The Integrated Microbial Genomes (IMG) System: A Case Study in Biological Data Management

Victor M. Markowitz[1], Frank Korzeniewski[1], Krishna Palaniappan[1], Ernest Szeto[1],
Natalia Ivanova[2], and Nikos C. Kyrpides[2]

[1] Biological Data Management and Technology Center
Lawrence Berkeley National Laboratory
1 Cyclotron Road, Berkeley, CA 94720, USA
{vmmarkowitz, frkorzeniewski, kpalaniappan, eszeto}@lbl.gov

[2] Microbial Genome Analysis Program
Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, CA 94598, USA
{nckyrpides, nnivanova}@lbl.gov

Abstract

Biological data management includes the traditional areas of data generation, acquisition, modelling, integration, and analysis. Although numerous academic biological data management systems are currently available, employing them effectively remains a significant challenge. We discuss how this challenge was addressed in the course of developing the Integrated Microbial Genomes (IMG) system for comparative analysis of microbial genome data.

1. Introduction

Problems related to biological data management systems have been examined extensively over the past decade. These problems are usually discussed in terms of novel methods and technologies needed for developing biological data management systems. For example, a recent report discusses the need for extending database technology to support biological data types, provenance, evolution, and integration [12]. In practice, hundreds of commercial and public biological databases have been developed using existing data management technology [8]. Most problems with these databases regard effective use of, rather than deficiencies with, existing technologies [27].

One of the key goals for biological data management systems is to provide support for data analysis, which often involves exploring data across multiple heterogeneous data sources. Data warehousing and data federation technologies have been employed for handling syntactic heterogeneity, that is, differences in data structure and formats, across diverse biological data sources, as discussed in [6], [17], and [19]. Effective data analysis, however, also needs to support seamless flow (composition) of analysis operations, while addressing semantic heterogeneity, that is, differences in the meaning of related data items (objects). Providing such support presents a significant challenge for biological data management systems, especially for those developed in academic settings.

Biological data management systems in academic settings were originally confined to relatively small individual scientific groups or laboratories: these systems were often limited to specialized data sets and analysis operations and were developed without considering data analysis workflows, heterogeneity, evolution, and scalability issues. Addressing such problems requires a systematic process for analyzing the data structure and operations for the application domain. This process entails substantial documentation which is especially difficult to maintain for biological data whose semantics are complex and tend to evolve. These data are generated via processes that involve multiple transformations between different levels of data granularity and are based on evolving technology platforms and computational methods. In spite of this complexity, a systematic application domain analysis process and comprehensive documentation are essential for providing effective data analysis support and

Proceedings of the 31st VLDB Conference,
Trondheim, Norway, 2005

thus address the frustration scientists often encounter in dealing with public biological data management systems [12]. In this paper we discuss how this challenge has been addressed in the development of the Integrated Microbial Genomes (IMG) system.

The development process for IMG is based on established practices and starts with application domain analysis, followed by abstract data model definition, system design and implementation. Application domain analysis is based on requirements gathered from biologists and entails detailed use case scenarios that serve as a vehicle for bridging the rather steep communication gap between these scientists and data management system developers. Application domain analysis is used for defining an abstract microbial genome data model in terms of data types and operations. This data model then serves as the foundation for the design and development of the data management system.

IMG is the result of the collaboration between the scientists of the Microbial Genome Analysis Program (MGAP) at the Joint Genome Institute (JGI) and members of the Biological Data Management and Technology Center (BDMTC) at Lawrence Berkeley National Laboratory. The IMG case study is instructive since it deals with genomic sequence data generated using established technologies and methods. Systems that deal with data generated using newer technology platforms and methods, such as gene expression and proteomic data, are likely to encounter similar or more complex challenges. Furthermore, MGAP scientists and BDMTC engineers had prior experience in developing both academic and commercial large scale biological data management systems. Their combined experience was not enough to avoid the communication problems mentioned above, but was essential in following the process required to address these problems.

A public version of IMG that supports microbial genome data analysis was released in March 2005 [11]. An enhanced version of IMG, with additional support for genome data curation (editing) is used at JGI for improving the quality of annotations for newly sequenced microbial genomes.

In the following sections we present a brief overview of the microbial genome data application, and then discuss gathering and analyzing application requirements for IMG. Next, we present the abstract data model that has resulted from analyzing these requirements, whereby microbial genome data are modelled as a multidimensional data space. Finally, we show how this data model was used for developing IMG analysis tools that support exploring microbial genome data along individual or across multiple dimensions.

2. Microbial Genome Application

According to the Genomes OnLine Database, about two hundred microbial genomes have been sequenced to date, with 530 other projects ongoing and more in the process of being launched [2]. Microbial genome analysis is a growing area that is expected to lead to advances in healthcare, environmental cleanup, agriculture, industrial processes, and alternative energy production [26].

2.1 Microbial Genome Data Types

Microbial genome data captures information about raw DNA sequence data, along with *genes* characterized in terms of *functions* and *pathways*.

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). Characterizing a gene consists of determining its biological context, including its location on a chromosome within a (species specific) *genome*, and its associated *functional* roles in cellular *pathway*s. A key characteristic for genome is its taxonomic (*phylogenetic*) lineage, including its *domain*, *phylum*, *class*, *order*, *family*, *genus*, *species* and *strain* [25].

Pathways can be viewed as ordered lists of reactions, whereby each reaction involves compounds which are reactants (substrates, products), catalyzed by enzymes. Pathways can be combined in pathway *networks*, whereby pathways can be associated via reactions that share common components. Pathways are associated with genes via gene products that function as enzymes that serve as catalysts for individual reactions of metabolic pathways [15]. Accordingly, pathways provide a biologically meaningful framework for examining functional relationships between genes, rather than individual gene functions.

2.2 Microbial Genome Annotation

Microbial genome annotation generally refers to a process of assigning biological meaning to the raw sequence data by identifying gene regions or functional features and determining their biological functions. Gene annotation is a combination of automated methods that generate a "preliminary" annotation in terms of predicted genes (also called Open Reading Frames or ORFs, which represent the sequence of DNA or RNA located between the start-codon and stop-codon sequence) and associated functions and pathways based on sequence similarity or profile searches.

The result of a preliminary (baseline) annotation is often sparse, with numerous genes not having associated functions or pathways. Consequently, several techniques are employed for further annotating genes as well as validate baseline annotations. The most effective annotation techniques involve comparative multi-genome analysis based on observed biological evolutionary phenomena: pairs of genes with related (coupled) functions (1) are often both present or both absent within genomes; (2) tend to be collocated (on chromosomes) in multiple genomes; (3) might be fused into a single gene in

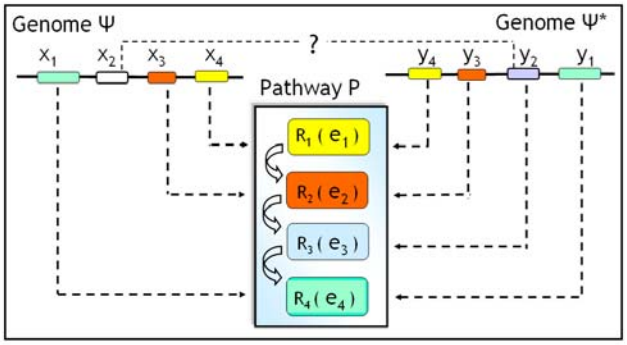


Figure 1. Sketch of Genomes Associated with a Pathway.

some genomes; or (4) are components of an operon (a set of genes transcribed as a unit under the control of an operator gene) [4].

Consider the example shown in Figure 1, where pathway P involves reactions R_1, R_2, R_3, and R_4: genes x_1, x_3, and x_4 of genome Ψ are associated with pathway P via enzymes e_1, e_3, and e_4, respectively; genes y_1, y_2, y_3, and y_4 of genome Ψ^* are associated with pathway P via enzymes e_1, e_2, e_3, and e_4, respectively; if gene x_2 is similar (i.e., determined to be related via significant sequence similarity) to gene y_2, then, following rules above, x_2 may be associated with P via enzyme e_2.

2.3 Microbial Genome Data Sources

Microbial genomes are sequenced by organizations worldwide, follow an annotation process similar to that mentioned above, and end up in one of several microbial genome data sources, such as EBI Genome Reviews [14], CMR[23], and RefSeq [24]. Furthermore, additional genome annotation details such as protein families and pathways reside in multiple specialized data sources such as UniProt (protein sequences and functions), InterPro (protein families and domains), COG (clusters of orthologous genes), and KEGG (pathway maps). Consequently, analyzing microbial genome data entails integration of data from diverse, usually heterogeneous, data sources.

It is important to distinguish between "shallow" and "deep" integration of biological data. The former amounts to "data sorting and collating" and does not address semantic problems between individual data items [5], while the latter involves identifying and matching data items (objects) in different data sources that may represent the same underlying biological objects, such as genes. Resolving semantic heterogeneity between diverse biological data sources is a complex problem. For example, a protein sequence is represented in data sources such as GenBank and SwissProt using different accession numbers to identify it and different terms to characterize it. Consequently, mapping objects across data sources may require expert scientific review of individual objects.

Effective comparative analysis of microbial genome data requires a coherent view of biological data and therefore involves "deep" data integration. Different microbial genome data sources provide a variety of alternative or fragmented views of an inherently incomplete and imprecise data domain. These sources share common goals but contain different collections of genomes or data with different degrees of resolution regarding the same genomes. These differences are the result of diverse annotation methods, curation techniques, and functional characterization employed across microbial genome data sources. An additional problem in dealing with these sources is the difficulty of determining the coherence and completeness of their data.

Data *coherence* regards the quality of annotations: although inherently imprecise, these annotations can be qualified in terms of "biological coherence" rules. For example, predicted genes with overlapping sequences often indicate errors in gene prediction and need to be manually reviewed and corrected. Problems related to data coherence are caused by the high cost in terms of time and expertise needed to validate and correct annotations manually.

Data *completeness* regards the extent and coverage of functional characterization and depends on the diversity of the genomes included in a data source and the depth of integration of genome annotations collected from diverse sources [20]. Problems related to data completeness are caused by the complexity of "deep" integration, which often requires complex expert scientific reviews to resolve semantic heterogeneity problems.

2.4 JGI Microbial Genome Data

The Joint Genome Institute (JGI) is one of the key sources of microbial genome sequence data, covering about 22% of the reported number of microbial genome projects worldwide. Individual microbial genomes are sequenced and assembled at JGI's production facility, producing data files with so called "draft" genome sequences [3]. Draft genomes are subsequently completed ("finished") by JGI's partners at Los Alamos National Lab and Stanford. Both draft and finished genomes pass through the automatic Genome Analysis Pipeline at Oak Ridge National Lab which identifies genes using gene prediction methods, and associates them with preliminary functional annotations, such as InterPro protein families and domains, COG categories, and KEGG pathway maps [10]. Finished genomes and their annotations are eventually published on individual genome portals [13].

Before publication, scientific groups interested in a specific genome further review and curate the microbial genome data in collaboration with JGI's Microbial Genome Analysis Program. The genome annotation and curation processes are greatly enhanced when individual microbial genomes can be analyzed in the comparative context of other genomes. Providing such a context is the main purpose of the Integrated Microbial Genomes (IMG) system. IMG aims at providing high levels of data diversity in terms of the number of genomes integrated in the system from public sources, data coherence in terms of the quality of the gene annotations, and data completeness in terms of breadth of the functional annotations. IMG also aims at providing a high level of comprehensibility in terms of documenting its data structural and operational semantics.

3. Microbial Genome System Requirements

We discuss below the process of analyzing system and application requirements for developing a biological data management system in the context of the Integrated Microbial Genomes (IMG) data management system [11].

Developing a biological data management system starts with the analysis of application domain requirements. This analysis is one of the most difficult problems for biological data management systems, and involves domain scientists who outline what they need in abstract, potentially ambiguous or vague, domain-specific terms. The key challenge is to translate the "what" of abstract application domain views into the "how" of data management system components. This process is prone to misinterpretation, may require reconciling conflicting views, and often involves numerous iterations. Furthermore, this process is time consuming and requires a reliable mechanism for clarifying questions between individuals who have different views of the application.

3.1 Data Content Requirements

Gathering and analysing requirements for IMG first involved its data content. A prototype database that included a representative set of microbial genome sequences and associated annotations from a variety of sources was developed for this purpose.

The key question addressed in analyzing data content requirements for IMG was finding a primary source of public microbial genomes with annotations that are not only extensive and accurate, but also amenable for integration with additional annotations available in other data sources. For example, the source initially considered for public microbial genome data, NCBI's RefSeq [24], had only sparse annotations (e.g., in terms of gene names, symbols, etc.), and poor cross references with additional sources of annotations, such as UniProt and InterPro. EBI's Genome Reviews [14] had better annotations and cross references than RefSeq, and therefore was selected as IMG's main source for public microbial genome data. It is worth noting that the quality of and issues with cross references between multiple biological data sources is not well documented and often requires extensive experimentation in collecting and integrating data from these sources. This problem is compounded by changes in the structure of biological data sources which range from occasional minor extensions to restructuring that may affect the semantics of the data. Furthermore, although correlated through mutual cross references, biological data sources tend to evolve on different schedules, which is another source of potential semantic inconsistencies.

3.2 Application Requirements

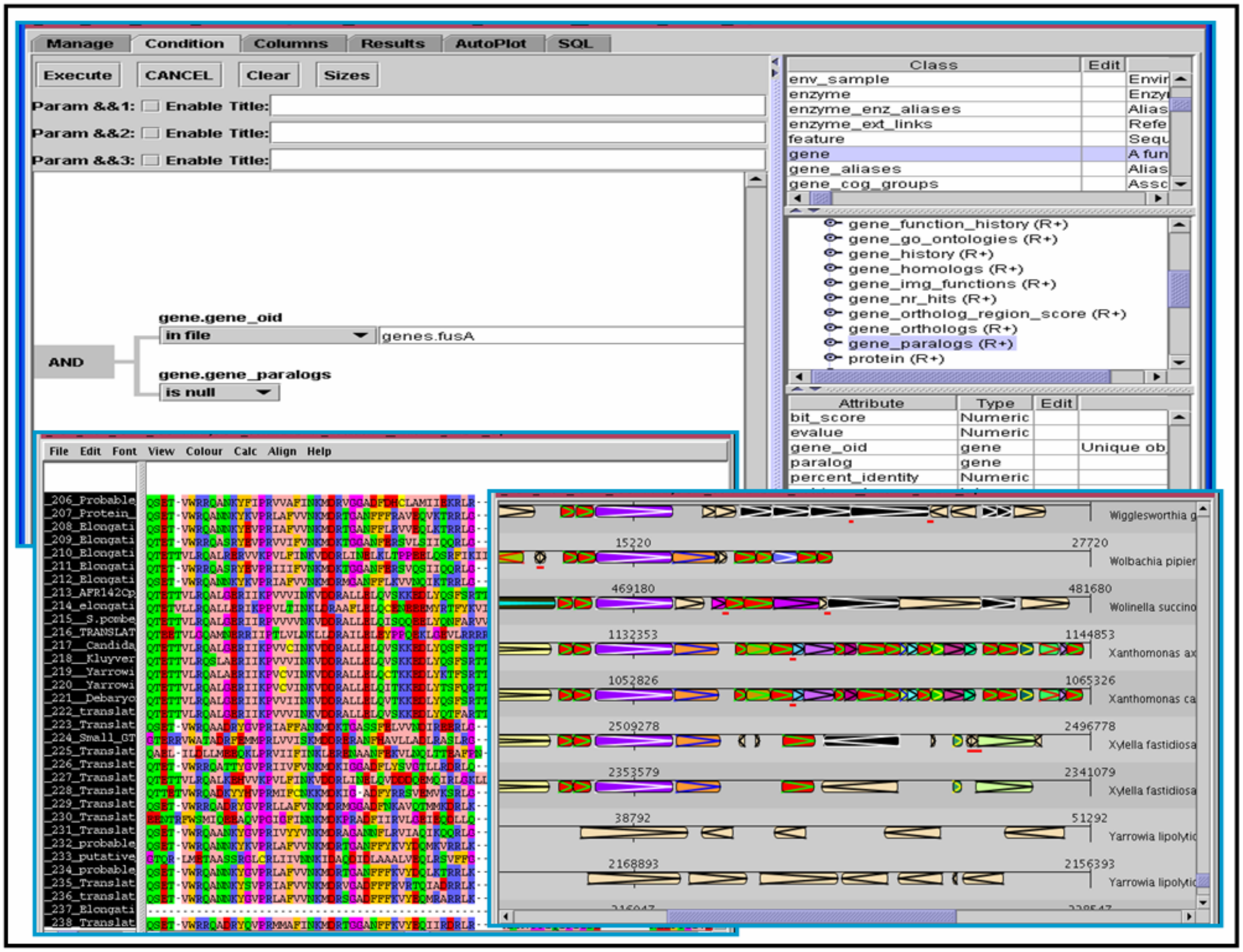
A second, equally important, aspect of analysing requirements for IMG regarded microbial genome data analysis. A prototype analytical tool was devised for examining, validating, refining, and documenting these requirements. This prototype was developed in the framework provided by the Apollo tool [16], and includes in addition to Apollo's native viewers additional visualization capabilities, for example for displaying genes on multiple genomes in a comparative context and for aligning DNA sequences. A key component of this prototype is a generic query constructor that allows experimenting with a variety of analysis workflows involving composition of individual operations.

For example, consider a typical microbial genome analysis that involves identifying and grouping genes that may belong to a particular protein family. Such an analysis entails: (a) finding the genes associated with a specific protein family, such as "*fusA*"; (b) identifying and eliminating so called "duplicate" genes associated with individual genomes - such genes may be *paralogs*, that is genes that result from gene duplication events and variation within the same species; (3) finding genes that have strong similarity with genes found in the previous steps - such genes may be *orthologs*, that is genes in different species that have the same evolutionary origin; (4) removing ortholog genes whose similarities are determined to be "false positives", by examining their aligned protein sequences.

Clarifying the requirements for this analysis involved using the query constructor as illustrated in the upper side of Figure 2, where class `gene` is first selected from a list of classes (see right upper side Class list), attributes such as `gene_oid` and `gene_paralogs` are then selected from the list of attributes associated with this class and added to the query tree, and finally conditions that involve selected attributes are specified. Queries can be saved, customized, and/or executed. Query results can be saved in local files and used in other queries, as shown in the example of Figure 2, where the condition for attribute `gene_oid` involves the result of a previous query, `genes.fusA`, that consists of genes associated with the *fusA* protein family.

Experimenting with alternative or related queries helps defining, validating, and documenting individual operations required to support genome data analysis, as well as defining analysis workflows in terms of individual operations. The documentation involves description of genome data analysis case scenarios, whereby specific operations are defined using set expressions as well as SQL queries underlying the query constructor, associated with concrete examples based on the prototype database.

Graphical visualization is critical in evaluating results



Param &&1: □ Enable Title:

Param &&2: □ Enable Title:

Param &&3: □ Enable Title:

AND

gene.gene_oid
in file genes.fusA

gene.gene_paralogs
is null

Class	Edit	
env_sample		Envir
enzyme		Enzy
enzyme_enz_aliases		Alias
enzyme_ext_links		Refe
feature		Sequ
gene		A fun
gene_aliases		Alias
gene_cog_groups		Assc

- gene_function_history (R+)
- gene_go_ontologies (R+)
- gene_history (R+)
- gene_homologs (R+)
- gene_img_functions (R+)
- gene_nr_hits (R+)
- gene_ortholog_region_score (R+)
- gene_orthologs (R+)
- gene_paralogs (R+)
- protein (R+)

Attribute	Type	Edit
bit_score	Numeric	
evalue	Numeric	
gene_oid	gene	Unique ob
paralog	gene	
percent_identity	Numeric	

Figure 2. IMG Prototype Analytical Tool.

of genome analysis operations: for example, examining genes that have strong similarity with (are orthologous to) the *fusA* genes mentioned above requires graphical display of orthologous genes across multiple genomes as shown in the lower right side of Figure 2, while removing genes whose similarities are problematic requires examining the graphical representation of the alignment of DNA sequences for the genomes involved in the analysis, as shown in the lower left side of Figure 2.

4. Microbial Genome Data Space

Requirements analysis provides the basis for specifying an abstract data model for microbial genome data. For IMG, data warehouse constructs were employed for specifying its data model in order to allow reasoning about genome data in an established framework that also provides helpful analogies to well understood traditional data applications. Consequently, the microbial genome data space is modelled in terms of primary (also known as *fact*) objects characterized in the context of other (also known as *dimension*) objects. Each dimension is further characterized by one or several *category* attributes which are sometimes organized in a classification hierarchy. Operations in such a framework can be then defined in a multidimensional data space.

4.1 Microbial Genome Data Model

Microbial genome data can be viewed as an abstract multidimensional data space, whereby *genes* form the primary class of objects and are characterized in the context of other classes of objects, in particular individual *genomes*, *functions* and *pathways*.

The definition for each class of objects must include specifications for the semantics of component objects and for the operations that can be applied on them. Defining the semantics of biological data objects is a daunting task and requires a thorough understanding of the process involved in their generation. Unlike traditional (e.g., financial) data, biological data are imprecise, generated via processes that involve transformations between different levels of data granularity and are based on evolving technology platforms and computational methods. Consequently, the semantics of biological objects often cannot be fully characterized without information about their generation (i.e., *provenance*), such as experimental conditions, methods, data transformation parameters.

For example, class *Gene* models objects that may represent hypothetical genes that are predicted using gene prediction methods, such as Glimmer or Critica (e.g., see [10]) or experimentally validated genes. Genomes are annotated using a variety of gene prediction methods that yield results with different precision and reflecting different (e.g., either under or over prediction) biases. Functional characterizations for genes are also generated using diverse methods with different degrees of confidence (e.g. sequence homology based methods, experimental evidence based methods etc.) and often employ different terms for identifying functions. Ontology *terms* have been traditionally used to specify the functions of gene products. For example, the Gene Ontology [9] provides terms to describe the attributes of gene products in three domains of molecular biology: molecular function, biological process, and cellular component. The Gene Ontology is not the only controlled vocabulary used for this purpose, nor is it used consistently for annotating different genomes. Furthermore, the association of a gene with a function may change because of amendments to the functional characterization of genes: for example, see [22] for a discussion of problems associated with gene and function nomenclature and association.

Every class of objects is associated with basic operations, such as comparing data items of the same type (e.g., two DNA sequences associated with genes), as well as complex operations, such as searches for certain patterns across sets of data items. For example, a typical microbial genome analysis operation involves comparing distribution patterns (known as *gene occurrence profiles* or simply *gene profiles*) for genes associated with a specific genome, across other genomes. This operation is described in more detail in the next section.

Relationships between biological classes of objects are usually specified using operations associated with each individual class. For example, associating a gene with a pathway may involve matching (via sequence similarity) the protein sequence of the gene with the protein sequence considered as representative for the enzyme that serves as catalyst for reactions involved in pathways. Alternatively, a gene may be associated with a pathway based on its functional coupling with another gene that is already associated with the pathway. Specific methods employed for associating genes with pathways affect the precision of the functional characterization for genes. Operations that are based on these characterizations, such as grouping genes based on their association with pathways, will be also affected by the choice of these methods.

4.2 Microbial Genome Data Analysis

Microbial genome data analysis is set mainly in the comparative context of multiple microbial genomes [21]. Comparative analysis is essential in the identification of similar or unique genes among different, potentially phylogenetically related, genomes, which provides the foundation for characterizing microbial genomes.

Analysis operations allow navigating the microbial genome data space along one or several dimensions and are often set in the context of specific genomes, pathways, and genes. Setting this context corresponds to reducing the dimensionality of the data for on-line analytical processing (OLAP) operations for traditional (e.g., financial) applications. Genome (organism) selections help focus the analysis on a subset of interest, especially in terms of phylogenetic relationships. For example, a set of interest may include the organisms for all the strains within a specified species. Similarly, pathway selections focus the analysis on a subset of interest, such as pathways involved in lipid metabolism. Gene selections reduce the scope of analysis to genes with certain properties, such as genes sharing a common gene symbol, function, or pathway.

Aggregation operations (usually called *summaries* or *statistics*) involving groups of objects, such as organisms, pathways, and genes, are commonly used over microbial genome data and are similar to analogous OLAP operations for traditional applications. For example, genes can be grouped over one level of the phylogenetic classification hierarchy associated with organisms. Such a grouping is then associated with a count of the group members and is employed to assess the extent of annotations over a selected set of organisms or genes. For example, the number of genes with functional characterization for a given set of organisms provides an assessment of the functional characterization across these organisms. Ranking (sorting) is another OLAP like operation that is employed over genome data summaries.

Additional operations in the microbial genome data space are domain-specific and do not have counterparts in the traditional data domain. Consider part of the genome data space that involves three dimensions: Genes, Genomes, and (sequence similarity) Methods. For a specific genome Ψ and specific method (e.g, sequence similarity comparison of genes, associated with a specific precision), the data space consists of either “p” or “a” for every pair (x, Ψ_i), whereby x is a Ψ gene (the Genes dimension), Ψ_i is a genome (on the Genomes dimension), and “p” / “a” indicates presence or absence of a Ψ_i gene that is similar to x, where the similarity is determined using the selected method. Figure 3 shows an example of the genome data space projected over Ψ and a specific method, with presence/ absence for a set of selected Ψ genes (x_1 to x_4) shown across a set of selected genomes (Ψ_1 to Ψ_8).

A typical operation in this simplified data space involves examining (computationally predicted) genes of a specific genome, Ψ, in the context of other related genomes, $\Psi_1,\ldots,\Psi_k$: this operation allows determining what genes genome Ψ may have in “common” with $\Psi_1,\ldots,\Psi_k$. Such an operation is sometimes called *gene profile* (also called phylogenetic profile [4] and occurrence profile in [21]), and involves first selecting a specific method (i.e., projection on the Method dimension), and then selecting the genomes of interest ($\Psi_1,\ldots,\Psi_k$), for example based on their phylogenetic relationship (on the Genomes dimension). Then Ψ genes with specific “profiles” can be examined.

For the example shown in Figure 3, genome Ψ has gene x_4 in “common” with genomes Ψ_1 to Ψ_8; and genes x_1 and x_2 of Ψ have the same profile across genomes Ψ_1 to Ψ_8. Note that although this operation seems to be similar to (and is often confused with) a set operation (e.g., intersection) over genes, it is not an operation over sets of genes associated with different genomes: from a data point of view each gene is unique although it may be similar (but not identical) to other genes in terms of their associated protein sequences.

Ψ	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6	Ψ_7	Ψ_8
x_1	*p*	*p*	*p*	*p*	*a*	*a*	*p*	*a*
x_2	*p*	*p*	*p*	*p*	*a*	*a*	*p*	*a*
x_3	*a*	*p*	*p*	*a*	*a*	*a*	*a*	*a*
x_4	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*

Figure 3. Example for Examining Gene Profiles.

Gene profile operations are used for analysing biological phenomena of interest, such as gene *conservation* or *gain*, for a specific genome (e.g., Ψ) in the context of other genomes (e.g., $\Psi_1,\ldots,\Psi_k$). For the example shown in Figure 3, gene x_4 of Ψ is conserved across Ψ_1 to Ψ_8, while gene x_3 of Ψ is gained with respect to Ψ_1 and Ψ_4 to Ψ_8.

The gene profile operations are also key components of the microbial genome functional characterization process, which, as mentioned above, is based on a number of assumptions regarding (phylogenetically) related genomes, whereby pairs of genes with similar functions are often both present or both absent (i.e, have similar profiles), tend to be collocated, and/or are components of conserved collocated genes across such genomes. For the example above (see also Figure 1), suppose that gene x_1 of Ψ is functionally characterized while x_2 is not; then the fact that genes x_1 and x_2 have the same profile across genomes Ψ_1 to Ψ_8, may help characterize x_2 which may participate in a similar biological process as gene x1.

5. The Integrated Microbial Genomes System

The abstract microbial genome data model discussed in the previous section helped design the IMG data management system.

5.1 System Architecture

Microbial genome data analysis involves large amounts of data distributed across diverse, usually heterogeneous, data sources. Effective data analysis requires providing a coherent view of the biological phenomena that may be concealed by the fragmentation or ambiguity of genomic data.

Integration of biological data has been considered extensively over the years because of the continuous proliferation of these sources and the need to access multiple sources inherent to biological data exploration. Data integration can be carried out using data warehouse or federated database approaches. Both approaches are based on a common (global) view of the data and involve transformation of the data from individual data sources to a common view. While data warehouses use extract-transform-load (ETL) tools for assembling and then regularly updating data in a centralized system, database federations extract and assemble data dynamically from individual data sources through data adapters [6].

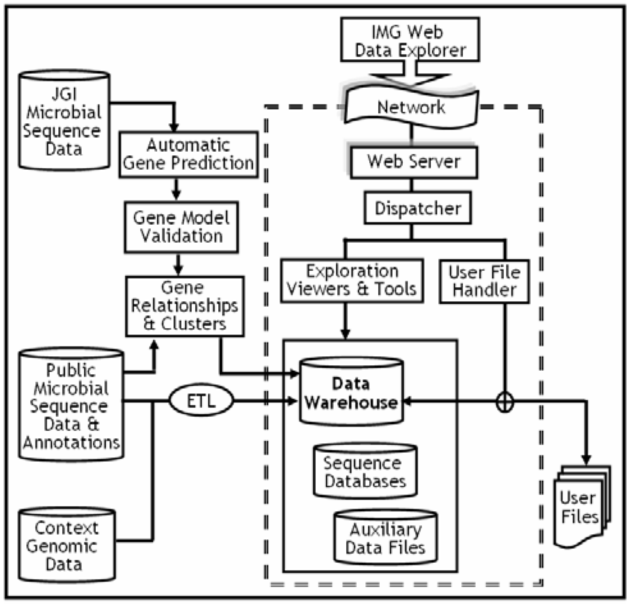


Figure 4. IMG System Architecture.

Data warehouse and analytical processing methodologies have proven to be successful in building academic systems (e.g., see [6], [19]), as well as commercial systems [17]. While the data federation methodology is more appealing and has been the subject of extensive research, the data warehouse approach has proven to be better suited for dealing with inherently imprecise biological data that require substantial manual data curation [6]. Consequently, a data warehouse approach has been adopted for IMG, as reflected in the system architecture shown in Figure 4.

A biological data management system often involves a mix of commercial off-the-shelve (COTS), open source, and custom tools, whereby available tools and methods need to be adapted in order to address application-specific characteristics. For IMG, proven open source tools developed under the Apache Software Foundation provide the components needed for the Web server, a high performance Oracle 9i DBMS is employed for the data warehouse, while custom tools have been developed in order to address problems specific to the microbial genome data application domain. A variety of open source bioinformatics tools have also been employed. Most prominent among these tools is BLAST which is widely used to identify homologous (similar) genes in different organisms [1].

IMG has a multi-tier architecture, as shown in Figure 4. Gene predictions (called gene models) are validated and corrected manually by expert scientists using custom tools. An ETL toolkit is employed for extracting, cleaning, integrating and loading data from public data sources into the IMG warehouse. Gene relationships and clusters are computed using custom tools and are then loaded into the IMG warehouse.

The IMG back-end consists of the IMG 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.

The IMG user interface follows the Web interface paradigm of transparent operations, whereby simplicity is preferred over flexibility. For example, preformatted (canned) queries support various Web forms, while dynamic construction of queries is not supported. A client web browser is used for accessing IMG: the Exploration Viewers and Tools component handles the data exploration operations and provides support for running application-specific bioinformatics tools (e.g., BLAST), while the User File Handler component handles files consisting of genes and organisms of interest to users. These files can be generated using IMG's data export capabilities or can be created locally, and allow users to save the results of their analysis.

Web based technologies have limitations in terms of user interactivity and the amount of data that can be efficiently transferred to web pages, which are reflected in certain restrictions on analysis workflows. Such limitations can be addressed by client applications. For example, a separate Java client application that is an extension of the analytical tool prototype mentioned in section 3.2 allows JGI scientists to analyze and curate (edit) IMG data. This tool has substantially more power than IMG's Web based interface, but requires users to tolerate higher complexity. Developing an equally powerful, but less complex interface remains a challenge that needs to be addressed for future versions of IMG.

5.2 Data Structure

The structure for the IMG data warehouse was defined after analyzing data content requirements and the characteristics of various public data sources considered for IMG. An outline of the IMG data warehouse and its main data sources are shown in figure 5.

*Gene*s are represented in IMG using several classes of related objects: in addition to class Gene that represents curated or predicted hypothetical genes, non-coding RNA genes, other related gene features such as mRNA transcripts and proteins are represented by classes Transcript and Protein respectively, while class Feature represents additional sub-sequence features such as promoters. Gene similarity relationships are represented by classes Ortholog and Paralog.

*Genome*s (also referred to as organisms and species) are represented in IMG with their taxonomic lineage (domain, phylum, order, class, family, genus, species, strain) using class Taxon, while *chromosomes* and *plasmids* are represented by class Chromosome.

Gene functions are represented by several classes in IMG: for example, class GO Ontology represents the vocabulary of terms used to describe gene functions following [9] and class Enzyme further characterizes the gene product in terms of molecular function.

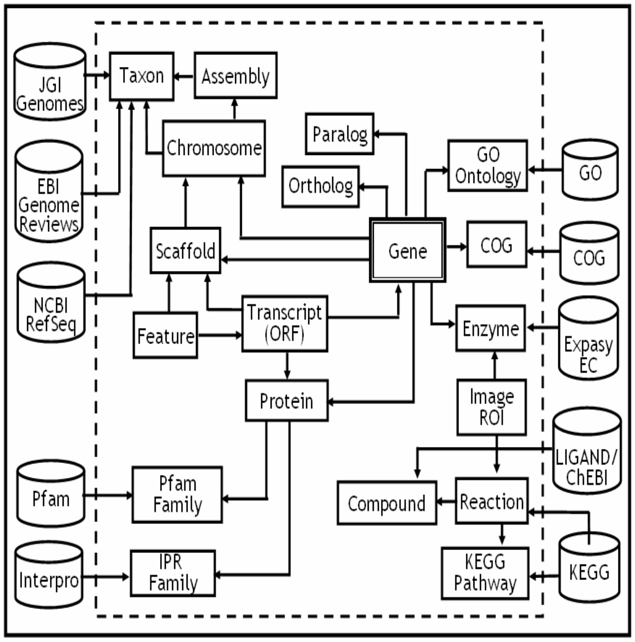


Figure 5. IMG Data Model Outline and Data Sources.

Finally, *pathways* are represented in IMG by several classes, including KEGG Pathways, Reaction and Compound, which represent pathways, reactions and compounds, respectively, with class Image ROI modelling the association of reactions with enzymes.

It is important to note that a data model such as that underlying IMG needs to be extensible, whereby data structure changes reflect the evolution of the biological application domain or respond to new system requirements. Keeping such changes documented as well as transparent to users, poses an additional challenge that needs to be addressed in the development of biological data management systems. In IMG, this problem has been addressed by using the OPM toolkit that has been used in developing several biological databases, such as GDB [7]. OPM tools allow describing database structures in terms of classes of objects and provide support for dealing effectively with rapid schema evolution [18].

5.3 Data Content

The process of data extraction, transformation, cleaning, integration and loading into the IMG warehouse is similar to analogous data warehousing processes: custom tools are used to extract and parse information from external sources and then transform the data into a structure consistent with the object structure in the IMG data warehouse; transformed data are eventually loaded into the Oracle database. In order to cope with the imprecision of the genomic data, expert manual data revision ("quality assurance") and correction check points have been included in the process. Lack of standard representation of data in different data sources and the idiosyncratic nature of genome annotations hamper the automation of the data revision and correction process.

For example, each time microbial genome data are extracted from a public source such as EBI's Genome Reviews, the taxonomic information characterizing these genomes, such as domain, phylum, order, class, and, family, needs to be reviewed based on different taxonomy sources, so that differences can be reconciled. An important stage is verifying the correctness of genome annotations, especially the automatically generated gene predictions and associated functional annotations. While the type and rate of errors for genome annotations are usually known, correcting them requires time consuming expert manual reviews, which are seldom performed. For example, while overlapping sequences of predicted genes are known to indicate errors and are easy to detect, full characterization of the errors and correcting them requires visual inspection and manual editing. For IMG, an annotation quality control and correction stage is part of the standard procedure for including data into IMG and is documented as part of the system documentation [11].

5.4 Data Analysis

The IMG Web Data Explorer allows exploring the microbial genome data space along the genomes (organism) and genes dimensions, via analysis workflows that are based on some of the operations discussed in section 4.2. A Methods dimension is built into the Phylogenetic Profiler (see below), while the Pathway and Function dimensions are not directly supported in the current version (IMG 1.1), but can be explored through a Gene Information interface described below.

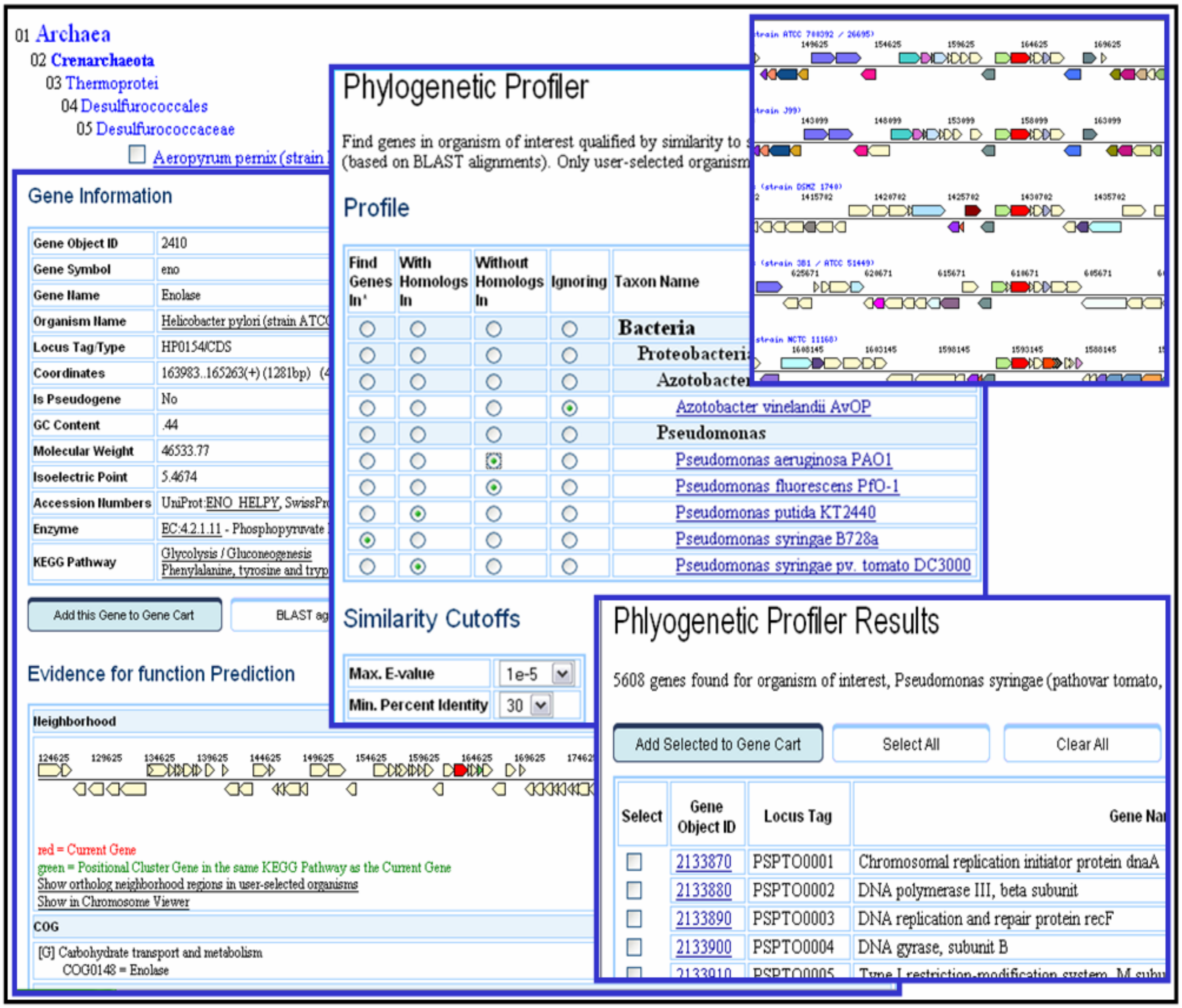
IMG data analysis is typically set in the context of specific genomes (organisms). The simplest form of defining this context is through key word search. Alternatively, organism selection can be carried out by browsing the list of all organisms whose genomes are available for analysis, ordered alphabetically or organized as a phylogenetic tree, as shown in the top left side of Figure 6. Organism selections can be refined by examining the organisms of interest individually or in a comparative context using summaries (statistics) for assessing the extent of the functional characterization for their genes.

Gene selections help focus the analysis on genes with certain properties, such as genes sharing a common gene symbol, function, or pathway. Gene selection can be carried out through keyword or sequence similarity searches, or using the Phylogenetic Profiler.

The Phylogenetic Profiler allows selecting genes of a genome (organism) of interest that are present (i.e, have similar genes) in other, usually phylogenetically related, organisms, or are absent (i.e., have no similar genes) in other, potentially related, organisms. Using *Phylogentic Profiler* is illustrated by the middle upper side screen of Figure 6. First, phylogenetically related organisms are selected using the phylogenetic hierarchy in the Taxon pane. Genes are selected for the organism identified as the "organism of interest" (OOI), based on their presence or absence in the other selected organisms. For example, the profile specified in Figure 6 is for *Pseudomonas syringae B* (OOI) genes that are present ("with homologs in") in *Pseudomonas putida* and *Pseudomonas syringae pv tomato,* and are absent ("without homologs in") in *Pseudomonas fluorescens* and *Pseudomonas aeruginosa*. Note that gene absence or presence is based on sequence similarity, whereby parameters (max e-value and min percent identity) for similarity comparison represent the methods dimension of the genome data space, and can be fine tuned as needed (see Similarity Cutoffs in Figure 6). The OOI genes with the specified profile are then retrieved, as shown in the right lower side screen of Figure 6.

Individual genes can be further analyzed by examining additional details regarding gene annotations, as shown in the Gene Information screen in the left side of Figure 6. These details include evidence for the functional characterization (prediction), including the "gene neighbourhood" shown in the lower left side of Figure 6. A gene neighbourhood displays the gene of interest in its location on the chromosome, together with other genes collocated on the same area of the chromosome. Gene neighbourhood is an example of the importance of graphical representation for genomic data: visual exploration of a gene in the context of other genes helps determining the accuracy of its functional annotation and its participation in positional clusters of genes that may represent operons.

One can also examine a gene of interest in the context of related, such as orthologous, genes in other related genomes, across multiple neighbourhoods as shown in the right upper side of Figure 6. This type of visualization



Gene Information

Gene Object ID	2410
Gene Symbol	eno
Gene Name	Enolase
Organism Name	Helicobacter pylori (strain ATCC
Locus Tag/Type	HP0154/CDS
Coordinates	163983..165263(+) (1281bp) (4
Is Pseudogene	No
GC Content	.44
Molecular Weight	46533.77
Isoelectric Point	5.4674
Accession Numbers	UniProt:ENO_HELPY, SwissPr
Enzyme	EC:4.2.1.11 - Phosphopyruvate
KEGG Pathway	Glycolysis / Gluconeogenesis Phenylalanine, tyrosine and tryp

Add this Gene to Gene Cart   BLAST ag

Evidence for function Prediction

Neighborhood

red = Current Gene
green = Positional Cluster Gene in the same KEGG Pathway as the Current Gene
Show ortholog neighborhood regions in user-selected organisms
Show in Chromosome Viewer

COG

[G] Carbohydrate transport and metabolism
COG0148 = Enolase

Phylogenetic Profiler

Find genes in organism of interest qualified by similarity to
(based on BLAST alignments). Only user-selected organism

Profile

Find Genes In*	With Homologs In	Without Homologs In	Ignoring	Taxon Name
○	○	○	○	Bacteria
○	○	○	○	Proteobacteri
○	○	○	○	Azotobacter
○	○	○	●	Azotobacter vinelandii AvOP
○	○	○	○	Pseudomonas
○	○	●	○	Pseudomonas aeruginosa PAO1
○	○	●	○	Pseudomonas fluorescens PfO-1
○	●	○	○	Pseudomonas putida KT2440
●	○	○	○	Pseudomonas syringae B728a
○	●	○	○	Pseudomonas syringae pv. tomato DC3000

Similarity Cutoffs

Max. E-value	1e-5
Min. Percent Identity	30

Phlyogenetic Profiler Results

5608 genes found for organism of interest, Pseudomonas syringae (pathovar tomato,

Add Selected to Gene Cart   Select All   Clear All

Select	Gene Object ID	Locus Tag	Gene Na
□	2133870	PSPTO0001	Chromosomal replication initiator protein dnaA
□	2133880	PSPTO0002	DNA polymerase III, beta subunit
□	2133890	PSPTO0003	DNA replication and repair protein recF
□	2133900	PSPTO0004	DNA gyrase, subunit B
□	2133910	PSPTO0005	Type I restriction-modification system, M subu

Figure 6. IMG Web Data Explorer: Data Analysis Example.

allows examining concordance with the biological evolutionary phenomena mentioned in section 2.2. For example, one can determine whether pairs of genes with related (color coded) functions are collocated on chromosomes in multiple genomes.

Another comparative analysis method for genes is based on the phylogenetic occurrence profile operations discussed in section 4.2. Given a gene of interest for a selected genome (organism), the phylogenetic occurrence profile shows the presence/absence pattern of this gene across selected (by default, all) organisms. Occurrence profiles are usually examined for multiple genes associated with the same organism, since if such genes have similar profiles then they may also have a similar evolutionary history and may potentially be functionally linked, or co-regulated in a pathway, as mentioned in section 2.2 [4]. In order to find such genes in IMG, a phylogenetic occurrence profile similarity search allows finding other genes in the same organism that have occurrence profiles that are similar to that for the gene of interest (see the example shown in Figure 3, where genes x_1 and x_2 of Ψ have the same profile across genomes Ψ_1 to Ψ_8). The occurrence profiles of multiple genes across these organisms can be then visually compared using a phylogenetic occurrence profile viewer.

6. Conclusion

Effective data analysis across biological data management systems involves providing support for seamless composition of analysis operations, which in turn requires a systematic process for analyzing the data structure and operations of the application domain. We discussed in this paper how this challenge has been addressed in the development of the Integrated Microbial Genomes (IMG) system.

The development process for IMG involved detailed application domain analysis based on use case scenarios subsequently used for defining an abstract microbial genome data model in terms of data types and operations. Data structure definitions and analysis operations are partly described in the online IMG documentation which also covers data content, data processing, and system related aspects (see AboutIMG [11]). As IMG evolves, the current level of documentation will be expanded and improved in terms of clarity and completeness.

A data warehouse framework was used in developing IMG, and was found to provide an effective environment for developing a system that needs to support the integration and management of data from diverse sources, where data are inherently imprecise and tend to evolve over time. The data warehouse approach has allowed dealing with data quality problems in a systematic manner, whereby data validation and cleansing steps have been included in the overall data acquisition and integration process. The data warehouse environment has also provided an established framework for modelling and reasoning about genomic data.

The first version of IMG was released in March 2005, and contained data for 296 genomes, including over 100 JGI sequenced microbes. A second version of IMG was released in June 2005, containing additional 32 public genomes and 9 JGI genomes, bringing the total to 337 genomes. The first two versions of IMG have focussed on data quality in terms of the coherence of annotations, based on sound validation and correction procedures, as well as corroboration of annotations from other public microbial genome data resources. There comparative analysis context provided by IMG facilitates the detection and correction of annotation errors.

IMG continues to be extended in terms of data content through quarterly updates, whereby it aims at continuously increasing the number of genomes integrated in the system from public and local resources, following the principle that the value of genome analysis increases with the number of genomes available as a context for comparative analysis.

IMG aims at increasing the coverage (breadth and depth) of functional annotations in the system, the result of providing scientists with tools that implement annotation techniques based on the functional coupling of genes, a hypothesis inspired by observed biological evolutionary phenomena. IMG will also be enhanced in terms of its data analysis capabilities. For example, operations for exploring data along the functional and pathway dimensions are developed for the next version of IMG scheduled to be released in September 2005.

There are several public systems, such as Genome Reviews [14], CMR [23], and MBGD [28], as well as a commercial system, ERGO [20], that share IMG's goal of providing microbial genome data in an integrated context. These systems are distinguished by the extent of their coverage in terms of genomes and annotations. From a content point of view, it is difficult to properly characterize these systems because accepted metrics and benchmarks for qualifying the accuracy of biological data are controversial. Nevertheless, these systems provide valuable sources of public genome data that are used by IMG to complement its own collection of JGI complete and draft genomes, as well as for corroboration of its annotation and content validation procedures.

As the main drivers for seamless data analysis, the biologists involved in the development of IMG have recognized the need for a detailed domain analysis process and have been actively involved both in this process and in specifying the documentation associated with it. The development of IMG also confirms the need for a tight collaboration between biologists and data management experts, a key recommendation in [12]. Lack of such collaborations often leads to poor use of data management technologies or misunderstood requirements which can result in "sterile pursuits of insignificant or misunderstood problems" [12].

A systematic development process, starting with requirements analysis, provides the framework needed for specifying analysis workflows, including documentation for data structure and operations. Following such a process is time consuming and requires resources that may not be available to academic groups. When such a process is followed the results are enduring as illustrated by GDB's data management infrastructure developed almost a decade ago [7].

Acknowledgements

Members of the Genome Data Systems group at JGI developed data content validation tools and provided software and content QA support for IMG. Phil Hugenholtz oversees the IMG taxonomy and has provided valuable suggestions for developing genome data viewers. IMG uses tools and data from a number of publicly available resources. Their availability and value is gratefully acknowledged. The work presented in this paper was supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

References

[1] S.F. Altschul, W. Gish,, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology,* . **215**. 403-410. 1990.

[2] A. Bernal, U. Ear, and N. Kyrpides, Genomes Online Database (GOLD): A Monitor pf genome projects world-wide. *Nucleic Acid Research* **29**, 126-127, 2001. See also: http://www.genomesonline.org/.

[3] T. Bloom and T. Sharpe, Managing Data from High-Throughput Genomic Processing: A Case Study, *Proc. of the 30th VLDB Conference*, 2004.

[4] P.M. Bowers, M. Pellegrini, M.J. Thompson, J. Fierro, T.O. Yeates, and D. Eisenberg, Prolinks: A

Database of Protein Functional Linkages Derived from Coevolution, , *Genome Biology* **5**, 2004.

[5] M.A. Branka, T.V. Venkatesh, and N. Goodman, Bioinformatics: Getting Results in the Era of High-Throughput Genomics, *Cambridge Healthtech Institute Report,* http://www.healthtech.com/, 2001.

[6] S.B. Davidson, J. Crabtree, B. Bunk, J. Schug, V. Tannen, and C. Stoeckert, K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources, *IBM Systems Journal*, **40**, 512-531, 2001.

[7] GDB Human Genome Database. http://www.gdb.org/. See data definitions at: http://gdbwww.gdb.org/gdb/schema.html.

[8] M.Y. Galperin, The Molecular Biology Collection: 2005 Update, *Nucleic Acids Research*, **33**, Database Issue, D5-dD24, 2005.

[9] Gene Ontology Consortium, The Gene Ontology Database and Informatics Resource, *Nucleic Acids Research*, **32**, 258-261, 2004. See also http://www.geneontology.org/.

[10] L. Hauser, F. Larimer, M. Land, M. Shah, and E. Uberbacher, Analysis and Annotation of Microbial Genome Sequences, *Genetic Engineering*, Vol. 26, Kluwer Academic/Plenum Publishers, 225-238, 2004.

[11] Integrated Microbial Genomes (IMG) System, http://img.jgi.doe.gov/.

[12] H.V. Jagadish and F. Olken, Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular an Cell Biology, *OMICS*, Vol. 7, No. 1, 2003.

[13] JGI Microbial Genome Portals. http://genome.jgi-psf.org/microbial/index.html

[14] P. Kersey, et al. Integr8 and Genome Reviews: Integrated Views of Complete Genomes and Proteoms, *Nucleic Acid Research* **33**, D297-D302, 2005. See http://www.ebi.ac.uk/GenomeReviews/

[15] L. Krishnamurty, J. Nadeau, G. Ozsoyoglou, M. Ozsoyoglou, G. Schaeffer, M. Tasan, and W. Xu, Pathways Database System: An Integrated System for Biological Pathways, *Bioinofrmatics*, **19**, 930-937, 2003.

[16] S. E. Lewis et al, Apollo: A Sequence Annotation Editor. *Genome Biology*, **3** (12), 2002. http://genomebiology.com/2002/3/12/.

[17] V.M. Markowitz, J. Campbell, I.A. Chen, A. Kosky, K. Palaniappan, and T. Topaloglou, Integration Challenges in Gene Expression Data Management. *Bioinformatics: Managing Scientific Data*, Morgan Kauffman Publishers (Elsevier Science), 277-301, 2003.

[18] V.M. Markowitz, I.A. Chen, A.S. Kosky, and E. Szeto, Object-Protocol Model Data Management Tools '97. Bioinformatics, Databases and Systems, Stan Letovsky (ed), Kluwer Academic Publishers, pp. 187-199, 1999.

[19] R. Nagarajan, M. Ahmed, and A. Phatak, Database Challenges in the Integration of Biomedical Data Sets, *Proc. of the 30th VLDB Conference*, 2004.

[20] R. Overbeek, R., et al. The ERGO Genome Analysis and Discovery System. *Nucleic Acid Research* **31**, 164-171, 2003. See also http://www.ergo-light.com/ERGO/

[21] A. Osterman and R. Overbeek, Missing Genes in Metabolic Pathways: A Comparative Genomic Approach, *Chemical Biology*, **7**, 238-251, 2003.

[22] H. Pearson, Biology's name game, *Nature*, 417, 631-632, 2001

[23] J.D. Peterson, L.A. Umayam, T. Dickinson, E.K. Hickey, and O. White, The Comprehensive Microbial Resource, *Nucleic Acid Research* **29**, 123-125, 2001. See http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl

[24] K.D. Pruitt, T. Tatusova, and D.R. Maglott, NCBI Reference Sequence (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts, and Proteins, *Nucleic Acid Research* **33**, D501-D504, 2005. See http://www.ncbi.nlm.nih.gov/RefSeq/.

[25] NCBI Taxonomy, http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/. See also Bergey's Manual of Systematic Bacteriology at http://www.cme.msu.edu/bergeys/.

[26] R.J. Roberts, P. Karp, S. Kasif, S. Linn, and M.R. Buckley, An Experimental Approach to Genome Annotation, *American Academy for Microbiology*, 2005, http://www.asm.org/Academy/index.asp?bid=32664.

[27] T. Topaloglou, Panel on Biological Data Management: Research, Practice, and Opportunities, *Proc. of the 30th VLDB Conference*, 2004.

[28] I. Uchiyama, I. MBGD: Microbial Genome Database for Comparative Analysis, *Nucleic Acid Research* **31**, 58-62, 2003.