



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

FY07 LDRD Final Report Comparative Analysis of Genome Composition with Respect to Genotype-to-Phenotype Mapping and Metabolic Capability

P. D'haeseleer

February 11, 2008

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 05-ERD-065.

FY07 LDRD Final Report

Comparative Analysis of Genome Composition with Respect to Genotype-to-Phenotype Mapping and Metabolic Capability

LDRD Project Tracking Code: 05-ERD-065

Patrik D'haeseleer, Principal Investigator

Abstract

Given the glut of sequence data, comparative genomics methods are essential to efficiently leverage existing knowledge. However, most current approaches are limited to comparisons between closely related species. We study a large collection of bacterial genomes at the level of gene content rather than precise sequence similarity, allowing us to take advantage of sequence data from even remotely related species. By linking genome content to phenotypic traits across hundreds of fully sequenced microorganisms, we intend to elucidate genotype-to-phenotype mapping, with particular emphasis on metabolic processes.

Our modeling tools to decompose the genome composition include non-negative matrix factorization, linear and logit models, class association rule mining, support vector machines, and other machine learning techniques, validated against published data.

The patterns we discover in gene composition across the spectrum of bacterial genomes will increase understanding of which genes, gene classes, pathways, etc. are associated with or required for specific bacterial phenotypes, as well as yielding computational predictions of function for many unknown genes. Based on a list of genes in a newly sequenced genome (or even an unassembled environmental "shotgun" sequence), we expect to predict the metabolic processes, and how the organism fits into its environment, which will give us insight on modifying or exploiting the organism(s) in question. Such a predictive capability for genotype-to-phenotype mapping is crucial for analyzing the flood of new sequence data.

Introduction/Background

The pace of sequencing and the accumulation of new genes vastly outstrip the ability of human experts to hand-annotate the newly sequenced genomes. However, cross-genomic comparisons offer the potential to leverage existing knowledge, and increase our ability to make sense of this flood of data. Currently, most comparative genomics approaches are based on sequence conservation, limiting them to closely related organisms. In this project, we studied genomes primarily at the level of the collection of genes they contain – the "Bag of Genes" model – rather than their exact sequence of base pairs. This allows us to leverage off a much wider range of organisms, bypassing issues of sequence conservation, genome rearrangements, even sequence assembly and genomic

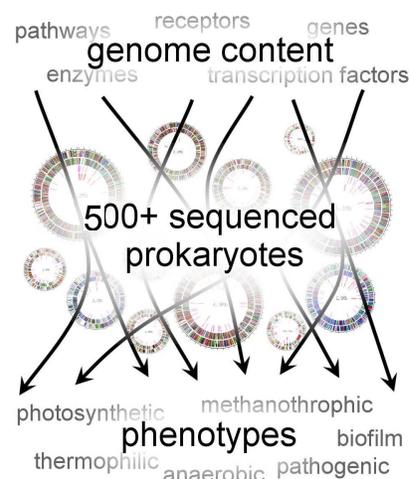


Figure 1. Finding patterns relating genotype to phenotype across 500+ sequenced prokaryotes

identity.

Genotype-to-phenotype mapping is the “holy grail” of 21st century biology. This is reflected in core aim of DOE’s Genomes to Life program to achieve predictive understanding of biological systems. The project presented here has the potential to make a substantial contribution to this aim. Our goal is to be able to take a newly sequenced genome, or even an unassembled set of environmental sequences containing genes from a whole community of species, and predict its behavior, metabolic capabilities, and thus eventually how this species or community interacts with and reacts to its environment.

This project is aligned with the Laboratory’s goals on wide range of levels: (1) The project directly addresses the core goal of DOE’s Genomes to Life program, to gain predictive mastery of the microbial world. (2) The project is closely aligned with the Laboratory S&T long range plans, especially with respect to the Systems Biology approach to the study of biological function and pathways, supporting LLNL's missions in homeland security, environmental assessment and management, and biosciences to improve human health. (3) It is in line with CMELS’s latest strategic planning, to develop transformational approaches to identify and characterize biological systems, as well as research in Systems Biology that will lead to a fundamental understanding of the energy metabolism in microbes and microbial communities. (4) The project also fits within the Computational Directorate’s latest Computational Biology strategic planning, by enabling the rapid identification, characterization, and exploitation of biological mechanisms through high-throughput bio-informatics, and (5) takes advantage of the Laboratory’s computational resources, as well as the sequencing and annotation capabilities available at JGI.

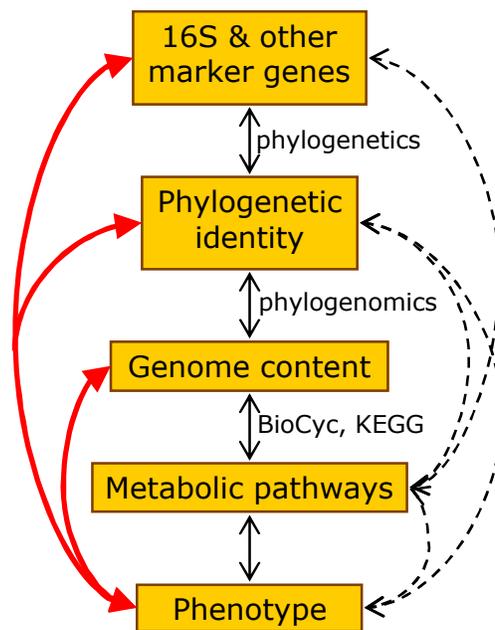


Figure 2. Red arrows: this study. Dashed: opportunities for future research, enabled by our current work.

Research Activities

FY05 Accomplishments and Results

After a mid-year start, we began gathering and organizing genotype and phenotype data on a large number of microbial species, initially focusing on integrating the European Molecular Biology Laboratory’s Search Tool for the Retrieval of Interacting Genes/Proteins database (110 species by 20,000 clusters of orthologous groups) with phenotypic descriptors from the Institute for Genomic Research’s genome properties. An initial decomposition of this data showed some groups of genes that are specific for (or specifically absent in) certain phylogenetic lineages. More interestingly, we also identified some components showing very strong correlations with phenotype rather than phylogeny, including growth temperature, human pathogenicity (obligate parasites), animal pathogenicity (sporeforming), methanogenesis, and photosynthesis. We also identified a number of promising additional sources for phenotype data.

FY06 Accomplishments and Results

In FY06, we (1) assembled a BioWarehouse database that encompasses genomes, taxonomy, bacterial genes and proteins, metabolic pathways, and phenotypes; (2) created a statistical phenotype model that uses genomic data to predict whether a genome represents a human

and/or animal pathogen, and its optimal growth temperature, outer membrane type, type III secretion, and type of flagella; (3) examined sets of genes linked to specific phenotypes, as well as a gene set involved in determining both optimal growth temperature and pathogenicity of a bacterial pathogen; and (4) began gathering transcription factor data from Genbank and other databases. Mapping of transcription factor binding sites in *Escherichia coli* have been delayed because of the departure of a collaborator.

FY07 Accomplishments and Results

In FY07, we (1) integrated a comprehensive database of microbial genotypes and phenotypes, containing >15,000 observations across 559 microbial organisms, including disambiguation of organism identifiers between datasets, (2) developed logistic regression models for probabilistic phenotype prediction, (3) developed a novel phenotype method for genotype-phenotype analysis based on class association rule mining method (4) analyzed 6 gene sets for phenotypes of interest; (5) adapted prediction to metagenomic data; (6) submitted a paper on genome decomposition by CAR mining; (7) developed a novel method to extrapolate phenotypes from evolutionary distance.

Key collaborator for ChIP-on-chip experiments left, leading us to drop the regulatory analysis in favor of the phenotype and metabolic analysis.

Overall, we achieved exciting results on microbial phenotypes, despite reduction in scope due to PD hiring delays. One paper submitted to a top journal, second one in preparation, 2 more were planned before premature cancellation of the project.

Results/Technical Outcome

Collection and integration of genotype and phenotype data

We have built a comprehensive database of microbial genotypes and phenotypes. The genotype database consists of the STRING database (von Mering *et al.*, 2003), and both the BioCyc and BioWarehouse projects from SRI International (Karp *et al.*, 2005; Lee *et al.*, 2006). The BioCyc project consists of 322 pathway genome databases (PGDBs) derived from annotated microbial genomes. BioWarehouse is the database back-end for the DARPA BioSPICE project, and incorporates several bioinformatic datasets into one comprehensive relational schema, including TIGR CMR, ENZYME DB, GenBank, Gene Ontology (The Gene Ontology Consortium, 2004), KEGG (Kanehisa and Goto, 2000), NCBI's Taxonomy DB, BioCyc, and UniProt.

The incorporation of data sources that identify organisms by various means, identifiers, and levels of specificity has led to a need to verify and disambiguate their assignments. We have developed a pipeline that analyses a candidate dataset, and maps the organism identifiers to specific and unambiguous identifiers in our database.

We assembled phenotype data of microbial organisms from a number of different publicly available sources, summarized in the table below. In total, this collection covers 559 sequenced strains, and 752 phenotypes, for a total of 15,011 individual phenotype annotations! Figure 3 shows the structure of the resulting phenotype database, and its links with the various data sources and genotype datasets.

- CMR – Comprehensive Microbial Resource (TIGR)
- GOLD – Genomes Online Database
- NCBI – Genome project
- Tavazoie – 6 curated phenotypes (Slonim, 2006)
- Minor datasets: PUMA2 (trophic ecology levels), PGTdb (growth temperatures)

- Manual curation by our group

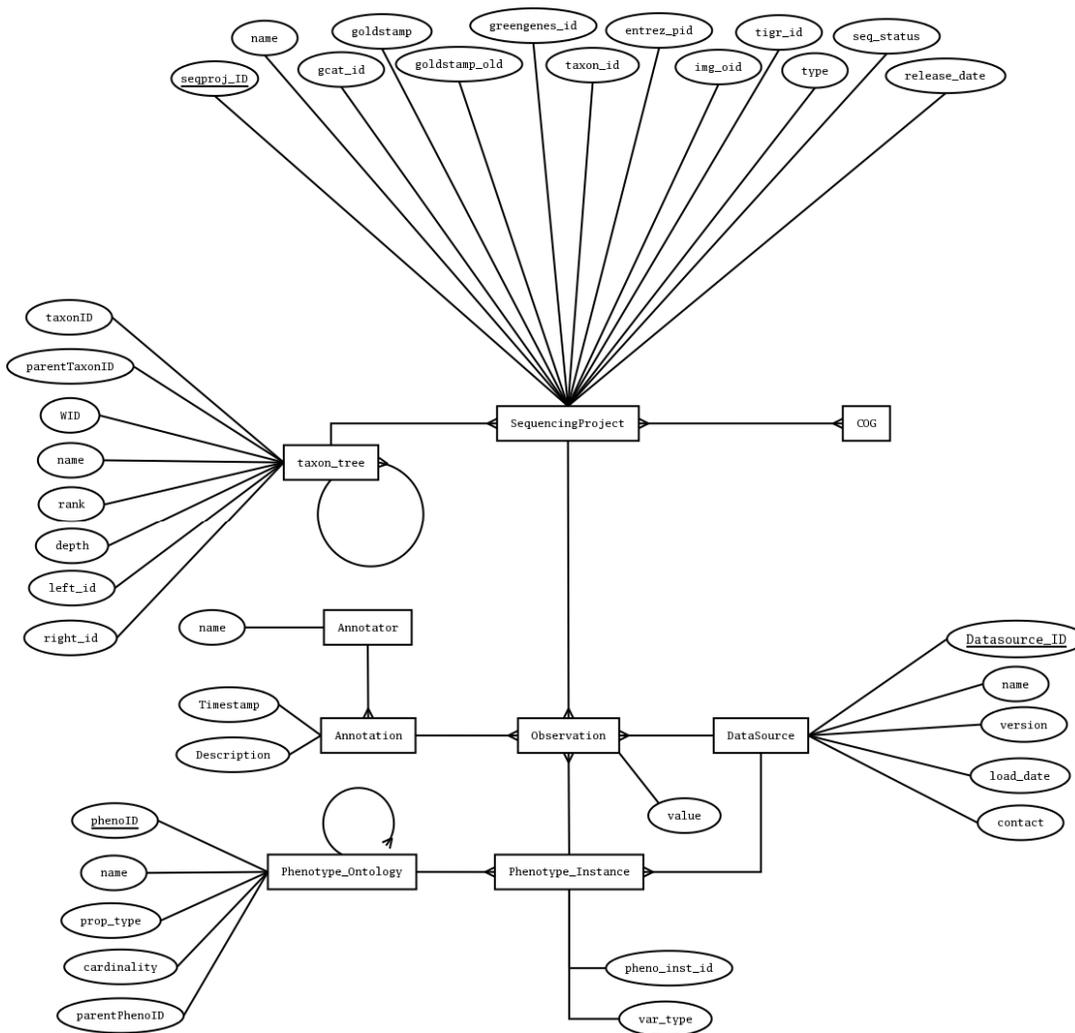


Figure 3. Structure of the phenotype database

These databases contain phenotypes of mixed relevance and quality. First, the perhaps greatest obstacle to be overcome is the extant lack of a controlled phenotype vocabulary. Even the NCBI database uses freestyle entries rather than NCBI's own taxonomic ID's to indicate the host range of pathogens. For instance, "Human pathogen" exist as the following entries: pathogen host: Homo Sapiens - Human - Humans - Hunan. Secondly, definitions for the phenotype traits are usually missing, or inconsistent. For example, one might be tempted to interpret the "Animal pathogen" phenotype in NCBI as "prokaryote with the ability to act as a pathogen in at least one animal". However, even this simple category is differently defined in each database. As can be seen from Figure 4, several human pathogens are not annotated as animal pathogens in the NCBI database and indeed "Animal pathogen" here actually means "Pathogen in at least one animal other than Human/Primate/Mammal". Note that the overlap between databases is sometimes small, see Figure 5, although there are relatively few outright disagreements.

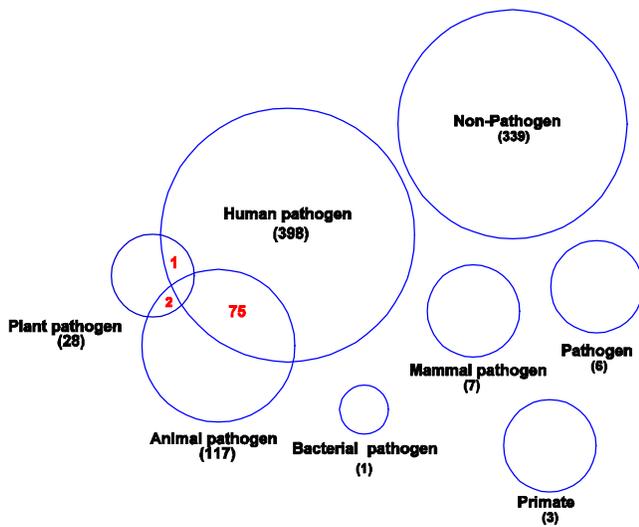


Figure 5. The overlap between different phenotypes within the NCBI phenotype database.

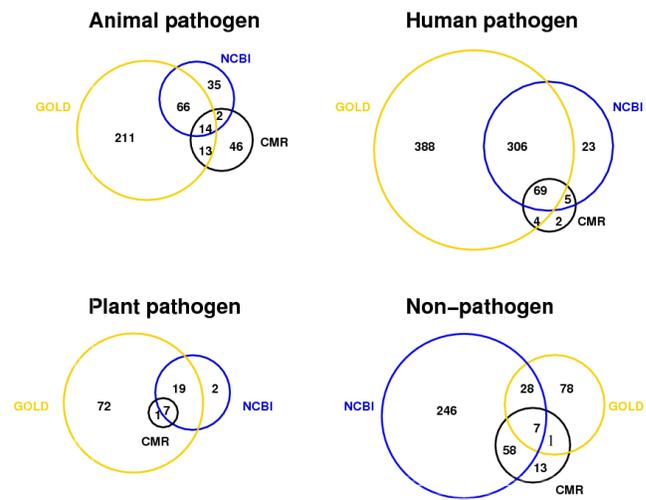


Figure 5. The overlap between different databases (Gold=yellow, NCBI=blue, CMR=black) for the same phenotypes

Another example of inconsistent definitions is for the temperature classes: psychrophile, mesophile, thermophile, and hyperthermophile. As Figure 6 shows, no two authoritative sources we consulted used the same temperature boundaries between temperature classes, resulting in multiple inconsistent annotations.

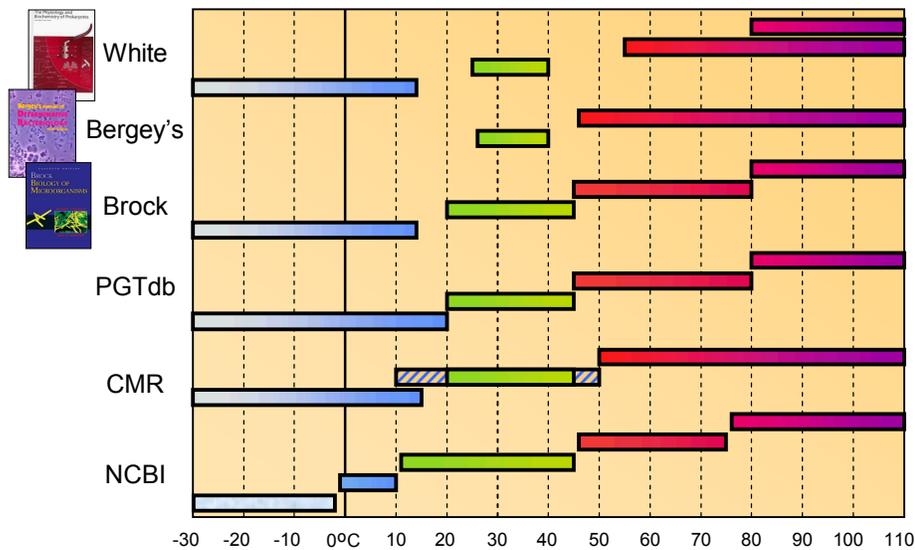


Figure 6. Inconsistent definitions in temperature classes psychrophile – mesophile – thermophile - hyperthermophile. CMR and NCBI are our main sources for temperature phenotypes, supplemented by data from the dedicated Prokaryotic Growth Temperature database (PGTdb). White, Bergey’s and Brock refer to standard microbiology reference works.

We constructed a composite database consisting of the three primary datasets (GOLD, NCBI and CMR), complemented by smaller datasets obtained from the literature. In the process, we corrected numerous annotation errors by manually curating disagreements between the datasets. We augmented the dataset where this was possible to do some with minimal effort. For example, we assigned taxonomic ID's to pathogen hosts, enabling the derivation of useful new categories such as "vertebrate pathogen". Further, 234 phenotypes, deemed to be clearly defined and of sufficient biological interest, were extracted and merged into one dataset consisting of 79 phenotypes, see Figure 7. Note that most phenotypes are very sparsely annotated. Another issue for our analysis is the relative lack of *negative* information: the two largest data sources, GOLD and NCBI only provide positive instances of phenotype observations, but omit, for example, information on negative results of experimental phenotype tests (e.g. a negative result on a standard sporulation assay).

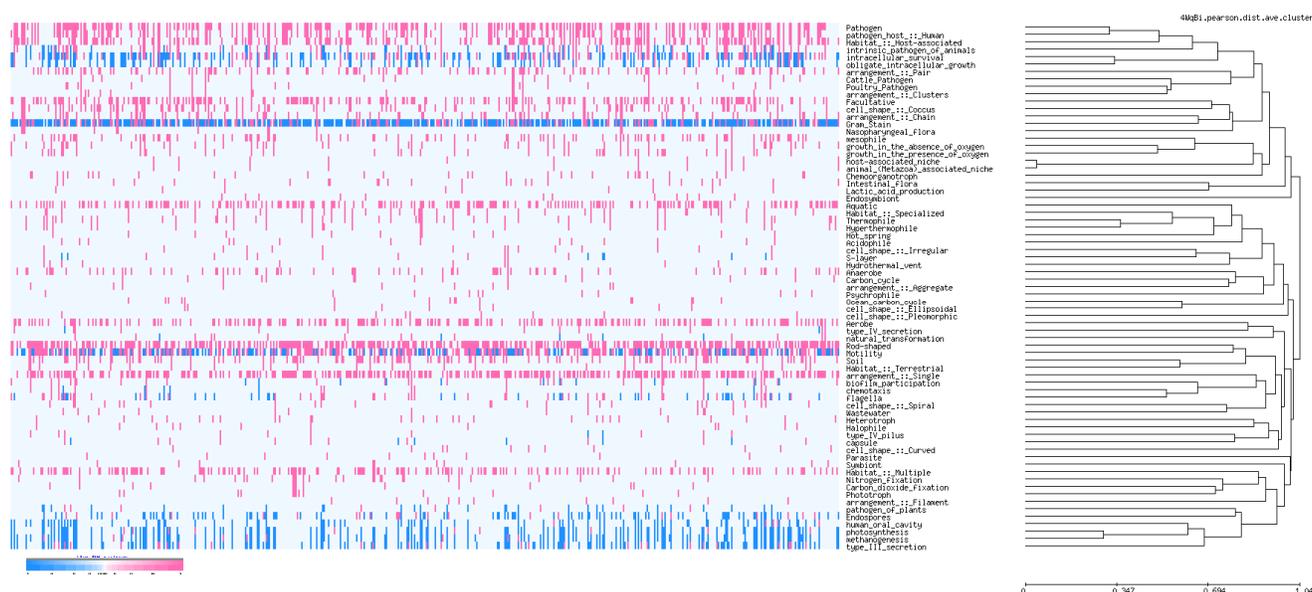


Figure 7. Presence (red) or absence (blue) of 79 phenotypes, across 559 sequenced prokaryotes. Light blue indicates no annotation is available for this combination of organism and phenotype (“don’t know”).

Genome decomposition using non-negative matrix factorization

In previous work on this project, we had performed a decomposition of the STRING database (von Mering *et al.*, 2003), containing 110 fully sequenced species and their genome composition with respect to 20,000 Clusters of Orthologous Genes (COGs, Tatusov *et al.*, 2001). The 110 x 20,000 STRING data set was decomposed automatically into 37 components using Non-Negative Matrix Factorization (Lee and Seung, 1999) – see Figure.8. Each component contains a specific subset of COGs that tend to segregate together across species (for example, all genes involved in

a specific pathway), and each species is modeled as a specific mixture of these components.

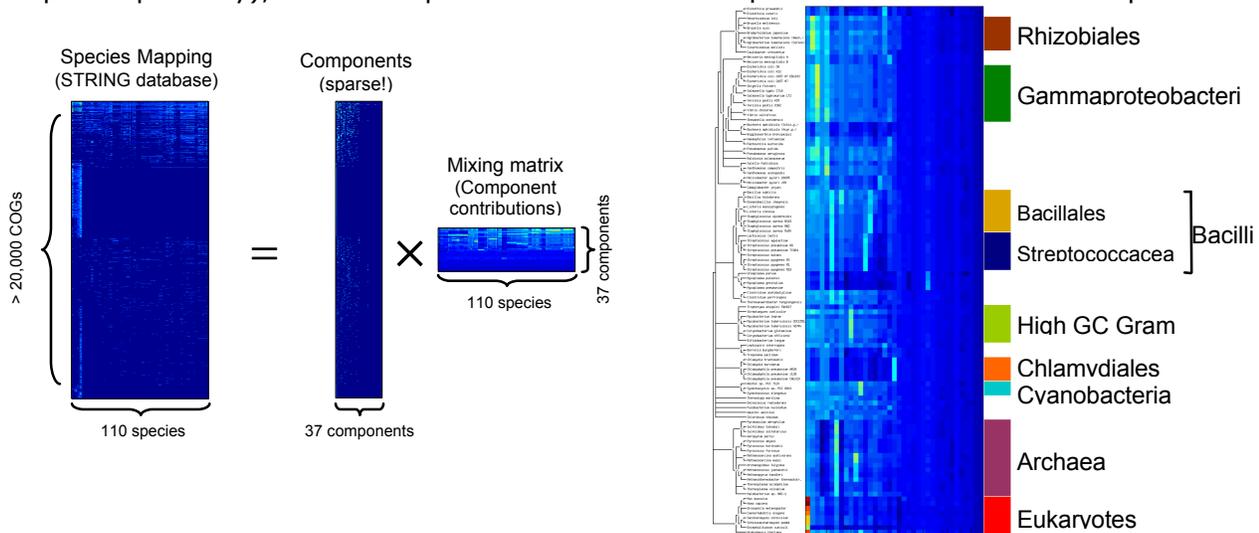


Figure 8. Decomposition of genome content using Non-Negative Matrix Factorization

A number of these components show significant correlations with phenotypes of interest. We can then design a classifier based on these components to predict the phenotype. One approach, which preserves a reasonably concise biological interpretation, is to build a regression model with stepwise variable selection, i.e. we take a weighted sum of a small number of components, and threshold the result to predict the phenotype. However, rather than thresholding, we can also attempt to predict the probability of the phenotype, by mapping the linear regression into probability values between zero and one, e.g. using a logistic function (logistic regression). A number of phenotypes can be predicted fairly well using this method, including human pathogens (misclassification rate = 16.3%, based on leave-one-out cross-validation) and outer membrane type (misclassification = 3.3%) – see Figure 9. Further work is needed to generalize these results to more phenotypes, and to understand the biological basis of these correlations.

The sparsity of many phenotypic categories still presents a hindrance to be able to achieve significant statistics of genotype-phenotype correlation, justifying our continued efforts to consolidate and unify phenotype datasets from different sources.

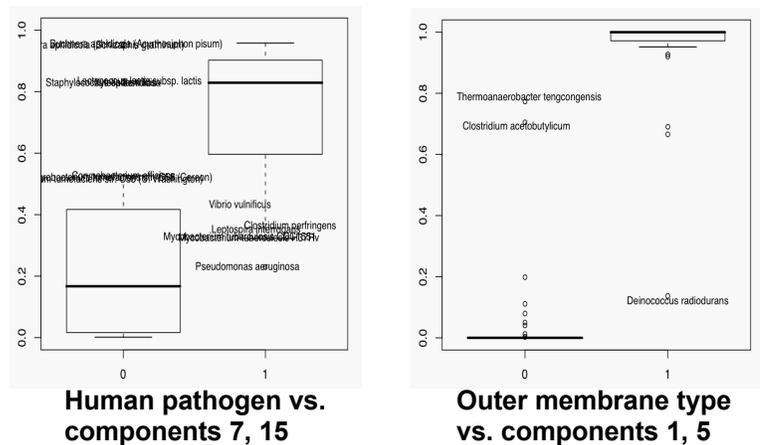


Figure 9. Prediction of "human pathogen" and "outer membrane type" phenotypes, based on selected components

Genotype-Phenotype Mapping by Class Association Rule Mining

Decomposition by NMF tends to result in components consisting of hundreds of genes. In order to be able to better pinpoint which genes are relevant to the phenotype, we explored an alternative approach, based on feature selection, followed by machine learning methods to extract combinatorial relationships between COGS and the phenotype.

Microbial phenotypes are typically due to the concerted action of multiple gene functions yet the presence of each gene may have only a weak correlation with the observed phenotype. Hence, it may be more appropriate to examine co-occurrence between sets of genes and a phenotype (many-to-one) instead of pairwise relations between genes and the phenotype. We propose an efficient Class Association Rule mining algorithm (Agrawal *et al.*, 1993), NetCAR, in order to extract sets of COGs (Clusters of Orthologous Groups of proteins) associated with a phenotype from COG phylogenetic profiles and a phenotype profile. NetCAR takes into account the phylogenetic co-occurrence graph between COGs to restrict hypothesis space, and uses mutual information to evaluate the biconditional relation.

We examined the mining capability of pairwise and many-to-one association by using NetCAR to extract COGs relevant to six microbial phenotypes (aerobic, anaerobic, facultative, endospore, motility, and Gram negative) from 11,969 unique COG profiles across 155 prokaryotic organisms. With the same level of False Discovery Rate (FDR), many-to-one association can extract about 10 times more relevant COGs than one-to-one association. We also reveal various topologies of association networks among COGs (modules) from extracted many-to-one correlation rules relevant with the six phenotypes; including a well-connected network for motility, a star-shaped network for aerobic, and intermediate topologies for the other phenotypes. NetCAR outperforms a standard Class Association Rule mining algorithm, CARapriori, while requiring several orders of magnitude less computational time for extracting 3-COG sets.

We developed a new class association rule mining algorithm, NetCAR that extracts many-to-one relationship between COGs and a phenotype with interest from a COG phylogenetic and the phenotype profile. NetCAR is much more efficient than standard CAR mining algorithm, CARapriori in computational time. The many-to-one association rules with stringent False Discovery Rate level for aerobic, anaerobic, facultative, endospore, and Gram staining phenotype contain significantly larger numbers of COGs than those by pairwise methods. We compiled association network from extracted 3-COG rules and revealed that the network can not only have Clique, for which previous pairwise methods implicitly assumed, but also Star type topology that contains large number of COGs whose occurrence is only weakly correlated with a phenotype observation. These results indicate that a gene module can be a combination of genes that span some depth in a biological network from a layer where we can see strong pairwise association. NetCAR algorithm is a powerful CAR mining algorithm to extract relevant entity (COG) with an observation (Phenotype) that cannot be elucidated by pairwise comparison. We also discuss the phenotype prediction capability of the extracted rule in supporting material. It is often the case in biological data that dimension (in our case, COGs) is much larger than samples (genomes), and the NetCAR algorithm may be also appropriate to extract for such cases. For example, NetCAR may also be applicable to mine co-regulatory gene network module relevant with a target physiological observation, from microarray data with many more genes than expression arrays.

A manuscript describing this work (Makio Tamura and Patrik D'haeseleer, Microbial Genotype-Phenotype Mapping by Class Association Rule Mining) is currently under revision for resubmission to *Bioinformatics*. Additional results and details of this approach can be found there.

Phenotype prediction for genomes with incomplete sequence coverage

Until a few years ago, almost all our knowledge about microbial genomes came from whole-genome sequencing of cultivated strains. However, comparisons between counts of microbial species by microscopy versus culturing plates indicate that the number of organisms we can study by cultivation is only a fraction of a percent of the total microbial diversity around us. With the advent of metagenomic shotgun sequencing of DNA extracted from environmental samples, and whole-genome amplification from single cells or small clusters of cells, we can now access the vast unexplored diversity of microbial genomes. However, the genomes recovered from these novel sequencing approaches are typically incomplete, either because of low abundance within

the microbial community, leading to a fractional sequencing coverage, or because of limitations of whole genome amplification from extremely small amounts of starting material.

Phenotype classifiers based on a small number of highly relevant genes, such as the Class Association Rules developed in the previous section, are very sensitive to lack of genome coverage. However, in the previous section we also developed a Support Vector Machine classifier, as a baseline "black box" model to predict phenotypes based on the entire genotype profile of the organism. A classifier such as this, where the classification decision is distributed across thousands of genes, is likely to be far more robust to missing data.

A Support Vector Machine (Vapnik, 1998) using a linear kernel finds an optimal linear separatrix between the positive and negative instances of the phenotype:

$$score_i = \sum_j g_{ij} w_j - b \quad (1)$$

$$\text{predict : } \begin{cases} phenotype = \text{YES for genome } i, \text{ iff } score_i > 0 \\ phenotype = \text{NO for genome } i, \text{ iff } score_i < 0 \end{cases}$$

where g_{ij} is a binary variable indicating the presence or absence of gene (COG) j in genome i , w_j indicates how much the presence of gene j contributes to the prediction of the phenotype, and b imposes a threshold that the summation in the formula above needs to exceed in order to predict the presence of the phenotype.

If we have incomplete genome data, some of the binary variables g_{ij} will be zero, i.e. the gene has not been observed, even though it does occur within the genome of the organism. Note that in this case, we cannot simply apply the classifier above without modification. For example, if we only have 50% of the complete genome, the expected value of the summation above will be only 50% of what it would be for the full genome:

$$E_{50\%} \left[\sum_j g_{ij} w_j \right] = 50\% \sum_j g_{ij} w_j$$

$$\Rightarrow E_{50\%} [score_i] = 50\% score_i - 50\% b$$

In other words, if we blindly apply the SVM classifier to an incomplete genome, we will get a score which is *lower* (if $b > 0$) than we would expect for the full genome, leading us to under-predict the phenotype.

If we could get rid of the constant threshold b , and distribute it across the weights w_j for the individual genes, we would achieve a classifier in which the expected score for a fractional genome is equal to that same fraction of the score of the full genome. In this case, if the full genome scores positive (phenotype = YES), the fractional genome will be expected to score positive as well:

$$score_i = \sum_j g_{ij} (w_j - c_j) \quad (2)$$

$$= \sum_j g_{ij} w_j - b, \text{ iff } \sum_j g_{ij} c_j = b \quad (3)$$

If we choose c_j such that (3) holds for all the reference genomes for which we know the phenotype classification, we will have achieved a classifier which is equivalent to the original SVM classifier on the training genomes, but which yields consistent phenotype classifications for fractional genomes. The choice of c_j is underdetermined, leading to a range of possible phenotype classifiers for incomplete genomes, which will all perform equally well on the complete training genomes. Below, we show the performance of the family of classifiers based on the pseudo-inverse of g_{ij} , which is the solution to (3) which minimizes the norm of the vector c_j . (Another

interesting choice might be to minimize the norm of the vector $(w_j - c_j)$, which would spread the weights more equally across all genes, possibly resulting in an even more robust classifier.)

Figure 10 shows the performance of the phenotype classifier on the six microbial phenotypes used in the previous section (see Slonim *et al.*, 2006), across 155 different organisms, with genome coverage from 100% down to as little as 1%. There are 62 aerobic, 31 anaerobic, 42 facultative, 11 endospore forming, 76 motile, and 95 Gram negative organisms in our dataset, out of a total of 155 organisms. As above, genome content is described as a binary vector indicating which genes (COGs) occur in each genome. For each phenotype, we perform 25-fold cross validation to get an unbiased estimate of the performance of the classifier. For each cross-validation run, we set aside $1/25^{\text{th}}$ of the organisms as a test set, train the SVM classifier and derive the new partial-genome classifier (2) on the remaining organisms, then test the performance on the organisms which were set aside. We calculate the average *Recall*, *Precision* and *F1* score as follows:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$F_1 = 2 \frac{Recall \times Precision}{Recall + Precision}$$

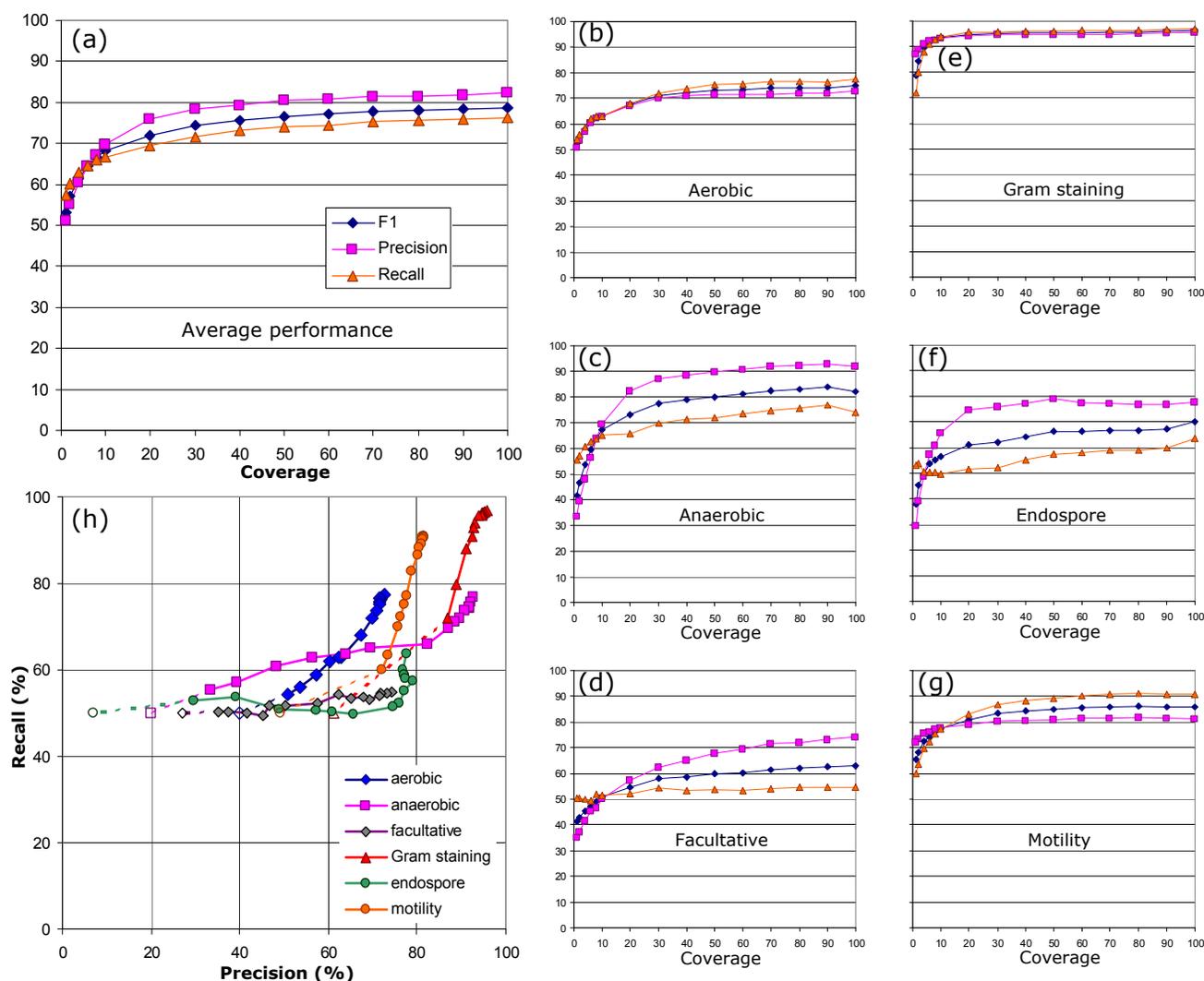


Figure 10. Performance of classifiers with incomplete genome coverage, for 6 selected phenotypes.

Figure 10(a) shows that the performance of the partial-genome classifier is affected only minimally compared to the whole-genome SVM classifier, down to 20-30% coverage. Figure 10(h) shows that *Recall* typically degrades faster than *Precision* (a desirable behavior if we want to make conservative phenotype predictions), eventually dropping to around 50%. Genome coverage can be reduced as low as 8% (for facultative anaerobic) to less than 1% (for Gram staining) before the performance of the classifier drops by 50%, compared to a hypothetical random classifier with 50% *Recall*.

These preliminary results are highly encouraging for the applicability of these types of classifiers to predict phenotypic traits of incompletely sequenced novel organisms from metagenomic sequencing surveys. More research will be needed to examine the effect of different choices of the weights c_j in equation (3). In particular, we expect that different choices of weights may lead to different tradeoffs between *Recall* and *Precision* (Figure 10(h)), or different degrees of robustness against missing data.

Assessment of phenotypic similarities between prokaryotes

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.

The 16S rRNA sequence alignments for all fully sequenced genomes were extracted from the GreenGenes database (DeSantis *et al*, 2003) and the evolutionary distances between the 488 organisms whose phenotypes were included in our composite phenotype database were calculated using ClustalW (Chenna *et al*, 2003). 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).

We have devised a phenotypic similarity measure between organisms, based on mutual information, $I(A;B)$, the information contained in A that remains when B is already given.

$$I(A;B) = H(A) + H(B) - H(A,B)$$

where I is mutual information, A and B are two vectors, and H(A) is the entropy of A.

$$H(A) = -\sum \Pr(A) \log \Pr(A)$$

Because of the large number of missing phenotype annotations, the mutual information between the phenotype vectors for two organisms is only calculated across that subset of phenotypes for which both organisms have annotations (either presence of the phenotype, or an explicitly annotated absence of the phenotype).

The correlation between evolutionary distance and phenotypic similarity is moderate (Figure 11(a)), although the correlation varies depending on marker gene and method used to derive the evolutionary distance. As an alternative measure, Figure 11(b) below shows the decrease in phenotypic similarity with increasing taxonomic rank.

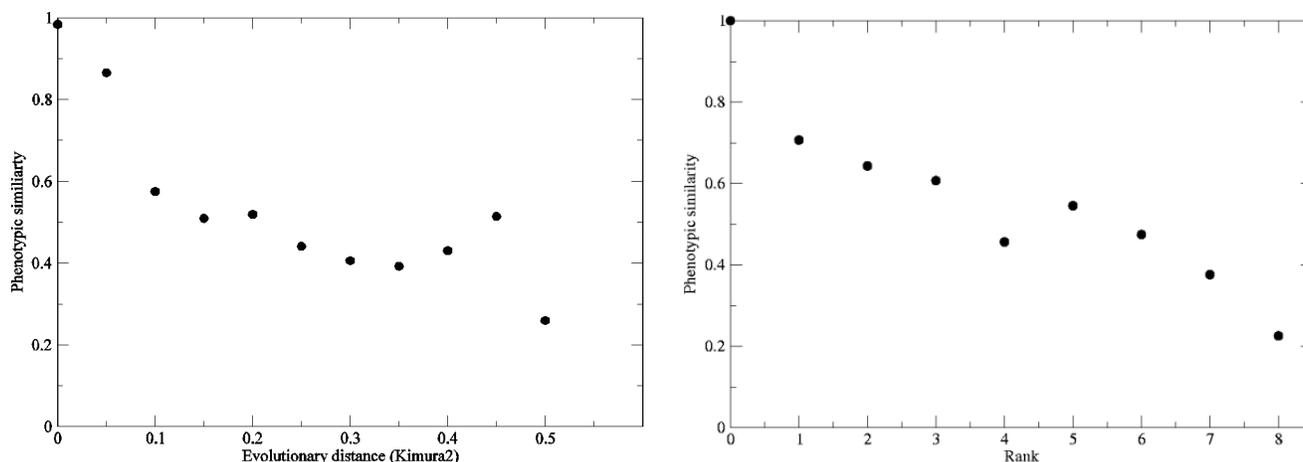


Figure 11. Mutual information based phenotypic similarity between pairs of organisms, versus 16S rDNA evolutionary distance (left) and taxonomic rank (right)

The observed relationship between phenotype and evolutionary distance varies according to lineage (Figure 12). Within the *Firmicutes* lineage, the correlation is comparatively weak ($r^2=0.2$), indicating that the overall phenotype profile of closely related species belonging to the

Firmicutes phylum varies considerably. On the other hand, for *Proteobacteria*, the correlation is best preserved ($r^2=0.4$). In addition, the phenotypes “facultative anaerobic”, “endospore formation”, “coccus shaped” and “intracellular growth” appear to be best conserved over evolutionary distance.

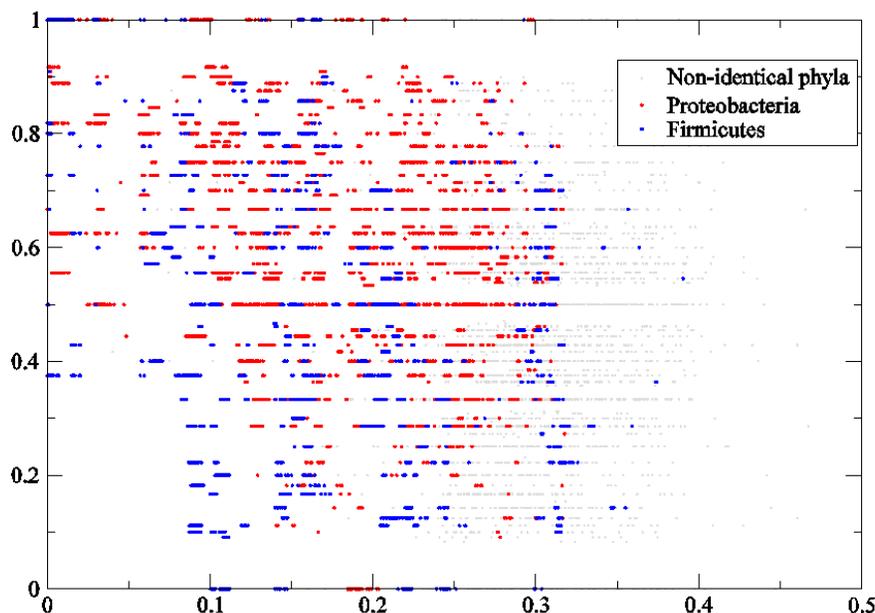


Figure 12. Scatterplot of the observed phenotypic similarity between pairs of organisms within the phyla Firmicutes (blue) and Proteobacteria (red) at different evolutionary distances

The results presented herein are based on the mutual information measure. However, under some circumstances, very phenotypically dissimilar organisms can have high mutual information scores. To avoid this problem, we developed an alternative phenotypic similarity measure, 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. Given the phenotype profile for the two organisms, a phenotype similarity score may be calculated in the following manner:

$$PhenSimScore(org_1, org_2) = -\sum_{i \in A} \log(\Pr(Ph_i = 0)^2 + \Pr(Ph_i = 1)^2) + \sum_{i \in D} \log(2 \Pr(Ph_i = 0) \Pr(Ph_i = 1)),$$

where $i \in A$ refers to those phenotypes for which organisms *org1* and *org2* have identical annotations (i.e. either both 1 or both 0), and $i \in D$ refers to those phenotypes on which their annotations disagree.

A manuscript describing these results – recalculated using the new log likelihood phenotypic similarity measure, and including a statistical analysis of the correlation length of different phenotypes with evolutionary distance – is currently in preparation.

Exit Plan

The type of large-scale cross-genome analyses presented here are essential for DOE/OBER’s Genomes to Life program. Based in part on the expertise we have developed during this LDRD project, we are participating in the Joint BioEnergy Initiative, a \$25M multi-institutional (LBNL, SNL, LLNL, Stanford and UCD) proposal for the GTL call for Bioenergy Research Center. We are

also participating in a project which has been proposed for external commercial funding, on Bioprospecting for interesting biomass degrading communities in hot springs, and are in negotiations with two other commercial partners to do microbial systems biology and microbial community work with applications to bioenergy. We are also collaborating with researchers at the JGI on a new strategically important LDRD project in metagenomics.

Within the area of human health, the NIH has recently started a Human Microbiome Project, in coordination with an international Human Microbiome Consortium, and we are currently preparing a grant proposal on metagenomics tools for and upcoming RFA.

This work also has possible applications for biodefense - we are currently participating in a LDRD Strategic Initiative proposal on Host-Pathogen interactions, and also intend to submit an application for participation in the Pacific-Southwest Regional Center of Excellence for Biodefense & Emerging Infections, with a focus on the role of metabolic and regulatory networks in virulence. In the past, we have also submitted two pre-proposals to DTRA calls in this area.

Given the sharp rise in interest in cross-genomic comparisons, and microbial communities since this project began, it is likely that this sort of work will continue to attract other sources of funding, and that demand for these types of methods - and hopefully funding to analyze this new data - will increase sharply.

Summary

The type of predictive capability for genotype - phenotype mapping presented here is crucial if we want to make sense of the flood of new sequence data. We expect that this will position the Laboratory as a leader for the study of bacterial systems, and a key player in the DOE's Genomes to Life program, as well as provide an invaluable resource and computational tool set for use by the rest of the scientific community.

References

- von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P., Snel B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258-61.
- Karp P.D., Ouzounis C.A., Moore-Kochlacs C., Goldovsky L., Kaipa P., Ahren D., Tsoka S., Darzentas N., Kunin V., Lopez-Bigas N.. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* 19:6083-89.
- Lee T.J., Pouliot Y., Wagner V., Gupta P., Stringer-Calvert D.W., Tenenbaum J.D., Karp P.D. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*. 7:170.
- Tatusov R.L., Natale D.A., Garkavtsev I.V., Tatusova T.A., Shankavaram U.T., Rao B.S., Kiryutin B., Galperin M.Y., Fedorova N.D., Koonin E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1):22-8.
- Lee D., Seung H., 1999. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* 401, 788-791.
- The Bergey's Manual Trust. (1994) *Bergey's Manual of Determinative Bacteriology* (ed. Holt J.G.). Williams & Wilkins, Baltimore, 9th Ed.
- Agrawal R., Imieliński T., and Swami A. (1993) Mining association rules between sets of items in large databases. In *Proc.1993 ACM SIGMOD international conference on Management of data*, 207-216

- Slonim N., Elemento O., Tavazoie S. (2006) Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks. *Molecular Systems Biology*, 2, 2006
- Ciccarelli F.D., Doerks T., von Mering C., Creevey C.J., Snel B. Bork P. (2006) Towards automatic reconstruction of a highly resolved tree of life. *Science*, 311:1283-1287.
- Kanehisa, M., Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, 27-30.
- The Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**: D258-D261.
- Vapnik, V. (1998) Statistical Learning Theory. Wiley, New York.
- DeSantis, T. Z., I. Dubosarskiy, S. R. Murray, and G. L. Andersen. 2003. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. *Bioinformatics* 19:1461-8.
- Chenna, Ramu, Sugawara, Hideaki, Koike, Tadashi, Lopez, Rodrigo, Gibson, Toby J, Higgins, Desmond G, Thompson, Julie D. Multiple sequence alignment with the Clustal series of programs. (2003) *Nucleic Acids Res* 31 (13):3497-500