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

SPONSORED BY THE Federal Ministry of Education and Research

Columba Protein Structure Annotation

www.columba-db.de

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 is beneficial for the researcher. On structures within this set special analysis can be performed , trying to find set-specific properties.

We have created the Columba database of annotations for protein structures to create such sets of proteins. 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.

Database Schema

The database contains records for all PDB structures, folding classification from SCOP and CATH, metabolic pathways from KEGG and the Boehringer maps, DSSP secondary structures and references to scientific publications.

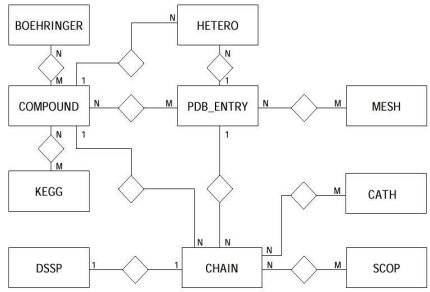

Figure 1: Basic entity-relationship model of the Columba database. PDB entries are stored in the PDB_ENTRY, CHAIN, COMPOUND and HETERO tables. Other sources are added thereafter.

We use the OpenMMS parser and database schema to read PDB entries. It 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).

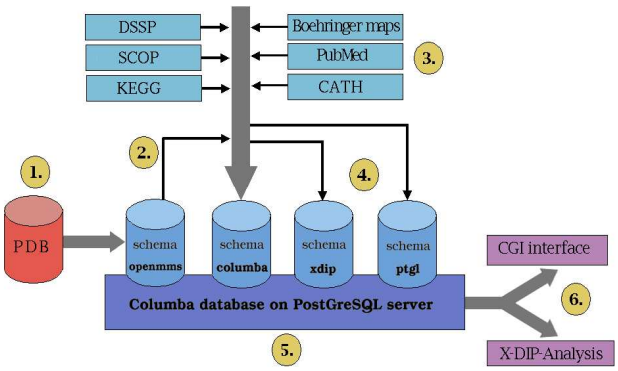

Figure 2: Database schema describing data fluxes and content relations: 1.) Load PDB data to OpenMMS database. 2.) Map OpenMMS schema to a simpler model. 3.) Add annotation from multiple sources of data. 4.) Add data to application-specific schemes. 5.) Have everything in one database. 6.) Query the data.

Web interface

Columba will be freely accessible through a web interface at www.columba-db.de. 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.

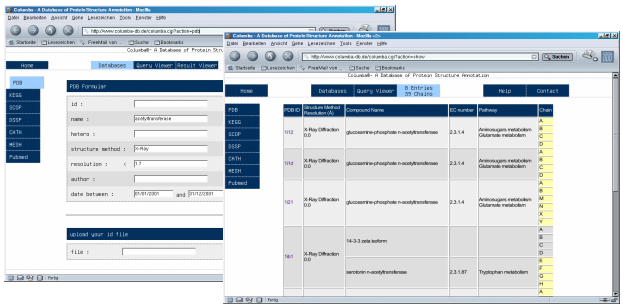

Figure 3: Web-interface for the Columba database with the query-form for information in PDB and a summary list of query results

Applications

We found that for metabolic pathways in KEGG, up to 60% of the participating enzymes are structurally resolved. For many pathways, no enzyme structures are known yet. The pathway diagrams in KEGG often cover several branches of a metabolite group, so the percentages of many sub-pathways can be expected to have a higher coverage.

pathway	enzymes	structures	enzymes with 1+ structures
Fatty acid biosynthesis	7	78	57 %
Val/Leu/Ile biosynthesis	15	57	47 %
Carbon fixation	23	504	43 %
Citrate cycle (TCA cycle)	23	181	39 %
Alkaloid biosynthesis I	34	325	12 %
Peptidoglycan biosynthesis	16	7	6 %
Biotin metabolism	8	0	0 %

Table 1: Coverage of metabolic pathways with structures in PDB

We have connected the PTGL protein topology graph library 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). 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.

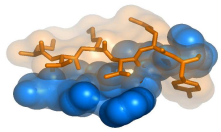

Figure 4: Molecular interface betweeen two beta-sheets from Subtilisin Carlsberg (1cse).

References

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

[2]May P. and Koch I. PTGL – protein topology graph library. ECCB Poster (2003)

[3]Preissner R., Goede A., Froemmel C. Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches J Mol Biol. 280(3):535-50. (1998)

Charité Medical School and University Hospital, Humboldt-Universität zu Berlin

Humboldt-Universität zu Berlin

Konrad-Zuse-Zentrum für Informationstechnik Berlin

University of Applied Sciences, Berlin

This work was done within the context of the NGFN Nationales Genomforschungsnetz

and is supported by Projektträger Jülich