Nucleic Acids Research, 2004, Vol. 32, Database issue ***D129–D133***
DOI: 10.1093/nar/gkh028

The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data

Craig T. Porter[1], Gail J. Bartlett[1,2] and Janet M. Thornton[1,*]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK

Received August 15, 2003; Revised August 21, 2003; Accepted September 3, 2003

ABSTRACT

The Catalytic Site Atlas (CSA) provides catalytic residue annotation for enzymes in the Protein Data Bank. It is available online at http://www.ebi.ac.uk/thornton-srv/databases/CSA. The database consists of two types of annotated site: an original hand-annotated set containing information extracted from the primary literature, using defined criteria to assign catalytic residues, and an additional homologous set, containing annotations inferred by PSI-BLAST and sequence alignment to one of the original set. The CSA can be queried via Swiss-Prot identifier and EC number, as well as by PDB code. CSA Version 1.0 contains 177 original hand-annotated entries and 2608 homologous entries, and covers ~30% of all EC numbers found in PDB. The CSA will be updated on a monthly basis to include homologous sites found in new PDBs, and new hand-annotated enzymes as and when their annotatation is completed.

INTRODUCTION

Enzymes are amongst the most studied biological molecules and are vital for all processes of life. The catalytic activity of an enzyme is performed by a small, highly conserved constellation of residues within the active site. Additionally, binding interactions in the active site allow the recognition and precise positioning of an enzyme's substrate in proximity to the chemically active catalytic residues and lower the energy of the transition state, which aids catalysis. Unlike the catalytic residues, residues responsible for binding the substrate are not as vital to the catalytic function of the enzyme and can change through evolution, sometimes allowing the enzyme to accommodate new substrates. Detailed information regarding enzyme active sites and the residues explicitly involved in catalysis is essential for understanding the relationship between protein structure and function, novel enzyme design and the design of inhibitors.

Databases such as Swiss-Prot (1) and BRENDA (2) contain an enormous wealth of data on enzymes. BRENDA currently has 3600 different EC numbers, i.e., 3600 different enzyme reactions. Swiss-Prot currently contains ~130 000 sequence entries, just over 42 000 of which have an assigned EC number. There are ~10 200 PDB entries with an assigned EC number. As each enzyme catalyses a reaction, it would be useful to have annotation which describes the residues implicated in catalysis. Incorporating this annotation with enzyme structure would additionally be invaluable. However, annotation of functional site residues both in the literature and other databases is variable, and subject to the author's interpretation of the word 'function'. Indeed, we find that the 'SITE' records in PDB (3) are used for many different types of functional annotation, such as substrate binding or cofactor binding residues, and allosteric sites.

In order to perform an analysis of residues involved in enzyme catalysis (4), we defined a classification of catalytic residues which includes only those residues which are thought to be directly involved in some aspect of the reaction carried out by an enzyme. For enzymes of known structure and catalytic mechanism, catalytic residues are defined by manual inspection of the primary literature. A feature of these catalytic residues is that they are highly conserved in sequence. Here we present a web server for our database of catalytic residues. The Catalytic Site Atlas (CSA) includes hand-annotated descriptions of these enzyme active sites, as well as equivalent sites in related proteins found subsequently by sequence alignment with the original set of enzymes.

ASSIGNMENT OF CATALYTIC RESIDUES

Principles of annotation by hand

A data set of non-homologous enzymes of known structure with a well-defined active site and plausible catalytic mechanism was constructed (4). Enzymes are chosen for this analysis primarily by EC number, obtained from the Enzyme Structures Database (5) (http://www.biochem.ucl.ac.uk/bsm/enzymes/index.html) and retained in the data set if there is an available X-ray crystal structure or NMR model, and if sufficient information concerning active site, overall reaction catalysed and catalytic mechanism can be obtained from the

*To whom correspondence should be addressed. Tel: +44 1223 494648; Fax: +44 1223 494468; Email: thornton@ebi.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

primary literature. Additional cross-checks are performed against Web of Knowledge (http://wok.mimas.ac.uk) to ensure that the most up-to-date information from the primary literature is incorporated for each enzyme. Residues are defined as catalytic if they fulfil any one of the following criteria:

(i) direct involvement in the catalytic mechanism, e.g. as a nucleophile;

(ii) alteration of the pK_A of a residue or water molecule directly involved in the catalytic mechanism;

(iii) stabilization of a transition state or intermediate, thereby lowering the activation energy for a reaction;

(iv) activation of the substrate in some way, e.g. by polarizing a bond to be broken.

Our classification excludes residues involved in ligand binding unless they also fulfil one of the above criteria.

Principles of inference for homologues in PDB

In order to reduce the need for manual annotation a protocol was developed to allow annotation of related structures in the PDB. The sequence of each enzyme which has been manually annotated was taken from the PDB sequences repository held at EBI (pdb_aa.fasta, available from the MSD group FTP site at ftp://ftp.ebi.ac.uk/pub/databases/msd/) and subjected to PSI-BLAST (6) analysis (using a cut-off of <0.0005 for inclusion in the developing profile) against a composite database consisting of the Non-Redundant DataBase (NRDB) from NCBI and protein sequences extracted from structures found in the PDB. The alignment for each enzyme was inspected and homologous PDB sequences identified (i.e. homologues with a protein structure). The equivalent residues to those residues annotated as catalytic in our previous work were taken from the multiple sequence alignment and if they were found to be identical to the catalytic residues (only one residue change is allowed per site, to account for the many single site mutants in PDB), a record for this enzyme structure was created, noting equivalent catalytic residues and the provenance of the information.

Technical implementation

Each enzyme in the data set and its corresponding catalytic residues were stored in a MySQL database, with assignments based on PDB code and PDB residue numbering. Links were generated to Swiss-Prot and to the ENZYME (9) database. An online version is produced 'on the fly' by querying this CSA database.

Assessment of coverage

We assessed the coverage of the database by comparing the annotation of our original 177 enzyme set with the 'ACT_SITE' annotation in Swiss-Prot and the SITE records found in PDB. The Swiss-Prot identifier for each original entry was taken either from the PDB file, or by a BLAST search of the enzyme sequence against Swiss-Prot and finding a 100% match. One hundred and seventy-four enzymes could be assigned a Swiss-Prot identifier. Each Swiss-Prot entry was examined and ACT_SITE annotations retrieved and compared with the CSA annotation for that enzyme. The SITE records of each enzyme PDB entry were retrieved and compared with the CSA annotation of the enzyme.

RESULTS

CSA version 1.0 data set

The database contains entries of two types, the 'original' set of enzymes and a 'homologous' set, identified by PSI-BLAST. For the 'original' set, there is good experimental knowledge of the reaction catalysed, and details of the catalytic mechanism, validated where possible by experimental data (e.g. site directed mutagenesis and kinetic data). For the 'homologous' set, we are inferring, via sequence analysis, the function of the enzyme and the residues that may be involved in catalysis. Each enzyme entry lists a number of sites, in the form of a list of residues. There are 177 'original' entries and 2608 'homologous' entries, with a total of 17 917 residues annotated. Each site has an evidence tag, which provides information on the source of the site. If the site is from an 'original' enzyme, the evidence tag is a literature reference. If the site is from a 'homologous' enzyme, the evidence tag is a PSI-BLAST hit to one of the 'original' enzymes. An example is shown schematically in Figure 1. Enolase (PDB code 5enl), one of the original enzyme set, is shown in the top box with its catalytic residues annotated. Its homologues, found by PSI-BLAST, are listed in the central box, and one of these homologues, another enolase (PDB code 1pdz) is shown in the bottom box, with its equivalent residues annotated. CSA Version 1.0 contains only well-understood enzymes in the original data set, hand annotated, with annotation of residue function. It is planned that Version 2.0 will be available in 2004, with 500 hand-annotated 'original' enzymes, plus their homologues as identified using the same protocol as described previously. In addition, Version 1.0 will be updated on a monthly basis, with PSI-BLAST runs against new PDBs so that new homologues can be identified.

The CSA webserver

The online version of the CSA is available at http://www.ebi.ac.uk/thornton-srv/databases/CSA/index.html. Queries can be made by PDB code, Swiss-Prot (1) identifier or EC number (7). Each entry is presented with the catalytic residues annotated, highlighted on the amino acid sequence and on the structure, via a link to a RasMol (8) script. Additional information on the function of annotated residues, references to literature used to produce the database, and additional notes on the enzyme function and the reaction catalysed are available for the original entries. Links are provided to the ENZYME database, PDBsum (5) and Swiss-Prot. Additionally, for each original entry, a link is available to a list of the homologues found by sequence as described above.

The CSA database will be updated automatically on a monthly basis, to include all new homologues found by a PSI-BLAST search of an updated composite database (NRDB + PDB + new PDBs included), and will be updated with new hand-annotated 'originals' as and when their annotation is complete.

Comparison with Swiss-Prot and PDB

A comparison of annotation of original enzymes in the CSA with Swiss-Prot and SITE records in PDB can be seen in Figure 2. Of 174 enzymes with a Swiss-Prot identifier, we found 180 annotated ACT_SITE residues, as compared with

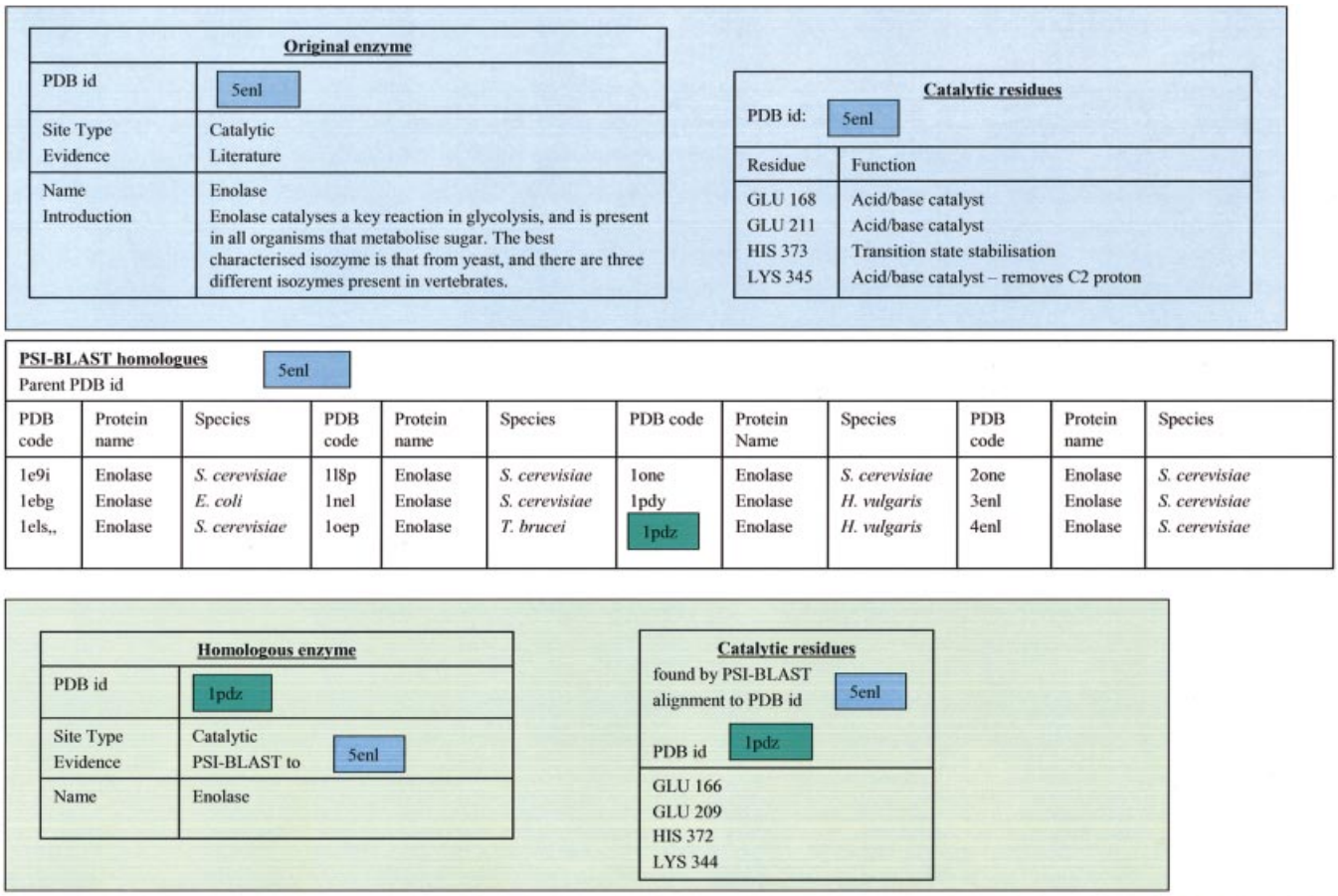

Figure 1. Schematic description of CSA database. The blue block demonstrates the 'original' enzyme data for enolase, PDB code 5enl. The central block shows 'homologous' enzymes identified by PSI-BLAST analysis, and one of these, PDB code 1pdz, is highlighted in green as an example homologue. The green block at the bottom demonstrates the 'homologous' enzyme datzyme data set, in Venn diagram form. The number of residues annotated by each database is given, along with the number of identical residues found in both databases.

614 in CSA. Of these 180 annotated in Swiss-Prot, 157 are also annotated as catalytic in the CSA. Additionally, 23 residues are annotated as ACT_SITE in Swiss-Prot, but are not annotated in the CSA. In some cases, these Swiss-Prot annotations represent binding sites for allosteric interactions and some ligands, in other cases they describe an activity not covered by the CSA annotation, for example, dehydroquinate synthase (1dqs) has the Swiss-Prot identifier ARO1_YEAST. This is a polyprotein with five separate activities. The ACT_SITE annotation refers to catalytic activity in the other four activities found in this polyprotein.

Of 177 enzymes, there are 611 residues annotated as SITE in the PDB, compared with 614 in the CSA. Of these, only 127 are annotated as catalytic in the CSA. The 'SITE' annotation in PDB lists a further 484 residues which are not annotated as catalytic in the CSA. These are predominantly binding sites and other functional sites that do not fall into the strict 'catalytic' classification to which the CSA adheres. The CSA annotates a further 487 residues which are not annotated in PDB. The basis of our annotation is therefore more similar to that adopted by Swiss-Prot, although the Swiss-Prot annotations are much more conservative. PDB annotation is much less well defined.

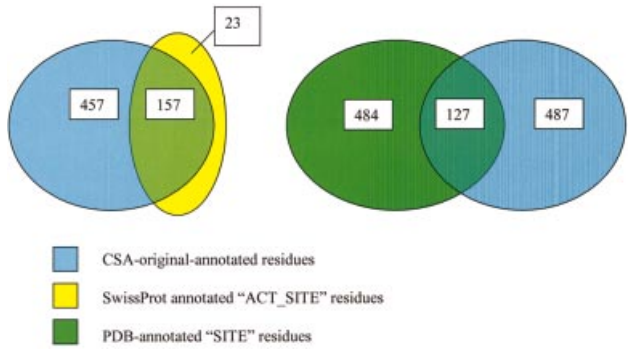

Figure 2. A comparison of residue annotation between the CSA and SwissProt (left) and PDB (right) for the 'original' enzyme dataset, in Venn diagram form. The number of residues annotated by each database is given, along with the number of identical residues found in both databases.

EC coverage

There are ~10 200 PDB entries with an assigned EC number. The CSA annotates 2785 of these, a coverage of 27%. This is partially a reflection of the size of the families annotated, and

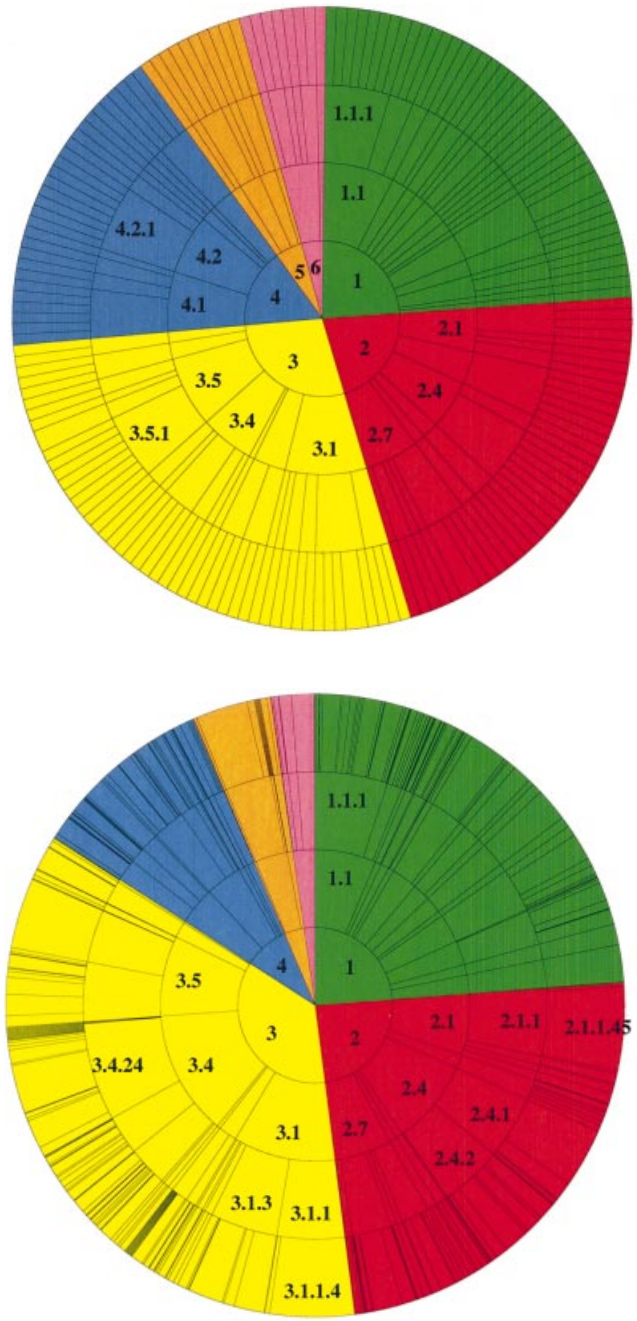

Figure 3. EC wheel functional description of the 'original' enzyme set (above) and the 'homologous' enzyme set (below) in the CSA database. The EC classification (7) assigns a four-digit number to the reaction catalysed by an enzyme, where the first digit denotes the class of reaction (1.-.-.- oxidoreductases, 2.-.-.- transferases, 3.-.-.- hydrolases, 4.-.-.- lyases, 5.-.-.- isomerases and 6.-.-.- ligases) and subsequent levels define the reaction in more detail by substrate, bond broken, etc. The outer wheel segments are proportional to the number of enzyme structures in the database with a particular EC number.

also the fact that PSI-BLAST misses the most distant relatives. We plan to improve the coverage by using 3D motifs to pull in these distant relatives (see Discussion). Additionally, many EC numbers in PDB are duplicates. There are 954 unique EC numbers in PDB, and the CSA annotates 286 of these, which is just less than half. In Swiss-Prot, there are ~42 000 entries with an assigned EC number, but again, there are many duplicates; just under 2000 of these are unique. Our coverage obviously reflects the EC coverage found in PDB rather than that found in Swiss-Prot. EC numbers covered by the 'original' enzyme data set and the 'homologous' set can be found in Figure 3.

DISCUSSION AND PERSPECTIVES

We have set up a catalytic site atlas which contains hand-curated core data on enzyme active sites and catalytic residues. The core data are supplemented by enzyme homologues with equivalent residues found using sequence searching. The core data are limited to a set of enzymes with well-defined structures and catalytic mechanisms. It is by no means an exhaustive collection of enzymes in PDB. However, our comparison of the core data with Swiss-Prot and PDB shows that it is at least as well-annotated as Swiss-Prot, and is much more comprehensive and specific to catalysis than the information provided by the SITE records in PDB.

We plan to extend the database to allow for classification of enzymes based on 3D motifs. Previous analysis (11) has shown that catalytic residues can be conserved in structure but not in sequence. We plan to address these factors by using 3D templates similar to those used in the PROCAT (12) website, and an improved search algorithm (13) alongside sequence conservation in future releases of this database of catalytic residues. Finding a good template match in 3D will provide additional validation that the functional inference made from sequence is correct. This will enable us to assign multi-chain sites, and find sites that are only identifiable through structure comparison, and additionally, examples of convergent evolution.

It is well known that enzymes that perform the same function can utilize a variety of different mechanisms to catalyse a particular reaction. We have recently started a collaborative project to classify enzyme reactions by mechanism (G. L. Holliday and G. J. Bartlett, unpublished results). It is hoped that ultimately, this information can be integrated into or linked with the CSA.

We plan to update the database automatically on a monthly basis to include sites found by PSI-BLAST to new release PDBs. Additionally, a large update is planned for 2004 to incorporate ~500 new 'original' hand-annotated enzyme active sites and their homologues. In the longer term, it is hoped that the data we have collated will eventually become part of the large Macromolecular Structural Database resource at EBI and hence of Swiss-Prot.

REFERENCES

1. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
2. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.
3. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D.*, **58**, 899–907.
4. Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.

5. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
6. Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.
7. Webb,E.C. (1992) *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York.
8. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
9. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
10. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
11. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
12. Wallace,A.C., Borkakoti,N. and Thornton,J.M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.
13. Barker,J.A. and Thornton,J.M. (2003) An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. *Bioinformatics*, **19**, 1644–1649.