

BIOZON: a system for unification, management and analysis of heterogeneous biological data

Aaron Birkland¹ and Golan Yona²

Keywords: integration, heterogeneous sources, unified schema, complex searches

Introduction.

Biological entities are strongly related and mutually dependent on each other. Therefore, there is a growing need to corroborate and integrate data from different resources and aspects of biological systems in order to analyze them effectively.

To identify entities, existing databases use explicit references by accession number or a mutual ontology. Some databases relate and cross link elements from other databases based on these identifiers. However, this information is very partial and is not readily available in some. Moreover, these links are not established in coordination with the other linked databases. With the source databases changing rapidly, this leads to problems of consistency and updatability. Furthermore, it is hard to query this wealth of data in ways that can benefit and exploit the mutual dependency between entities.

Biozon is a unified biological database that integrates heterogeneous data types and the relationships between them, such as nucleic acid sequences, proteins, structures, protein domains and protein families, protein-protein interactions and cellular pathways, into a single extensive schema. This schema allows one to see each data instance in its full biological context. More importantly it allows for complex searches that span multiple data types from a heterogeneous set of sources and for arbitrary computations on that data. Biozon can also rank results, the same way Google ranks web documents, and uses similarity relationships to extend query results to similar biological entities.

Our Approach

The data in Biozon is composed of two main types: **source data** that are gleaned from established online databases (such as SWISS-PROT, Genbank, BIND, KEGG and others), and unique **derived (computed) data** that includes similarity relationships between objects and predicted information about their functional role. The derived data introduces another level of complexity to our model, but also allows for even more powerful methods of data querying, management and manipulation to the extent of biological theorem verification and computation.

The information in Biozon is logically represented as graph in which nodes represent some unit of data, and edges indicate a relationship between two nodes. Each graph node or edge is given a classification as part of a hierarchy of data types. A constituent data set (for example, from a source database), therefore, maps onto some subset of this graph. Graph nodes that are instances of fundamental biological objects, such as protein sequences, are required to be non-redundant.

The schemata for different data sets can and do share nodes that represent the same fundamental biological type of object. As a result, our graph ends up becoming highly

¹Cornell University, Ithaca, NY, United States. E-mail: birkland@cs.cornell.edu

²Cornell University, Ithaca, NY, United States. E-mail: golan@cs.cornell.edu

connected and centered around hubs of such objects. This connectivity allows for efficient formulation and execution of complex queries that span multiple data types.

Updates on such a tightly integrated graph require that consistency of the data be defined and upheld. We define consistency in terms of the nature of the objects in database, how they reflect their published counterparts and relate to the derived (computed) data, and how they are depended on by others in the database. We designed a framework that executes actions and applies rules to enforce consistency that is based on the action of a collection of small, simple subunits called authorities, each one responsible for decision making and task completion.

Results

The Biozon database currently stores extensive information about more than 37,000,000 protein and DNA sequences, integrating sequence, structure, protein-protein interactions, pathways and expression data, totaling over 60 million documents from more than 20 different databases. It also stores information about 2.5 billion relations between documents, including explicit relations between objects, and derived or computed relations based on sequence similarities, structural similarities and more.

The Biozon database is accessible now at biozon.org, and serves as a useful proof of concept that the ideas expressed in our approach are practical. Indeed, the following functionality is provided as a direct result of our efforts:

- The data graph may be browsed from any entry as a starting point. The effects of our non-redundant object model are made apparent as each entry has all annotations and links to other related objects listed.
- Complex queries that span multiple data types may be created through our query-building interface and executed in real-time. These queries can specify desired interrelationships between types (e.g. searches like '3D structures of proteins that are involved in phosphorylation interactions and are part of the Prostaglandin and leukotriene metabolism pathway').
- First-of-a-kind fuzzy searches that extend complex queries to include homologous sequences or structures as a search step. Queries may be extended by incorporating materialized similarity data in any appropriate query step. For example, querying for structures of proteins in enzyme family 1.1.1.1 and involved in an interaction returns no results. Incorporating similarity into the search transforms the query into one that searches for structures of proteins that are involved in interactions and *similar* to proteins that are members of the 1.1.1.1 family. This query does return significant results from a very large search space in less than a minute.
- First-of-a-kind biological ranking system which resembles the methods implemented in Google. Results may be ranked by an algorithm similar to PageRank that takes into account the link structure of the data graph.
- Various prediction and analysis tools, such as domain structure prediction are available for general use

All the above features may be used in real time by visiting biozon.org.