homologs of DENV have been cloned in a paramecium virus and phage RB70, but the activities of these are not known. DENV homologs are not present in any of the complete genome sequences but they have been found in two other viruses.

AP Endonucleases (Abasic site endonucleases)

AP endonucleases are required for the BER process and also for other processes. They cleave the DNA backbone at sites at which bases are missing (see (82) for review).

There are at two distinct families of AP endonucleases. One includes the Xth protein of *E. coli* , RRP1 of *D. melanogaster*, and the APE1/BAP1/HAP1 proteins of mammals (83). The other family includes the Nfo protein of *E. coli* and the APN1 proteins of yeast. Some other proteins can serve as AP endonucleases, but usually as part of base-glycosylase (e.g., Nth and DENV have AP endonuclease activity).

Our comparative analysis shows that members of the Xth/APE1 family are found in almost every species (with the exception of the two mycoplasmas and *M. jannascii*). Thus, Xth/APE1 is probably an ancient protein and the absence from these species is likely due to gene loss. Members of the Nfo/APN1 family have a more limited distribution, although members are found in all domains of life, suggesting that these proteins are also ancient. Thus the species that are missing either Nfo or Xth very likely have lost these genes sometime in their history. All species encode a homolog in at least one of the two families. Thus while there have been gene losses of AP endonucleases, no species has lost both AP endonuclease genes. Thus AP endonuclease activity is universal. This is not surprising in view of the high frequency of spontaneous depurination of DNA.

Recombination and Recombinational Repair

Homologous recombination is required for a variety of DNA repair and repair related activities (1,84,85). Before discussing the role of homologous recombination in repair, it is useful to review some of the details of homologous recombination in general. Homologous recombination can be divided into four main steps: (1) initiation (during which the substrate for recombination is generated); (2) strand pairing and exchange; and (3) branch migration and (4) branch resolution. In many species, there are multiple pathways for homologous recombination, although there is frequently a great deal of overlap among the pathways. Pathways often differ in the mechanism of (and proteins used for) initiation, but then use the same mechanism (and proteins) for the pairing and exchange step. In some species, there are also multiple pathways for the branch migration and resolution steps. For example, in *E. coli*, there are at least four pathways for the initiation of recombination - the RecBCD, RecE, RecF, and SbcCD pathways. All of these pathways generate substrates that are used by RecA to catalyze the pairing and exchange steps. The branch migration and resolution steps can be carried out by at least three pathways - the RuvABC, RecG and Rus pathways. Thus homologous recombination in *E. coli* revolves around RecA but has many ways that it can feed into and out of the recombinase step.

Homologous recombination is required for the repair of a few different types of DNA lesions. Perhaps it is best understood in its role in the repair of double-strand breaks. Double-strand breaks (DSBs) can be caused by many agents including reactive oxygen species, restriction enzymes and normal cellular processes like VDJ recombination in mammals. Such DSBs can be repaired through homologous recombination with intact chromosomes. In addition, DSBs can also be repaired without the use of homologous recombination in a process termed non-homologous end joining (NHEJ) (discussed in more detail below). In *E. coli* and yeast, the majority of the repair of DSBs is carried out by homologous recombination pathways. Yeast are able to perform double-strand end repair but this process has a limited role in the repair of DSBs. In contrast, in humans, most of the repair of DSBs is carried out by non-homologous end joining, although some homologous recombination based repair is also performed.

Another type of DNA abnormality that can be repaired by homologous recombination is the post-replication daughter strand gap. When DNA is being replicated, if the polymerase encounters a DNA lesion, it has three choices - replicated the DNA anyway, and risk that the lesion might be miscoding, stop replication and wait for repair, or leave a gap in the daughter strand and continue replication a little but further downstream. In *E. coli*, the choice depends on the type of lesion, but frequently gaps are left in the daughter strand. In such cases, it is no longer possible to perform excision repair on the lesion because there is no intact template to allow for the repair synthesis step. However, such daughter strand gaps can be repaired by homologous recombination in a process known as daughter-strand gap repair (DSGR) (86). DSGR uses an undamaged homologous section of DNA to provide a patch for the unreplicated daughter strand section. Thus, although DSGR does not remove the instigating DNA damage, it is still a form of DNA repair. Once the recombinational strand exchange has taken place, then excision repair processes may be able to deal with the original lesion.

Homologous recombination can also be used to repair some other types of DNA abnormalities including interstrand cross-links. In all cases, as with recombination in general, the repair processes that use homologous recombination differ from each other mostly in the initiation steps. Below we discuss different pathways for homologous recombination, focusing on those known to be involved in some type of DNA repair. In the Table, and below, the proteins are categorized by the stage in which they participate in the recombination process.

Recombination I: Initiation Pathways

RecBCD pathway (DSBR initiation in bacteria)

The primary pathway for the initiation of homologous recombination in *E. coli* is the RecBCD pathway (see (87) for review). This pathway is used for the majority of chromosomal recombination (such as during Hfr mating) and for DSBR. The initiation steps for this pathway require primarily the RecB, RecC and RecD proteins, although other proteins such as PriA may also be required. Together, RecB, RecC and RecD make up an exonuclease/helicase complex that is used to assemble a substrate for RecA-mediated recombination. Functionally similar complexes have been described and isolated from many bacterial species. Many of these are composed of proteins not homologous to RecB, RecC or RecD (88). Only a few complexes homologous to the RecBCD complex have been described.

Our comparative analysis shows a limited distribution of RecB, RecC and RecD orthologs (they are only found in some enterobacteria, *M. tuberculosis*, and possibly in *B. borgdorferi*). Based on this, we conclude that the RecBCD pathway has evolved relatively recently within bacteria. The finding that particular species either have orthologs of all three or of none of these proteins suggests that these proteins have a conserved affiliation with each other. Analysis of the individual proteins suggests that this complex may have an ancestry in recombination and repair functions. RecB and RecD are both in the helicase superfamily of proteins and both are closely related to proteins with recombination or repair roles (RecB is related to UvrD proteins and the AddA proteins of gram-positive bacteria, RecD is related to the TraA proteins involved in DNA transfer in *Agrobacterium tumefaciens* (see (89,90) for review). The timing of the origin of the RecBCD pathway is somewhat ambiguous. The pathway could have evolved within the Proteobacteria and *M tuberculosis* could have received it by lateral transfer. Alternatively, the pathway could have been present in the common ancestor of highGC Gram-positive species and Proteobacteria. If that were true then the absence of the pathway in many Proteobacteria and possibly the lowGC Gram positive species would have to be due to gene loss. Nevertheless, since RecBCD orthologs are not found in many deep branching bacterial species, it is likely that this pathway is a recently evolved system.

RecF pathway - DSGR initiation in bacteria

In *E. coli*, the RecF pathway is responsible for most plasmid recombination, for daughter-strand gap-repair, for some replication related functions (91) and for a process known as thymineless death (92,93). This pathway has only a limited role in "normal" homologous recombination accounting for less than 1% of the recombination in *E. coli*. The proteins involved in recombination initiation in this pathway are RecF, RecJ, RecN, RecO, RecR and RecQ (see (85) for more detail about what each of these proteins do). Interestingly, RecQ was originally isolated as a mutant of *E. coli* that was resistant to thymineless death (92). However, not all of the genes in the RecF pathway are required for every function of the pathway. For example, RecF, RecJ and RecQ are required for thymineless death while RecF, RecR, and RecO are evidently required for replication restart functions (91). In addition, some of the genes in this pathway are involved in other repair pathways. For example, RecJ can be used as an exonuclease in MMR if other exonucleases are defective.

Homologs of some of the proteins in the RecF pathway have been characterized in a variety of species. RecF and RecJ homologs in many bacteria have similar functions to the *E. coli* proteins. RecQ homologs have been characterized in many eukaryotic species. The yeast RecQ homolog SGS1 is involved in the maintenance of chromosome stability, possibly through interaction with topoisomerases during recombination (94). Humans encode at least three RecQ homologs. Werner's syndrome is caused by a defect in one of these (95) and Bloom's syndrome is caused by a defect in another (96). Many of the eukaryotic RecQ homologs have been shown to be helicases (97,98), like the *E.*

coli RecQ, but their cellular functions are not known.

Our comparative analysis of proteins in the RecF pathway was somewhat limited by difficulty identifying orthologs of some of the proteins in the pathway, mostly because the degree of conservation among some of the proteins, even between close relatives was somewhat low. RecF and RecN are both part of the SMC family of proteins (99,100) and almost all species encode some members of this gene family. Therefore homologs of RecN and RecF are found in most species, and it was necessary to attempt to divide this gene family into groups of orthologs. Identification of RecF orthologs was relatively simple since the degree of conservation among RecF proteins is quite high. However, RecN proteins are less conserved and we were unable to distinguish whether the RecN-like proteins of the *Mycoplasmas* and *B. borgdorferi* were RecN orthologs or paralogs. RecO proteins were also poorly conserved. Since RecO is not part of a large protein family, we were able to use more liberal motif searches to search for RecO homologs, but we still did not find likely RecO homologs in many species. Our analysis shows that, in striking contrast to the RecBCD pathway, the proteins in the RecF pathway do not have perfectly correlated distribution patterns. For example, orthologs of RecF are not found in *N. gonorhoeae* and *H. pylori* while orthologs of RecJ are. In addition, eukaroytes encode orthologs of RecQ but not of any other proteins in this pathway. Thus in other species, if they do have a RecF-like pathway, it cannot work the same way as it does in *E. coli.* It is possible that similar pathways exist in other species but that they have coopted alternative proteins for some of the functions. For example it is known that some of the functions of RecJ in *E. coli* can be complemented by other 5’ exonucleases such as RecD. Perhaps, as with MMR, the exact details of the RecF pathway are not conserved between species but the general scheme is. Thus species without orthologs of certain genes in the RecF pathway may use other genes to carry out those functions. Alternatively, it is possible that the functions of the RecF pathway are specific for *E. coli* and that other species do not have a similar pathway. The RecF example illustrates the limitation that the presence/absence information cannot always be used to reliably predict the capabilities of a particular species.

RecE pathway – alternative initiation pathway in bacteria

The RecE recombination pathway of *E. coli* is only activated in *recBrecC, sbcA* mutants. This pathway requires many of the proteins in the RecF pathway, as well as two additional proteins RecE and RecT (101-103). These additional proteins are both encoded by a cryptic lambda phage. RecE is an exonuclease that can generate substrates for recombination either by RecT or by RecA. RecT may be able to catalyze strand invasion without RecA (104). Our comparative analysis shows that the species distribution of these proteins extremely limited. RecT is found in some lowGC gram positive bacteria. RecE is not found in any species other than *E. coli.* The presence of these genes on a cryptic phage may reflect a recent lateral transfer between species.

SbcBCD pathway

The SbcB, SbcC and SbcD proteins were all identified as genes that, when defective, led to the suppression of the phenotype of recBC mutants (see (85) for review). SbcB is an exonuclease (also known as exonuclease I, exoI, of Xon). When it is defective, the RecE and RecF pathways are revealed. SbcC and SbcD together make up an exonuclease that cleaves hairpin structures and thus functions to eliminate long cruciform or palindromic sequences and thereby remove sequences that may interfere with DNA synthesis (105). Homologs of SbcC and SbcD have not been characterized in many bacteria. However, these proteins do share some sequence similarity to Rad50 and MRE11 and may be homologs of these proteins (discussed below) (106).

Our comparative analysis shows that SbcB homologs are found only in *H. influenzae* and thus this protein apparently originated within the gamma-Proteobacteria. Homologs of SbcC and SbcD are present in many bacteria and are always present together. Thus the interaction of these proteins appears to have been conserved over time. Given the likely homology of these proteins to MRE11/ RAD50 (which are found in eukaryotes and Archaea) we believe SbcC and SbcD are ancient proteins. Thus, the species missing SbcC and SbcD homologs likely lost these genes during their evolutionary history. In addition, since the function of MRE11/Rad50 is similar to that of SbcCD it is likely that the species with SbcCD homologs have similar activities.

Rad52 pathway - DSBR in eukaryotes

The primary pathway for homologous recombination in yeast is the Rad52 pathway (107). This pathway is used for mitotic and meiotic recombination as well as for double-strand break repair. Although the exact biochemical details of this pathway are not completely worked out, the initiation step depends on three proteins - MRE11, Rad50 and XRS2 which form a distinct complex. The exact biochemical activity of this complex is not well characterized but clues have come from the identification of Rad50 and MRE11 as distant relatives of SbcC and SbcD (106). The MRE11-Rad50-XRS2 complex has exonuclease activity and MRE11 can act as an exonuclease on its own (108) in certain conditions. No function is known for the XRS2 protein. It is believed that the MRE11-Rad50-XRS2 complex functions to induce DSBs for mitotic and meiotic recombination and that it may alter other DSBs to allow them to be repaired by homologous recombination. Interestingly, genetic studies have found that the MRE11-Rad50-XRS2 genes are also involved in the non-homologous end-joining pathway (see below).

Homologs of MRE11 and Rad50 have been identified in humans as part of a five protein complex (108). This complex has some similar activities as the yeast MRE11-Rad50-XRS2 complex. As with the yeast complex, this complex is also likely involved in recombination and DSBR (109) although, as mentioned above, much of the repair of DSBs in mammals is carried out by NHEJ. It is not known whether the human complex plays any role in non-homologous end joining. The human MRE11 is a 3'-5' exonuclease (108). No homolog of XRS2 has yet been identified in humans. Interestingly, defects in one of the other proteins in the human complex (NBS1) lead to the Nijmegen breakage syndrome (109). Thus there are some significant differences between the human and yeast complexes.

Our comparative analysis shows that homologs of MRE11 and Rad50 are found in all the Archaea analyzed (although the Archaeal Rad50 homologs may not be orthologs of the eukaryotic Rad50s – our phylogenetic analysis was ambiguous). Since these genes are related to the SbcC and SbcD genes of bacteria, we conclude that the SbcC/Rad50 and SbcD/MRE11 proteins are ancient proteins.

Recombination II: Strand Pairing and Recombinases

The RecA protein of *E. coli* is the recombinase for all homologous recombination pathways. Thus RecA is absolutely required for homologous recombination in *E. coli*. Comparative studies have revealed that homologous recombination depends on RecA homologs in many other bacterial species as well as in Archaea and eukaryotes (110-112). The comparison of RecA homologs between species is somewhat complicated by the fact that many species encode multiple homologs of RecA. Our phylogenetic analysis suggests that there was an ancient duplication in this gene family into two lineages. Those in one lineage are recombinases (bacterial RecA, Archaeal RadA, eukaryotic Rad51 and DMC1) and those in the other lineage have alternative roles (bacterial SMS, Archaeal RadC, eukaryotic Rad55 and Rad57). In Table 4, we only list the number of genes a species encodes in the RecA lineage. We refer to these as orthologs of RecA.

Our comparative analysis shows that all species for which complete genomes are available encode orthologs of RecA. RecA is the only repair gene for which homologs are found in all the species analyzed. Since all characterized genes in this lineage are recombinases, this suggests that all these species have recombinase activity. The universal presence in these species suggests that recombinase activity is fundamental to life. However, there have been reports of some mycoplasma species encoding defective RecA proteins (and possibly thus being defective in all homologous recombination). The presence of multiple orthologs of RecA in eukaryotes is likely due to a duplication early in eukaryotic evolution (112). These genes have diverged somewhat in function. Rad51 is the recombinase for the majority of mitotic recombination. Both Rad51 and DMC1 are used for aspects of meiotic recombination. Thus both Rad51 and DMC1 are recombinases. Some bacteria also encode multiple RecA orthologs (e.g., *Myxococcus xanthus*). The functions of these two are not well understood (113). Interestingly, phage T4 encodes a RecA homolog, UvsX, which also has recombinase activity (114,115). The origins of this gene are unknown.

Recombination III: Branch Migration and Resolution

In *E. coli*, at least three pathways have been identified that can perform branch migration and resolution: RuvABC, RecG and Rus. The RuvABC pathway may be the main branch migration and resolution pathway. It works in the following way: RuvA binds to Holliday junctions, RuvB is a helicase that catalyzes branch migration and RuvC is a resolvase. The RecG protein catalyzes branch migration and Holliday junction resolution (116). RusA is a Holliday junction resolvase that is normally suppressed (117-119). It is encoded by a defective prophage DLP12 and is similar to protein in phage82. Genetic studies suggest that these proteins are somewhat interchangeable. For example, RecG can substitute for some of the proteins in the RuvABC pathway (120). The functions of RuvABC appear to be conserved in other species of bacteria. Little is known about the proteins required for resolution in eukaryotes. One protein, CCE1, appears to be involved in resolution in yeast (121). It has been suggested that Rad54 may be involved in branch migration in the Rad52 pathway.

Our comparative analysis suggests that the RuvABC, RecG and Rus pathways all evolved within bacteria – no Archaeal or eukaryotic species encodes an ortholog of any of these. The RuvABC and RecG proteins are found in a wide diversity of bacterial species are likely evolved early in the history of bacteria. Rus on the other hand, has a very limited distribution and probably evolved quite recently. The distribution patterns of the RuvABC and RecG proteins are somewhat surprising. RuvA and RuvB are universal within bacteria – all bacteria encode orthologs of these proteins, suggesting that all bacteria can bind to Holliday junctions and catalyze branch migration. However, many species do not encode RuvC orthologs. This separation of the functions of RuvAB and RuvC is not that surprising. In *E. coli*, and many other species RuvA and RuvB are cotranscribed suggesting a tight functional link. Two of these species do encode RecG orthologs and thus may be able to catalyze branch resolution but three species (the two Mycoplasmas and *B. burgdorferi*) do not encode orthologs of any known resolvase. Whether these species encode alternative resolvases remains to be determined. It is possible that they can resolve junctions non-enzymatically.

RuvA has little similarity to any other proteins. RuvB and RecG are both members of the helicase superfamily of proteins, and thus arose as gene duplications from ancestral helicase motif containing proteins. It is of some interested that RecG is particularly closely related to Mfd and UvrB. RuvB is particularly closely related to an uncharacterized group of RuvB-like proteins. RuvC is somewhat similar in structure to RnaseH1 (see (122) for review) so it is possible that these proteins share a common ancestor.

Non-Homologous End Joining

In mammals, most of the repair of double-strand breaks is carried out without homologous recombination by non-homologous end joining (NHEJ) (123,124). In this process, DSBs are simply restitched back together. Thus this is in essence a form of direct repair. Genetic studies have shown that there are at least four proteins specifically required for this pathway in humans: XRCC4-7. The nature of XRCC4 is not known. Together XRCC5-7 make up the DNA-dependent protein kinase complex composed of Ku80/86 (XRCC5), Ku70 (XRCC6) and DNA-PKcs (XRCC7). These four proteins likely function by binding to DNA ends and stimulating DNA ligase activity. The proteins are involved in the repair of DSBs induced by irradiation and other DNA damaging agents, as well as by cellular processes such as VDJ recombination. Therefore, mutants in these proteins show are not only sensitive to DSB causing agents, but also have immunodeficiencies. In addition, recent results have shown that the human homologs of MRE11 and Rad50 are also involved in NHEJ (108). Putative homologs of Ku70 and Ku86 have been identified in yeast and these have been found to be involved in the repair of DSBs by NHEJ (125). As with humans, the yeast MRE11 and Rad50 (as well as XRS2) are also involved in NHEJ in yeast. However as mentioned earlier, most of the repair of DBSs in yeast is carried out by homologous recombination based pathways (126).

Our comparative analysis shows that there are no homologs of XRCC4 or any of the three subunits of DNA-PK in Archaea or bacteria. Therefore this pathway most likely evolved in eukaryotes (see (127) for more information about the evolution of some of these proteins). Our analysis also shows that the sequence similarity between the yeast and mammalian proteins is very limited. Although it is likely that these proteins are homologous, the low level of sequence similarity suggests that they also may have many functional differences. No ortholog of DNA-PKcs is found in yeast.

DNA Replication

Most repair pathways require some DNA synthesis as part of the repair process. In some cases, specific polymerases are used only for repair. In other cases, the normal replication polymerases are used for repair synthesis. Since the evolution of polymerases has been reviewed elsewhere it will not be discussed in detail here (128). Obviously, all species are able to replicate their DNA in some way and thus should be able to perform repair synthesis. The specific types of polymerases used may help determine the accuracy of repair synthesis.

Inducible Responses

LexA and the SOS system in bacteria

The SOS system in *E. coli* is an inducible response to a variety of cellular stresses, including DNA damage (129). A key component of the SOS system is the LexA transcription repressor. In response to stresses such as DNA damage, the RecA protein is activated to become a coprotease and assists the autocatalytic cleavage of LexA. When LexA is cleaved, it no longer functions as a transcription repressor, and the genes that it normally represses are induced. The induction of these LexA-regulated SOS genes is a key component of the SOS system. SOS-like processes have been documented in a wide variety of bacterial species. Those that have been characterized function like the *E. coli* system, with regulation of SOS genes by LexA homologs, although sometimes different sets of genes are repressed by LexA in the other species (129). Our comparative analysis suggests that LexA appeared near the origin of bacteria since it is found in a wide diversity of bacterial species including many not analyzed here. Nevertheless, many species do not encode a LexA homologs (Table 4). Thus the LexA gene was likely lost from these lineages sometime in the past. Since the role of LexA is conserved in the species in which it has been characterized, we conclude that those species that do not encode LexA do not have a standard SOS system. However, since there are many ways to regulate responses to external stimuli, it is possible that these species have co-opted another type of transcription regulator to control an SOS like response.

Interestingly, LexA is part of a multigene family that includes the UmuD protein and some phage repressors. One thing these proteins all have in common is that all undergo RecA assisted autocatalytic cleavage. In Table 4 we list presence and absence of LexA and UmuD separately.

P53 in animals

Inducible responses have also been found in some eukaryotes. One gene that is involved in inducible responses in animals is p53. One of p53's activities is transcriptional activation, and this activity is stimulated by the presence of DNA damage. Homologs of p53 have only been found in animals, suggesting that this inducible system evolved after animals diverged from other eukaryotes.

The Big Picture: Examining the Evolution of All Repair Processes

In the preceding sections we have focused the discussion on what the phylogenomic analysis reveals about specific repair proteins and pathways. We believe it is now important to take a "big picture" approach and consider all of the pathways together. One reason to take such a global approach is that the different pathways overlaps a great deal in their specificity. For example, cyclobutane pyrimidine dimers can be repaired by PHR, NER, BER (by T4EV), and can be tolerated through recombinational repair. In fact, it is rare for a particular lesion to be repaired only by one pathway. Yet another reason for the big picture approach is that some repair genes function in multiple pathways. Thus to understand the evolution of DNA repair processes and to make predictions about the repair capabilities of species from genome sequences, it is necessary to consider all processes and pathways together.

Distribution patterns and the "universality" of particular genes

Examination of the distribution of all DNA repair genes together reveals some interesting patterns. For example, only one DNA repair gene, RecA, is found in every species analyzed here. The universality of RecA suggests both that it is an ancient gene, and that its activity is irreplaceable (at least for these species). Since many DNA repair genes have important cellular functions, we were surprised that RecA there was only one gene that was present in all species. We chose a few other methods to examine distribution patterns. For example, in Table 5a we list those repair genes found in all or most bacteria. This zooming in on the bacteria shows that most bacteria have a large number of DNA repair genes in common and also that DNA repair is relatively highly conserved among bacteria. In addition, since the lists in Table 5a could include genes from outside the bacteria, in Table 5b we list genes that are found in bacteria but not eukaryotes and the converse, genes that are found in eukaryotes but not bacteria. We did not include Archaea in this particular list because the Archaea encode homologs of only a limited number of the DNA repair genes of bacteria or eukaryotes (discussed in more detail below).

We were also interested in obtaining a more objective measure of the "universality" of repair genes. To do this we created a crude universality measure (Table 5c). In this weighting scheme, we calculated the percentage of species within each domain (Archaea, Bacteria, Eukarya) in which a gene is present, and averaged these percentages. Thus a gene that is found in all bacteria but not in Archaea or eukaryotes (e.g., RuvA) would have a lower score than a gene found in some members of each domain of life (e.g., Xth). While this universality measure is biased because we do not have a random sampling of species, we believe it is still a useful way to compare the distribution patterns of different genes.

Summarizing evolutionary events

While the distribution and universality analysis described above are useful, we were more interested in comparing and contrasting (both between species and between the different classes of repair) the evolution of repair genes and pathways. To simplify this analysis, we have summarized the results of the evolutionary analyses discussed in the respective sections on each pathway. In Figure 4 we have traced the inferred gain and loss of repair genes onto an evolutionary tree of the species. In Table 6 we have sorted the repair genes by pathway and by the inferred timing of the origin of each gene. In the following sections we discuss some of the different features of the evolution of repair pathways.

Origins of DNA repair genes and processes I. timing

Comparing the timing of the origins of the different DNA repair genes and processes can be very informative. For example, the list of "ancient" genes serves to identify the repair genes and activities present in a common ancestor of all organisms. From this list we conclude that early in the evolution of life, many DNA repair activities were already present. These include PHR, alkyltransfer, recombination, AP endonuclease, a few DNA glycosylases and MMR. Interestingly, our analysis shows further that most of these ancient repair pathways have been lost from at least one evolutionary lineage. Thus these ancient activities are not absolutely required for survival in all species. Many of the other repair genes are actually quite old, even though they originated after the time of the last common ancestor of all organisms (Table 6). A large number of these old genes originated at or near the origins of major evolutionary groups (these are the genes which are listed as gained in Figure 4 near but not at the base of the tree). Interestingly, in many cases, genes with similar functions originated separately at the origins of bacteria and eukaryotes (e.g., UvrABCD vs. XPs, RuvABC vs. CCE1, LigI vs. LigII). In our analysis of the timing of the origin of the different repair genes we were also surprised to find that many repair genes are of a much more recent origin (e.g., MutH, SbcB, Rus, RecBCD, RecE, and AddAB). Thus repair processes are continuing to be originated in different lineages.

Origins of DNA repair genes and processes II. mechanism of origin

It is also interesting to determine the actual mechanism of origin of particular genes. How are new DNA repair genes created? One common means is by gene duplication (Table 7). Perhaps the best example of this comes from the helicase superfamily of proteins. This gene family is defined by the presence of the seven so-called helicase motifs. It is important to note, however, that not all proteins in this superfamily have helicase activity. Thus the presence of helicase motifs does not guarantee the presence of helicase activity. However, all proteins in this gene family do have helicase-like activities (e.g., some strip proteins from DNA and thus work as a protein-DNA helicase). Members of the helicase superfamily are involved in almost every repair pathway. In some cases, the same gene product is used for different pathways (e.g., UvrD in MMR and NER). In most cases however, distinct genes are used and sometimes multiple members of this gene family are used in the same pathway (e.g., UvrB, UvrD, and Mfd in NER). Since all members of the helicase superfamily are related to each other, there must have been dozens of gene duplication events over the history of this gene family. Two very useful pieces of information come from the finding of so many helicases involved in DNA repair. First, helicase activity, or related activities such as protein-DNA helicase activity is clearly required for most repair pathways. In addition, these activities are apparently difficult to develop from scratch because rather than invent new helicases, pathways "steal" helicases from other pathways by gene duplication, and then use them in a slightly different way. The fact that there have been many duplications of other repair genes suggests that many of the activities required for repair have only evolved once and then these activities have been incorporated into new pathways following gene duplication. Closer examination of the nature of particular gene duplications can be even more revealing. For example, many of the eukaryotic helicase superfamily genes involved in repair are members of one helicase family - the SNF2 family. As with the helicase superfamily, the fact that many repair genes are from within this particular family suggests that this particular family has an activity very useful for repair in eukaryotes (19).

Gene duplication is not the only way that new repair activities have been acquired over evolutionary time. Some repair genes have apparently been co-opted from other pathways without any gene duplication event (e.g., MutH may have descended from a restriction enzyme). New repair genes have also originated by gene fusion (e.g., SMS is a fusion between Lon and RecA, Ada is a fusion between an alkyltransferase and a transcription regulator (Figure 3). New repair activities can be acquired by a particular lineage without the creation of a new gene by the process of lateral transfer. There is only one well established case of lateral transfer of a DNA repair gene - that of RecA from the chloroplast to the plant nucleus (110). We have identified a few other possible cases of lateral transfer (listed with a "t" in Figure 4). However, these are only suggestions and need more detailed phylogenetic analysis to be confirmed.

Gene loss

Our analysis has identified many cases of loss of DNA repair genes (Figure 1). As discussed in the phylogenomics methods section above, identifying gene loss events is useful for understanding gene and pathway functions. From the global point of view, there have been many different types of gene loss events. In some cases, it appears that whole pathways have been lost as a unit (e.g., MutLS, SbcC). However, in other cases single genes or only parts of pathways are lost (e.g., components of the RecF pathway). Overall there has been a large amount of gene loss in the history of DNA repair processes. In some lineage the gene loss is extensive (Table 8). Why is this? In part, it is methodological, we have analyzed a biased sample of species. Many of these species for which complete genome sequences are available were chosen for sequencing because they have small genome sizes. Thus we are looking at a sample of species that have undergone large scale genome size reductions, possibly in the recent past

Origins of differences within a class of repair between species

Comparisons of the evolution of the different classes of repair reveal a great deal of diversity in how well conserved the classes of repair are. In addition, the ways in which classes of repair differ between species are also variable. The conservation between species can be classified according to the level of homology of the pathways. Some pathways are completely homologous between species (they make use of homologous genes in all species). However, this is only the case for some of the single enzyme pathways (PHR and alkyltransfer). Other pathways are partially homologous. For example, some of the proteins involved in MMR are homologous between *E. coli* and eukaryotes (e.g., MutS and MutL), but others are not (e.g., MutH and UvrD). Finally, there are some pathways that are not homologous at all between species despite performing the same functions. The best example of this is NER in bacteria compared to that in eukaryotes. These systems are clearly of completely separate origins.

Another means by which pathways differ between species is by functional divergence of homologs. Examples of this include the divergence of 6-4 and CPD photolyases and the divergence of MSH genes for MMR in eukaryotes. Related to this, it is interesting that while the most striking evolutionary difference in MMR between species is in different mechanisms of strand recognition, the main functional differences between species involves slight functional divergence among MutS homologs.

Prediction of species phenotypes and universal DNA repair activities

Our analysis shows that predicting a species' phenotype from its genome sequence is not completely simple. In essence, the difficulty in predicting phenotypes can be reduced to two problems. First, the presence of a homolog of a gene does not necessarily mean the presence of an activity because not all homologs have the same function (see above). The second and more difficult problem with functional predictions is that absence of a homolog does not necessarily imply absence of an activity. One reason for this is the overlap between repair pathways discussed above. In addition, it is always possible that a species may have novel genes that carry out an activity and thus these would not be detected by searches with characterized repair genes. This is one of the reasons why it is useful to identify how frequently particular functions evolve. For example, the fact that pathways for recombination initiation have evolved multiple times in different lineages suggests that such pathways may have also evolved in many unstudied lineages. In contrast, the fact that all generalized MMR systems use homologs of MutS and MutL for MMR suggests that the absence of *mutL* and *mutS* genes means the absence of general MMR.

Related to all of the problems described above, it is particularly difficult to make phenotypic predictions for those species that are not closely related to any of the well characterized model repair species. For example, there has been very little experimental work on DNA repair in Archaea and what has been done is usually the characterization of homologs of known repair genes (see Appendix F). Thus the nearest "template" species is very distant and one runs a great risk of missing novel repair pathways in these lineages. One can easily see the "bias" of model systems by following the gain of repair genes in Figure 1. Essentially all of the gain events are in the lineages leading up to *E. coli*, *B. subtilis*, yeast, and humans. This is not surprising because almost all the repair genes we analyzed are from these species. Clearly, repair genes must have originated in other lineages - especially given the evidence that new repair genes have originated relatively recently (see above).

Despite all these potential problems, we have still tried to make phenotypic predictions (Table 9). We tried to control for some of the prediction problems. For example, we did not make predictions for the presence of some activities in Archaea for those pathways that were significantly different between bacteria and eukaryotes. It should be remembered that all predictions need to be confirmed by experimental studies. We believe such predictions are a useful starting point for designing experiments on these species and for determining if the predicted presence or absence of particular repair activities can be correlated with any interesting biological properties. For example, the predicted absence of many repair pathways from mycoplasmas is consistent with the high mutation and evolutionary rates of mycoplasmas. Thus we can use the absence of certain genes to make some predictions. For example, the presence of UvrABCD but the absence of Mfd from the two *Mycoplasmas* and *A. aeolicus* suggests that these species can perform NER but not the TCR component of it.

One generalization that can be made from our phenotypic predictions is that, despite the lack of many universal genes, it appears that there are many universal activities. For example, we predict that all species have AP endonuclease activity. However, no AP endonuclease gene is universal because there are two evolutionarily unrelated AP endonuclease families (Nfo or Xth). All species encode at least one of these genes. Similarly, all species encode at least one of the two ligase genes.

Summary and Conclusions

We believe that the analysis reported here can serve as a starting point for experimental studies of repair in species with complete genome sequences and for understanding the evolution of DNA repair proteins and processes. However, it is important to restate some of the caveats to this type of analysis. First, it should be remembered that all functional and phenotypic predictions are just that - predictions and should be followed up by experimental analysis. Another source of bias is that the species for which complete genome sequences are available is not a random sampling of ecological and evolutionary diversity. In particular, many are somewhat degenerate species which have likely undergone large scale gene loss events in the recent past. This is one of the reasons they were sequenced. Thus this may give a misleading picture about what an average bacterium or Archaeon is like.

Despite these limitations, the phylogenomic analysis of DNA repair proteins presented here reveals many interesting details about DNA repair proteins and processes and the species for which complete genome sequences were analyzed. We have identified many examples of gene loss, gene duplication, functional divergence and recent origin of new pathways. All of this information helps us to understand the evolution of DNA repair as well as to predict phenotypes of species based upon their genome sequences. In addition, our analysis helps identify the origins of the different repair genes and has provided a great deal of information about the origins of whole pathways. We believe our analysis also helps identify potentially rewarding areas of future research. There are some unusual patterns that require further exploration such as the presence of UvrABCD in some Archaea and the only limited number of homologs of known repair genes in any of the three Archaea. In addition, the areas with empty spaces in the tree tracing the origin of repair genes may be of interest to determine if novel pathways exist in such lineages. In summary, we believe that this composite phylogenomic approach is an important tool in making sense out of genome sequence data and in understanding the evolution of whole pathways and genomes. Combining genomics and evolutionary analysis into phylogenomics is useful because genome information is useful in inferring evolutionary events and evolutionary information is useful in understanding genomes.

REFERENCES

1. Camerini-Otero, R. D. and Hsieh, P. (1995) *Annu Rev Genet*, **29,** 509-552.
2. Sancar, A. (1996) *Annu Rev Biochem*, **65,** 43-81.
3. Dybvig, K. and Voelker, L. L. (1996) *Annu Rev Microbiol*, **50,** 25-57.
4. Labarére, J. (1992) In Maniloff, J. (ed.), Mycoplasmas: Molecular Biology And Pathogenesis. American Society For Microbiology, Washington, D. C., pp. 309-323.
5. Modrich, P. and Lahue, R. (1996) *Annu Rev Biochem*, **65,** 101-133.

6. Promislow, D. E. (1994) *J Theor Biol*, **170,** 291-300.
7. Cortopassi, G. A. and Wang, E. (1996) *Mech Ageing Dev*, **91,** 211-218.
8. LeClerc, J. E., Li, B., Payne, W. L. and Cebula, T. A. (1996) *Science*, **274,** 1208-1211.
9. Matic, I., Radman, M., Taddei, F., Picard, B., Doit, C., Bingen, E., Denamur, E. and Elion, J. (1997) *Science*, **277,** 1833-1834.
10. Taddei, F., Vulic, M., Radman, M. and Matic, I. (1997) *Experientia*, **83,** 271-290.
11. Eyre-Walker, A. (1994) *Mol Biol Evol*, **11,** 88-98.
12. Sueoka, N. (1995) *J Mol Evol*, **40,** 318-325.
13. Sharp, P. M., Shields, D. C., Wolfe, K. H. and Li, W. H. (1989) *Science*, **246,** 808-810.
14. Battista, J. R. (1997) *Annu Rev Microbiol*, **51,** 203-224.
15. Sniegowski, P. (1998) *Curr Biol*, **8,** R59-61.
16. Matic, I., Rayssiguier, C. and Radman, M. (1995) *Cell*, **80,** 507-515.
17. Cleaver, J. E., Speakman, J. R. and Volpe, J. P. (1995) *Cancer Surv*, **25,** 125-142.
18. Kanai, S., Kikuno, R., Toh, H., Ryo, H. and Todo, T. (1997) *J Mol Evol*, **45,** 535-548.
19. Eisen, J. A., Sweder, K. S. and Hanawalt, P. C. (1995) *Nucleic Acids Res*, **23,** 2715-2723.
20. Eisen, J. A. (1998) *Nucleic Acids Res*, **26,** 4291-4300.
21. Henikoff, S., Greene, E. A., Pietrovsky, S., Bork, P., Attwood, T. K. and Hood, L. (1997) *Science*, **278,** 609-614.
22. Eisen, J. A. (1998) *Genome Res*, **8,** 163-167.
23. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) *Nucleic Acids Res*, **25,** 3389-3402.
24. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucleic Acids Res*, **22,** 4673-4680.
25. Swofford, D. (1991). *Phylogenetic Analysis Using Parsimony (PAUP) 3.0d.* Illinois Natural History Survey, Champaign, Ill.
26. Maddison, W. P. and Maddison, D. R. (1992). *MacClade 3*. Sinauer Associates, Inc., Sunderland, MA.
27. Maidak, B. L., Larsen, N., McCaughey, M. J., Overbeek, R., Olsen, G. J., Fogel, K., Blandy, J. and Woese, C. R. (1994) *Nucleic Acids Res.*, **22,** 3485-3487.
28. Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) *Science*, **278,** 631-637.
29. Zhao, S. and Sancar, A. (1997) *Photochem Photobiol*, **66,** 727-731.
30. Li, Y. F., Kim, S. T. and Sancar, A. (1993) *Proc Natl Acad Sci U S A*, **90,** 4389-4393.
31. Sutherland, J. C. and Griffin, K. P. (1980) *Radiat Res*, **83,** 529-536.
32. Chen, J., Huang, C. W., Hinman, L., Gordon, M. P. and Deranleau, D. A. (1976) *J Theor Biol*, **62,** 53-67.
33. Helene, C., Toulme, F., Charlier, M. and Yaniv, M. (1976) *Biochem Biophys Res Commun*, **71,** 91-98.

34. Toulme, J. J. and Helene, C. (1977) *J Biol Chem*, **252,** 244-249.

35. Pegg, A. E. and Byers, T. L. (1992) *Faseb J*, **6,** 2302-2310.

36. Pegg, A. E., Dolan, M. E. and Moschel, R. C. (1995) *Prog Nucleic Acid Res Mol Biol*, **51,** 167-223.

37. Leclere, M. M., Nishioka, M., Yuasa, T., Fujiwara, S., Takagi, M. and Imanaka, T. (1998) *Mol Gen Genet*, **258,** 69-77.

38. Skorvaga, M., Raven, N. D. H. and Margison, G. P. (1998) *Proc Natl Acad Sci U S A*, **95,** 6711-6715.

39. Kolodner, R. (1996) *Genes Dev*, **10,** 1433-1442.

40. Taddei, F., Matic, I., Godelle, B. and Radman, M. (1997) *Trends Microbiol*, **5,** 427-428.

41. Sniegowski, P. D., Gerrish, P. J. and Lenski, R. E. (1997) *Nature*, **387,** 703-705.

42. Mizrahi, V. and Andersen, S. J. (1998) *Mol Microbiol*, **29,** 1331-1340.

43. Matic, I., Taddei, F. and Radman, M. (1996) *Trends Microbiol*, **4,** 69-72.

44. Ban, C. and Yang, W. (1998) *EMBO J*, **17,** 1526-1534.

45. Glasner, W., Merkl, R., Schellenberger, V. and Fritz, H. J. (1995) *J Mol Biol*, **245,** 1-7.

46. Hoeijmakers, J. H. (1993) *Trends Genet*, **9,** 173-177.

47. Hoeijmakers, J. H. (1993) *Trends Genet*, **9,** 211-217.

48. Van Houten, B. and Snowden, A. (1993) *Bioessays*, **15,** 51-59.

49. Selby, C. P. and Sancar, A. (1995) *J Biol Chem*, **270,** 4882-4889.

50. Mellon, I. and Hanawalt, P. C. (1989) *Nature*, **342,** 95-98.

51. Ayora, S., Rojo, F., Ogasawara, N., Nakai, S. and Alonso, J. C. (1996) *J Mol Biol*, **256,** 301-318.

52. Zalieckas, J. M., Wray, L. V., Jr., Ferson, A. E. and Fisher, S. H. (1998) *Mol Microbiol*, **27,** 1031-1038.

53. Koonin, E. V. (1993) *J Mol Biol*, **229,** 1165-1174.

54. Linton, K. J. and Higgins, C. F. (1998) *Mol Microbiol*, **28,** 5-13.

55. Lin, C. G., Kovalsky, O. and Grossman, L. (1997) *Nucleic Acids Res*, **25,** 3151-3158.

56. Lomovskaya, N., Hong, S. K., Kim, S. U., Fonstein, L., Furuya, K. and Hutchinson, R. C. (1996) *J Bacteriol*, **178,** 3238-3225.

57. Wood, R. D. (1996) *Annu Rev Biochem*, **65,** 135-167.

58. Lieber, M. R. (1997) *Bioessays*, **19,** 233-240.

59. McCready, S. (1996) *Mutat Res*, **364,** 25-32.

60. Krokan, H. E., Standal, R. and Slupphaug, G. (1997) *Biochem J*, **325,** 1-16.

61. Meyer-Siegler, K., Mauro, D. J., Seal, G., Wurzer, J., deRiel, J. K. and Sirover, M. A. (1991) *Proc Natl Acad Sci U S A*, **88,** 8460-8464.

62. Muller, S. J. and Caradonna, S. (1991) *Biochim Biophys Acta*, **1088,** 197-207.

63. Slupphaug, G., Eftedal, I., Kavli, B., Bharati, S., Helle, N. M., Haug, T., Levine, D. W. and Krokan, H. E. (1995) *Biochemistry*, **34,** 128-138.

64. Nilsen, H., Otterlei, M., Haug, T., Solum, K., Nagelhus, T. A., Skorpen, F. and Krokan, H. E. (1997) *Nucleic Acids Res*, **25,** 750-755.

65. Koulis, A., Cowan, D. A., Pearl, L. H. and Savva, R. (1996) *FEMS Microbiol Lett*, **143,** 267-271.

66. Labahn, J., Scharer, O. D., Long, A., Ezaz-Nikpay, K., Verdine, G. L. and Ellenberger, T. E. (1996) *Cell*, **86,** 321-329.

67. Manuel, R. C., Czerwinski, E. W. and Lloyd, R. S. (1996) *J Biol Chem*, **271,** 16218-16226.

68. Piersen, C. E., Prince, M. A., Augustine, M. L., Dodson, M. L. and Lloyd, R. S. (1995) *J Biol Chem*, **270,** 23475-23484.

69. McGoldrick, J. P., Yeh, Y. C., Solomon, M., Essigmann, J. M. and Lu, A. L. (1995) *Mol Cell Biol*, **15,** 989-996.

70. Slupska, M. M., Baikalov, C., Luther, W. M., Chiang, J. H., Wei, Y. F. and Miller, J. H. (1996) *J Bacteriol*, **178,** 3885-3892.

71. Aspinwall, R., Rothwell, D. G., Roldan-Arjona, T., Anselmino, C., Ward, C. J., Cheadle, J. P., Sampson, J. R., Lindahl, T., Harris, P. C. and Hickson, I. D. (1997) *Proc Natl Acad Sci U S A*, **94,** 109-114.

72. Horst, J. P. and Fritz, H. J. (1996) *EMBO J*, **15,** 5459-5469.

73. Duwat, P., de Oliveira, R., Ehrlich, S. D. and Boiteux, S. (1995) *Microbiol*, **141,** 411-417.

74. Mikawa, T., Kato, R., Sugahara, M. and Kuramitsu, S. (1998) *Nucleic Acids Res*, **26,** 903-910.

75. Jiang, D., Hatahet, Z., Blaisdell, J. O., Melamede, R. J. and Wallace, S. S. (1997) *J Bacteriol*, **179,** 3773-3782.

76. Jiang, D., Hatahet, Z., Melamede, R. J., Kow, Y. W. and Wallace, S. S. (1997) *J Biol Chem*, **272,** 32230-32239.

77. van der Kemp, P. A., Thomas, D., Barbey, R., de Oliveira, R. and Boiteux, S. (1996) *Proc Natl Acad Sci U S A*, **93,** 5197-5202.

78. Arai, K., Morishita, K., Shinmura, K., Kohno, T., Kim, S. R., Nohmi, T., Taniwaki, M., Ohwada, S. and Yokota, J. (1997) *Oncogene*, **14,** 2857-2861.

79. Radicella, J. P., Dherin, C., Desmaze, C., Fox, M. S. and Boiteux, S. (1997) *Proc Natl Acad Sci U S A*, **94,** 8010-8015.

80. Rosenquist, T. A., Zharkov, D. O. and Grollman, A. P. (1997) *Proc Natl Acad Sci U S A*, **94,** 7429-7434.

81. Takao, M., Aburatani, H., Kobayashi, K. and Yasui, A. (1998) *Nucleic Acids Res*, **26,** 2917-2922.

82. Barzilay, G. and Hickson, I. D. (1995) *Bioessays*, **17,** 713-719.

83. Kuo, C. F., Mol, C. D., Thayer, M. M., Cunningham, R. P. and Tainer, J. A. (1994) *Ann N Y Acad Sci*, **726,** 223-234.

84. Clark, A. J. and Sandler, S. J. (1994) *Crit Rev Microbiol*, **20,** 125-142.

85. Kowalczykowski, S., Dixon, D., Eggleston, A., Lauder, S. and Rehrauer, W. (1994) *Microbiol Rev*, **58,** 401-465.

86. Hanawalt, P. C., Cooper, P. K., Ganesan, A. K. and Smith, C. A. (1979) *Annu Rev Biochem*, **48,** 783-836.

87. Eggleston, A. K. and West, S. C. (1997) *Curr Biol*, **7,** R745-749.

88. el Karoui, M., Ehrlich, D. and Gruss, A. (1998) *Proc Natl Acad Sci U S A*, **95,** 626-631.

89. Farrand, S. K., Hwang, I. and Cook, D. M. (1996) *J Bacteriol*, **178,** 4233-4247.

90. Alt-Morbe, J., Stryker, J. L., Fuqua, C., Li, P. L., Farrand, S. K. and Winans, S. C. (1996) *J Bacteriol*, **178,** 4248-4257.

91. Courcelle, J., Carswell-Crumpton, C. and Hanawalt, P. C. (1997) *Proc Natl Acad Sci U S A*, **94,** 3714-3719.

92. Nakayama, H., Nakayama, K., Nakayama, R., Irino, N., Nakayama, Y. and Hanawalt, P. (1984) *Mol Gen Genet*, **195,** 474-480.

93. Nakayama, K., Shiota, S. and Nakayama, H. (1988) *Can J Microbiol*, **34,** 905-907.

94. Watt, P. M., Hickson, I. D., Borts, R. H. and Louis, E. J. (1996) *Genetics*, **144,** 935-945.

95. Yu, C. E., Oshima, J., Fu, Y. H., Wijsman, E. M., Hisama, F., Alisch, R., Matthews, S., Nakura, J., Miki, T., Ouais, S., et al. (1996) *Science*, **272,** 258-262.

96. Ellis, N. A., Groden, J., Ye, T. Z., Straughen, J., Lennon, D. J., Ciocci, S., Proytcheva, M. and German, J. (1995) *Cell*, **83,** 655-666.

97. Gray, M. D., Shen, J. C., Kamath-Loeb, A. S., Blank, A., Sopher, B. L., Martin, G. M., Oshima, J. and Loeb, L. A. (1997) *Nat Genet*, **17,** 100-103.

98. Karow, J. K., Chakraverty, R. K. and Hickson, I. D. (1997) *J Biol Chem*, **272,** 30611-30614.

99. Hirano, T. (1998) *Curr Opin Cell Biol*, **10,** 317-322.

100. Jessberger, R., Frei, C. and Gasser, S. M. (1998) *Curr Opin Genet Dev*, **8,** 254-259.

101. Clark, A. J., Sharma, V., Brenowitz, S., Chu, C. C., Sandler, S., Satin, L., Templin, A., Berger, I. and Cohen, A. (1993) *J Bacteriol*, **175,** 7673-7682.

102. Kolodner, R., Hall, S. D. and Luisi-DeLuca, C. (1994) *Mol Microbiol*, **11,** 23-30.

103. Kusano, K., Takahashi, N. K., Yoshikura, H. and Kobayashi, I. (1994) *Gene*, **138,** 17-25.

104. Noirot, P. and Kolodner, R. D. (1998) *J Biol Chem*, **273,** 12274-12280.

105. Connelly, J. C., Kirkham, L. A. and Leach, D. R. F. (1998) *Proc Natl Acad Sci U S A*, **95,** 7969-7974.

106. Sharples, G. J. and Leach, D. R. F. (1995) *Mol Microbiol*, **17,** 1215-1217.

107. Petrini, J. H. J., Walsh, M. E., Dimare, C., Chen, X. N., Korenberg, J. R. and Weaver, D. T. (1995) *Genomics*, **29,** 80-86.

108. Paul, T. T. and Gellert, M. (1998) *Mol Cell*, **1,** 969-979.

109. Carney, J. P., Maser, R. S., Olivares, H., Davis, E. M., Le Beau, M., Yates , J. R., Hays, L., Morgan, W. F. and Petrini, J. H. J. (1998) *Cell*, **93,** 477-486.

110. Eisen, J. A. (1995) *J Mol Evol*, **41,** 1105-1123.

111. Gruber, T. M., Eisen, J. A., Gish, K. and Bryant, D. A. (1998) *FEMS Microbiol Lett*,

162, 53-60.

112. Stassen, N. Y., Logsdon, J. M., Jr., Vora, G. J., Offenberg, H. H., Palmer, J. D. and Zolan, M. E. (1997) *Curr Genet*, **31,** 144-157.

113. Norioka, N., Hsu, M. Y., Inouye, S. and Inouye, M. (1995) *J Bacteriol*, **177,** 4179-4182.

114. Griffith, J. D. and Harris, L. D. (1988) *CRC Crit Rev Biochem*, **23,** S43-86.

115. Bianco, P. R., Tracy, R. B. and Kowalczykowski, S. C. (1998) *Front Biosci*, **3,** d570-603.

116. Muller, B. and West, S. C. (1994) *Experientia*, **50,** 216-222.

117. Mandal, T. N., Mahdi, A. A., Sharples, G. J. and Lloyd, R. G. (1993) *J Bacteriol*, **175,** 4325-4334.

118. Sharples, G. J., Chan, S. N., Mahdi, A. A., Whitby, M. C. and Lloyd, R. G. (1994) *EMBO J*, **13,** 6133-6142.

119. Chan, S. N., Vincent, S. D. and Lloyd, R. G. (1998) *Nucleic Acids Res*, **26,** 1560-1566.

120. Ishioka, K., Iwasaki, H. and Shinagawa, H. (1997) *Genes Genet Syst*, **72,** 91-99.

121. Schofield, M. J., Lilley, D. M. and White, M. F. (1998) *Biochemistry*, **37,** 7733-7740.

122. West, S. C. (1997) *Annu Rev Genet*, **31,** 213-244.

123. Jeggo, P. A., Taccioli, G. E. and Jackson, S. P. (1995) *Bioessays*, **17,** 949-957.

124. Ramsden, D. A. and Gellert, M. (1998) *EMBO Journal*, **17,** 609-614.

125. Wilson, T. E., Grawunder, U. and Lieber, M. R. (1997) *Nature*, **388,** 495-498.

126. Chu, G. (1997) *J Biol Chem*, **272,** 24097-24100.

127. Keith, C. T. and Schreiber, S. L. (1995) *Science*, **270,** 50-51.

128. Edgell, D. R. and Doolittle, W. F. (1997) *Cell*, **89,** 995-998.

129. Shinagawa, H. (1996) *Experientia*, **77,** 221-35.

130. Blattner, F. R., Plunkett, G. I., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997) *Science*, **277,** 1453-1462.

131. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., et al. (1995) *Science*, **269,** 496-498, 507-512.

132. Roe, B. A., Clifton, S. and Dyer, D. W., personal communication.

133. Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., et al. (1997) *Nature*, **388,** 539-547.

134. Kunst, A., Ogasawara, N., Moszer, I., Albertini, A., Alloni, G., Azevedo, V., Bertero, M., Bessieres, P., Bolotin, A., Borchert, S., et al. (1997) *Nature*, **390,** 249-256.

135. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., et al.

(1995) *Science*, **270,** 397-403.

136. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. and Herrmann, R. (1996) *Nucleic Acids Res*, **24,** 4420-4449.

137. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry III, C. E., et al. (1998) *Nature*, **393,** 537-544.

138. Fraser, C., Casjens, S., Huang, W., Sutton, G., Clayton, R., Lathigra, R., White, O., Ketchum, K., Dodson, R., Hickey, E., et al. (1997) *Nature*, **390,** 580-586.

139. Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., et al. (1998) *Science*, **281,** 375-388.

140. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. (1996) *DNA Res*, **3,** 109-136.

141. Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Grahams, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M., et al. (1998) *Nature*, **392,** 353-358.

142. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., et al. (1996) *Science*, **273,** 1058-1073.

143. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., et al. (1996) *J Bacteriol*, **179,** 7135-7155.

144. Klenk, H.-P., Clayton, R., Tomb, J.-F., White, O., Nelsen, K., Ketchum, K., Dodson, R., Gwinn, M., Hickey, E., Peterson, J., et al. (1997) *Nature*, **390,** 364-370.

145. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996) *Science*, **274,** 546, 563-567.

Table 1. Completely or nearly completely sequenced genomes.

Species	Classification	Size (mb)	# Orfs	Ref.
Bacteria				
Escherichia coli K-12	Proteobacteria (γ)	4.60	4288	(130)
Haemophilus influenzae Rd KW20	Proteobacteria (γ)	1.83	1743	(131)
Neisseria gonorrhoeae	Proteobacteria (β)	2.20	n/a	(132)
Helicobacter pylori 26695	Proteobacteria (ε)	1.67	1590	(133)
Bacillus subtilis 169	Low GC Gram +	4.20	4100	(134)
Streptococcus pyogenes	Low GC Gram +	1.98	n/a	(132)
Mycoplasma genitalium G-37	Low GC Gram +	0.58	470	(135)
Mycoplasma pneumoniae M129	Low GC Gram +	0.82	679	(136)
Mycobacterium tuberculosis H37rV	High GC Gram +	4.41	~4000	(137)
Borrelia borgdorferi B31	Spirochete	1.44	1283	(138)
Treponema pallidum Nichols	Spirochete	1.14	1041	(139)
Synechocystis sp. PCC6803	Cyanobacteria	3.57	3168	(140)
Aquifex aeolicus	Aquificaceae	1.55	1512	(141)
Archaea				
Methanococcus jannascii DSM 2661	Euryarchaeota	1.66	1738	(142)
Methanobacterium thermoautotrophicum ΔH	Euryarchaeota	1.75	1855	(143)
Archaeoglobus fulgidus VC-16, DSM4304	Euryarchaeota	2.18	2436	(144)
Eukaryote				
Saccharomyces cerevisiae	Fungi	13.0	5885	(145)

Table 2. Components of phylogenomic analysis

Component	How is it Determined?	Uses of This Component
Gene Analysis		
1. Database of genes of interest.	Personal choice, characterized genes.	Similarity searches (2).
2. Searching for homologs.	Blast, PSI-blast, BLOCKS. Set homology threshold.	Presence/absence (4); gene tree (7).
3. Functional predictions.	Overlay known functions of genes onto gene tree.	Prediction of phenotypes (6); functional evolution.
Genome Analysis		
4. Gene presence/absence in species.	Searches (2) of complete genome sequences. Some refinement from evolutionary analysis (7, 10).	Evolutionary analysis (8, 10)
5. Correlated presence/absence.	Analyze presence/absence (4) in different species.	Functional predictions (3), pathway evolution (11).
6. Phenotype predictions.	Combine functional predictions (3), presence/absence (4) and pathway evolution (11).	Identify universal activities.
Evolutionary Analysis		
7. Gene trees.	Set homology threshold for searches (2) and use phylogenetic analysis of all homologs.	Presence/absence (4); identifying evolutionary events (10), functional predictions (3).
8. Evolutionary distribution patterns.	Overlay gene presence/absence (4) onto species tree.	Identifying gene evolutionary events; pathway evolution.
9. Congruence.	Compare gene tree (7) to species tree.	Distinguish lateral transfer from other events (8).
10. Gene evolution events.	Analysis of gene tree (7), congruence (9) and evolutionary distribution patterns (8).	Pathway evolution (11), correlated and convergent events, presence/absence (4); functional predictions (3)
11. Pathway evolution.	Integrate gene evolution (10), evolutionary distribution (8), correlated presence/absence (5).	Phenotype predictions (6); functional predictions (3).

Table 3. Evolutionary distribution patterns.

Type of pattern[1]	Description	Likely explanations	How resolve ambiguities?
Universal	All species have the gene.	Gene is ancient	n/a
Uniform presence	Gene is in only one evolutionary lineage.	Gene originated in that lineage.	n/a
Uniform absence	Gene is missing from one lineage.	Gene lost in that lineage.	n/a
Uneven	Presence/absence scattered through tree.	Gene loss or lateral transfer.	Compare gene tree vs. species tree.
Multicopy	Multiple homologs in some species.	Gene duplication or lateral transfer.	Compare gene tree vs. species tree.

[1] Determined by overlaying presence/absence of genes onto evolutionary tree of species

Table 4. Presence and absence of homologs of repair genes in different species.

Pathway Protein Name(s)	Biochemical Activity(s).	|-------------------------Bacteria--------------------------|														|--Archaea---|				|-Eukarya-|			Comments
		ECOLI	HAEIN	NEIGO	HELPY	STRPY	BACSU	MYCGE	MYCPN	SYNSP	MYCTU	BORBO	TREPA	AQUAE	Any	METTH	METJA	ARCFU	Any	YEAST	HUMAN	Any	
Direct Repair																							
Photoreactivation																							
PhrI	Photolyase (CPDs or 6-4s)	+	-	+	-	+	-	-	-	++	-	-	-	-	+	-	-	-	+	+	+	+	Homologous to PhrII. Not all have photolyase activity.
PhrII	Photolyase (CPDs or 6-4s)	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	-		+	Homologous to PhrI. Present in *M. xanthus*.
Alkylation reversal																							
Ogt (MGMT)	Alkyltransferase	+	+	-	+	-	++	-	-	-	+	-	+	+	+	+	+	+	+	+	+	+	Single domain. Called DAT1 in *B. subtilis*.
Ada	Alkyltransferase, adaptive response	+	+	±	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Two domains – (1) Ada transcription regulation (2) alkyltransferase.
Base Excision Repair[1]																							
Ung	Glycosylase (Uracil)	+	+	+	+	+	+	+	+	-	+	+	-	-	+	-	-	-	-	+	+	+	Also in many viruses. May have been lateral transfer to eukaryotes.
GT mismatch glycosylase	Glycosylase (T:G, T:U)	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	Also in *S. pombe*, *Serratia*.
Ogg1, Ogg2	Glycosylase (8-oxoG)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	+	++	+	Distantly related to MutY-Nth family, AlkA.
MutY-Nth family	Glycosylase (many)	++	++	++	++	+	++	-	-	++	++	+	++	++	++	++	++	++	++	++	++	++	Cannot identify distinct subfamilies. Distantly related to Ogg1, AlkA.
Fpg/MutM	Glycosylase (8-oxoG, FAPY)	+	+	+	-	+	+	+	+	+	+	-	-	-	+	-	-	-	-	-		-	Homologous to Nei. Also has dRPase and nicking activity.
Nei	Glycosylase (damaged C or Y)	+	-	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	-	-		-	Homologous to Fpg.
MPG (3MG, AAG)	Glycosylase (3-MeA)	-	-	-	-	-	+	-	-	-	+	+	-	-	+	-	-	-	-	-	+	+	Human protein also repairs 7-MeG, 3-MeG. Found in *A. thaliana*.
TagI, 3MG1	Glycosylase (3-MeA)	+	+	-	-	+	-	-	-	-	+	-	-	-	+	-	-	-	-	-		-	A.KA. 3-Me-A glycosylase I. Some activity for 3-Et-A, 3-Me-G.
AlkA (3MG2/TagII/MAG)	Glycosylase (3-MeA, many others)	+	-	-	-	-	++	-	-	+	+	-	-	-	+	-	-	+	+	+		+	Wide specificity (many alkyl-base lesions). Distantly related to Ogg1, Nth. Two domain protein in gram + species (1 - Ada, 2- AlkA).
AP Endonucleases[2]																							
Xth (APE1,ExoA)	5' AP endonuclease	+	+	+	+	+	+	-	-	+	+	+	+	-	+	+	-	+	+	-	+	+	*E. coli* protein has some exonuclease activities as well (aka ExoIII)
Nfo (APN1)	5' AP endonuclease	+	-	-	-	-	+	+	+	-	+	-	-	+	+	+	+	-	-	+		+	Also found in *S. pombe*, *C. elegans*, and some viruses.
Mismatch Excision Repair																							
Mismatch Recognition																							
MutS1 (MSH1,2,3,6)	Binds mismatches and loops	+	+	+	-	+	+	-	-	++	-	+	+	+	+	-	-	-	-	++	++	++	Part of MutS family (see MutS2 below). Heterodimers in euks.
MutL (PMS1, MLH1)	Binds MutS	+	+	+	-	+	+	-	-	+	-	+	+	+	+	-	-	-	±	++	++	++	Different versions used for heterodimer in eukaryotes.
Vsr	T:G mismatch endonuclease	+	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Also in some *Xanthomonas* and some *Haemophilus* species
Strand Recognition																							
MutH	GATC endonuclease	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Related to Sau3A restriction enzyme.
Dam	GATC methylase	+	+	+	-	-	-	-	-	+	-	-	+	-	+	+	±	-	±	-		-	Methylation activity used in other pathways in many species.
Exonucleases[3]																							
ExoI (SbcB)	3'-5' ssDNA exonuclease	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Also involved in recombination.
RecJ	5'-3' ssDNA exonuclease	+	+	+	+	+	+	-	-	+	-	+	+	+	+	-	-	-	-	-		-	Also involved in recombination.
XseA	5'-3' ssDNA exo. w/ XseB	+	+	+	+	+	+	-	-	-	+	-	-	-	+	-	-	-	-	-		-	Large subunit of exo VII.
XseB	5'-3' ssDNA exo. w/ XseA	+	+	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	-	-		-	Small subunit of exo VII. Small size limits homology searches.
DHS1 (ExoI)	Exonuclease	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	±	±	±	+	+	+	FEN1 family. Called Hex1 in humans, tosca in flies.
Exision Helicase																							
UvrD/Helicase II	Excision helicase	+	+	+	+	+	++	+	+	+	++	+	+	+	+	++	-	-	-	+		-	Helicase superfamily, related to Rep, RadH, PcrA. Used in NER.

Table 4. Presence and absence of homologs of repair genes in different species.

Pathway Protein Name(s)	Biochemical Activity(s).	|-------------------------Bacteria													---------------------------|	|--Archaea---|				|-Eukarya-|			Comments
		ECOLI	HAEIN	NEIGO	HELPY	STRPY	BACSU	MYCGE	MYCPN	SYNSP	MYCTU	BORBO	TREPA	AQUAE	Any	METTH	METJA	ARCFU	Any	YEAST	HUMAN	Any	
<u>Nucleotide Excision Repair</u>																							
<u>Bacterial NER</u>																							
UvrA	Binds damaged DNA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-		-	ABC transporter superfamily. Called mtrAB in *D. radiodurans*.
UvrB	Helicase, 3' incision endonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-		-	Helicase superfamily, related to RecG, MFD.
UvrC	5' incision endonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-		-	Many bacteria have a second UvrC-like protein.
UvrD	Excision helicase	+	+	+	+	+	++	+	+	+	++	+	+	+	+	++	-	-	-	+		+	Helicase superfamily, related to Rep, RadH, PcrA. Used in MMR.
MFD	Transcription repair coupling	+	+	+	+	+	+	-	-	+	+	+	+	-	+	-	-	-	-	-		-	Helicase superfamily, related to UvrB, RecG.
<u>Eukaryotic NER</u>																							
Recognition																							
Rad14 (*XPA*)	Binds damaged DNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
RFA1/RPA1	ssDNA binding w/ RFA2,3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	-	-	-	+	+	+	Metth protein is distantly related.
RFA2/RPA2	ssDNA binding w/ RFA1,3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	++	+	Human RFA4 is very similar to RFA2.
RFA3/RPA3-human	ssDNA binding w/ RFA1,2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	
RFA3/RPA3-yeast	ssDNA binding w/ RFA1,2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+		+	
Initiation																							
Rad3 (*XPD*) (ERCC2)	TFIIH component – helicase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	+	Helicase superfamily. Related to DinG. *S. pombe rad15*.
Rad25 (*XPB*) (ERCC3)	TFIIH component – helicase	-	-	-	-	-	-	-	-	-	+	-	+	-	+	-	-	+	+	+	+	+	Helicase superfamily. Also in a Halophilic Archaea.
SSL1 (*p44*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
TFB1 (*p62*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
TFB2 (*p52*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
TFB3 (*MAT1*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
TFB4 (*p34*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	
CCL1 (*CyclinH*)	TFIIH component	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Cyclin family.
Kin28 (*CDK7*)	TFIIH component - protein kinase	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	CDK-like kinase
Incision																							
Rad2 (*XPG*) (ERCC5)	3' incision (flap endonuclease)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	FEN1 family. *rad13* in *S.pombe*.
Rad10 (ERCC1)	5' incision endonuclease w/ Rad1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Shares motif w/ UvrC, ligases. *swi10* in *S.pombe.* Involved in recomb.
Rad1 (*XPF*) (ERCC4)	5' incision endonuclease w/ Rad10	-	-	-	-	-	-	-	-	-	-	-	-		-	+	+	+	+	+	+	+	*Rad16* in *S.pombe*, mei-9 in fly. Also involved in recombination.
Specificity																							
Rad4 (*XPC*)	Repair of inactive DNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	XPC forms complex with Rad23. XPC/Rad4 similarity is limited.
Rad23 (*HHRAD23*)	Repair of inactive DNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	++	+	Contains ubiquitin motif. Human Rad23 complexes with XPC.
Rad7	Repair of inactive DNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+		+	No known human homolog..
Rad16	Repair of inactive DNA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Helicase superfamily, SNF2 family.
Rad26 (*CSB*) (ERCC6)	Transcription-repair coupling	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Helicase superfamily, SNF2 family.
CSA (ERCC8)	Transcription-repair coupling	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	WD repeat containing protein. Not true ortholog of Rad28.

Table 4. Presence and absence of homologs of repair genes in different species.

Pathway Protein Name(s)	Biochemical Activity(s).	|-------------------------Bacteria---------------------------|														|--Archaea---|				|-Eukarya-|			Comments
		ECOLI	HAEIN	NEIGO	HELPY	STRPY	BACSU	MYCGE	MYCPN	SYNSP	MYCTU	BORBO	TREPA	AQUAE	Any	METTH	METJA	ARCFU	Any	YEAST	HUMAN	Any	
<u>Recombinational Repair</u>																							
<u>Initiation</u>																							
RecBCD pathway																							
RecB	ExoV[4] Helicase	+	+	+	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-	-		-	Helicase superfamily, related to AddA, UvrD, PcrA.
RecC	ExoV Nuclease	+	+	+	-	-	-	-	-	-	+	±+	-	-	+	-	-	-	-	-		-	
RecD	ExoV Helicase	+	+	+	-	±	±	-	-	-	+	±+	-	-	+	-	-	-	-	-		-	Helicase superfamily, related to TraI, TraA.
RecF pathway																							
RecF	Assists RecA filamentation	+	+	-	-	+	+	-	-	+	+	-	+	-	+	-	-	±	-	±		±	SMC family.
RecJ	5'-3' ssDNA exonuclease	+	+	+	+	+	+	-	-	+	-	+	+	+	+	-	-	-	-	-		-	Also used in MMR and RecE pathway.
RecO	Binds ssDNA, assists RecF?	+	+	+	-	+	+	-	-	+	+	-	-	-	+	-	-	-	-	-		-	
RecR	ATP binding, assists RecF?	+	+	+	±+	+	+	-	-	+	+	-	+	+	+	-	-	-	-	-		-	
RecN	ATP binding	+	+	+	+	+	+	-	-	+	+	-	+	+	+	-	-	±	-	-		-	SMC family
RecQ	3'-5' DNA helicase	+	+	+	-	±	+	-	-	+	-	-	+	-	+	-	-	-	-	+	++	+	Helicase superfamily, Dead family. Human homologs are defective in Werner's, Bloom's syndromes
RecE pathway																							
RecE/ExoVIII	5'-3' dsDNA exonuclease	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Encoded by cryptic rac prophage.
RecT	Binds ssDNA, promotes pairing	+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Encoded by cryptic rac prophage. Also in phage PVL, SPP1
SbcBCD pathway																							
SbcB/ExoI	3'-5' ssDNA exonuclease	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-		-	
SbcC	dsDNA exonuclease (w/ sbcD)	+	-	-	-	±+	+	-	-	+	-	+	+	+	+	±	±	±	±	±	±	±	SMC family. May be homolog of Rad50.
SbcD	dsDNA exonuclease (w/ sbcC)	+	-	-	-	-	+	-	-	+	-	+	+	+	+	±	±	±	±	±	±	±	May be ortholog of MRE11.
AddAB Pathway																							
AddA/RexA	Exonuclease + helicase w/ AddB	-	-	+	-	+	+	-	-	-	-	-	+	-	+	-	-	-	-	-		-	Helicase superfamily. Related to UvrD, PcrA, RecB.
AddB/RexB	Exonuclease + helicase w/ AddA	-	-	+	-	+	+	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Distantly related to AddA, may be in helicase family.
Rad52 pathway																							
Rad52, Rad59	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	++	+	+	Rad52 and Rad59 are homologs of each other.
Mre11/Rad32	Nuclease w/ Rad50	±	-	-	-	±	±	-	-	±	-	±	±	±	±	+	+	+	+	+	+	+	May be homolog of SbcD.
Rad50	Nuclease w/ Mre11	±	-	-	-	±	±	-	-	±	-	±	±	±	±	+	+	+	+	+	+	+	SMC family. May be ortholog of SbcC
<u>Recombinase</u>																							
RecA, Rad51	DNA binding, strand exchange	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	++	Homolog also in T4 phage (UvsX). Related to SMS, Rad55, Rad57.
<u>Branch migration/resolution</u>																							
Branch migration																							
RuvA	Binds junctions. Helicase w/ RuvB	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	-		-	
RuvB	5'-3' junction helicase w/ RuvA	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	-		-	Helicase superfamily.
RecG	Resolvase, 3'-5' junction helicase	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-		-	Helicase superfamily, related to UvrB, Mfd.
Resolvases																							
RuvC	Junction endonuclease	+	+	+	+	-	-	-	-	+	+	-	+	-	+	-	-	-	-	-		-	
RecG	Resolvase, 3'-5' junction helicase	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-		-	Helicase superfamily, related to UvrB, Mfd.
Rus	Junction endonuclease	+	-	-	-	-	-	-	-	-	±+	-	-	±+	+	-	-	-	-	-		-	Encoded by prophage DLP12. Also in phage 82.
CCE1	Junction endonuclease	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+		+	May function in mitochondria.
<u>Other recombination proteins</u>																							
Rad54	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Helicase superfamily, SNF2 family.
Rad55	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Distant relative of RecA/Rad51.
Rad57	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	Distant relative of RecA/Rad51.
Xrs2	Assists Rad50/MRE11?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+		+	

Table 4. Presence and absence of homologs of repair genes in different species.

Pathway Protein Name(s)	Biochemical Activity(s).	Bacteria														Archaea				Eukarya			Comments
		ECOLI	HAEIN	NEIGO	HELPY	STRPY	BACSU	MYCGE	MYCPN	SYNSP	MYCTU	BORBO	TREPA	AQUAE	Any	METTH	METJA	ARCFU	Any	YEAST	HUMAN	Any	
Non-homologous end joining																							
Ku70	Subunit of DNA-PK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	Yeast and human proteins distantly related. Similar to Ku86.
Ku86	Subunit of DNA-PK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	Yeast and human proteins very distantly related. Similar to Ku70.
DNA-PKcs	Catalytic subunit of DNA-PK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	PIK/ATM/DNA-PK family. No clear yeast ortholog.
XRCC4	Recruits ligase?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	
DNA Ligases																							
DnlI	NAD-dependent DNA ligase	++	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-		-	Distantly related to replication factor C of eukaryotes
LigII	ATP-dependent DNA ligase	-	-	-	-	-	+	-	-	-	++	-	-	+	±	+	+	+	+	++	++	++	Also in many viruses. *B. subtilis* protein in prophage.
Nucleotide pools																							
MutT family	Repairs 8-oxy-dGTP, GTP	+	+	+	+	±	+	-	-	+	++	-	-	+	+	+	++	+	+	+	++	+	Not all proteins in the MutT family have this activity.
Dut	Keeps dUTP pool low	+	+	+	+	+	+	-	-	-	+	-	+	+	+	-	-	-	-	+	+	+	Eukaryotic forms may be of mitochondrial origin.
Replication																							
PolA family (Pols A,γ)	DNA polymerase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	Polγ is a mitochondria protein. Homologs in many phage.
PolB family (PolsB, α, δ, ζ)	DNA polymerase	+	-	-	-	-	-	-	-	-	-	-	-	-	-	++	+	+	+	++	++	++	Polζ = Rev3. Homologs in many eukaryotic viruses.
PolC family (DnaE)	DNA polymerase	+	+	+	+	+	+	+	+	+	++	+	+	+	+	-	-	-	-	-		-	
PCNA	Sliding clamp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	
Other Repair Proteins																							
UmuC family	SOS mutagenesis	+	-	+	-	+	+	+	+	+	++	-	-	-	+	-	-	-	+	+	+	+	
UmuD	SOS mutagenesis	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	Related to LexA.
LexA	SOS regulon transcription repressor	+	+	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-		-	Also in Thermotoga and Deinococcus.
SMS, RadA	Unknown	+	+	+	+	+	+	-	-	+	+	-	+	+	+	±	±	±	±	±	±	±	Two domains. 1) RecA-like. 2) Lon-like.
PriA	Primosome assembly helicase	+	+	+	+	+	+	-	-	+	+	+	+	±	+	-	-	-	-	-		-	
P53	Txn. regulation, tumor suppressor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	++	+	
SSB	Binds ssDNA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	+	+	+	Eukaryotic SSBs function in mitochondria and of mitochondrial origin.
HepA1	Subunit of RNA-pol holoenzyme	+	+	-	-	-	+	-	-	-	+	-	-	?	+	-	-	-	+	++	++	++	Helicase superfamily, SNF2 family.
HepA2	Unknown functions	-	-	-	-	+	+	+	+	+	+	-	-	?	+	-	-	+	+	++	++	++	Helicase superfamily, SNF2 family.
Spl	Repairs spore UV dimers	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-		-	Also in Clostridia.
Lon	Protease	++	+	+	+	-	++	+	+	-	-	++	++	+	+	+	+	+	+	+	+	+	Related to domain2 of SMS. Euk. forms function in mitochondria.
MutS2	Chromosome segregation?	-	-	-	+	+	+	-	-	+	-	+	-	+	+	±	-	±	±	++	++	++	MutS family. Includes MSH4, MSH5 of eukaryotes.
XRCC1	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	Also found in *S. pombe*.
XRCC2	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	Distantly related to RecA/Rad51.
XRCC3	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	±	+	+	Related to RecA/Rad51 (closer than XRCC2).
XRCC9	n/a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	

[*] In those cases in which a species encoded a gene for which homology to the gene of interest was ambiguous, we indicated ±. If a gene was found in any other species within bacteria, Archaea or eukaryotes, this is listed in the "ANY" column. For those genes that were part of multigene families, we used phylogenetic analysis to divide the family into subfamilies and groups of orthologs and paralogs (see Results and Discussion). If subfamilies could be determined unambiguously, we only identify presence and absence of a homolog within the same subfamily as the search gene. If subfamilies could not be determined unambiguously, we listed the number of homologs of a particular gene (e.g., MutY-Nth). In cases of relatively recent gene duplications, presence of multiple homologs (++) was indicated for a few species a limited number of species encoded multiple orthologs of a gene. If lateral transfers were identified, this is indicated in the Comments column. Additional details can be found in the discussion.

[1] The first step in BER involves glycosylases. See text for details on other steps. Some of these glycosylases also have AP lyase or dRPase activity.

[2] Functions similarly to AP-Endonuclease but biochemical activity is AP lyase (in conjunction with role in base excision repair).

[3] Many exonucleases can serve this role in mismatch repair.

[4] RecBCD complex (ExoV) has many activities including dsDNA and ssDNA exonuclease and endonuclease, ATPase, helicase, and Chi-site recognition.

Table 5a. DNA repair genes present in all or most bacteria.

Process	In all bacteria	In most bacteria
Nucleotide excision repair	UvrABCD	UvrABCD
Holliday junction resolution	-	RuvABC
Recombination	RecA	RecA; RecJ, RecG
Replication	PolA, C	PolA,C; PriA; SSB
Ligation	LigaseI	LigaseI
Transcription-coupled repair	-	Mfd
Base excision repair	-	Ung, MutY-Nth
AP endonuclease	-	Xth
Single-strand binding protein	SSB	SSB

Table 5b. DNA repair genes present in bacteria or eukaryotes but not both.

Process	Only in Bacteria	Only in Eukaryotes
Transcription-coupled repair	Mfd	CSB, CSA
Mismatch strand recognition	MutH	-
Nucleotide excision repair	UvrABCD	XPs, TFIIH, etc.
Recombination initiation	RecBCD, RecF	KU, DNA-PK
Holliday junction resolution	RuvABC	CCE1
Base excision	Fpg-Nei, TagI	-
Inducible responses	LexA	P53

Table 5c. Universality of DNA repair genes.

Universality[1]	Gene
0-0.1	RFA3, CSA, Ku70, Ku86, DNA-PKcs, XRCC4, P53, XRCC1, XRCC2, XRCC3, XRCC9, GT glyc, RecE, UmuD, Spl, Ada, Nei , Vsr, MutH, SbcB, MPG, RecT, AddB, Rus
0.1-0.2	TagI, 3MG1, AddA, LexA, PhrII, RecB, RecC, RecD, XseB, XseA, RecF, RecO, RuvC
0.2-0.3	Fpg, Dam, RecJ, MFD, RecJ, RecR, RecN, SMS, RadA, PriA, RecG, RecG
0.3-0.4	RuvA, RuvB, DHS1, Rad14, RFA1, RFA2, RFA3, Rad3, SSL1, TFB1, TFB2, TFB3, TFB4 , CCL1, Kin28, Rad10, Rad4, Rad23, Rad7, Rad16, Rad26, Rad52, Rad59, CCE1, Rad54, Rad55, Rad57, Xrs2, DnlI, PolC family
0.4-0.5	PhrI, HepA1, Ogg1, Ogg2, UvrA, UvrB, UvrC, Xth, RecQ, MutS2, Rad25
0.5-0.6	UmuC family, AlkA, MutS1, MutL, Dut, Ung, HepA2
0.6-0.7	Rad2, Rad1, PCNA, PolA family, SSB, PolB family
0.7-0.8	Nfo, LigII, UvrD
0.8-0.9	Ogt , Mre11, Rad50, MutT family
0.9-0.99	Lon, MutY-Nth family
1	RecA

[1] Universality was calculated by calculating the frequency of species in which a gene was found within a particular domain of organisms (bacteria, eukarya, Archaea) and averaging this frequency between the three domains. It is important to note that this is a highly biased estimate since the species represented are not a random sample of each domain.

Table 6. Origin of DNA repair genes and pathways.

Pathway	Ancient	Within Bacteria	Arch-Euk Lineage	Within Archaea	Within Eukaryota	Ambiguous Origin	General Mechanisms Conserved?	Comments
Photoreactivation	PhrI PhrII	-	-	-	-	-	Yes	Specificity varies between species. PhrI and PhrII genes lost many times. Also some lateral transfer and duplication.
Alkyltransfer	Ogt	Ada	-	-	-	-	Yes	Addition of Ada domain to Ada protein occurred in bacteria.
Base Excision Repair	Ung? MutY/Nth AlkA	Fpg/Nei TagI	Ogg	-	-	3MG GT MMR	Yes	Ung may have originated in bacteria. Specificity varies greatly between species for MutY-Nth, AlkA, and others. Many cases of gene loss.
AP Endonucleases	Xth Nfo	-	-	-	-	-	Yes	Many cases of gene loss of Xth and Nfo. All species have one or the other.
Nucleotide Excision Repair	-	UvrABCD	Rad1 Rad2	-	All euk. NER prots except Rad1,2,25	Rad25	Yes/No	UvrABCD in *M. thermoautotrophicum* (Archaea) probably by lateral transfer.
Transcription-Coupled Repair	-	Mfd	-	-	CSA, CSB	-	?	Mfd missing from some bacteria.
General Mismatch Repair	MutLS?	MutH Dam Vsr	-	-	dup MutS dup MutL	-	Yes/No	Strand recognition systems and exonucleases differ between species. Many cases of loss of MutLS genes. Duplication in eukaryotes allows use of heterodimers.
Recombination Initiation	SbcCD	AddAB RecBCD RecFJNOR RecET SbcB	-	-	dup RecQ	RecQ	No	Many cases of gene loss in bacteria. RecF pathway genes not always present together.
Recombinase	RecA	RecT?	-	-	dup RecA	-	Yes	Lateral transfer from chloroplast to plant nucleus has occurred. RecT is of phage origin.
Branch Migration	-	RuvAB RecG	-	-	-	-	Yes/No	RuvAB and RecG missing from some bacteria.
Branch Resolution	-	RuvC Rus RecG	-	-	CCE1	-	Yes/No	CCE1 may function in mitochondria. Rus is likely of phage origin and is only found in a few species.
Other Recombination	-	-	-	-	Rad52-59 XRS2	-	-	-
Non-homologous end joining	-	-	-	-	XRCC4 Ku70, 86 DNA-PKcs	-	-	-
Ligation	-	LigI	LigII	-	-	-	Maybe	-
Induction	-	LexA	-	-	P53	-	No	-
Other	MutT UmuC SMS?	SSB	-	-	RFAs	Dut	-	Eukaryotic SSB came from mitochondria.

Table 7. Gene duplications in the history of DNA repair genes

Ancient	SNF2
	MutS1-MutS2
	RecA-SMS
	PhrI-PhrII
	MutY-Nth
	Early helicase evolution
In eukaryotes	Rad23a-Rad23b in animals
	RecQL-Blooms-Werner's in animals
	SNF2 family massive duplication
	Rad51-DMC1
	MSH1-6 (MutS family)
	PMS1-MLH1-MLH2 (MutL family)
	Rad52-Rad59
	polB family
	Ligase family II
In bacteria	Fpg-Nei,
	UvrB-Mfd-RecG
	UvrA
	LexA-UmuD
	Ada-Ogt in Proteobacteria
	Phr in some cyanobacteria
	UvrD-Rep-RecB
	RecA1-RecA2 in *Myxococcus xanthus*

Table 8. DNA repair genes that were lost in the mycoplasmal lineage

Process	Protein
Base excision repair	MutY/Nth, AlkA
Recombination initiation	RecF pathway, SbcCD
Recombination resolution	RecG, RuvC
Mismatch repair	MutLS
Transcription coupled repair	MFD
Induction	LexA
Direct repair	PhrI, Ogt
AP endonuclease	Xth
Other	MutT, Dut, PriA, SMS

Table 9. Predicted DNA repair phenotypes of different species.

Pathway	Proteins with Activity	|--Bacteria--|														|-------Archaea--------|				|---Eukarya---|	
		ECOLI	HAEIN	NEIGO	HELPY	STRPY	BACSU	MYCGE	MYCPN	MYCTU	SYNSP	BORBO	TREPA	AQUAE	ALL	METTH	METJA	ARCFU	ALL	YEAST	HUMAN	ALL
Photoreactivation	PhrI,PhrII	+	-	+	-	+	-	-	-	-	+	-	-	-	-	+	-	-	-	+	-	-
Alkyltransfer	Ada/Ogt/MGMT	+	+	-	+	-	+	-	-	+	-	-	+	+	-	+	+	+	+	+	+	+
Nucleotide excision repair	UvrABCD or XPs	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	?	?	+	+	+
Transcription-coupling	Mfd or CSA/CSB	+	+	+	+	+	+	-	-	+	+	+	+	-	+	?	?	?	?	+	+	+
Base excision repair		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Uracil glycosylase	Ung, GT	+	+	+	+	+	+	+	+	+	-	+	-	-	-	-	-	-	-	+	+	+
Alkylation glycosylase	AlkA, TagI, MPG	+	+	-	+	+	+	-	-	+	+	-	-	-	-	-	-	+	-	+	+	+
Misc. damaged bases	MutY/Nth, Fpg/Nei, Ogg	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
AP endonucleases	Xth/APE1, Nfo/APN1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Mismatch repair	MutLS	+	+	+	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	+	+	+
Recombination																						
Initiation																						
RecBCD-DSBR	RecBCD	+	+	+	-	-	-	-	-	+	-	±	-	-	-	-	-	-	-	-	-	-
RecF-DSGR	RecF	+	+	±	±-	±	+	-	-	±	+	-	±	±	-	-	-	-	-	-	-	-
AddAB-pathway	AddAB	-	-	+	-	+	+	-	-	-	-	-	±	-	-	-	-	-	-	-	-	-
SbcCD	SbcC/MRE11, SbcD/Rad50	+	-	-	-	-	+	-	-	-	+	+	+	+	-	+	+	+	+	+	+	+
Recombinase	RecA, RadA, Rad51	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Branch migration	RuvAB, RecG	+	+	+	+	+	+	+	+	+	+	+	+	+	+	?	?	?	?	?	?	?
Resolution*	RuvC, Rus, RecG, CCE1	+	+	+	+	+	+	-	-	+	+	-	+	+	-	?	?	?	?	+	?	?
Non-homologous end joining	Ku, DNA-PK	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+
Ligation	LigI, LigII	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Figure 1. Schematic diagram of phylogenomic methodology.

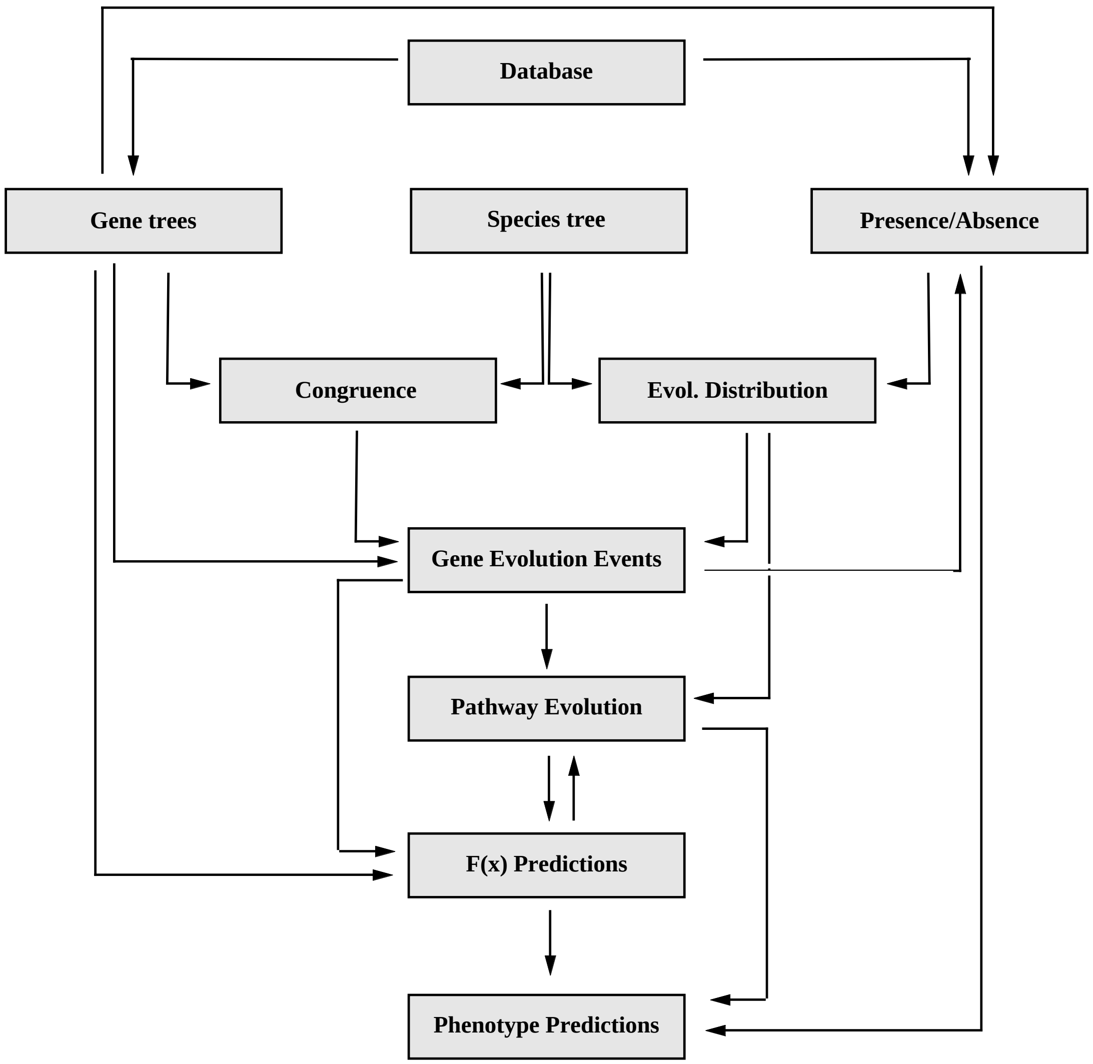
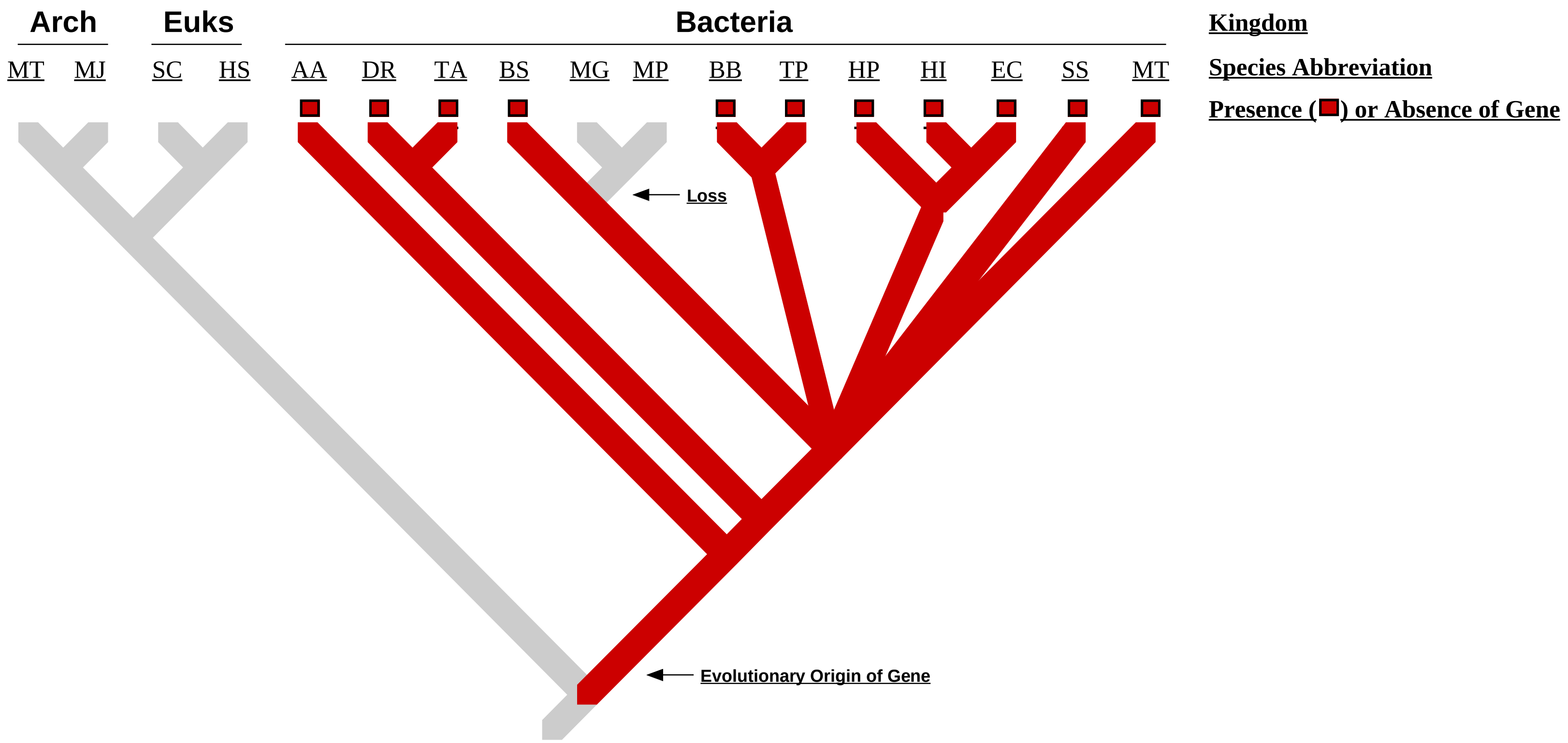

Database
Gene trees
Species tree
Presence/Absence
Congruence
Evol. Distribution
Gene Evolution Events
Pathway Evolution
F(x) Predictions
Phenotype Predictions

Figure 2. Demonstration of using evolutionary distribution patterns to trace gene gain and loss.

An evolutionary tree of the relationships among some representatives of the bacteria, Archaea, and eukaryota is shown. Presence of genes in these species is indicated by a colored box at the tip of the terminal branches of the tree. Gain and loss of the gene is inferred through parsimony reconstruction techniques.



Arch
Euks
Bacteria
Kingdom
MT
MJ
SC
HS
AA
DR
TA
BS
MG
MP
BB
TP
HP
HI
EC
SS
MT
Species Abbreviation
Presence (🟥) or Absence of Gene
Loss
Evolutionary Origin of Gene

Figure 3. Schematic diagram of an alignment of alkyltransferase genes.

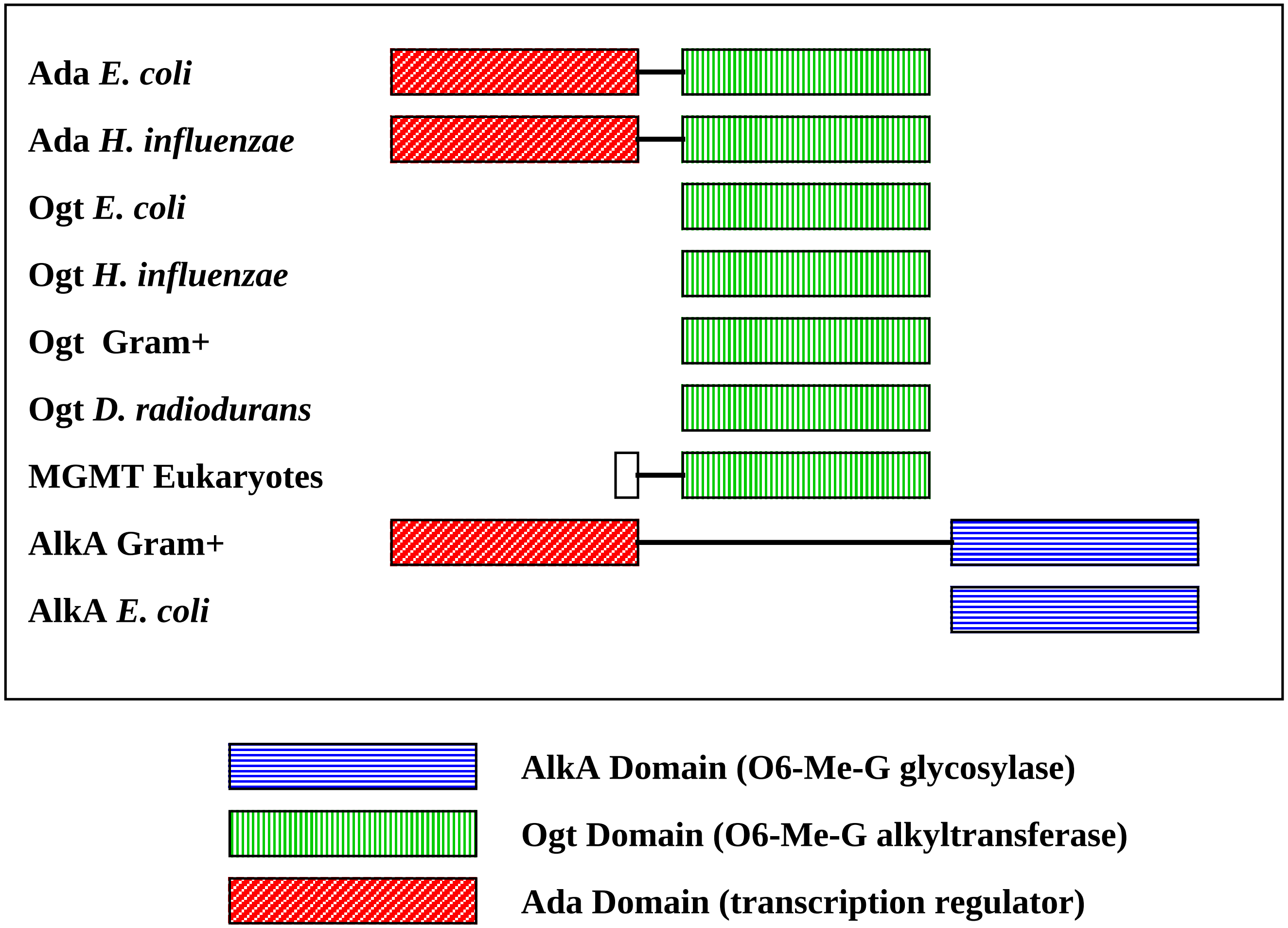
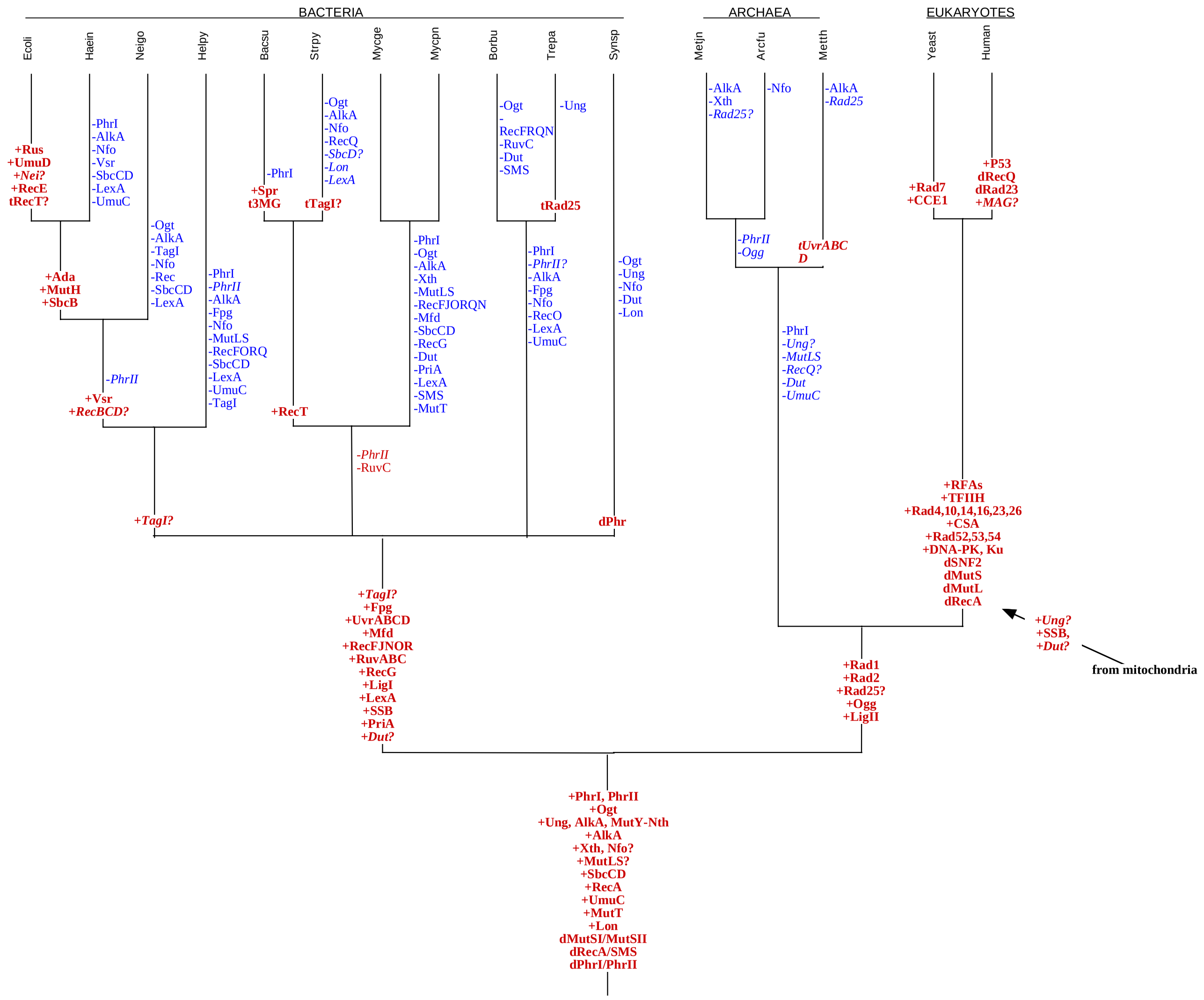

Ada *E. coli*
Ada *H. influenzae*
Ogt *E. coli*
Ogt *H. influenzae*
Ogt Gram+
Ogt *D. radiodurans*
MGMT Eukaryotes
AlkA Gram+
AlkA *E. coli*
AlkA Domain (O6-Me-G glycosylase)
Ogt Domain (O6-Me-G alkyltransferase)
Ada Domain (transcription regulator)

Figure 4. Evolutionary gain and loss of DNA repair genes.

The gain and loss of repair genes is traced onto an evolutionary tree of the species for which complete genome sequences were analyzed. Gain and loss was inferred by methods described in the main text. Origins of repair genes (+) are indicated on the branches while loss of genes (-) is indicated along side the branches. Gene duplication events are indicated by a "d" while possible lateral transfers are indicated by a "t".


BACTERIA
ARCHAEA
EUKARYOTES
Ecoli
Haein
Neigo
Helpy
Bacsu
Strpy
Mycge
Mycpn
Borbu
Trepa
Synsp
Metjn
Arcfu
Metth
Yeast
Human
+Rus
+UmuD
+Nei?
+RecE
tRecT?
-PhrI
-AlkA
-Nfo
-Vsr
-SbcCD
-LexA
-UmuC
+Ada
+MutH
+SbcB
-Ogt
-AlkA
-TagI
-Nfo
-Rec
-SbcCD
-LexA
-PhrII
+Vsr
+RecBCD?
-PhrI
-PhrII
-AlkA
-Fpg
-Nfo
-MutLS
-RecFORQ
-SbcCD
-LexA
-UmuC
-TagI
+TagI?
-PhrI
+Spr
t3MG
-Ogt
-AlkA
-Nfo
-RecQ
-SbcD?
-Lon
-LexA
tTagI?
+RecT
-PhrII
-RuvC
-PhrI
-Ogt
-AlkA
-Xth
-MutLS
-RecFJORQN
-Mfd
-SbcCD
-RecG
-Dut
-PriA
-LexA
-SMS
-MutT
-Ogt
-
RecFRQN
-RuvC
-Dut
-SMS
-Ung
tRad25
-PhrI
-PhrII?
-AlkA
-Fpg
-Nfo
-RecO
-LexA
-UmuC
-Ogt
-Ung
-Nfo
-Dut
-Lon
dPhr
+TagI?
+Fpg
+UvrABCD
+Mfd
+RecFJNOR
+RuvABC
+RecG
+LigI
+LexA
+SSB
+PriA
+Dut?
-AlkA
-Xth
-Rad25?
-Nfo
-AlkA
-Rad25
-PhrII
-Ogg
tUvrABC
D
-PhrI
-Ung?
-MutLS
-RecQ?
-Dut
-UmuC
+Rad7
+CCE1
+P53
dRecQ
dRad23
+MAG?
+RFAs
+TFIIH
+Rad4,10,14,16,23,26
+CSA
+Rad52,53,54
+DNA-PK, Ku
dSNF2
dMutS
dMutL
dRecA
+Ung?
+SSB,
+Dut?
from mitochondria
+Rad1
+Rad2
+Rad25?
+Ogg
+LigII
+PhrI, PhrII
+Ogt
+Ung, AlkA, MutY-Nth
+AlkA
+Xth, Nfo?
+MutLS?
+SbcCD
+RecA
+UmuC
+MutT
+Lon
dMutSI/MutSII
dRecA/SMS
dPhrI/PhrII

APPENDIX A

DNA Turnover,
Thymineless Death,
and
Stationary Phase Mutagenesis

"It may seem a deplorable imperfection of nature that mutability is not restricted to changes that enhance the adaptedness of their carriers. However, only a vitalist Pangloss could imagine that genes know how and when it is good for them to mutate."

T.H. Dobzhansky (1970)

SUMMARY

My initial thesis research involved experiments on DNA turnover, thymineless death (TLD), and stationary phase mutagenesis (SPM, also known as directed evolution). I suggested that these phenomena were all related as different aspects of DNA turnover in non-dividing cells. In general it is thought that DNA turnover (the replacement of small stretches of DNA without genome wide replication) is dependent on DNA repair processes. Since earlier studies showed that DNA turnover is dependent on transcription, and since some forms of DNA repair are coupled to transcription, we thought that transcription-coupled repair might be involved in DNA turnover. Unfortunately, attempts in the Hanawalt lab to study the role DNA repair processes play in DNA turnover had been hampered because turnover is difficult to quantify biochemically. I proposed to use TLD and stationary phase mutagenesis as phenotypic screens for genes that affected DNA turnover (e.g., genes involved in transcription-coupled repair) because both of these processes are thought to depend on DNA turnover. In this appendix I describe some of the reasons we were interested in these phenomena and the results of some of my experiments regarding this subject area.

INTRODUCTION TO STATIONARY PHASE MUTATIONS

The generation of heritable variation is an integral part of evolution by natural selection. Thus, even before the chemical nature of heredity was understood, there was debate over the origin of variation. Even after it was accepted that variation could arise spontaneously by mutation (as emphasized by Hugo de Vries) there were still many unanswered questions about their origin such as whether the process is random or whether it was biased towards changes beneficial for the organism (1). With better understanding of the nature of mutation, it became generally accepted that they were random and unbiased (2). This fit well with the notion that it is selection that provided the only direction to evolution. However, this notion was not actually experimentally tested until 1943.

The Dogma - Luria and Delbrück and Pre-Existing Mutations

In 1943 Luria and Delbrück published a now classic paper in which they applied a statistical test to try to determine whether mutations arose spontaneously without regard to their potential advantage or whether they arose in response to selection pressure (3). In their test they used a particular strain of the bacteria *Escherichia coli* that was sensitive to killing by a bacteriophage. The strain was grown in liquid media and a small amount of this culture was used to inoculate multiple "sister" tubes. These tubes were then incubated and the bacteria were allowed to grow to a high density. A sample of bacteria from each tube was mixed with the phage and plated. The number of colonies that grew on the plates was used as an indication of the number of resistant cells. They argued that, if the resistance to the phage arose in the sister tubes prior to exposure to the phage, then the number of resistant colonies per plate should vary greatly because it would be determined by the time in the growth of the culture that the mutation arose. If instead the mutations arose after exposure to the phage then the distribution of mutants on the plates should be narrow because all the plates would be roughly equivalent. The results of this fluctuation test were conclusive - the distribution matched the Jackpot pattern expected if the mutations arose spontaneously prior to selection. Subsequent experiments by Cavalli-Sforza (4) and Lederberg (5) using replica plating and sib-selection were able to show conclusively that mutations existed in populations prior to selection. These results led to the general belief that mutations were spontaneous and random, that they arose in dividing cells only, and that could not be directed by their potential benefit.

Over the years some of these ideas have been shown to be somewhat inaccurate. Mutations are not truly random in that they vary with many factors such as type of mutation (6) genome position (7), sequence context (8), transcriptional activity, proximity to other genes (9), and genotype of the individual. In addition, mutations are not solely spontaneous because they can be induced by environmental agents such as X-rays and ultraviolet irradiation. However, the main tenet, that mutations are random in relation to their potential benefit was generally accepted for many years.

The Heresy - Ryan, Cairns, and Directed Mutations

One person who led the challenge to the randomness of mutations was Francis Ryan (10-13). He did a series of elegant experiments studying mutation processes in non-dividing bacterial cells. He challenged many of the conclusions of the Luria and Delbrück experiments, but his results were largely ignored. That is, until a paper in Nature in 1988 by John Cairns, Julie Overbaugh, and Stephen Miller (14). These authors suggested that the experiments by Luria and Delbrück were flawed and could not possibly have detected directed mutations because any cells without pre-existing mutations were killed. Cairns et al. performed new fluctuation tests using non-lethal selections and concluded that directed mutations do occur. Their experimental plan was quite simple. In their main experiment they grew up cells with an amber mutation in the lacZ gene that prevents the normal utilization of lactose. These $lacZ_{am}$ cells were grown up in sister cultures as in Luria and Delbrück and plated onto lactose-minimal plates. Revertants to $lacZ^{+}$ would grow into colonies while $lacZ_{am}$ cells would stay in stationary phase on the plates. The distribution of colonies from sister cultures was intermediate between the jackpot pattern expected for only pre-selection mutations and the narrow pattern expected for post-selection mutations. This, and the fact that colonies continue to appear after many days on the plates led Cairns et al. to conclude that some of the mutations were arising after selection. Since the cells should be in stationary phase on the plates, they thus concluded that the mutations were arising without replication or division. They also conducted a few controls to try to better understand the process. One such control was the plating of sister cultures onto media with no lactose (nor any other sugar). They then overlayed these plates at different times with lactose agar and counted the colonies that grew. The number of colonies corresponded to the amount of time which the plates had been exposed to lactose, not the amount of time without it. From this they concluded that the mutations occurred only when the selection was present. In addition, they compared the number of lac revertants over time with the number of cells with mutations in other genes. They did this by overlaying plates with valine media. Any colonies would be from cells that had mutations to valine resistance (val^{R}). The number of valine resistant colonies was much lower than the number of lac^{+} colonies. From this they concluded that adaptive mutations were increased specifically and thus that this represented a case of directed evolution.

This paper caused an enormous furor in the scientific community. Major science journals wrote news articles about it giving it the status of heresy in evolutionary thought (15-17). It did not lessen the controversy that Cairns was willing to suggest that this represented some type of neo-Lamarckian inheritance. Cairns proposed a model to explain the phenomena involving reverse transcription of the RNAs that coded for useful proteins. Soon after that paper, other researchers presented data with similar patterns of beneficial mutations apparently only accumulating in the presence of selection and at a higher rate than non-beneficial mutations (e.g., (18,19)). The phenomenon also was shown to occur in other bacteria (20) and in yeast (21,22). Cairns and Foster followed up their work with new experiments in which directed mutations are still proposed to occur (23-25).

The Skeptics

Despite the support of many researchers, many others remained skeptical. Some proposed mechanisms that explain Cairns results without invoking directionality to the mutation process itself. For example, Stahl proposed that some DNA changes were occurring in stationary phase and that most such changes would be repaired by correction mechanisms. However, changes that altered the genotype of the cell to something that could better use lactose might out run the repair system by allowing for replication of the genome (26,27). Even Cairns original mechanism if it were to be occurring would not be truly directed mutation but selection at the molecular level. Many other potential problems existed with the Cairns and Foster experiments

Mutations in Non-Dividing Cells: DNA Turnover and Thymineless Death

However, regardless of whether the nature of Cairnsian mutations is directed or spontaneous, one thing that was clear to us at the time was that little was known about the origin of mutations in non-dividing cells. Ryan was the first to provide evidence that mutations could accumulate in non-dividing cells. He showed that DNA replication was limited in these conditions and thus suggested that the mutations arose by some type of DNA turnover. He never did show that turnover was responsible for the mutations.

However, soon after, DNA turnover was documented in response to DNA damage (28,29). It is now known that many repair processes result in the turnover of small stretches of DNA. In addition it is known that DNA turnover can occur without any known DNA damage (30-32). What causes this turnover, and how it varies within a genome or under different conditions is unknown. Also unknown is whether this type of turnover is responsible for mutations.

DNA turnover may be involved in a variety of biological phenomena. One such phenomenon is thymineless death (TLD). TLD is the loss of viability in growing cells when the availability of thymine is inhibited. This phenomenon was first documented in *E. coli* (33). It has subsequently been shown to occur in many species of bacteria and eukaryotes (e.g., mycoplasmas (34); *Deinococcus radiodurans* (35,36), B. subtilis (37-40), *S. aureus* (41), lactobacilli (42), yeast (43,44), Candida (45), and human cells (46-49)).

Despite the universality of TLD, its mechanism is not well understood. Much of the original information concerning TLD involved correlating it with other things going on in the cell. Simultaneous to TLD many thing occur including synthesis of a variety of proteins (50), decrease in RNA synthesis (51), DNA damage accumulates (52), mutation increases (53), recombination increase (54) colicin production (55), prophage induction (56), DNA turnover increase (57). Overall, many of the phenomena associated with TLD are also associated with UV irradiation (58). Other studies focused on the factors that were required for TLD to occur which included RNA synthesis (59), presence of all required amino-acids in the media (60) and a carbon source (33) and active growth.

A great deal of information about the mechanism of TLD has come from genetic studies. TLD can be induced by mutations in genes relating to thymine metabolism (e.g., *thyA* (60)) and by chemicals that inhibit thymine incorporation into DNA (e.g., cytosine arabinoside (61)). Such chemicals have proven useful as anticancer therapies and antibiotics. TLD can be inhibited by mutations in other genes such as uracil glycosylase (as shown for *B. subtilis* (62)) and some genes involved in DNA replication and recombination (e.g., (63)) especially some of those in the RecF pathway (*recQ, recF, recJ* and *recO* but not *recN* (54,64-66).

The commonly accepted model is that TLD results from the incorporation of excessive amounts of uracil into DNA (from dUTP in the absence of dTTP) and the subsequent overcleavage of the genome by uracil-DNA glycosylase (a model outlining a likely scenario for TLD, based in part on (67,68) is shown in Figure 1). Genes known to be involved in TLD with steps that they may be involved in are identified. Despite the potential importance of this process it is not known what causes the excessive incorporation of uracil into the DNA. It was our belief that this incorporation of uracil is dependent upon DNA turnover. One reason for this belief was that, like DNA turnover, TLD is dependent upon transcription (30). Also, as shown by Nakayama and Hanawalt, the size distribution of DNA fragments in alkaline sucrose gradients decreased during the period of incubation of thymine-requiring *E. coli* without thymine (69). Because TLD and SPM were both thought to involve DNA turnover, we believed that TLD and SPM could be used as genetic screens to identify molecular mechanisms underlying DNA turnover.

SUMMARY OF TLD EXPERIMENTS

We were particularly interested in whether genes involved in transcription-coupled repair were involved in TLD. Other studies have shown that in *E. coli* transcription-coupled repair is dependent on the *mfd* gene (70), the genes in the uvrABC pathway, and the *mutL* and *mutS* genes (71). I created appropriate strains with mutations in *mfd*, *uvrC* , *mutL* and *mutS* (see Table 1 for a listing of strains). None of these strains showed any significant changes in the sensitivity to thymine deprivation (see Fig. 2 for outline of method used to study TLD) relative to their isogenic parent strains (see Fig. 3). In addition, sensitivity to thymine deprivation (Phil Hanawalt, unpublished). Possible explanations for the absence of an effect of these genes on TLD include 1) TLD is not dependent on DNA turnover (it may be replication dependent) 2) TLD is dependent on DNA turnover but under these conditions transcription coupled repair is not a significant contributor to turnover 3) transcription-coupled repair operates under a different pathway in the conditions used in this experiments.

SUMMARY OF STATIONARY PHASE MUTAGENESIS EXPERIMENTS

I created strains to study the role transcription coupled repair plays in SPM. I also began to conduct a variety of other experiments designed to determine if the SPM phenomenon was real or an artifact. During this time a series of papers was published pointing out serious flaws in the SPM experiments suggesting that the phenomena may be an artifact. Papers continue to be published pointing out problems in the initial ideas about SPM (72-76). I decided that I did not want a large part of my thesis to represent control experiments for someone else research. Since the TLD and SPM experiments seemed to have stalled, and since attempts to develop a biochemical assay for DNA turnover had not progressed any further, I decided to tackle a new project. However, I present here the results of one interesting discovery I made concerning SPM and UV irradiation (Fig. 4). In these experiments, I irradiated *E. coli* cells after plating onto selective media. Since the cells were not supposed to be growing or dividing on these plates unless they reverted to lac$^+$, I thought that the irradiation should have little effect on the number of mutants that arose. To my surprise, irradiation led to an increase in the number of revertants (as long as the levels of UV did not kill too many cells) and this increase was dose dependent.

ACKNOWLEDGEMENTS

Although I never made enormous progress on solving the mechanisms of DNA turnover, thymineless death and/or stationary phase mutagenesis, I never would have gotten anywhere without the help of many people. In particular I wish to thank Phil Hanawalt, Shi-Kau Liu, Allan Campbell, and Richard Lenski for helpful discussions, everyone who provided me strains to use for these studies, and Ann Ganesan for help constructing strains.

REFERENCES

1. Mayr, E. (1982) The growth of biological thought, The Belknap Press of Harvard University Press, Cambridge, MA.
2. Lederberg, J. (1989) *Genetics*, **121,** 395-399.
3. Luria, S. and Delbrück, M. (1943) *Genetics*, **28,** 491-511.
4. Cavalli-Sforza, L. and Lederberg, J. (1956) *Genetics*, **41,** 367-381.
5. Lederberg, J. and Lederberg, E. (1952) *J. Bacteriol.*, **63,** 399-406.
6. Li, W.-H., Wu, C.-I. and Luo, C.-C. (1984) *J. Mol. Evol.*, **21,** 58-71.
7. Sharp, P. M. and Li, W. H. (1989) *J Mol Evol*, **28,** 398-402.
8. Kunkel, T. (1992) *Bioessays*, **14,** 303-308.
9. Liu, S.-K. and Tessman, I. (1990) *J. Bacteriol.*, **172,** 6135-6138.
10. Ryan, F., Nakada, D. and Schneider, M. (1961) *Z. Verebungsl.*, **92,** 38-41.
11. Ryan, F. (1954) *Proc. Natl. Acad. Sci. U.S.A.*, **40,** 178-186.
12. Ryan, F. (1955) *Genetics*, **40,** 726-738.
13. Ryan, F. (1959) *J. Gen. Microbiol.*, **21,** 530-549.
14. Cairns, J. (1988) *Nature*, **336,** 527-528.
15. Symonds, N. (1989) *Nature*, **337,** 119-120.
16. Lewin, R. (1988) *Science*, **241,** 1431.
17. Cherfas, J. (1988) *New Scientist*, **119,** 34-35.
18. Hall, B. G. (1990) *Bioessays*, **12,** 551-558.
19. Hall, B. G. (1991) *Proc Natl Acad Sci U S A*, **88,** 5882-5886.
20. Gizatullin, F. and Lyozin, G. (1992) *Res. Microbiol.*, **143,** 711-719.
21. Steele, D. and Jinks-Robertson, S. (1992) *Genetics*, **132,** 9-21.
22. Hall, B. G. (1992) *Proc Natl Acad Sci U S A*, **89,** 4300-4303.
23. Cairns, J. and Foster, P. (1991) *Genetics*, **128,** 695-701.
24. Foster, P. (1992) *J. Bacteriol.*, **174,** 1711-1716.
25. Foster, P. and Cairns, J. (1992) *Genetics*, **131,** 783-789.
26. Stahl, F. (1992) *Genetics*, **132,** 865-867.
27. Stahl, F. (1988) *Nature*, **335,** 112.
28. Pettijohn, D. and Hanawalt, P. (1964) *J. Mol. Biol.*, **9,** 395-410.
29. Couch, J. and Hanawalt, P. (1967) *Bichem. Biophys. Res. Commun.*, **29,** 779-784.
30. Grivell, A., Grivell, M. and Hanawalt, P. (1975) *J. Mol. Biol.*, **98,** 219-233.
31. Hanawalt, P., Grivell, A. and Nakayama, H. (1975) *Basic Life Sci.*, **5,** 47-51.
32. Tang, M.-S., Wang, T.-C. V. and Patrick, M. (1978) *Photochem. Photobiol.*, **29,** 511-520.
33. Cohen, S. and Barner, H. (1954) *Proc. Natl. Acad. Sci. U.S.A.*, **40,** 885-893.
34. Smith, D. W. and Hanawalt, P. C. (1968) *J Bacteriol*, **96,** 2066-2076.

35. Little, J. G. and Hanawalt, P. C. (1973) *J Bacteriol*, **113,** 233-240.
36. Rauko, P., Budayova, E. and Sedliakova, M. (1976) *Folia Microbiol*, **21,** 438-443.
37. Buick, R. N. and Harris, W. J. (1972) *Biochem J*, **129,** 49.
38. Buick, R. W. and Harris, W. J. (1975) *J Gen Microbiol*, **88,** 115-122.
39. Ephrati-Elizur, E., Yosuv, D., Shmueli, E. and Horowitz, A. (1974) *J Bacteriol*, **119,** 36-43.
40. Rolfe, R. (1967) *Proc Natl Acad Sci U S A*, **57,** 114-121.
41. Mathieu, L. G., Repentigny, J. d., Turgeon, S. and Sonea, S. (1968) *Can J Microbiol*, **14,** 983-987.
42. Reich, J. and Soska, J. (1967) *Biochem Biophys Res Commun*, **29,** 62-67.
43. Barclay, B. J., Kunz, B. A., Little, J. G. and Haynes, R. H. (1982) *Can J Biochem*, **60,** 172-184.
44. Brendel, M. and Langjahr, U. G. (1974) *Mol Gen Genet*, **131,** 351-358.
45. Henson, O. E. and McClary, D. O. (1979) *Antonie Van Leeuwenhoek*, **45,** 211-223.
46. Houghton, J. A., Harwood, F. G. and Tillman, D. M. (1997) *Proc Natl Acad Sci U S A*, **94,** 8144-8149.
47. Ingraham, H. A., Dickey, L. and Goulian, M. (1986) *Biochemistry*, **25,** 3225-3230.
48. Kyprianou, N. and Isaacs, J. T. (1989) *Biochem Biophys Res Commun*, **165,** 73-81.
49. Seno, T., Ayusawa, D., Shimizu, K., Koyama, H., Takeishi, K. and Hori, T. (1985) *Basic Life Sci*, **31,** 241-263.
50. Dankberg, F. and Cummings, D. J. (1973) *J Bacteriol*, **113,** 711-717.
51. Medoff, G. (1972) *J Bacteriol*, **109,** 462-464.
52. Mennigmann, H. D. and Carmona, M. T. (1976) *J Bacteriol*, **125,** 1232-1234.
53. Smith, M. D., Green, R. R., Ripley, L. S. and Drake, J. W. (1973) *Genetics*, **74,** 393-403.
54. Nakayama, K., Kusano, K., Irino, N. and Nakayama, H. (1994) *J Mol Biol*, **243,** 611-620.
55. Mennigmann, H. D. (1964) *Biochem Biophys Res Commun*, **16,** 373-378.
56. Melechen, N. E., Go, G. and Lozeron, H. A. (1978) *Mol Gen Genet*, **163,** 213-221.
57. Pauling, C. and Hanawalt, P. C. (1965) *Proc Natl Acad Sci U S A*, **54,** 1728-1735.
58. Gallant, J. and Suskind, S. R. (1961) *J Bacteriol*, **82,** 187–194.
59. Hanawalt, P. C. (1963) *Nature*, **198,** 286.
60. Barner, H. D. and Cohen, S. S. (1954) *J Bacteriol*, **68,** 80-88.
61. Atkinson, C. and Stacey, K. A. (1968) *Biochim Biophys Acta*, **166,** 705-707.
62. Makino, F. and Munakata, N. (1978) *J. Bacteriol.*, **134,** 24-29.
63. Bouvier, F. and Sicard, N. (1975) *J Bacteriol*, **124,** 1198-204.
64. Nakayama, H., Nakayama, K., Nakayama, R. and Nakayama, Y. (1982) *Can. J. Microbiol.*, **28,** 425-430.
65. Nakayama, H., Nakayama, K., Nakayama, R., Irino, N., Nakayama, Y. and Hanawalt, P. (1984) *Mol Gen Genet*, **195,** 474-480.

66. Nakayama, K., Shiota, S. and Nakayama, H. (1988) *Can J Microbiol*, **34,** 905-907.
67. Kunz, B. (1982) *Environ. Mutagen.*, **4,** 695-725.
68. Haynes, R. (1985) In de Sorres, F. (ed.), Genetic Consequences of Nucleotide Pool Imbalances. Plenum Publishing Corp., New York.
69. Nakayama, H. and Hanawalt, P. (1975) *J Bacteriol*, **121,** 537-547.
70. Selby, C. P., Witkin, E. M. and Sancar, A. (1991) *Proc Natl Acad Sci U S A*, **88,** 11574-8.
71. Mellon, I. and Champe, G. N. (1996) *Proc Natl Acad Sci U S A*, **93,** 1292-1297.
72. Lenski, R., Slatkin, M. and Ayala, F. (1989) *Proc. Natl. Acad. Sci. U.S.A.*, **86,** 2775-2778.
73. Lenski, R., Slatkin, M. and Ayala, F. (1989) *Nature*, **337,** 123-124.
74. Sniegowski, P. D. (1995) *J Mol Evol*, **40,** 94-101.
75. Lenski, R. E. and Sniegowski, P. D. (1995) *Curr Biol*, **5,** 97-99.
76. MacPhee, D. G. and Ambrose, M. (1996) *Genetica*, **97,** 87-101.

Table 1. Strains used in study of thymineless death and stationary phase reversion.

Name	Main Genotype	Origin	Comments
HL758	uvrA::tn10		Checked uvrA genotype with UV sensitivity.
SM195 (HL681)		P. Foster	
SM196 (HL682)		P. Foster	
SM195 uvrA::tn10	uvrA::tn10, ΔuvrB-bio	P1 from HL758	Confirmed Bio-
SM196 uvrA::tn10	uvrA::tn10	P1 from HL758	Confirmed Bio+
HL660	zcf117::tn10 (linked to *mfd*-)		
J1-1 to J1-10	thyA, mfd?	P1 from HL660 x AB2497	Attempted to confirm mfd with UV sensitivity but inconclusive.
CAG12156	uvrC::tn10	Carol Gross	
J2	thyA, uvrC::tn10	P1 CAG12156 x AB2497	Confirmed uvrC genotype with UV sensitivity
HL789	mutS::tn5	M.Marinus(ES1574)	
HL785	mutS::tn10	M.Marinus (GM2165)	
HL786	mutL::tn10	M.Marinus(GM2166)	
J3	thyA, mutS::tn5	P1 HL789 x AB2497	Confirmed pro-, thy-
J5	thyA, mutS::tn10	P1 HL785 x AB2497	Confirmed pro-, thy-, mutator phenotypes.
J6	thyA, mutL::tn10	P1 HL786 x AB2497	Confirmed pro-, thy-, mutator phenotypes.
NR10121, NR10122	zcf117:tn10, mfd-	from A. Oller	Confirmed thy+
NR10125, NR10126	zcf117:tn10, Mfd+	from A. Oller	Confirmed thy+
J4-1 to J4-15	thyA, mfd?	P1 NR10121 x AB2497	Confirmed pro-, thy-. Attempted to confirm mfd with UV sensitivity but inconclusive.
J7	thyA, Mfd+?	P1 NR10125 x AB2497	Attempted to confirm mfd with UV sensitivity but inconclusive.
HL771	ΔthyA::kanR	from M. Belfort	
J8	ΔthyA::kanR, *mfd*-	P1 HL771 x NR10121	Confirmed thy-, pro-, leu+
J9	ΔthyA::kanR, *mfd*+	P1 HL771 x NR10125	Confirmed thy-, pro-, leu+

Figure 1. Model of thymineless death.

When thymine is removed from the growth media of *thyA* mutants, they lose viability over time (thymineless death). The first step in this process is likely the depletion of thymine pools in the cell. One way for this to occur is by mutations in the *thyA* gene. After this, uracil begins to accumulate in the DNA, either by DNA turnover (and incorporation of dUTP at sites where dTTP would have been used) or by deamination of cytosine. TLD only occurs if cells are actively growing and transcribing when the thymine is removed from the media. We believe that these processes contribute to DNA turnover. With more and more turnover, uracil will continue to accumulate in the DNA. It is believed that the cleavage of this uracil leads to TLD because uracil glycosylase mutants (*ung*) are resistant to TLD. Similarly, AP endonuclease mutants are also resistant to TLD. Genes in the RecF pathway (including *recF*, *recJ* and *recQ*) are also involved in TLD. We hypothesis that these genes are involved in making the Ung-Ape induced nicks particularly lethal to the cell. Our main focus was determining if genes involved in transcription-coupled repair might also affect this turnover.

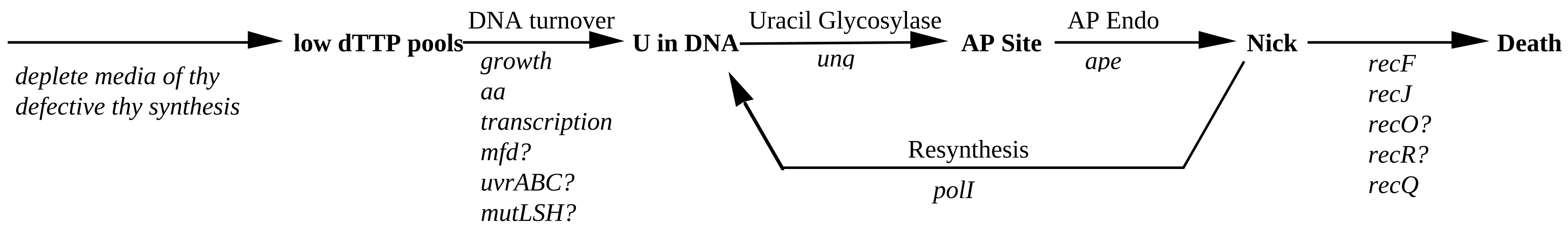

low dTTP pools
deplete media of thy
defective thy synthesis
DNA turnover
growth
aa
transcription
mfd?
uvrABC?
mutLSH?
U in DNA
Uracil Glycosylase
unq
AP Site
AP Endo
ape
Nick
Resynthesis
polI
Death
recF
recJ
recO?
recR?
recQ

Figure 2. Protocol for studying the effects of thymine deprivation (e.g., thymineless death).

Thymine-less Death

Jonathan A. Eisen - Hanawalt Lab

1) grow thy- cells to **mid-log** (very important) phase in thy rich media (+ gluc, other reqmts)
2) filter 5.0 mls cells through 0.2 um filter
3) resuspend cells in minimal media plus

    -glucose 0.4% (very important)
    -required a.a. (very important)
    -no thymine

4) incubate x 37°C x shaker
5) remove 100 ul cells at time points (including t=0, t=20 min., and t=50 min.)
6) make serial dilution

	A	B	C	D	E	F
cells	100 ul	5 ul A	5 ul B	5 ul C	5ul D	5 ul E
media (as in 3)	0.0 ul	95 ul	95 ul	95 ul	95 ul	95 ul

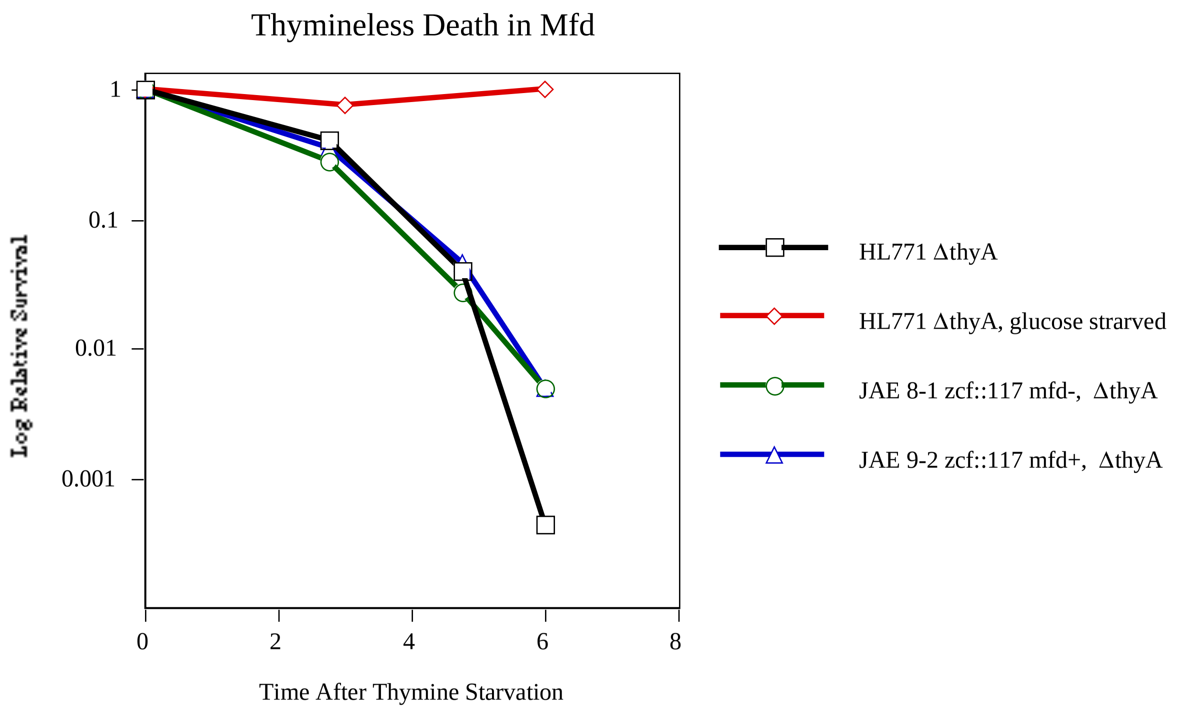
7) pipette 10 ul of each dilution 3X on non-selective plates (e.g. LB + thy) and selective plates (minimal without thy)
8) dry and incubate @ 37°C
9) count # of colonies per drop. Reversions to thy+ will show up on selective plates.

Controls:

    1) score death/growth in presence of thymine
    2) check for reversion to thy+

Figure 3. Defects in *mutL*, *mutS* and *mfd* do not affect thymineless death.

Strains are described in more detail in Table 1. Protocol for thymineless death was used as outlined in Figure 2. A. Thymineless death in *mfd* mutants. Note, if glucose is not added during thymine starvation, cells do not die TLD. B. Thymineless death in *mutL* and *mutS* mutants.


Thymineless Death in Mfd
Log Relative Survival
1
0.1
0.01
0.001
0
2
4
6
8
Time After Thymine Starvation
HL771 ΔthyA
HL771 ΔthyA, glucose strarved
JAE 8-1 zcf::117 mfd-, ΔthyA
JAE 9-2 zcf::117 mfd+, ΔthyA

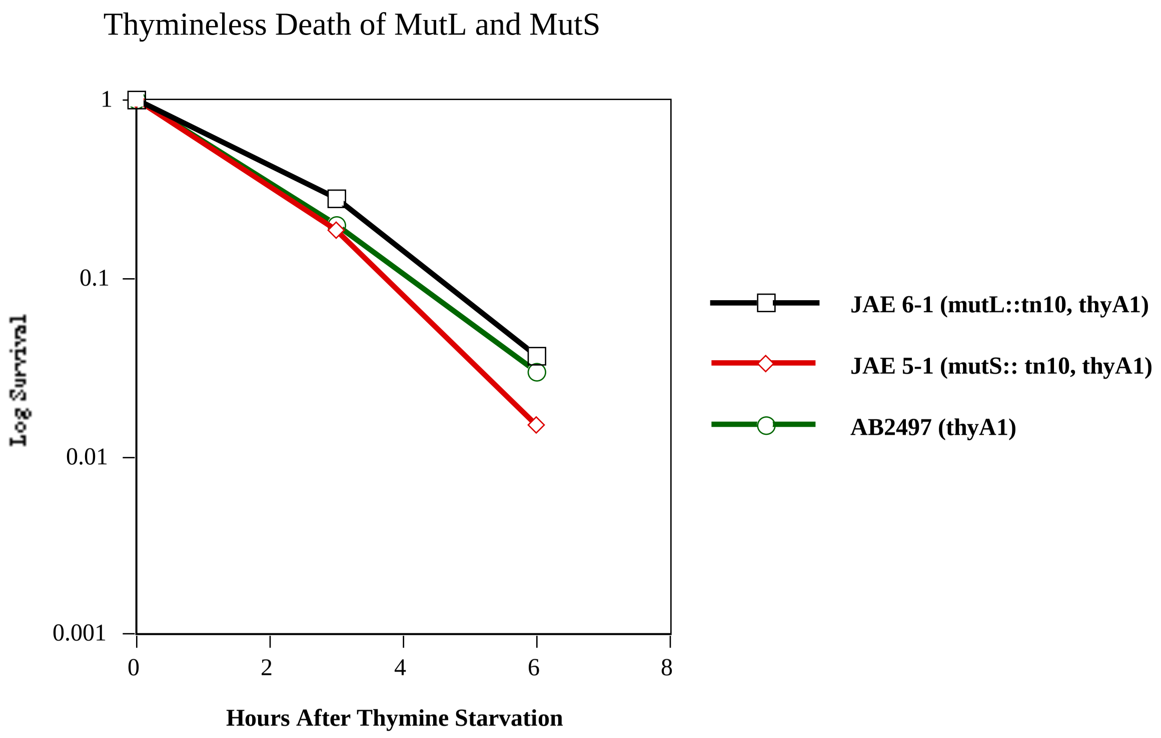
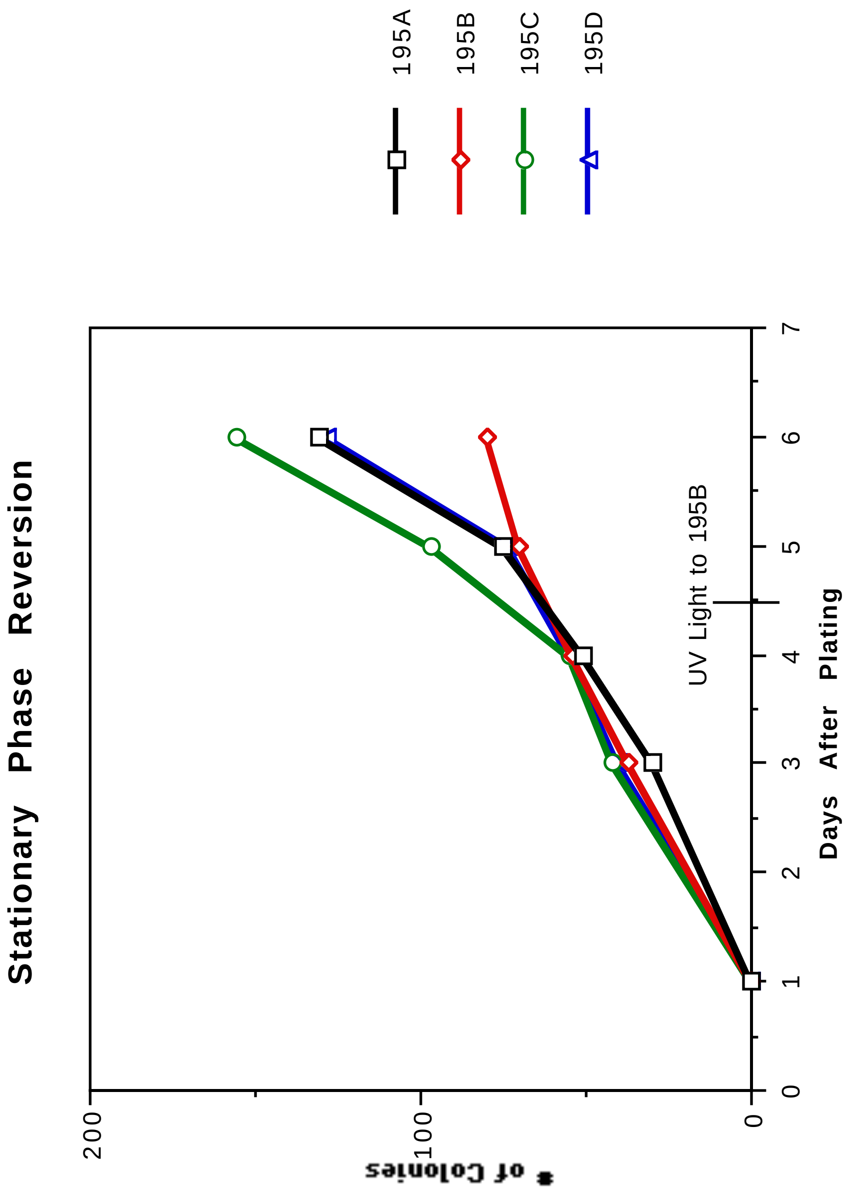
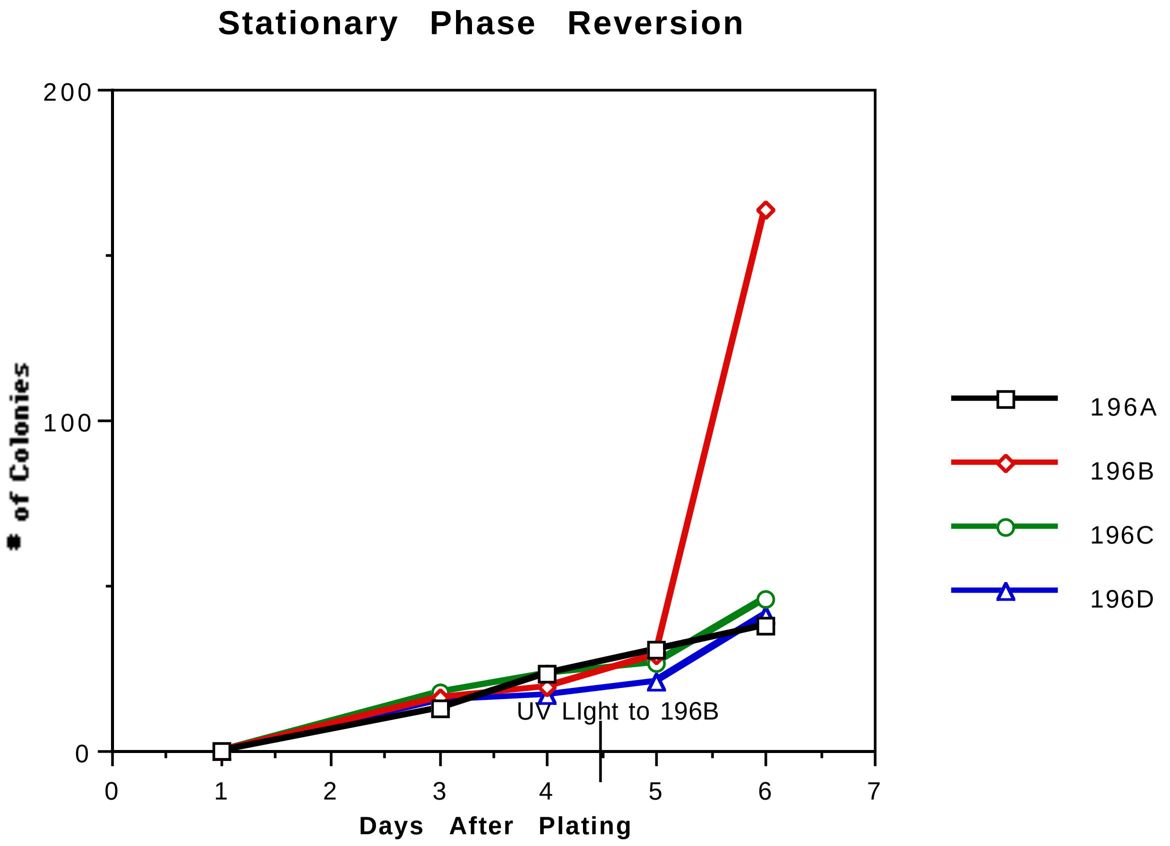
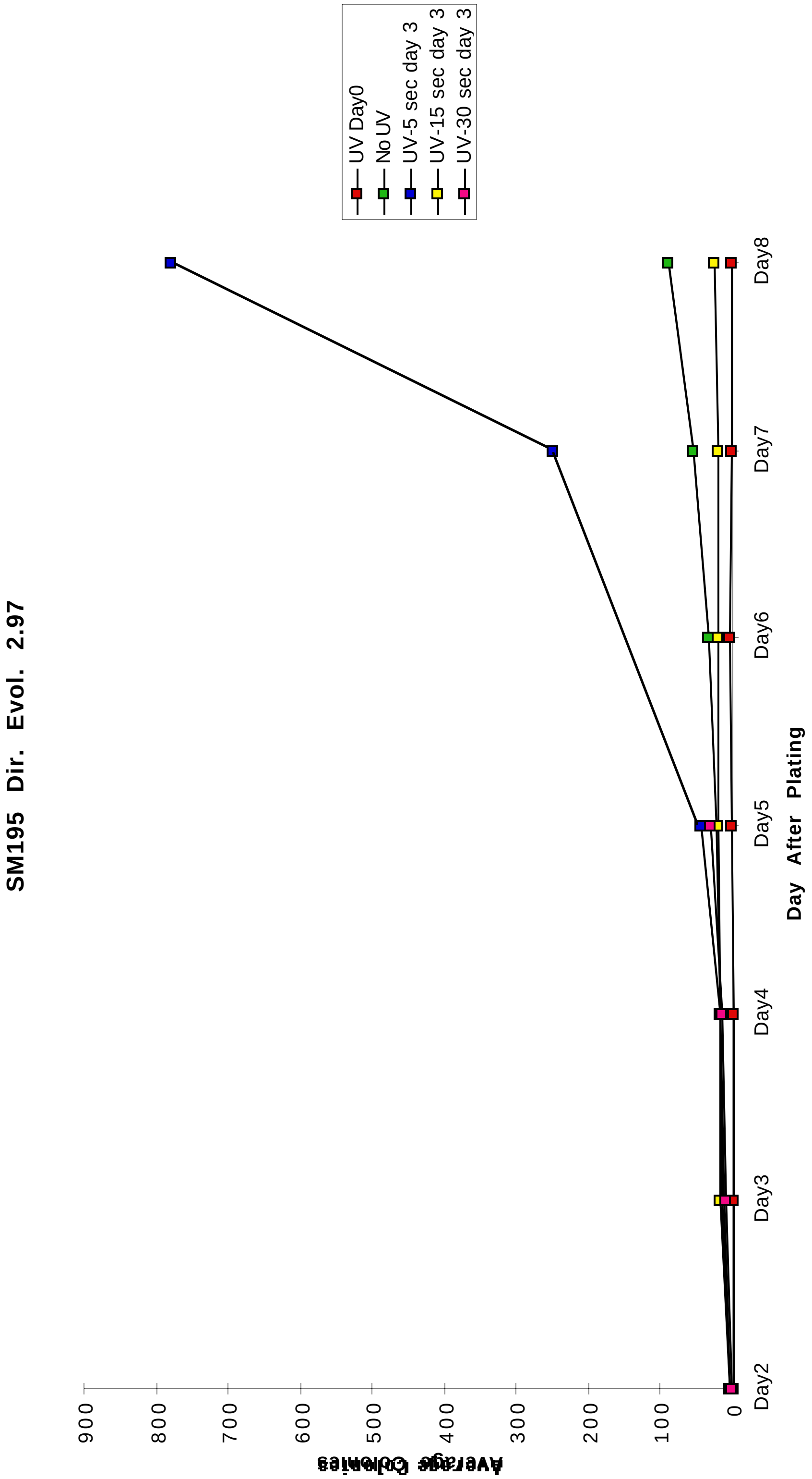

Thymineless Death of MutL and MutS
1
0.1
0.01
0.001
Log Survival
0
2
4
6
8
Hours After Thymine Starvation
JAE 6-1 (mutL::tn10, thyA1)
JAE 5-1 (mutS:: tn10, thyA1)
AB2497 (thyA1)

Figure 4. Effects of UV irradiation after plating on stationary phase reversion of SM195 and SM196.

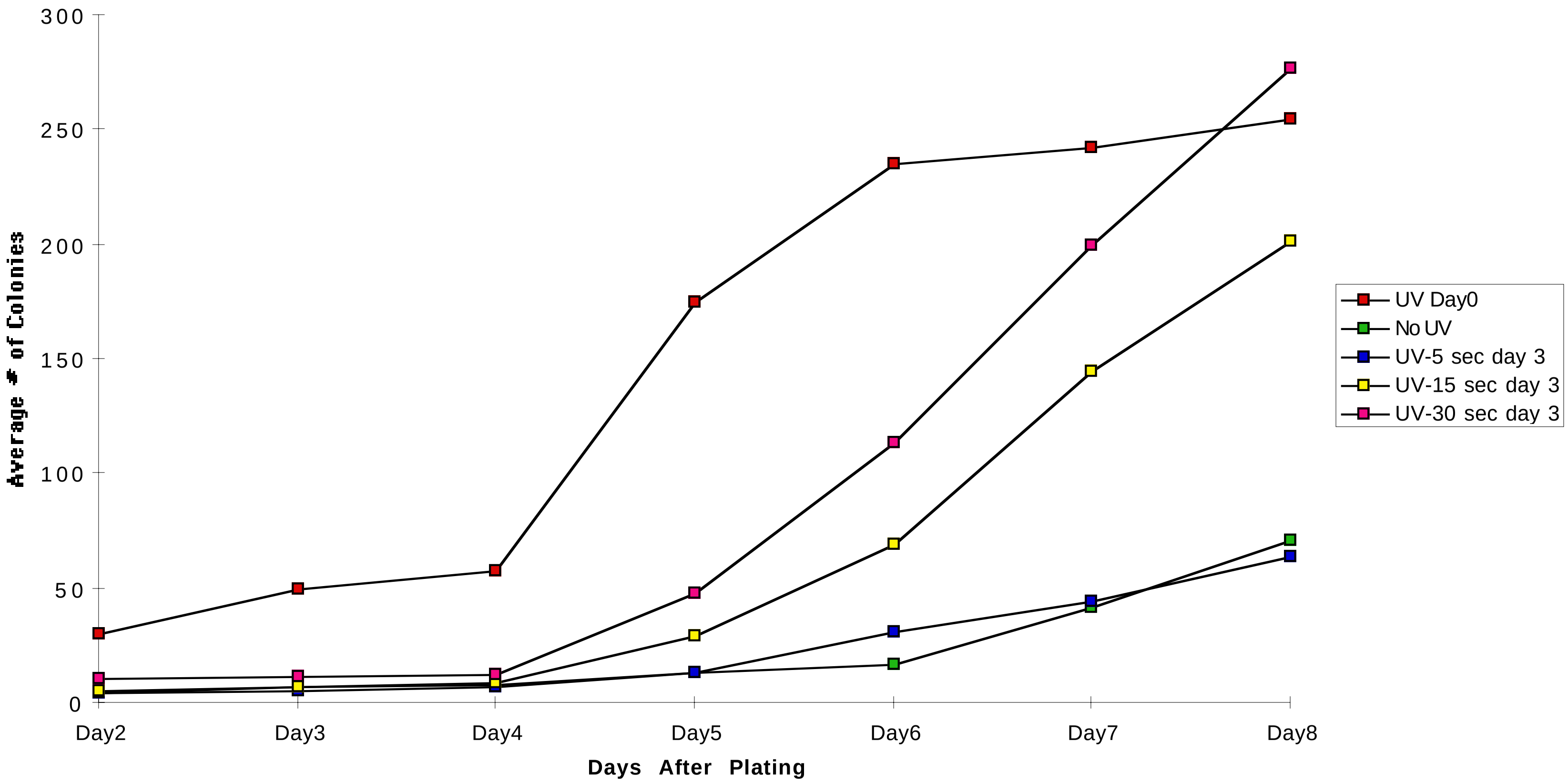
Strains are described in more detail in Table 1. A. B. Four samples of 200 ul of a single stationary phase culture of SM195 or SM196 were plated onto Davis-Lac-Thi-Bio plates (without glucose). Cells were allowed to grow at 37°C for many days. At day 4.5, one plate (labeled B) was exposed to ~45 J/m^2 UV irradiation and then returned to the incubator (wrapped in foil to prevent photoreactivation). C. D. Five samples of 200 ul of a single stationary phase culture of SM195 were plated onto Davis-Lac-Thi-Bio plates (without glucose). Cells were allowed to grow at 37°C for many days. Plates were exposed to either no UV, 30 seconds at day 0 (30 seconds = ~22 J/m^2); 5 seconds at day 3; 15 seconds at day 3; or 30 seconds at day3. Note how the 15 and 30 second doses killed the *uvrB*- SM195 but the low-dose of 5 seconds (~ 3.75 J/m^2) did not and led to an increase in number of mutants. Note how the 5 second dose did not lead to a large increase in the number of revertants per day, but the 15 and 30 second doses led to an increase at this was dose dependent. These results show that even while in "stationary phase" UV irradiation can stimulate mutagenesis.


Stationary Phase Reversion
# of Colonies
200
100
0
Days After Plating
0
1
2
3
4
5
6
7
UV Light to 195B
195A
195B
195C
195D


Stationary Phase Reversion
200
100
0
# of Colonies
0
1
2
3
4
5
6
7
Days After Plating
UV LIght to 196B
196A
196B
196C
196D


SM195 Dir. Evol. 2.97
UV Day0
No UV
UV-5 sec day 3
UV-15 sec day 3
UV-30 sec day 3
Average Colonies
900
800
700
600
500
400
300
200
100
0
Day2
Day3
Day4
Day5
Day6
Day7
Day8
Day After Plating

SM196 Dir. Evol. 2.97



APPENDIX B

Supplements to *recA1202* Study

Figure 1. Diagram of RecA structural information.

Numbers at the top refer to *E. coli* amino acid residue. First row - *E. coli* RecA secondary structure from crystal - dark boxes are β-sheets (numbered 0-10), medium shaded boxes are α-helices (lettered A-J), and lightly shaded regions are disordered. Second row - residues involved in intermonomer (IM) contact within a filament. Third row - conservation of RecA sequence within bacteria - dark shading >90% of all bacteria are identical at that residue, light shading >90% have similar amino acids. Fourth row - residues involved in contact between filaments (IF). At the bottom are all the reported single site mutants in *E. coli* RecA with arrows pointing to the residue mutated. For residues mutated in multiple alleles, the number of different alleles is indicated after dashes.

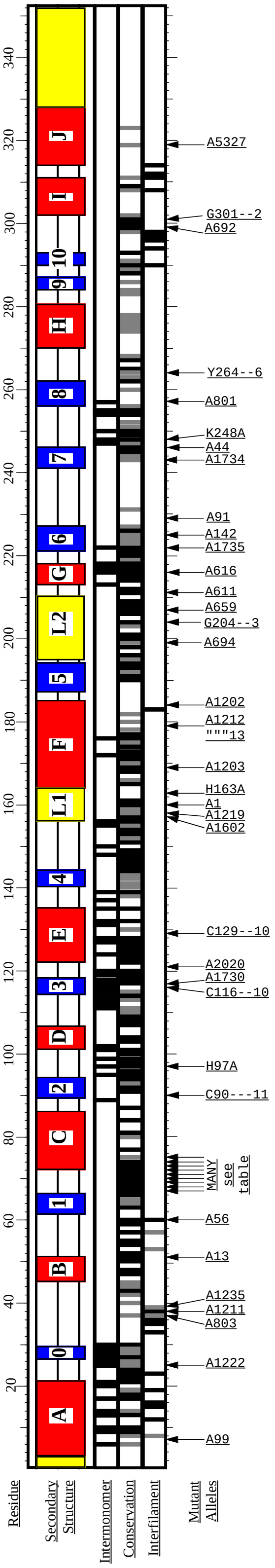
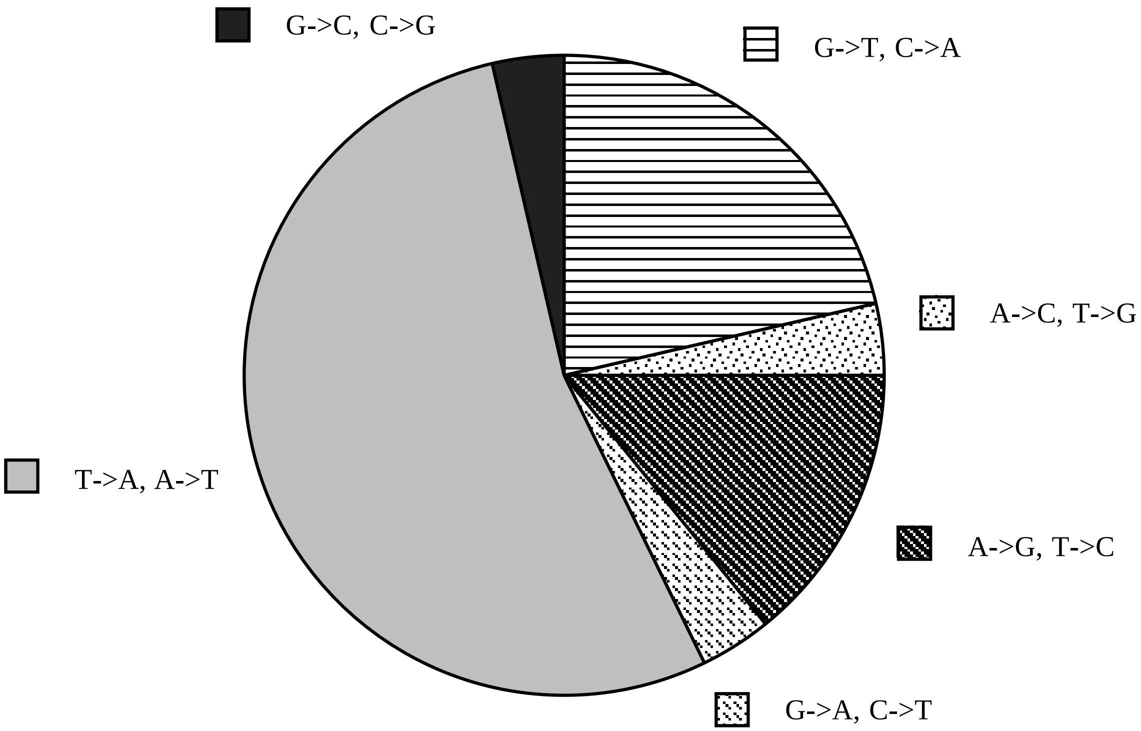

Residue
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
Secondary Structure
A
0
B
1
C
2
D
3
E
4
L1
F
5
L2
G
6
7
8
H
9
10
I
J
Intermonomer
Conservation
Interfilament
Mutant Alleles
A5327
G301--2
A692
Y264--6
A801
K248A
A44
A1734
A91
A142
A1735
A616
A611
A659
G204--3
A694
A1202
A1212
"""13
A1203
H163A
A1
A1219
A1602
C129--10
A2020
A1730
C116--10
H97A
C90---11
MANY see table
A56
A13
A1235
A1211
A803
A1222
A99

Figure 2. Mutation spectrum of 2nd site suppressor mutations of *recA1202*.

The DNA sequence changes of the second site suppressor mutations of *recA1202* are summarized here. These may indicate the mutation patterns of proximal mutagenesis.



RecA1202 Second Site Mutations

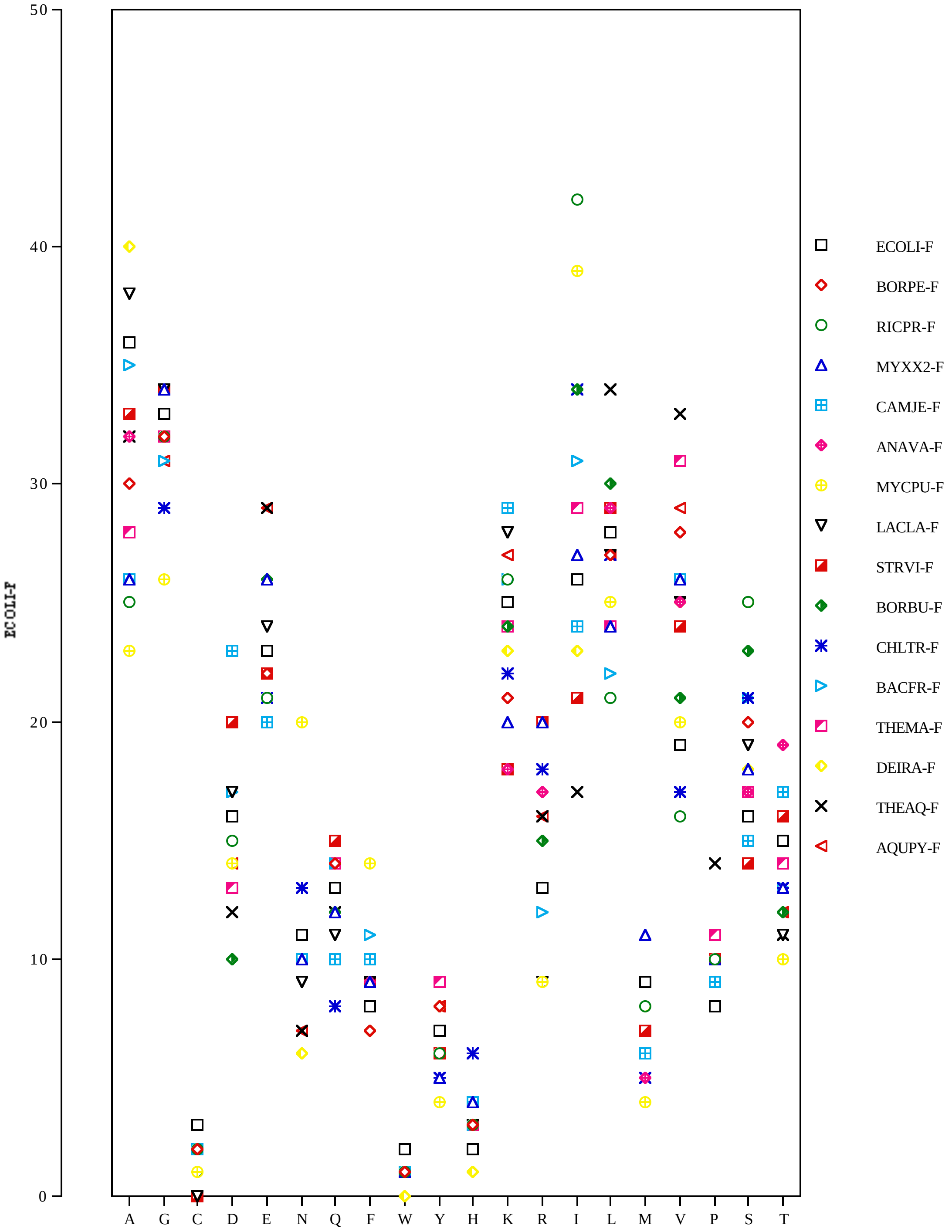
APPENDIX C

RecA Structure-Function Analysis

SUMMARY

This appendix reports unpublished results concerning evolutionary analysis of RecA sequences. The main point of this analysis was to use studies of evolutionary substitution patterns to better understand structural and functional properties of the RecA protein. The sequences analyzed are the same as those in Chapter 2a. Amino-acid substitution analysis was done as described in that chapter. Structural information came from the structures of Story and Steitz (see Chapter 2a for references). In Figure 1 and 2 information on the frequencies of different amino-acids in RecA sequences is presented. This is important because the amino-acid frequencies give a picture of the total "phenotype" of the RecA proteins in different species and it turns out that this phenotype is correlated with some other cellular features. In particular, some aspects of the frequencies of different amino-acids are correlated with the organism's GC content. Specifically, classes of amino-acids are kept at almost constant frequency in different RecAs (Figure 3) but the choice of which amino-acid to use within these classes is correlated with GC content in many cases (Figures 4 and 5). I believe that this is due to selection for GC content driving amino-acid evolution. In these cases, the selection for GC content appears to lead to using amino-acids whose codons have the right GC content. Another way to look at amino-acid evolution is to study amino-acid changes over evolutionary time (Figures 6-8). Such studies provide useful information because they examine how proteins have changed over time and not just which parts are kept conserved (which is what standard sequence comparisons reveal). In this case, perhaps the most interesting result is that the ratio of conservative to non-conservative amino-acid changes varies greatly within the RecA primary and secondary structure. When this ratio is high, amino-acids are changing but only among similar amino-acids (high number of conservative changes). Thus, this analysis allows one to distinguish sites at which the selection is for which amino-acid to use from those for which the selection is for the class of amino-acid (e.g., hydrophobic) but for which the particular amino-acid within that class does not matter.

Figure 1. Frequencies of amino-acids in representative RecA sequences.



50
40
30
20
10
0
ECOLI-F
A
G
C
D
E
N
Q
F
W
Y
H
K
R
I
L
M
V
P
S
T
ECOLI-F
BORPE-F
RICPR-F
MYXX2-F
CAMJE-F
ANAVA-F
MYCPU-F
LACLA-F
STRVI-F
BORBU-F
CHLTR-F
BACFR-F
THEMA-F
DEIRA-F
THEAQ-F
AQUPY-F

Figure 2. Average frequencies of amino-acids in all RecA sequences.

The average frequencies were calculated across all of the RecA sequences used in Eisen (1995).

Amino Acids in All RecAs

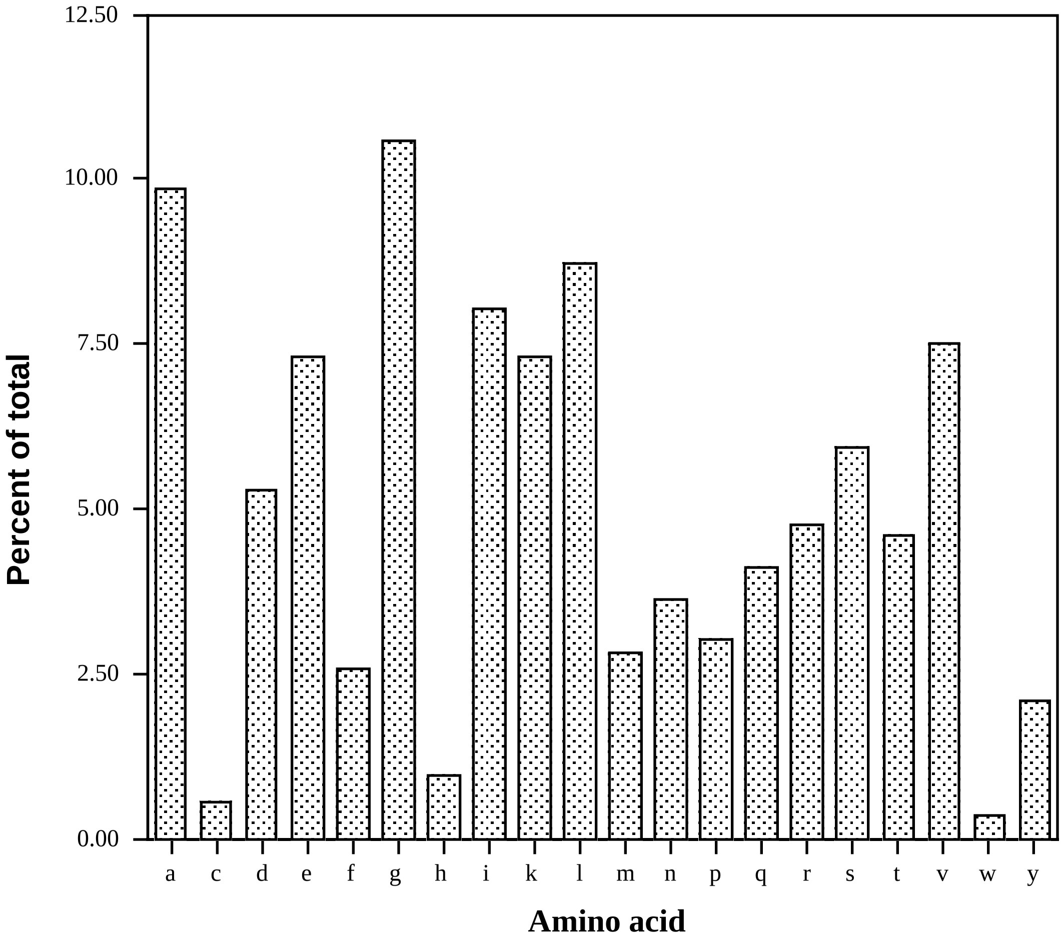


Figure 3. Frequencies of different classes of amino-acids in all RecA sequences, compared to genome GC content.

Note that, despite significant variation in frequencies of specific amino-acids, the frequency of amino acid classes is relatively constant.

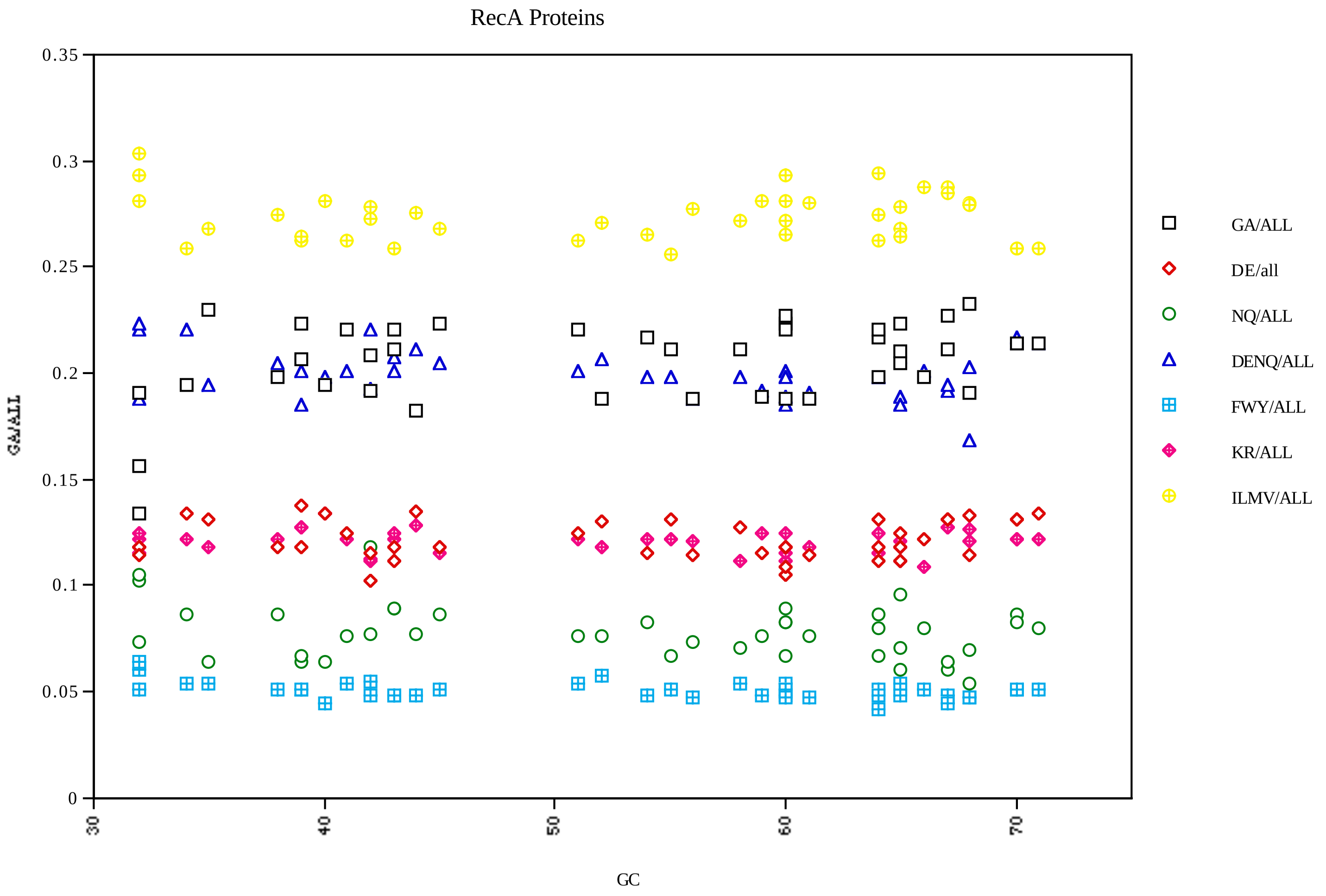

RecA Proteins
GA/ALL
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
30
40
50
60
70
GC
□ GA/ALL
◇ DE/all
○ NQ/ALL
△ DENQ/ALL
⊞ FWY/ALL
◆ KR/ALL
⊕ ILMV/ALL

Figure 4. Proportion of different types of hydrophobic amino-acids compared to species GC content

Note positive correlation of valine and negative correlation of isoleucine. All valine codons start with a G while all isoleucine codons start with an A. Thus it appears that while overall hydrophobicity is conserved, the choice of which hydrophobic amino-acid to use depends on genome GC content.

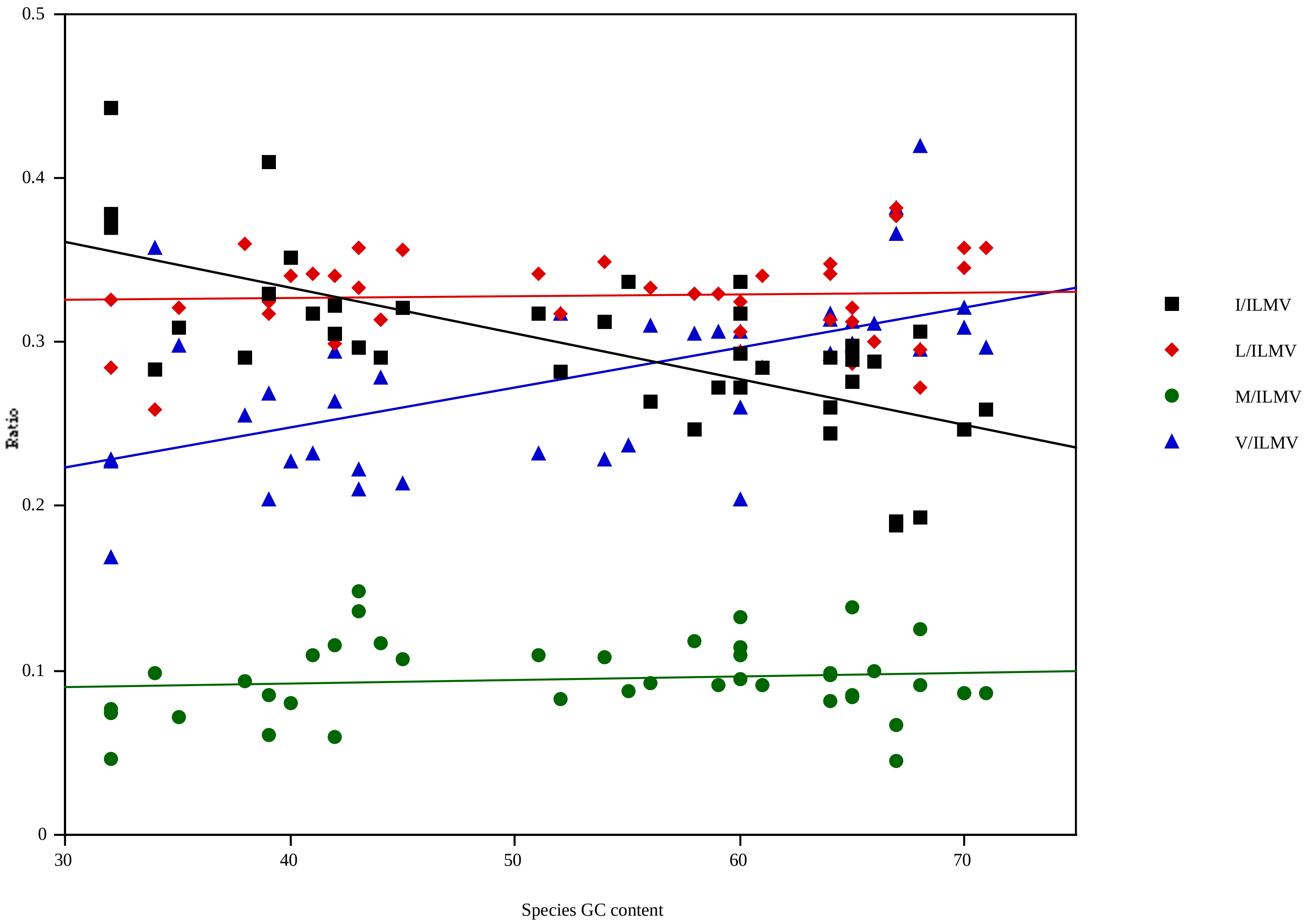

0.5
0.4
0.3
0.2
0.1
0
Ratio
30
40
50
60
70
Species GC content
I/ILMV
L/ILMV
M/ILMV
V/ILMV

Figure 5. Proportion of arginine out of total basic amino acids.

Note strong positive correlation. Arginine codons have more GC than Lysine codons. It is likely that selection for certain GC content has determined which basic amino acid is used when a basic amino-acid is needed. Selection for lowGC content leads to the use of lysine.

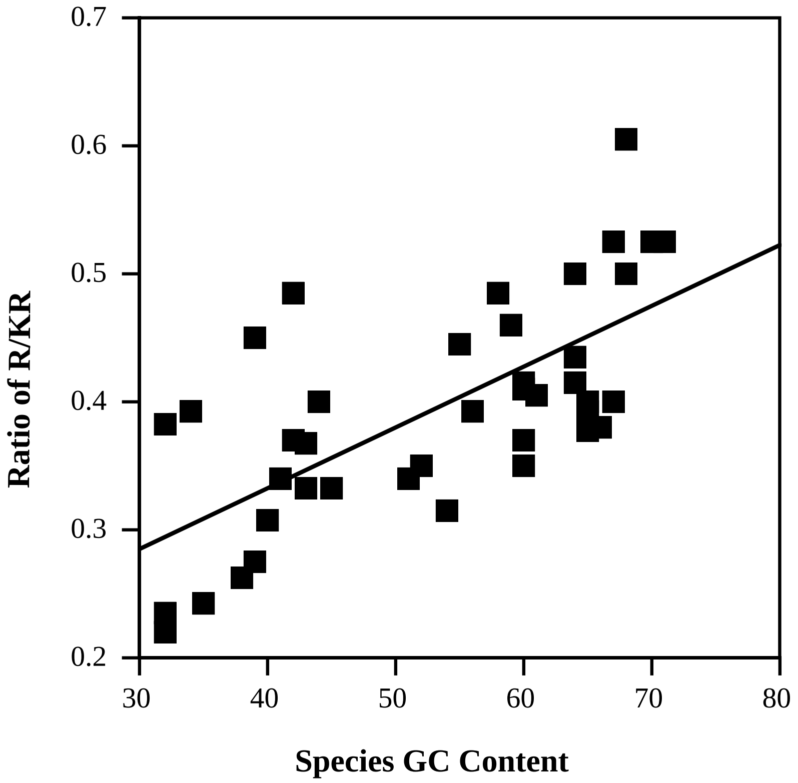
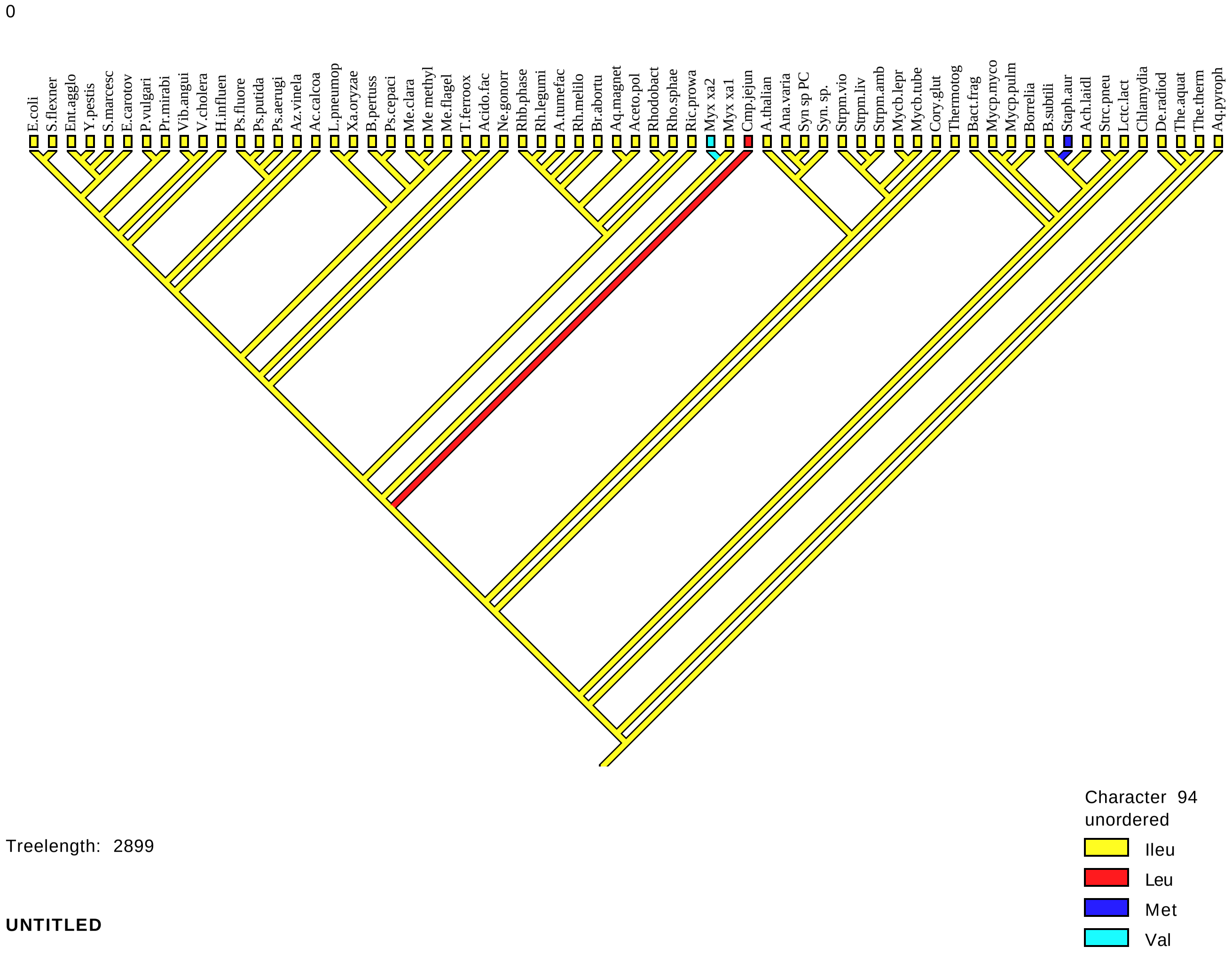
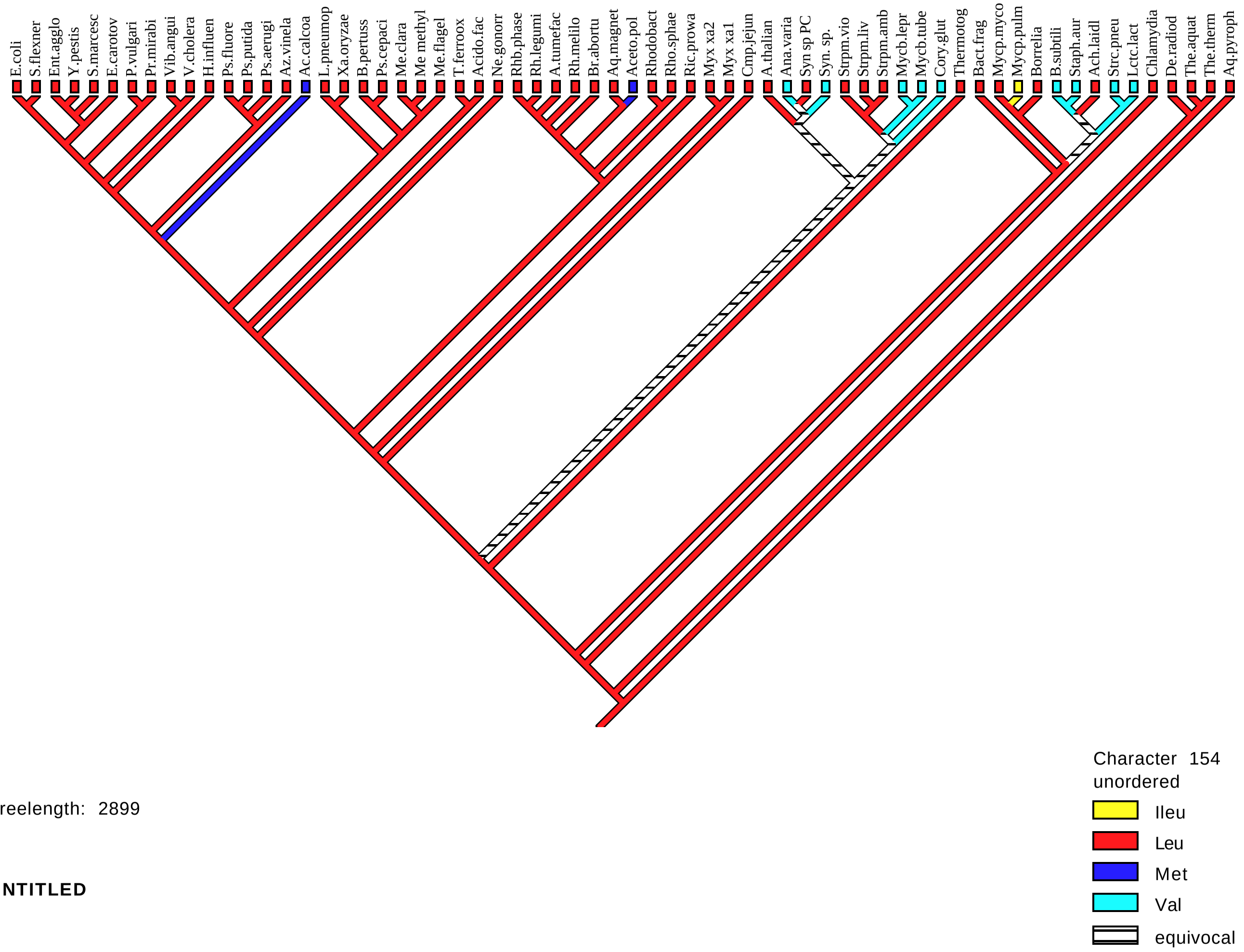

0.7
0.6
0.5
0.4
0.3
0.2
Ratio of R/KR
30
40
50
60
70
80
Species GC Content

Figure 6. Examples of tracing amino-acid substitution patterns at different positions in RecA.

A. Position 94. B. Position 154. C. Position 219. Amino-acid substitutions over evolutionary time were calculated at each alignment position using parsimony character state analysis (by MacClade 3.0). Substitutions were counted on the Fitch-Margoliash tree.


0
E.coli
S.flexner
Ent.agglo
Y.pestis
S.marcesc
E.carotov
P.vulgari
Pr.mirabi
Vib.angui
V.cholera
H.influen
Ps.fluore
Ps.putida
Ps.aerugi
Az.vinela
Ac.calcoa
L.pneumop
Xa.oryzae
B.pertuss
Ps.cepaci
Me.clara
Me methyl
Me.flagel
T.ferroox
Acido.fac
Ne.gonorr
Rhb.phase
Rh.legumi
A.tumefac
Rh.melilo
Br.abortu
Aq.magnet
Aceto.pol
Rhodobact
Rho.sphae
Ric.prowa
Myx xa2
Myx xa1
Cmp.jejun
A.thalian
Ana.varia
Syn sp PC
Syn. sp.
Strpm.vio
Strpm.liv
Strpm.amb
Mycb.lepr
Mycb.tube
Cory.glut
Thermotog
Bact.frag
Mycp.myco
Mycp.pulm
Borrelia
B.subtili
Staph.aur
Ach.laidl
Strc.pneu
Lctc.lact
Chlamydia
De.radiod
The.aquat
The.therm
Aq.pyroph
Treelength: 2899
UNTITLED
Character 94
unordered
Ileu
Leu
Met
Val

0



Treelength: 2899

UNTITLED

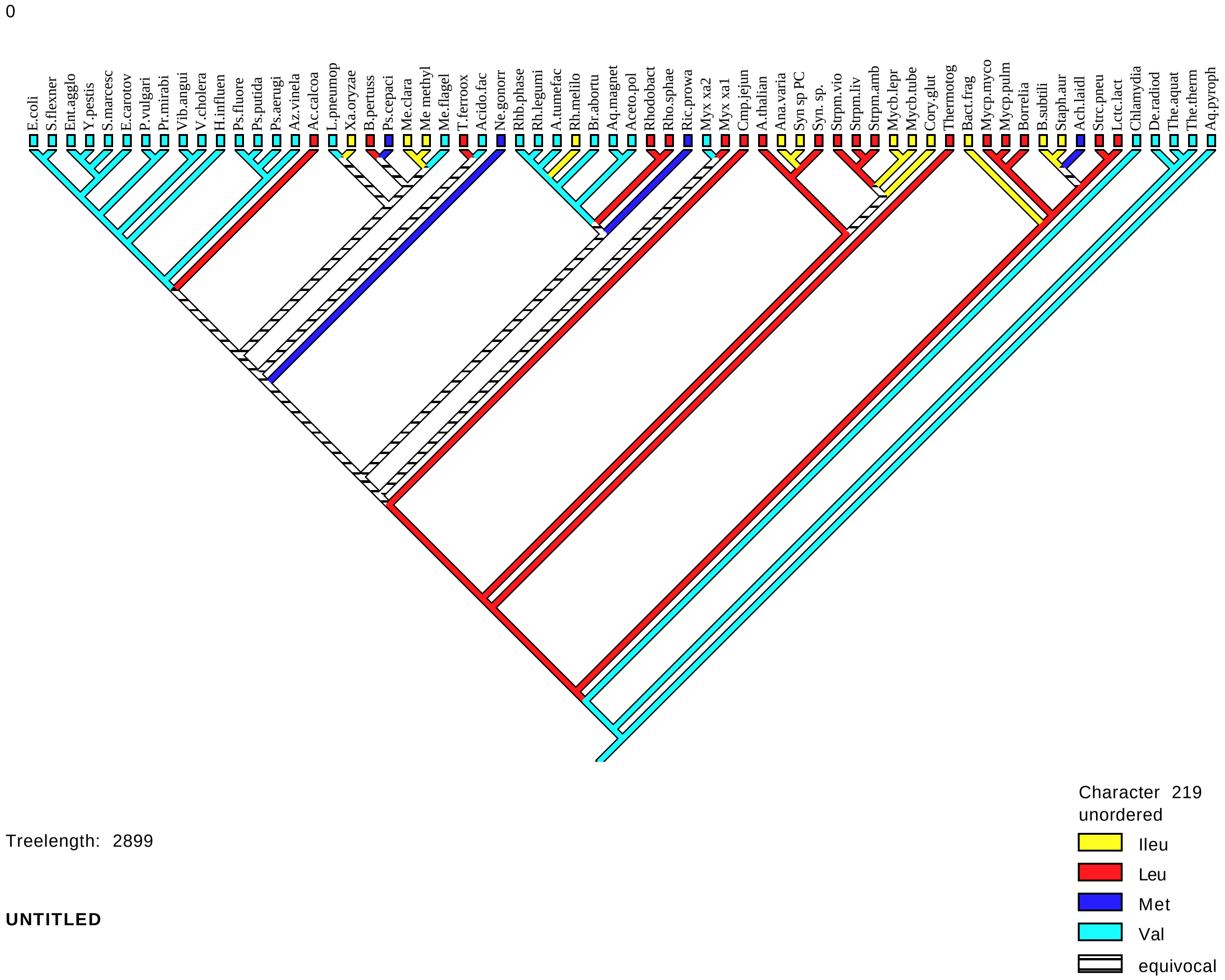


Figure 7. Correlation of amino-acid states and number of evolutionary substitutions.

Amino-acid substitutions over evolutionary time were calculated at each alignment position using parsimony character state analysis (by MacClade 3.0). Substitutions were counted on the Fitch-Margoliash tree.

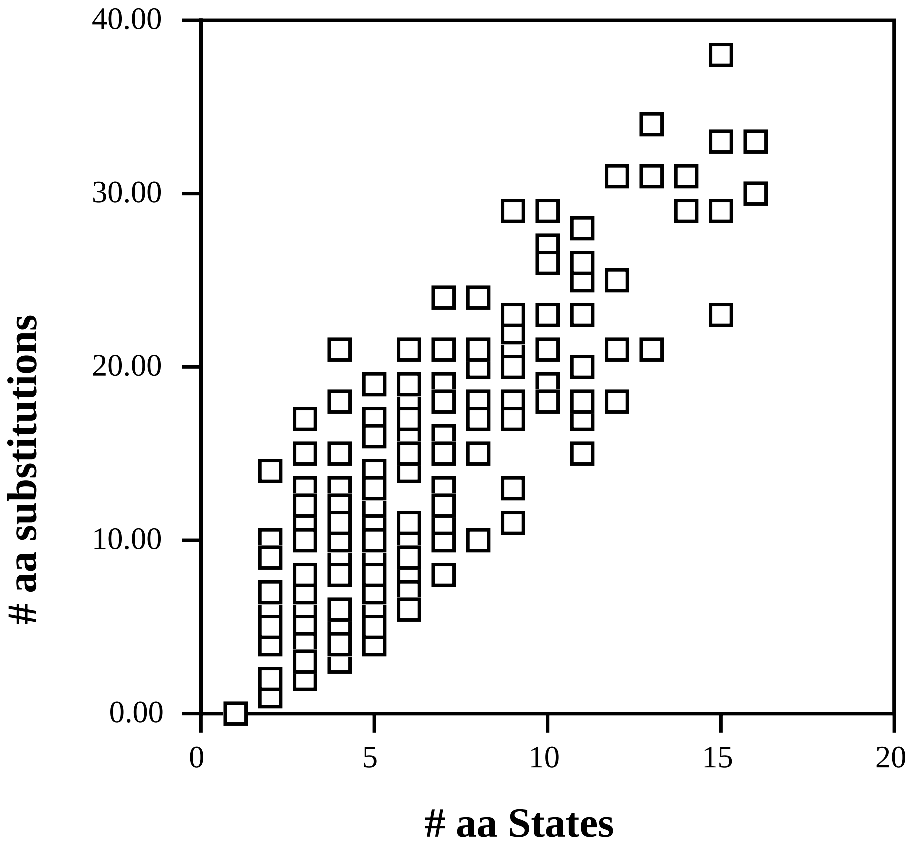
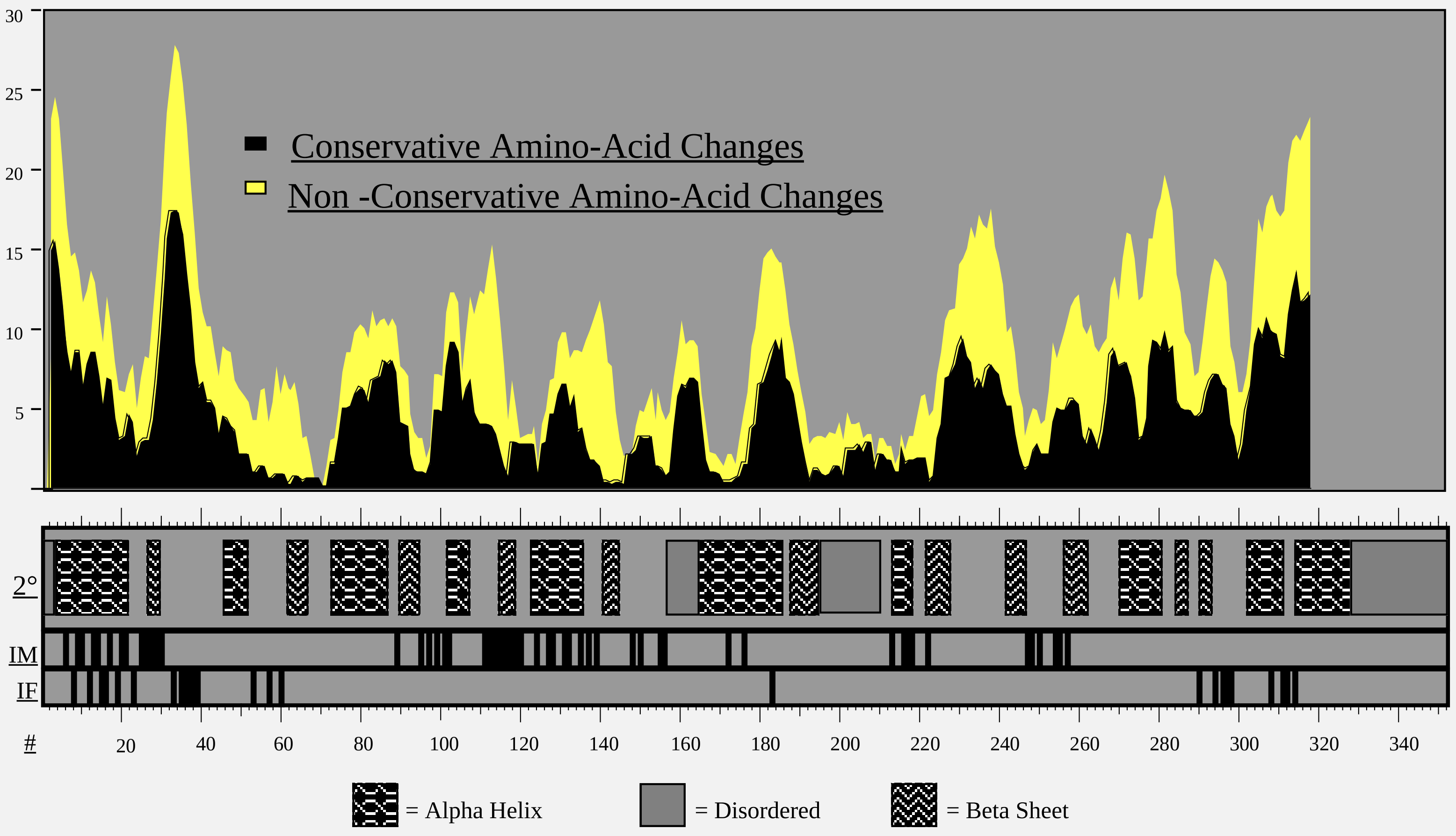

40.00
30.00
20.00
10.00
0.00
0
5
10
15
20
# aa substitutions
# aa States

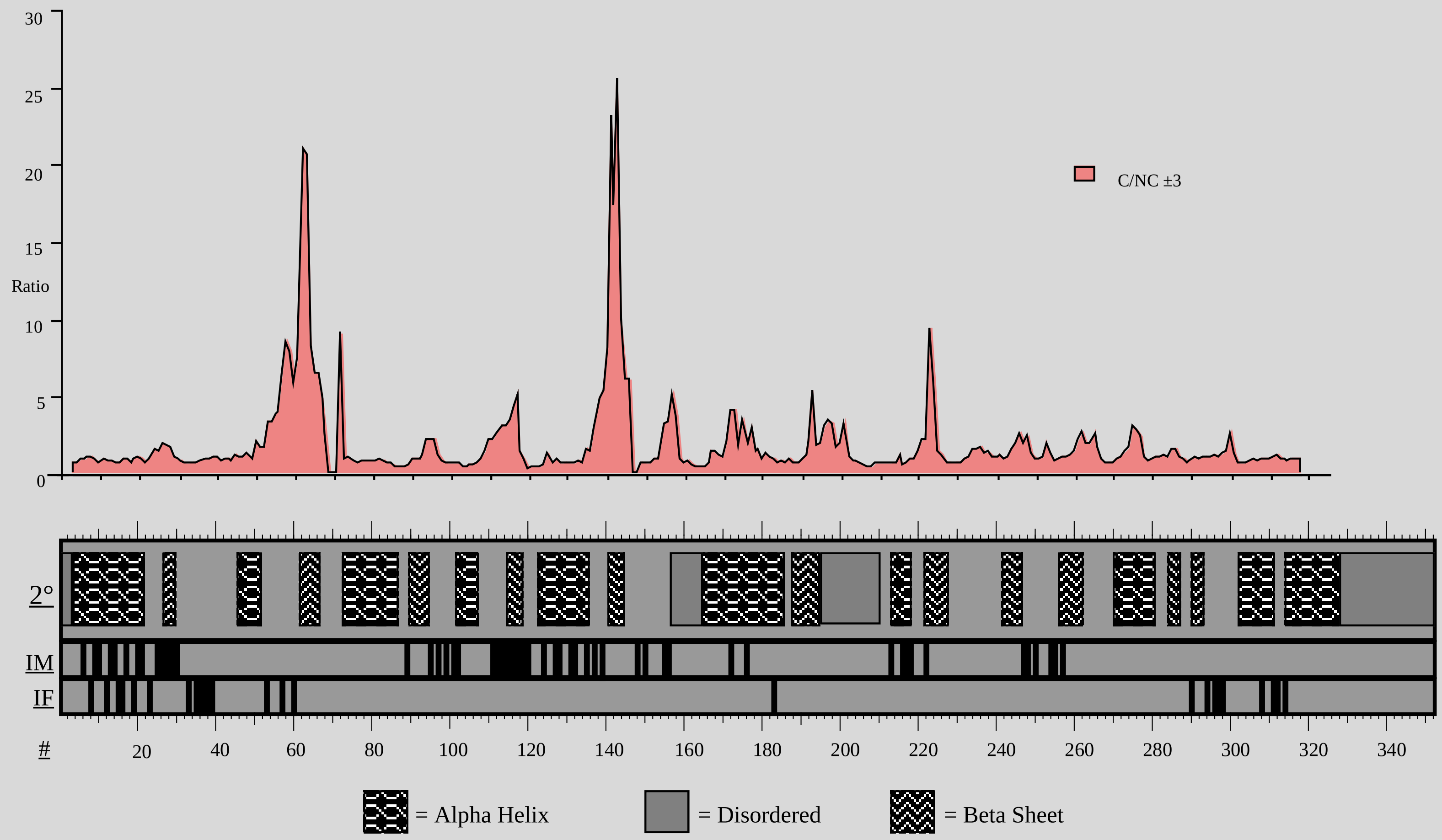
Figure 8. Conservative and non-conservative substitutions over evolutionary time in RecA sequences.

Amino-acid substitutions over evolutionary time were calculated at each alignment position using parsimony character state analysis (by MacClade 3.0). Substitutions were counted on the Fitch-Margoliash tree. Substitutions were considered conservative if within the following amino acid groups: (F, W, Y), (D, E, N, Q), (K, R), (S, T), (G, A), (M, I, L, V). All other substitutions were considered non-conservative. A. Conservative and non-conservative substitutions vs. *E. coli* primary structure. B. Ratio of conservative to non conservative substitutions vs. *E. coli* primary structure. C. Conservative and non conservative substitutions vs. secondary structural element in *E. coli* RecA. The average number of each type of substitution was calculated for different secondary and tertiary structural elements based on the *E. coli* RecA crystal structure. D. Ratio of conservative to non conservative substitutions vs. secondary structural element in *E. coli* RecA.

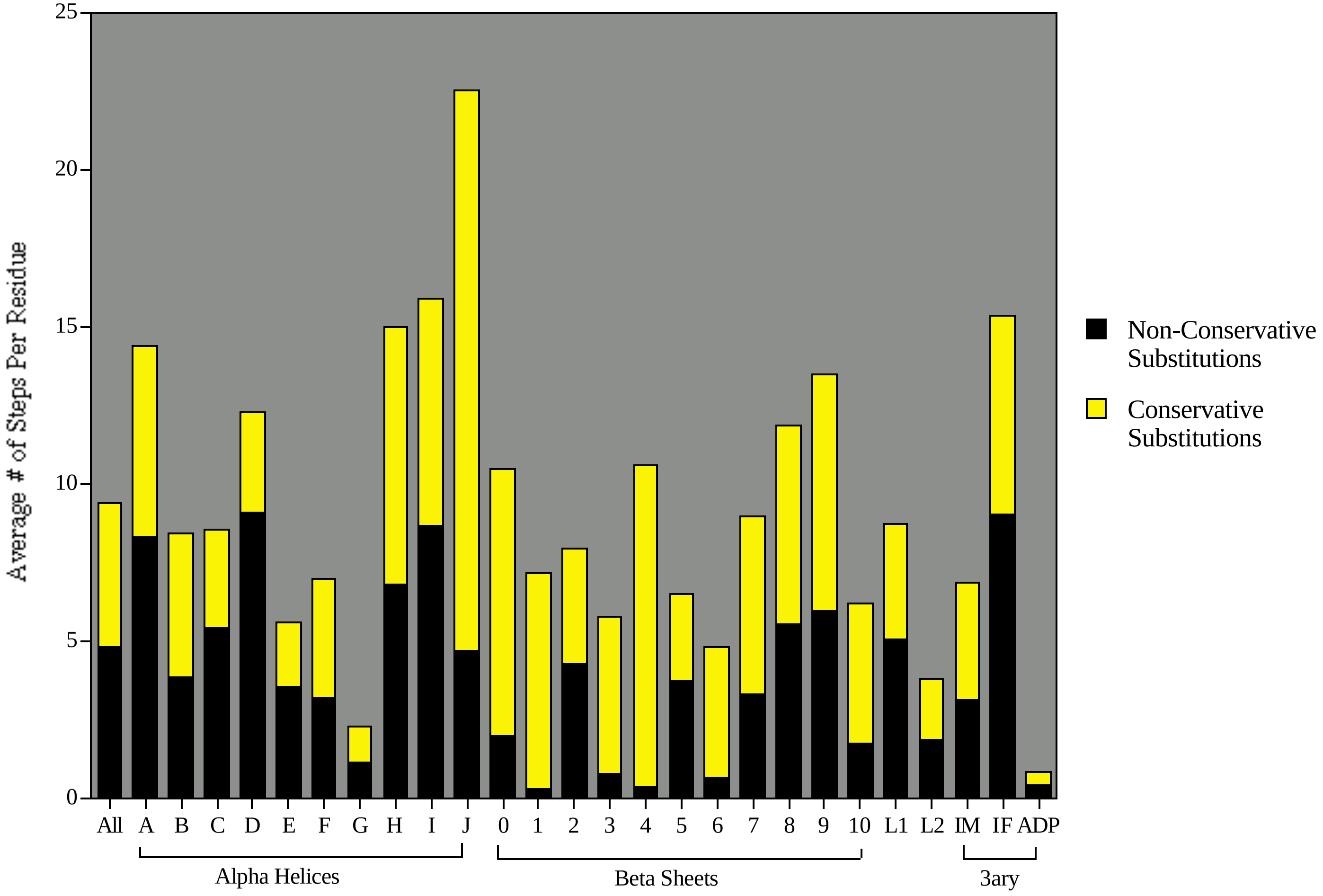
RecA Structural Evolution



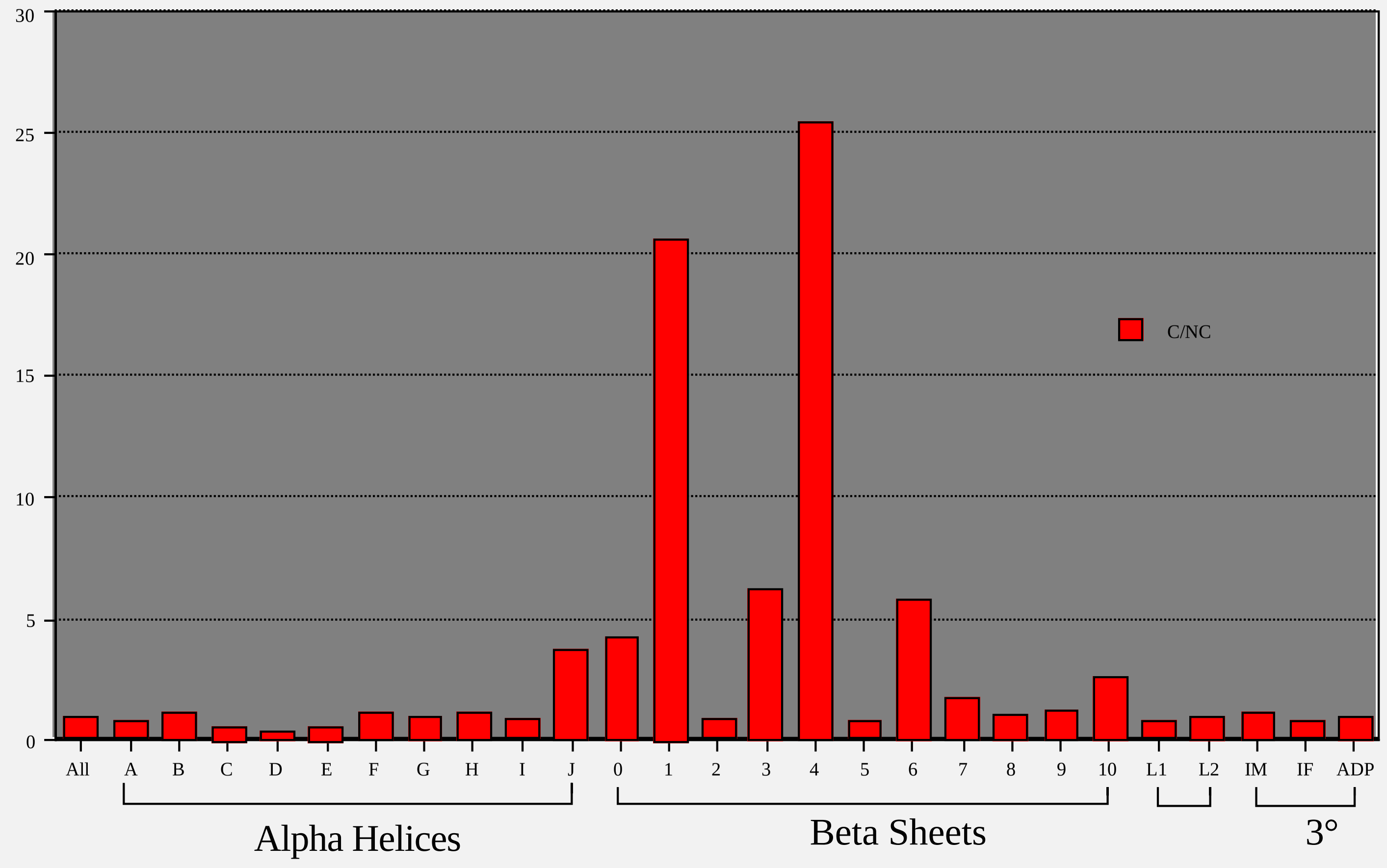
Ratio of Conservative to Non-Conservative Amino-Acid Substitutions



Amino Acid Substitutions in RecA by Secondary and Tertiary Structure Element



Ratio of Conservative to Non-Conservative Substitutions in Different Regions of RecA Protein



APPENDIX D

PCR Primers

Table 1. PCR primers.

Primer	AA Sequence	Primer Sequence
RecA1F	GPESSGKTT	5' GGNCCNGAYWSNWSNGGNAARACNACN
RecA2F	AF(I/V)DAEHALDP	5' GCITTYRTIGAYGCIGARCAYGCIYTIGAYCC 3'
RecA3F	GEQALEI	5' GGIGARCARGCIYTIGARAT 3'
RecA3R	GEQALEI	5' ATYYCIARIGCYTGYTCICC 3'
RecA4F	DSVAAL	5' GAYWSNGTNGCNGCNYT 3'
RecA-proteos-5R	FINQIRMKIGVM	5' CATNACNCCDATYTTCATNCKDATYTGRTTDATRAA 3'
RecA5R	IFINQ(I/V/L)R	
RecA6.1R	PETT(T/P)GG	5' CCICCIGKIGTIGTRTCIGG 3'
RecA6.2F	ALKFY	5' GCNYTNAARTTYTAY
RecA7R	KVVKNK	5' YTTRTTYTTIACIACYTT 3'
RECA-EUK1	I(V/I/T)E(L/M/I/V)(F/Y)G	gggagctcAAHRYIGARITITWYGG
RECA-EUK2	DS(V/A/C)(A/T)AL	ggctgcagIARIGCIGKIVMISWRCT
Ung1F	GQDPYH	5' gggagctcGCICARGAYCCITAYCA 5'
Ung1F#2	GQDPYH	5' GCICARGAYCCITAYCA 3'
Ung1F-Halo	V(K/R)VVI(V/I/L)GQDPYH	5' GTSMRNGTSGTSATYVTBGGNCARGACCCSTACCA
Ung1F+LF1	GQDPYH	5' atatggtaccgcgggggGCICARGAYCCITAYCA 3'
Ung1.5F	QA(H/Q)GL(C/A/S)FSV	
Ung3F+LF1	QGVLLLN	5' atatggtaccgcgggggCARGGIGTIYTIYTIYTIAA 3'
Ung3R	QGVLLLN	3' GTYCCICAIRAIRAI---TTRcgacgtcggg 5'
Ung3R#2	QGVLLLN	3' GTYCCICAIRAIRAIRAITTR 5'
Ung3R+LR1	QGVLLLN	3' GTYCCICAIRAIRAIRAITTRggggagctcttaagaaaa 5'
Ung3R-Halos	WA(K/S/R/E)QGVLLLN	5' GTTVAGVAGVAGVACNCCYTG
Ung3.5	(G/I)WE(Q/T/K/P)FT(D/K)	
Ung4R	(V/I/L)FMLWG	3' CAIAARIAIRAIACCCCgacgtcggg
Ung4R+LR1	(V/I/L)FMLWG	3' CAIAARIAIRAIACCCCggggagctcttaagaaaa
Ung4F-Halos	(L/V)VF(L/M/I)LWG	5' STBGTBTTCMTBCTBTGGGGG
Ung5R	HPSPL	3' GTRGGISWIGGIRAcgacgtcggg
Ung5R+LR1	HPSPL	3' GTRGGISWIGGIRAggggagctcttaagaaaa
Ung5F-halos	HPSPLS	5' CACCCSWSSCCSCTBWSS
MutL1F	N(Q/R/K)IAAGE	5' ggggagctcAAYMRIATHGCIGCIGGIGA
MutL3F	GFRGEA	5' ggggagctcGGITTYMGIGGIGARGC
MutL3R		5' gggctgcagcGCYTCICCICKTAAICC
MutL4R	VDVNVHP	5' gggctgcagcGGRTGIACRTTIACRTC
MutS1F	ITGPNMG	5' ggggagctcATHACNGGNCCNAAYATGGG
MutS2R	TFM(V,E)E	5' gggcygcagcTCNSCCATRAAIGT
MutS3R	DE(V,I,L)GRGT	5' gggctgcagcGTNCCNCKNCCNANYTCRTC
SNF2-Micro-1F	LAD(D,E)(V,M)GLGKT	5' CTBGCNGACGAVRTBGGNCTBGGNAARAC
SNF2-micro-4R	(K,E)AGG(F,V,E,T)G(I,L)NL	5' AGGTT(AGC)AKNCCNRBNCCNCCNGCYTY
SNF2-prok2	(L/V/I)(V/I/L/F)(V/I/L)DEA(H/Q)	
SNF2-prok3	LT(G/A)TP(I/V/E)(E/Q)(N)	
SNF2-prok5	(V/M)I(H/N/L)(F/Y)D(L/R/V)(W/P)WNP	
MFD-humhomo1		GACCATGACGGTTGATGGTGGC
MFD-humhomo2		CCCAGCTTCCGCTTCCCGTTGGG
MFD-R		
MFD-F		

APPENDIX E

Cloning the *recA* Gene of *Caulobacter crescentus*

SUMMARY

This appendix shows some of the results of experiments concerning cloning the *recA* gene of *Caulobacter crescentus* using degenerate PCR. These were done in collaboration with Rob Wheeler in Lucy Shapiro's laboratory at Stanford. Analysis of the sequence is reported in Chapter 2b.

Figure 1. Degenerate PCR amplification of fragments of the *Caulobacter crescentus recA.* gene.

PCR done using reaction conditions as described in Gruber et al except 100 pmoles of each primer were used. Thermal cycling was done using a Perkin Elmer 2400 with the following temperature parameters: 97°C x 2 minutes; 30 cycles of (94°C x 0.5 minutes, 53°C x 1 minutes, 72°C x 1 minute); and 72°C x 10 minutes. See Appendix D Table 1 for a list of PCR primer sequences.

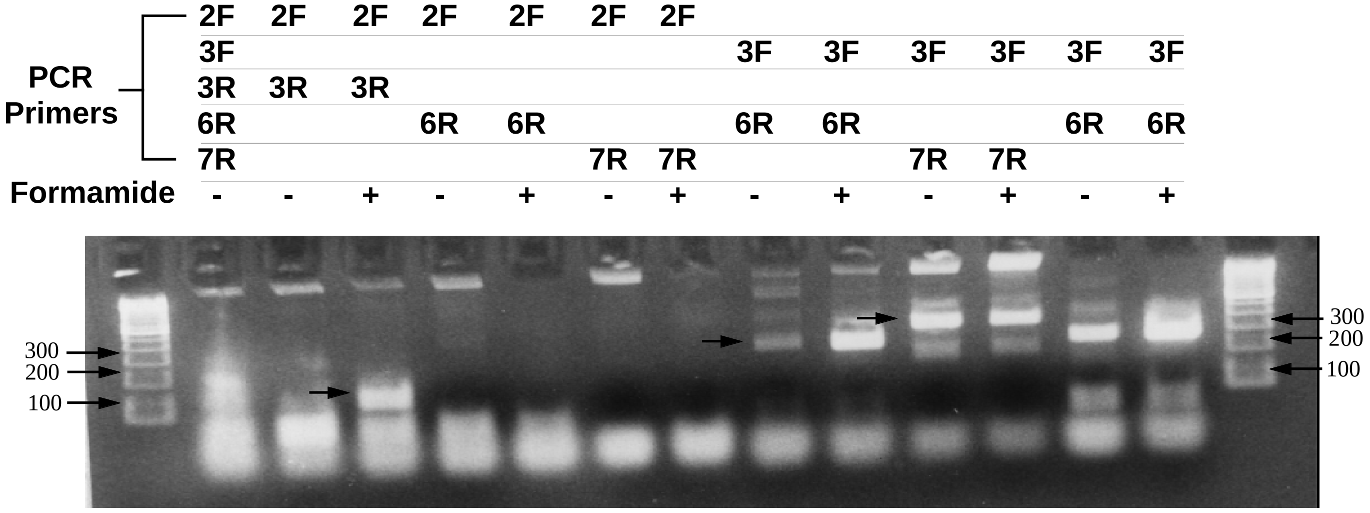
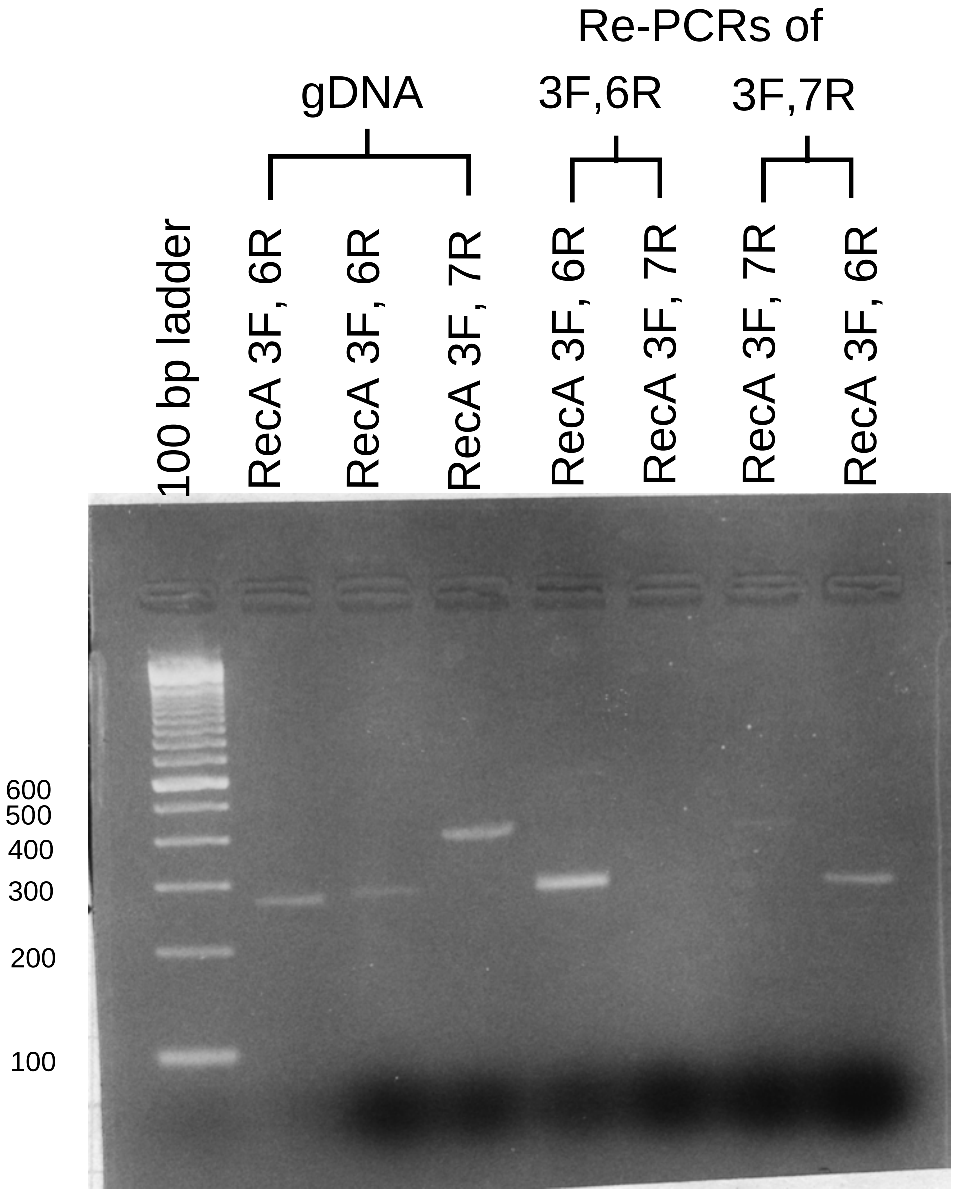

PCR
Primers
2F 2F 2F 2F 2F 2F 2F
3F 3F 3F 3F 3F 3F 3F
3R 3R 3R
6R 6R 6R 6R 6R 6R 6R
7R 7R 7R 7R 7R
Formamide - - + - + - + - + - + - +
300
200
100
300
200
100

Figure 2. Purification of *Caulobacter crescentus recA* PCR products.

Appropriate bands were cut out a purified using the Wizard kit and used for cycle sequencing and as probes to pull out full-length clones.


Re-PCRs of
gDNA
3F,6R
3F,7R
100 bp ladder
RecA 3F, 6R
RecA 3F, 6R
RecA 3F, 7R
RecA 3F, 6R
RecA 3F, 7R
RecA 3F, 7R
RecA 3F, 6R
600
500
400
300
200
100

APPENDIX F

Repair of UV Induced Cyclobutane Pyrimidine Dimers in the Extremely Halophilic Archaea *Haloferax volcanii*

SUMMARY

There have been very few studies of DNA repair in Archaea. Repair studies in Archaea would be particularly interesting since they are ecologically and evolutionarily distinct from other species in which repair processes have been well characterized. Such information will be useful to better understand the evolution of DNA repair processes, as well as to understand the mechanisms by which Archaea grow and thrive in the extreme environments in which they live. In this appendix I present the results of experiments on the repair of UV irradiation induced cyclobutane dimers in the extremely halophilic Archaea *Haloferax volcanii*. In addition, I give a brief introduction to some of the features of Archaea and provide arguments as to why repair studies in Archaea in general, and *H. volcanii* in particular would be of interest. It is hoped that these studies will help lay the foundation for developing *H. volcanii* into a new model species for the studies of DNA repair.

INTRODUCTION AND BACKGROUND

DNA repair processes have been documented in a wide variety of species. However, the ecological and evolutionary diversity of DNA repair studies is somewhat limited (see Table 1) and the majority of studies have been done in a few model organisms like *Escherichia coli*, yeast, and mammals. While a great deal of mechanistic information has been gained by focusing on such model species, there are many reasons to expand this listing to include other species. Although there are many organisms and even groups of organisms that are poorly represented in studies of repair, I believe it is particularly important to expand the studies of DNA repair in the "third" domain of life, the Archaea.

DNA repair studies in Archaeal species

Before discussing why I believe studies of repair in Archaea should be of interest it is helpful to review what is known about repair in Archaea. As suggested above only limited studies of DNA repair have been conducted in Archaea. Photoreactivation has been documented in a few different species of Archaea including the methanotrophic thermophile *Methanobacterium thermoautotrophicum* (1) and two species of extreme halophiles *Halobacterium halobium* and *H. cutirubrum* (2). Photolyase enzymes have been cloned and characterized from *M. thermoautotrophicum* ((1,3) and *H. halobium* (4) and both are homologous to previously characterized photolyases, although one is a classI photolyase (*H. halobium*'s) and the other is a classII photolyase. The ability to repair ionizing radiation damage to DNA has been shown in some thermophiles (5) and in the halophile *H. mediterranei* (6). Uracil glycosylase activity is found in some thermophilic Archaea (7). Suggestions of recombinational repair in Archaea come from studies of a recombination defective strain of *H. volcanii* (8) which is UV sensitive, and from the ability of some *Pyrococcus* species to repair double strand breaks (5). Other information about repair genes that have been found in Archaea, especially in complete genome sequences, can be found in Chapter 7.

With one exception (cited below), the only attempts to study nucleotide excision repair in Archaea have been made in halophiles. Initially, *H. cutirubrum* was shown to be extremely resistant to UV irradiation (9). Hescox and Carlsberg showed that survival increases markedly after exposure to photoreactivating light but that survival does not increase over time when cells are left in the dark. In *E. coli* and other species, survival increases in certain non-growing conditions in the dark due to light-independent DNA repair processes, especially excision repair. This recovery in the dark is also known as liquid-holding recovery. Hescox and Carlsberg suggested that the lack of liquid holding recovery in *H. cutirubrum* was due to a lack of excision repair. Grey and Fitt also noted the lack of liquid holding recovery and suggested the lack of excision repair (10). Subsequent attempts to detect liquid holding recovery in halophiles also failed (11,12). Fitt and Sharma pursued this suggestion by conducting an experiment to study the removal of cyclobutane pyrimidine dimers (CPDs) in the dark (13). No removal was detected and thus they have concluded that excision repair, at least of CPDs does not exist in halophiles. After I began my work on repair in *Haloferax volcanii*, two studies relating to nucleotide excision repair in Archaea have been published. These are discussed below.

Why study DNA repair in Archaea?

One reason to study repair in Archaea is that they are evolutionary unique. Archaea are single celled anuclear organisms that were originally grouped into the kingdom Monera with other anuclear organisms (the so-called prokaryotes). However, on the basis of molecular phylogenetic studies, the Monera have been recognized as an assemblage of distinct evolutionary groups. Initial studies, based on phylogenetic analysis of rRNA sequences, identified three major domains of life - the eukaryotes, the "true" bacteria, and a third group, which is now referred to as the Archaea (14). Recent studies have cast some doubt on whether the Archaea, as defined by Woese and others, represent a single monophyletic group (e.g., (15,16)). However, whichever point of view one takes on Archaeal evolution, all studies have confirmed that each of the major Archaeal groups is evolutionary distant from other groups of organisms. Thus, whether the Archaea can be considered a single group is not of particular importance here. What is important is that any species of Archaea one chooses will be evolutionary distant from any of the species in which repair has been well characterized.

Studies of repair in Archaea would also be of interest because of the ecological novelty of these species. The Archaea tend to grow in extreme environments such as high salt (5M KCl), high temperature (110°C), or high pressure (~300 ATM at 13000 feet below sea level). If ecology influences DNA repair processes and their evolution then the Archaea are likely to have significant differences from *E. coli*, yeast, and mammals.

What is known about other molecular processes in Archaea suggests that studies of repair will be useful and interesting. For example, it has been shown that excision repair is coupled to transcription in *E. coli,* yeast and mammals. Archaeal transcription is similar to eukaryotes in some ways (e.g., RNA polymerase sequence and structure (17) and promoter sequence (18)) but similar to bacteria in other ways (e.g., the use of a single RNA polymerase for all transcription and the use of operons (19)). Thus it would be of interest to determine if they have transcription-coupled DNA repair and if so, whether it is like the Mfd based system of bacteria or the CSB-CSA based system of eukaryotes (see Chapter 7). In addition, the DNA of Archaea appears to be packaged into a type of chromatin/nucleosomal structure with histone-like proteins as seen in eukaryotes (20-23). In eukaryotes it is thought that this structure plays a role in regulating repair processes

(24). Since the packaging in Archaea is apparently less complex than that in eukaryotes, it may be useful to study the effects of packaging on repair in a simpler Archaeal system. Another cellular factor that is thought to affect repair is attachment to the nuclear matrix (25). While Archaea do not have a nucleus there have been suggestions that they may have cytoskeletal-like features (26,27). This and other features of Archaea have led some researchers to propose that the eukaryotic nucleus is a remnant of an endosymbiotic Archaea (28). Overall there are many features in which the Archaea are similar to eukaryotes and many others in which they are similar to bacteria (28,29). In general the similarities to eukaryotes tend to be in things that are general molecular processes while the similarities to bacteria are in aspects of life that are thought to be adaptations to "streamlining" (like operons), so Archaea are thought to be a sister group of eukaryotes. Other evidence for an evolutionary relationship between eukaryotes and Archaea comes from phylogenetic trees of duplicated genes (30,31). Thus Archaea may be a more relevant model for eukaryotes than *E. coli* and other bacteria, yet they have much of the simplicity that makes *E. coli* preferable to yeast for many basic studies.

Finally, another reason to study repair in the Archaea is that comparative genomics reveals that Archaea only encode homologs of a limited number of repair genes that are found in bacteria and eukaryotes (Chapter 7). In addition, analysis of the evolutionary history of repair genes suggest that there have been multiple origins for many types of repair processes (e.g., the nucleotide excision repair processes of bacteria and eukaryotes appear to be of separate origin). Thus it is likely that novel repair processes will exist in Archaea and these can only be discovered by experimental studies.

Haloferax volcanii as a model for studies of DNA repair

How does one go about choosing the Archaeal species in which to study repair? The Archaea are divided into three main evolutionary and ecological groups: the extreme halophiles, the extreme thermophiles, and the methanogens (32). I chose to work on repair in the extreme halophiles (see Table 2 for a listing of some features of halophilic Archaea) for a few reasons including that (a) halophiles are the easiest to grow and manipulate of the Archaea - they can be grown at 37°C on minimal medium plates (with high salt) unlike other Archaeons which tend to need anaerobic, high temperature (33) (b) the previous results (see above) suggested that halophiles lack nucleotide excision repair despite being extremely radiation resistant (c) studies in halophiles would give ecological breadth to studies of repair because there are few studies of repair in photosynthetic species which are exposed to very high levels of DNA damaging light; (d) of the Archaea, halophiles are the group that has been best characterized at the molecular level (33); and (e) there may be some interesting effects on repair imposed by the high internal salt concentrations (~4M) found in halophiles.

Of the Halobacteria, I chose to work on *Haloferax volcanii*. *H. volcanii* is preferable because it has many properties (such as a low level of transposable elements and natural transformation techniques) that make it amenable to use in molecular studies (33). In particular, it is being developed into a model Archaeal species for general molecular studies and as a result of this there are many molecular tools available for experimental studies in this species including a large number of mutants (33); cosmid libraries (34), transformation (35-37), shuttle vectors (38,39), a physical map (34,40). All of these tools will facilitate future repair studies in this species (see Table 3 for a listing of some characteristics of *H. volcanii*). Finally, I believe the development of this species as a model for molecular studies will benefit from characterizing its repair processes.

Excision repair in Haloferax volcanii

I was particularly interested in studying nucleotide excision repair in *H. volcanii* because of the previous reports of an evident lack of NER in halophilic Archaea. It seemed unlikely to me that these previous reports were correct in concluding that halophiles lacked NER. First, the TLC method used in the earlier studies is not very sensitive and would have been unable to detect low levels of ER. More importantly, to use the TLC method, the researchers had to expose cells to incredibly high doses of UV irradiation in order to detect enough CPDs to study repair. Repair processes may have been inactivated at such high doses. It seemed possible that repair might occur after lower doses. In addition, the previous studies only looked for repair at time points similar to those at which repair was studied in *E. coli* despite the cell doubling time of over 10 hours in most halophiles. It seemed like it would be better to study repair at time points that coincided with the slow growth of these species. Finally, it did not make much intuitive sense that halophiles would lack excision repair. On the contrary there was suggestive evidence that halophiles have some form of repair, including: (a) halophiles, like most Archaea, have relatively slow rates of molecular evolution (b) halophiles are relatively resistant to mutagens such as MNNG (41) and EMS (42) and (c) most halophiles are aerobic making them highly prone to oxidative damage in DNA and thus in need of some form of repair (43,44). Thus I set out to re-examine excision repair in Halophiles. My suspicions of the presence of repair in halophiles were confirmed by my own work and by a study that was published after I began my research (45). This study represents the first evidence for excision repair in an Archaea. In addition, a very recent study has reported that extracts of *M.thermoautotrophicum* contain activities that incise DNA containing a site-specific 6-4 photoproduct (46). The excised segment was similar to that in *E. coli*. Although this study did not identify any of the genes involved or whether this activity occurs in vivo, it does suggest the presence of some for of nucleotide excision repair in this species. This is not entirely surprising since *M .thermoautotrophicum* encodes homologs of the bacterial nucleotide excision repair genes (*uvrABCD*).

SUMMARY OF EXPERIMENTS AND METHODS

Growth, strains, and DNA extraction

All experiments were done on *Haloferax volcanii* strain WFD11 unless otherwise noted. Growth media was made as described in X and Y. Cells were grown aerobically at 37°C. DNA was isolated using the *E. coli* miniprep method of (47) except without the addition of lysozyme.

UV irradiation, survival curves, and repair conditions

UV irradiation was done using essentially the same strategy as in (48). Cells were spun down, resuspended in minimal media, and irradiated in glass dishes. UV survival curves were determined by plating serial dilutions of each time point (in 10 ul drops) and counting the resulting colonies. Unless otherwise noted, after UV irradiation, cells were exposed only to yellow non-photoreactivating light (include during DNA extractions and during growth of cells on plates). Conditions for repair are described in the figure legends.

T4EV assay

The amount of CPDs in DNA after UV irradiation was measured by an assay described by Spivak and Hanawalt (49). In this assay, the DNA is either treated or mock treated with T4 endonuclease V which cuts the DNA at sites of pyrimidine dimers, and then the DNA is electrophoresed on a denaturing alkali agarose gel. The average size of the DNA decreases with increasing numbers of CPDs.

Whole genome DNA repair assay

Repair in unreplicated DNA was measured by a modification of (50). First, to pre-label *Haloferax volcanii* DNA, cells were grown in 25-30 ml minimal media from a single colony (all growth was at 37°C in a shaking water bath). After a few hours, ^{3}H-thymine (>3 uCi/ml final) was added. Cells were then grown at least three generations (>15 hours). Cells were spun down and resuspended in an equal volume minimal media w/o label and grown for 0.5-1 hour. At the start of the repair experiment 5 mls of cells were removed for a zero time point and placed on ice. The remaining cells were placed in a glass dish and exposed to UV irradiation. Immediately after UV irradiation, a sample was removed for a "no repair" time point and placed on ice. The remainder of the cells were placed in a flask and incubated (37°C x shaking). At various time points, cells were removed and placed on ice and DNA was isolated. DNA was quantified using a spectrophotometer and ^{3}H counts in 5-10 ul DNA by dropping onto filter paper, TCA washing, and counting. This DNA was then used in the T4 assay described above. The gels were stained with EtBR and photographed with a ruler (for calculating the average size of the DNA). The gel was partially dried (with no heat) using a vacuum blotter. A grid was then drawn on the dried gel and each "fraction" was excised and placed in a scintillation vial. 250-500 ul 0.2N HCl was added to each vial and then the tubes were autoclaved for 1 minute to melt the gel. ^{3}H in each sample was counted using aqueous counting solution.

The average molecular weight for each fraction was calculated using molecular weight markers based on the average migration distance of each fraction. Repair was visualized by comparing the average molecular weight of each fraction versus the percent of total CPM for that fraction. Total percent repair was calculated by the following formula based on (50). First, the average molecular weight of each sample was calculated by the following ratio

$$\frac{\text{sum CPM}}{\text{sum (CPM/mol. wt.)}}$$

where each parameter is summed for all fractions. The number of enzyme sensitive sites (ESS) was calculated by

$$\frac{\text{A(w/o T4)}}{\text{A (w/ T4)}} - 1$$

The number of ESS per base pair (E) was calculated by ESS / A (w/o T4). The inverse of this gives you X where X is the distance between cuts. Percent repair at time X calculated by dividing $(E_0\text{-}E_X)/E_0$.

RESULTS AND DISCUSSION

Growth characteristics

An important component of DNA repair studies is information concerning the growth parameters of the species of interest. In general, it is important to conduct experiments in time frames relevant to the cell cycle of the organism of interest. As a first part of characterizing the growth of *H. volcanii*, I wanted to be able to estimate number of cells from OD measurements. First, I wanted to determine which wavelength to use and whether the OD measurement was affected by factors other than the number of cells. Therefore I measured absorption of a dense culture as well as a 1:10 dilution of this dense culture (at many wavelengths) and calculated the ratio of the ODs at the two densities of cells (Figure 1). If the OD was influenced only by number of cells then the ratio of the ODs at the two different densities should equal the dilution ratio. As can be seen, the ratio is only around 1:10 at wavelengths above 500 nm and thus wavelengths less that 500 nm should not be used to estimate number of cells in a *H. volcanii* culture. I then compared OD measurements to number of colony forming units to get an estimate of the number of cells per OD (Figure 2). At 500 nm, 1 OD corresponds to about 2 x 10^6 cells per ul or 2 x 10^9 per ml. In addition, I generated growth curves using a few different ODs and growth conditions (Figure 3). These curves were used to calculate doubling times (~ 7.5 hours in log phase) and to identify OD levels of different phases of the growth cycle for later experiments on UV resistance.

UV survival

The first step in characterizing the repair processes of *H. volcanii* was to study the lethality of different doses of UV irradiation. A few different survival curves are shown in Figure 4. As can be seen from Figure 4a, *H. volcanii* is much more resistant to UV irradiation than *E. coli*. Such a high level of resistance has also been found in other halophiles (e.g., (8,11)). In theory, such extreme UV resistance could be due to a few mechanisms including protection from damage, tolerance of damage, or DNA repair. It is unlikely that the extreme UV resistance is due to protection from damage because, surprisingly pigmentless strains are no more sensitive to UV (51); cell density has little effect on UV survival curves (not shown) and the amount of damage per dose of UV is comparable to that for *E. coli*. Therefore I believed it was likely that part of the explanation was efficient DNA repair. The first step in examining the potential for DNA repair was to characterize UV survival curves in more detail. In other species, the ability to recover viability after UV is a good indication of repair under non-growing conditions. In particular, the ability to recover viability after UV irradiation with incubation in the absence of photoreactivating light is a good indication of excision repair. I found that *H. volcanii* is able to recover viability with incubation in the dark as well as with incubation in photoreactivating light (Fig. 4b,c). This ability to recover after incubation in the dark suggests some form of dark repair. Another indication of likely DNA repair activities in this species is the difference in UV sensitivity of log and stationary-phase cells (Fig. 4d). Such a difference generally represents the same phenomenon as liquid-holding recovery to permit repair to operate before DNA replication is attempted.

Characterizing DNA repair

Since the results of the survival experiments suggested that some form of dark repair was occurring, I set out to determine if I could detect DNA repair. First, I examined the loss in the DNA over time of sites sensitive to cutting by the T4EV enzyme (Fig. 5). T4EV cuts the DNA backbone at sites of CPDs, and thus the loss of enzyme sensitive sites (ESS) is considered to be equivalent to the repair of CPDs. This analysis showed that ESS disappeared over time after UV irradiation suggesting that the CPDs were removed from the DNA. However, this reduction in ESS could be due to DNA replication since the assay that was used examines all the DNA in the cells at the time points of interest and not just DNA that had been irradiated. I used radiolabelling of the DNA to estimate the amount of replication after UV. These experiments suggested that little replication was occurring at these doses (Fig. 6). However, it was still important to determine how much ESS removal was occurring in the unreplicated DNA. Therefore I used a method in which ESS were measured only in the irradiated DNA. An outline of the method is given in Fig. 7. First, I used this method to measure repair at 180 J/m^2 (Fig. 8). After 24 hours, there was only a little repair at this dose, as seen with earlier studies in other halophiles. However, when the dose was lowered to 45 and 90 J/m^2, extensive repair was detected (Fig. 9-11). These results show clearly that repair of CPDs does occur in halophiles. However, the methods used do not reveal whether this repair is a form of NER or some other type of DNA repair. I attempted to determine if this repair was coupled to transcription by studying repair in the trpCBA operon. However my initial results were inconclusive. In Fig. 12, I list some of the plasmids constructed for these experiments with the hope that they will be used by someone else to characterize transcription-coupled repair in this species.

Conclusions

The results of my analysis show that the extreme UV resistance of *H. volcanii* is explained at least in part by efficient repair. Earlier studies of repair in halophiles did not detect repair probably because too high doses of UV were used and because the time points examined were too soon after irradiation. This shows that DNA repair experiments should be done in coordination with cellular duplication period. Additional studies are needed to determine the mechanism of the observed repair. In addition it would be of interest to study some other aspects of repair in this species. For example, since genetic methods are available in *H. volcanii* it would be of interest to try to isolate UV sensitive mutants. Such mutants provided a wealth of information about repair in species such as *E. coli* and *S. cerevisiae* and humans and would be of interest in Archaea as well. In addition, it would be useful to try to isolate mutator strains of this species, since such strains in other species frequently contain defects in DNA repair processes. It would also be interesting to study whether there are any unusual forms of DNA damage due to the extremely high intracellular salt conditions found in this species and whether there are novel forms of repair to deal with such damage. Another possible area of research is in the desiccation resistance of *H. volcanii*, since such resistance has been found to be linked to DNA repair processes in other extremely radiation resistant species like *Deinococcus radiodurans* (52). Finally, since targeted disruptions are possible in this species, it would be useful to search for homologs in this species of repair genes that have been characterized in other species and to make knockouts of any such genes. Along these lines, I used degenerate PCR to try and clone some such homologs and was able to clone a MutL homolog but have not yet made a knockout (Appendix F).

ACKNOWLEDGEMENTS

My work on halophile repair benefited from help from many people. All the repair experiments were made possible by help from the members of the Hanawalt Lab, in particular Alan Smith and Ann Ganesan. Working with *H. volcanii* would not have been possible without help from W. Ford Doolittle and Patrick Keeling.

REFERENCES

1. Kiener, A., Husain, I., Sancar, A. and Walsh, C. (1989) *J Biol Chem*, **264,** 13880-13887.
2. Iwasa, T., Tokutomi, S. and Tokunaga, F. (1988) *Photochem Photobiol*, **47,** 267-270.
3. Jorns, M. S. (1990) *Biofactors*, **2,** 207-211.
4. Takao, M., Kobayashi, T., Oikawa, A. and Yasui, A. (1989) *J Bacteriol*, **171,** 6323-6329.
5. DiRuggiero, J., Santangelo, N., Nackerdien, Z., Ravel, J. and Robb, F. T. (1997) *J Bacteriol*, **179,** 4643-4645.
6. Mukherjee, A. (1994) *Ind. J. Exp. Biol*, **32,** 565-566.
7. Koulis, A., Cowan, D. A., Pearl, L. H. and Savva, R. (1996) *FEMS Microbiol Lett*, **143,** 267-271.
8. Woods, W. G. and Dyall-Smith, M. L. (1997) *Mol Microbiol*, **23,** 791-797.
9. Hescox, M. A. and Carlberg, D. M. (1972) *Can J Microbiol*, **18,** 981-985.
10. Grey, V. L. and Fitt, P. S. (1976) *Biochem J*, **156,** 569-575.
11. Fitt, P. S., Sharma, N. and Castellanos, G. (1983) *Biochim Biophys Acta*, **739,** 73-78.
12. Eker, A. P. M., Formenoy, L. and De Wit, L. E. A. (1991) *Photochem Photobiol*, **53,** 643-652.
13. Fitt, P. S. and Sharma, N. (1987) *Biochim Biophys Acta*, **910,** 103-110.
14. Woese, C., Kandler, O. and Wheelis, M. (1990) *Proc Natl Acad Sci U S A*, **87,** 4576-4579.
15. Forterre, P. (1997) *Curr Opin Genet Dev*, **7,** 764-770.
16. Brown, J. R. and Doolittle, W. F. (1997) *Microbiol Mol Biol Rev*, **61,** 456-502.
17. Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. and Miyata, T. (1991) *J Mol Evol*, **32,** 70-78.
18. Bucher, P. and Trifonov, E. (1986) *Nucleic Acids Res*, **13,** 1009-1026.
19. Zillig, W., Palm, P., Reiter, W., Gropp, F., Puhle, G. and Klenk, H.-P. (1988) *Eur J Biochem*, **173,** 473-482.
20. Sandman, K., Krzycki, J. A., Dobrinski, B., Lurz, R. and Reeve, J. N. (1990) *Proc Natl Acad Sci U S A*, **87,** 5788-5791.
21. Forterre, P., Charbonnier, F., Marguet, E., Harper, F. and Henckes, G. (1992) *Biochem Soc Symp*, **58,** 99-112.
22. Pereira, S. L., Grayling, R. A., Lurz, R. and Reeve, J. N. (1997) *Proc Natl Acad Sci U S A*, **94,** 12633-12637.
23. Reeve, J. N., Sandman, K. and Daniels, C. J. (1997) *Cell*, **89,** 999-1002.
24. Surralles, J., Puerto, S., Ramirez, M. J., Creus, A., Marcos, R., Mullenders, L. H. and Natarajan, A. T. (1998) *Mutat Res*, **404,** 39-44.
25. Koehler, D. R. and Hanawalt, P. C. (1996) *Nucleic Acids Res*, **24,** 2877-2884.
26. Searcy, D. G. and Hixon, W. G. (1991) *Biosystems*, **25,** 1-11.

27. Trent, J. D., Kagawa, H. K., Yaoi, T., Olle, E. and Zaluzec, N. J. (1997) *Proc Natl Acad Sci U S A*, **94,** 5383-5388.

28. Sogin, M. (1991) *Curr Opin Genet Dvlp*, **1,** 457-463.

29. Zillig, W. (1991) *Curr Opin Genet Dvlp*, **1,** 544-551.

30. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. and Miyata, T. (1989) *Proc Natl Acad Sci U S A*, **86,** 9355-9359.

31. Gogarten, J., Kibak, H., Dittrich, P., Taiz, L., Bowman, E., Bowman, B., Manolson, M., Poole, R., Date, T., Oshima, T., et al. (1989) *Proc Natl Acad Sci U S A*, **86,** 6661-6665.

32. Woese, C. (1987) *Microbiol Rev*, **51,** 221-271.

33. Doolittle, W. F., Lam, W. L., Schalkwyk, L. C., Charlebois, R. L., Cline, S. W. and Cohen, A. (1992) *Biochem Soc Symp*, **58,** 73-78.

34. Charlebois, R. L., Schalkwyk, L. C., Hofman, J. D. and Doolittle, W. F. (1991) *J Mol Biol*, **222,** 509-524.

35. Cline, S. W., Schalkwyk, L. C. and Doolittle, W. F. (1989) *J Bacteriol*, **171,** 4987-4991.

36. Cline, S. W., Lam, W. L., Charlebois, R. L., Schalkwyk, L. C. and Doolittle, W. F. (1989) *Can J Microbiol*, **35,** 148-152.

37. Cline, S. W. and Doolittle, W. F. (1992) *J Bacteriol*, **174,** 1076-1080.

38. Holmes, M. L., Nuttall, S. D., Dyall, S. and M, L. (1991) *J Bacteriol*, **173,** 3807-3813.

39. Holmes, M. L., Dyall, S. and M, L. (1991) *J Bacteriol*, **173,** 642-648.

40. Charlebois, R. L., Hofman, J. D., Schalkwyk, L. C., Lam, W. L. and Doolittle, W. F. (1989) *Can J Microbiol*, **35,** 21-29.

41. Bonelo, G., Megías, M., Ventosa, A., Nieto, J. J. and Ruiz-Berraquero, F. (1984) *Curr Microbiol*, **11,** 165-170.

42. Fernandez, C. R., Nieto, J. J., Megias, M. and Ruiz, B. F. (1990) *Curr Microbiol*, **21,** 83-90.

43. Joshi, P. and Dennis, P. P. (1993) *J Bacteriol*, **175,** 1561-1571.

44. Joshi, P. and Dennis, P. P. (1993) *J Bacteriol*, **175,** 1572-1579.

45. McCready, S. (1996) *Mutat Res*, **364,** 25-32.

46. Ogrunc, M., Becker, D. F., Ragsdale, S. W. and Sancar, A. (1998) *J Bacteriol*, **180,** 5796-5798.

47. Sambrook, J., Fritsch, E. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2 Ed., Cold Sring Harbor Laboratory, Cold Spring Harbor, New York.

48. Crowley, D. J. and Hanawalt, P. C. (1998) *J Bacteriol*, **180,** 3345-3352.

49. Spivak, G. and Hanawalt, P. C. (1995) *Methods*, **7,** 147-161.

50. van Zeeland, A. A., Smith, C. A. and Hanawalt, P. C. (1981) *Mutat Res*, **82,** 173-189.

51. Sharma, N., Hepburn, D. and Fitt, P. S. (1984) *Biochim Biophys Acta*, **799,** 135-142.

52. Battista, J. R. (1997) *Annu Rev Microbiol*, **51,** 203-224.

Repair Studies in Different Organisms

(determined by Medline searches)

Humans	7028
E. coli	3926
S. cerevisiae	988
Drosophila	387
B. subtilits	284
S. pombe	116
Xenopus sp.	56
C. elegans	25
A. thaliana	20
Methanogens	16
Haloferax sp.	5
Giardia	0

Table 2. Salient features of Halophiles

-First described as contaminants on salted meats
-Square shaped cells
-Grow in saline lakes and other high salt environments (1.5 to 4.5 M; sea water = 0.5M)
-Accumulate inorganic ions (usu. K+) to maintain osmolarity

-e.g., for *H. salinarum*

Ion	Outside	Inside
Na^+	3.30 M	0.80 M
K^+	0.05 M	5.30 M
Mg^{++}	0.13 M	0.12 M
Cl^-	3.30 M	3.30 M

-Some grow in very high pH (e.g., Mono Lake)
-Most are aerobic (most other Archaea are anaerobic)
-Some species are photosynthetic (use bacteriorhodopsin to synthesize ATP). These are the only photosynthetic Archaea.
-Membranes and proteins have unique adaptations to high salt conditions (e.g. low in hydrophobic residues)
-Most use a.a. for carbon

Table 3. Haloferax volcanii Notes.

Model halophile for molecular biology and genetics.

- Physical map available.
- Genetic map available.
- Ordered cosmid library available.
- Transformation, shuttle vectors available.
- Transcription maps available.
- Grows aerobically at 37-45°C.
- Faster growing than most Halophiles.

Other features

- Isolated from Dead Sea.
- Fastest grower of Halophiles.
- Optimal growth 1.7-2.5 M NaCl.
- Requires >> 1.0M NaCl in media.
- GC content = 64%.
- Genome = 4140 kbp.
    - main chromosome =2920 kbp
    - pHV4 = 690 kbp
    - pHV3 = 442 kbp

Figure 1. Absorption ratio of a culture of *H. volcanii* and a 1:10 dilution of this culture.

A culture of *H. volcanii* was grown to high density in minimal media. The absorption of this culture, and a 10x dilution of this culture, was measured using a spectrophotometer (using growth media without cells as a blank). The figure shows a plot of the ratio of the absorption of the 1x versus the 1:10x culture. Wavelengths at which the ratio is 1:10 indicate that the absorption a this wavelength corresponds well to density of cells.

H.volcanii   Absorbance vs. Cell Density

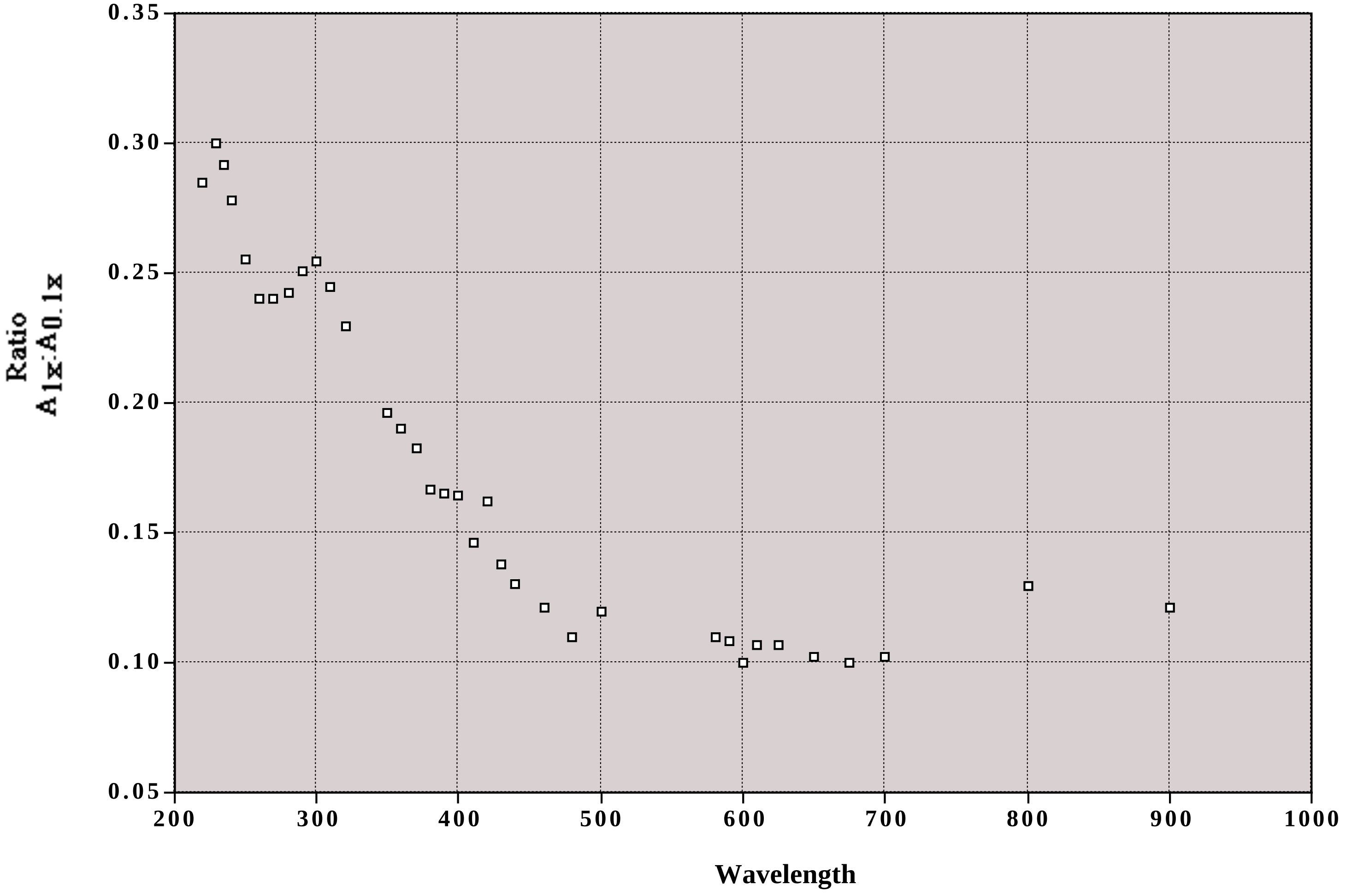


Figure 2. Relationship between number of colony forming units and optical density.

The absorption of cultures of *H. volcanii* was measured as described in Figure 1. The number of colonies was determined by plating serial dilutions of each time point in 10 ul drops onto minimal plates, incubating at 37°C and counting the resulting colonies. The graph includes data from cultures in both minimal and rich media.

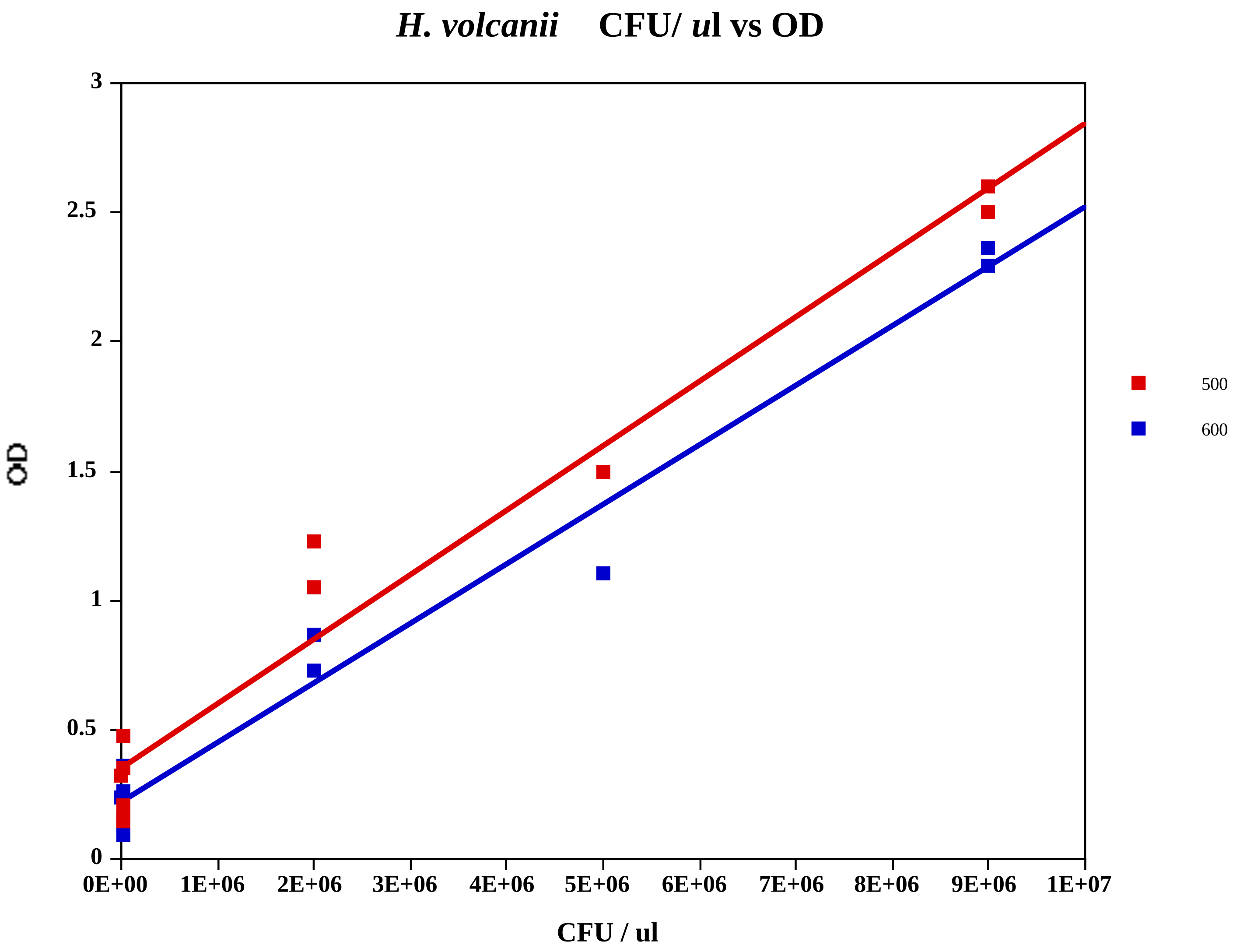
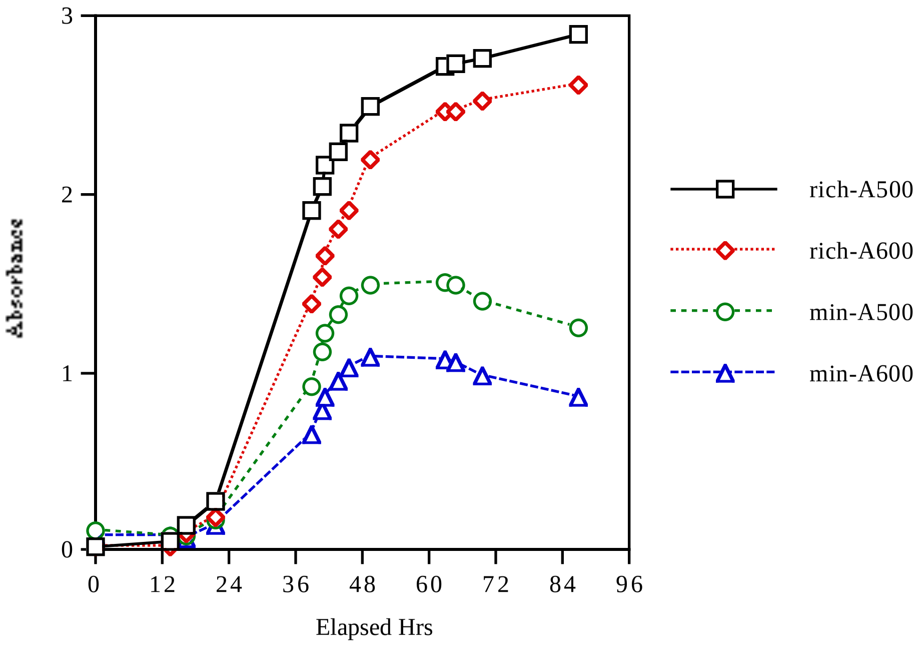

H. volcanii CFU/ *u*l vs OD
OD
3
2.5
2
1.5
1
0.5
0
0E+00
1E+06
2E+06
3E+06
4E+06
5E+06
6E+06
7E+06
8E+06
9E+06
1E+07
CFU / ul
500
600

Figure 3. Growth of *Haloferax volcanii*

H. volcanii cells were grown in rich and minimal media and absorption measurements were taken over time as described in Figure1. A. Growth in rich media. B. Growth in rich and minimal media.

H. volcanii growth
6.96



H.volcanii growth in Rich Media (9.95)

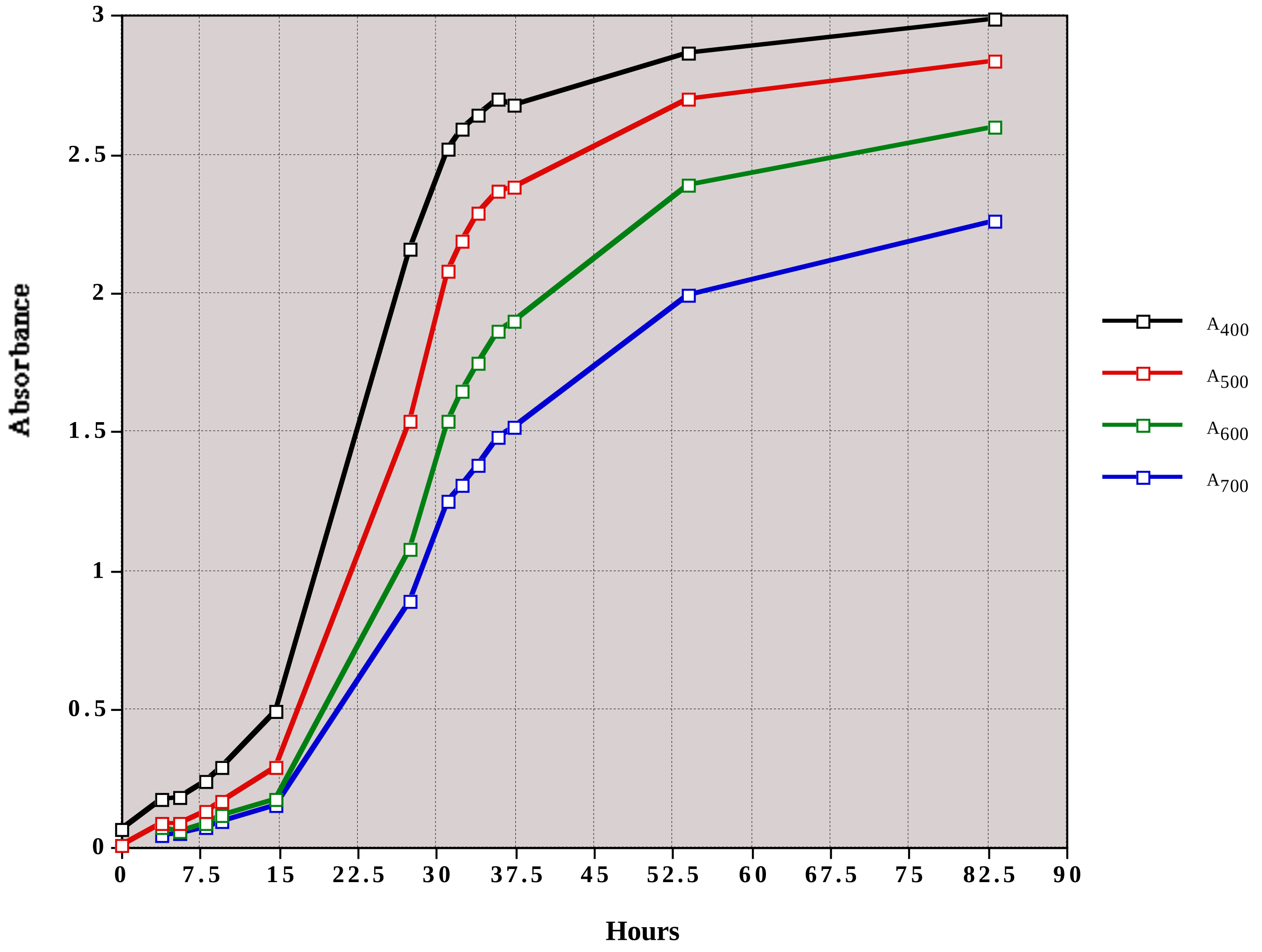
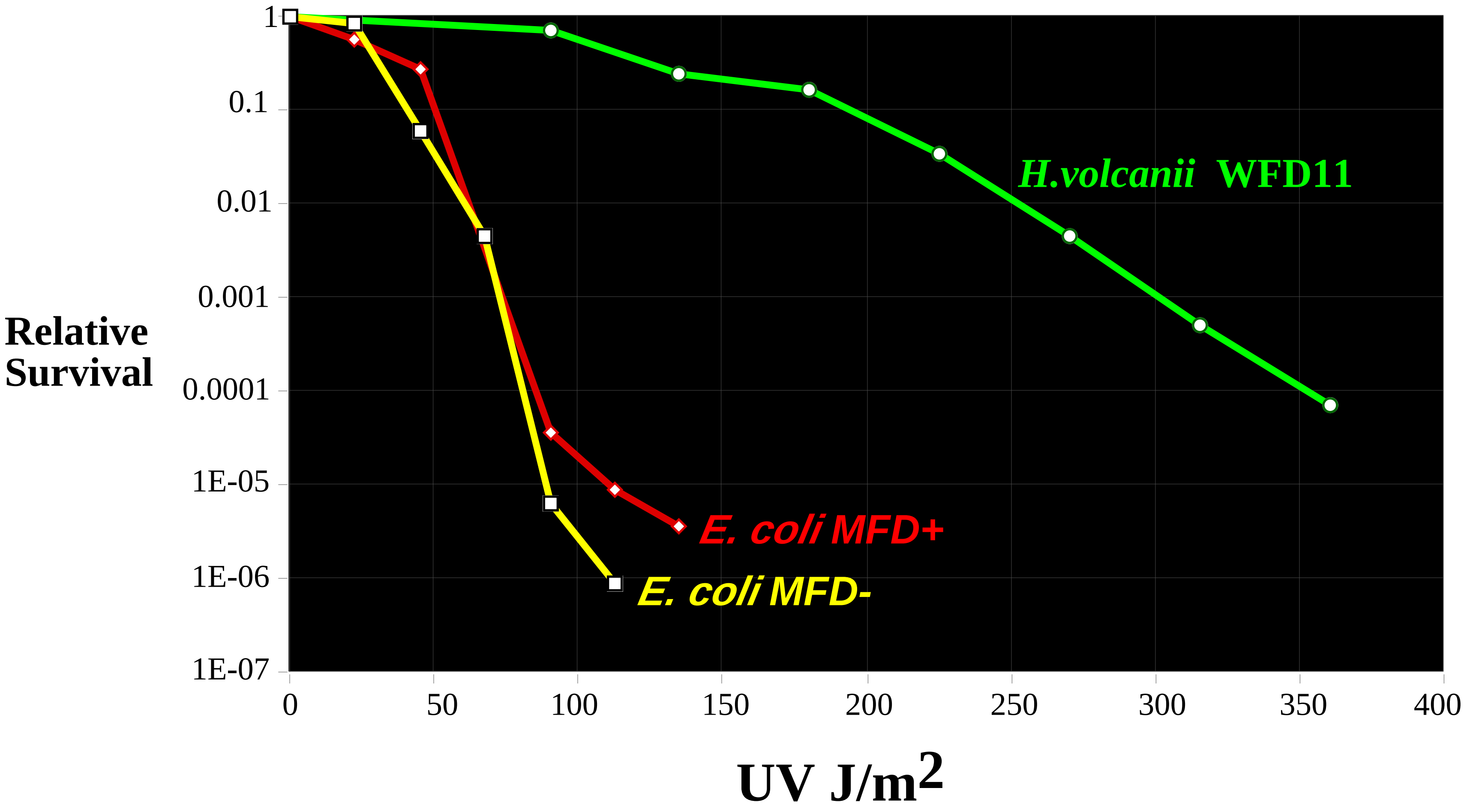


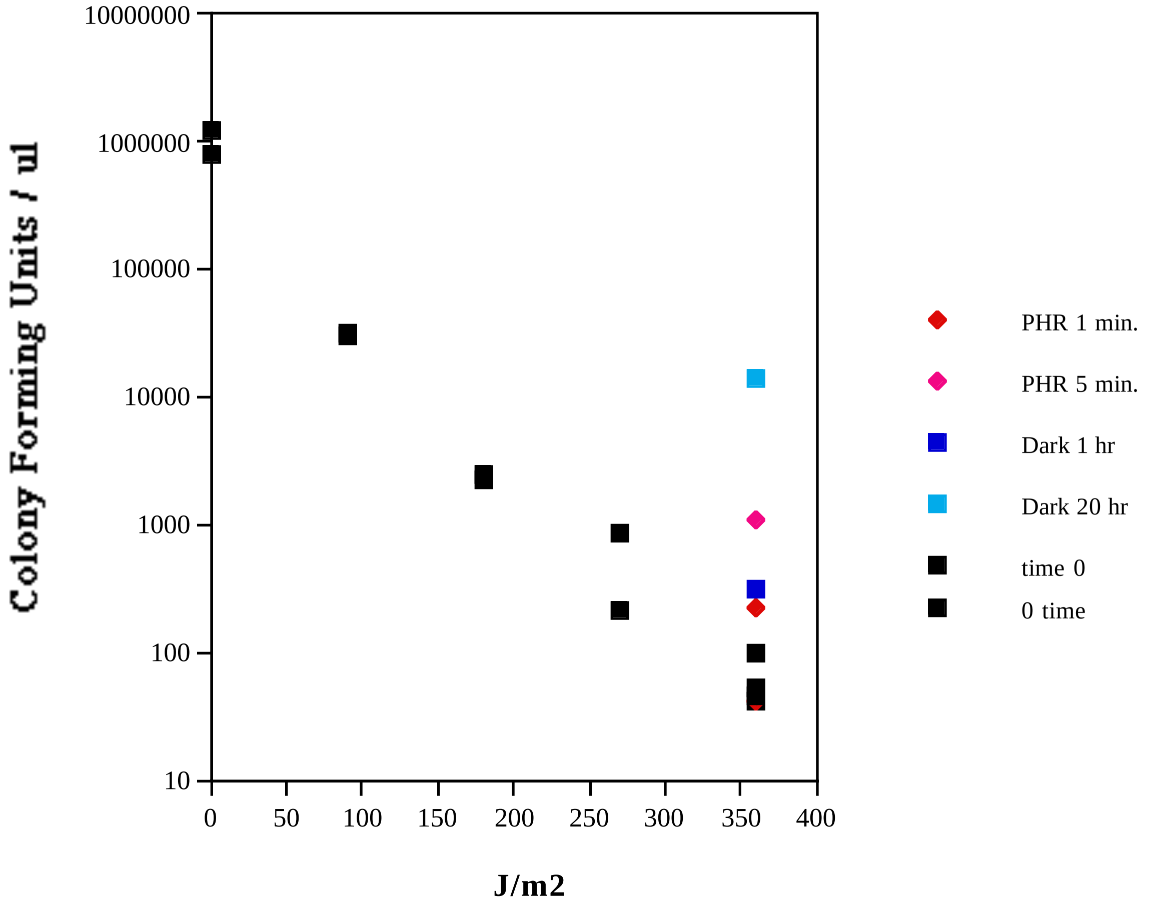
Figure 4. UV survival of *Haloferax volcanii* .

H. volcanii cells were grown to different phases of the cells cycle, spun down, resuspended in minimal media, and were exposed to UV irradiation. The number of colony forming units was determined by plating serial dilutions of each time point in 10 ul drops onto minimal plates, incubating at 37°C and counting the resulting colonies. A. UV survival of *H. volcanii* and *E. coli* in mid-log phase. B. UV survival of *H. volcanii* in mid-log phase with photoreactivation (exposure to white light) and liquid holding recovery (incubation at 37°C in the dark with shaking). C. Liquid holding recovery of *H volcanii.* Based on the data shown in Figure 4B. D. UV survival of *H volcanii* in log and stationary phase.

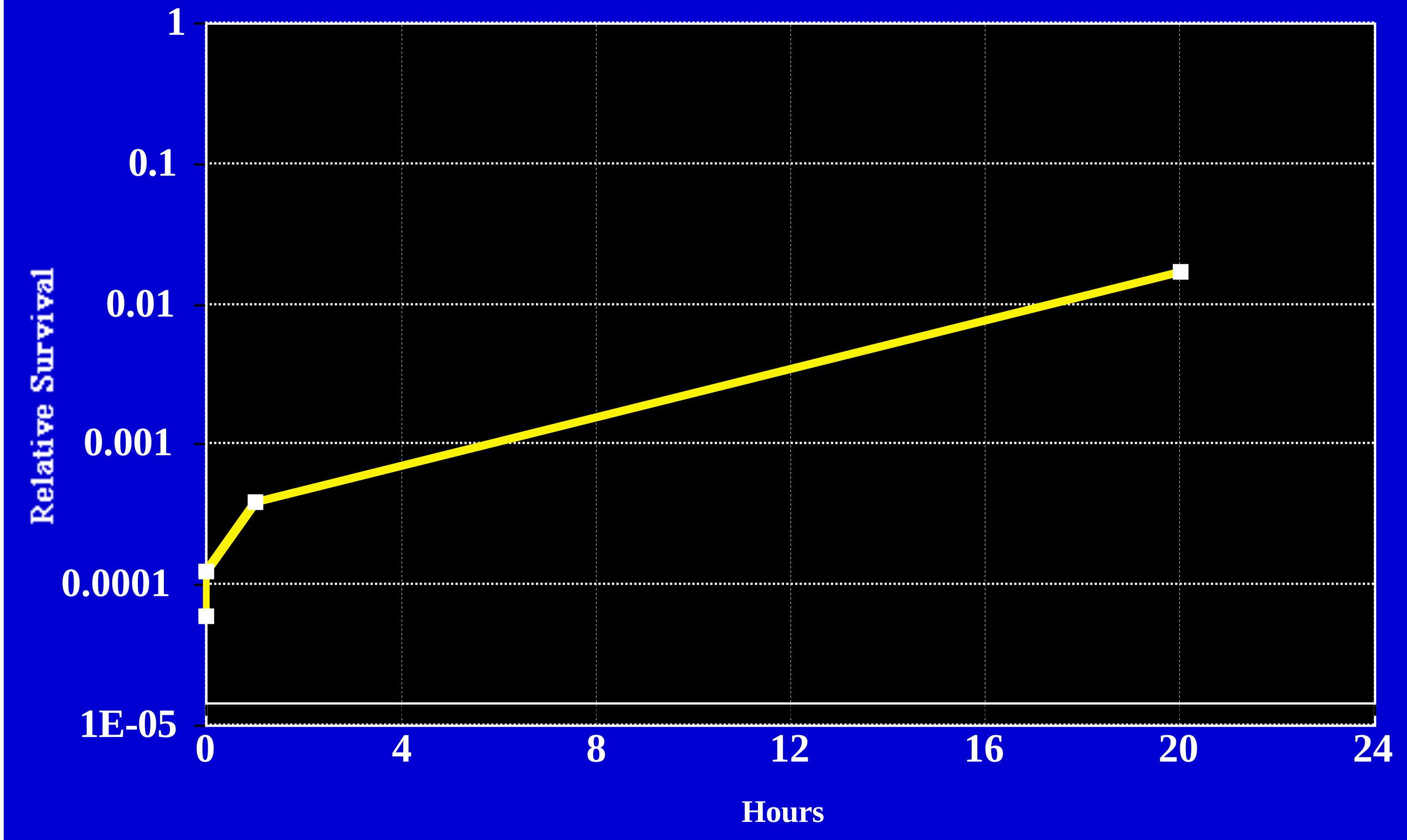
UV Survival *E. coli* vs. *H. volcanii*



H. volcanii **UV survival**
11.18.94



H. volcanii Liquid Holding Recovery (263 J/m2)



H. volcanii UV Survival

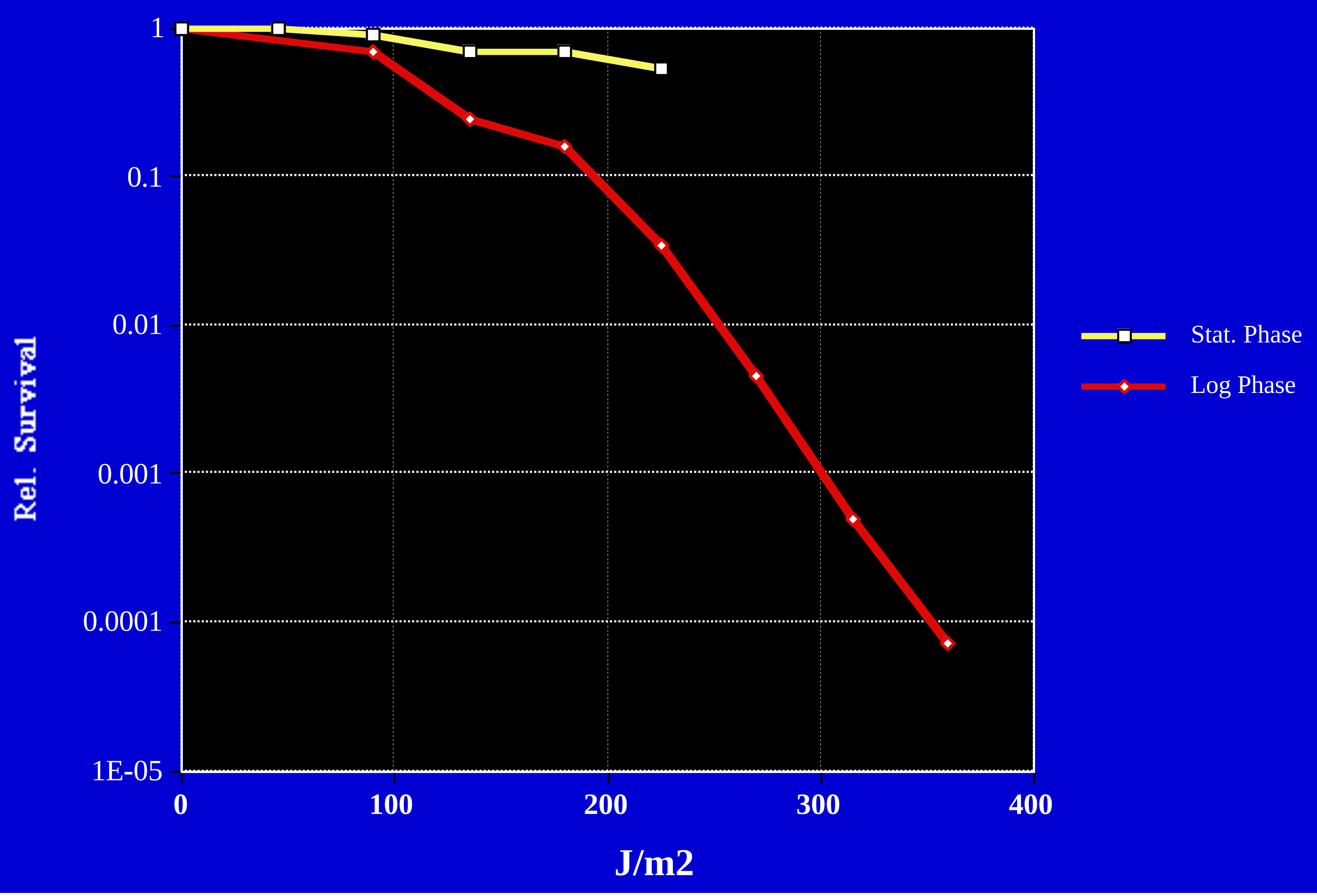


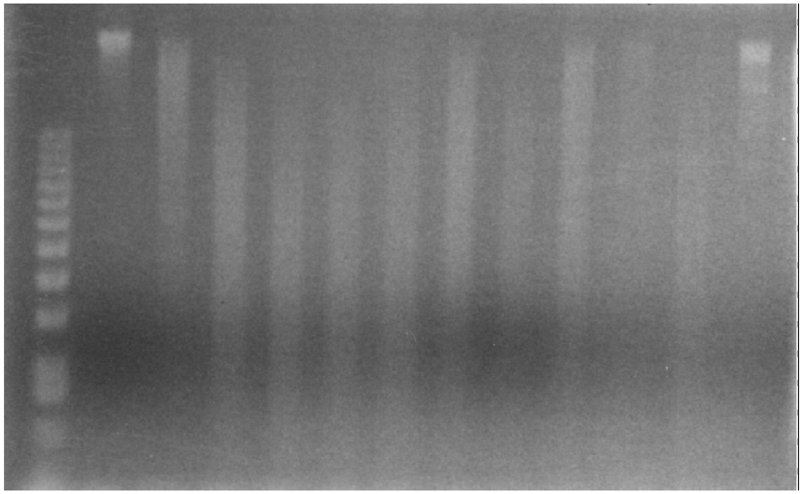
Figure 5. Removal of T4EV enzyme sensitive sites by *Haloferax volcanii* by photoreactivation and dark repair.

H. volcanii cells were grown to mid-log phase, spun down, resuspended in minimal media, exposed to UV irradiation, and incubated under different conditions. DNA was extracted, treated with T4EV and electrophoresed on alkali agarose gels. A. Experiment UV3. Doses of 0-180 J/m^2. PHR = photoreactivation after UV. LHR = liquid holding recovery after UV in minimal media at 37°C with shaking. B. Experiment Label10. Time points were for time growing after UV in minimal media at 37°C with shaking.

H. volcanii UV3

UV (min)*	0	1	2	3	4	4	4	2	2	2	4	4
PHR(min)	-	-	-	-	-	5	60	-	-	-	-	1
LHR (hrs)	-	-	-	-	-	-	-	1	4	8	1	24

* 1 min= 45 J/m^2



H. volcanii Label10

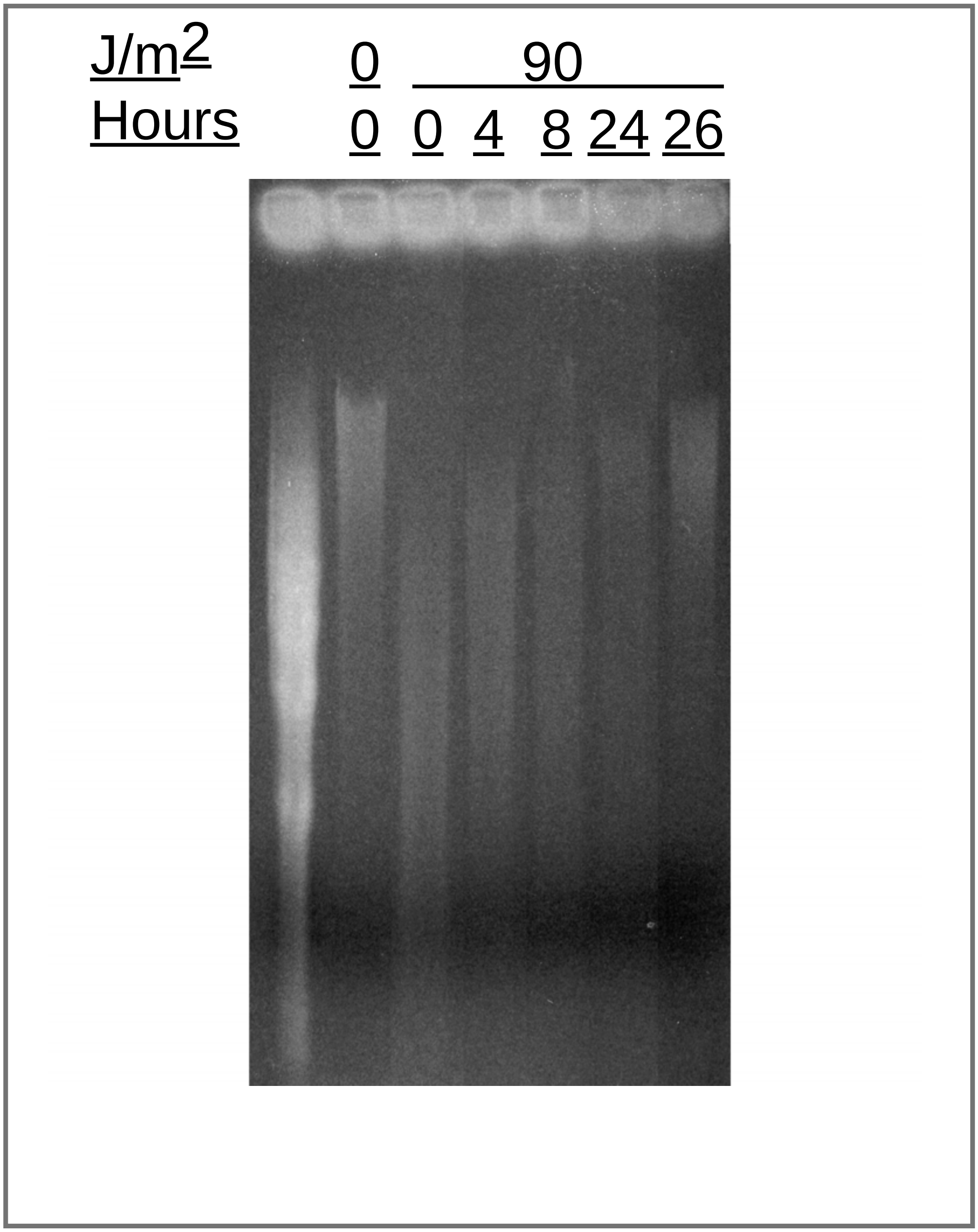
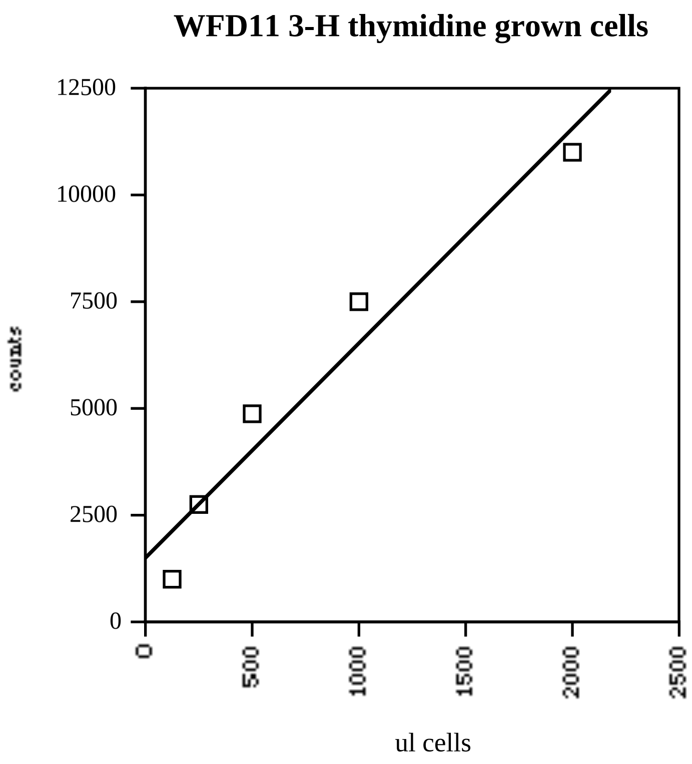
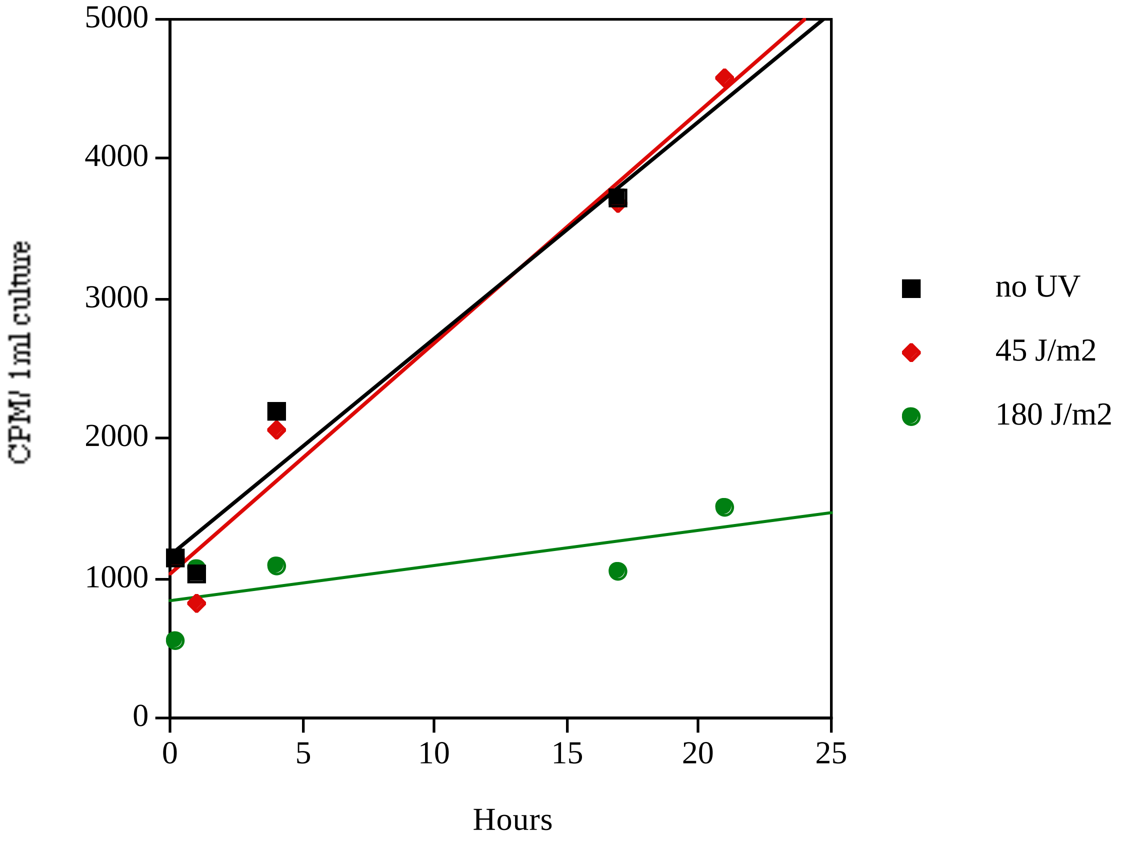


Figure 6. Incorporation of radiolabel in *Haloferax volcanii* DNA with and without UV irradiation.

H. volcanii cells were grown in minimal media with radiolabel added. Counts were determined for different volumes of cells by TCA precipitation. A. ^{3}H thymine vs. number of cells. ^{3}H thymine (1 uCi/ml) was added to 20 mls minimal media culture at mid log phase. B. Inhibition of ^{3}H thymine incorporation by UV irradiation. *H. volcanii* cells were grown to mid-log phase, spun down, resuspended in minimal media, and were exposed to different doses of UV irradiation. ^{3}H thymine was then added (1 uCi/ml) and cells were taken out at different time points. C. Inhibition of ^{3}H BrDU incorporation by UV irradiation. *H. volcanii* cells were grown to mid-log phase, spun down, resuspended in minimal media, and were exposed to different doses of UV irradiation. ^{3}H BrDU was then added (1 uCi/ml) and cells were taken out at different time points.


WFD11 3-H thymidine grown cells
counts
12500
10000
7500
5000
2500
0
0
500
1000
1500
2000
2500
ul cells

3-H-Thymine Incorporation
H. volcanii ± UV



WFD 11 3-H BrDU Incorporation ± UV

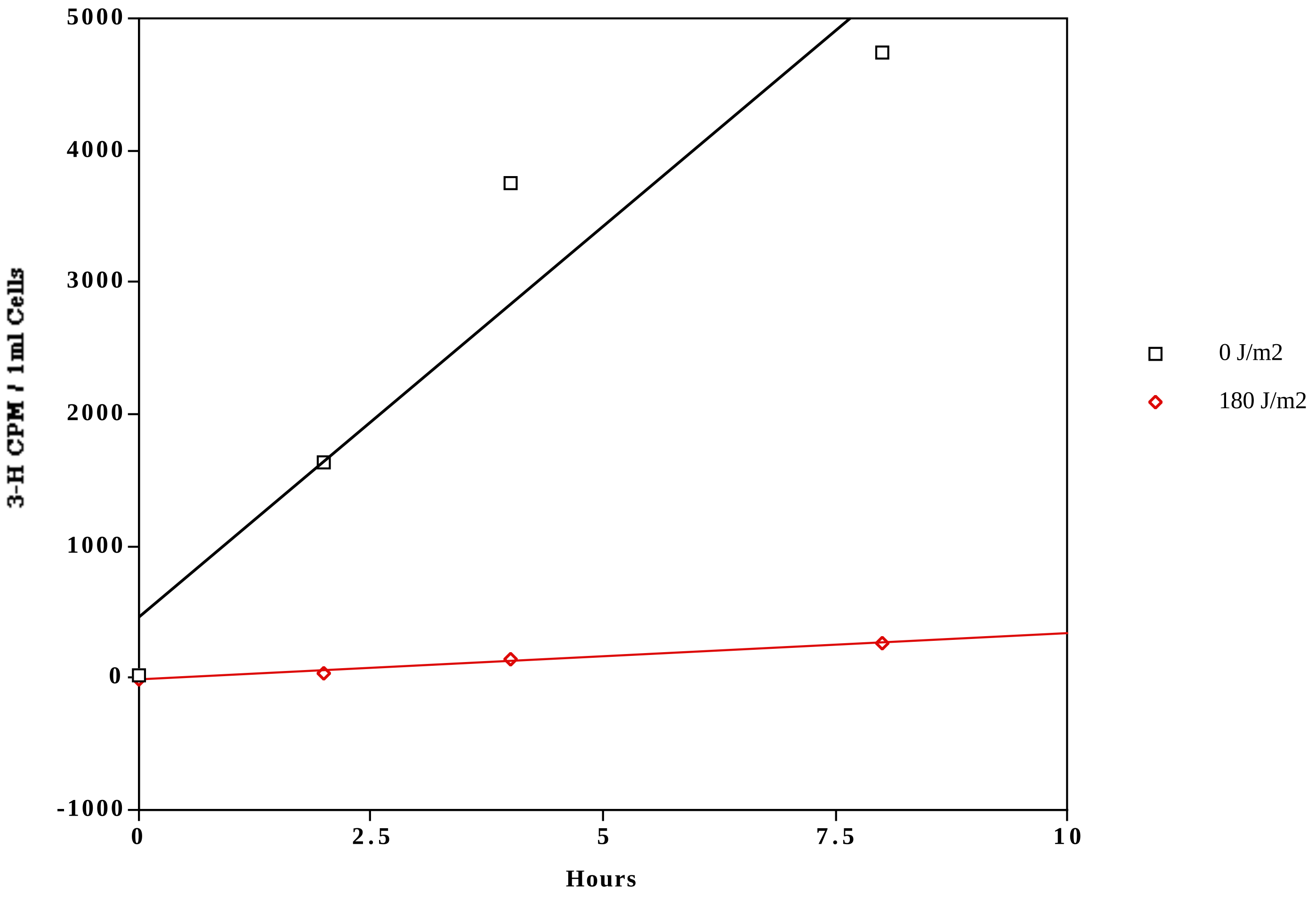
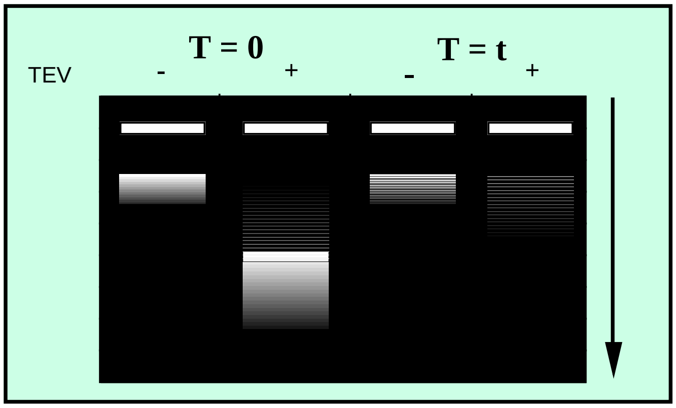


Figure 7. Whole genome DNA repair assay.

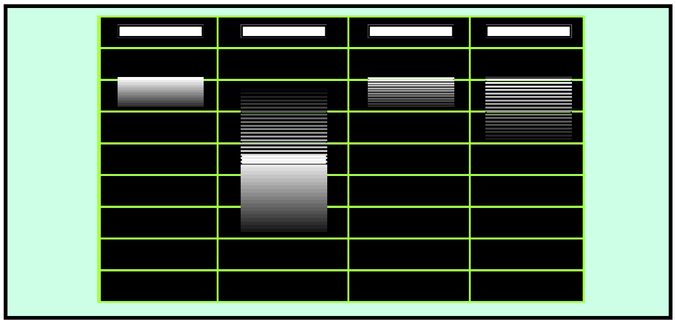
Repair in unreplicated DNA was measured by a modification of Spivak and Hanawalt (50). First, cells are grown in radioactively labeled media to pre-label the DNA. Cells are then exposed to UV and allowed different amounts of time for repair. DNA is extracted from these cells and either treated (lanes labeled +) or mock treated (lanes labeled -) with T4EV (which cuts the DNA backbone at sites of cyclobutane pyrimidine dimers), and then electrophoresed in an alkali agarose gel which separates out single stranded DNA by size. The gel cut into a grid and the amount of pre-label in each fraction is counted. Alternatively, the gel can be exposed directly to a phosphorimager. Percent repair can then be calculated from the fraction vs. CPM information as previously described (50).

Whole Genome CPD Assay

Alkali Agarose Gel (± T4EV)



Cut into grid



Count each fraction

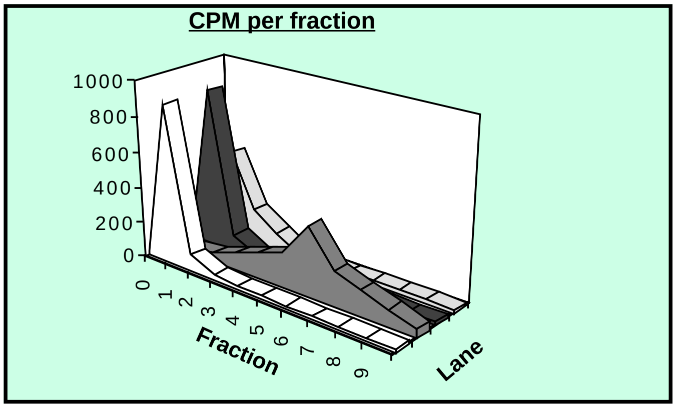
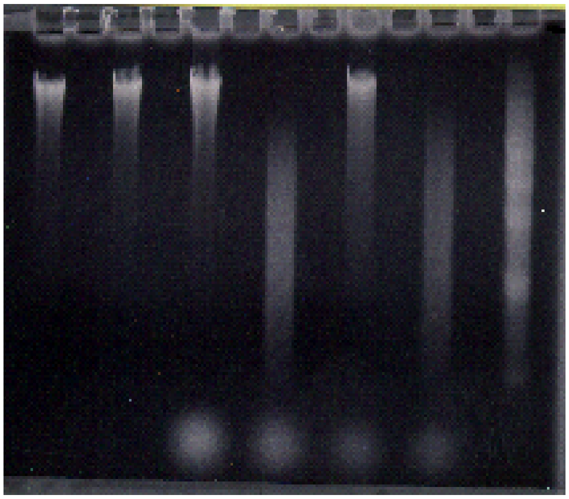


Figure 8. Dark repair of UV irradiation induced cyclobutane dimers by *Haloferax volcanii* - 180 J/m^2.

H. volcanii cells were grown to mid log phase in the presence of ^{3}H thymine, spun down, resuspended in minimal media, were exposed to UV irradiation, and allowed to repair for different amounts of time. DNA was extracted, treated with T4EV and electrophoresed on alkali agarose gels. A. Alkali agarose gel. B. Plot of CPM versus fraction. Counts were determined as described in Figure 7. JAE experiment Label 5.

H. volcanii Label5

T4EV	-	+	-	+	-	+
Hours	0	0	0	0	24	24
J/m2	0	0	180	180	180	180



H. volcanii UV repair (Label 5 - 180 J/m2)

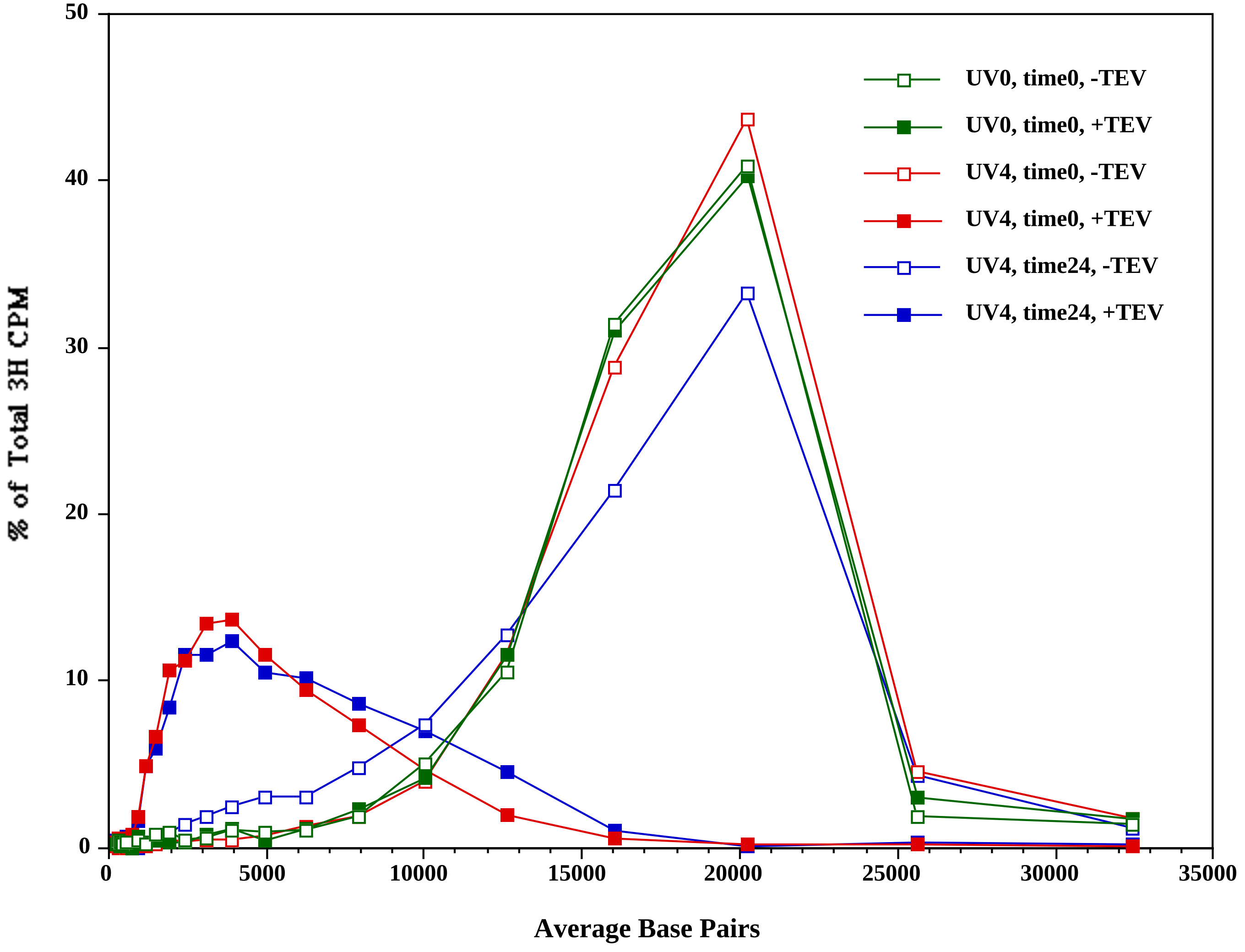
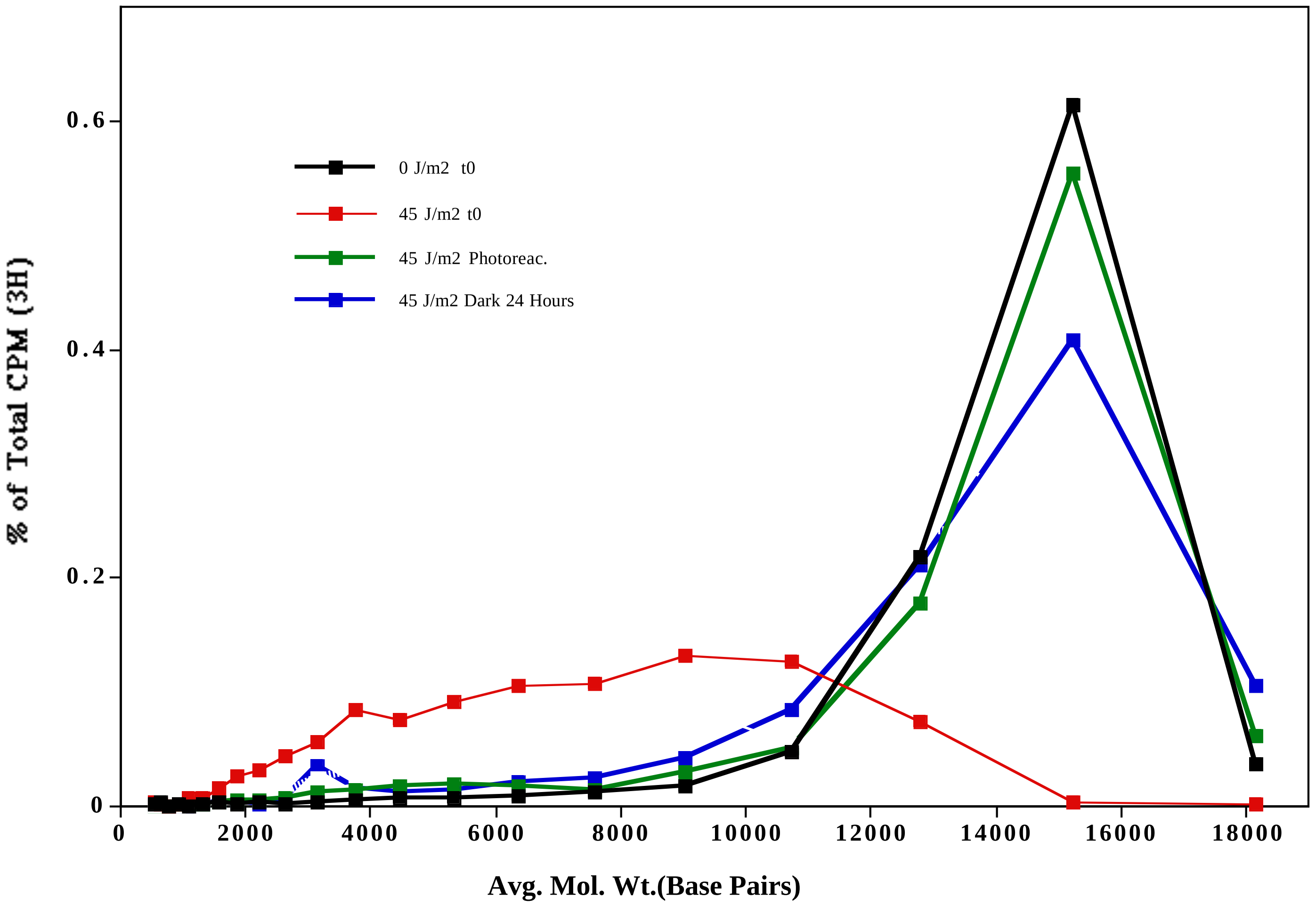


Figure 9. Dark repair and photoreactivation of UV irradiation induced cyclobutane dimers by *Haloferax volcanii* - 45 and 90 J/m^2.

H. volcanii cells were grown to mid log phase in the presence of ^{3}H thymine, spun down, resuspended in minimal media, exposed to UV irradiation, and allowed to repair for different amounts of time. DNA was extracted, treated with T4EV and electrophoresed on alkali agarose gels. A. Plot of CPM versus average molecular weight showing both plus and minus T4EV lanes. Counts were determined as described in Figure 7. B. A. Plot of CPM versus average molecular weight showing only both plus T4EV lanes. Counts were determined as described in Figure 7. JAE experiment Label 7.

H. volcanii UV Repair Label 7 - 45J / m2)



H. volcanii UV Repair (Label 7 - 45 & 90 J/ m2)

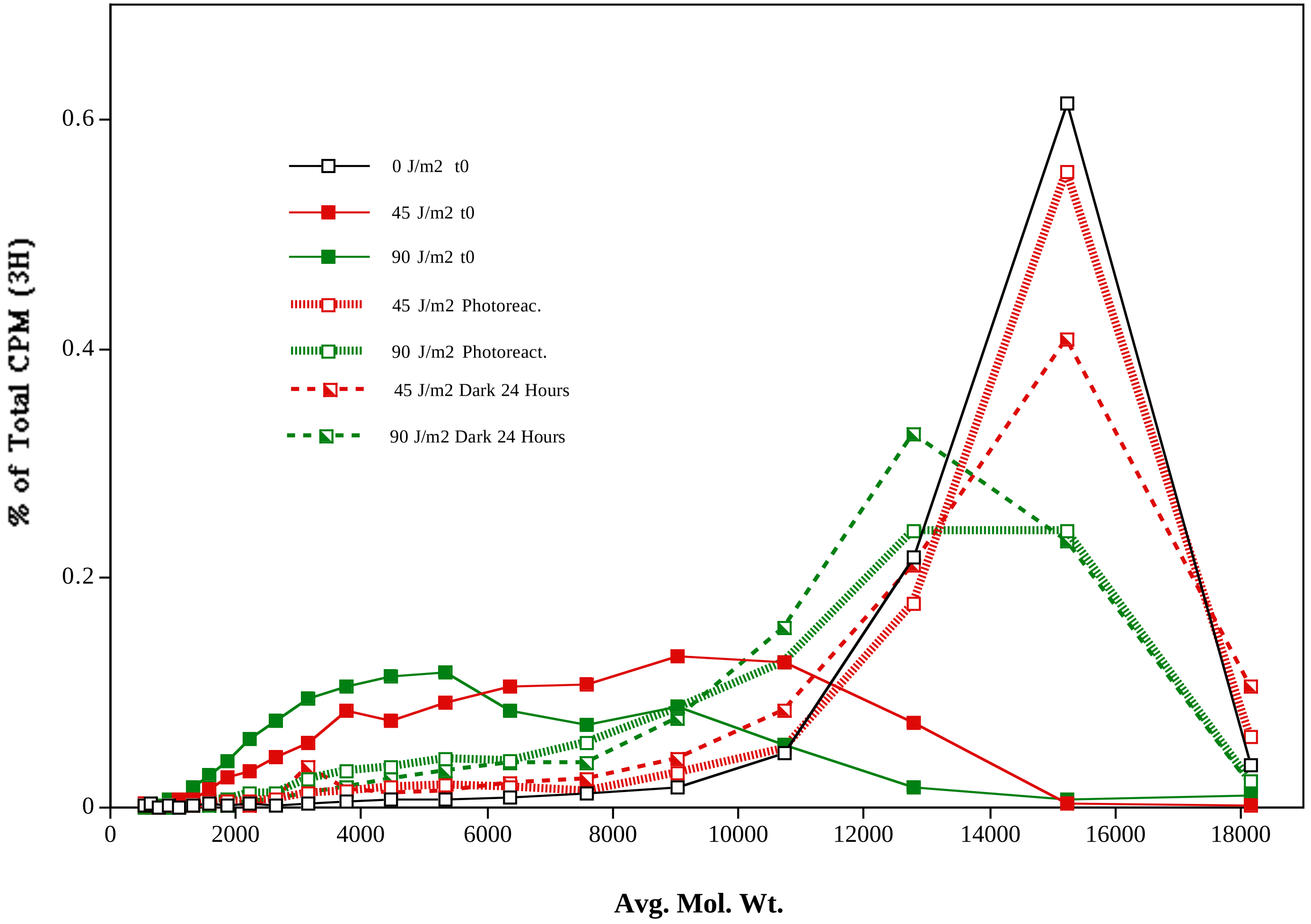


Figure 10. Dark repair of UV irradiation induced cyclobutane dimers by *Haloferax volcanii* - 45 J/m^2.

H. volcanii cells were grown to mid log phase in the presence of ^{3}H thymine, spun down, resuspended in minimal media, exposed to UV irradiation, and allowed to repair for different amounts of time. DNA was extracted, treated with T4EV and electrophoresed on alkali agarose gels. Plot of CPM versus average molecular weight showing only both plus T4EV lanes. Counts were determined as described in Figure 7. JAE experiment Label 11.

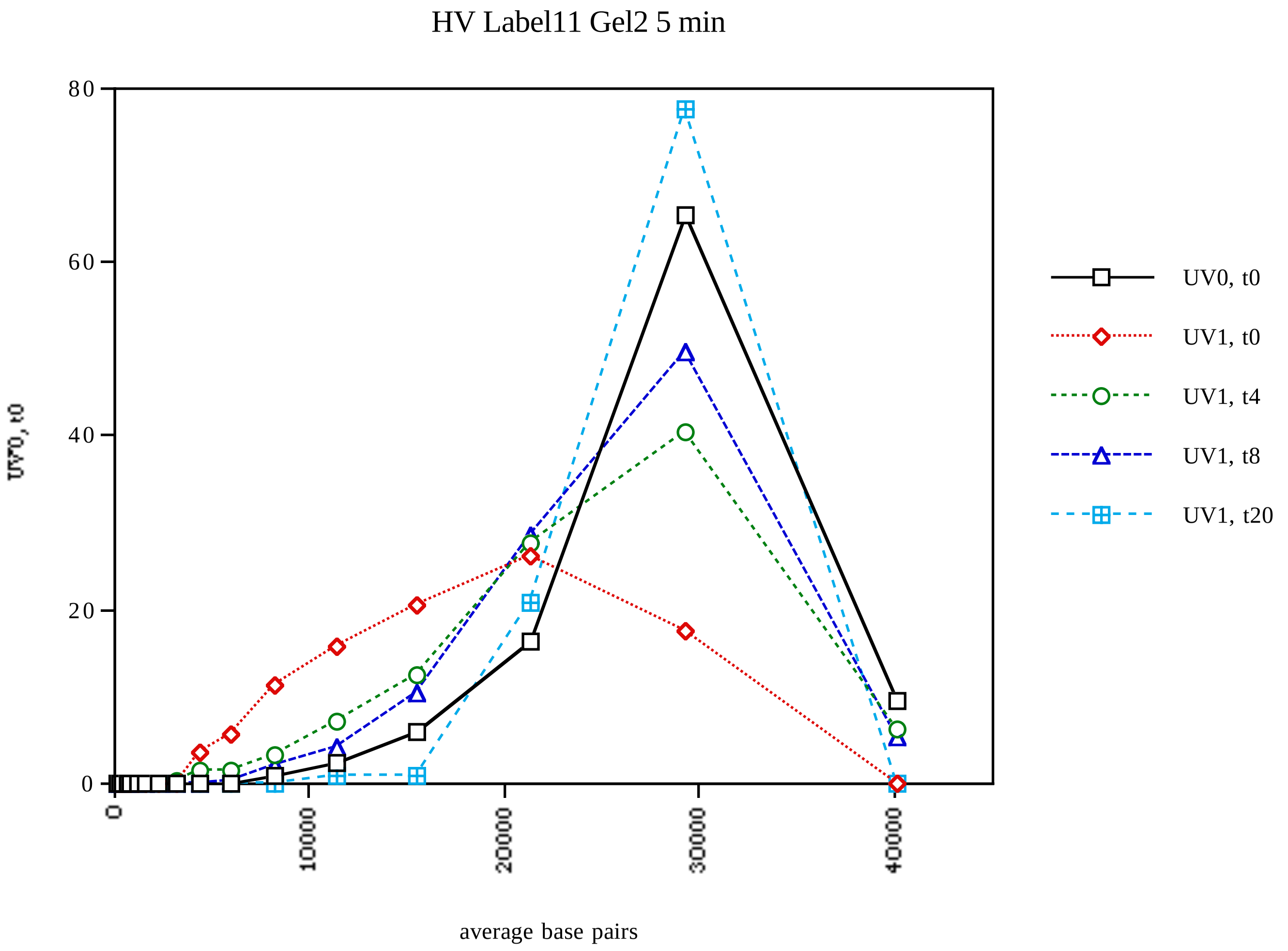

HV Label11 Gel2 5 min
80
60
40
20
0
UV0, t0
0
10000
20000
30000
40000
average base pairs
UV0, t0
UV1, t0
UV1, t4
UV1, t8
UV1, t20

Figure 11. Dark repair (in percent) of UV irradiation induced cyclobutane dimers by *Haloferax volcanii* - 45 and 90 J/m^2.

H. volcanii cells were grown to mid log phase in the presence of ^{3}H thymine, spun down, resuspended in minimal media, exposed to UV irradiation, and allowed to repair for different amounts of time. DNA was extracted, treated with T4EV and electrophoresed on alkali agarose gels. Plot of CPM versus average molecular weight showing only both plus T4EV lanes. Counts were determined as described in Figure 7. JAE experiments Label 11 and 12.

H. volcanii CPD Repair

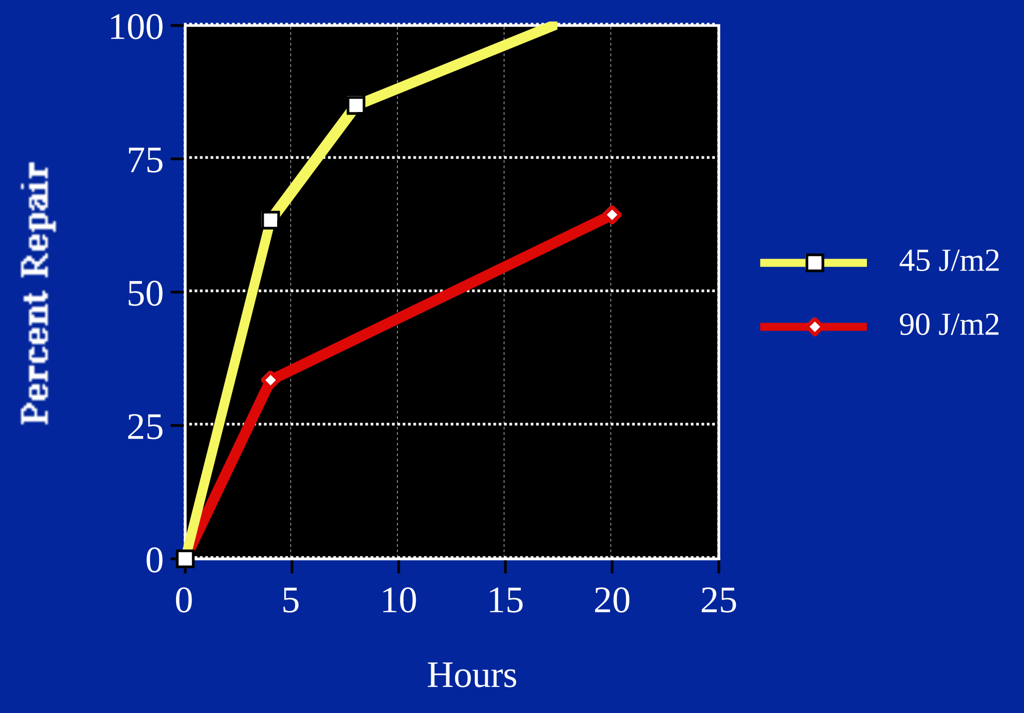
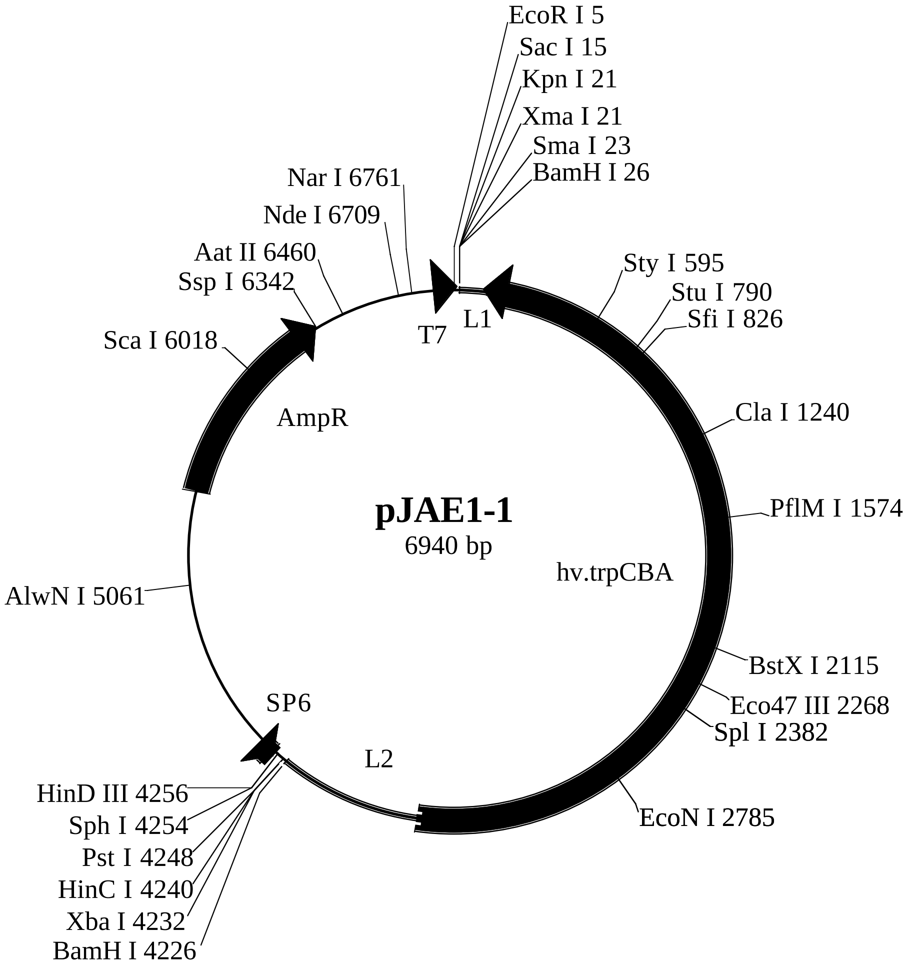


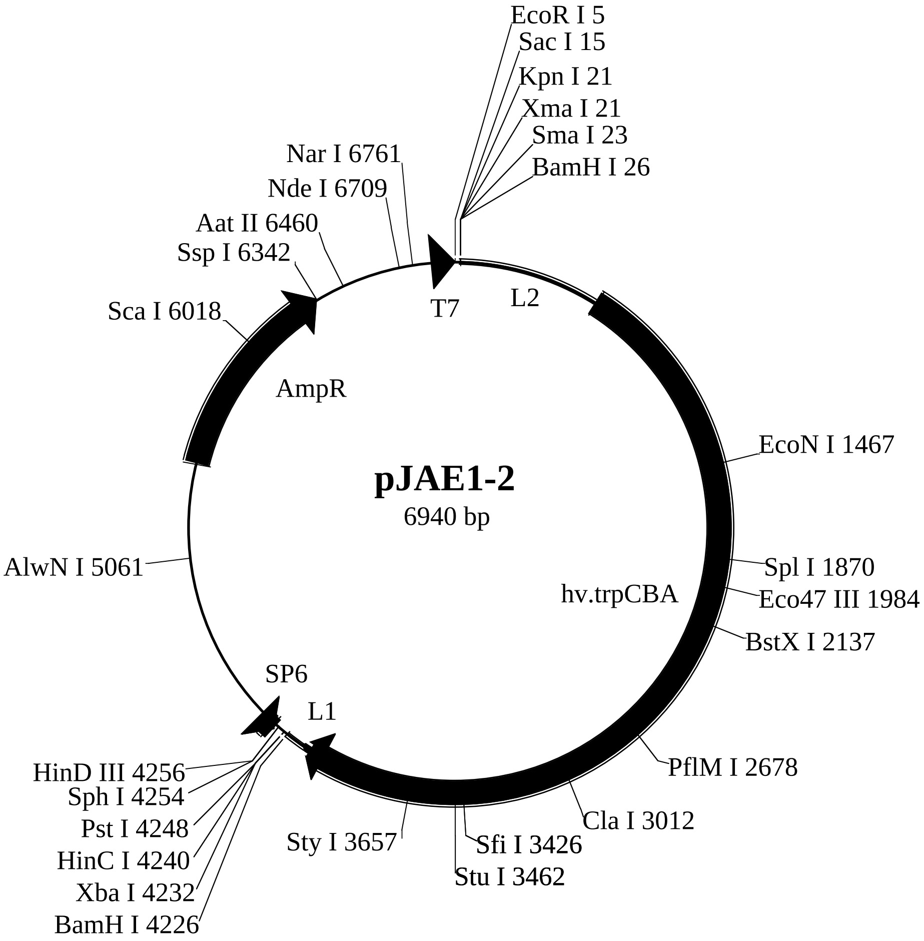
Figure 12. Maps of plasmids for transcription-coupled repair assays.

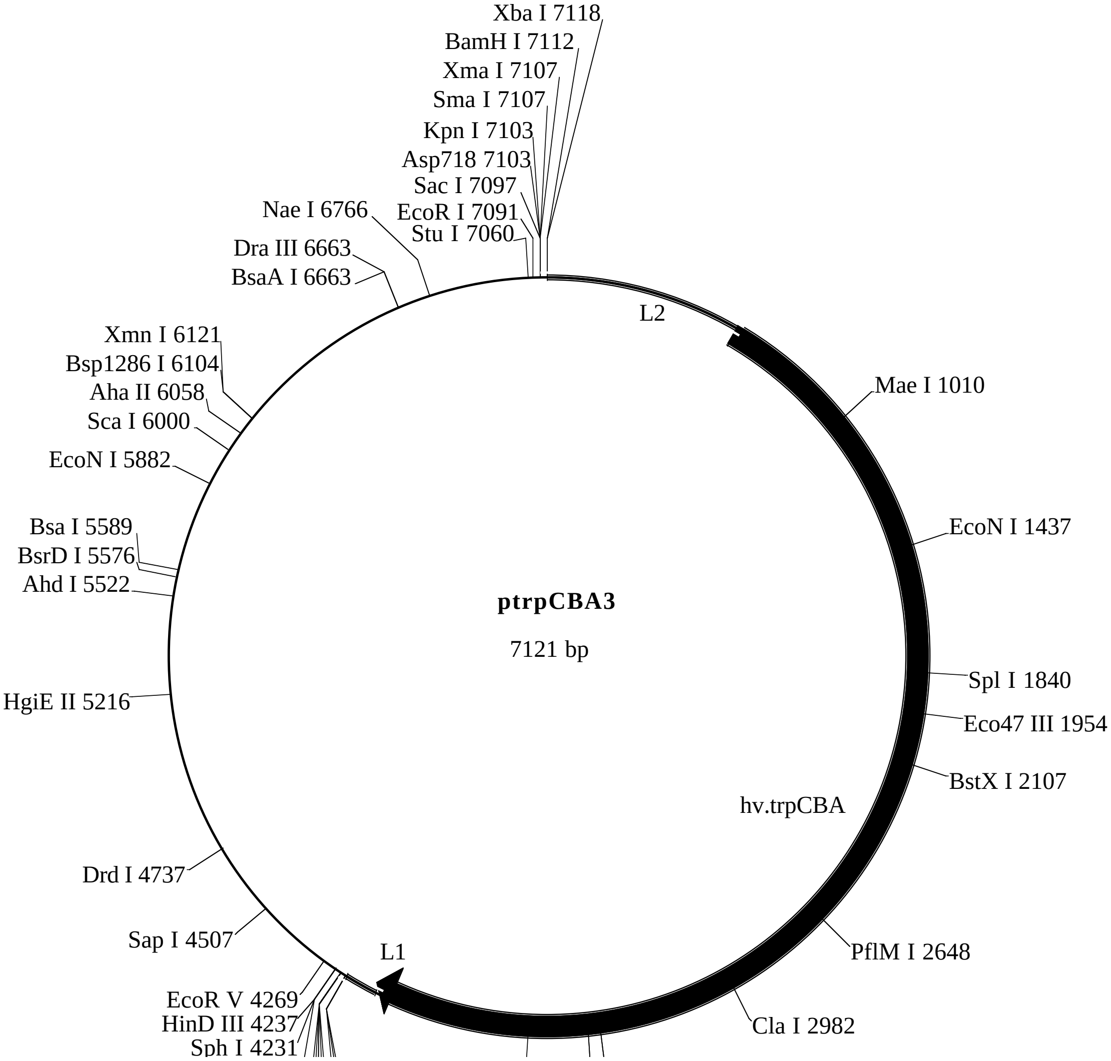
A. Map of pJAE1-1. PAS124 was cut with BamHI. The band corresponding to the trpCBA operon was extracted from a low melting point agarose gel, purified with a Wizard Kit, and cloned into BamHI cut PGEM3Z site. This plasmid is also known as Hanawalt Lab plasmid 143 and is carried in *E. coli* strain HL828.

B. Map of pJAE1-2. Cloned as for pJAE1-1. Inverse orientation relative to pJAE1-1. This plasmid is also known as Hanawalt Lab plasmid 144 and is carried in *E. coli* strain HL829.

C. Map of ptrpCBA


EcoR I 5
Sac I 15
Kpn I 21
Xma I 21
Sma I 23
BamH I 26
Nar I 6761
Nde I 6709
Aat II 6460
Ssp I 6342
Sca I 6018
Sty I 595
Stu I 790
Sfi I 826
T7
L1
AmpR
Cla I 1240
PflM I 1574
pJAE1-1
6940 bp
hv.trpCBA
AlwN I 5061
BstX I 2115
Eco47 III 2268
Spl I 2382
SP6
L2
EcoN I 2785
HinD III 4256
Sph I 4254
Pst I 4248
HinC I 4240
Xba I 4232
BamH I 4226


EcoR I 5
Sac I 15
Kpn I 21
Xma I 21
Sma I 23
BamH I 26
Nar I 6761
Nde I 6709
Aat II 6460
Ssp I 6342
Sca I 6018
T7
L2
AmpR
pJAE1-2
6940 bp
EcoN I 1467
Spl I 1870
Eco47 III 1984
BstX I 2137
hv.trpCBA
AlwN I 5061
SP6
L1
PflM I 2678
Cla I 3012
HinD III 4256
Sph I 4254
Pst I 4248
HinC I 4240
Xba I 4232
BamH I 4226
Sty I 3657
Sfi I 3426
Stu I 3462


Xba I 7118
BamH I 7112
Xma I 7107
Sma I 7107
Kpn I 7103
Asp718 7103
Sac I 7097
EcoR I 7091
Stu I 7060
Nae I 6766
Dra III 6663
BsaA I 6663
Xmn I 6121
Bsp1286 I 6104
Aha II 6058
Sca I 6000
EcoN I 5882
Bsa I 5589
BsrD I 5576
Ahd I 5522
HgiE II 5216
Drd I 4737
Sap I 4507
EcoR V 4269
HinD III 4237
Sph I 4231
L2
Mae I 1010
EcoN I 1437
ptrpCBA3
7121 bp
Spl I 1840
Eco47 III 1954
BstX I 2107
hv.trpCBA
PflM I 2648
L1
Cla I 2982

APPENDIX G

Cloning of a MutL Homolog from *Haloferax volcanii* by Degenerate PCR

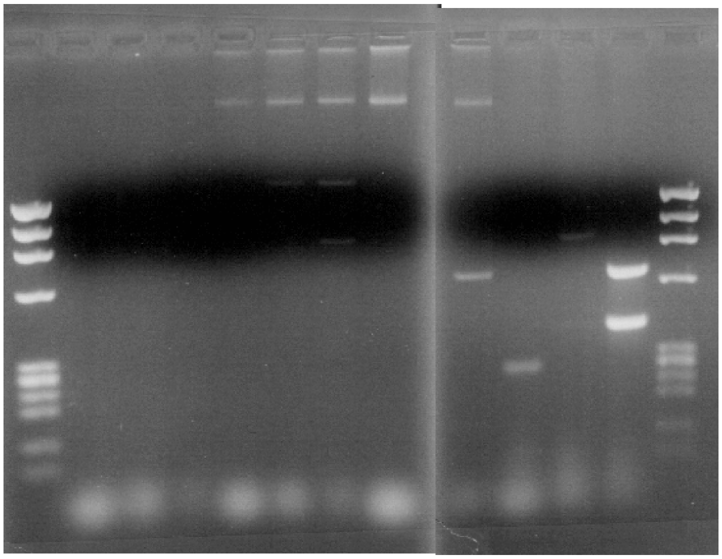
SUMMARY

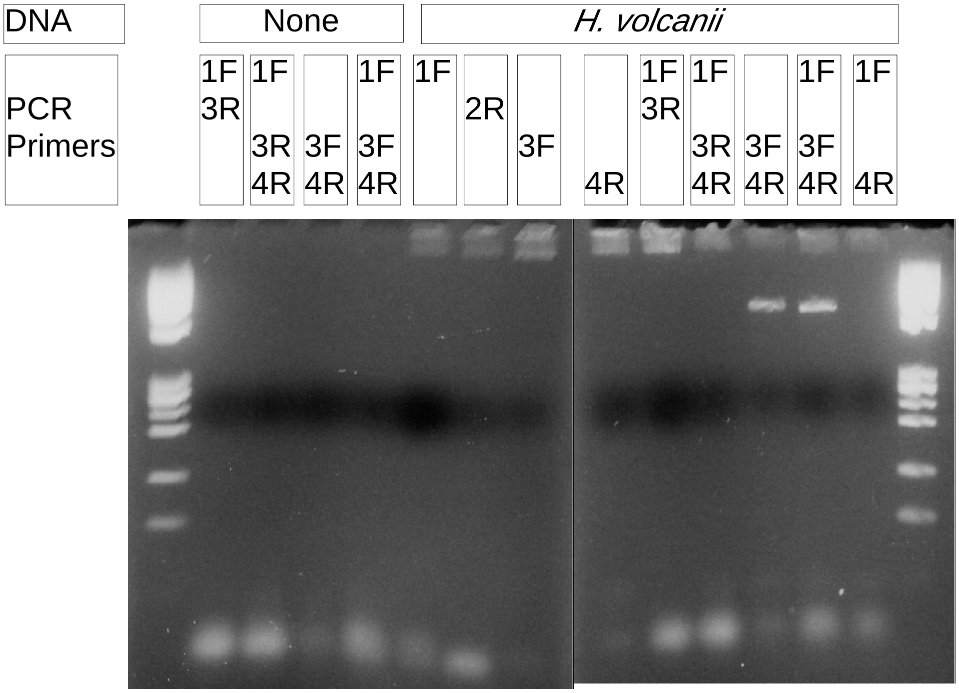
In this appendix I present results of cloning a portion of a gene encoding a homolog of MutL in *Haloferax volcanii.* The cloning of this MutL homolog is of interest since MutL homologs have not yet been found in any Archaea species (see Chapter 6 for more details on MutL in different species). All methods are described in the figure legends. DNA for the degenerate PCR came from *H. volcanii* WFD11. Figure 1 shows the results of degenerate PCR experiments in which a portion of the *H. volcanii* mutL was amplified. The PCR product corresponding to primers 3F and 4R was cloned into pGEM3Z (described in Figure 2). This insert was sequenced and the sequence clearly encodes a MutL homolog. Figure 3 shows the sequence, Figure 4 the results of blastx searches and Figure 5 and alignment with other MutL homologs. The genome position of this sequence was determined by probing a blot of the ordered cosmid library of the *H. volcanii* genome, kindly provided by W. Ford Doolittle (Figure 6).

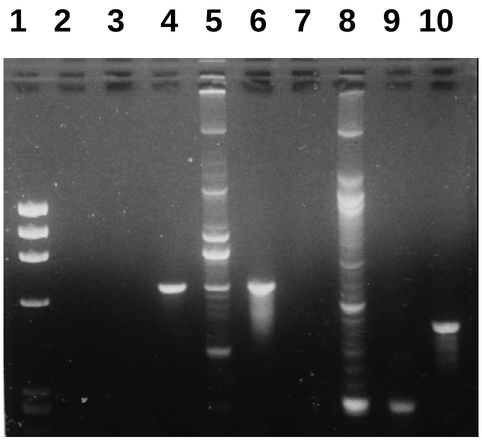
Figure 1. Degenerate PCR amplification of *Haloferax volcanii mutL*.

The degenerate PCR primers used are described in Appendix D Table1. 60 pmoles of each primer was used in a total volume of 50 ul. Thermal cycling was done on a Perkin Elmer 2400. A. Temperature parameters were 94°C x 5 minutes; 30 cycles of (94°C x 30 seconds, 50°C x 30 seconds, 72°C x 1 minute); and 72°C x 7 minutes. B. Temperature parameters were 94°C x 5 minutes; 30 cycles of (94°C x 30 seconds, 55°C x 30 seconds, 72°C x 1 minute); and 72°C x 7 minutes. C. Re PCR of *Haloferax volcanii mutL*. Temperature parameters were 94°C x 5 minutes; 30 cycles of (94°C x 30 seconds, 55°C x 30 seconds, 72°C x 1 minute); and 72°C x 7 minutes.

DNA	None			*H. volcanii*					Ecoli		
PCR Primers	1F 3R	1F 4R	3F 4R	1F 3R	1F 4R	3F 4R	1F 3R 4R	1F 3F 4R	1F 3R	1F 4R	3F 4R




DNA
None
H. volcanii
PCR Primers
1F 3R
1F 3R 4R
3F 4R
1F 3F 4R
1F
2R
3F
4R
1F 3R
1F 3R 4R
3F 4R
1F 3F 4R
1F 4R



DNA	PCR Primers
1 = phiX174 HaeIII	
2 = negative control	MutL3F, MutL4R
3 = negative control	MutS1F, MutS3R
4 = gDNA	MutL3F, MutL4R
5 = gDNA, Mg++	MutL3F, MutL4R
6 = MutL	MutL3F, MutL4R
7 = gDNA	MutS1F, MutS3R
8 = gDNA, Mg++	MutS1F, MutS3R
9 = MutS PCR	MutS1F, M
10 = phiX174 HaeIII	

Figure 2. Map of pJAE2-1 and cloning of *Haloferax volcanii mutL* PCR product.

PCR products from MutL PCR in Figure 2 (using primers F and 4R) were cut out form a low-melting point agarose gel and purified using a Wizard PCR product purification kit. The purified product was cut with SstI and PstI (restriction sites for these enzymes were part of the PCR primers) and then cloned into PGEM3Z (also cut with same enzymes). A map of the resulting plasmid is shown here. Another plasmid, pJAE12-2 should be identical to this one. These plasmids are also known as Hanawalt Lab plasmids 145, 146. *E. coli* strains carrying this plasmid are HL830, HL831.

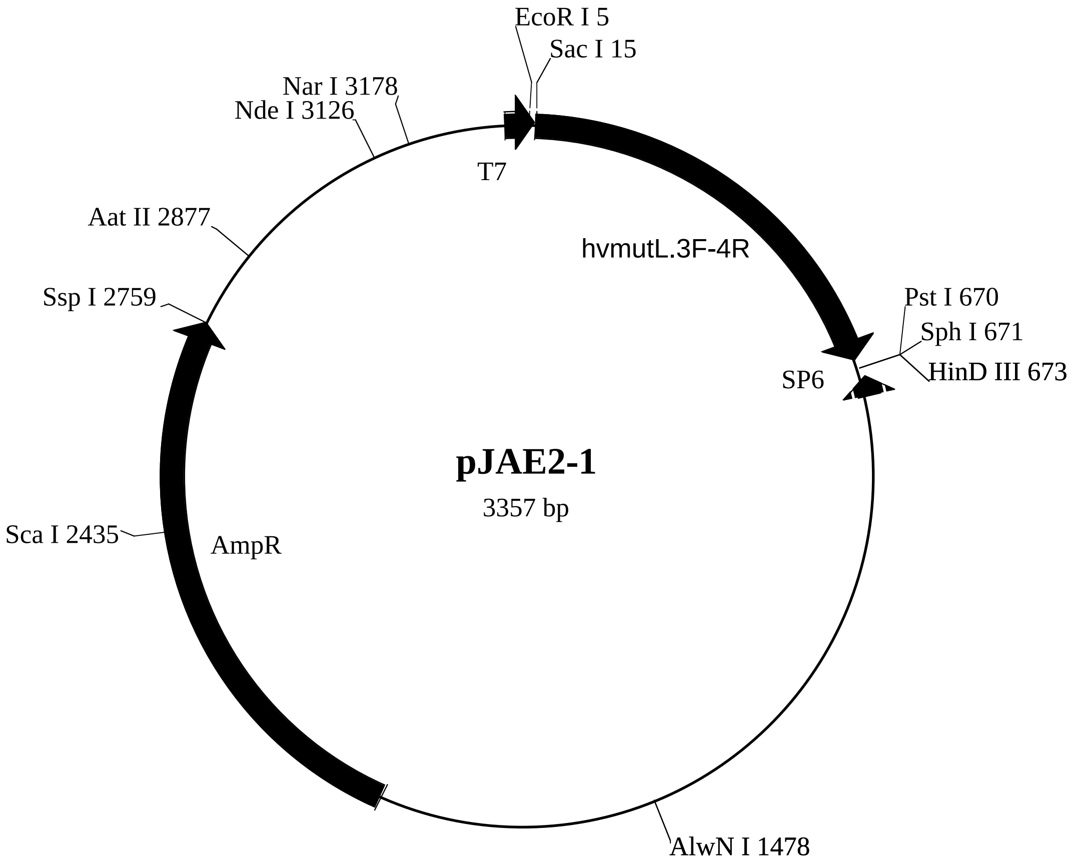

EcoR I 5
Sac I 15
Nar I 3178
Nde I 3126
T7
hvmutL.3F-4R
Aat II 2877
Ssp I 2759
Pst I 670
Sph I 671
HinD III 673
SP6
pJAE2-1
3357 bp
Sca I 2435
AmpR
AlwN I 1478

Figure 3. Partial sequence of *Haloferax volcanii mutL* gene and encoded protein.

DNA and predicted protein sequence of *Haloferax volcanii mutL* PCR product are shown. Sequence was determined from pJAE1-1 by Kurt Gish at DNAX.

```
          ttcgggggggaagcgCTCCACACCATCGGNGCGGTGTCGCGG
          f--g--g--e--a--L--H--T--I--G--A--V--S--R--

CTGACCATCCGGTCGAAGCCCCGCGGCGGCGACGTGGGCACCGAGTTGCAG
L--T--I--R--S--K--P--R--G--G--D--V--G--T--E--L--Q--

TACGAGGGCGGCGAGGTCGAGTCGATTCGACCCGCCGGCTGTCCCAAGGGG
Y--E--G--G--E--V--E--S--I--R--P--A--G--C--P--K--G--

ACGGTCGTCGAGGTCGACGACCTGTTTTACAACACGCCCGCCCGCCGGAAG
T--V--V--E--V--D--D--L--F--Y--N--T--P--A--R--R--K--

TTCCTCAAGACGACGGCGACCGAGTTCGACCACGTCAACGCGGTCGTCACG
F--L--K--T--T--A--T--E--F--D--H--V--N--A--V--V--T--

CACTACGCCCTCGCCAACCCGGACGTGGCCGTCTCGCTCGAACACGACGAC
H--Y--A--L--A--N--P--D--V--A--V--S--L--E--H--D--D--

CGCGAGGTGTTCGCCACCGAGGGCCGCGGCGACCTCCAGTCGACCGTGCTC
R--E--V--F--A--T--E--G--R--G--D--L--Q--S--T--V--L--

TCGGTGTACGSCCGCGAGGTCGCGGAGTCGATGGTCCCCGTGGACCACGAC
S--V--Y--R--E--V--A--E--S--M--V--P--V--D--H--D--A--

GCCCCCGGCGTCTCCGTCTCGGGGCTCGTGAGCCACCCCGAGACGACCCGG
P--G--V--S--V--S--G--L--V--S--H--P--E--T--T--R--S--

AGCACCCGCGACTACCTCTCGACGTTCGTCAACGACCGCTACGTCACCGAC
T--R--D--Y--L--S--T--F--V--N--D--R--Y--V--T--D--R--

CGCGTGCTCCGCGAGNCCGTCCTCGACGCCTACGGCGGCCAACTCGACGCG
V--L--R--E--X--V--L--D--A--Y--G--G--Q--L--D--A--D--

GACCGCTACCCCTTCGCGGTGCTGTTCGTCGAGGTCGCGCCCGACgccgtg
R--Y--P--F--A--V--L--F--V--E--V--A--P--D--A--v--d--

gatgtcaacgtcc
v--n--v--h--p
```

Figure 4. Blast search results of *Haloferax volcanii mutL* PCR product.

The DNA sequence of the *Haloferax volcanii mutL* PCR product was used for a blastx search of the NCBI non-redundant database. All high hits were to members of the MutL gene family indicating that this sequence encodes a MutL homolog.

```
                                                                   Score     E
Sequences producing significant alignments:                        (bits)  Value

sp|P14160|HEXB_STRPN  DNA MISMATCH REPAIR PROTEIN HEXB >gi|98033...   171  3e-42
sp|P49850|MUTL_BACSU  DNA MISMATCH REPAIR PROTEIN MUTL >gi|10025...   148  3e-35
sp|P44494|MUTL_HAEIN  DNA MISMATCH REPAIR PROTEIN MUTL >gi|10740...   135  3e-31
sp|P23367|MUTL_ECOLI  DNA MISMATCH REPAIR PROTEIN MUTL >gi|78712...   131  4e-30
sp|P14161|MUTL_SALTY  DNA MISMATCH REPAIR PROTEIN MUTL >gi|96808...   130  1e-29
gi|1575786  (U71053) DNA mismatch repair protein [Thermotoga mar...   116  1e-25
gi|1575784  (U71052) DNA mismatch repair protein [Aquifex pyroph...   106  1e-22
gi|2983934  (AE000746) DNA mismatch repair protein MutL [Aquifex...   105  3e-22
gnl|PID|e1298231  (Z92813) predicted using Genefinder; similar t...    97  9e-20
gi|466462  (U07418) human homolog of E. coli mutL gene product, ...    95  3e-19
gi|1724118  (U80054) mismatch repair protein [Rattus norvegicus]       95  5e-19
sp|P40692|MLH1_HUMAN  MUTL PROTEIN HOMOLOG 1 (DNA MISMATCH REPAI...    95  5e-19
gi|604369  (U17857) hMLH1 gene product [Homo sapiens]                  95  5e-19
gi|460627  (U07187) Mlh1p [Saccharomyces cerevisiae]                   94  6e-19
gi|3192877  (AF068257) mutL homolog [Drosophila melanogaster]          93  1e-18
sp|P38920|MLH1_YEAST  MUTL PROTEIN HOMOLOG 1 (DNA MISMATCH REPAI...    93  2e-18
gi|3329017  (AE001328) DNA Mismatch Repair [Chlamydia trachomatis]     92  2e-18
gi|2688099  (AE001131) DNA mismatch repair protein (mutL) [Borre...    84  7e-16
gi|3322578  (AE001210) DNA mismatch repair protein (mutL) [Trepo...    72  3e-12
gnl|PID|d1018113  (D90905) DNA mismatch repair protein MutL [Syn...    72  4e-12
sp|P54278|PMS2_HUMAN  PMS1 PROTEIN HOMOLOG 2 (DNA MISMATCH REPAI...    70  1e-11
gi|557470  (U14658) similar to S. cerevisiae PMS1 Swiss-Prot Acc...    70  1e-11
sp|P54279|PMS2_MOUSE  PMS1 PROTEIN HOMOLOG 2 (DNA MISMATCH REPAI...    68  6e-11
gi|1777768  (U50453) Hexb/MutL homolog [Thermus aquaticus]             66  2e-10
sp|P54280|PMS1_SCHPO  DNA MISMATCH REPAIR PROTEIN PMS1 >gi|12468...    62  4e-09
sp|P54277|PMS1_HUMAN  PMS1 PROTEIN HOMOLOG 1 (DNA MISMATCH REPAI...    61  1e-08
gi|3193291  (AF069298) Similar to DNA mismatch repair protein; T...    58  5e-08
gnl|PID|e1313318  (AL031135) putative protein [Arabidopsis thali...    50  1e-05
gi|172203  (M29688) DNA mismatch repair protein [Saccharomyces c...    48  5e-05
sp|P14242|PMS1_YEAST  DNA MISMATCH REPAIR PROTEIN PMS1 >gi|10770...    48  5e-05
gi|887629  (X89016) ORF N2317 [Saccharomyces cerevisiae]               44  8e-04
```

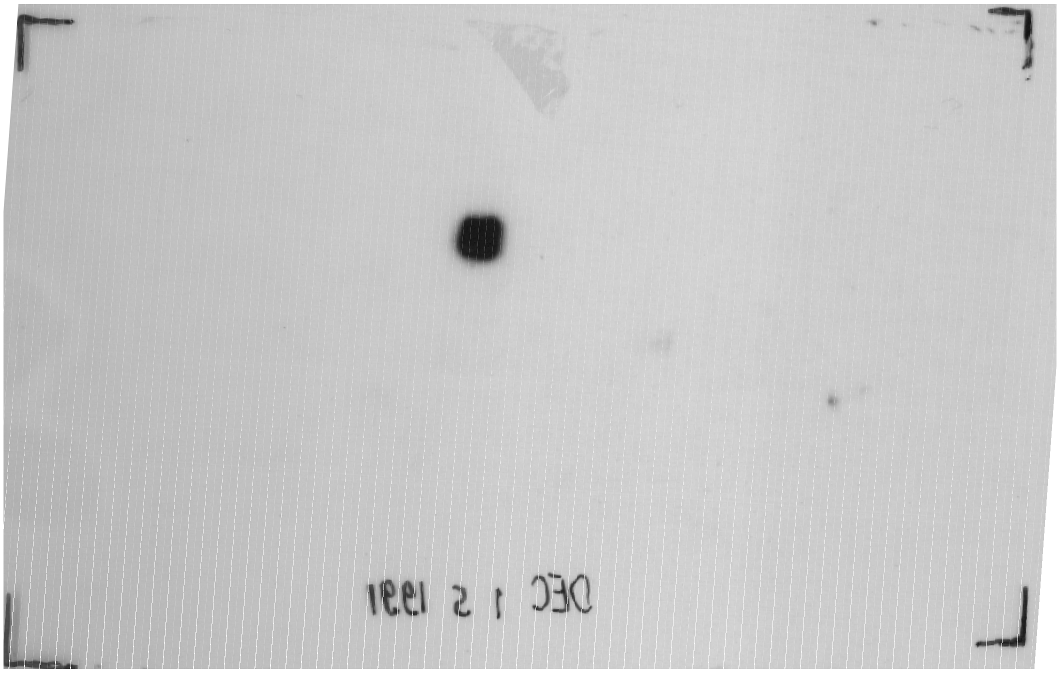
Figure 5. Alignment of amino-acid sequences of MutL homologs from bacteria and *H. volcanii.*

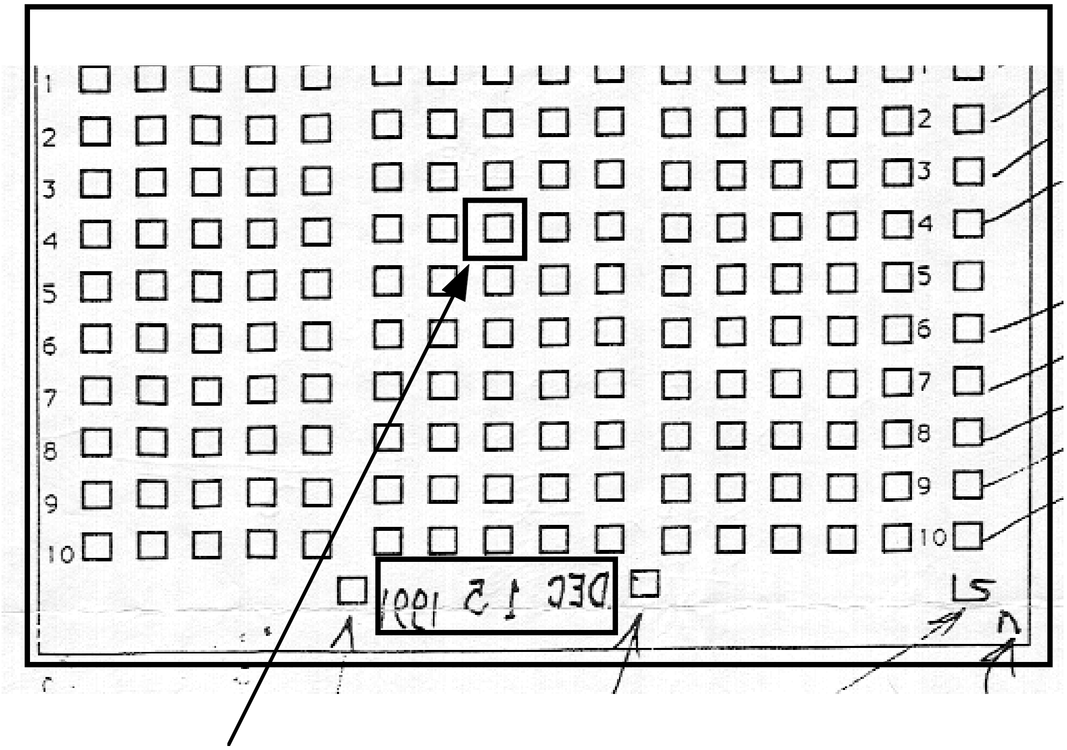
The predicted protein sequence of the *Haloferax volcanii mutL* PCR product was aligned to bacterial MutL homologs. Conserved amino-acids are shadowed.

```
H.VOLCANII                      LHTIG----A  VSRLTIRSKP  RGGD--VGTE  LQYEGGEVES
mutl_Bsubt         VRTLGFRGEA  LPSIA----S  VSHLEITTST  GEG---AGTK  LVLQGGNIIS
mutL_Ecoli         IISLGFRGEA  LASIS----S  VSRLTLTSRT  AEQQ--EAWQ  AYAEGRDMNV
mutl.neigo         VASMGFRGEG  LASIA----S  VSRLTLTSRQ  EDSS--HATQ  VKAEDGKLSS
mutL.Synsp         IKTLGFRGEA  LHSLA----Q  VARLTISSRS  VASPG-CGWR  ITYSPQGSPE
mutl.borbu         IETLGFRGEA  LSSIA----I  CSNISITSST  TSNE---SYQ  IEVENGIEKC
mutl.thermotoga    IRTYGFRGEA  LASIV----Q  VSRAKIVTKT  EKDA--LATQ  LMIAGGKVEE
mutl.deira         VTTLGFRGEA  LWAAA----Q  AGELELTTRP  AAQV--GAAR  XRAQGDAVEV
mutL.T_aquaticus   IATLGFRGQA  LYALR----Q  AATLKIRSRP  RGQV--GGGL  LLARGERVEL
mutl.Aquifex       VETYGFRGEA  LYSIS----S  VSKFRLRSRF  YQEK--EGRE  IEVEGGTLKS

H.VOLCANII         -IRPAGCPKG  TVVEVDDLFY  NTPAR-RKFL  ---KTTATEF  DHVNAVVTHY
mutl_Bsubt         -ESRSSSRKG  TEIVVSNLFF  NTPAR-LKYM  ---KTVHTEL  GNITDVVNRI
mutL_Ecoli         TVKPAAHPVG  TTLEVLDLFY  NTPAR-RKFL  ---RTEKTEF  NHIDEIIRRI
mutl.neigo         -PTAAAHPVG  TTIEAAELFF  NTPAR-RKFL  ---KSENTEY  AHCATMLERL
mutL.Synsp         QIEPVAIAMG  TRVEVRQLFA  NFPQR-RQAF  ---AKSQQFW  RPMVTYLQQL
mutl.borbu         -FKKQPAING  TIVDVTKIFH  NFPAR-KRFL  ---KQEPIET  KMCLKVLEEK
mutl.thermotoga    -ISETHRDTG  TTVEVRDLFF  NLPVR-RKSL  ---KSSAIEL  RMCREMFERF
mutl.deira         --SRTSAPAG  TTVTVSQLFA  RLPAR-LRTQ  ---ASAAAEV  RDITALLGRY
mutL.T_aquaticus   --RPAPAPPG  TRVEVLGLFA  GEGRD-----  ----P-KAEA  RGVLDLLKRY
mutl.Aquifex       -VRRVGMEVG  TEVEVYDLFF  NLPAR-KKFL  ---RKEDTER  RKITELVKEY

H.VOLCANII         ALANPDVAVS  LEHDDREVFA  TEGR------  --GDLQSTVL  SVYXREVAES
mutl_Bsubt         ALAHPEVSIR  LRHHGKNLLQ  TNGN------  --GDVRHVLA  AIYGTAVAKK
mutL_Ecoli         ALARFDVTIN  LSHNGKIVRQ  YRAVPEG---  --GQKERRLG  AICGTAFLEQ
mutl.neigo         ALAHPHIAFS  LKRDGKQVFK  LPAQ------  ---SLHERIA  AIVGDDFQTA
mutL.Synsp         ALCHPQVTWQ  LWQDERLRLS  LSPGPNPEAI  LLQCLKSLQA  GQLGYTQQSL
mutl.borbu         IITHPEINFE  INLNQKLRKI  YFKE------  ---SLIDRVQ  NVYGNVIENN
mutl.thermotoga    VLVRNDVDFV  FTSDGKIVHS  FPRT------  --QNIFERAL  LILEDLRKG-
mutl.deira         VLHHSALHWR  LTVDGDPRLT  HAPA------  ---DHRGAVA  TVYGPLSANR
mutL.T_aquaticus   LLHHPHLSLV  LFLEGEARLL  FPGA------  ---GLKEAAR  QAFGGLLAER
mutl.Aquifex       AITNPQVDFH  LFSEGKETLN  LKKK------  ---DLKGRIE  EIFESIFEE-

H.VOLCANII         MVPVDHDAP-  ----------  ----------  --GVSVSGLV  SHPETTR-S-
mutl_Bsubt         MLPLHVSS--  ----------  ----------  -LDFEVKGYI  ALPEITR-A-
mutL_Ecoli         ALAIEWQH--  ----------  ----------  -GDLTLRGWV  ADPNHTTPA-
mutl.neigo         SLEIDSGN--  ----------  ----------  -SALRLYGAI  AKPTFAK-G-
mutL.Synsp         SLPVDLEN--  ----------  ---------Q  ATSAQLSLTF  GYPDRCHRP-
mutl.borbu         KFRVLKKEH-  ----------  ----------  -DNIKIEIFL  APDNFSK-K-
mutl.thermotoga    YITFEEELS-  ----------  ----------  --GLRIKGIV  SSREVTR-S-
mutl.deira         VLTLDTPG--  ----------  ----------  -----VRGVV  SRPELTR-A-
mutL.T_aquaticus   LFPLEKGG--  ----------  ----------  --AFALEGLL  TGPQVSR--T
mutl.Aquifex       ----ESSE--  ----------  ----------  -----REGIK  VRAFISRNQ-

H.VOLCANII         ----TRDYLS  TFVNDRYVTD  R-VLREXVLD  AYGGQLDAD-  --RYPFAVLF
mutl_Bsubt         ----SRNYMS  SVVNGRYIKN  F-PLVKAVHE  GYHTLLPIG-  --RHPITFIE
mutL_Ecoli         ----LAEIQY  CYVNGRMMRD  R-LINHAIRQ  ACEDKLGAD-  --QQPAFVLY
mutl.neigo         ----KTDKQY  CFVNHRFVGD  K-VMLHAVKQ  AYRDVLHNA-  --LTPAFVLF
mutL.Synsp         ----RPDWLI  IAINGRPVNV  P-ELTQTILA  VFHRTLPRQ-  --RYPLCFAH
mutl.borbu         ----SKRHIK  TFVNRRPIDQ  K-DLLEAITN  GHSRILSPG-  --NFPICYLF
mutl.thermotoga    ----SRTGEY  FYVNGRFVVS  E-ELHEVLMK  VYDLPKR---  --SYPVAVLF
mutl.deira         ----RRDRMH  FAVNGRPIVA  PPELERAVID  AYAELLPAG-  --TAPLCVLD
mutL.T_aquaticus   ----RPDLLF  LAVNGRPVAL  PEGVLRAVRR  AYRELLPEG-  --HYPVGVLN
mutl.Aquifex       ----KRGKYY  LFVNSRPVYN  K-NLKEYLKK  TFGYKT----  -----IVVLF

H.VOLCANII         VEVAPDA
mutl_Bsubt         ITMDPILVDV  NVHPSKLEVR  L
mutL_Ecoli         LEIDPHQVDV  NVHPAKHEVR  F
mutl.neigo         LELPPEAVDV  NVHPTKTEIR  F
mutL.Synsp         WQLPPQCIDW  HRHPAKTEIY  L
mutl.borbu         LEINPEYIDF  NVHPQKKEVR  F
mutl.thermotoga    IEVNPEELDV  NIHPSKIVVK  F
mutl.deira         LTVAPEDYDP  NIHPAKQVVA  L
mutL.T_aquaticus   LSLPPGAYRL  RLDARKEEVA  L
mutl.Aquifex       IDIPPFLVDF  NVHPKKKEVK  F
```

Figure 6. Mapping of *Haloferax volcanii mutL* PCR product.

The *Haloferax volcanii mutL* was end-labeled and used to probe the *H. volcanii* ordered cosmid library (Cohen et al. 1992. Proc. Natl. Acad. Sci. U. S. A. 89: 1602-1606). Only one cosmid on the blot showed any hybridization signal - cosmid 455 - which maps to ~1450 on the map. Subsequent PCRs and Southern's using this cosmid alone confirm that this cosmid contains the PCR product.





Cos455

THE END