A6

PHYLOGENETIC AND PHYLOGENOMIC APPROACHES TO STUDIES OF MICROBIAL COMMUNITIES[†]

Jonathan A. ***Eisen***[25]

Note: this paper is based on a transcript of a talk given at the IOM Forum on Microbial Threats in March, 2012. Only minor modifications have been made (e.g., additional of section headers, addition of references, removal of side comments) in order to as accurately as possible reflect the presentation. A recording of the talk with slides is available on YouTube at http://www.youtube.com/watch?v=ddGyEExi-FI&feature=share&list=PL3E32A3B8B2642F62. Because my presentation was in essence a review of my work in the area, this should not be viewed as a review of the field but rather of my work in this area.

Acknowledgements and Introduction

Thank you. I guess I have the awkward after-lunch talk here, so I will try not to use the most complicated slide I used, although I am not so sure about that.

[25] University of California, Davis, California.

[†] Adapted from remarks presented at the IOM Forum on Microbial Threats March 2012 Workshop, The Social Biology of Microbial Communities.

Since I frequently don't get to my last slide, I just want to do like some other people have been doing, acknowledgements at the beginning. And what I am going to talk about is work funded by a lot of different agencies that has gone on in my lab for about 10, 15 years, including in particular work funded by the Department of Energy, the National Science Foundation, the Gordon and Betty Moore Foundation, and recently Homeland Security, all related to phylogenetic analysis of genomes and metagenomes. And there are a lot of people I will mention, many of the people involved in this. But this is the trans-disciplinary type of work. It hurts my head a lot of time to think of all the people involved in some of these projects, but I will try to acknowledge as many of them as possible.

So what I am going to do is give just a quick introduction to phylogeny and then talk about three examples of the uses of phylogeny in studying microbial communities via DNA sequencing—phylotyping, functional prediction (just a tiny bit, because I want to raise the point as [the topic] has come up a few times [at this meeting], and then selection of organisms for study. And then I will end with just a couple of things about future directions.

What Is Phylogeny?

I assume most people here know what phylogeny is, but just [a quick reminder]: phylogeny is a representation of the history of entities, and that could be the history of genes, the history of genomes, the history of species. And in many cases, people have represented this history by a bifurcating tree-like structure. Phylogeny doesn't have to be represented as a bifurcating tree-like structure. We can have reticulation events, like recombination and lateral gene transfer. I include all those complexities within the concepts of phylogeny, so I am not trying to discriminate between vertical evolution versus lateral evolution, but really this sort of representation of the history of organisms. I am also not going to get into the debates about what that exact history is. People are still trying to resolve the evolutionary history of microbes as well as other organisms, and it is a constant area of research.

Whatever your belief of the latest model is, in my opinion if you incorporate phylogenetic approaches in your analysis of genome and metagenome and other data, it can improve what you are doing relative to not trying to incorporate phylogenetic approaches. And what I am going to do is try and walk you through a couple of examples of this.

Example I: Phylotyping

The first one I want to talk about is phylotyping, which we have heard either directly or indirectly a lot about at this meeting. Phylotyping, I was exposed to as a young, budding scientist in the lab of Colleen Cavanaugh. I was an undergraduate at Harvard and ended up in Colleen's lab, and I spent a year and a

half sequencing one 16s ribosomal RNA gene. But I got a paper out of that one 16s ribosomal RNA gene (Eisen et al., 1992). And the point of sequencing that 16s ribosomal RNA gene, as well as the point, even today in many cases, of ribosomal RNA sequencing, is to try and figure out what the organism is related to where that 16s came from.

And the way phylotyping works, this is basically developed by Norm Pace and colleagues (e.g., Hugenholtz et al., 1988). You collect DNA from your sample, you clone out some sequence like ribosomal RNA, you build an evolutionary tree of that sequence. So this is where the phylo part comes in *phylo*typing, a phylogenetic tree of that sequence. And you compare your unknowns to known things that are out there. And this is the tree from our *Solemya velum* chemosynthetic symbiont 16s, which by the way was accepted 20 years ago tomorrow, I think, my first scientific paper (see Figure A6-1).

Ribosomal RNA phylotyping has been amazing at revolutionizing our understanding of microbes in the world. I assume most people here appreciate the vast diversity of things that have been discovered by using phylogenetic trees of ribosomal RNAs from the environment to understand what the organisms are that those ribosomal RNAs came from. I am not going to go into the whole history of this. What I really want to talk about is three challenges that now come up with phylogenetic typing that largely relate to this issue of the cost of sequencing dropping and dropping and dropping and getting easier and easier and easier, accelerating at a rate faster than Moore's Law.

And with new sequencing machines being announced every 2 or 3 days, not that all of them work, but there are all sorts of cool things coming out there. And so this affects things like PCR amplification of ribosomal RNA sequences. We now have literally trillions of ribosomal RNA sequences to analyze, as opposed to that one that I got a paper out of. It also is really important in terms of revolutionizing metagenomic approaches. And I appreciate what Jo [Handelsman] was talking about, with metagenomics is not everything about a community, but the cheaper sequencing is, the more data we are going to have for metagenomics. Even though it doesn't tell us everything we need to analyze that data. And a third challenge is that most of the DNA sequencing technologies that people have been using generate short sequence reads, as opposed to long contigs that are easier to analyze.

And so it is sort of obvious that when metagenomic data was generated, you could scan through the metagenomic data to build evolutionary trees of sequences that were in that data. And this is what I did with Craig Venter in the analysis of the Sargasso Sea data (Venter et al., 2004). You can scan through the metagenomes, find ribosomal RNA sequences, and build evolutionary trees of those ribosomal RNAs just like we did with PCR amplified data (Figure A6-2).

The great thing about metagenomic data is, we can build phylogenetic trees of other genes that are good phylogenetic markers that we never could really get a good sample of most of these because PCR amplification of protein coding genes

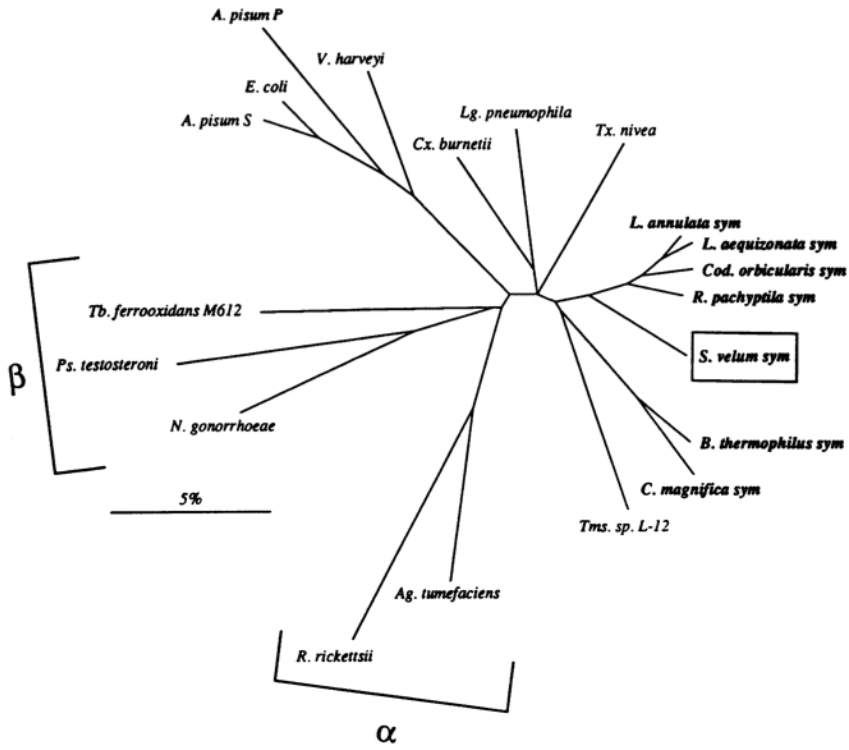


FIGURE A6-1 Unrooted phylogenetic tree showing the position of the *S. velum* symbionts in relation to that of other Proteobacteria species on the basis of 16s rRNA gene sequences. The tree was constructed from evolutionary distances in Table 1. Members of the alpha and beta subclasses of the Proteobacteria are bracketed; all others are of the gamma subclass. Chemoautotrophic symbionts (sym) are listed in boldface type. Full species names listed in Table A6-1. Scale bar represents percent similarity.
SOURCE: Eisen et al. (1992).

across broad diversity does not work very well. So now with metagenomic data we can look at protein coding genes and compare and contrast the results with those to the results with ribosomal RNA. I have been obsessed with the RecA gene for a long time (e.g., Eisen, 1995[26]), so I always end up working on RecA (Figure A6-3).

But there are lots of others genes that you can analyze, and we did this in the Sargasso Sea analysis. And if you compare and contrast the results that you

[26] 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. *Journal of Molecular Evolution* 41(6):1105-23.

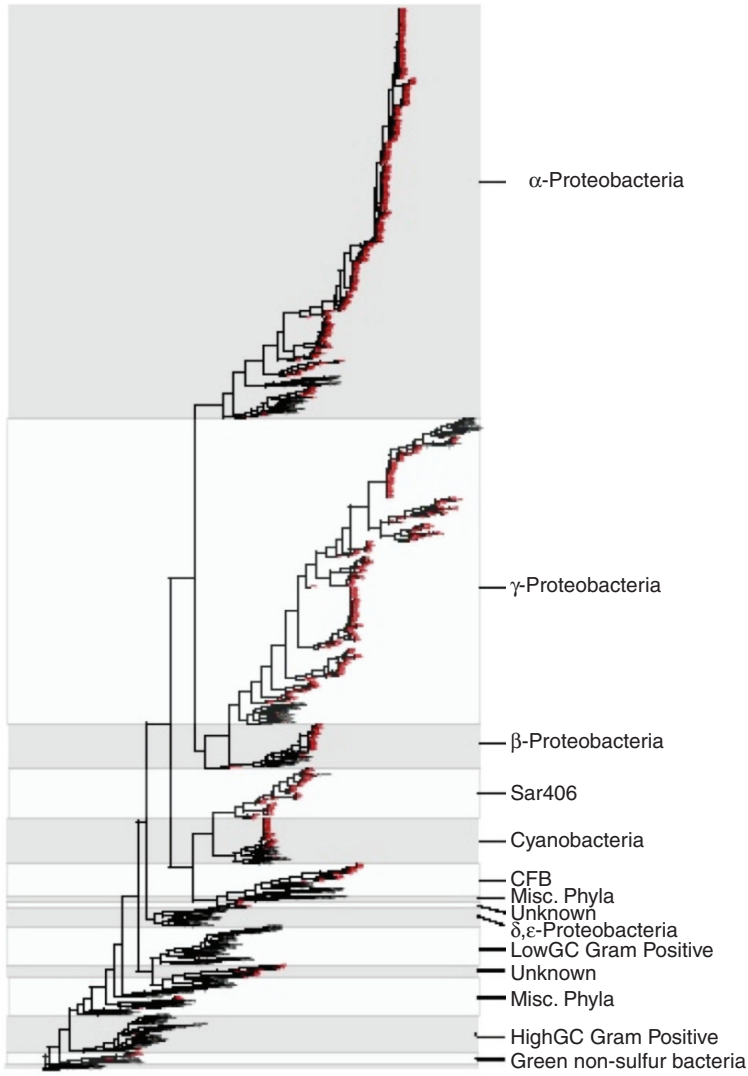


FIGURE A6-2 rRNA tree. Phylogenetic tree of 16s rRNA. Phylogenetic trees are shown for this gene, with sequences from this study colored red, and with major phylogenetic groups outlined (clades of sequences that could not be assigned to any group are labeled as "Unknown"). Only the bacterial portions of the tree are shown. The phylogenetic tree of rRNAs was generated by (1) aligning each Sargasso Sea rRNA of greater than 400 bp against its closest match in the alignments from the Ribosomal Database Project – II (RDP II) database and then using that alignment to align the new sequence to the complete RDP database; (2) a phylogenetic tree was generated using the dnapars algorithm of the Phylip package in which all new Sargasso sequences were included as were all sequences from complete genomes and sequences from representatives of major phylogenetic groups. Only complete genomes were used for comparison so that each tree can be compared to the others without differences in species sampling complicating the comparison.
SOURCE: Venter et al. (2004).

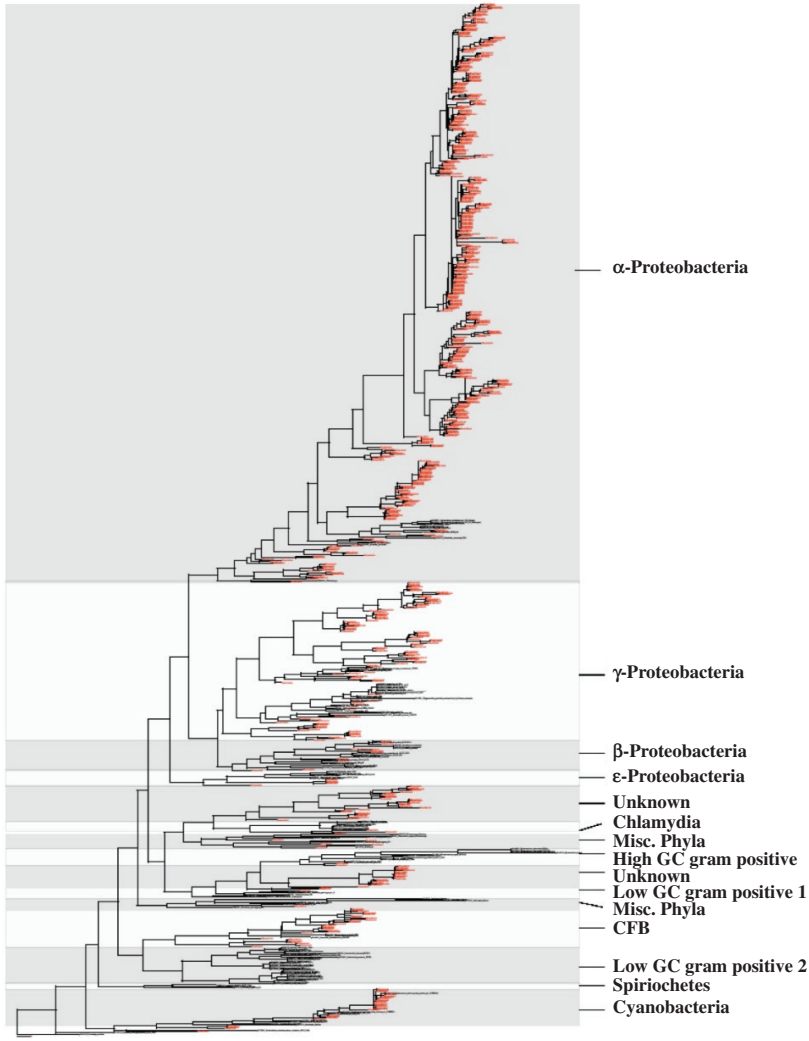


FIGURE A6-3 RecA tree. Phylogenetic tree of 16s rRNA. Phylogenetic trees are shown for this gene, with sequences from this study colored Red, and with major phylogenetic groups outlined (clades of sequences that could not be assigned to any group are labeled as "Unknown"). Only the bacterial portions of the tree are shown. The phylogenetic tree was generated in the following way: (1) homologs of each protein were identified in the Sargasso predicted protein set and in complete genome sequences using blastp and Hidden Markov Model (HMM) searches; (2) distant paralogs of each protein were excluded using a reciprocal-top match filter; (3) all sequences were aligned to each other using the HMM as a template; (4) poorly aligned regions were identified and removed using a conservation-score based filter; (5) all sequences that did not have >50% overlap with the *E. coli* ortholog were excluded; and (6) phylogenetic trees were generated using the protein parsimony algorithm in Phylip (parsimony was used to better deal with the limited overlap between many pairs of sequences). Only complete genomes were used for comparison so that each tree can be compared to the others without differences in species sampling complicating the comparison. SOURCE: Venter et al. (2004).

get with phylogenetic typing of protein coding genes and assigning those types into phyla, into bins that correspond to the phyla of organisms, you see some interesting patterns (Figure A6-4).

There are some differences between what you get with ribosomal RNA and what you get with protein coding genes. I think a lot of this is due to the differences in copy number. So if you estimate relative abundance of organisms from ribosomal RNA, the copy number of ribosomal RNA varies a lot between taxa, but the copy number of many protein coding genes does not vary a lot between taxa. So the protein coding genes, even though they are not as richly sampled, are

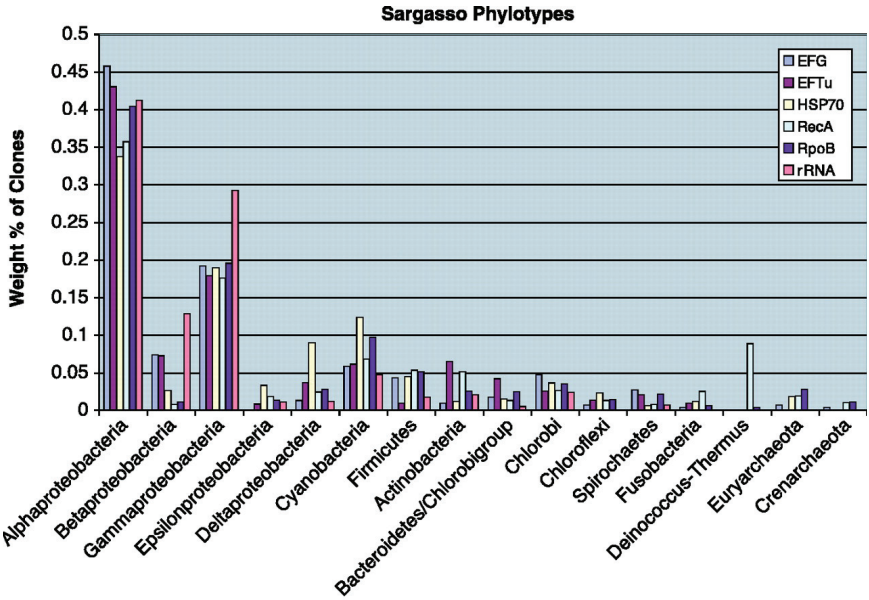


FIGURE A6-4 Phylogenetic diversity of Sargasso Sea sequences using multiple phylogenetic markers. The relative contribution of organisms from different major phylogenetic groups (phylotypes) was measured using multiple phylogenetic markers that have been used previously in phylogenetic studies of prokaryotes: 16S rRNA, RecA, EF-Tu, EF-G, HSP70, and RNA polymerase B (RpoB). The relative proportion of different phylotypes for each sequence (weighted by the depth of coverage of the contigs from which those sequences came) is shown. The phylotype distribution was determined as follows: (1) Sequences in the Sargasso data set corresponding to each of these genes were identified using HMM and BLAST searches. (2) Phylogenetic analysis was performed for each phylogenetic marker identified in the Sargasso data separately compared with all members of that gene family in all complete genome sequences (only complete genomes were used to control for the differential sampling of these markers in GenBank). (3) The phylogenetic affinity of each sequence was assigned based on the classification of the nearest neighbor in the phylogenetic tree.
SOURCE: Venter et al. (2004).

probably better markers for estimating relative abundance than ribosomal RNA sequences. People have been doing this now with metagenomic data in many different contexts.

Need for Automation

I am not going to cover all of the phylogenetic approaches to metagenomic data. But what I want to talk about is this issue of automation. I think as we get more and more sequence data, we can't look at trees any more. We can't look at sequence alignments. We can't even handle all the data at all. But we certainly need to automate everything.

There are multiple strategies to trying to automate phylogenetic typing of ribosomal RNA or metagenomic data. And one of them has been to use the BLAST program (Altschul et al., 1990), or analogs of the BLAST program, which basically looks at sequence similarity of your sequence to sequences in databases. This is not the best approach to analyzing data. Percent similarity or other measures of similarity are not a good indicator of evolutionary relatedness and can produce misleading patterns about the taxonomy and other parts of information that you want to analyze (Eisen, 1998). There are also approaches that look at compositional and word frequencies. Now both of these approaches are very fast, so you can generate a lot of results very rapidly, and that can be an advantage in many cases. But phylogenetic analysis is generally better than most of these approaches, and the challenge is, how do you implement phylogenetic analysis on a massive scale?

And so what I am going to do is just give you sort of four examples of some of the issues related to implementing automated phylogenetic analysis on a large scale.

Method 1: Each Sequence Is an Island

You can scan through the data and say I am going to take each individual sequence, each individual new thing that I get, and build an evolutionary tree of it relative to known things. So in essence, each sequence is an island in and of itself. So we have done this with a variety of tools. We built our automated ribosomal RNA tool called STAP (Wu et al., 2008), which goes through and takes a reference alignment of known ribosomal RNAs and then for each new sequence aligns your new sequence to that and builds an evolutionary tree in a completely automated manner, and then can scan through the tree to look at the taxonomy results from the tree (Figure A6-5).

We also built a tool that will do this with protein coating genes, called AMPHORA (Wu and Eisen, 2008). So it can automatically scan through metagenomic data, find homologues of particular protein families, build an alignment of them, build an evolutionary tree of them (Figure A6-6). And if you have a good reference alignment from known organisms, you can identify a

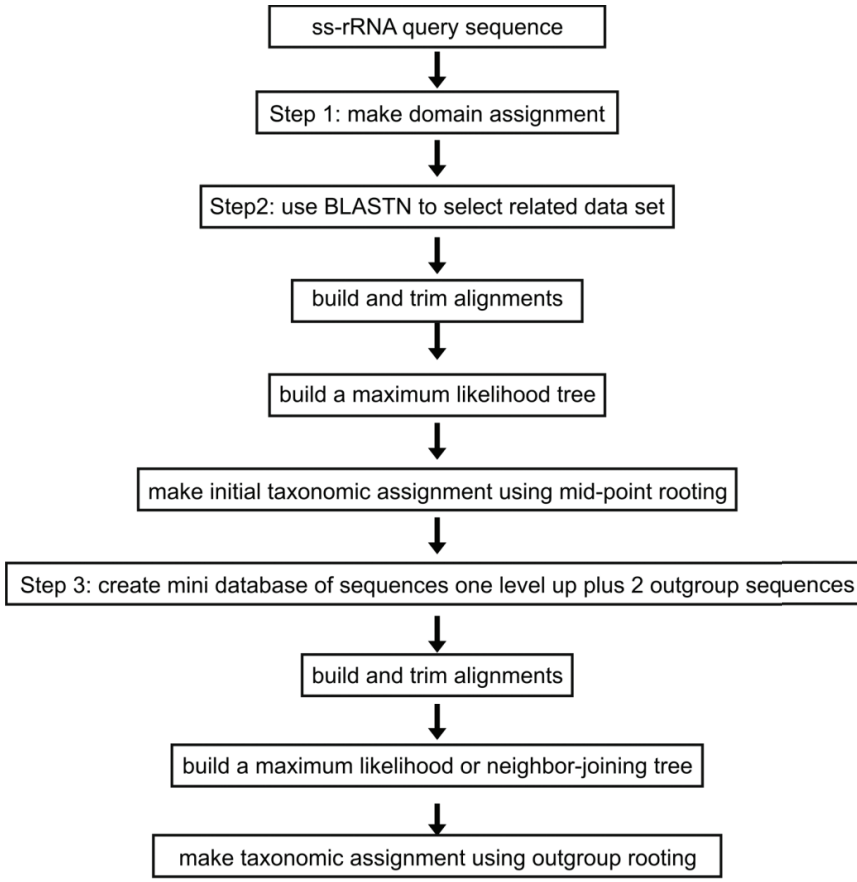


FIGURE A6-5 A flow chart of the STAP pipeline.
SOURCE: Wu et al. (2008).

candidate—sort of identify what those protein coating genes come from in your environmental sample (Figure A6-7).

Now having a good reference database is challenging for protein coating sequences, whereas we have now trillions of ribosomal RNA sequences and tens of thousands of complete ribosomal RNA sequences. We don't have good databases of protein coating sequences. All of the good data are now coming from genome sequencing projects. So whatever has been sequenced in terms of genomes is basically our source of protein coating genes for building these evolutionary trees. And so you build a reference tree from the genomes, you take your new data, stream them against the reference tree, build a new tree with that, and assign your new sequence to somewhere compared to the reference

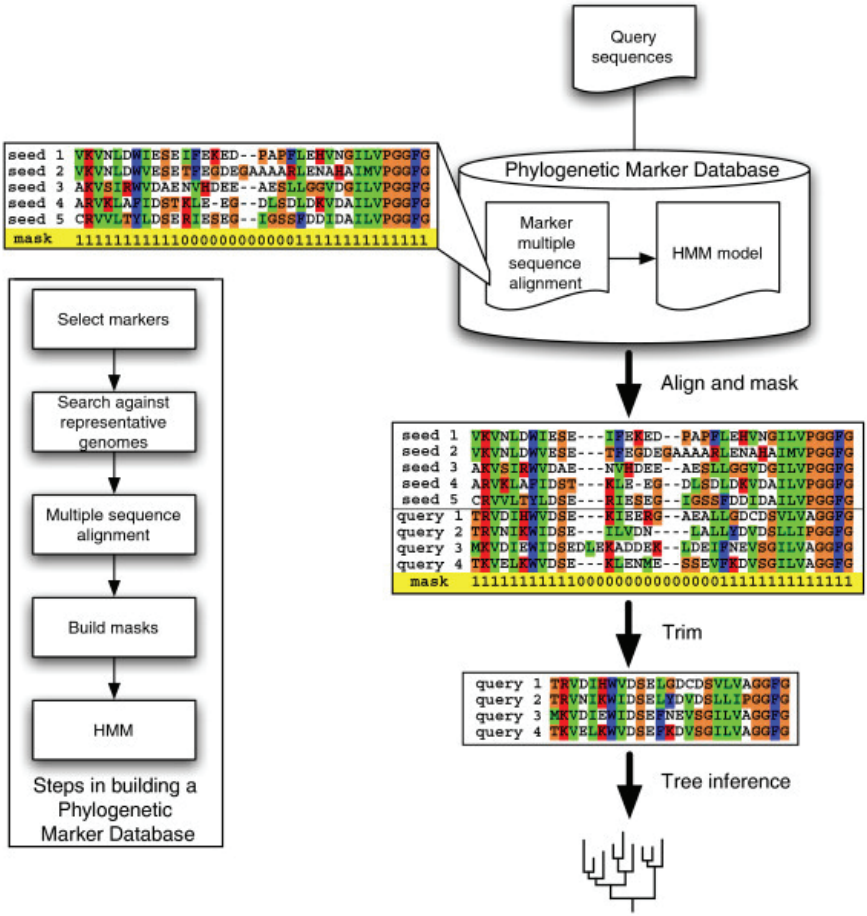


FIGURE A6-6 A flow chart illustrating the major components of AMPHORA. The marker protein sequences from representative genomes are retrieved, aligned, and masked. Profile hidden Markov models (HMMs) are then built from those "seed" alignments. New sequences of interest are rapidly and accurately aligned to the trusted seed alignments through HMMs. Predefined masks embedded within the "seed" alignment are then applied to trim off regions of ambiguity before phylogenetic inference. Alignment columns marked with "1" or "0" were included or excluded, respectively, during further phylogenetic analysis.
SOURCE: Wu and Eisen (2008).

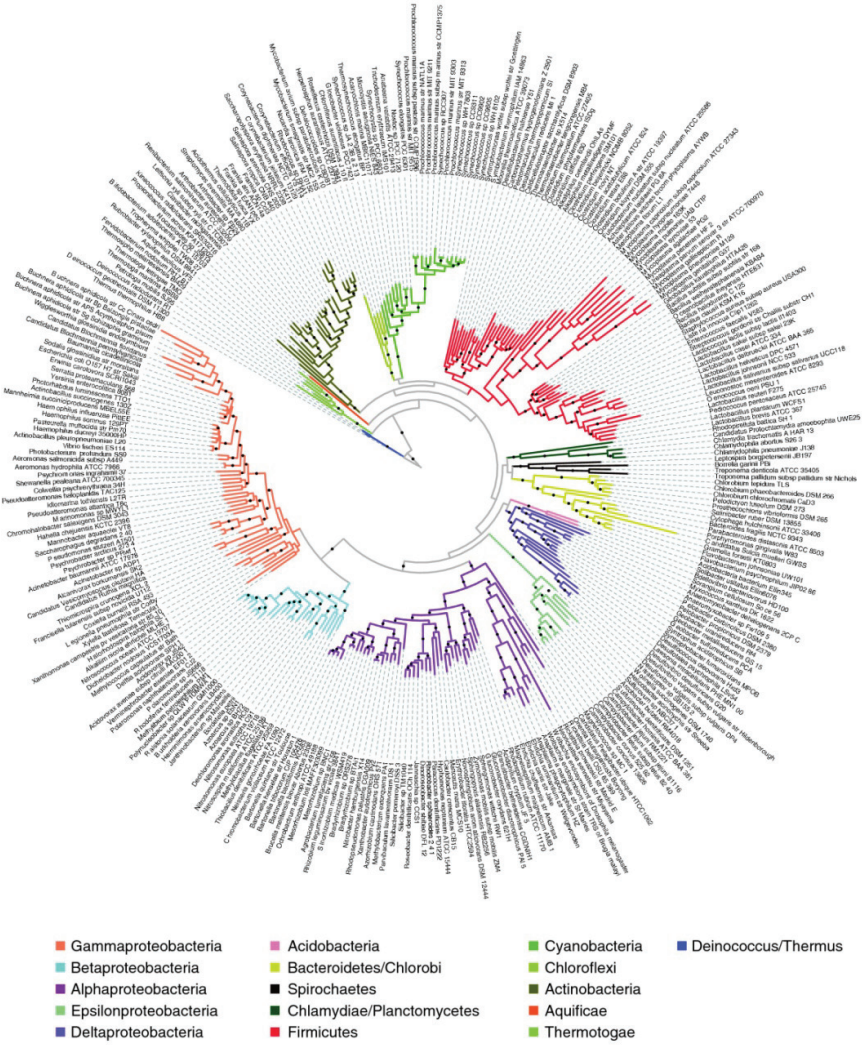


FIGURE A6-7 An unrooted maximum likelihood bacterial genome tree. The tree was constructed from concatenated protein sequence alignments derived from 31 housekeeping genes. All major phyla are separated into their monophyletic groups and are highlighted by color. The branches with bootstrap support of more than 80 (out of 100 replicates) are indicated with black dots. Although the relationships among the phyla are not strongly supported, those below the phylum level show very respectable support. The radial tree was generated using iTOL.
SOURCE: Wu and Eisen (2008).

data. And we have done this with this AMPHORA. It can allow you to stream through massive metagenomic data sets, and do taxonomic assignments for a suite of protein coating genes, just again like you would do with ribosomal RNA sequences (Figure A6-8).

And again, I think this is very advantageous in particular because of the copy number variation with ribosomal RNA. We have shown that it is better than similarity-based approaches (Figure A6-9).

Method 2: Most in the Family

This [approach involves] analyzing each individual sequence on its own. But of course, when you sequence from a new environment, you also want to compare the new sequences to each other. You don't want to compare each one individually to the reference data. And so there are a lot of methods that people have been trying to develop to build evolutionary trees of all the new sequences compared to each other.

One of the challenges with this is when you have metagenomic data in particular, the new sequences that you get might not correspond to the entire length of the reference sequences that you are analyzing. So you might have an alignment that looks like this (Figure A6-10).

One solution to this is to just trim the alignment and only pick out regions from the metagenomic data that overlap with everything in your reference database. We and other people have built methods to do this, to go through, take all the new sequences, align them to the reference data, and chop out a core region that everything has, and build an evolutionary tree of that region (Figure A6-11).

I did this by hand, to analyze the Sargasso data with ribosomal RNAs and a variety of other sequences including RecA, et cetera. All of that was done by hand. It is much better to automate this. So we have added a step in this for this ribosomal RNA pipeline. There are many other tools to do this with ribosomal RNA. Qiime (Caporaso et al., 2010), mother (Schloss et al., 2009), a lot of tools out there will build alignments for you with ribosomal RNA and help you build trees of everything. Usually these work best when you have all the sequences overlap with each other. So the challenge again is, what do you do in cases where the sequences don't overlap with each other completely, as you would get with metagenomic data.

Again, I did this by hand but we have developed methods that can allow you to do this for protein coating sequences and compare them all to each other. So in the Sargasso data, in red were sequences from the Sargasso Sea and in black were sequences that were from genomes. So you can see how those new sequences relate to each other, in addition to how they relate to the reference data. (See Figures A6-2 and A6-3.)

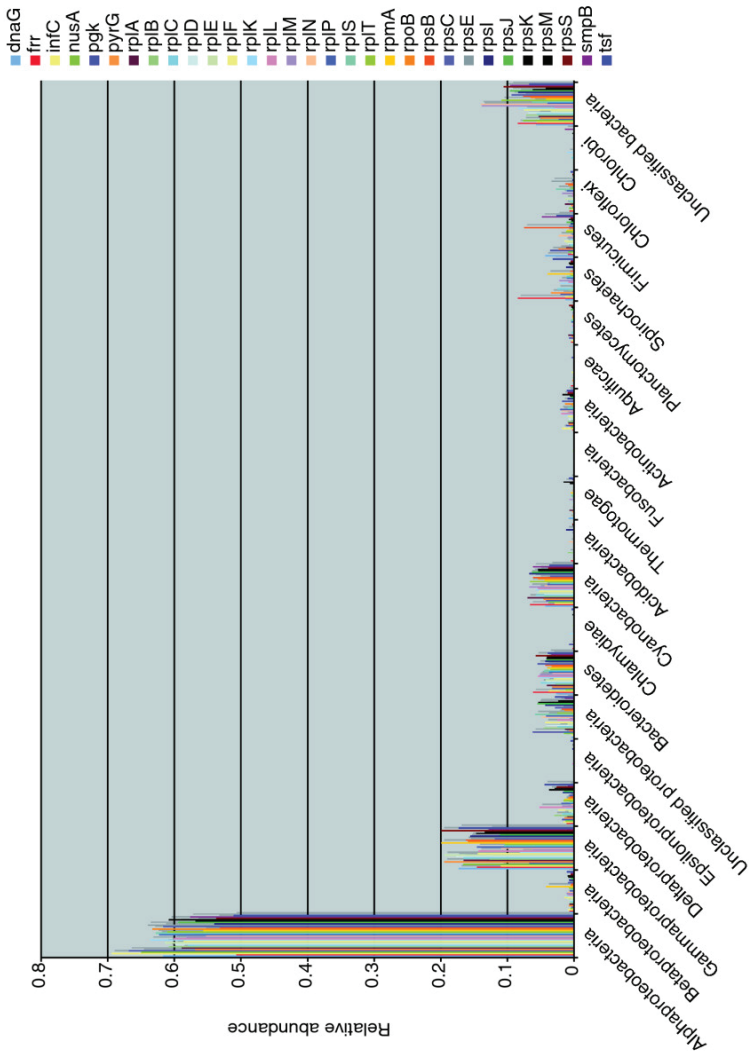


FIGURE A6-8 Major phylotypes identified in Sargasso Sea metagenomic data. The metagenomic data previously obtained from the Sargasso Sea was reanalyzed using AMPHORA and the 31 protein phylogenetic markers. The microbial diversity profiles obtained from individual markers are remarkably consistent. The breakdown of the phylotyping assignments by markers and major taxonomic groups is listed in Additional data file 5. SOURCE: Wu and Eisen (2008).

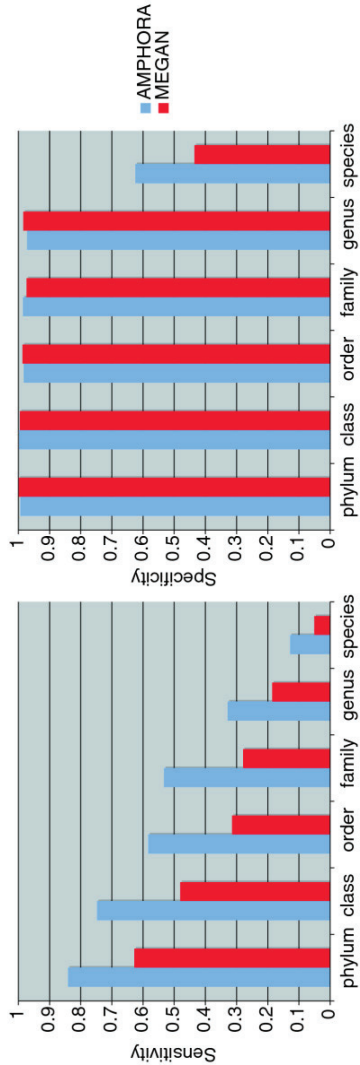


FIGURE A6-9 Comparison of the phylotyping performance by AMPHORA and MEGAN. The sensitivity and specificity of the phylotyping methods were measured across taxonomic ranks using simulated Sanger shotgun sequences of 31 genes from 100 representative bacterial genomes. The figure shows that AMPHORA significantly outperforms MEGAN in sensitivity without sacrificing specificity. SOURCE: Wu and Eisen (2008).

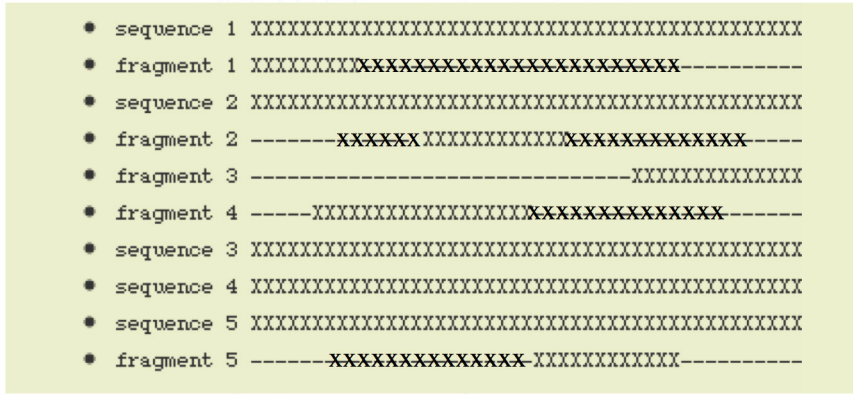


FIGURE A6-10 Hypothetical multiple sequence alignment including full length "reference" sequences as well as fragmentary sequences from metagenomic data. Xs represent areas where a sequence lines up with other sequences. Dashes represent gaps in the alignment (e.g., due to some sequences being fragments). Figure by J. A. Eisen.

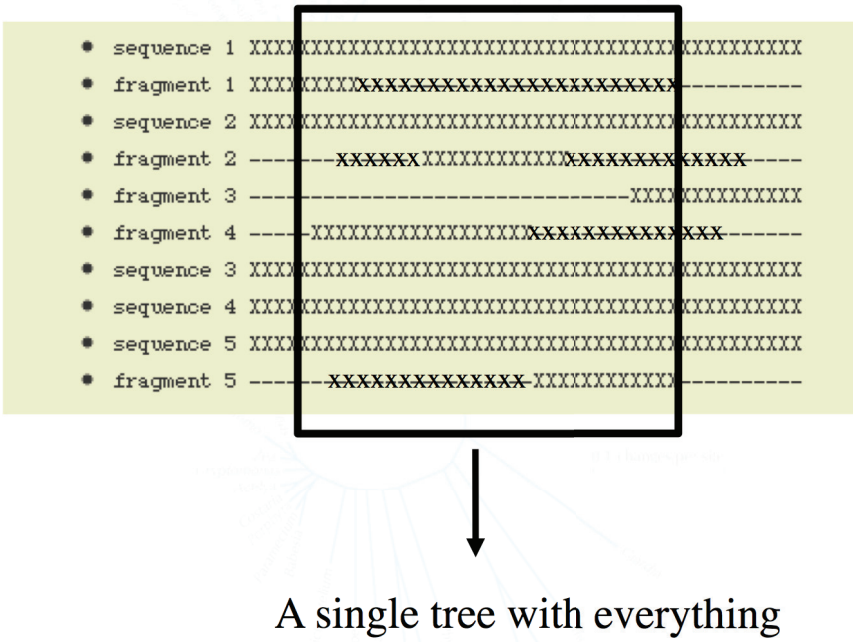


FIGURE A6-11 Hypothetical multiple sequence alignment showing one approach to carrying out phylogenetic analysis of metagenomic data—to extract a "core" region of the alignment and only analyze sequences that contain most of this core. Xs represent areas where a sequence lines up with other sequences. Dashes represent gaps in the alignment (e.g., due to some sequences being fragments). Figure by J. A. Eisen.

Method 3: All in the Family

Method 2 is limited in that it involves constraining yourself to this core region of the sequence alignment. But there are methods available that people have used, primarily in analysis of morphological data or of express sequence tag (EST) data, where you can build an evolutionary tree of sequences that don't overlap with each other at all (Figure A6-12).

So if you have good reference data, and you have a sequence that matches the left hand of the reference data and a sequence that matches the right hand of the reference data, that is sort of like if you went to an archaeological or paleontological dig and you had a femur bone from one organism and maybe some teeth from another. And you can figure out in essence whether or not they might have come from the same organism by comparing them to references. You can do the same thing with sequences. And the latest in the phylogenetic analysis of metagenomic data has been to try and build methods that will build evolutionary trees even when sequences don't overlap with each other at all, by using the reference sequences as your anchor.

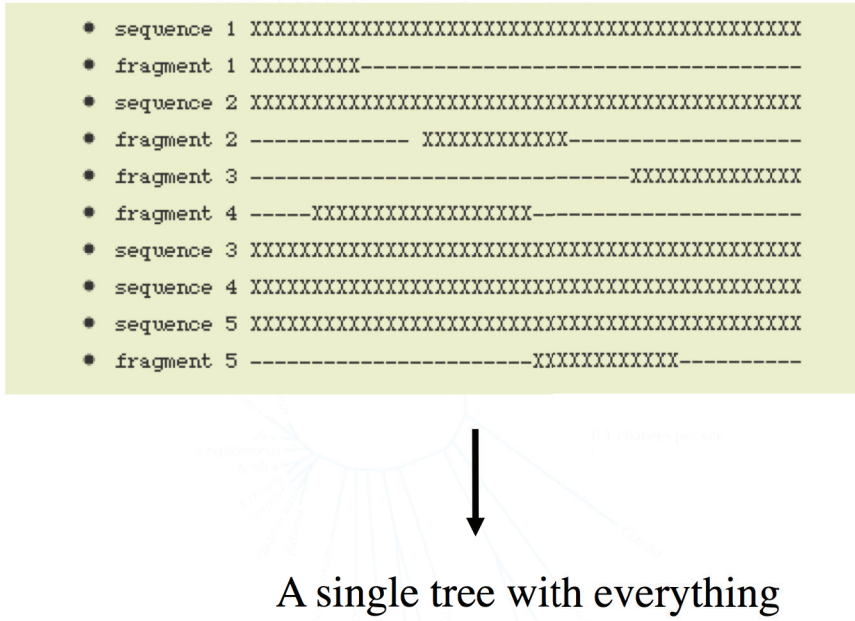


FIGURE A6-12 Hypothetical multiple sequence alignment showing an alternative strategy for phylogenetic analysis of metagenomic data—to analyze everything even if some sequences do not overlap with each other. Xs represent areas where a sequence lines up with other sequences. Dashes represent gaps in the alignment (e.g., due to some sequences being fragments). Figure by J. A. Eisen.

We have developed a few tools in my lab, in collaboration with Katie Pollard and Jessica Green that take this "all in the family" approach. One is called PhylOTU (Sharpton et al., 2011), which identifies operational taxonomic units (OTUs) using this approach (Figure A6-13).

We have another one called PhyloSift, which is like the new version of AMPHORA, that will do this for protein coating genes (see https://github.com/gjospin/PhyloSift).

There is a great method called pplacer (Matsen et al., 2010) that we have integrated in PhyloSift from Erick Matsen. That has been developed to do this exact type of thing. Again, you can build trees for sequences even if they don't overlap with each other.

Method 4: All in the Genome

So the final frontier in this is to try and build trees, even with different genes, when they do not overlap with each other. We did a little test case of this in collaboration with Stephen Kembel and Jessica Green in Oregon, where we basically took all of the genes that we had been analyzing in that AMPHORA package,

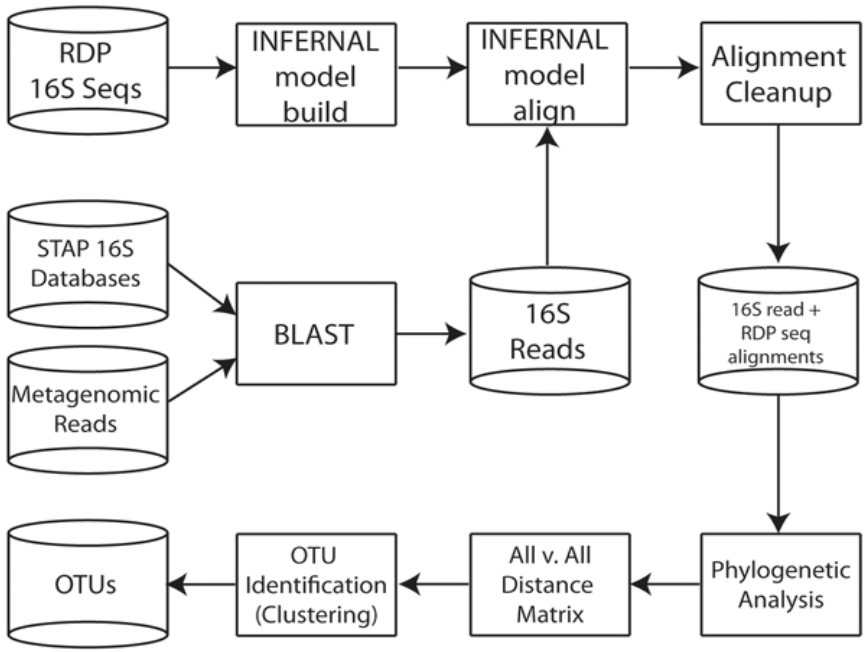


FIGURE A6-13 Computational processes are represented as squares and databases are represented as cylinders in this generalized workflow of PhylOTU.
SOURCE: Sharpton et al. (2011).

found homologues of those, and now built a reference tree of a concatenation of all of those sequences (Kembel et al., 2011). For each sequence that matched any of those individual sequences, different protein families in the environmental data, we can build an evolutionary tree that fits them relative to this anchor of the concatenated alignment of all sequences (Figures A6-14 and A6-15).

So in the long run, I think this is what we are going to want to do with environmental data, is for all genomes, build up a reference tree of all the gene families in those genomes, and then anchor environmental data to that reference tree. And you can figure out much more precisely where those sequences came from, even if they are not part of a traditional sort of evolutionary marker gene family.

Steve Kembel did this analysis, not to do phylotyping, but because he wanted an evolutionary tree to do what is called phylogenetic ecology. Many people are probably familiar with UniFrac analysis to compare the diversity of communities by their overlap in the amount of phylogenetic tree that they cover from the two communities, the unique fraction of the evolutionary tree (Figure A6-16).

That is an approach that could generally be called phylogenetic ecology. And Steve Kembel was really interested in comparing phylogenetic diversity between communities with metagenomic data. And the reason he wanted to concatenate all of these different genes was, we didn't have enough sequences from individual

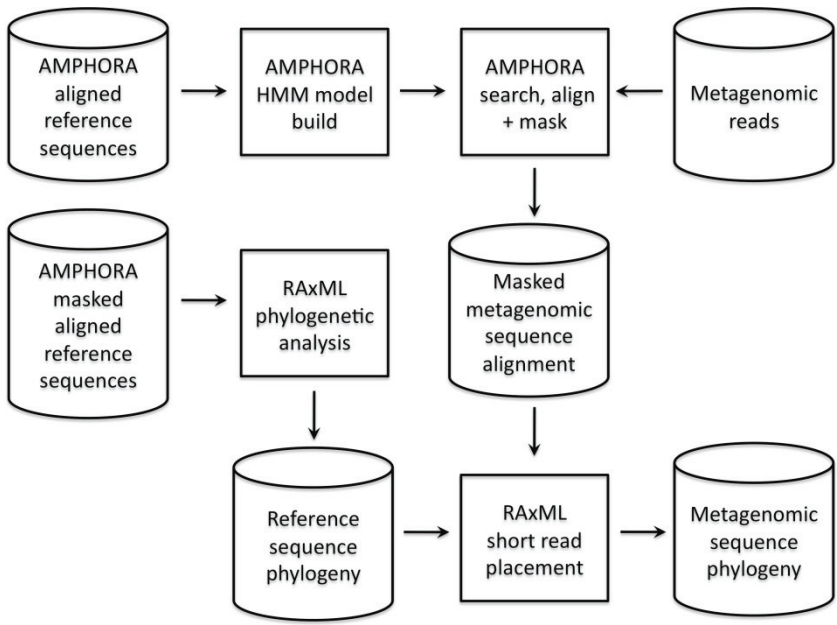


FIGURE A6-14 Conceptual overview of approach to infer phylogenetic relationships among sequences from metagenomic data sets.
SOURCE: Kembel et al. (2011).

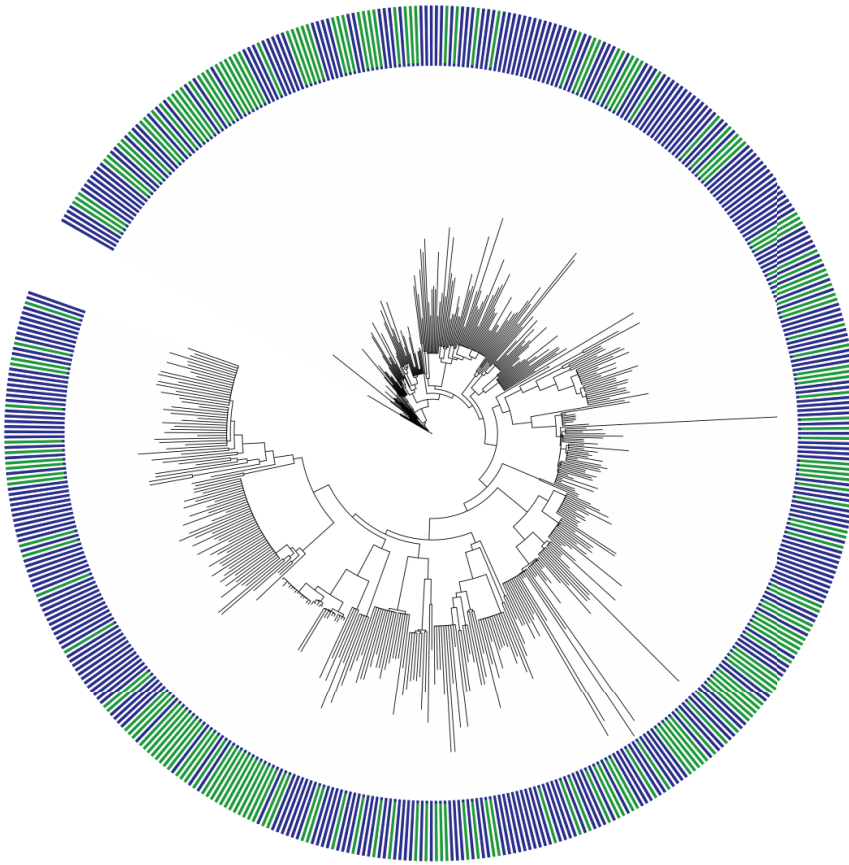
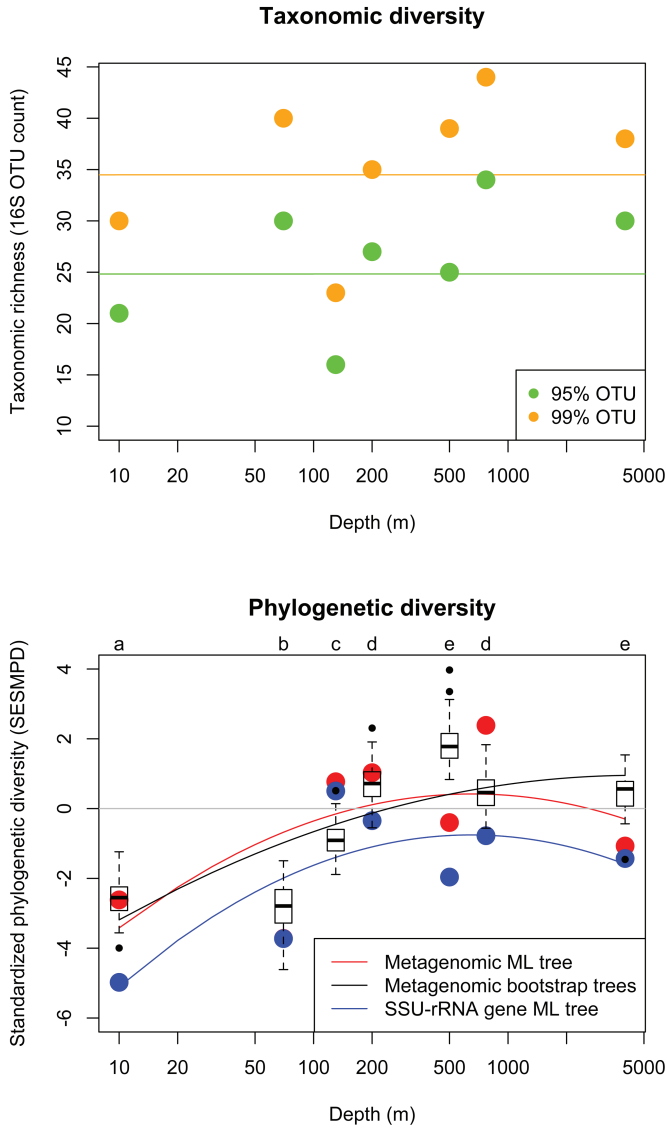


FIGURE A6-15 Phylogenetic tree linking metagenomic sequences from 31 gene families along an oceanic depth gradient at the HOT ALOHA site (DeLong et al., 2006). The depth from which sequences were collected is indicated by bar color (green = photic zone (< 200 m depth), blue = nonphotic zone). The displayed tree is the one that was identified as having the maximum likelihood by placing metagenomic reads on a reference phylogeny inferred with a WAG + G model partitioned by gene family in RAxML (Stamatakis, 2006). The phylogeny is arbitrarily rooted at Thermus for display purposes. SOURCE: Kembel et al. (2011).

FIGURE A6-16 Taxonomic diversity and standardized phylogenetic diversity versus depth in environmental samples along an oceanic depth gradient at the HOT ALOHA site. Taxonomic diversity is calculated as OTU richness (number of OTUs) based on binning of SSU-rRNA gene sequences into OTUs at a 95% and 99% similarity cutoff. Phylogenetic diversity is calculated as the standardized effect size of the mean pairwise phylogenetic distances (SESMPD) among SSU-rRNA gene sequences (blue symbols) and metagenomic sequences from the 31 AMPHORA gene families (red symbols). Standardized phylogenetic diversity values less than zero indicate phylogenetic clustering (sequences more closely related than expected); values greater than zero indicate phylogenetic evenness (sequences more distantly



related than expected). Phylogenetic diversity was estimated from the maximum likelihood phylogenies for SSU-rRNA gene and metagenomic data, as well as for 100 replicate phylogenies inferred from the metagenomic data with a phylogenetic bootstrap (black symbols). Lines indicate best-fit from quadratic regressions of diversity versus depth; the slopes of regressions of taxonomic diversity versus depth were not significantly different than zero ($P > 0.05$). At all depths, standardized phylogenetic diversity across 100 bootstrap phylogenies differed significantly from the null expectation of zero (t-test, $P < 0.05$). Phylogenetic diversity based on the 100 bootstrap phylogenies differed significantly among samples that do not share a letter label at the top of the panel (Tukey's HSD test, $P < 0.05$).
SOURCE: Kembel et al. (2011).

genes to have enough signal. But if you have 100 genes that you can analyze at once from across the genomes, you can build up enough signal to ask questions about beta diversity, et cetera, in ecological communities.

Method 5: Novel Lineages and Decluttering

So another thing that I am very interested in, and have been interested in for awhile, is to look for novel lineages in metagenomic data. We wanted to do this a long time ago, and we ran into some bioinformatic roadblocks. So what I really wanted to do was scan through metagenomic data to find whether or not there was evidence for a fourth branch in the tree, or something to that effect, things that were really phylogenetically novel compared to other sequences (Figure A6-17).

We had this problem, which was, if we did this for any type of data set, like RecA, at the time, we had something like 10,000 RecA sequences from bacteria, 200 from Archaea, and 200 from Eukaryotes. And at the time, building evolutionary trees of 10,000 sequences or more were challenging. So we wanted to sort

BOX A6-1
Questions During Talk

PARTICIPANT: (off microphone) question about long branches in Figure A6-15.

Answer: The question is what is the meaning of the especially long branches? In this case I think it is taxa that actually evolve rapidly. So in the reference data here we have some organisms like endosymbionts, mycoplasmas, et cetera, that every one of their genes evolves on a long branch. You can get artifacts in some of these cases where you have too small of a fragment, and the phylogenetic methods just get confused by that and give you a really long branch length. I don't think that's the case here. So I think most of the cases here are where taxa are known to evolve more rapidly. And the branch length is in essence a representation of evolutionary rate.

PARTICIPANT: (off microphone) question about meaning of colors in Figure A6-15.

Answer: The colors were different. Sorry, I didn't want to go into the ecological detail here. What Steve analyzes was Ed DeLong's Hawaii Ocean Time Series Data, re-analyzed that, and the colors correspond to different samples from Ed's data. What he was asking basically was primarily whether or not phylogenetic approaches to calculating beta diversity gave different answers than taxonomic approaches to calculating beta diversity, where you just count organisms as opposed to comparing the phylogenetic relatedness of organisms. So again, it is analogous to UniFrac, but now you can do it with metagenomic data, not just with ribosomal RNA data.

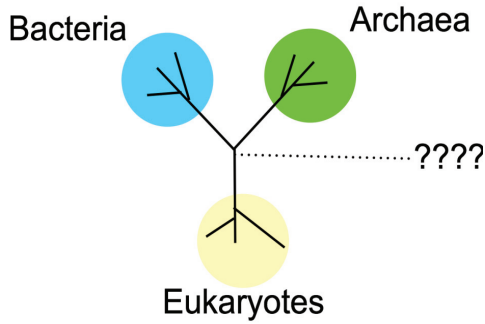


FIGURE A6-17 Searching for novel phylogenetic lineages. One key question we have been trying to answer in my lab is "Are there sequences out there that fall in between the current main groups on the tree of life?" Figure by J. A. Eisen.

of remove many of the bacterial sequences and only analyze a couple of them, as opposed to analyzing the 10,000 bacterial sequences. And what we did was develop a method that is an analog of something called Lek clustering, where you take sequences, you group them together, in essence into subfamilies, and you can identify sets of sequences that are related to each other really rapidly. And that is what we use to find, to flag different subgroups in these massive data sets of 10,000, 15,000, 100,000 sequences (Wu et al., 2011).

And when we do this for RecA or RNA polymerase or other protein family sequences, and scan through metagenomic data, you find lineages that don't group into any known current lineages of organisms (Figures A6-18 and A6-19).

These novel sequences easily could be coming from viruses that are out there in the environment, they could be new paralogs of RecA that are previously uncharacterized. Or they could be coming from phylogenetically very novel lineages that are out there in the environment. And the way to find phylogenetically novel lineages is to build evolutionary trees of all the sequences that you can get from environmental samples. I don't know if Jill [Banfield] is going to talk about this. I know Jill has done this and found novel archaeal lineages, for example, in metagenomic data, within the archaea. What we were looking for here was basically is there anything that can show up between bacteria, archaea, and eukaryotes. My guess is these are not cellular organisms. They are just DNA sequences, and they are probably from viruses or something to that effect. But again, phylogenetic approaches are the way to scan through this type of data to find novel lineages.

I am not going to talk about this, but phylotyping is also very useful for binning metagenomic data, for trying to pull things together into one group that corresponds to a particular organism. We have done this previously with endosymbionts, for example (Wu et al., 2006). But you can use it with any type of data.

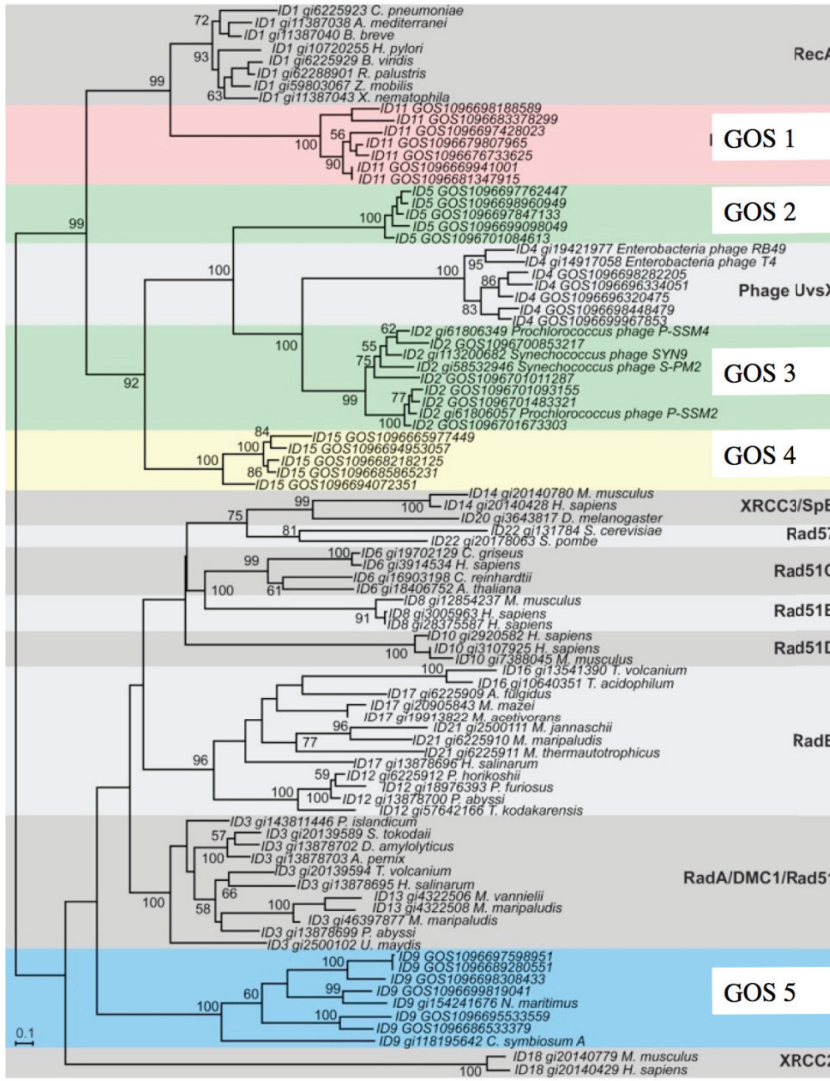


FIGURE A6-18 Phylogenetic tree of the RecA superfamily. All RecA sequences were grouped into clusters using the Lek algorithm. Representatives of each cluster that contained >2 members were then selected and aligned using MUSCLE. A phylogenetic tree was built from this alignment using PHYML; bootstrap values are based on 100 replicas. The Lek cluster ID precedes each sequence accession ID. Proposed subfamilies in the RecA superfamily are shaded and given a name on the right. Five of the proposed subfamilies contained only GOS sequences at the time of our initial analysis (RecA-like SAR, Phage SAR1, Phage SAR2, Unknown 1 and Unknown 2) and are highlighted by colored shading. As noted on the tree and in the text, sequences from two Archaea that were released after our initial analysis group in the Unknown 2 subfamily.
SOURCE: Wu et al. (2011).

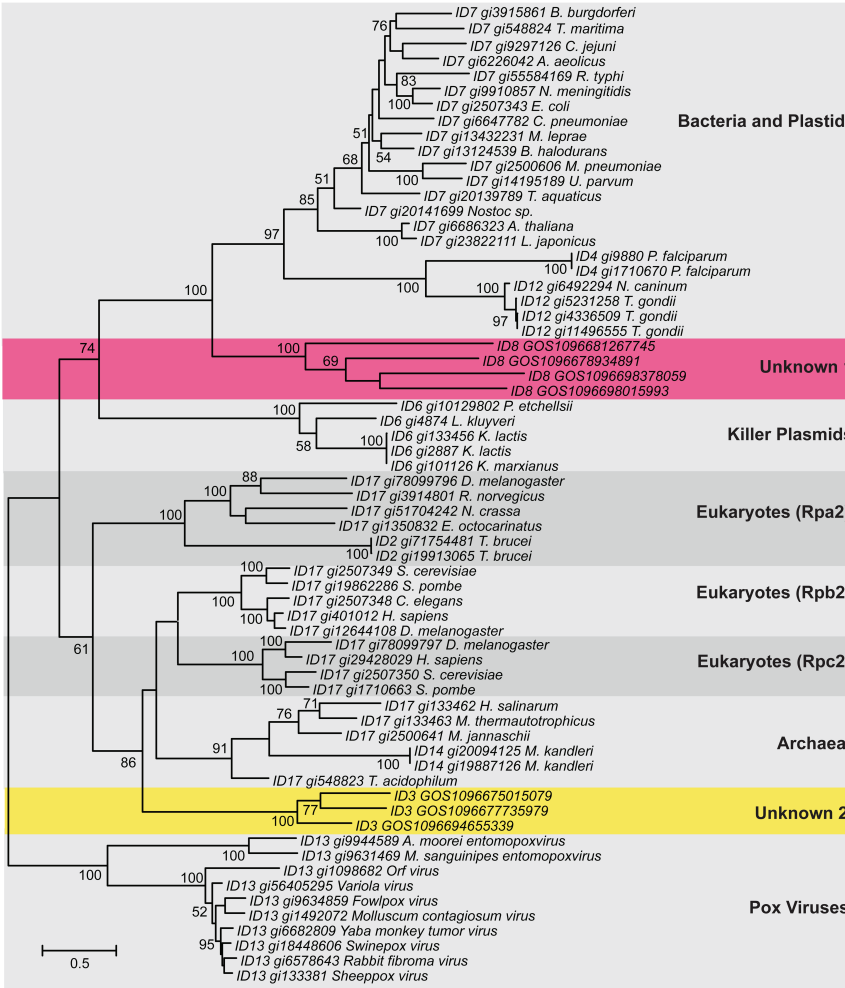


FIGURE A6-19 Phylogenetic tree of the RpoB superfamily. All RpoB sequences were grouped into clusters using the Lek algorithm. Representatives of each cluster that contained >2 members were then selected and aligned using MUSCLE. A phylogenetic tree was built from this alignment using PHYML; bootstrap values are based on 100 replicas. The Lek cluster ID precedes each sequence accession ID. Proposed subfamilies in the RpoB superfamily are shaded and given a name on the right. The two novel RpoB clades that contain only GOS sequences are highlighted by the colored panels.
SOURCE: Wu et al. (2011).

Example II: Function

So I want to shift gears. I think phylotyping is a really important area, and we are doing a lot of research on it, as are many others. And I want to shift gears just quickly to talk about a few other uses of phylogenetic analysis for studying microbial communities or microbes. And one relates to functional diversity and functional prediction, and I have spent a lot of time working on this over the years, and I have been very interested in basically when you have new sequence data, how do you make a prediction of the function of that gene.

We have talked about this a little bit at this meeting with the examples of the cytochrome, oxidases, and a few other examples here. How do you make a robust prediction for a sequence of a gene? And I think just like analyzing ribosome RNA sequences to understand an organism by building an evolutionary tree of those sequences, you should build an evolutionary tree of protein family sequences in order to understand the functional diversity in a family.

I developed an approach that I originally called phylogenomics (Figure A6-20), to do this many years ago. And [the approach] is basically: you take a sequence, you compare it to its homologues, you build an evolutionary tree of that sequence and its homologues, you overlay experimentally determined functions onto the tree. And then you use character state reconstruction methods to predict the functions of unknowns.

Sound familiar, anybody? It's phylotyping. I in essence co-opted this from the ribosome RNA world (Figure A6-21). But you can apply it to functions as opposed to organisms. And it is a very powerful tool in predicting functions for uncharacterized genes.

Again, placing them in their phylogenetic context is incredibly powerful. I think in the interest of time I won't go into the multiple examples that I have of this. I would be happy to talk to people about this. This is routinely used now in many genome analysis projects to build evolutionary trees of various sequences. You could do it with whatever sequences you want from environmental data as well as from sequence genomes.

The latest thing in functional prediction, which I think is really interesting, is to use non-homology functional prediction methods, which look at things like distribution of patterns of genes across organisms (e.g., see our use of phylogenetic profiling [Wu et al., 2005]). You can also use distribution patterns of genes across environments to try and help you make predictions of functions of genes. This has been done in a variety of metagenomic projects. Exactly how you group genes and analyze the correlation between the distribution pattern of a gene and the taxa present in a sample, and the metadata of the sample, is still sort of a work in progress.

We have now been collaborating with Simon Levin and others as part of a project that Simon Levin was in charge of. In this "DARPA fundamental laws of biology" project we have been working to apply non-homology methods to metagenomic data (Jiang et al., in press).

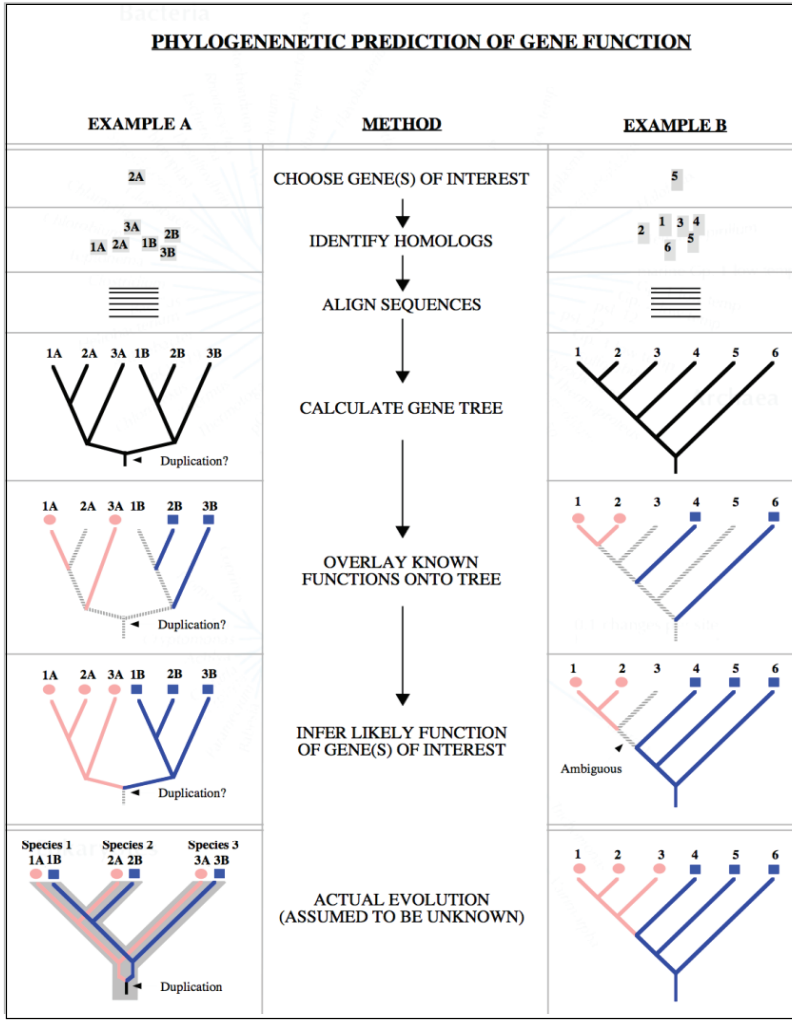


FIGURE A6-20 Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (A) A gene family has undergone a gene duplication that was accompanied by functional divergence. (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 (bottom) 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; gray (with hatching) represents either unknown or unpredictable functions. SOURCE: Eisen (1998).

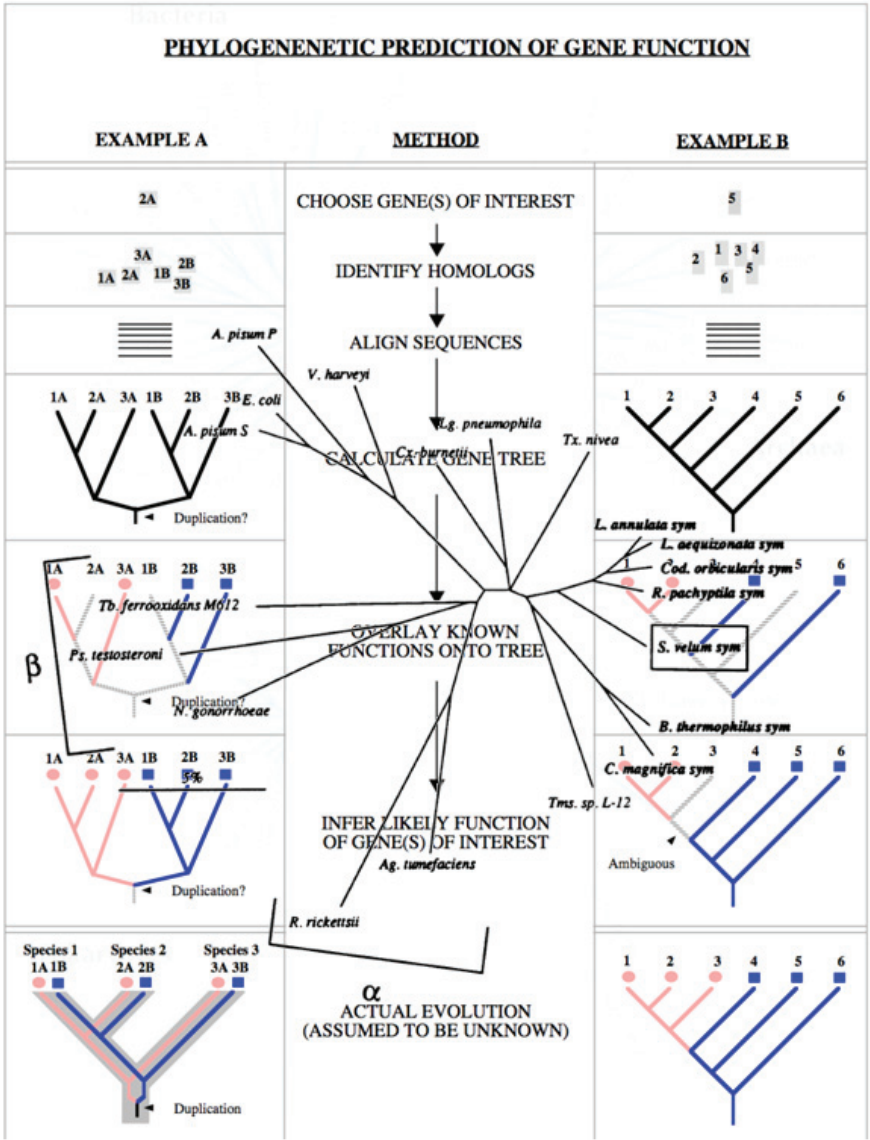


FIGURE A6-21 Phylogenomic functional prediction is based on the concept of phylotyping. This figure is a merging of Figure 1 and Figure 20. By J. A. Eisen.

Example III: Selecting Organisms to Study

The last thing I want to talk about is this issue of selecting organisms for study. A different use of phylogenetics is to try and understand what we have known about in the diversity of life. And so if you go through the ribosomal RNA tree of life there are many different lineages of bacteria (Figure A6-22). In 2000, when we first sort of noted this ourselves, most of the genomes came from three phyla. That is still basically true. There were some genome sequences available from other lineages, but most lineages were poorly sampled. The same trend is true for eukaryotes, true for archaea, and true for viruses (e.g., see Eisen [2000]).

We had a project when I was at TIGR to sequence eight representatives of novel phyla for which genomes were not available that was funded by the National Science Foundation. More recently I have been coordinating a project called the genomic encyclopedia of bacteria and archaea at the Department of Energy (DOE) Joint Genome Institute where we have been really filling in the tree of bacteria and archaea, of cultured organisms with representative genome sequences.

This is one of these massive projects with hundreds of people involved, and it has been this amazing collaboration with the DOE JGI and the DSMZ culture collection. What we did is basically go through the tree of life, select the genomes to sequence by their phylogenetic novelty, and then ask questions about whether or not phylogeny ended up being a useful guide in selecting genome sequences. And we have shown now about five or six areas where phylogenetic sampling has improved our analysis of genome or metagenome data. So one is in functional predictions of genes, another is in discovery of genetic diversity. So a phylogenetically novel organism, if all else is equal, is more likely to have new protein families than a phylogenetically not novel organism (Figure A6-23).

And we also showed that these phylogenetically novel organisms could help you analyze metagenomic data, by providing more reference data across the tree in essence. But when we did this, there was very little benefit to analyzing the metagenomic data, from the first 50 or even 100 genomes that we sequenced from this genomic encyclopedia project (Wu et al., 2009). And the reason for this is we need to adapt many of the methods that we are doing to improve our ability to make use of this phylogenetically diverse data. So we need to design new phylogenetic methods, new metagenomic methods, to take into account this environmental data.

I have been involved in this great collaboration with Jessica Green and Katie Pollard and Martin Wu to try and develop methods to take advantage of this. It just ended. We called it ISEEM (see http://iseem.org).

This is one of the products of iSEEM, which is a tree of 2,500 genomes, that Jenna Morgan and Aaron Darling in my lab generated. We have been building new protein family markers from all of these genomes, so we can improve that AMPHORA pipeline by having hundreds to thousands of new phylogenetic markers to use to scan through metagenomic data.

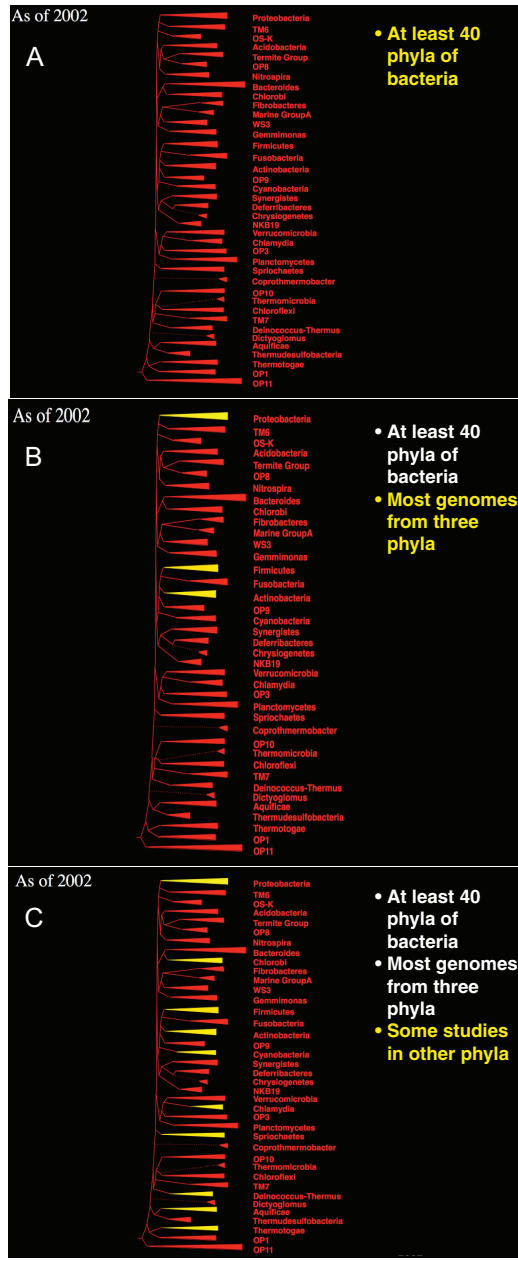


FIGURE A6-22 Phylogenetically biased genome sequencing. Phylogenetic tree is based on one from Hugenholtz 2002.
SOURCE: Adapted from Hugenholtz (2002).

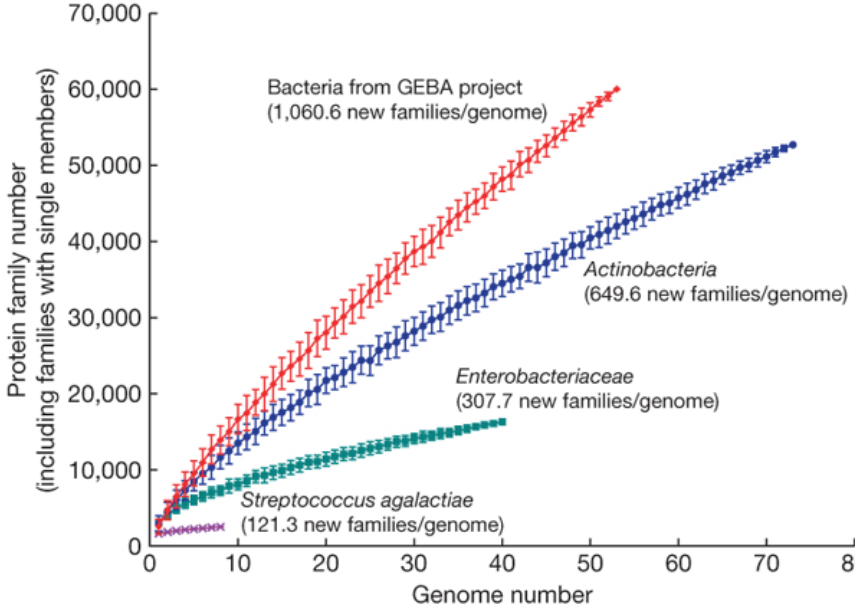


FIGURE A6-23 For each of four groupings (species, different strains of Streptococcus agalactiae; family, Enterobacteriaceae; phylum, Actinobacteria; domain, GEBA bacteria), all proteins from that group were compared to each other to identify protein families. Then the total number of protein families was calculated as genomes were progressively sampled from the group (starting with one genome until all were sampled). This was done multiple times for each of the four groups using random starting seeds; the average and standard deviation were then plotted.
SOURCE: Wu et al. (2009).

And the last thing I want to leave you with is the reason why I think this project did not really help analyze metagenomic data, which is we haven't even begun to scratch the surface of microbial diversity in terms of reference genome data. So if you go through the tree of life, and you count the branch length in the tree, it is something called PD, or phylogenetic diversity (Figure A6-24). If you sum up the total length of the branches, for all the genomes sequenced before our project, it came to 25 units. Each genome that we added, added a lot of new units of phylogenetic diversity. It better have, because that is how we picked them. If you go through cultured organisms that are known and described—so the 8,000 or so described bacteria in archaea—we would need about 1,000 genomes to capture half of that diversity. This will be done in the next year or two. However, if you go through all the environmental data, we would need about 10,000 genomes to capture half of the diversity known 5 years ago in full-length ribosomal RNA sequences. So the vast majority of genomic diversity out there is uncaptured in most studies of cultured organisms.

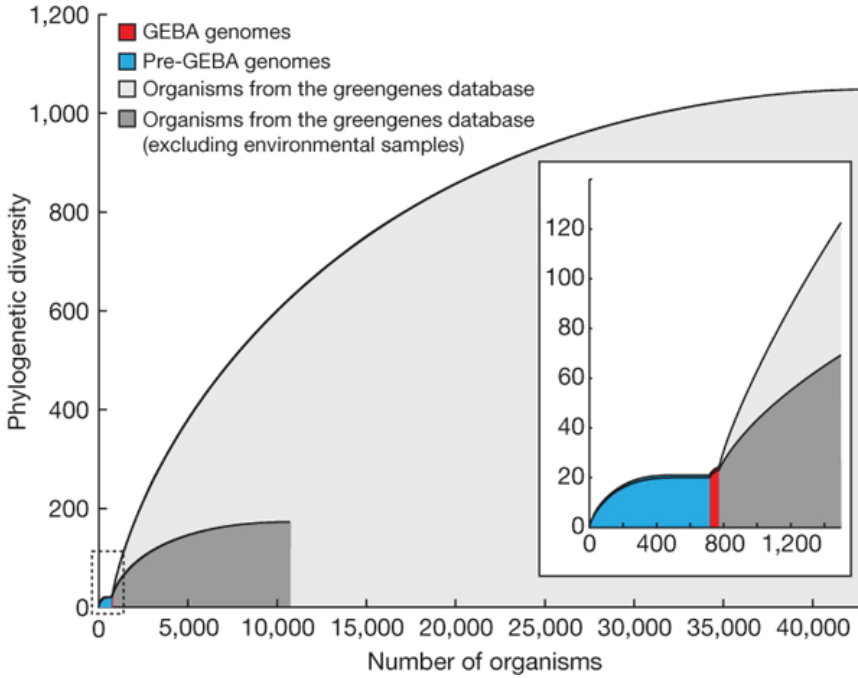


FIGURE A6-24 Using a phylogenetic tree of unique SSU rRNA gene sequences, phylogenetic diversity was measured for four subsets of this tree: organisms with sequenced genomes pre-GEBA (blue), the GEBA organisms (red), all cultured organisms (dark grey), and all available SSU rRNA genes (light grey). For each subtree, taxa were sorted by their contribution to the subtree phylogenetic diversity and the cumulative phylogenetic diversity was plotted from maximal (left) to the least (right). The inset magnifies the first 1,500 organisms. Comparison of the plots shows the phylogenetic "dark matter" left to be sampled. SOURCE: Wu et al. (2009).

I think the solution to this is to do genomes of uncultured organisms, either via single cell capture or, as I think [meeting participants] will hear from Jill [Banfield] and other people, metagenomic sequencing and assembly of those metagenomes can generate genome data from uncultured organisms. [Note added after talk: see Narasingarao et al. (2012) for an example of this.] And by doing that, we will really fill in the tree. And that will enable all sorts of different uses of phylogeny in analysis of environmental data.

And I will leave it at that and say that of course we need experiments from across the tree, too. Sequencing is great. I love sequencing, but it doesn't tell us everything. And what we need is an organized effort like this genomic encyclopedia to target functional diversity from across the tree of life, too. And I will leave it there.

[Update. I realize in retrospect that my "conclusion" for my talk was pretty minimal so I am adding a few sentences below to wrap up this paper.]

Conclusions

All biological entities have a history. Making sense out of biological data is best done in the context of that history. What I have tried to show in this paper are examples of how phylogenetic approaches can be useful in studies of microbial diversity. I gave three main kinds of examples—phylotyping, functional prediction, and identifying gaps in our genomic reference data. There are hundreds more, some developed by myself, most developed by others. To best understand the present, and even predict the future, we need to understand the past and how things changed over time.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5):335-336.

DeLong E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. Rodriguez Brito, S. W. Chisholm, and D. M. Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496-503.

Eisen, J. A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* 8(3):163-167.

Eisen, J. A. 2000. Assessing evolutionary relationships among microbes from whole genome analysis. *Current Opinion in Microbiology* 3:475-480.

Eisen, J. A., S. W. Smith, and C. M. Cavanaugh. 1992. Phylogenetic relationships of chemoautotrophic bacterial symbionts of *Solemya velum* say (Mollusca: Bivalvia) determined by 16S rRNA gene sequence analysis. *Journal of Bacteriology* 174(10):3416-3421.

Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biology* 3(2): REVIEWS0003.

Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace. 1998. Novel division level bacterial diversity in a Yellowstone hot spring. *Journal of Bacteriology* 180(2):366-376.

Jiang, X., M. G. I. Langille, R. Y. Neches, M. A. Elliot, S. A. Levin, J. A. Eisen, J. S. Weitz, and J. Dushoff. In press. Functional biogeography of ocean microbes: Dimension reduction of metagenomic data identifies biological patterns across scales. *PLoS One*.

Kembel, S. W., J. A. Eisen, K. S. Pollard, and J. L. Green. 2011. The phylogenetic diversity of metagenomes. *PLoS ONE* 6(8):e23214. doi:10.1371/journal.pone.0023214.

Matsen, F. A., R. B. Kodner, and E. V. Armbrust. 2010. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538.

Narasingarao, P., S. Podell, J. A. Ugalde, C. Brochier-Armanet, J. B. Emerson, J. J. Brocks, K. B. Heidelberg, J. F. Banfield, and E. E. Allen. 2012. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME Journal* 6(1):81-93. doi: 10.1038/ismej.2011.78.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied Environmental Microbiology* 75(23):7537-7541.

Sharpton, T. J., S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O'Dwyer, J. L. Green, J. A. Eisen, and K. S. Pollard. 2011. PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computation Biology* 7(1): e1001061.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, H. O. Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.

Wu, D., S. C. Daugherty, S. E. Van Aken, G. H. Pai, K. L. Watkins, H. Khouri, L. J. Tallon, J. M. Zaborsky, H. E. Dunbar, P. L. Tran, N. A. Moran, and J. A. Eisen. 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology* 4(6):e188. doi:10.1371/journal.pbio.0040188.

Wu, M., and J. A. Eisen. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology* 9:R151. doi:10.1186/gb-2008-9-10-r151.

Wu, D., A. Hartman, N. Ward, and J. A. Eisen. 2008. An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP). *PLoS ONE* 3(7):e2566. doi:10.1371/journal.pone.0002566.

Wu, D., M. Wu, A. Halpern, D. B. Rusch, S. Yooseph, M. Frazier, J. C. Venter, and J. A. Eisen. 2011. Stalking the fourth domain in metagenomic data: Searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS ONE* 6(3):e18011. doi:10.1371/journal.pone.0018011.

Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk, J. A. Eisen. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462(7276):1056-1060.

Wu, M., Q. Ren, A. S. Durkin, S. C. Daugherty, L. M. Brinkac, R. J. Dodson, R. Madupu, S. A. Sullivan, J. F. Kolonay, D. H. Haft, W. C. Nelson, L. J. Tallon, K. M. Jones, L. E. Ulrich, J. M. Gonzalez, I. B. Zhulin, F. T. Robb, and J. A. Eisen. 2005. Life in hot carbon monoxide: The complete genome sequence of Carboxydothermus hydrogenoformans Z-2901. *PLoS Genetics* 1(5):e65.