BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 23 2005, pages 4315–4316
doi:10.1093/bioinformatics/bti693

Databases and ontologies

MACiE: a database of enzyme reaction mechanisms

Gemma L. Holliday[1], Gail J. Bartlett[2,†], Daniel E. Almonacid[1], Noel M. O'Boyle[1], Peter Murray-Rust[1], Janet M. Thornton[2] and John B. O. Mitchell[1,*]

[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK and [2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received on July 21, 2005; revised on September 22, 2005; accepted on September 23, 2005
Advance Access publication September 27, 2005

ABSTRACT

Summary: MACiE (mechanism, annotation and classification in enzymes) is a publicly available web-based database, held in CMLReact (an XML application), that aims to help our understanding of the evolution of enzyme catalytic mechanisms and also to create a classification system which reflects the actual chemical mechanism (catalytic steps) of an enzyme reaction, not only the overall reaction.
Availability: http://www-mitchell.ch.cam.ac.uk/macie/
Contact: jbom1@cam.ac.uk

A great deal of knowledge about enzymes, including structures, gene sequences, mechanisms, metabolic pathways and kinetic data, now exists. However, it is spread between many different databases and throughout the literature. Here we announce the completion of the initial version of MACiE, a unique database of the chemical mechanisms of enzymatic reactions.

Web resources such as BRENDA (Schomburg *et al*., 2004), KEGG (Kanehisa *et al*., 2004) and the International Union of Biochemistry and Molecular Biology (IUBMB) Enzyme Nomenclature website (IUBMB, 2005, http://www.chem.qmul.ac.uk/iubmb/enzyme/) contain descriptions of the overall reactions performed by enzymes, accompanied in some cases by a textual or graphical description of the mechansim. MACiE is unique in combining detailed stepwise mechanistic information (including 2D animations), a wide coverage of both chemical space and the protein structure universe, and the chemical intelligence of CMLReact (Holliday,C.L., Murray-Rust,P., and Rzepa,H.S., 2005, manuscript submitted to *J. Chem. Inf. Modeling*). MACiE usefully complements both the mechanistic detail of the Structure–Function Linkage Database (SFLD) for a small number of enzyme superfamilies (Pegg *et al*., 2005) and the wider coverage with less chemical detail provided by EzCatDB (Nagano, 2005) which also contains a limited number of 3D animations.

*To whom correspondence should be addressed.
†Present Address: Bioinformatics Support Service (Biochemistry Building), Centre for Bioinformatics, Division of Molecular Biosciences, Faculty of Life Sciences, Imperial College London, London, SW7 2AZ, UK

DESIGN

The MACiE dataset evolved from that published in the Catalytic Site Atlas (CSA) (Bartlett *et al*., 2002; Porter *et al*., 2004), and each entry is selected so that it fulfils the following criteria:

(1) There is a 3D crystal structure of the enzyme deposited in the Protein Databank (PDB) (Berman *et al*., 2000).
(2) There is a relatively well-understood mechanism available. Taken from the literature, these cover a variety of methodologies, including chemical and biochemical studies, quantum mechanical calculations and structual biology reports.
(3) The enzyme is unique at the H level of the CATH classification—a hierarchical classification system of protein domain structures (Orengo *et al*., 1997)—unless there is a homologue with a significantly different chemical mechanism.
(4) Where there are a number of possible PDB codes available the entry should be, if possible, a wild-type enzyme.

All MACiE enzymes are also contained in the Enzyme Commission (EC) classification system (IUBMB, 2005, http://www.chem.qmul.ac.uk/iubmb/enzyme/), that is, they all have four number codes describing their overall reaction. The first level (Class) describes the basic reaction type. The second and third levels (subclass and sub-subclass, respectively) describe the reaction in further detail and the final level (serial number) describes substrate specificity. For example, the β-lactamases (Fig. 1) are assigned the EC number 3.5.2.6, i.e. a hydrolase (3) acting on a C–N bond (5) in a cyclic amide (2) with a β-lactam as the substrate (6).

In MACiE, the data centre on the catalytic steps involved in the chemical mechanism as well as the overall reaction. Each entry includes the following steps:

- Enzyme name and EC number
- PDB code and CATH codes of all domains in the enzyme
- Diagram and annotation of the overall reaction
- Primary literature references

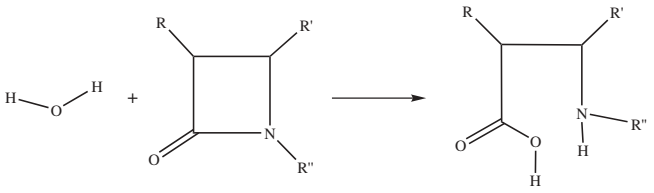

Fig. 1. The overall reaction for a β-lactamase.

- Diagram and annotation of all reaction steps, including:
  —The Ingold mechanism (Ingold, 1969)
  —Diagram and function of catalytic amino acid residues
  —Information on the reactive centres and bond changes
- Comments on the reaction (where applicable).

CONTENT

The criteria defined in the Design section initially produced a dataset of 100 entries. A single EC number may cover a plurality of MACiE entries when different mechanisms bring about the same overall chemical transformation, as with the two types of 3-dehydroquinate dehydratase, and thus 100 MACiE entries span only 96 EC numbers.

The 100 enzymes in Version 1 of MACiE incorporate domains from 140 CATH homologous superfamilies. MACiE currently covers 56 of the 174 EC sub-subclasses present in the PDB, thus, we feel that we have a representative coverage of EC reaction space (comparative EC wheels are available at URL http://www-mitchell.ch.cam.ac.uk/macie/ECCoverage/). We anticipate that all 158 sub-subclasses for which both structures and reliable mechanisms are available will be represented in the forthcoming MACiE Version 2.

SOFTWARE

The data are initially entered in MDL's ISIS/Base, a database package for chemical reactions, validated by at least two people, and then converted into CMLReact using the Jumbo Toolkit (Wakelin *et al*., 2005) to create an information and semantically rich database. At this stage we add extra fields of information to the CMLReact version of MACiE that are unavailable in the ISIS version, including the CATH code. Jumbo is a set of Java-based software which converts the MDL file format produced from ISIS/Base into CMLReact. The MacieConverter section of Jumbo performs the following functions:

- Integration of the files in the ISIS/Base version of MACiE
- Identification of reactant, product and spectator molecules
- Splitting of groups of molecules
- Automatic mapping of atoms within the reaction
- Checking for mass and charge conservation throughout the reaction (stoichiometry)
- Integration and checking of MACiE Dictionary entries.

Once the conversion process has been completed, a further tool in the Jumbo Toolkit, called CMLSnap (Holliday *et al*., 2004), can be used to create an animation of the reaction. This animation includes all of the atoms and bonds involved as well as the electron movements, which are calculated automatically. It is expected that CML will become our primary method of data entry and storage.

CURATION

The annotation process involves input and validation steps. Terms have been rigorously defined either from the IUPAC Gold Book (McNaught *et al*., 1997), such as chemical terms like hydrolysis, or from primary literature, such as mechanism, which is defined using Ingold's terminology (Ingold, 1969), originally put forward in the 1930s. All of the technical and scientific terms used in MACiE are contained in the MACiE dictionary, which is available at the URL http://www-mitchell.ch.cam.ac.uk/macie/glossary.html and is also available as a raw XML file.

The entries online are accessed via an HTML look-up table and include all of the information available in the database. The original ISIS/Base format file and the raw CML files can be supplied.

FUTURE WORK

Future work includes expanding the dataset to include a representative set of EC numbers (at the sub-subclass level), creating a search interface for MACiE and developing authoring tools for MACiE in CML. Ongoing research focuses on the evolution of enzyme catalysis and the classification of enzyme reaction mechanisms.

ACKNOWLEDGEMENTS

G.J.B. would like to thank Dr Jonathan Goodman for his invaluable help with organic chemistry queries. We would also like to thank the EPSRC (G.L.H. and J.B.O.M.), the BBSRC (G.J.B. and J.M.T.—CASE studentship in association with Roche Products Ltd; N.M.O.B. and J.B.O.M.—grant BB/C51320X/1), the Chilean Government's Ministerio de Planificación y Cooperación and Cambridge Overseas Trust (D.E.A.) for funding and Unilever for supporting the Centre for Molecular Science Informatics.

Conflict of Interest: none declared.

REFERENCES

Bartlett,G.J. *et al*. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol*., **324**, 105–121.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Holliday,G.L. *et al*. (2004) CMLSnap: animated reaction mechanisms. *Internet J. Chem*., **7**, Article 4.

Ingold,C.K. (1969) *Structure and Mechanism in Organic Chemistry*. 2nd edn, Cornell University Press, Ithaca, NY, Chapters 5–15.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.

McNaught,A.D. and Wilkinson,A. (1997) International Union of Pure and Applied Chemistry Compendium of Chemical Terminology (''The Gold Book''). 2nd edn, ISBN 0-8-654-26848.

Nagano,N. (2005) EzCatDB: the Enzyme Catalytic-mechanism Database. *Nucleic Acids Res*., **33**, D407–D412.

Orengo,C.A. *et al*. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Pegg,S.C-H. *et al*. (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. *Pac. Symp. Biocomput*., 358–369.

Porter,C.T. *et al*. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*., **32**, D129–D133.

Schomburg,I. *et al*. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*., **32**, D431–D433.

Wakelin,J. *et al*. (2005) CML tools and information flow in atomic scale simulations. *Mol. Simul*., **31**, 315–322.