Theoretical Population Biology **61**, 481–487 (2002)
doi:10.1006/tpbi.2002.1594

TPB

Phylogenetic Analysis and Gene Functional Predictions: Phylogenomics in Action

Jonathan A. Eisen and Martin Wu

Department of Microbial Genomics, The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850
E-mail: jeisen@tigr.org,URL: http://www.tigr.org/~jeisen

Received January 15, 2002

Making accurate functional predictions for genes is a key step in this era of high throughput gene and genome sequencing. While most functional prediction methods are comparative in nature, many do not take advantage of the power that an evolutionary perspective provides to any comparative biology analysis. Here we review how evolutionary analysis can greatly benefit both homology-based and non-homology-based functional prediction methods. Examples that are discussed include phylogenetic determination of orthology, the use of character state reconstruction analysis of gene function, and evolutionary analysis of rates and patterns of gene evolution. © 2002 Elsevier Science (USA)

INTRODUCTION

One of the key steps in all genome sequencing projects is the prediction of function of genes present in that genome. While this process should be considered a form of educated guesswork (and thus one should never assume that the predicted function is correct), the predictions that are made are very important in interpreting the genome sequence data and in guiding future experimental work. Many computational methods have been developed to aid in the prediction of gene function. These methods can be divided into two main classes—the homology (or similarity) methods and the non-homology methods. In the homology/similarity methods, functional predictions are based on identifying and characterizing similarity in sequence or structure of the gene or its encoded product to genes of known function. The similarity that has been used includes that at different levels: motifs, domains, entire genes/proteins, and even possibly secondary or tertiary structure. In the non-homology methods (which have only come to the forefront recently), properties of a gene other than its similarity to other gene/products are used to aid in functional predictions. Properties that have been used include distance from origins of replication, distribution patterns across species, analysis of neighboring genes (Overbeek *et al*., 1999; Huynen *et al*., 2000a, b), domain patterns (Snel *et al*., 2000), and codon usage or nucleotide composition (Karlin, 2001; Karlin and Mrazek, 2001; Karlin *et al*., 2001). In the non-homology methods, genes in a genome or even across genomes are grouped by these properties and then the function of unknowns can be predicted if they group with genes with known functions.

The homology and non-homology functional prediction methods are both comparative in nature. That is, they rely on comparing genes and genomes within and between species. In theory, any comparative biological analyses such as this could simply focus on quantifying and characterizing the similarities and differences between species. However, for a deeper understanding of the biology being studied, it is helpful to go beyond this cataloguing of similarities and differences to try to understand how and even why those similarities and differences came to be. This is what is known as the evolutionary perspective to comparative biology and the

0040-5809/02 $35.00

basis for the now famous quote of Dobzhansky (1973), "Nothing in biology makes sense except in the light of evolution."

This "evolutionary perspective" has become an integral part of some biological fields, such as physiology and developmental biology (Harvey *et al*., 1996). However, the applications of this perspective have been slower to take hold in other fields, such as molecular biology and genomics. This is also true of efforts in functional prediction. Here we discuss how and why information on phylogeny can improve the functional predictions made with these methods. In addition, we suggest additional ways in which functional analysis in the future might take advantage of the benefits of an evolutionary perspective.

EVOLUTIONARY PERSPECTIVE AND HOMOLOGY METHODS

Similarity Is Not a Reliable Indicator of Evolutionary Relatedness or Function

A key step in all homology-based functional prediction methods is the determination of whether there is similarity of the gene of interest to characterized genes. This similarity can be measured at any or all of many levels: primary sequence of the DNA or protein, secondary structure, or three-dimensional structure (Bork and Koonin, 1998; Doerks *et al*., 1998). Of course, if no similarity is detected, or if similarity is detected only to genes with no known function, then no functional prediction can be made. If similarity is detected to a gene with known function, then the unknown gene can possibly be assigned the function of the known gene. While in principle this seems simple—there are many potential difficulties. First, one must determine if the similarity is biologically significant. In addition, there are many search result patterns that produce somewhat ambiguous results. The query gene can have significant matches to genes with known functions, but much better matches to genes with no known function. In addition, the query gene may have significant matches, all of equal or similar value, to genes with different functions. In these cases, it is necessary to choose which gene to use as the putative function of the query gene. While there are ways to make this choice based on the levels of similarity observed, they are all flawed since similarity itself is not a reliable indicator of evolutionary relatedness (Li, 1997). Therefore, what is needed are measures of relatedness not similarity (Eisen *et al*., 1997; Eisen, 1998b).

Distinguishing Orthologs from Paralogs

One important aspect of gene relatedness is the issue of orthology and paralogy (Fitch, 1970). Since functional divergence frequently accompanies gene duplication, determining whether the query gene is an ortholog of a gene with known function or a paralog can help in deciding whether to assign the unknown gene the function of the known gene. For example, if one determined that the gene of interest was similar to hemoglobin genes, it would be important to know whether it is an ortholog of any of the known hemoglobin. There have been many studies looking at individual gene families trying to determine orthology and paralogy as an aid in functional predictions (e.g., Eisen *et al*., 1995; Saier *et al*., 1999).

We give a few examples here to show the potential utility of such analyses. Analysis of the complete genome of *Deinococcus radiodurans* (White *et al*., 1999), the most radiation-resistant species known, revealed the presence of two homologs of the *uvrA* gene, a gene that in many other bacterial species is involved in nucleotide excision repair. At the time, *uvrA* homologs were found in all complete bacterial genomes and this was the first species found to encode two *uvrA* homologs (Eisen and Hanawalt, 1999).

uvrA homologs between species are very highly conserved and from the blast searches alone it was difficult to determine if either of the *D. radiodurans*' *uvrA* homologs was more similar to the characterized *uvrA*s than the other. However, phylogenetic analysis of the *uvrA* gene family revealed a clear distinction, one of *D. radiodurans* genes was most closely related to the normal *uvrA*s and the other was actually in a novel *uvrA* orthologous group which we called *uvrA*2. Interestingly, the other genes in this family were involved in antibiotic resistance rather than DNA repair. This fact, coupled with the fact that *uvrA* genes are members of the ABC transporter superfamily of proteins led us to suggest that the *uvrA*2s may be involved in transporting DNA damaging agents out of the cell. We have used similar analysis in many of our genome analysis papers, identifying a new RNA polymerase subfamily in *Arabidopsis thaliana* (AGI, 2000) (Fig. 1, available at http://www.nature.com/nature/journal/v408/n6814/suppinfo/408796a0.html) and unusual photolyase homologs in *Vibrio cholerae* (Heidelberg *et al*., 2000). While analysis of interesting genes can continue by detailed, curated efforts, for genome analysis in the future, with

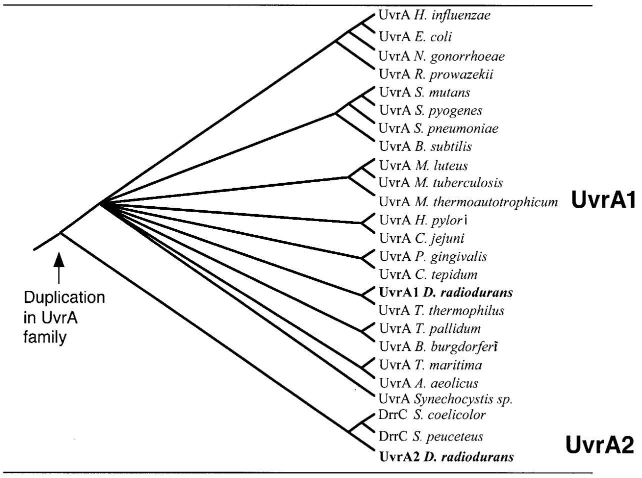

FIG. 1. Evolution of the *UvrA* family of proteins within the ABC transporter superfamily. (A) Phylogenetic tree of the ABC transporter superfamily showing the origin of the *UvrA* family. Representative genes of other families are also shown. (B) Phylogenetic tree of *UvrA* family showing two subfamilies (*UvrA*1 and *UvrA*2). *D. radiodurans* is so far the only species with genes in both subfamilies. The *UvrA*1 genes are all involved in nucleotide excision repair. The only *UvrA*2 gene with a known function is the *DrrC* gene of *Streptomyces peucetius* which is required for resistance to daunorubicin and may be involved in transport of this antibiotic from the cell. Trees were generated from sequence alignments using the neighbor-joining method and a PAM-based protein distance calculation.

more and more genomes being sequenced, automated methods of phylogenetic analysis are going to be more and more important.

The first attempt to develop a rapid method for determining orthology was the COG method (Tatusov *et al.*, 1997, 2001), for which there is an online resource at http://www.ncbi.nlm.nih.gov/COG/. This method uses a clustering algorithm to identify genes in genomes that are mutually more similar to each other than to any other genes in those genomes. The method then assumes that these are orthologs. While this method is very useful, it is not the ideal way to identify orthology since it still relies on pairwise similarity scores and not on phylogeny (Xie and Ding, 2000).

Another approach to allow rapid identification of orthologs has been to first divide gene families up into groups of orthologs and paralogs (or at least into distinct subfamilies), and to then build models that represent the sequence diversity in those groups of genes. Unknown genes can then be searched against these ortholog models and would be assigned to the group for which it is most similar to their model. Examples of this approach include the Panther system (Venter *et al.*, 2001), TIGRfams (Haft *et al.*, 2001), and some of the models in PFAM (Bateman *et al.*, 2002). These methods are perhaps one step better than COGs since they make use of explicit phylogenetic determination of subfamilies. However, since the placement of a gene into a group is still based on a similarity score (to the subfamily model), they are still potentially affected by evolutionary rate variation.

Since orthology is defined based on phylogeny, it makes sense that phylogeny would be the way to identify orthologs. However, it is only recently that methods have been developed to allow high throughput, accurate placement of query genes into phylogenetic groups (e.g., Sicheritz-Ponten and Andersson, 2001; Zmasek and Eddy, 2001). Thus manual phylogenetic analysis is still an important tool in determination or orthology.

Orthology Is Not Enough

While the identification of orthologs is helpful, it is still not a sufficient tool since orthologs sometimes have different functions. Because of this, we have proposed that an important tool in the prediction of gene function is a phylogenetically based reconstruction of functional evolution for a gene family (Eisen *et al.*, 1997; Eisen, 1998b). To do this, function can be treated as a character state and standard character state reconstruction methods can then be used to trace the changes in function over time. Ancestral states can then be inferred for the gene family and the function of genes with unknown function can be predicted from the tree. While initially we suggested that a parsimony-based character state reconstruction be used, a likelihood approach would probably be more accurate. A likelihood model could take into account information on the probabilities of functional change between and within orthologous groups (Gu, 2001a, b; Knudsen and Miyamoto, 2001).

EVOLUTIONARY PERSPECTIVE AND NON-HOMOLOGY METHODS

Orthology Determination Is Needed for Non-homology Methods

Non-homology methods attempt to group genes based on features other than sequence/structure that they share in common. The function of a gene of interest can possibly be predicted if it groups with genes with known functions. Even those methods claim to be non-homology based, they can be improved using analysis of homology. This is perhaps most clearly understood in relation to one of the non-homology methods—phylogenetic profiling (Pellegrini *et al.*, 1999). In phylogenetic profiling, genes are grouped based on their distribution patterns across species. The use of gene distribution patterns has been suggested as a useful tool in studies of evolution and protein function approach (e.g., Tatusov *et al.*, 1997; Eisen, 1998a, b, 1999). It was first described in detail by Gaasterland and Ragan (1998a, b; Ragan and Gaasterland, 1998) and has now become known as phylogenetic profiling (Pellegrini *et al.*, 1999). Since genes that function together in the same cellular process are frequently inherited or lost as a unit, their distribution patterns across species are often similar. Thus, the function of a query gene can possibly be predicted if its presence/absence pattern across species is the same as genes with known functions. In the initial phylogenetic profiling study (Pellegrini *et al.*, 1999), the profile for a gene was binary in nature. A gene was considered present in another genome if there was a match better than some threshold using a similarity search tool such as BLAST. While this profiling method is powerful, it is limited by the fact that evolutionary rates vary greatly among proteins. Therefore, it is impossible to select a versatile cutoff value to define presence or absence across species. Using normalized measures of sequence similarity (Marcotte *et al.*, 2000) can relieve this problem somewhat but still does not deal well with the issue of orthology and paralogy. We and others have previously suggested that the co-distribution of orthologs can be an important tool in functional predictions (Tatusov *et al.*, 1997; Eisen, 1998a; Eisen and Hanawalt, 1999). Thus, for a phylogenetic profile to be most useful, it should be able to distinguish orthologs from paralogs, and then genes could be grouped based on the presence and absence of orthologs. For example, we have found (Wu and Eisen, submitted) that orthology-based phylogenetic profiling is vastly superior to similarity-based profiling for ABC transporter complexes (made up of ATPases, permeases, and substrate-binding proteins). In the similarity-based profile analysis, the ATPases belonging to distinct ABC transporters tend to cluster together. This is not surprising since even paralogous ATPases are highly similar to each other. When orthology is used instead, the ATPases are resolved into separate groups, each containing components of the same ABC transporter (not shown). Other examples of orthology-based phylogenetic profiles are shown in Fig. 2. In this figure we show a comparison of the phylogenetic profile based clustering of *HisA* (Fig. 2A and B) and *PurK* (Fig. 2C and D) with an orthology-based profiling versus a homology-based profiling. In each case, the orthology-based analysis (using COGs as the measure of orthology) is vastly superior to homology-based analysis. In general, it is likely that all so-called non-homology methods can be improved by improved analysis of homology.

Phylogenetic Contrasts

The non-homology methods of functional prediction rely upon the identification or correlations among genes in properties other than sequence. Since these correlations are of biological data, they are biased by the phylogenetic background of the taxa being compared. Thus it is beneficial to remove the phylogenetic component of the correlation and study the residual correlations that remain (Felsenstein, 1985). Thus in the future, non-homology-based functional predictions would likely benefit from application of phylogenetic contrast methods.

RATES AND PATTERNS OF EVOLUTION

Since the function of a gene influences its rate and pattern of evolution, it is possible to use information on the evolutionary patterns of a gene to aid in functional predictions. For example, genes in pathogens that are antigenic are frequently under directional selection by the immune system to change their amino-acid sequences. Therefore, it is possible to identify genes in a genome that are more likely to be antigenic by examining synonymous and non-synonymous substitution patterns and identifying genes with excessively high ratios of non-synonymous:synonymous changes (e.g., Black and Coppel, 2000). Similarly, it is possible to identify genes that are likely under strong purifying selection (the non-synonymous:synonymous ratio should be low). Such analysis requires multiple genome sequences of closely related strains and can be quite powerful. Analyzing the patterns of evolution of a gene (compared to a closely related homolog/ortholog) is particularly helpful for genes for which none of the homologs have a known function. While this does not help assign specific functions to genes, it can help identify groups of genes that have different functional and structural constraints. Phylogenetic analysis can be helpful in these cases because it can allow improved determination of the rates of non-synonymous and synonymous changes (Muse and Gaut, 1994; Yang and Nielsen, 1998) and also allow the determination of the actual direction of sequence changes rather than just the

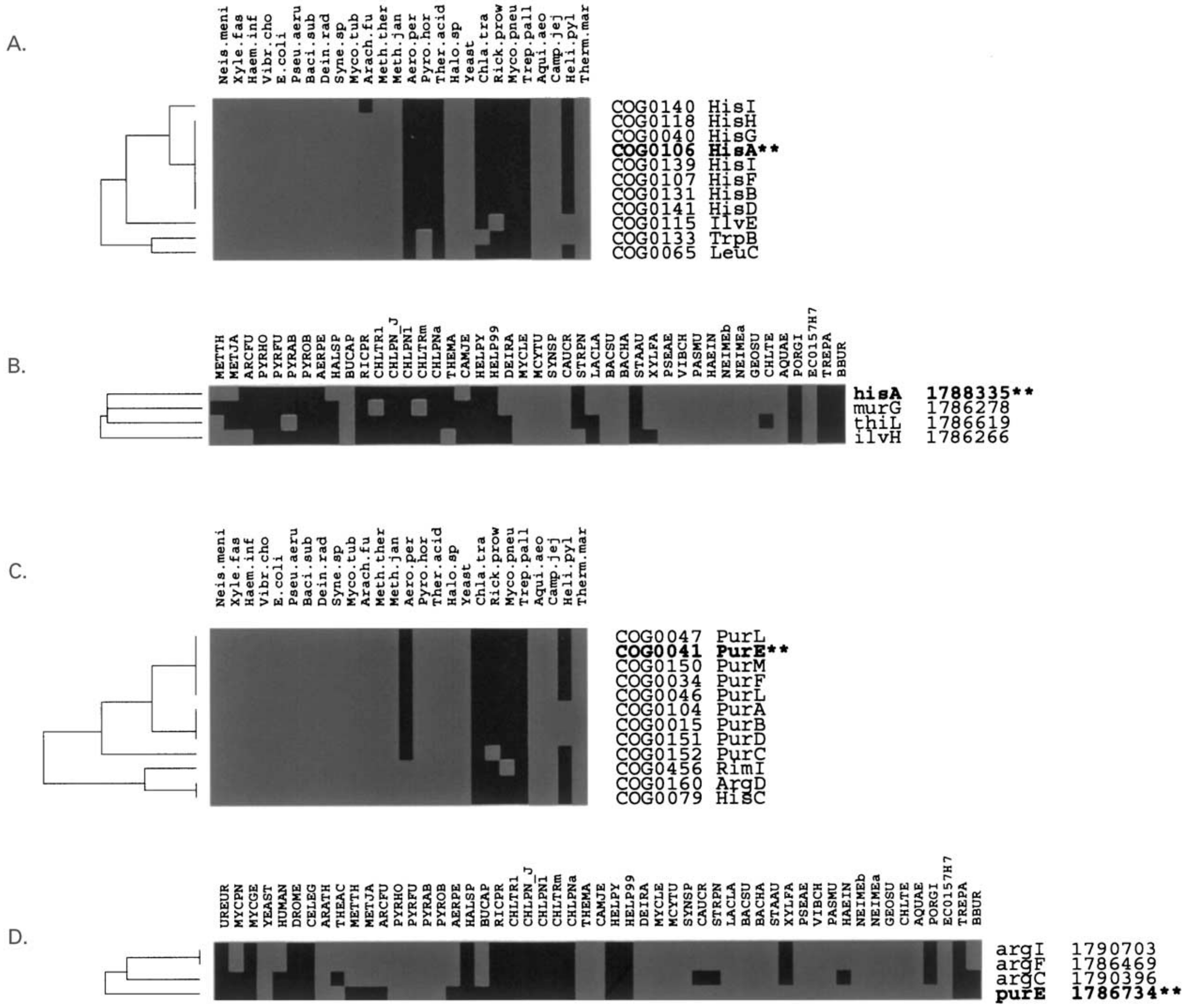

FIG. 2. Phylogenetic profiling using orthology versus homology. Phylogenetic profiles were generated from the COGs dataset of putative orthologs (A and C) or from homology search results for the *E. coli* K12 genome (B and D). Rows correspond to gene or COG presence/absence (Black = present) and columns correspond to different complete genomes. Profiles were clustered using single linkage clustering of the CLUSTER program and viewed using TREEVIEW (both available at http://rana.lbl.gov). Genes were considered present in a species if they were identified in the COGs dataset (http://www.ncbi.nlm.nih.gov/COG) or if there was a blast match with an *E*-value of better than 10^{-15} to an *E. coli* protein. Comparison shows that the orthology-based analysis (A and C) groups the reference genes (*hisA* and *purE*) with other genes in their pathways better than the homolog method even though fewer species were used for the COG analysis.

differences between genomes. Phylogenetic analysis would also be needed to distinguish orthologs from paralogs for this type of analysis.

CONCLUSIONS

Most methods of making functional predictions involve some type of comparative biology. Thus, just as in other fields of biology, functional predictions can be improved through evolutionary analysis. We believe that this is true for most areas of genome analysis. Similarly, there are some evolutionary studies that cannot be done without complete genome sequences (e.g., analysis of genome structure or gene loss). In some cases, the genome analysis and evolutionary reconstructions have feedback loops between them such that they cannot be conducted separately. For this reason, we believe evolutionary and genome analysis should be combined into a single composite approach, which we refer to as phylogenomics (Eisen and Hanawalt, 1999). While such a composite approach can certainly benefit functional predictions, we believe it will also prove very useful to many other areas of genomics and evolutionary biology.

ACKNOWLEDGMENTS

This paper was supported in part by NIH Grant U01-AI47409 and NSF Grant MCB-0084211.

REFERENCES

The Arabidopsis Genome Initiative (AGI). 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. The Arabidopsis Genome Initiative, *Nature* **408**, 796–815.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall M., and Sonnhammer, E. L. 2002. The Pfam protein families database, *Nucleic Acids Res*. **30**, 276–280.

Black, C. G., and Coppel, R. L. 2000. Synonymous and non-synonymous mutations in a region of the *Plasmodium chabaudi* genome and evidence for selection acting on a malaria vaccine candidate, *Mol. Biochem. Parasitol*. **111**, 447–451.

Bork, P., and Koonin, E. V. 1998. Predicting functions from protein sequences—where are the bottlenecks?, *Nat. Genet*. **18**, 313–318.

Dobzhansky, T. 1973. Nothing in biology makes sense except in the light of evolution, *Am. Biol. Teacher* **35**, 125–129.

Doerks, T., Bairoch, A., and Bork, P. 1998. Protein annotation: Detective work for function prediction, *Trends Genet*. **14**, 248–250.

Eisen, J. A. 1998a. A phylogenomic study of the MutS family of proteins, *Nucleic Acids Res*. **26**, 4291–4300.

Eisen, J. A. 1998b. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis, *Genome Res*. **8**, 163–167.

Eisen, J. A. 1999. "Evolution of DNA Repair Genes, Proteins, and Processes," Stanford University, Stanford, CA.

Eisen, J. A., and Hanawalt, P. C. 1999. A phylogenomic study of DNA repair genes, proteins, and processes, *Mutat. Res*. **435**, 171–213.

Eisen, J. A., Kaiser, D., and Myers, R. M. 1997. Gastrogenomic delights: A movable feast, *Nature* (*Medicine*) **3**, 1076–1078.

Eisen, J. A., Sweder, K. S., and Hanawalt, P. C. 1995. Evolution of the SNF2 family of proteins: Subfamilies with distinct sequences and functions, *Nucleic Acids Res*. **23**, 2715–2723.

Felsenstein, J. 1985. Phylogenies and the comparative method, *Am. Nat*. **125**, 1–15.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins, *Syst. Zool*. **19**, 99–113.

Gaasterland, T., and Ragan, M. A. 1998a. Constructing multigenome views of whole microbial genomes, *Microb. Comp. Genomics* **3**, 177–192.

Gaasterland, T., and Ragan, M. A. 1998b. Microbial genescapes: Phyletic and functional patterns of ORF distribution among prokaryotes, *Microb. Comp. Genomics* **3**, 199–217.

Gu, X. 2001a. Mathematical modeling for functional divergence after gene duplication, *J. Comput. Biol*. **8**, 221–234.

Gu, X. 2001b. Maximum-likelihood approach for gene family evolution under functional divergence, *Mol. Biol. Evol*. **18**, 453–464.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, L. T., and White, O. 2001. TIGRFAMs: A protein family resource for the functional identification of proteins, *Nucleic Acids Res*. **29**, 41–43.

Harvey, P. H., Leigh Brown, A. J., Marynard Smith, J., and Nee, S. Eds. 1996. "New Uses for New Phylogenies," Oxford University Press, Oxford.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L. *et al*. 2000. The genome sequence of *Vibrio cholerae*, the etiologic agent of cholera, *Nature* **406**, 477–484.

Huynen, M., Snel, B., Lathe 3rd, W., and Bork, P. 2000a. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences, *Genome Res*. **10**, 1204–1210.

Huynen, M., Snel, B., Lathe, W, and Bork, P. 2000b. Exploitation of gene context, *Curr. Opin. Struct. Biol*. **10**, 366–370.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes, *Trends Microbiol*. **9**, 335–343.

Karlin, S., and Mrazek, J. 2001. Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage, *Proc. Natl. Acad. Sci. USA* **98**, 5240–5245.

Karlin, S., Mrazek, J. Campbell, A., and Kaiser, D. 2001. Characterizations of highly expressed genes of four fast-growing bacteria, *J. Bacteriol*. **183**, 5025–5040.

Knudsen, B., and Miyamoto M. M. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins, *Proc. Natl. Acad. Sci. USA* **98**, 14,512–14,517.

Li, W. H. 1997. "Molecular Evolution," Sinauer Associates, Inc., Sunderland, MA.

Marcotte, E. M., Xenarios, I., van Der Bliek, A. M., and Eisenberg, D. 2000. Localizing proteins in the cell from their phylogenetic profiles, *Proc. Natl. Acad. Sci. USA* **97**, 12,115–12,120.

Muse, S. V., and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome, *Mol. Biol. Evol*. **11**, 715–724.

Overbeek, R., Fonstein, M., D'Souza, M. Pusch, G. D., and Maltsev, N. 1999. The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288.

Ragan, M. A., and Gaasterland, T. 1998. Microbial genescapes: A prokaryotic view of the yeast genome, *Microb. Comp. Genomics* **3**, 219–235.

Saier Jr., M. H., Eng, B. H., Fard, S., Garg, J., Haggerty, D. A., Hutchinson, W. J., Jack, D. L., Lal, E. C., Liu, H. J., Nusinew, D. P., *et al*. 1999. Phylogenetic characterization of novel transport protein families revealed by genome analyses, *Biochim. Biophys. Acta*. **1422**, 1–56.

Sicheritz-Ponten, T., and Andersson S. G. 2001. A phylogenomic approach to microbial evolution, *Nucleic Acids Res*. **29**, 545–552.

Snel, B., Bork, P., and Huynen, M. 2000. Genome evolution. Gene fusion versus gene fission, *Trends Genet*. **16**, 9–11.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. 1997. A genomic perspective on protein families, *Science* **278**, 631–637.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res*. **29**, 22–28.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. 2001. The sequence of the human genome, *Science* **291**, 1304–1351.

White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., *et al*. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1, *Science* **286**, 1571–1577.

Xie, T., and Ding, D. 2000. Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale, *Gene* **261**, 305–310.

Yang, Z., and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals, *J. Mol. Evol*. **46**, 409–418.

Zmasek, C. M., and Eddy, S. R., 2001. ATV: Display and manipulation of annotated phylogenetic trees, *Bioinformatics* **17**, 383–384.