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

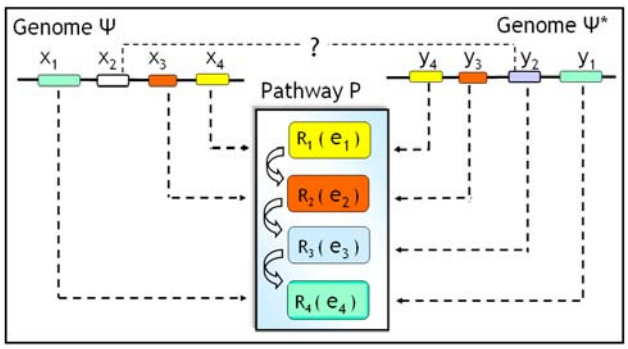

Figure 1. Sketch of Genomes Associated with a Pathway.

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

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