Columba - A Database of Annotations of Protein Structures

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

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

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

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, protein-fold classification, secondary structure, and taxonomic information.

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 compile sets of protein structures sharing certain properties and conduct research on them.

Columba can answer questions such as

- Which mammalian proteins do have a TIM barrel fold?
- For which enzymes in the citric acid cycle exists a resolved structure?
- Which proteins with transporter activity have a resolution less than 2.0 Å?

with a single query.

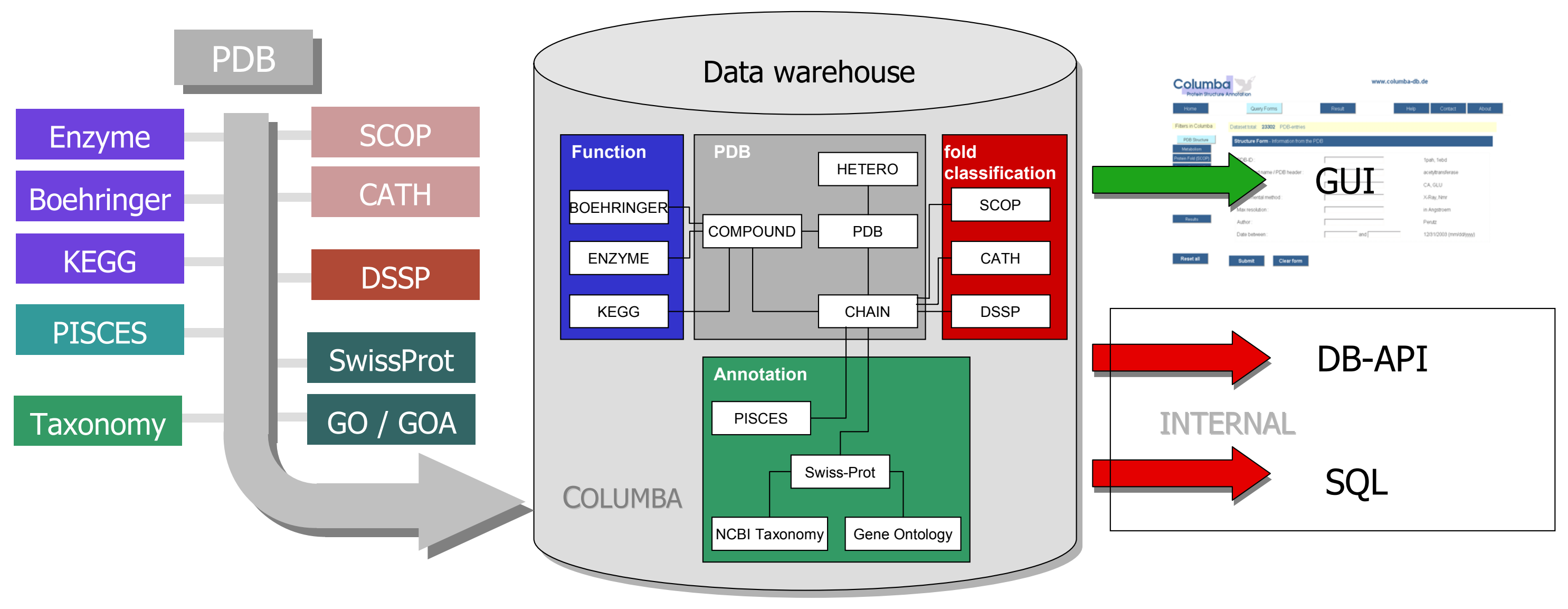
Database Schema

At the moment we integrate information from 11 different sources apart from the PDB.

We annotate each compound from a PDB entry 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. Each chain is assigned to a PISCES (Wang et al. 2003) cluster, according to sequence similarity and experimental properties.



The production workflow (left), like the schema (center), 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 it is hosted at the Zuse Institute Berlin (ZIB).

References

Rother, K., Mueller, H., Trissl, S., Koch, I., Steinke, T., Preissner, R., Froemmel, C., and Leser, U. COLUMBA: Multidimensional Data Integration of Protein Annotations. *In Data Integration in the Life Sciences, (ed. Erhard Rahm.).* pp. 156 – 171. Springer, Leipzig, Germany. (2004)

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)

Wang, G., and Dunbrack, R.L. Jr. PISCES: a protein sequence culling server. *Bioinformatics.* 19: 1589 – 1591. (2003)

Web interface

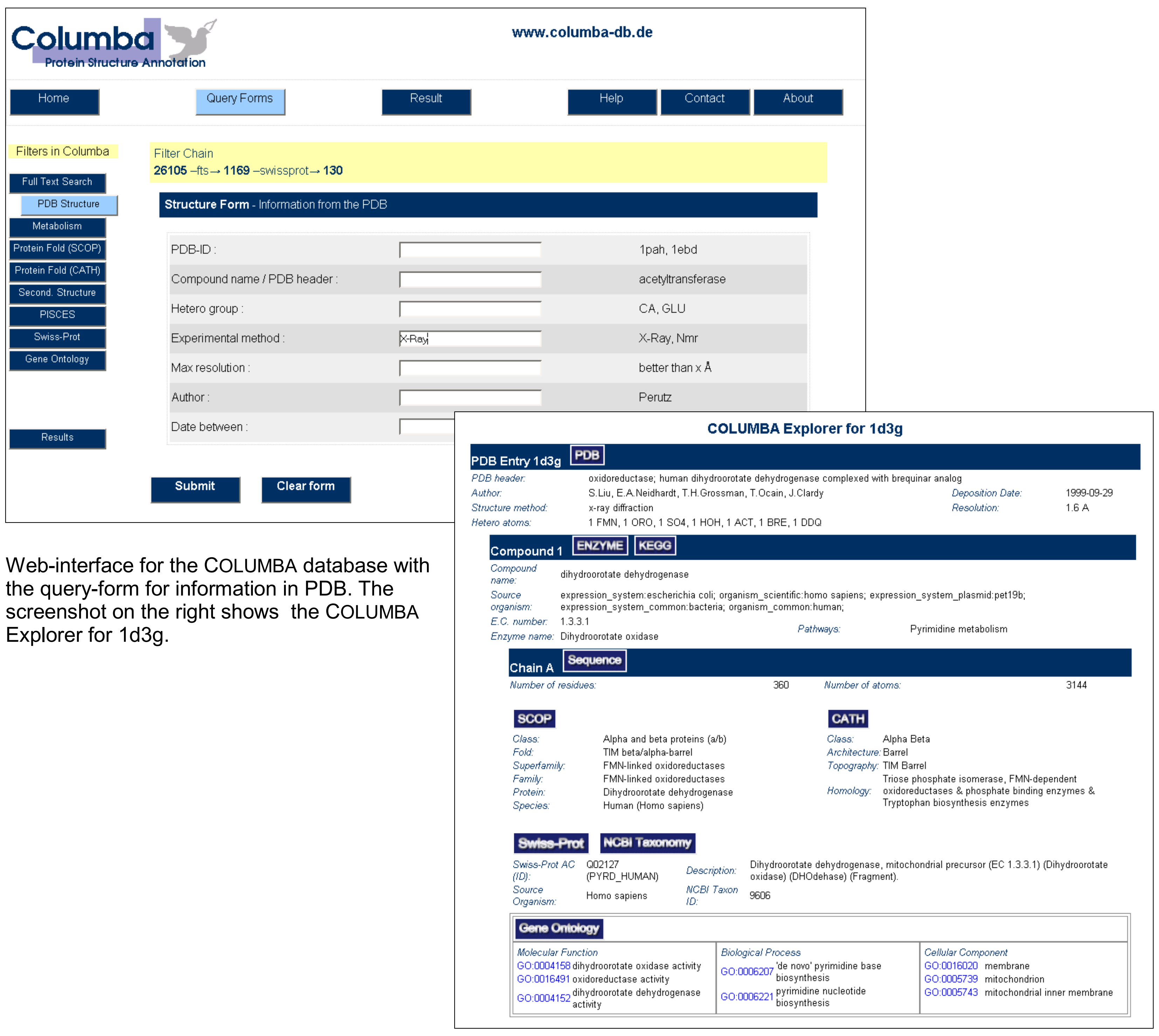
We have created a web interface, which is publicly available at www.columba-db.de. The web interface supports

- full text search and
- data source specific queries.

After entering conditions in a form, the result set is implicitly restricted to those PDB entries that fulfill the stated conditions. Several forms can be used consecutively to restrict the original set of PDB entries by conditions on multiple data sources. In addition to the source specific forms, the full-text search can be used as an additional restriction condition on the result set.

To guide the user, COLUMBA constantly shows the number of qualifying PDB entries after each query step in the header of the page. This demonstrates the consequences of adding, deleting, or changing conditions.

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.



www.columba-db.de

Filter Chain
26105 –fts→ 1169 –swissprot→ 130

Structure Form - Information from the PDB

PDB-ID : 1pah, 1ebd
Compound name / PDB header : acetyltransferase
Hetero group : CA, GLU
Experimental method : X-Ray | X-Ray, Nmr
Max resolution : better than x Å
Author : Perutz
Date between :

COLUMBA Explorer for 1d3g

PDB Entry 1d3g
PDB header: oxidoreductase; human dihydroorotate dehydrogenase complexed with brequinar analog
Author: S.Liu, E.A.Neidhardt, T.H.Grossman, T.Ocain, J.Clardy | Deposition Date: 1999-09-29
Structure method: x-ray diffraction | Resolution: 1.6 A
Hetero atoms: 1 FMN, 1 ORO, 1 SO4, 1 HOH, 1 ACT, 1 BRE, 1 DDQ

Compound 1
Compound name: dihydroorotate dehydrogenase
Source organism: expression_system:escherichia coli; organism_scientific:homo sapiens; expression_system_plasmid:pet19b; expression_system_common:bacteria; organism_common:human;
E.C. number: 1.3.3.1
Enzyme name: Dihydroorotate oxidase | Pathways: Pyrimidine metabolism

Chain A
Number of residues: 360 | Number of atoms: 3144

SCOP
Class: Alpha and beta proteins (a/b)
Fold: TIM beta/alpha-barrel
Superfamily: FMN-linked oxidoreductases
Family: FMN-linked oxidoreductases
Protein: Dihydroorotate dehydrogenase
Species: Human (Homo sapiens)

CATH
Class: Alpha Beta
Architecture: Barrel
Topography: TIM Barrel
Homology: Triose phosphate isomerase, FMN-dependent oxidoreductases & phosphate binding enzymes & Tryptophan biosynthesis enzymes

Swiss-Prot AC (ID): Q02127 (PYRD_HUMAN) | Description: Dihydroorotate dehydrogenase, mitochondrial precursor (EC 1.3.3.1) (Dihydroorotate oxidase) (DHOdehase) (Fragment).
Source Organism: Homo sapiens | NCBI Taxon ID: 9606

Molecular Function	Biological Process	Cellular Component
GO:0004158 dihydroorotate oxidase activity	GO:0006207 'de novo' pyrimidine base biosynthesis	GO:0016020 membrane
GO:0016491 oxidoreductase activity	GO:0006221 pyrimidine nucleotide biosynthesis	GO:0005739 mitochondrion
GO:0004152 dihydroorotate dehydrogenase activity		GO:0005743 mitochondrial inner membrane

Web-interface for the COLUMBA database with the query-form for information in PDB. The screenshot on the right shows the COLUMBA Explorer for 1d3g.

Applications see also Poster C35

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	
Carbon fixation	23	354	82.6
Glycolysis / Gluconeogenesis	38	519	76.3
Citrate cycle (TCA cycle)	23	140	52.1
Glycerolipid metabolism	75	591	37.3
Vitamin B6 metabolism	20	24	30.0
Alkaloid biosynthesis I	34	105	14.7
Retinol metabolism	10	0	0.0

Coverage of selected pathways with structurally resolved enzymes.

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

Nationales Genomforschungsnetz

and is supported by

sponsored by bmb+f