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, Germany*,
[2]*Institute for Biochemistry, Charite-Campus Mitte, Monbijoustr. 2, 10098 Berlin, Germany*

Researchers interested in the analysis of protein structures often encounter both technical and data quality problems on collecting information about proteins. We have created a relational database, called Columba, which integrates many resources on protein structures.

Columba is centered around proteins whose structure has been resolved and adds as much annotations as possible to them by describing their properties. We obtained the structurally resolved proteins from the Protein Data Bank (PDB) and extracted annotations from seven external data sources, including folding classification from SCOP and CATH, secondary structures calculated with DSSP, metabolic pathways from KEGG, and enzyme annotation from the ENZYME database. We introduced a flexible software framework, implemented in Python, to populate the database and establish links between the various data sources.

Columba does not try to remove redundancies and overlaps in data sources, but views each source as a proper dimension describing a protein. This allows the user to easily identify the concept and terminology of the original source. That is also reflected in the web interface, available at www.columba-db.de. The web interface uses a 'query refinement' paradigm to return a subset of PDB entries, where a query is defined by entering restriction conditions on the data source specific annotations. The user can combine several queries, acting as filters, to obtain the desired subset of PDB entries.

One interesting question, which we can answer very easily, is the coverage of metabolic pathways with PDB structures. We found that some pathways are highly covered, e.g. the Carbon fixation pathway with 83% coverage, while others have no structure assigned to them at all. This example shows that Columba can be a very useful tool for protein researchers to find sets of proteins, sharing certain properties.