Columba - A Database of Annotations of Protein Structure

Kristian Rother [(1)], Silke Trissl, Heiko Müller, Patrick May, Rene Heek,
Robert Preissner, Thomas Steinke, Ina Koch, Ulf Leser, Cornelius Frömmel

Berlin Center for Genome Based Bioinformatics (BCB)
http://www.bcbio.de
[(1)] Corresponding author: kristian.rother@charite.de

Keywords: Protein structures, Data integration, Annotation, Databases

Introduction

Researchers on protein structures are often interested in sets of proteins sharing certain properties, such as subfolds, protein function, organism source, or pathways. Being able to generate such sets quickly has two major applications. First, researchers may select such a set to perform some special analysis only on the structures within this set, trying to find set-specific properties. Second, researchers may generate groups of proteins by doing some analysis, and are then interested in correlating those groups with common properties of the protein structures contained.

We have created the Columba database of annotations for protein structures to support both types of research. Columba physically combines resources on protein structures into a single relational database ready for queries. Columba can answer questions such as

- Which protein chains have a TIM-barrel fold?
- Which proteins in the citric acid cycle have a resolved structure available?
- Which protein chains have a helix-turn-helix motif matching a given regular expression?

with just a single query. In contrast, databases such as PDB or MSD provide only pointers to related resources, which requires a user to manually click through large numbers of web pages to achieve a similar result. Columba is intended to shorten the distance structural biologists have to take when creating sets of structures for their purposes.

Database content

Currently, the database contains records for all structures from the PDB, annotation on folding classification from SCOP and CATH, metabolic pathways from KEGG and the Boehringer maps, DSSP secondary structures and references to scientific publications parsed from PDB. We are working on extending the database with functional classification of proteins using the Gene Ontology and keywords and medical terms taken from related Pubmed entries.

Applications

Using Columba we already obtained results like:

- We found that for the metabolic pathways in KEGG, between 0% and 60% of the participating enzymes are structurally resolved. For many pathways, no enzyme structures are known yet. However, the pathway diagrams in KEGG often cover several branches of a metabolite group, so the percentages of many sub-pathways can be expected to be higher.
- We have connected the PTGL protein topology graph library (2) describing 3D-relationships between secondary structures to the Columba database. Using SQL queries, we can search and compare topologies and sub-topologies. This opens the opportunity to search for correlations between topological properties and functional characterizations of proteins.
- In the same manner, we integrated 3D-virtual screening data comparing patches of protein surfaces (DIP) (3). Querying the database, we can now cross-correlate DIP patches with folding and pathway data. As a preliminary result, we found that the RMSD between patches correlates well with different levels of SCOP classification, allowing to predict the protein family from local similarity data.

Implementation

We use the OpenMMS parser and database schema (1) to read PDB entries in the mmCif format into a relational database. The OpenMMS schema is mapped to a simpler representation containing four instead of 120 tables by using schema transformation rules written in SQL. Additional tables in our data model are used to annotate the structure entries with data from the sources mentioned above. The resulting data model contains 26 tables. Columba is implemented on PostGreSQL 7.3 running on a server at the Konrad Zuse Institute Berlin (ZIB). We have created a Python application for constant data updates, both in case of new structures as well as new versions of annotation sources.

Availability

Columba will be freely accessibly through a web interface. The interface follows a "query refinement" paradigm, where the intended set of structures is interactively obtained by defining conditions on structural properties, thus limiting the space of matching structures iteratively. Each query results in a list of PDB entries with summarized information on an entry. This way, queries for sets of protein structures with certain properties are easily performed. We are planning to provide dumps of the Columba database that can be loaded into a PostGreSQL server for in-house data integration and for executing queries which exceed the scope of the web interface.

Ackknowledgements

This research was supported by the German Ministry of Education and Research (BMBF), grant no. 0312705B.

References

[1] D. S. Greer, J. D. Westbrook, and P. E. Bourne. An ontology driven architecture for derived representations of macromolecular structure. *Bioinformatics*, 18(9):1280–1281, 2002.

[2] P. May and I. Koch. Ptgl - protein topology graph library. *ECCB Poster*, 2003.

[3] R. Preissner, A. Goede, and C. Froemmel. Dictionary of interfaces in proteins (dip). *J Mol Biol*, 280(3):535–550, 1998.