Columba - A Database of Annotations of Protein Structures

[1]Silke Trissl, [2]Kristian Rother, [1]Heiko Müller, [1]Ulf Leser

[1]Institute for Informatics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin
[2]Institute for Biochemistry, Charite-Campus Mitte, Monbijoustr. 2, 10098 Berlin

Corresponding author: trissl@informatik.hu-berlin.de

SPONSORED BY THE Federal Ministry of Education and Research

Columba Protein Structure Annotation

www.columba-db.de

Introduction

The number of protein structures deposited in the Protein Data Bank, PDB (Berman et al. 2000) is increasing rapidly. This growth allows researchers in life science to study complex relationships between macromolecular structures and their properties, such as biological function, folding classification, secondary structure, or protein-protein interaction.

The PDB is only used as a repository for resolved protein structures. Explicit information about folding classification or the participation in metabolic pathways is completely absent from the records stored.

For this reason, we have created an integrated database, called COLUMBA, where we annotate protein structures from the PDB with several other data sources to conduct further research.

Database Schema

At the moment we integrate information from 11 different sources apart from the PDB. For each entry we annotate the compounds with information about the enzyme classification from ENZYME, the participation in molecular pathways from the Kyoto Enzyclopedia of Genes and Genomes, KEGG, and the Roche Biochemical Pathways.

Structural classification from SCOP as well as CATH is added for each chain. In addition to that, we calculate the secondary structure with the DSSP program.

To get controlled vocabulary for each chain, links to SwissProt entries, the NCBI Taxonomy and the Gene Ontology (GO) are established, but not yet publicly available. Two different cluster methods are available in COLUMBA, the culled set of protein sequences, PISCES (Wang et al. 2003), and SYSTERS (Krause et al. 2000).

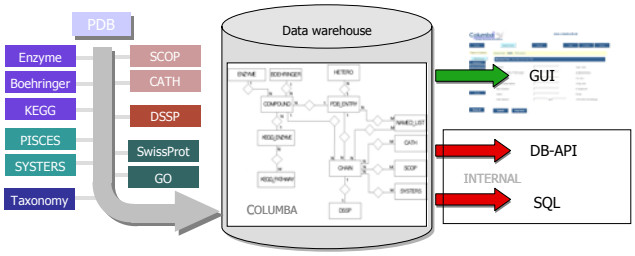

Figure 1: The production workflow, like the schema, is centred around PDB entries. The annotation pipeline, written in Python, generates connections between PDB entries and objects from the other data sources.

The schema COLUMBA has in total 38 tables, of which five are reserved for metadata and manual annotation, eight tables link the data sources to PDB entries, while the remaining tables store information from the original data sources. We use the open source database system PostGreSQL 7.4 which serves as the data warehouse for the integration of the different sources and is hosted at the Zuse Institute Berlin (ZIB).

Source		Number of entries	
PDB	total	24 461	
	compounds	31 536	
	chains	56 182	
ENZYME		4 242	(1 023)
KEGG	pathways	119	(108)
	enzymes	1 903	(643)
Boehringer		1 113	(625)

Table 1: This table shows the number of entries in the COLUMBA database. For the selected data sources the number of entries is given. The number in brackets represents the number of entries, which could be linked to at least one PDB entry.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E.. The Protein Data Bank. *Nucleic Acids Research.* 28: 235-242. (2000)

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)

Wang, G., and Dunbrack, R.L. Jr. PISCES: a protein sequence culling server. *Bioinformatics.* 19:1589-91. (2003)

Krause, A., Stoye, J., and Vingron, M. The SYSTERS protein sequenc cluster set. Nucleic Acids Res. 28(1):270-272. (2000)

Web interface

We have created a user friendly web interface, which is publicly available at www.columba-db.de. The web interface uses a "query refinement" paradigm to return a set of PDB entries, which fulfill the conditions stated. A query is defined by entering restriction conditions in the form for the data source specific annotation. The user can combine queries from different data sources, which act as filters, to obtain the desired subset of PDB entries.

The interface supports interactive and exploratory usage by straightforward adding, deleting, restricting, or easing of conditions. The user is supported by a header, called "filter chain", where the number of PDB entries after each filter step is stated.

The user can see the full scope of COLUMBA in the Explorer for a single entry, where all the associated annotation for that particular entry is shown.

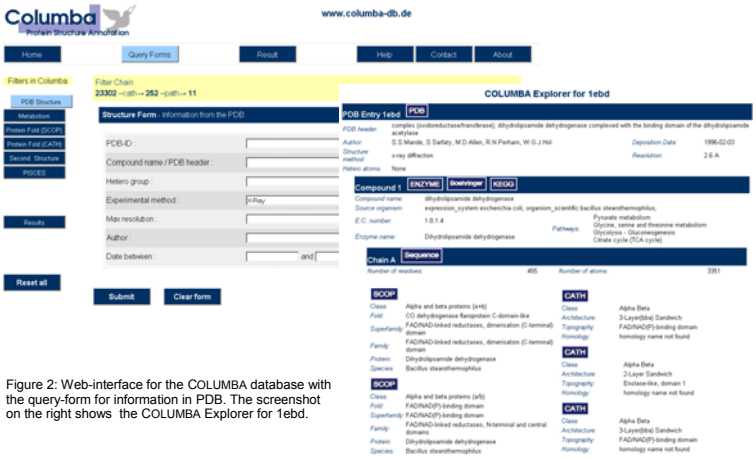

Figure 2: Web-interface for the COLUMBA database with the query-form for information in PDB. The screenshot on the right shows the COLUMBA Explorer for 1ebd.

Applications

We found that for metabolic pathways in KEGG, on average 43.6 % of the participating enzymes are structurally resolved. Several pathways have no or insignificant coverage. This result might help to direct the research on structural genomics towards a total determination of enzyme structures.

Pathway	Total number of		Enzymes with one or more structures in %
	Enzymes	Structures	
Erythromycin biosynthesis	6	48	100.0
Carbon fixation	23	354	82.6
Pyrimidine metabolism	60	709	63.3
Citrate cycle (TCA cycle)	23	134	47.8
Glycerolipid metabolism	74	542	33.7
Vitamin B6 metabolism	20	22	30.0
Alkaloid biosynthesis I	34	101	14.7
Retinol metabolism	10	0	0.0

Table 2: Coverage of selected pathways with structurally resolved enzymes.

We integrated 3D-virtual screening data comparing patches of protein surfaces, DIP (Preissner et al. 1998). 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.

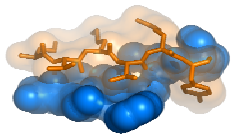

Figure 3: Molecular interface betweeen two beta-sheets from Subtilisin Carlsberg (1cse).

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 sponsored by bmb+f