LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-410988

FY08 LDRD Final Report Probabilistic Inference of Metabolic Pathways from Metagenomic Sequence Data

P. D'haeseleer

March 2, 2009

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Auspices Statement

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was funded by the Laboratory Directed Research and Development Program at LLNL under project tracking code 08-ERD-011.

FY08 LDRD Final Report

Probabilistic Inference of Metabolic Pathways from Metagenomic Sequence Data

LDRD Project Tracking Code: 08-ERD-011

Patrik D'haeseleer, Principal Investigator

Abstract

Metagenomic "shotgun" sequencing of environmental microbial communities has the potential to revolutionize microbial ecology, allowing a cultivation-independent, yet sequence-based analysis of the metabolic capabilities and functions present in an environmental sample. Although its intensive sequencing requirements are a good match for the continuously increasing bandwidth at sequencing centers, the complexity, seemingly inexhaustible novelty, and "scrambled" nature of metagenomic data is also proving a tremendous challenge for analysis. In fact, many metagenomics projects do not go much further than providing a list of novel gene variants and over- or under-represented functional gene categories. In this project, we proposed to develop a set of novel metagenomic sequence analysis tools, including a binning method to group sequences by species, inference of phenotypes and metabolic pathways from these reconstructed species, and extraction of coarse-grained flux models. We proposed to closely collaborate with the DOE Joint Genome Institute to align these tools with their metagenomics analysis needs and the developing IMG/M metagenomics pipeline. Results would be cross-validated with simulated metagenomic data using a testing platform developed at the JGI.

Introduction/Background

Metagenomic sequence data gives us an insight into the genes and pathways that drive the behavior and metabolic capabilities of a microbial community. Just as whole-genome sequencing has radically changed 21st century microbiology, we expect that metagenomic sequencing will likewise cause a sea change in microbial ecology and environmental microbiology. The first meta-genome of a microbial community was sequenced only three years ago (Tyson *et al*, 2004), today there are more than two dozen metagenome sequence data sets available, and more than a hundreds of new projects are expected over the next five years, indicating a potentially exponential increase, roughly eight years behind the exponential rise in sequenced single genomes.

However, despite the intense interest in this new approach, making sense of these large sequence data sets has been very challenging, in large part because the "shotgun" sequencing destroys the information of which piece of sequenced DNA belongs to which species in the community. We have a unique opportunity today, to create the tools that will make this data accessible, and shape the field for years to come.

In particular, the extraction of metabolic pathways from metagenomic data – a key step in determining the repertoire of metabolic capabilities present in the community – is hindered by the mixture of enzyme sequences from hundreds or thousands of different species. Naively treating the entire community as a single "bag of enzymes" will lead to the spurious inference of large numbers of mosaic pathways, consisting of enzymes from multiple species. Separating the mixture of shotgun sequences into separate species presents a serious challenge for all but the least complex of communities. Various binning methods

have been proposed, from coarse metrics such as GC content and sequence coverage, frequencies of oligonucleotides or more complex sequence patterns, to overly specific criteria such as closest BLAST hits. We proposed to develop a novel binning method that takes advantage of the specificity of BLAST hits, but locates each species bin based on its phylogenetic relationship with fully sequenced reference species.

SRI's Pathway Tools software (Karp *et al*, 2002) provides ontology, inference, and visualization services for organism-specific pathway-genome databases (PGDBs). To reconstruct the metabolic network of an organism, the PathoLogic program (Karp *et al*, 2002) uses an organism's genome annotation and a reference set of known, experimentally validated pathways from MetaCyc (Caspi *et al*, 2006) or from the PGDB of a known, closely related species. However, it assumes a complete and fully annotated genome, and will make incorrect inferences when faced with the incomplete species coverage, poorly separated bins, and uncertain functional annotations typical of most metagenomics projects. We proposed to provide the capability to deal with this type of data by providing a probabilistic representation of each predicted gene, enzyme, pathway, and species in the community, allowing us to analyze the collective genome and metabolic capabilities of an entire microbial community.

Our proposed binning, enzyme annotations and pathway inference pipeline would take advantage of JGI's developing metagenomic analysis pipeline, and the IMG/M metagenomics portal (Markowitz *et al*, 2006a), which is currently lacking in effective pathway analysis. We will additionally take advantage of the metagenomic data simulation system that has been developed at JGI (Markowitz, 2006a) for evaluation of metagenomics tools and comparison with state-of the art existing tools (*e.g.* alternative binning methods).

In addition to leveraging our early access to the continuous stream of new metagenomics data being generated from the growing number of metagenomic sequencing projects under way at JGI, we also planned to demonstrate the power of this approach on a microbial community of interest to JGI, LLNL and DOE.

Research Activities

FY2008 Accomplishments and Results

Due to a reduction in funding, we eliminated sections on remote homology detection, and a NASA collaboration present in the original proposa (see Raymond *et al*., 2007)l. We more than accomplished all other goals for FY08, achieving excellent results and attracting interest from collaborators at JGI and SRI. Specifically, we illustrated a simplified form of metabolic pathway inference on a real-world metagenomic dataset, developed a method to map reference species and gene sequences into a Euclidean phylogenetic space, investigated the conservation of phenotypic traits and metabolic pathways with evolutionary distance, and developed and validated a Nearest Neighbor algorithm for prediction of metabolic pathways from 16S marker genes.

In addition, we also made significant progress on work that had originally been planned for FY09, including development of a phylogenomic algorithm for pathway prediction, design of a standalone web-based tool, and demonstration on real-world datasets. Unfortunately, despite the outstanding results in this first year, the project was canceled one week before the start of FY09 because of budget problems at the funding directorates. Due to this loss of funding, and long-term illness of the postdoc, it may be difficult to leverage any of the highly promising results that have come out of this project.

Results/Technical Outcome

Naïve inference of metabolic pathways from Acid Mine Drainage metagenome bins

Banfield *et al*. sequenced the metagenome of a biofilm floating on top of extremely acidic water in the abandoned Richmond iron mine at Iron Mountain, California (Tyson *et al*., 2004). The community composition of this biofilm was dominated by a small number of bacteria and archaea, presumably due to the extreme, pH<1.0 environment. Because of the low complexity, the 4 top species represented in the metagenome could be easily binned based on differences in GC content (38% G+C for the Archaea, 55% G+C for the Bacteria) and depth of coverage (10X local read depth for the two most abundant species, 10X for the less abundant ones).

The metagenome sequence for this Acid Mine Drainage community is available at JGI's

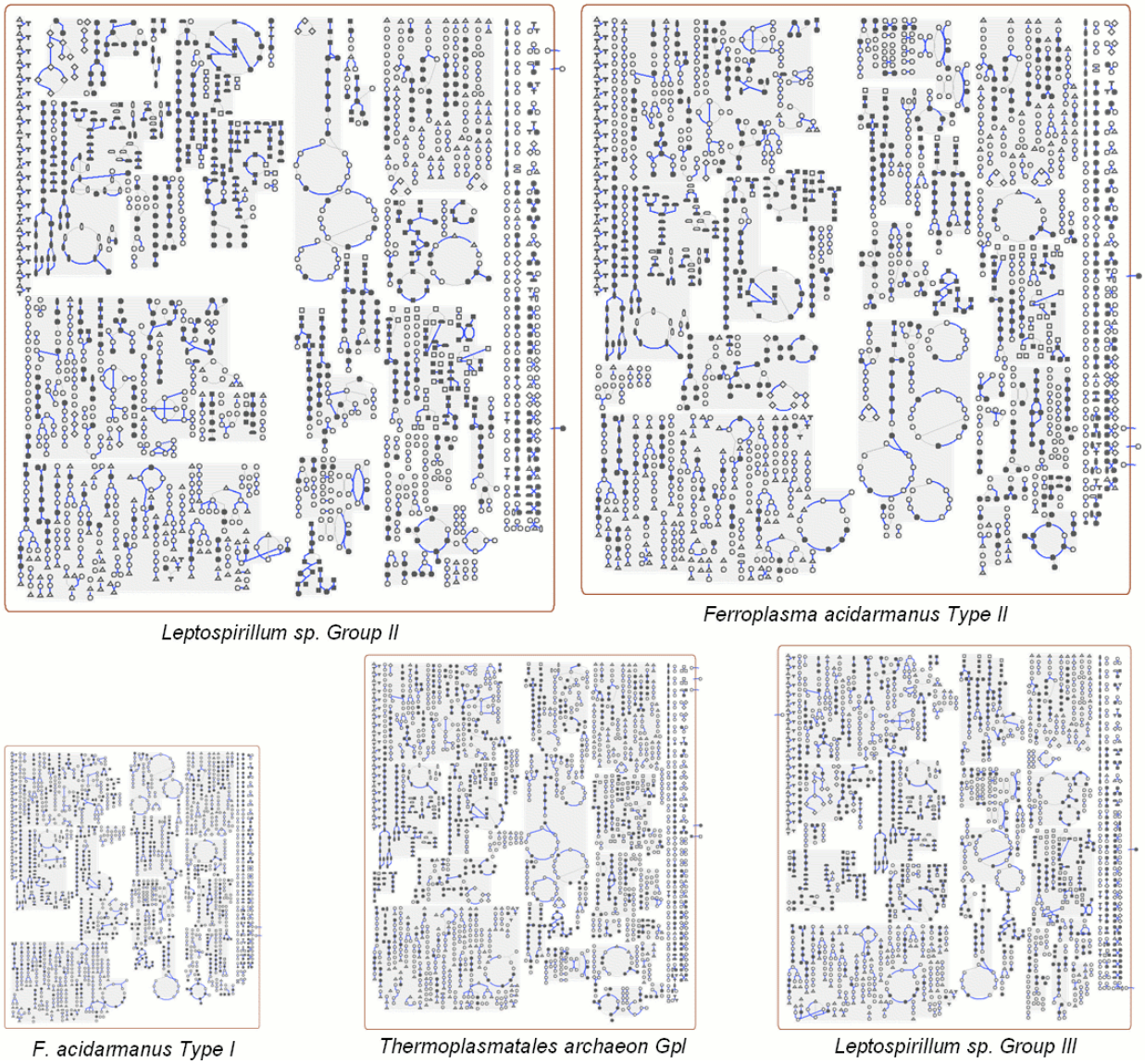

Figure 1. "Naive" reconstruction of metabolic pathways within the 5 sequence bins of the Acid Mine Drainage microbial community. Overview diagrams for the five dominant species were drawn proportional to their representation in the metagenome (10X coverage for *Leptospirillum* group II and Ferroplasma type II, around 3X or less for the others)

IMG/M microbiome portal (Markowitz *et al*., 2008), separated into five species bins based on tetranucleotide frequences. As a proof of principle, we downloaded each of these five annotated species bins, and processed them using the PathoLogic algorithm of Pathway Tools to extract the metabolic pathways. Each bin was treated as if it were a complete annotated genome, separated into numerous "Unassigned Contigs" - ranging from 59 scaffolds for the abundant and almost fully sequenced Ferroplasma acidarmanus Type II (average scaffold length 31Kbp), to 474 scaffolds for the lower-coverage and therefore harder to assemble Leptospirillum sp. Group III (average scaffold length 5.6Kbp).

Figure 1 shows the Pathway Tools overview diagrams reconstructed from these bins. This shows that the platform currently can handle the unusual nature of the data – hundreds of separate sequence scaffolds, each of which requires two separate input files in the current implementation. It also illustrates what a multi-species overview diagram might look like.

Sequence binning in phylogenetic space

The 16S rRNA sequence alignments for all fully sequenced microbial genomes were extracted from the GreenGenes database (DeSantis *et al*., 2006), and their pairwise evolutionary distances were calculated using ClustalW. We used Principal Coordinate Analysis to map the corresponding species into a high-dimensional Euclidean phylogenetic space. The most significant dimension in this space cleanly separates the Archaea from Bacteria, while the next most significant components separate the major bacterial lineages (Figure 2). The internal structure within these high-level clusters is also consistent with the evolutionary relationships at the lower taxonomic ranks.

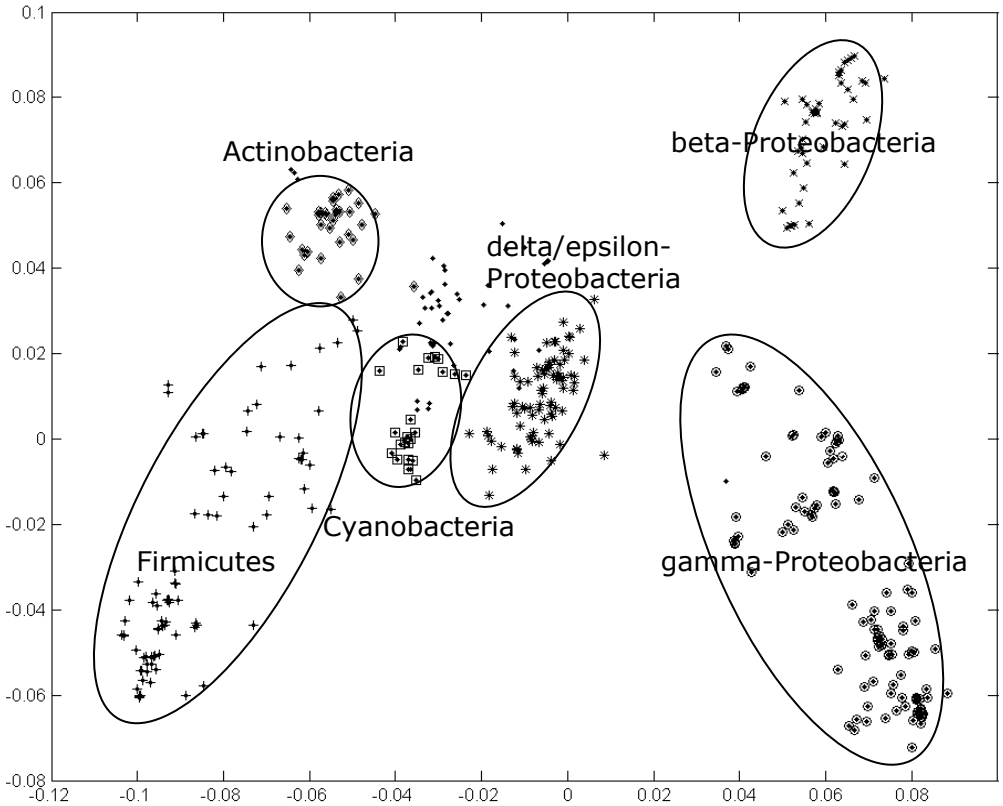

Figure 2. Principal coordinate mapping of 488 reference species, based on pairwise 16S distances. Only the second and third principal coordinates are shown. The first principal coordinate separates Bacteria from Archaea (not shown).

Note that the goal of this exercise is *not* to propose this as a novel tool for the construction of phylogenetic trees. In fact, since the distances within this space are a close approximation of the evolutionary distances between the species, applying a hierarchical clustering within this space would be roughly equivalent to the classical neighbor joining or UPGMA methods of phylogeny reconstruction – methods which are known to be highly computationally efficient, but tend to generate sub-optimal trees. Rather, the purpose of this phylogenetic space is to provide an efficient structure into which we can map – and cluster – hundreds of thousands of metagenomic sequences, based on their evolutionary distance with respect to the reference species which define the space.

The next step in this part of the project would be to map known individual genes from known species into this space, based on their BLAST distance to other reference species. Initial experiments showed that BLAST scores between genes can be correlated to evolutionary distances between species, allowing us to calibrate BLAST scores to distances to known reference points in the Euclidean phylogenetic space. The location of sample genes mapped into this space corresponded well with the location of the species the genes belonged to, indicating the feasibility of mapping and clustering metagenomic sequence reads into the phylogenetic space.

Conservation of phenotypic traits with evolutionary distance

The evolutionary relationship of an organism to other known organisms, typically measured by similarities in the nucleotide sequence of the 16S ribosomal subunit, is often used implicitly as a first estimate of the role of an organism within its larger environment, i.e. its phenotype. However, the utility of 16S as an approximation of phenotype is in question, as even strains from the same species can exhibit a substantial amount of diversity with regard to both genome content and phenotype. Here, we have critically assessed the relationship between phenotype and evolutionary distance across the prokaryotic kingdom.

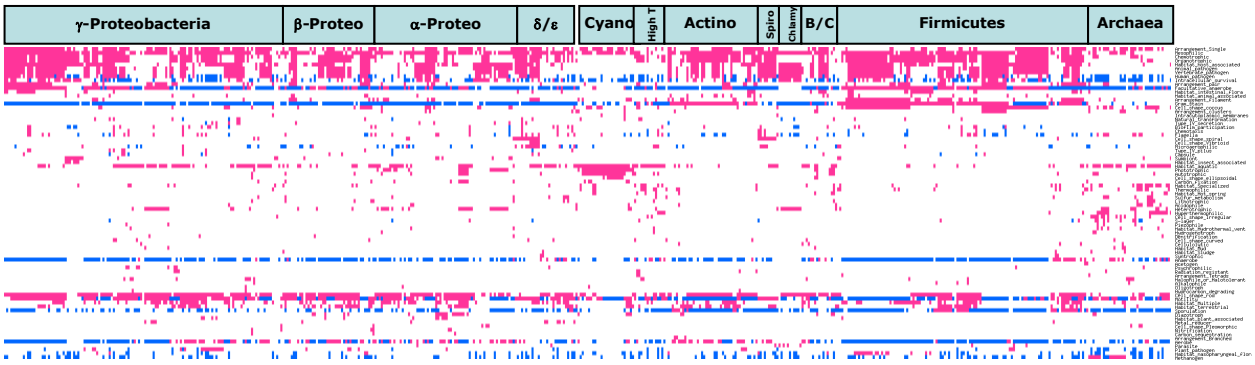

Figure 3. Phenotype profiles for 558 prokaryotic organisms x 79 phenotypes. Each column represents one organism, each row represents one phenotype. Red, blue, or white dots indicate the phenotype is present, absent, or unknown for that particular organism.

The 16S rRNA sequence alignments for all fully sequenced genomes were extracted from the GreenGenes database (DeSantis *et al*., 2006) and the evolutionary distances between the 488 organisms were calculated using ClustalW. As an alternative measure of evolutionary distance we also used distances between organisms derived from a phylogenetic tree based on universal protein sequences (Ciccarelli *et al*., 2006).

In an earlier project, we assembled and curated a large phenotype dataset, consolidated into 79 well defined and biologically relevant phenotypes across 558 sequenced prokaryotic organisms (Figure 3), and developed a phenotypic similarity measure between organisms, based on the log likelihood of agreements for the phenotype annotations they have in

common, minus a penalty for disagreements for the phenotypes for which they have opposite annotations. In the current project, we examined in greater detail the conservation of phenotypic profiles with evolutionary distance, in order to build a method to extrapolate phenotypic traits from well-studied model organism to novel organisms for which we may only know their evolutionary relationship.

We showed that the correlation between evolutionary distance and phenotypic similarity is moderate, although the correlation varies depending on marker gene and method used to derive the evolutionary distance. We have used both 31 universal marker genes and 16S sequences (results shown here). Figure 4 below shows the decrease in phenotypic similarity with increasing taxonomic rank.

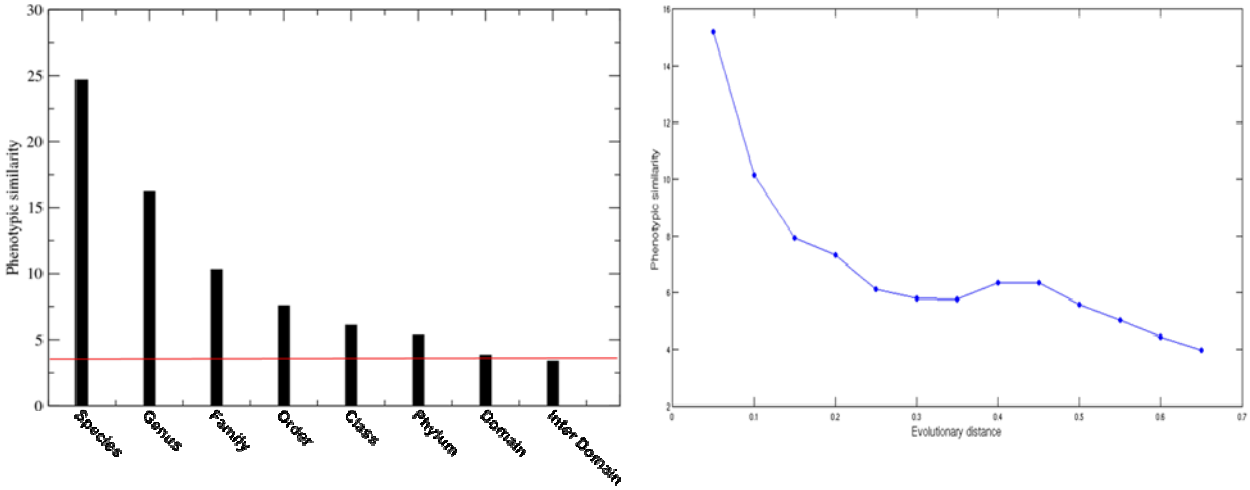

Figure 4 Log likelihood based phenotypic similarity between pairs of organisms, versus taxonomic rank (left), or evolutionary distance (right).

Some lineages are better preserved than others, such as for instance, the Proteobacteria are better preserved than the Firmicutes. Further, it appears that Archaea have a far better conserved overall phenotypic similarity, though this is likely due to poor sampling and less collective knowledge about the phenotypic traits of Archaea. Alternatively, the greater variance for well sequenced phyla may be the result of selective sequencing of strains that are phenotypically disparate. The most conserved phenotype in our study is methanogens.

Conservation of metabolic pathways with evolutionary distance

The analysis in the previous section can be equally well applied to pathway co-occurrence profiles, instead of phenotype profiles. We have applied the method on 788 pathways (see Figure 5), covering 343 organisms, collected from BioCyc (Caspi *et al*., 2006).

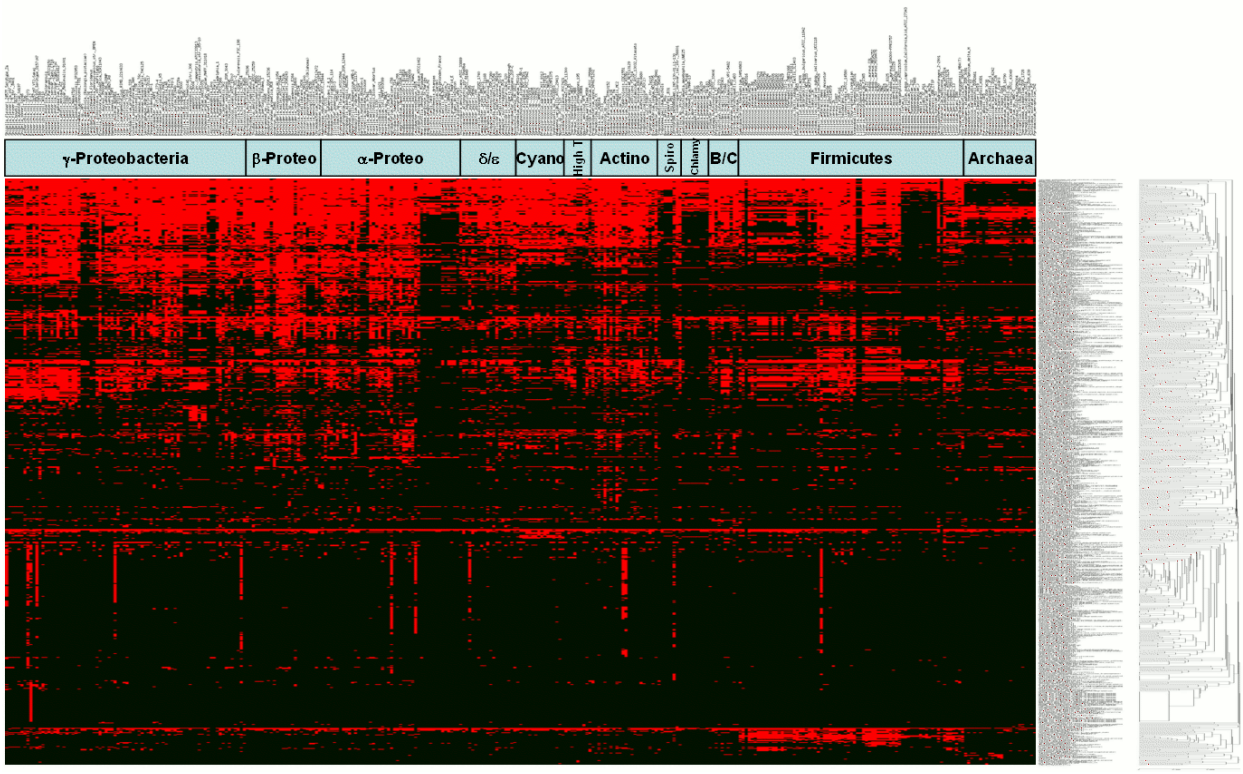

Figure 5. Pathway co-occurrence profiles for 343 prokaryotic genomes x 778 pathways. A red dot indicates the pathway is present in the BioCyc pathway genome database for that organism.

As with the microbial phenotypes, metabolic profiles show the highest degree of similarity between organisms of the same species or genus, and there is a moderate negative correlation between metabolic similarity and evolutionary distance, see Figure 6. The average pathway similarity is higher than for phenotype similarity, because there are a number of almost universal metabolic pathways, and, additionally, this dataset contains no missing values.

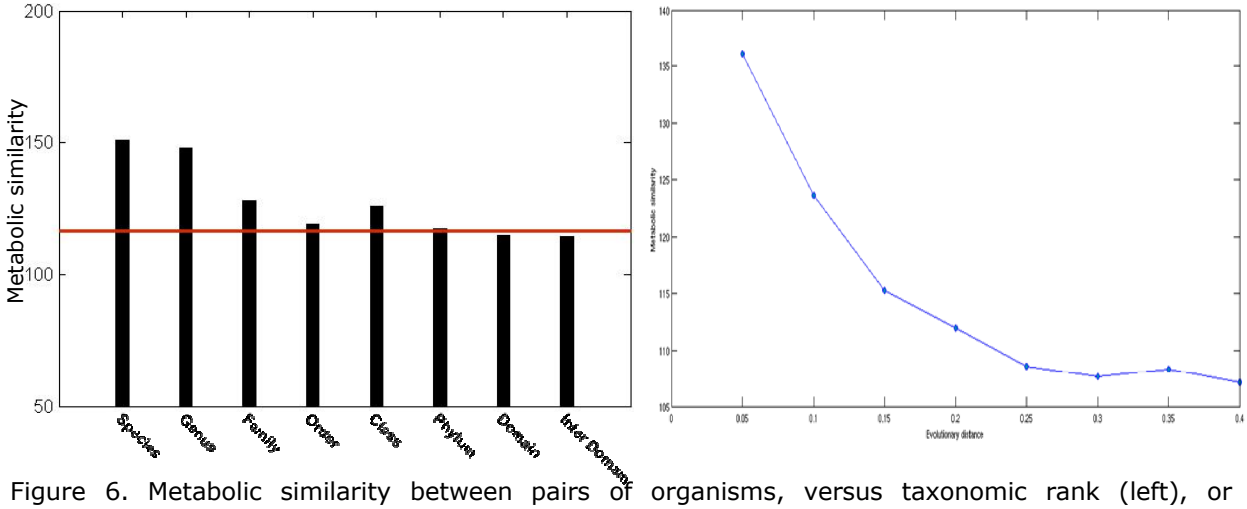

Figure 6. Metabolic similarity between pairs of organisms, versus taxonomic rank (left), or evolutionary distance (right).

Further, there are clear differences between lineages. Noticeably, some lineages, such as Cyanobacteria, Chlamydiae and Archaea, have a significantly higher metabolic similarity within the lineage compared to other phyla (Figure 7), although this may be in part because

their metabolic pathways remain less well explored. Furthermore, there are many different pathway conservation patterns, ranging from pathways that are only conserved at distances of 0.05 or less, while the core pathways are conserved to distances over 0.3 substitutions per site.

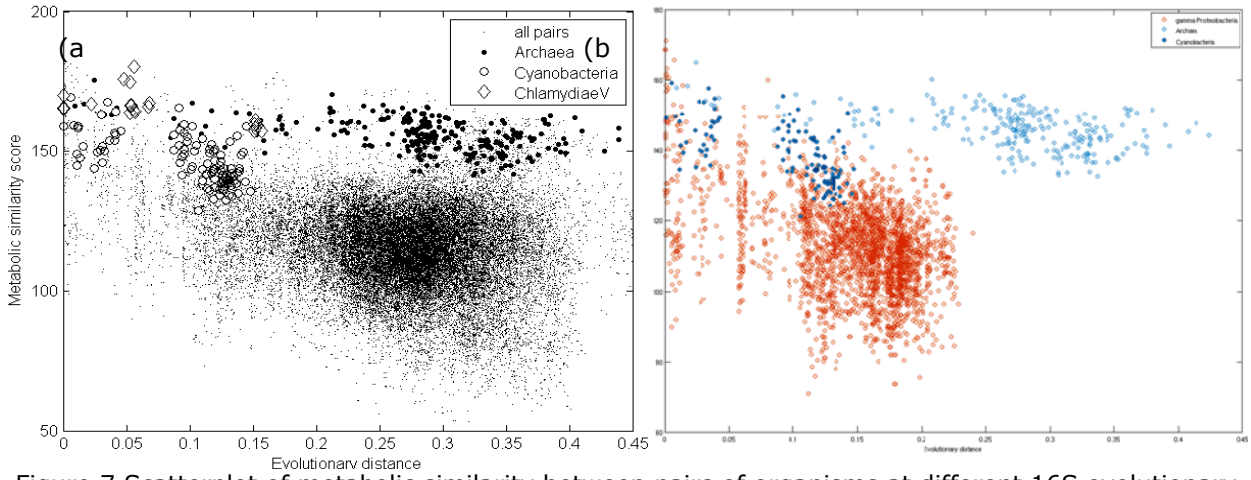

Figure 7 Scatterplot of metabolic similarity between pairs of organisms at different 16S evolutionary distance. Right: pairs of Archaea and Cyanobacteria show significantly higher metabolic similarity than γ-Proteobacteria at similar evolutionary distances.

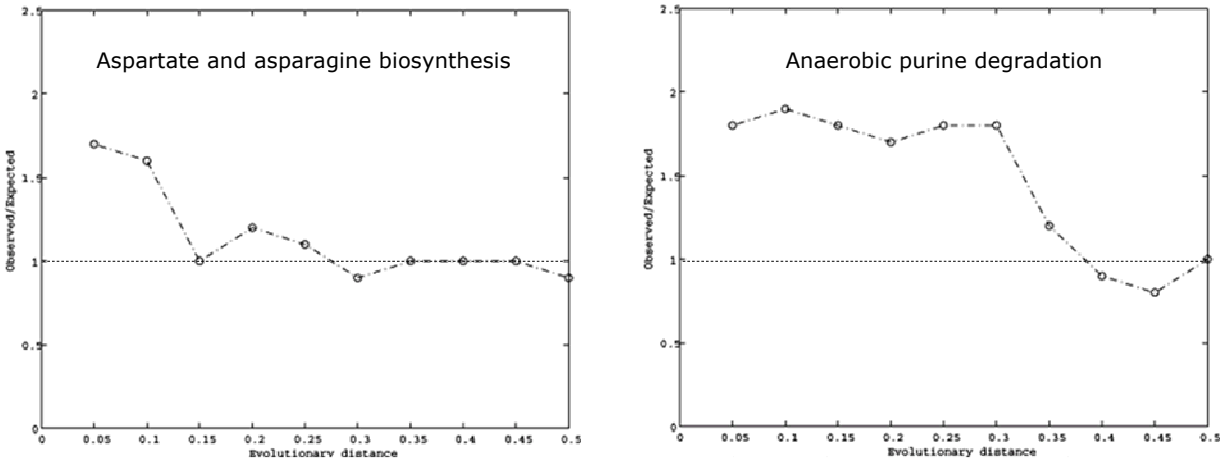

Figure 8. Some pathways are conserved over significantly greater distances than others.

Metabolic pathway prediction using Nearest Neighbors

We have developed a pathway prediction method based on nearest neighbors. Random sampling from the GreenGenes database indicates that >70% of randomly chosen 16S sequences have three nearest neighbors at a distance of 0.2 or less. Therefore, we may be able to derive a preliminary, probabilistic assessment of the metabolic capability of an organism, extrapolated from its closest relatives.

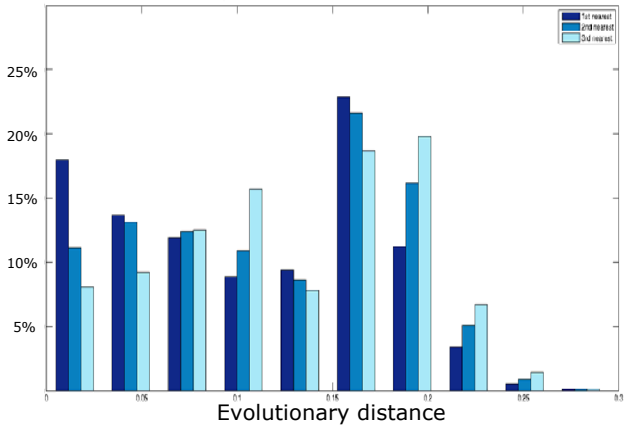

Figure 9. Fraction of organisms with 1, 2 or 3 sequenced nearest neighbors within a given evolutionary distance.

The microbial tree of life has been sequenced very unevenly, depending on ease of cultivation of the organisms, their significance as a human pathogen or model organism, etc. For example, there are currently 28 sequenced strains of E. coli, but only one other sequenced non-coli species of Escherichia. If we were to count all of these closely related E. coli strains with equal weight, then a standard Nearest Neighbors approach would predict E. coli pathways for any Escherichia species. To account for this, we down-weighted the nearest neighbors based on the local density of sequenced genomes, in such a way that *n* identical strains would each receive a weight of 1/*n*.

We can now calculate the pathway content of an unknown organism as a weighted sum of the pathway contents of its nearest neighbors with known pathway content. We multiply the pathway content of each organism with a distance kernel, such that closer neighbors are weighted more heavily than remote neighbors in the weighted sum. We have evaluated weighting according to an inverse exponential, square, or triangular function, and preliminary results suggest that the former is preferable.

To validate our predictions, we performed a leave-one-out-crossvalidation experiment on our entire metabolic pathway dataset, cycling through each of the 343 organisms, and predicting their metabolic pathway profile based only on their nearest neighbors. Figure 10 shows the results of the cross-validation predictions. Each pathway prediction is a value between 0.0 and 1.0, and van be directly related to the probability that the pathway in question is present in that organism. Figure 11 shows that for pathways which are known to be absent in a specific organism, the predicted value tends to be very close to 0.0 (blue histogram), while for the pathways which are known to be present, the predicted value tends to be very close to 1.0 (red histogram). If we set a classification threshold such that the number of predicted pathways is the same as in the original dataset, the method correctly predicts 80% of pathway presence and 90% of pathway absence (r=0.74).

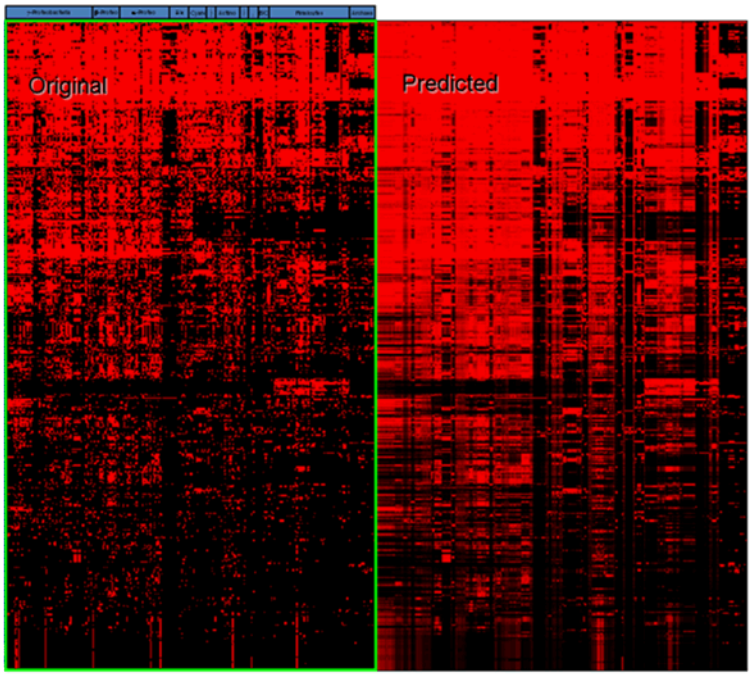

Figure 10. Original metabolic pathway dataset (left), *vs* predicted pathways based on cross-validation (right).

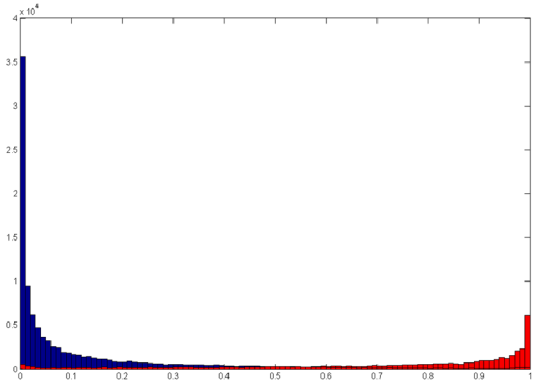
Figure 11. Histogram of predicted values for pathways known to be absent (blue) and present (red) in the dataset.

Planned improvements on this basic approach included:

1. As shown above, there is considerable variation in conservation between pathways. Therefore, we intended to include weighting according to individual pathways.

2. Similarly, certain lineages are better preserved than others. Hence, lineage variations can be

included in the calculation.

3. The genomes that have been partially manually annotated (Tier 2 in BioCyc) should be up-weighted in comparison to the automatically generated Tier 3 organisms.

4. Explore the possibility of weighing according to the reliability of the BioCyc pathway predictions. Ultimately, instead of pathway absence or presence, we would like to include the reliability scores for the presence of the pathways.

In addition to the cross-validation on known genomes explored above, we have also applied metabolic pathway prediction to sets of 16S sequences accompanying published metagenomics studies, with the intent to compare our predictions to findings derived from metagenomic sequence data. In particular, we have predicted the metabolic pathway content of the Sargasso sea, farm soil and whalefall datasets (Tringe, et al, 2005), see Figure 12.

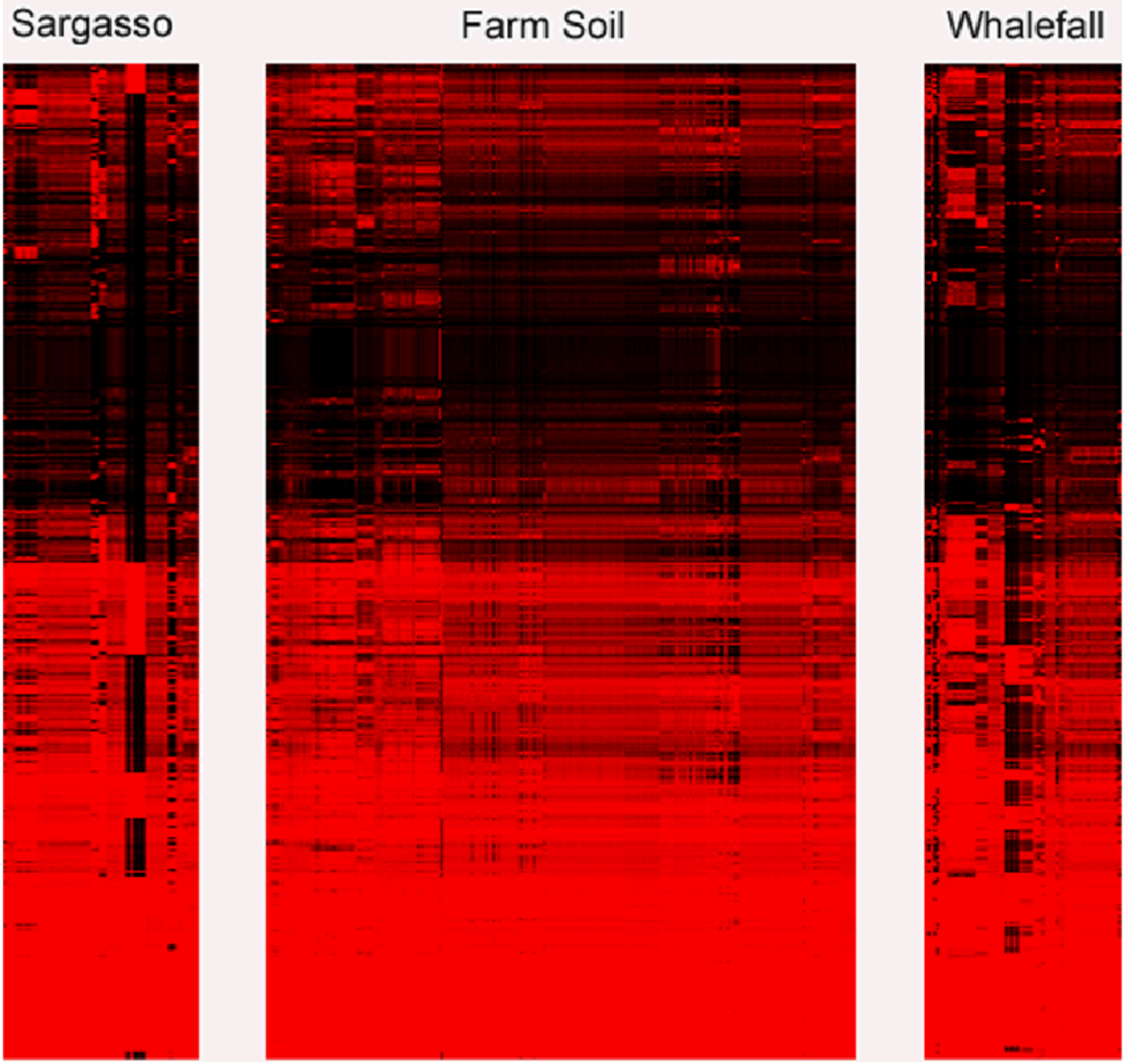

Figure 12. Metabolic pathway prediction for three real-world metagenomic datasets, based on their 16S rRNA community composition.

Metabolic pathway prediction using phylogenomics

Distance-based methods are known to exhibit certain types of errors, especially when the relative evolutionary rate between the closest related organisms varies. For example, in Figure 13, organism E may have organism C as closest neighbor in terms of evolutionary distance (sum of branch lengths), even though evolutionarily it is more closely related to D and F. By tracing the evolutionary inheritance of a trait, we can make the most likely or most parsimonious inference of the presence of the trait in novel organisms. Here, we use phylogenomic ancestral reconstruction to predict the presence or absence of known metabolic pathways in microbes.

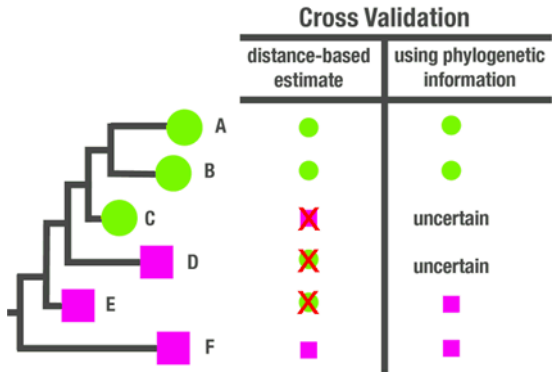

Figure 13. By tracing the evolutionary inheritence of a trait, phylogenomics can avoid some of the errors inherent with distance-based estimates.

To infer microbial metabolic capability, we created an algorithm, called meta-pup (metabolic pathway predictor, using phylogenomics). Meta-pup begins with four pieces of information: a list of known taxa, the metabolic pathways known to exist in those taxa, a list of novel taxa whose metabolic functionality is unknown, and a phylogeny expressing the evolutionary relationship between the known and novel taxa. For each of the targeted metabolic pathways, meta-pup uses a binary Markov model to reconstruct the presence/absence of that pathway in all internal (i.e. ancestral) nodes of the given phylogeny. The binary Markov model (see Figure 14) assumes a constant rate of loss ("death") or emergence ("birth") of metabolic pathways per unit of time. If we know the probability of presence/absence of a pathway at an ancestral node, the birth and death rates, and the branch length (evolutionary time) to a descendant node, we can calculate the probability of the pathway at that descendant node. Conversely, if we know probabilities at all the descendants of a node, we can infer the most likely probabilities at the ancestral node. The process of ancestral reconstruction therefore can be phrased as an optimization problem: we need to simultaneously optimize the birth and death rate parameters, and all ancestral probability estimates, to best fit the known presence/absence of the pathways at the leaf nodes. Using this ancestral reconstruction, Meta-pup then infers the probability of each pathway for each novel taxa.

$$p_{ij}(t) = \begin{bmatrix} \pi_0 + \pi_1 e^{-\mu t} & \pi_1 - \pi_1 e^{-\mu t} \\ \pi_0 - \pi_0 e^{-\mu t} & \pi_1 + \pi_0 e^{-\mu t} \end{bmatrix}$$

t = time interval
π_0 = death rate
π_1 = birth rate
μ = branch length scaling factor

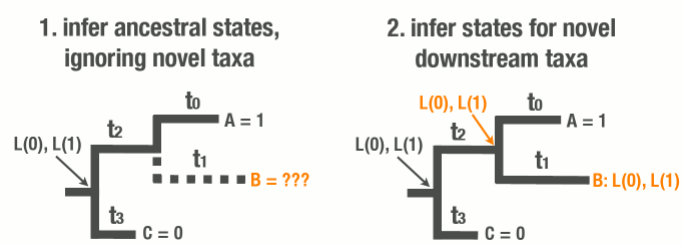

Figure 14. We first infer pathway probabilities at all internal nodes, by fitting a simplified birth/death model to the pathway content of all known taxa, then infer the probabilities for novel downstream taxa.

To explore the efficacy of this approach, we cross-validated meta-pup against 307 microbial species and 509 known pathways in the BioCyc database. Figure 15 shows the prediction error for each organism, mapped across the phylogenetic tree. Meta-pup's accuracy varies, depending on underlying phylogenetic accuracy, taxonomic sampling bias, and the quality of genomic annotation. A surprising result is that meta-pup shows usefulness for discovering poor or incomplete metabolic annotations. This especially shows up in cases where we have pairs of very closely related organisms with very different metabolic pathway content. In this case, the pathway prediction for each will be heavily influenced by the other member of

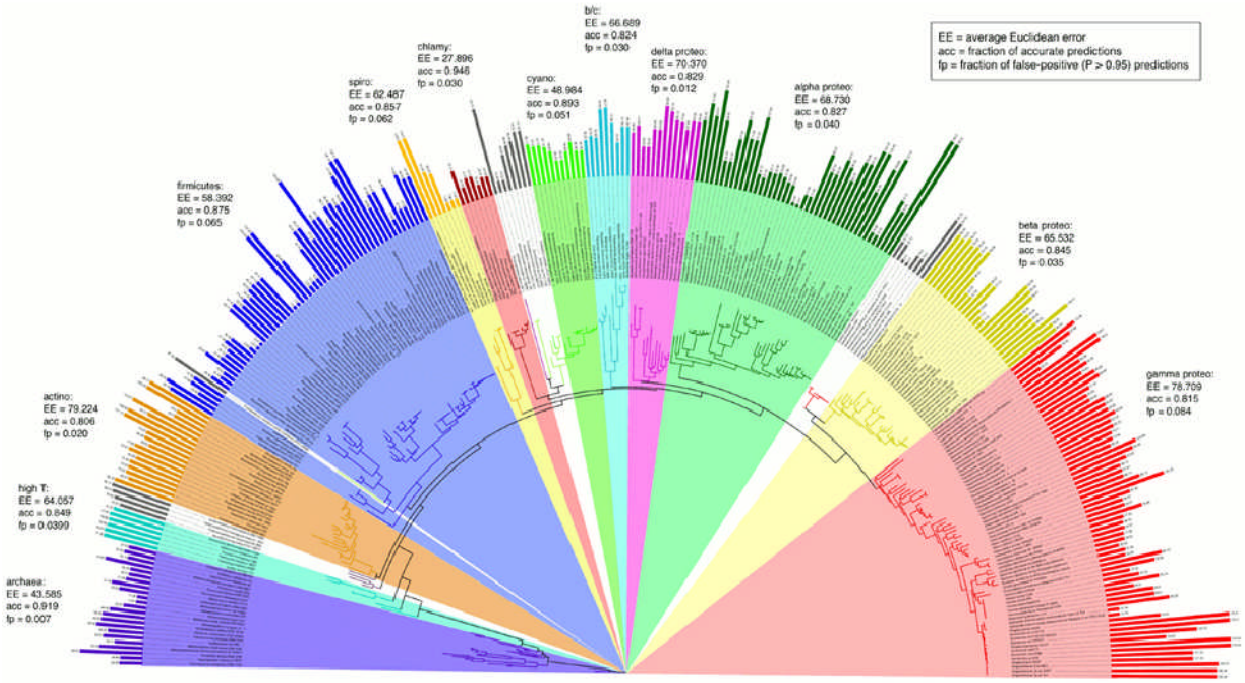

Figure 15. A phylogeny relating 307 known microbes. The colored wedges sort the microbes into taxonomic groups. The bars along the circumference indicate the Euclidean error of meta-pup's cross-validated predictions against 509 known metabolic pathways (with short bars indicating low error).

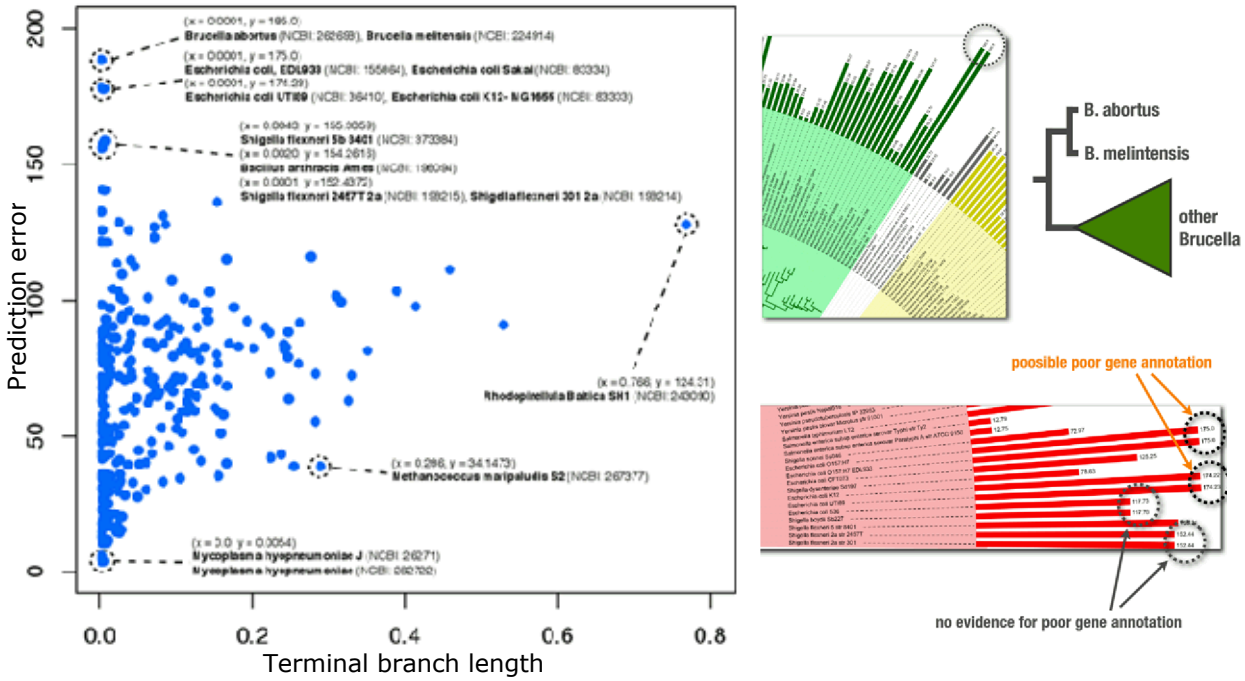

Figure 16. Prediction error versus terminal branch length in the phylogenetic tree. As expected, more remotely diverged organisms are harder to predict. The outliers at top left are due to inconsistencies in annotation quality between closely related strains. These outliers are also clearly visible in Figure 15, as pairs of organisms with very small evolutionary distance yet high prediction error (right).

the pair, so both organisms will be assigned a high prediction error (see Figure 16). At this point, the performance of this basic meta-pup algorithm is comparable to that of the Nearest Neighbor based method, although there room for significant improvement. Future work includes exploring alternative evolutionary models, and using meta-pup to make finer-grained predictions about the presence of individual enzymes within a given pathway.

Interface design for a standalone metabolic pathway prediction tool

With the help of Jason Conkey, we designed a standalone web tool, "metap" for metabolic pathway prediction from 16S rRNA sequence data, implemented primarily in Python, using the Cheetah templating package. 16S ribosomal RNA is a marker gene often used to characterize the phylogenetic position of an organism in the tree of life, and metagenomic studies of microbial communities are typically accompanied by sequencing of 16S sequences to characterize the species composition of the community. The user can upload or paste in a set of 16S sequences of organisms of interest, which need to be pre-aligned with the GreenGenes 16S phylogeny (e.g. using tools available at greengenes.lbl.gov). The system will then predict metabolic pathway probabilities for each pathway in MetaCyc and each organism of interest, and display the probability for each pathway that exceeds a specified threshold for at least one organism. Figure 17 shows the front page of the web-based tool, with entry fields for the 16S file and analysis parameters, as well as the resulst page with the pathway probabilities, grouped based on the pathway classes defined in BioCyc.

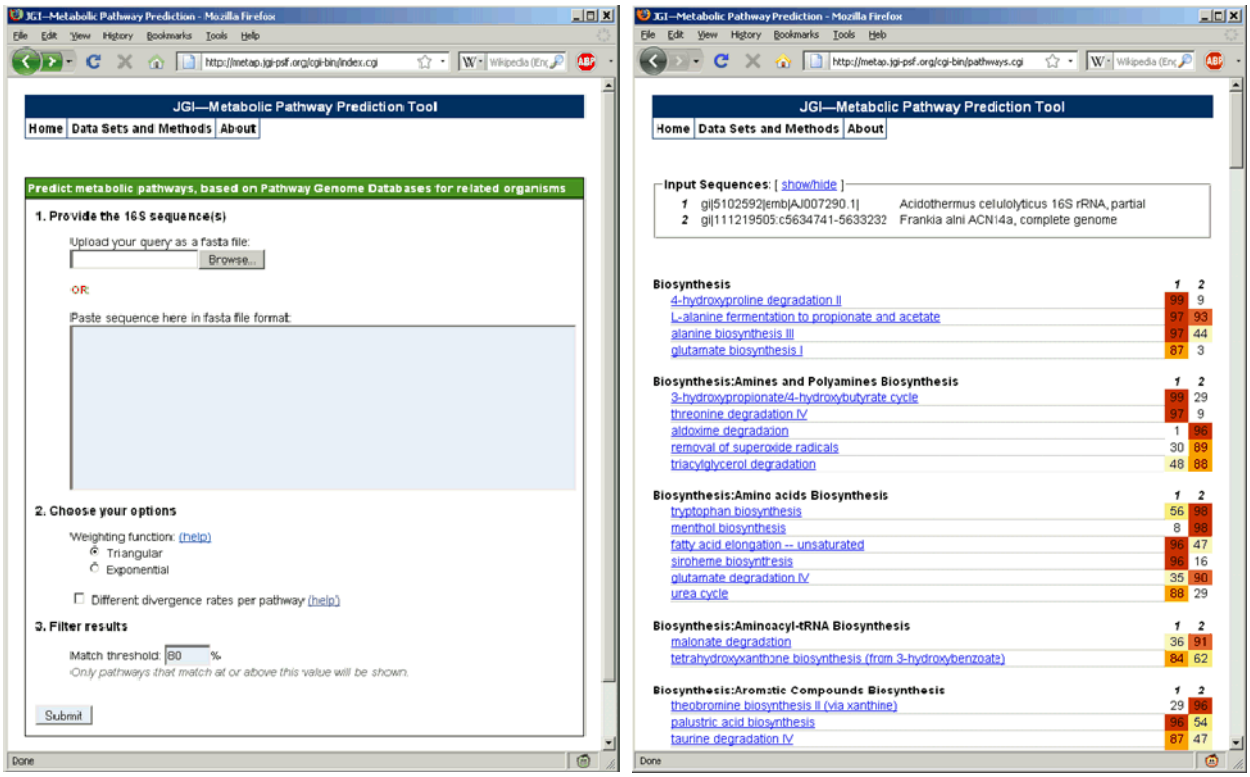

Figure 17. Interface for a web tool for metabolic pathway prediction based on 16S rRNA sequences.

At the moment, the web tool still uses a stub that generates random pathway probabilities, rather than the nearest neighbor or phylogenomic algorithms, although we intend to tie in the real algorithms soon.

Exit Plan

Due to the sudden and unexpected cancellation of the project after only one year and loss of the postdoc working on the project, there are a large number of loose ends which need to be tied up, without funding or personnel available to do so. The PI does intend to complete the web-based tool, although it may not be possible to add further refinements to the nearest neighbor and phylogenomic algorithms, which were expected to improve the performance significantly. We had started a manuscript on the conservation of phenotypic traits and metabolic pathways with evolutionary distance, although significant additional effort would be required for a high-profile scientific publication. Due to time restrictions, we may include some of this material in a Bioinformatics paper on the web tool instead.

We intend to salvage as much as possible of the other work done in this project, although sequence binning in phylogenetic space, as well as the planned integration of metagenomic binning, 16S-based prediction, and pathway prediction based on the metagenomic bins will likely require securing additional, reliable funding to complete.

Ever since this project was originally proposed in 2007, the need for these types of tools has become increasingly urgent. The number of metagenomic sequencing projects has more than doubled, following the exponential trend we initially predicted. Next generation sequencing technologies are poised to flood the field with an even greater volume of metagenomic and 16S sequence data. Completion of the tools developed in this project would have put us at the forefront of this rapidly growing field.

This research is directly in line with the goals of DOE/OBER's Genomics:GTL program, and would have positioned us to tackle issues related to bioenergy, bioremediation, carbon cycling and sequestration, mediated by metabolic processes in microbial communities. For example, our contribution to the DOE Joint Bioenergy Institute (JBEI) depends critically on our ability to analyze metagenomic and 16S community composition data of biomass degrading communities from compost and rainforest soil. Some of the research presented here may be continued under existing DOE GTL funding, either as part of LLNL's Scientific Focus Area in BioFuels, or as part of JBEI, although their support for development of computational analysis tools is limited. We will apply for additional funding within the framework of the planned GTL KnowledgeBase Center, which the PI has helped to plan.

We may be able to apply for follow-up funding, jointly with our current collaborators: (a) with JGI to continue developing metagenomics tools, and integrating these into the IMG/M portal, (b) with SRI to build further support for metagenomic datasets and multicellular metabolic models into the Pathway Tools software, or (c) with collaborators to further characterize their communities. Some of these continued collaborations might be funded through GTL, although other sources may be available as well. We already submitted one proposal in collaboration with Peter Karp at SRI to the NIH Human Microbiome Project last year, and we have established contacts with researchers at the USDA, UC Berkeley, and UC Davis, interested in studying a variety of microbial communities.

Summary

During the first year of our project, we established the feasibility of microbial species and gene sequences into a Euclidean space for metagenomic sequence binning, investigated the conservation of phenotypic traits and metabolic pathways with evolutionary distance, developed algorithms to predict metabolic pathways from 16S rRNA data based on Nearest Neighbors or phylogenomic approaches, and designed a web-based tool to make these tools available online.

The early cancellation of the project—despite excellent results and despite increasing need for such tools in the scientific community—has significantly impacted its potential, but we

intend to salvage as much of the research as possible and looks for external funding to continue development of these essential tools.

Acknowledgements

Most of the work presented here was accomplished in close collaboration between the PI and postdoctoral fellow Sara Light. We also thank Victor Hanson-Smith, who developed the "meta-pup" phylogenomic approach to metabolic pathway prediction during a very productive stay as a summer student, and Jason Conkey, who filled in to produce the web tool interface after Sara Light had to leave the project.

References

P.D. Karp, S. Paley and P. Romero (2002). The Pathway Tools software. *Bioinformatics* **18 Suppl 1**: S225-32.

R. Caspi, H. Foerster, C.A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S.Y. Rhee, C. Tissier, P. Zhang, and P.D. Karp (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **34**(Database issue): D511-6.

F.D. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, and P. Bork (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**:1283-7.

V.M. Markowitz, N. Ivanova, K. Palaniappan, E. Szeto, F. Korzeniewski, A. Lykidis, I. Anderson, K. Mavromatis, V. Kunin, H. Garcia Martin, I. Dubchak, P. Hugenholtz, and N.C. Kyrpides (2006a). An experimental metagenome data management and analysis system.. *Bioinformatics* **22**: e359-e367.

V.M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, and N.C. Kyrpides (2006b) The integrated microbial genomes (IMG) system. *Nucleic Acids Res*, **34**, D344–D34

V.M. Markowitz., et al., IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*, 2008. **36**(Database issue): p. D534-8.

M.L. Green. and P.D. Karp (2004). A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**: 76.

A.C. McHardy, H.G. Martin, A. Tsirigos, P. Hugenholtz and I. Rigoutsos (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*. **4**(1):63-72.

G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. **428**(6978):25-6

J. Raymond, B. Bebout, P. D'haeseleer, and T. Hoehler (2007) Connecting functional versatility and microbial diversity in a hypersaline microbial mat. *JGI Laboratory Science Program Small-Scale Sequencing Proposal*.

DeSantis, T.Z., *et al*., Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, 2006. **72**(7): p. 5069-72.