Descriptions of a Selected Sampling of Databases Having Relevance to the GTL Knowledgebase

Several existing databases have created effective systems for storing and analyzing genomic, metagenomic, proteomic, and other data. Having implemented successful data analysis tools, information management strategies, user capabilities, and architectures, these systems can provide viable examples of components envisioned for the GTL Knowledgebase. Moreover, many of these databases will provide important supplements to and links with other GKB capabilities. Descriptions of several such systems and their features follow.

Integrated Microbial Genomes (IMG, http://img.jgi.doe.gov). Developed through collaboration between the Department of Energy's (DOE) Joint Genome Institute (JGI) and the Biological Data Management and Technology Center at DOE Lawrence Berkeley National Laboratory, DOE's IMG is a data management, analysis, and annotation platform that enables the efficient comparative analysis of all complete public microbial genomes, draft or finished, produced at JGI and throughout the world. IMG currently integrates data from 4570 genomes (1155 bacteria, 56 archaea, 40 eukaryotes, 932 plasmids, and 2387 viruses), consisting of more than 4.9 million genes, with publicly available metabolic pathway collections and protein family information. IMG offer various tools for comparing genes, pathways, and functions across genomes; visualizing the physical distribution of genes within genomes; investigating the evolutionary history of genes; and developing user-defined pathways and functional categories to aid the analysis of poorly characterized genomes.

In addition to supporting the analysis of complete genomic sequences data from microbial isolates, the **Integrated Microbial Genomes with Microbiome Sampling (DOE IMG/M, http://img.jgi.doe.gov/m)** portal supports comparative analyses of more than 40 community sequences generated with various metagenomic sequencing technologies and data processing methods. IMG/M allows examination of profiles of functional annotations across microbial communities and isolate organisms of interest as well as analysis of strain-level heterogeneity within a species population in metagenomic data.

Phytozome (http://www.phytozome.net). This tool for green plant comparative genomics is a joint project of the DOE Joint Genome Institute and the Center for Integrative Genomics at the University of California, Berkeley. Phytozome provides access to nine sequenced and annotated green plant genomes, including poplar, grape, sweet sorghum, rice, soybean, green algae, moss, spikemoss, and the small flowering plant *Arabidopsis*. Clusters of orthologous and paralogous genes that represent the modern descendents of ancestral gene sets can be analyzed to explore genes associated with significant evolutionary events related to the development of green plants.

SEED (http://www.theseed.org). The SEED Project is a cooperative effort to annotate ever-expanding genomic data so researchers can conduct effective comparative analyses of genomes. Launched in 2003 by the Fellowship for Interpretation of Genomes (FIG), the project is one of several initiatives in ongoing development of data curation systems. SEED is designed to be used by scientists from numerous centers and with varied research objectives. As such, several institutions have since joined FIG in a consortium, including the University of Chicago, DOE's Argonne National Laboratory (ANL), the University of Illinois at Urbana-Champaign, and others. As one example, ANL has used SEED to develop the National Microbial Pathogen Data Resource. Other agencies and institutions have used the project to discover genome components and clarify gene functions such as metabolism. SEED also has enabled researchers to conduct comparative analyses of closely related genomes and has supported derivation of stoichiometric models to understand metabolic processes.

The SEED Project has been extended to support metagenomic samples and concomitant analytical tools. Moreover, the number of genomes being introduced into SEED is growing very rapidly. Building a framework to support this growth while providing highly accurate annotations is centrally important to SEED. The project's subsystem-based annotation strategy has become the technological foundation for addressing these challenges.

Proteomics Research Information Storage and Management (DOE PRISM, http://ncrr.pnl.gov/about/process.stm). The PRISM system enables the management and analysis of high-throughput proteomic data at DOE's Pacific Northwest National Laboratory (PNNL). The laboratory's proteomics capability, based on high-performance mass spectrometry, combines commercially available informatics systems with those developed at PNNL. Streams of raw data from 15 mass spectrometers are joined and analyzed via an automated workflow designed to transform raw instrument data into useful biological information, such as which proteins an organism produces under different growth and stress conditions.

PRISM—which contains more than 100,000 datasets and 1 billion mass spectra—manages data, metadata, and analysis workflows; maintains a sample request and queuing system, reports research results; and provides a user interface for data searches and queries. Since its initial development in 2000, DOE's PRISM has undergone several advancements, including (1) addition of a mass tag system, high-capacity storage servers, and a plug-in architecture for automated analysis tools; (2) improved peptide and protein identification; (3) use of archived data from DOE's Environmental Molecular Sciences Laboratory; and (4) automated interaction between computers for PRISM's data management system and liquid chromatography cart.

MicrobesOnline (http://www.microbesonline.org). Since 2003, this integrated Web portal has enabled scientists to conduct comparative genomic analyses. MicrobesOnline currently contains more than 700 complete microbial genome sequences and offers a suite of analysis tools, including (1) a multispecies genome browser, (2) operon and regulon prediction methods, (3) a combined gene and species phylogeny browser, (4) a gene expression data browser with expression profile searches, (5) a Gene Ontology browser, and (5) sequence analysis capabilities (e.g., sequence motif detection, motif searches, sequence alignment, and phylogeny reconstruction).

Funded by DOE, the database also offers resources for community annotation, integrates functional genomic data, and provides novel Web-based viewing and analysis tools for proteomic, gene expression microarray, and phenotype microarray data. Interactive heat maps allow users to compare microarray data for microbes under multiple stress conditions. Users also can analyze correlations between gene expressions from different experiments. Among the major new features of MicrobesOnline is the ability to search the data compendium for genes with expression profiles similar to those resulting from query profiles.

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA, http://camera.calit2.net). The ability to explore large metagenomic datasets can enhance the research of microbial ecologists. Until recently, large-scale metagenomic analysis has been limited by the availability of computational resources that provide scientists with easy-to-use, scalable, and fully integrated Web frameworks. One such resource—CAMERA—features a rich data repository, associated bioinformatic tools, and cyberinfrastructure for conducting such analyses.

CAMERA was launched in 2007 as a collaboration between the University of California at San Diego's Calit2 division and the J. Craig Venter Institute (JCVI). With funding from the Gordon and Betty Moore Foundation, CAMERA contains 12 metagenomic datasets consisting of 14 million genomic fragment sequences. Genomic data are layered with associated geographical, temporal, and physicochemical metadata to assist in metagenomic analyses. Additional capabilities enable homology identifications using Basic Local Alignment Search Tool (BLAST), an algorithmic resource for sequence comparisons. Furthermore, CAMERA provides graphical tools for viewing sequence regions that indicate genomic conservation and divergence and for correlating such regions with environmental factors. This new platform allows microbial researchers to begin to analyze large-scale sampling and sequencing endeavors such as JCVI's Global Ocean Sampling expedition.

Comprehensive Microbial Resource (CMR, http://cmr.jcvi.org). Containing more than 600 sequenced prokaryotic genomes, the CMR database provides researchers with information on inter- and intragenomic relationships for comparative genomics, genome diversity, and evolutionary studies. CMR—which is operated by JCVI—enables a wide variety of data retrievals and offers scientists numerous analytical tools for exploring the system's prokaryotic genomes. These data retrievals can be based on different gene properties that include molecular weight, hydrophobicity, guanine-cytosine (GC) content, functional-role assignments, and taxonomy. The system also has special Web-based analysis tools for precomputed homology searches, whole-genome dot plots, batch downloads, and searches across genomes using various data types.

Since consistent annotation is essential for robust genomic comparisons, CMR features primary annotations, as assigned by GenBank, and secondary annotations provided by JCVI. In addition, CMR provides comprehensive views of genes and gene annotations, genome-level structures, pathway maps, codon usage tables, GC plots, the ability to generate and visualize whole-genome alignments between two bacteria, and tabulated summary data from both individual genomes and CMR's entire genome collection.

Pathema (http://pathema.jcvi.org). As the Web resource for JCVI's Bioinformatics Resource Center, Pathema provides detailed curation of six target pathogens: *Bacillus anthracis*, *Clostridium botulinum*, *Burkholderia mallei*, *Burkholderia pseudomallei*, *Clostridium perfringens*, and *Entamoeba histolytica*. Funded by the National Institute of Allergy and Infectious Disease (NIAID), the center is one of eight designed to support biodefense and infectious disease research. Initially developed at JCVI, Pathema is co-maintained by the Institute of Genome Sciences (IGS) at the University of Maryland School of Medicine.

The Pathema website is separated into four main taxonomic clades: *Bacillus*, *Burkholderia*, *Clostridium*, and *Entamoeba*, allowing developers to customize clade-specific sites to each research community's needs. Pathema's dataset includes Gene Ontology assignments, metabolic pathway identification, transporter characterization, and predicted ortholog analysis and identification. The center's overarching goal is to provide a core online resource to accelerate scientific progress in understanding, detecting, diagnosing, and treating several categories of NIAID priority pathogens and other agents involved in new and re-emerging infectious diseases. Bioinformatics software with significant new capabilities, novel data types, Web resources, and analysis tools specifically geared toward biodefense are available on Pathema. Such capabilities—including intergenomic comparisons—help identify potential targets for vaccine development, therapeutics, and diagnostics. The site also serves as a focal point for the biodefense research community by disseminating data from bacterial genome-sequencing projects and by providing access to results of intergenomic comparisons.