A Bioinformatics Platform for Studying Metabolic Functions of Unculturable Microbes

Allan Dickerman[1] and Brett Tyler[2]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24060

[2]Department of Plant Pathology, 155 Robbins Hall, U.C. Davis, Davis, CA 95616

Abstract

The processes of natural ecosystems are strongly affected by the lifestyles of their most numerous inhabitants, the bacterial and archaeal prokaryotes. Examples of phenomena that affect local ecosystems and the biosphere at large include cycling of carbon and nitrogen, and both positive and negative interactions with plants and animals. Recent advances in molecular techniques such as 16S rDNA sequencing have revealed that natural ecosystems are dominated by bacteria which are not culturable and hence remain largely unknown. The growing wealth of completely sequenced microbial genomes does not directly illuminate these natural ecosystems. Direct amplification or cloning from environmental samples of genes involved in metabolic processes of interest does allow us to investigate molecular processes at the genetic level among these poorly known systems. A major informatics challenge is to integrate data from the sequencing of multiple different genes from a mixture of unknown organisms and assemble a model of which genes co-occur within individual species. We propose to develop an informatics-based solution to this problem based on concordance of gene phylogenies and sequence attributes such as codon bias and oligonucleotide word frequencies. The system will take anonymous DNA sequences from multiple loci and estimate which ones come from the same species. A successful solution will enable a new level of sophistication in the way we can study the functional genomics of the largely unknown, unculturable microbes that dominate terrestrial, subterranean, and aquatic ecosystems.

Introduction

Microbes have profound impacts on soil and aquatic ecosystems. These range from cycling of carbon, nitrogen and other major nutrients, to direct positive (mutualistic) and negative (pathogenic) impacts on plants, animals and humans. Microbes occupy habitats as diverse as soil, the open ocean, subterranean rocks, deep ocean floor vents and sediments, volcanic pools, and specialized symbiotic organs of plants and animals. Yet from 99% to 99.999% of bacterial species cannot be cultured for the purpose of study.

As result, we are profoundly ignorant of how these organisms affect the functioning of the ecosystems they inhabit (Cowan, 2000). The advent of PCR has enabled DNA sequences to be amplified for study directly from microbial communities without culturing the organisms. Sequencing 16S rRNA genes in particular has confirmed that there exists a vast phylogenetic diversity of microbes in these ecosystems. Over 8000 16S rRNA sequences reveal at least 36 major divisions of eubacteria alone; of these 13 are represented entirely by unculturable organisms, and several others have only a small number of culturable representatives (Hugenholtz, 1998). Some of these novel divisions are abundant and ubiquitous in soil and aquatic ecosystems, and hence their functions must be important, but they remain unknown. Despite the success of 16S rRNA analysis in revealing the diversity of the microbes, this approach reveals little about the ecological functions of those microbes. An alternative approach has been to amplify key functional genes from ecosystem DNA. Examples of functionally tartgetted environmental samples include nitrogenase (involved in nitrogen fixation) (Zani, 2000), nitrous oxide reductase (involved in denitrification) (Scala, 1999), ribulose-bisphosphate carboxylase (involved in carbon dioxide fixation) (Elsaied, 2001) or methane oxidase (involved in methane utilization) (Murrell, 2000). This approach has

been valuable in identifying the diversity of organisms catalyzing those reactions and quantifying their contributions. However, because only a single gene is amplified and sequenced, it is difficult or impossible to predict what other genes and what other ecological functions are possessed by the bacteria carrying the gene of interest. Even a close match to a gene from a known species is not fully reliable as the phylogenetic trees of individual proteins are not always congruent with 16S rRNA trees, often due to horizontal gene transfer (Cowan, 2000). Fluorescence in situ hybridization (FISH) and in situ PCR have been helpful in this regard, enabling different genes to be traced to particular bacterial cell morphologies (Reysenbach, 1994). However, the in situ approaches are very labor-intensive and are not always technically feasible, especially complex substrates such as soil.

An alternative, more feasible approach to determining the function of unculturable microbes in ecosystems might be to develop bioinformatics approaches that estimate which versions of different genes amplified from environmental DNA belong to the same species. For example, one might test whether a community contains any nitrogen-reducing, methane utilizing species by estimating whether any of the multiple versions of a nitrate reductase gene amplified from community DNA derive from the same species as one of the methane oxidase genes amplified from the same community DNA. The putative nitrogen-reducing, methane utilizing species might then be placed on a molecular phylogenetic tree by estimating which 16S rRNA sequence is most likely to be associated with the nitrate reductase and methane oxidase genes (Petri, 2000). Essentially, this approach would assemble specific genes of interest into hypothetical genome units without the expense of fully sequencing multiple genomes. The question we will address is whether this shortcut is possible. Here we propose to develop and test, in silico, a method for partitioning sampled sequences among a set of unknown species based on the concordance of phylogenetic trees and on DNA sequence attributes such as codon usage.

Approach

The problem we propose to solve is this: given a set of nucleotide sequences sampled from the natural ecosystem consisting of many unknown species, and assuming the sequences are sampled to target multiple genes of interest, assemble a model which partitions sequences among a set of hypothetical genomes. Further desirable inferences will include estimates of the phylogeny of the unknown species in the context of known genomes and possibly distinct separate gene trees possibly representing horizontal transfer. This problem is generic to cases where multiple genes are sampled from a mixture of unknown species and a robust solution should be of wide applicability. The sequences may be either short and represent subregions of the target genes, or they may be longer and include at least one target gene each in addition to other genes. We will include homologous sequences from known genomes to anchor the trees. We will align sequences with Clustal (Jeanmougin et al., 1998) and then infer phylogenetic trees with PAUP* (Swofford, 2000) and PHYLIP (Felsenstein, 1993) using distance, parsimony, and likelihood criteria. In the case of parsimony-based trees, we will have branch lengths fit by maximum likelihood using PAUP*. Trees will be generated from both nucleotides and inferred protein translations. The use of a variety of tree methods is planned to explore the tradeoffs between speed and rigor and to hedge against the vulnerabilities of any single method.

However generated, output trees will contain some tips labeled by species (those from complete genomes) and others with only anonymous sequence designations. The task will be to take two or more such trees and match up their anonymous tips in hypotheses of occurrence within the same species. The matching of tree tips need not be one-to-one as some genomes may lack (or have missed through sampling) one or more genes or have multiple paralogous copies. The overall problem consists of the subproblems of generating hypotheses (labelings of the anonymous tree tips) and evaluating these on a useful optimality criterion. Developing the appropriate optimality measure will be the core task. One component of the criterion will be a measure of tree congruence. One way to score tree congruence is to label the internal nodes with the set of terminal labels subtended by the node and then to score the number of internal nodes that have identical (or possibly similar) sets. Tree branch lengths below each corresponding internal node

can be incorporated in the form of a weighted correlation. We will apply such scores to the multiple trees generated by different methods to yield a set of means and variances for each score on each tree method to yield a multivariate measure of congruence.

Sequence attributes other than those used to infer the trees will also be drawn on, namely codon usage, frequencies of dinucleotides spanning codon boundaries and noncoding word frequencies. Codon usage can be affected by gene-specific effects such as the level of expression and the direction of transcription relative to DNA replication (Tillier, 2000). Therefore we will focus on genes on a fragment that have the most biased usage and if possible, include genes transcribed in opposite directions. Dinucleotide frequencies at codon junctions incorporate sequence biases not present within codons (De Amicis, 2000). Frequencies of nucleotide words of different lengths (di-, tri-, tetra-nucleotides etc) also reflect genome-wide biases. We will test a variety of word lengths from 2 to 6 nucleotides in both non-coding and all regions for their ability to group non-homologous genes within species. Each of these measures will be used to create a distance matrix among the sequences. The overall optimality criterion used to evaluate hypothetical partitions of sequences into genomes will be a weighted sum of the tree congruence and the sequence attribute scores. When drawing our "unknown samples" from complete genomes, we can also score how close each hypothesis matches reality. This will allow us to estimate optimal weights for each component measure by multiple regression against the measure of success as well as determine how well the algorithm will work.

The goal is to use estimates of genome partitioning to make biological interpretations and put some confidence measure on them. This could mean inferring that a hypothetical species is missing certain genes in the analysis. e.g. If a cluster contains only closely related DNA sequences spanning a nitrate reductase gene, but no methane oxidase genes, then we will tentatively infer that the species containing that particular nitrate reductase gene region lacks methane oxidase genes, with the concomitant implications for community metabolic capabilities. The information on sequences flanking target genes can be used to design experimental tests that a certain gene is indeed missing in a particular genome. Other possible feedback between analysis and further laboratory experiments that can emerge are hypotheses that a particular gene, while present, is not actually expressed as suggested by codon usage or that a particular gene occurs on a plasmid and may be readily transferred among species.

Current Status

We have performed a global all-versus-all alignment search between all pairs of proteins in 71 completely sequenced prokaryote genomes using BLASTP with a expectation threshold of 10^{-6}. This yielded over 20 million pairwise alignments, or "hits". We have implemented an algorithm to sort through these pairwise hits to find regions in which multiple sequences will be mutually alignable. We call such a cluster a homology group. This algorithm found 9703 such homology groups with three or more members, and 1565 groups with 30 or more members. We limited subsequent analyses to groups with 30 or more members to focus on taxonomic breadth. Performing multiple sequence alignment resulted in 1259 successful alignments with a minimum length of 50 amino acids. These were subjected to phylogenetic inference using protein parsimony (the Protpars program of PHYLIP) and the resulting trees stored to the database. We are analyzing these by comparison to a canonical species tree, for which we are using the RDB projects small-subunit prokaryote rRNA tree (website http://rdp.cme.msu.edu/html/). We are comparing the tree length of each alignment on the most-parsimonious tree to the length of the alignment forced onto the given species phylogeny. We are also comparing the distance matrices of each protein alignment to the distances on the rRNA tree. Both measures are being used to find gene trees that are in strong agreement with the species phylogeny (or it's 16S rRNA surrogate). Various statistical measures of robustness are also being applied. The goal here is to know the relative utility of each of these genes for phylogeny inference if found environmental samples of large DNA clones.

Subsequent work will focus on using complete genomes to do "blind" simulations of environmental samples using large-insert clones. We want to design a system that can reliably sort environmental sequence samples into well-supported hypothetical genomes and to characterize the metabolic repertoire of each species found, or to put limits on incomplete knowledge and predict how much more data will be necessary to achieve robust estimates.

Literature Cited

Cowan, D. A. (2000). "Microbial genomes: The untapped resource." Trends in Biotechnology 18(1): 14-16.

De Amicis, F. and S. Marchetti (2000). "Intercodon dinucleotides affect codon choice in plant genes." Nucleic Acids Research 28(17): 3339-3345.

Elsaied, H. and T. Naganuma (2001). "Phylogenetic diversity of ribulose-1,5- bisphosphate carboxylase/oxygenase large-subunit genes from deep-sea microorganisms." Applied and Environmental Microbiology 67(4): 1751-1765.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Hugenholtz, P., B. M. Goebel and N. R. Pace (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J. Bacteriology 180(18): 4765-4774.

Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998). Multiple sequence alignment with Clustal X. Trends Biochem Sci, 23, 403-5.

Murrell, J. C. and S. Radajewski (2000). "Cultivation-independent techniques for studying methanotroph ecology." Research in Microbiology 151(10): 807-814.

Petri, R. and J. F. Imhoff (2000). " The relationship of nitrate reducing bacteria on the basis of narH gene sequences and comparison of narH and 16S rDNA based phylogeny." Systematic and Applied Microbiology 23(1): 47-57.

Reysenbach, A.-L., G. S. Wickham and N. R. Pace (1994). "Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park." Applied and Environmental Microbiology 60(6): 2113-2119.

Scala, D. J. and L. J. Kerkhof (1999). "Diversity of nitrous oxide reductase (nosZ) genes in continental shelf sediments." Applied and Environmental Microbiology 65(4): 1681-1687.

Swofford, D. L. (2000). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tillier, E. R. M. and R. A. Collins (2000). "The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes." Journal of Molecular Evolution 50(3): 249-257.

Zani, S., M. T. Mellon, J. L. Collier and J. P. Zehr (2000). "Expression of nifH genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR." Applied and Environmental Microbiology 66(7): 3119-3124.