

Biological Data Management in a Dataspace Framework

Victor M. Markowitz

Biological Data Management and Technology Center

Lawrence Berkeley National Laboratory

Biological data reside in specialized databases that represent different data interpretation stages or different facets of biological phenomena. For example, microbial genomes are sequenced by organizations worldwide, follow an annotation process (gene prediction, functional characterization) that is often specific to each sequencing center, and end up in one of the *primary* archival public sequence data repositories, such as GenBank. Genome sequence data include information on gene coordinates, locus identifiers, gene names and protein functions. Analyzing microbial genomes requires however additional functional annotations, such as motifs, domains, and pathways, which are provided by diverse, usually heterogeneous, *auxiliary* annotation sources, such as Pfam¹, InterPro², COG³, and KEGG⁴. *Secondary* public resources such as EBI's Genome Reviews⁵ and NCBI's RefSeq⁶ integrate such additional functional annotations with the sequences from the primary sequence data sources, sometimes together with a review and curation of the associated annotations. Such secondary resources 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. *Tertiary* resources such as the Integrated Microbial Genomes (IMG) system [4] aim 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. Such a data context is critical for multi genome *comparative* analysis used in the functional characterization of microbial genomes,

The increasing number of biological databases⁷, the emergence of new types of data that need to be captured, as well as evolving technologies, methods and biological knowledge add to the complexity of data management required to support biological data analysis. A typical biological data management system involves accessing or gathering data from multiple sources, followed by data correlation, classification, review, and curation using domain specific tools (e.g., functional clusters, ontologies) and expertise. In practice, biological data management is less daunting when it is considered in the context of an iterative strategy based on gradual data integration while accumulating domain specific knowledge throughout the integration process. The recently proposed *dataspace* abstraction [1] provides the framework for such a strategy which has proved to be effective in devising systems such as IMG. For example, IMG's *dataspace* includes several primary and secondary microbial genome data sources

¹ Bateman, A. et al. 2004. The Pfam Protein Families Database. *Nucleic Acids Research* **32**.

² Mulder, N.J. et al. 2005. InterPro, Progress and Status in 2005. *Nucleic Acids Research* **33**.

³ Tatusov, R.L. et al. 1997. A Genomic Perspective on Protein Families, *Science*, **278**.

⁴ Kanehisa, M. et al. 2004. The KEGG Resource for Deciphering the Genome. *Nucleic Acids Research* **32**.

⁵ Kersey, P. et al. Integr8 and Genome Reviews: Integrated Views of Complete Genomes and Proteoms. 2005. *Nucleic Acid Research* **33**.

⁶ Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): A Curated Non-redundant Sequence Database of Genomes, Transcripts, and Proteins, *Nucleic Acid Research* **33**.

⁷ Over 850 databases (more than 130 compared to 2005) are listed in: Galperin, M.Y. 2006. The Molecular Biology Collection: 2006 Update, *Nucleic Acids Research* **34**.

together with a number of system components⁸ implementing integration that gradually increases the coherence and completeness of microbial genome functional annotations. This strategy allows analyzing genomes with “coarse granularity” annotations available in one component while these annotations are refined via further, often time consuming, computations⁹ and expert review and curation in another component. This strategy was also followed in the development of IMG/M [5], which handles aggregate genomes (called *metagenomes*) of microbial communities in the context of other metagenome and isolate microbial genomes, by extending IMG’s dataspace with metagenome specific data sources and additional components for metagenome data management, in particular metagenome annotation. It is worth noting that the analytical tools provided by IMG and IMG/M involve compositions of individual browsing, search, query, and application-specific operations and thus are consistent with the dataspace “search and query” capabilities outlined in [1].

A fundamental part of a dataspace is a *catalog* containing information about various system components and their relationships, with associated mechanisms providing support for gradually extending this catalog as the dataspace expands. As noted in [3], biological data sources are poorly specified, with semantics of data structures, individual data items and operations often represented implicitly and informally via shared assumptions, rather than specified explicitly and formally. Accordingly, setting the foundation of a biological data management system in a dataspace framework requires a substantial specification (i.e., documentation) effort¹⁰. Nevertheless, the dataspace abstraction helps reasoning about biological data management, with different systems seen as related components in an evolving framework rather than isolated data sources, and where various (rather than a single) integration techniques¹¹ address component specific needs. A dataspace framework would also facilitate *leveraging past experience* as suggested in [2], via mechanisms supporting the correlation and reuse of various data management system elements, such as schemas, ontologies, and analytical tools. Consequently, applying the dataspace abstraction to the biological domain could immediately benefit the development of new biological data management systems as well as improve the organization of existing systems and resources.

References

1. Franklin, M., Halevy, A., and Maier, D. 2005. From Databases to Dataspaces: A New Abstraction for Information Management, *SIGMOD Record*.
2. Halevy, A. 2005. Why Your Data Won’t Mix: Semantic Heterogeneity, *ACM Queue*
3. Jagadish, H.V., and Olken, F., 2003. Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular and Cell Biology. *OMICS*, **7** (1).
4. Markowitz, V.M. et al. 2005. The Integrated Microbial Genomes (IMG) System, A Case Study in Biological Data Management, *Proc. of 31st Int. Conf. on Very Large Data Bases*. See also: <http://img.jgi.doe.gov>.
5. Markowitz, V.M. et al. 2006. An Experimental Metagenome Data Management and Analysis System, *Bioinformatics*, **22**(14), 359-367. See also: <http://img.jgi.doe.gov/m>.

⁸ Each component can be seen as a standalone system in the dataspace spectrum of [1].

⁹ For example, so called “all vs. all” gene similarity relationships and gene clustering.

¹⁰ Systematic documentation of biological data sources is difficult because defining the semantics of biological data and operations requires marshaling domain-specific scientific expertise together with data management specific technical expertise.

¹¹ An overview of biological data integration approaches is provided in: Hernandez, T. and Kambhampati, S. 2004. Integration of Biological Sources: Current Systems and Challenges Ahead, *SIGMOD Record*, **33** (3).