BIOINFORMATICS

Vol. 22 no. 14 2006, pages e359–e367
doi:10.1093/bioinformatics/btl217

An experimental metagenome data management and analysis system

Victor M. Markowitz[1,*], Natalia Ivanova[2], Krishna Palaniappan[1], Ernest Szeto[1], Frank Korzeniewski[1], Athanasios Lykidis[2], Iain Anderson[2], Konstantinos Mavrommatis[2], Victor Kunin[3], Hector Garcia Martin[3], Inna Dubchak[2], Phil Hugenholtz[3] and Nikos C. Kyrpides[2]

[1]Biological Data Management and Technology Center, Lawrence Berkeley National Lab, USA, [2]Genome Biology Program, Joint Genome Institute, USA and [3]Microbial Ecology Program, Joint Genome Institute, USA

ABSTRACT

The application of shotgun sequencing to environmental samples has revealed a new universe of microbial community genomes (metagenomes) involving previously uncultured organisms. Metagenome analysis, which is expected to provide a comprehensive picture of the gene functions and metabolic capacity for microbial communities, needs to be conducted in the context of a comprehensive data management and analysis system. We present in this paper IMG/M, an experimental metagenome data management and analysis system that is based on the Integrated Microbial Genomes (IMG) system. IMG/M provides tools and viewers for analyzing both metagenomes and isolate genomes individually or in a comparative context. IMG/M is available at http://img.jgi.doe.gov/m.

Contact: vmmarkowitz@lbl.gov

1 INTRODUCTION

Environmental microbial community (microbiome) genome analysis, also known as *metagenome* analysis, (Riesenfeld *et al*., 2004) is expected to lead to advances in environmental cleanup, agriculture, industrial processes, and alternative energy production. Similarly, human metagenome analysis could provide new insights into the variation of microbial populations associated with the human body, ascribe qualitative and quantitative changes in human microbiota as risk/causative factors of disease and lead to the development of new treatment strategies (Gordon *et al*., 2005).

The application of shotgun sequencing to microbiome samples has enabled the study of metagenomes involving previously uncultured and unculturable organisms. Comparative analysis of the metagenomes in the context of available reference isolate genomes could potentially reveal large-scale patterns of biochemical interactions and habitat-specific correlations in the host environment that might otherwise be missed (DeLong and Karl, 2005). Studies of environmental microbiomes, such as acid mine drainage biofilms (Tyson *et al*., 2004) and Sargasso Sea samples (Venter *et al*., 2004), as well as studies of human microbiomes, such as the human gut microbiome (Gordon *et al*., 2005), are examples of a rapidly expanding area of metagenome analysis applications.

Unlike microbial genome data from isolate organisms, the generation and interpretation of metagenome data is in early stages of development. Metagenomes sequenced by organizations such as the Joint Genome Institute (JGI), TIGR, and the Venter Institute, follow an assembly and annotation process that is specific to each sequencing center. Although traditional assembly and annotation algorithms do not perform as well on metagenome sequences as they do on isolate microbial genomes (see (Chen and Pachter, 2005) for an overview of metagenome sequence assembly and gene prediction problems), they yield data that are amenable to valuable comparative analysis and interpretation as illustrated by the studies published in (Tringe *et al*., 2005) and (Tyson *et al*., 2004). Thus, the metagenome sequences of simple microbiomes can be assembled into sizable scaffolds and for highly abundant (dominant) member organisms the quality of the assembly and annotation may approach that of draft isolate genomes. For such metagenomes, it is possible to infer the metabolic capabilities of dominant organisms and identify the key member organisms that perform community-essential tasks.

Although metagenome sequence data processing poses numerous challenges due to the complex nature and inherent incompleteness of the data, and the lack of methods designed specifically for processing such data, successful analysis can be carried out on existing metagenomic data. As initial methods are improved or new methods emerge, metagenome data sets will be revised, thus leading to better quality data and annotations. However, metagenome data analysis needs to be conducted in the context of a comprehensive data management and analysis system that provides support for data review and revision. We have addressed this need by developing an experimental metagenome data management and analysis system, IMG/M, based on the Integrated Microbial Genomes (IMG) system (Markowitz *et al*., 2006).

Like IMG, IMG/M is based on the principle that integration of available genomic data is essential for understanding the biology of newly sequenced genomes, as the efficiency of genome analysis increases substantially when it is conducted in a comparative

*To whom correspondence should be addressed.

context. Such an integrated context is even more critical for analyzing the inherently incomplete metagenome data. IMG/M has been successfully used for the study of biological phosphorus removing (EBPR) sludge communities (Martin *et al*., 2006), and is currently used for analyzing several metagenomes sequenced at JGI.

In the following sections, we first discuss the main metagenome data processing challenges. Next, we briefly review metagenome data modeling and analysis. Finally, we present the IMG/M metagenome data analysis tools and discuss our plans to extend these tools.

2 METAGENOME DATA PROCESSING

There are two general sequencing strategies to obtain genome sequence data from microbiome samples: directed sequencing and shotgun sequencing of random clones. Directed sequencing is either (i) function-driven, whereby clone libraries from a microbiome sample are sequenced after being screened for a desired function; or (ii) driven by phylogenetic markers, whereby the DNA flanking taxonomic anchors, such as 16S rDNA, is sequenced in large-insert libraries. Conversely, shotgun sequencing of microbiome sample clone libraries follows a relatively unbiased approach, which provides a broad survey of the gene content and metabolic capabilities of a microbiome. A combination of shotgun and directed sequence approaches may emerge in the future and thus combine the advantages of the broad coverage provided by shotgun sequencing with the ability of sampling specific genome areas in low abundance organisms without over-sequencing more abundant members of the microbiome. The discussion below pertains to metagenome data generated using shotgun sequencing.

Metagenome sequence data processing follows assembly and annotation procedures that are specific to each sequencing center. Assemblers, such as the Celera Genome Assembler, PHRAP, and JAZZ (Aparicio *et al*., 2002) have been used with mixed results (Chen and Pachter, 2005). Assembly of shotgun-sequenced microbiome samples poses a serious challenge to traditional assembly methods, due to a fundamental difference between the sequences derived from cultivated microbes and microbial communities. While the genome sequence of a cultivated microbe is derived from a clonal isolate, where all cells are descendants of one cell and therefore genetically identical or nearly identical, the aggregated genome sequence of a microbiome is derived from a heterogeneous pool of cells, some of which are genetically related and probably correspond to different strains of the same species, while others are genetically distinct. Although co-assembly of the sequences derived from different species does not seem to be a problem, traditional methods are not consistent in assembling the sequence reads belonging to different strains of the same species: depending on the assembly algorithm and sequencing read depth they can be resolved into strain-specific scaffolds or co-assembled into a composite species population scaffold. In the later case the strain-specific variations appear as single nucleotide polymorphisms (SNPs) in the sequence.

Annotation of the assembled metagenomes is also currently carried out using traditional approaches developed for isolate genomes. For instance, protein-coding genes (CDSs) are predicted on scaffolds and/or so called *shrapnel* sequences (single reads that are not incorporated into scaffolds) using microbial gene finders, such as Glimmer (Delcher *et al*., 1999) or Fgenesb (Soft-Berry, 2006). Performance of traditional gene prediction methods is affected by the inevitable fragmentation of metagenomic sequences, which in turn leads to fragmentation of the genes, and therefore sometimes gene prediction is limited to BLASTx of all open reading frames against protein sequence databases. Functional annotation of predicted CDSs is generally carried out using COG (Tatusov *et al*., 1997), Pfam (Bateman *et al*., 2004), InterPro (Mulder *et al*., 2005), and KEGG (Kanehisa *et al*., 2004); functional annotations can also be marred by gene fragmentation in the metagenome datasets.

Sometimes an additional stage of scaffold *binning* is included in order to assign scaffolds and *shrapnel* sequences to organism types (phylotypes) that could range from coarse-level groupings such as domain (*Bacteria*, *Archaea*) down to fine-level groupings such as individual strains of a given species. It is highly desirable that all sequence fragments are assigned to a particular strain in the community; however, this is usually not feasible due to the different abundance of the strains and variation of sequence coverage. Consequently, the highest resolution grouping for metagenome data can be achieved at the species level, that is, grouping together genomic fragments that are likely derived from members of a given *species population*, whereby each bin represents a snapshot of a *composite genome* of a *species population*. Some regions of such a *composite genome* are represented by sequences originating from only one strain (usually, the most abundant one), while others are covered by sequences from multiple strains. The latter may exhibit different types of strain-level heterogeneity, from SNPs to extensive genome rearrangements. Binning algorithms rely on measuring the oligonucleotide frequency in different scaffolds, depth of sequence coverage or phylogeny of conserved protein markers; thus, binning accuracy depends on the sequence coverage, quality of the assembly, scaffold size, complexity of the microbiome, and available reference isolate microbial genomes (Chen and Pachter, 2005). While it is expected that binning will be difficult in the case of highly fragmented metagenomes of complex microbiomes, such as those from soil samples (Tringe *et al*., 2005), for simpler microbiomes with sufficient sequence coverage it is possible to reconstruct more than 95% of the individual genomes of the dominant community members (Tyson *et al*., 2004).

Despite the metagenome data processing challenges mentioned above, analysis of metagenomes does not need to wait for the development of optimal data generation and annotation methods: such analysis can be carried out with existing methods with the results of these analyses serving as a basis for improving the methods in an iterative process.

3 METAGENOME DATA MODEL AND ANALYSIS

Similar to isolate microbial genome data, *metagenome* data captures information about DNA sequences along with *genes* that can be further characterized in terms of *functional roles*. A *gene* represents an ordered sequence of nucleotides located on a particular *chromosome* that encodes a specific product (i.e., a protein or RNA molecule); its protein product can be characterized in terms of sequence similarity to other protein products, presence or absence of *conserved motifs* and *domains*. *Functional* roles of genes can be characterized in the context of *pathways*, whereby pathways are associated with genes via gene products that can function as

enzymes catalyzing individual reactions of metabolic pathways. Similar to isolate microbes, the metabolic capacity of a whole *microbiome* can be characterized by analyzing the *metabolic maps* inferred from the gene content and distribution of its composite genome.

Metagenome data have an additional level of complexity reflecting the complex nature of microbiomes, which, unlike clonal isolates, consist of heterogeneous pools of cells belonging to different strains and species. Therefore metagenome scaffolds can be further characterized in terms of their *bin* assignment, whereby a *bin* could correspond to a *composite genome* of a *species population* or another higher-level taxonomic group. If a *bin* corresponds to the *species population*, it could be characterized by strain-level heterogeneity (e.g., SNPs or genome rearrangments). Similar to a metagenome which represents a random sample of the aggregate microbiome genome, a *bin* may represent only a subset of the aggregate genome of a *species population*, and therefore may not reflect all the diversity of this species population in terms of strain-level heterogeneity.

Another important difference between metagenome data and isolate genome data is that metagenome data are representative of a microbiome in a specific *host* environment and a specific *sample* of this environment. Sample (meta) data characterizing the biological material collected for sequencing, are specific to an application domain. For example, for biomedical applications samples are collected from human donors and therefore are associated with attributes that describe donor host data (e.g., demographic and clinical record), sample structural and morphological characteristics (e. g., site and time of collection) and sample processing protocol. Sample metadata are critical in metagenome comparative data analysis.

Comparative data analysis plays an important role in understanding the biology of isolate microbial genomes (Bowers *et al*., 2004). Similar to isolate genomes, the analysis of metagenomes in the comparative context of other (e.g., phylogenetically related) genomes is substantially more efficient than analyzing each metagenome in isolation. Metagenome data analysis is set in a multidimensional data space, whereby microbiome samples form one of the dimensions and are analyzed in the context of other dimensions, such as component species populations, gene families represented by homolog/ortholog clusters, COG groups or Pfam families, and pathways and networks.

For example, microbiome samples can be compared in terms of presence and abundance of certain gene families. This type of analysis is based on the assumption that the genes important for adaptation to a particular environment will be found in many (if not all) organisms in the microbiome; moreover, such genes might be present in multiple copies, therefore, they are more likely to be found among the abundant gene families. Gene family abundance profiles can be analyzed at higher resolution, when bins within the same microbiome rather than microbiome samples are compared; this type of analysis allows to verify directly the assumption that abundant gene families are indeed present in many members of a microbiome.

Another emerging method of analyzing metagenomic data involves detection of presence and abundance of certain metabolic pathways in a specific microbiome sample or across samples of the same microbiome or different microbiomes. Such analysis typically involves examining *occurrence profiles* (Osterman and Overbeek, 2003) of functions and pathways of interest across samples associated with a specific microbiome or across diverse microbiomes. Alternatively, the bins within the same metagenome dataset can be compared in terms of presence/abundance of functions and pathways. This analysis helps to infer the metabolic capabilities of the component organisms in the community, and thus identify the key members of the microbiome that perform community-essential tasks and pinpoint the metabolic interactions within the microbiome and between the microbiome and its host environment.

Both examples discussed above are focused on the analysis of metagenome data *per se*, however, an efficient analysis of metagenomes is not possible without the context of reference genomes. Similar to comparisons of microbiome samples and bins within metagenome datasets, metagenome sequences can be compared to isolate microbial genomes in terms of gene family abundance, presence or absence of functions and pathways, and so on.

4 AN EXPERIMENTAL METAGENOME DATA MANAGEMENT AND ANALYSIS SYSTEM

We have developed an experimental metagenome data management and analysis system, IMG/M, based on the Integrated Microbial Genomes (IMG) system (Markowitz *et al*., 2006). The IMG/M system and data analysis tools are briefly overviewed below.

4.1 System Overview

The content of IMG/M can be seen as a superset of IMG's content. IMG integrates bacterial, archaeal and selected eukaryotic genomic data collected from multiple data sources. Thus, IMG 1.3 (as of December 1st, 2005) contains a total of 678 genomes consisting of 377 bacterial, 26 archaeal, 15 eukaryotic genomes and 260 bacterial phages. IMG's extensive collection of microbial genomes (both draft and finished) provides the foundation for analyzing the fragmented inventory of genes, functions, and organisms in microbiomes and their component populations.

In addition to the isolate genomes in IMG 1.3, the first experimental version of IMG/M (as of March 1st, 2006) includes metagenome sequences generated from an acid mine drainage (AMD) biofilm (Tyson *et al*., 2004), an agricultural soil sample (Tringe *et al*., 2005), three isolated deep sea ''whale fall'' carcasses (Smith and Baco, 2003), and two biological phosphorus removing (EBPR) sludge samples (Martin *et al*., 2006). These microbiomes comprise a representative set in terms of species diversity, abundance of dominant organism(s) and sequencing depth. For instance, species diversity ranges from very low in the case of the AMD sample to extremely high in the soil sample, while abundance of dominant organism(s) ranges from less than 1% in the soil sample to more than 80% in EBPR sludge samples. Furthermore, two EBPR sludge samples represent an example of microbiomes inhabiting similar environments in two distinct geographical locations. Consequently, the metagenome data in IMG/M can be employed to test use case scenarios, formulate and test various hypothesis, assess performance of available tools and develop new methods for metagenome analysis.

The IMG/M back-end consists of a data warehouse, sequence databases for similarity (BLAST) searches, and various auxiliary data files containing scaffold DNA sequences, pathway map images, and cached data for improving performance, such as pre-computed statistics and homolog results. An additional

BLAST database supports similarity searches based on the sequencing reads for analysis of strain level single nucleotide polymorphisms (SNPs). The data generated by microbial genome and metagenome data processing pipelines serve as input for a custom ETL (Extract-Transform-Load) toolkit that loads data into the IMG/M data warehouse. This toolkit is also employed for extracting, cleaning, integrating, and loading additional genomic and contextual data from external resources into the data warehouse. Additional custom tools are employed to compute gene relationships and clusters and load these data into the data warehouse.

The data model for the IMG/M data warehouse allows integrating primary genomic sequence information, computationally predicted and curated gene models, pre-computed sequence similarity relationships, and functional annotations and pathway information in a coherent biological context. Isolate organisms are identified via their taxonomic lineage (domain, phylum, class, order, family, genus, species, strain). For each genome, the primary DNA sequence and its organization in scaffolds and/or contigs, are recorded. Genomic features, such as predicted coding sequences (CDSs) and some functional RNAs, are also recorded. Protein-coding genes are further characterized in terms of molecular function and participation in pathways. Proteins are grouped into protein families based on sequence similarity. Pathways, reactions, and compounds are included from KEGG and LIGAND. Additional functional annotations according to Gene Ontology terms (Gene Ontology Consortium, 2004) are provided by EBI Genome Reviews (Kersey *et al*., 2005), while COG provides clusters of orthologous groups of genes. Ortholog and paralog gene relationships for isolate microbial organisms are computed based on bidirectional best hit (BBH) with clusters formed using Markov Clustering method (MCL) (Enright *et al*., 2002). Isolate organisms are characterized in terms of phenotypes (e.g., morphology, geochemistry), ecotype (including geographical coordinates) and disease.

Microbiome samples are treated as ''meta'' organisms with the collection of their associated genes forming their respective metagenomes. The sequences of a microbiome sample together with their associated genes and annotations are organized in bins when possible, with multiple bins providing support for recording data generated using different binning methods. Similar to isolate organisms, microbiome samples are characterized in terms of phenotypes, ecotype, disease, and relevance. These data are only minimal in coverage, reflecting the current scarcity of such data for microbiome samples.

4.2 Data Analysis

We review below the IMG/M data exploration and comparative analysis tools, with special emphasis on the support for metagenome analysis. IMG/M tools can be also employed for analyzing isolate microbial genomes in the same way as their IMG counterparts.

4.2.1 Data Exploration Data exploration tools in IMG/M help selecting and examining genomes, genes, and functions of interest. Metagenomes as well as isolate genomes can be selected using a keyword based *Genome Search* in conjunction with a number of filters or an alphabetically or phylogenetically organized *Genome Browser*. Microbiomes can be further examined using the *Microbiome Details*, where a user can find relevant metadata, such as geographical location, along with various summaries of interest, such as the total number of scaffolds and genes or the number of genes associated with functional characterizations (eg., COG, Pfam), as shown in the right pane of Figure 1. *Microbiome Details* also provides an estimate of phylum level assignment (*Phylogenetic Mapping*) of metagenomic fragments in the sample based on sequence comparison to isolate genomes. This overview consists of the distribution of the best BLAST hits at different percent identity thresholds of a metaproteome (i.e., the collection of all the proteins encoded in the metagenome) of interest against the proteomes of all isolate genomes in the system, as shown in the left pane of Figure 1. For each metagenome one can also examine the associated list of scaffolds and contigs, and information on individual bins and their scaffolds when bins are available.

Genes can be selected using a keyword based *Gene Search*, sequence similarity search tools, or a gene profile based selection tool, the *Phylogenetic Profiler*, discussed in more detail below. The functional role of genes in IMG/M is characterized by a variety of annotations, including their COG membership, association with Pfam domains, Gene Ontology (GO) assignments, and association with enzymes in KEGG pathways. Functional annotations can be searched using keywords and filters, with the selected functions leading to a list of associated genes either directly or via a list of organisms. COG functional categories and KEGG pathways can be searched and browsed separately. The lists of genes and functional annotations that are of interest for further exploration can be maintained using various *Analysis Carts*, which are similar to shopping carts of commercial websites.

Individual genes can be analyzed using *Gene Details* pages, as illustrated in Figure 1. A Gene Information table includes gene identification, locus information, biochemical properties of the product, and associated KEGG pathways. *Gene Details* also includes evidence for the functional prediction: gene neighborhood, COG, InterPro, and Pfam, and pre-computed lists of homologs, orthologs and paralogs (for isolate organisms), or intra-metagenome homologs as well as homologs to other genomes and metagenomes (for microbiomes). The gene neighborhood displays the target gene and its homologs in user selected related genomes with its neighboring genes in a 25kb chromosomal window: for example, the gene neighborhood in the *Gene Details* in Figure 1 shows the target gene (centered, in red) and other genes within a 25kb window. The *Gene Neighborhoods* in Figure 1 shows the neighborhood of a target gene of the *Ferroplasma acidarmanus Type I* bin of the AMD metagenome, compared to homologous genes of the *Ferroplasma acidarmanus fer*1 isolate genome: each gene's neighborhood appears above and below a single line showing the genes reading in one direction on top and those reading in the opposite direction on the bottom; genes with the same color indicate association with the same COG. For each gene, locus tag, scaffold coordinates, and COG number are provided locally (by placing the cursor over the gene), while additional information is available in the *Gene Details* associated with each gene. A gene can be also examined in the context of its associated pathways, whereby the link embedded in the pathway name listed in the *Gene Information* table allows the KEGG map associated with the gene to be displayed. On such a map, EC numbers are color-coded and linked to the *Gene Details* for the associated genes.

Individual COG categories can be further explored with *COG Category Details* that lists the COGs of a given category and

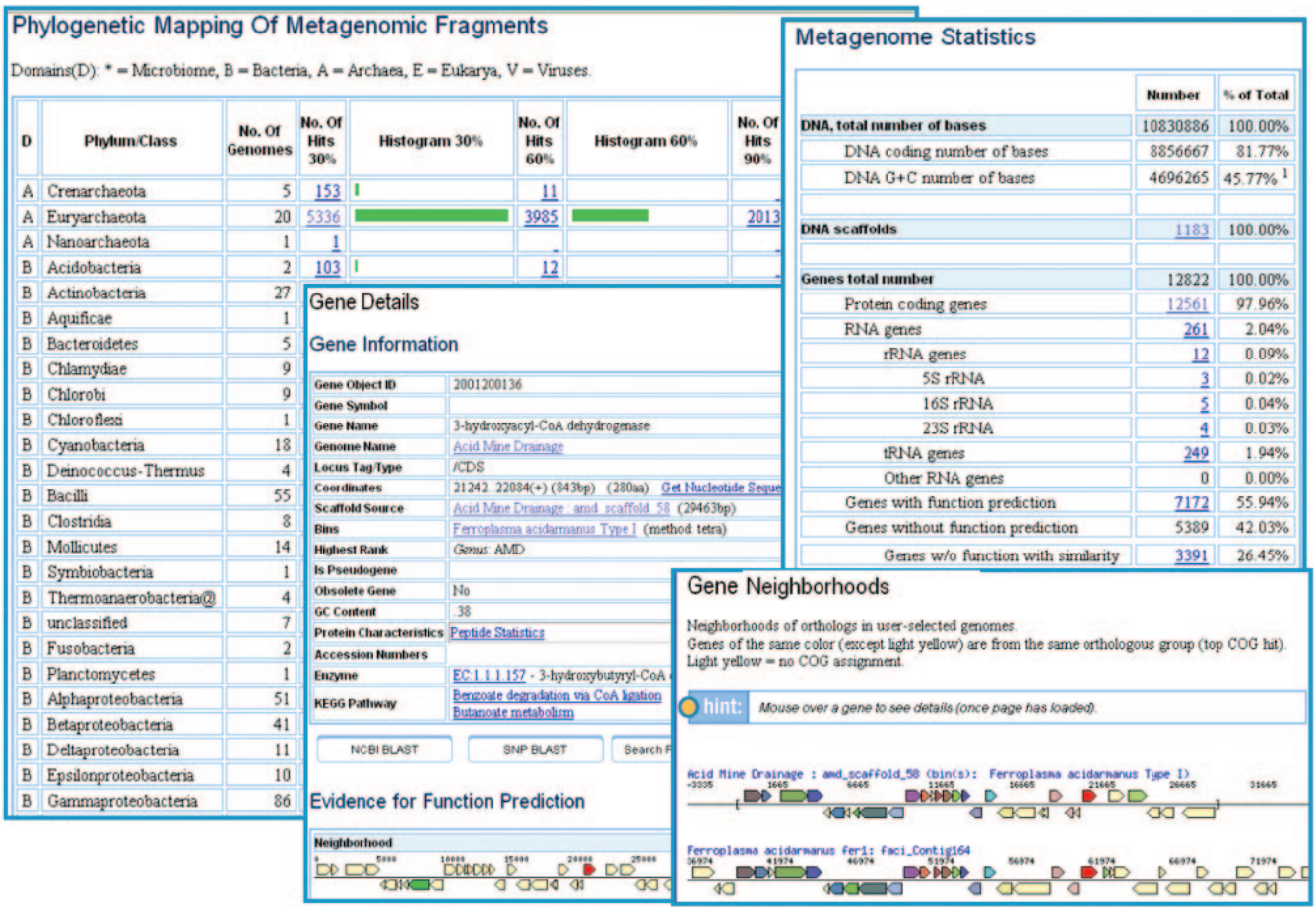

Fig. 1. AMD Microbiome Details: Metagenome Statistics and Phylogentic Mapping of Fragments. Gene Details and Gene Neighborhoods Example for an AMD Metagenome Gene.

the number of organisms that have genes belonging to each COG. For a given COG, the ''organism counts'' are linked to a list of organisms and their associated ''gene counts''. Gene counts for all COGs in a given category can be displayed for multiple organisms using *COG Profile*. KEGG pathways can be explored in a similar manner using *KEGG Pathway Details* and *Enzyme Profile*. *COG Profile* and *Enzyme Profile* are further discussed below.

4.2.2 Comparative Data Analysis The gene content of metagenomes and genomes can be examined with a profile-based selection tool, gene neighborhood analysis tools, and multiple sequence alignment tools. Functional annotations can be examined with several occurrence and abundance profile-based tools. We discuss below in more detail the profile based selection, occurrence profile, and abundance profile tools.

The *Phylogenetic Profiler* tool allows comparing the gene content of a target entity (microbiome, bin, or isolate organism) to that of other entities (microbiomes, bins or organisms) by defining a *profile* for the genes of the target entity in terms of presence or absence of homologs in other entities. Similarity cutoffs can be used to fine-tune the selection. Similar to isolate genomes, differences in gene content between metagenomes can be correlated with a specific phenotype or environment, while the comparison of the gene content of bins within the metagenome helps inferring the metabolic capabilities of the component populations and identify the organisms that may be responsible for community-essential tasks. The example shown in Figure 2 illustrates how the *Phylogenetic Profiler* helps finding differences in gene content between the component populations in the Acid Mine Drainage (AMD) microbiome. In this example, genes in the bin corresponding to *Leptospirillum* sp. group III that have no homologs in other bins in this metagenome are identified. Among the ''unique'' genes in *Leptospirillum* sp. group III one can find those responsible for nitrogen fixation, shown in the *Phylogenetic Profiler Results* pane of Figure 2, which makes this organism a keystone species in the AMD microbiome due to limitation of external nitrogen sources (Tyson *et al*., 2004).

Occurrence profile tools allow examining profiles of genes and functions across microbiomes, bins, and isolate organisms. Gene occurrence profiles usually involve genes within the same bin or organism: if such genes have similar occurrence profiles across other bins or organisms, then they may also have a similar evolutionary history and may potentially be functionally linked, or co-regulated in a pathway (Bowers *et al*., 2004). The profile for a gene x, across bins or organisms y_1 to y_n has the form of a vector $(L_1, \ldots, L_n)$ where L_i represents a set of y_i genes that are

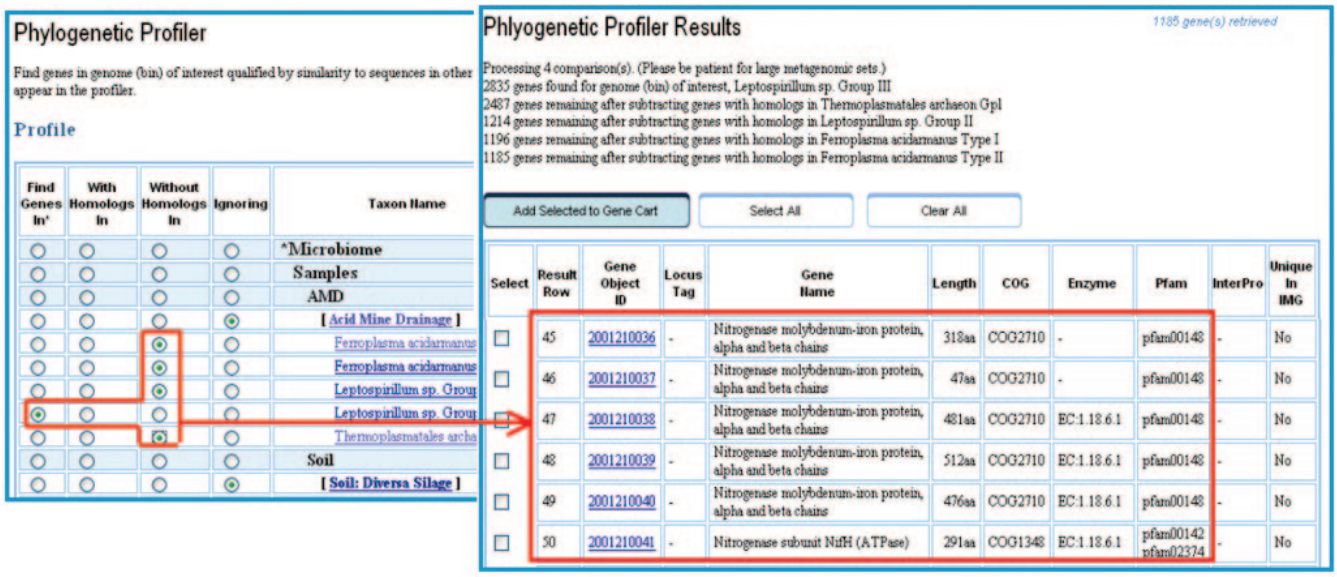

Fig. 2. Finding Gene Content Differences with the Phylogenetic Profiler Between AMD *Leptospirillum* sp. group III Bin and other AMD Bins.

associated with x, where the association of y_i genes with x is based on a specific sequence similarity method.

Functional occurrence profile tools, such as *COG Profile*, *Pfam Profile*, and *Enzyme Profile*, show the occurrence profiles for functional characterizations such as COGs, Pfam families, or enzymes involved in pathways across the selected entities (microbiomes, bins and organisms). Individual COGs, Pfam families, or enzymes are selected using a variety of search and browse tools and are maintained using *COG*, *Pfam*, and *Enzyme Carts*, respectively.

The occurrence profile for a specific function, f, shows the pattern of f across the selected entities, y_1 to y_n, in the form of a vector of the form $(L_1, \ldots, L_n)$, where L_i represents the set of y_i genes that are associated with f. Functional occurrence profiles provide an estimate of the similarity between entities in terms of association with a specific pathway or functional characterization.

The example shown in Figure 3 illustrates how occurrence profiles for a custom list of Pfam families can be used to predict the presence of a pathway for CO_2 fixation in metagenome data sets. The first step in one of CO_2 fixation pathways is catalyzed by anaerobic carbon monoxide dehydrogenase. A keyword search on expression ''CO dehydrogenase'' with Pfam as a filter (see *Search Terms and Pathways* pane of Figure 3) retrieves a list of six Pfam families, as shown in the *Function Search Results* pane of Figure 3. Four of these Pfam families correspond to different subunits of anaerobic carbon monoxide dehydrogenase, and therefore are selected and saved using the *Pfam Cart*. The occurrence profiles for these Pfam families are then computed and displayed in a tabular form as shown in the *Pfam Profile* pane of Figure 3, with each row displaying the profile of a specific Pfam across three whale-fall microbiomes and five bins of the AMD microbiome. Each cell in the profile result table contains a link to the associated list of genes and displays the count (*abundance*) of genes in the list. Colors are used to represent visually gene abundance, whereby white, bisque and yellow represent gene counts of 0, 1-4, and over 4 respectively. The occurrence profiles shown in Figure 3 indicate that, despite the presence of several spurious hits, anaerobic CO dehydrogenase is most likely absent from the organisms in the AMD microbiome and therefore these organisms probably rely on some other pathway of CO_2 fixation. Surprisingly, the genes coding for anaerobic CO dehydrogenase appear to be present in 2 out of 3 whale-fall microbiomes, as shown in in Figure 3. Occurrence profile tools provide two (functions vs. genomes, genomes vs. functions) display options for data visualization purposes.

An *Abundance Profile* tool allows comparing functional occurrence profiles for all COGs, Pfam families, or KEGG enzymes across microbiomes, bins, and isolate organisms of interest. This tool is especially useful for analysis of datasets obtained from the communities with high species diversity, where little or no sequence assembly can be achieved: for such datasets identification of predominant protein families allows users to infer habitat-specific biological traits.

The example in Figure 4 shows the abundance profiles of COGs displayed using a heat map, across the low-complexity AMD microbiome and the highly complex soil and whale-fall microbiomes. Arrows indicate COGs that are clearly overrepresented in the soil microbiome (bright red) as compared to other microbiomes (pink, orange, yellow and green); both COGs correspond to *glycosyl hydrolases* of different specificity. One would indeed expect to find glycosyl hydrolases abundant in microbiomes, such as those found in soil, that perform degradation of plant-derived carbohydrate polymers.

IMG/M also provides a tool for analysis of *strain-level heterogeneity* within a species population in metagenome data. *SNP BLAST* allows users to run BLASTn of nucleotide sequence of genes or scaffolds of interest in a metagenome, against a database of sequencing reads that were assembled to produce a composite species genome sequence comprised of multiple strains sequence types.

5 CONCLUSION

We have presented in this paper IMG/M, an experimental metagenome data management and analysis system. IMG/M provides support for the exploration and comparative analysis of

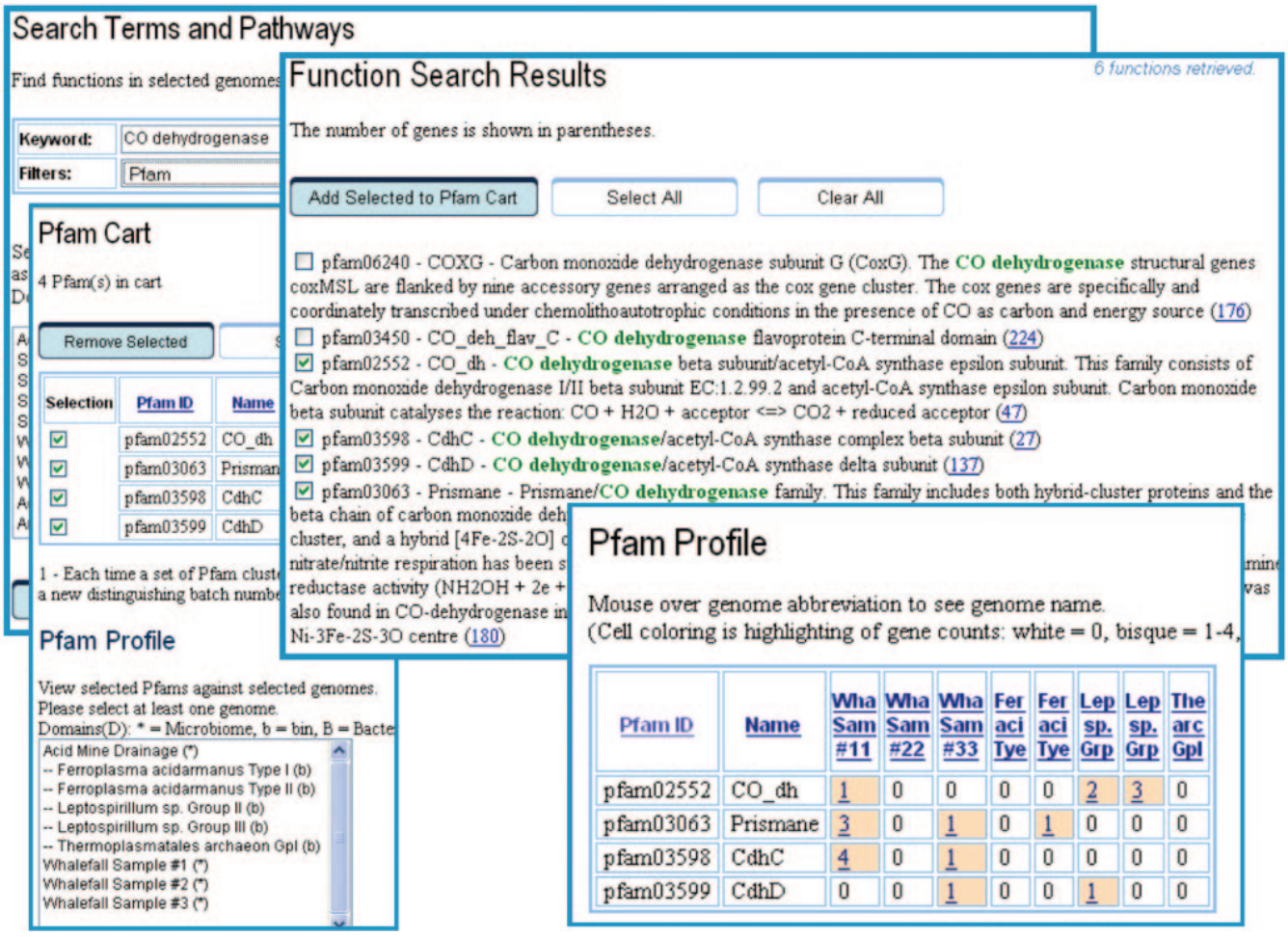

Fig. 3. Exploring the Presence of a Pathway for CO_2 Fixation Across Several Metagenomes with *Pfam Profile*.

metagenomes and their component populations in the context of other metagenomes and isolate genomes. IMG/M has been successfully used for the study of EBPR sludge communities (Martin *et al.*, 2006), and continues to be used for analyzing metagenomes sequenced at JGI, such as the *Olavius algarvensis* symbionts[1] and the termite gut microbial community[2]. Although IMG/M seems to be best suited for the analysis of low-complexity microbiomes, the system can be also used to infer the presence of important physiological characteristics in any microbiome and its species populations.

We plan to extend the tools provided by IMG/M in order to address several metagenome data analysis challenges. The first challenge regards the size and complexity of some metagenome data sets. Additional viewers need to be developed in order to improve the efficiency of analyzing such data sets via graphical representation of phenomena of interest set in a biological context.

A second challenge is posed by existing methods for binning metagenome scaffolds. These methods are in an early stage of development and have not been properly tested on metagenomes of complex microbiomes. We have found that some of these methods do not perform well even when applied to low diversity microbiomes in IMG/M, resulting in a significant number of unclassified or misclassified scaffolds. While analysis of environmental microbiomes is often function-driven and focuses on the genes and metabolic pathways of interest regardless of their assignment to a certain species, binning of scaffolds is essential for drawing a connection between the presence of certain genes (e.g. pathogenicity factors) and species composition of a microbiome. Consequently, there is an immediate need for tools that would provide support for comparing different binning methods and for assessing their accuracy, as well as for revising bins in terms of scaffold composition and gene content. We plan to develop tools for reviewing and curating the content of bins in IMG/M.

Finally, metagenome analysis tools need to be extended in order to account for the stochastic nature of metagenome data and variations in data quality due to incomplete sequence coverage. In most microbiomes a few dominant species tend to get the most sequencing coverage, sometimes approaching that of draft isolate genomes, while low abundance organisms can be represented by a small number of scaffolds or even single sequencing reads. Accordingly,

[1]http://www.jgi.doe.gov/sequencing/why/CSP2005/algarvensis.html
[2]http://www.jgi.doe.gov/sequencing/why/CSP2006/termitegut.html

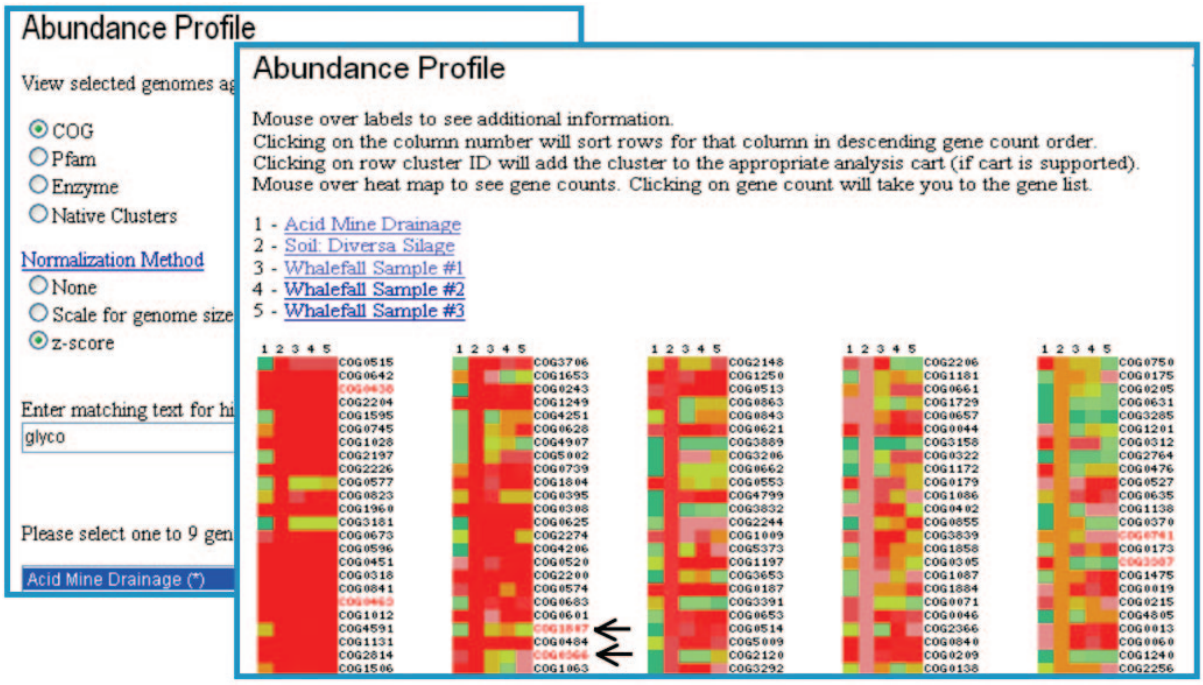

Fig. 4. Use of Abundance Profiles to Identify COG Families Overrepresented in the Soil Metagenome.

statistical tests need to be devised to estimate the sequence coverage of the bins and whether it is adequate for certain types of comparative analyses, such as metabolic reconstruction of pathways. Additionally, when metagenomes are compared to each other or to isolate genomes, statistical tests are needed for estimating the *statistical significance* of the observed differences. For example, the analysis of *Abundance Profiles* described above requires testing whether the differences in abundance can be ascribed to chance variation or not.

We also plan to extend the data model underlying the system in order to enhance its ability to capture *metadata* characterizing microbiome samples. Such metadata are often specific to an application (e.g., biomedical, ecological) domain. Samples are associated with properties used for metagenome analysis, such as sample structural and morphological characteristics (e.g. sample site, time of collection) and donor or host data (e.g. demographic and clinical record, including diagnosis, disease, stage of disease, and treatment information for human donors). Samples may also be involved in clinical studies and therefore can be grouped into several time/treatment study groups. In addition to extending the data model for supporting sample metadata, we plan to improve the coherence and completeness of these annotations via manual curation. In IMG/M, metadata such as disease, phenotype, ecotype and relevance for the isolate genomes were collected from sources such as GOLD (Liolios *et al*., 2006), while the microbiome sample metadata have been collected from published supplemental information and manually curated. The scarcity of metadata for isolate organisms and microbiome samples is a well known problem (Field and Hughes, 2005). We plan to collaborate with community standardization efforts in the metagenome data domain in order to ensure high coverage and consistence of microbiome sample metadata.

ACKNOWLEDGEMENTS

We thank all our colleagues who have contributed to the development of IMG and IMG/M. Special thanks to Eddy Rubin and James Bristow for their encouragement throughout this project. The work presented in this paper was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J. *et al*. (2002) Whole Genome Shotgun Assembly and Analysis of the Genome of Fugu rubripes. *Science*, **297**, 1301–1310.

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V. *et al*. (2004) The Pfam Protein Families Database. *Nucleic Acids Research*, **32**, D138–D141.

Bowers,P.M., Pellegrini,M., Thompson,M.J., Fierro,J., Yeates,T.O., Eisenberg,D. *et al*. (2004) Prolinks: A Database of Protein Functional Linkages Derived from Coevolution. *Genome Biology*, **5**.

Chen,K. and Pachter,L. (2005) Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Computational Biology*, **1**(2), e24.

Delcher,A.L., Harmon,D., Kasif,S., White,O. and and Salzberg,S.L. (1999) Improved Microbial Gene Identification with Glimmer. *Nucleic Acids Research*, **27**(23), 4636–4641.

DeLong,E.F. and Karl,D.M. (2005) Genomic Perspectives in Microbial Oceanography. *Nature*, **437**, 336–342.

Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An Efficient Algorithm for Large-Scale Detection of Protein Families. *Nucleic Acids Res*., **30(7)**, 1575–1584.

Field,D. and Hughes,J. (2005) Cataloguing our Current Genome Collection. *Microbiology*, **151(4)**, 1016–1019.

Gene Ontology Consortium. (2004), The Gene Ontology Database and Informatics Resource. *Nucleic Acids Research*, **32**, 258–261.

Gordon,J.I., Ley,R.E., Ruth,E., Ley, Wilson,R., Mardis,E., Xu,J., Fraser,C. and Relman,D.A. (2005) Extending Our View of Self: the Human Gut Microbiome Initiative (HGMI), http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/HGMISeq.pdf

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y., Hattori,M. *et al.* (2004) The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research*, **32**, D277–D280.

Kanz,C., Aldebert,P., Althorpe,N. *et al.* (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acid Research*, **33**, D29–D33.

Kersey,P., Bower,L., Morris,L. *et al.* (2005) Integr8 and Genome Reviews: Integrated Views of Complete Genomes and Proteoms. *Nucleic Acid Research*, **33**, D297–D302.

Liolios,K., Tavernarakis,N., Hugenholtz,P., Kyrpides,N.C. *et al.* (2006) The Genomes On Line Database (GOLD) v.2: A Monitor of Genome Projects Worldwide. *Nucleic Acid Research*, **34**, D332–D334.

Markowitz,V.M., Korzeniewski,F., Palaniappan,K., Szeto,E. *et al.* (2006) The Integrated Microbial Genomes (IMG) System, *Nucleic Acids Research*, **34**, D344–D348.

Martin,H.G., Ivanova,I., Kunin,V., Warnecke,F. *et al.* Genetic Blueprints for Phosphorus Removal from Sludge Based on Metagenomic Sequencing. Submitted for publication. See http://www.jgi.doe.gov/sequencing/why/CSP2005/PO4accum.html.

Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A. *et al.* (2005) InterPro, Progress and Status in 2005. *Nucleic Acids Research*, **33**, D201–D205.

Osterman,A., Overbeek,R. *et al.* (2003) Missing Genes in Metabolic Pathways: A Comparative Genomic Approach. *Chemical Biology*, **7**, 238–251.

Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts, and Proteins. *Nucleic Acid Research*, **33**, D501–D504.

Riesenfeld,C.S., Schloss,P.D. and Handelsman,J. (2004) Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, **38**, 525–552.

Smith,C.R. and Baco,A.R. (2003) Ecology of Whale Falls at the Deep-Sea Floor. *Oceanography and Marine Biology: an Annual Review*, **41**, 311–354.

SoftBerry, FGENESB Suite of Bacterial Operon and Gene Finding Programs, http://www.softberry.com/berry.phtml

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631–637.

Tringe,S., von Mering,C., Kobayashi,A., Salamov,A., Chen,K. *et al.* (2005) Comparative Metagenomics of Microbial Communities. *Science*, **308**, 554–557.

Tyson,G.W., Chapman,J., Hugenholtz,P. *et al.* (2004) Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. *Nature*, **428**, 37–43.

Venter,J.C., Remington,K., Heidelberg *et al.* (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304**, 66–74.