Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites

António J. M. Ribeiro[1],[*],[†], Gemma L. Holliday[1],[†], Nicholas Furnham[2], Jonathan D. Tyzack[1], Katherine Ferris[1] and Janet M. Thornton[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 1HT, UK

ABSTRACT

M-CSA (Mechanism and Catalytic Site Atlas) is a database of enzyme active sites and reaction mechanisms that can be accessed at www.ebi.ac.uk/thornton-srv/m-csa. Our objectives with M-CSA are to provide an open data resource for the community to browse known enzyme reaction mechanisms and catalytic sites, and to use the dataset to understand enzyme function and evolution. M-CSA results from the merging of two existing databases, MACiE (Mechanism, Annotation and Classification in Enzymes), a database of enzyme mechanisms, and CSA (Catalytic Site Atlas), a database of catalytic sites of enzymes. We are releasing M-CSA as a new website and underlying database architecture. At the moment, M-CSA contains 961 entries, 423 of these with detailed mechanism information, and 538 with information on the catalytic site residues only. In total, these cover 81% (195/241) of third level EC numbers with a PDB structure, and 30% (840/2793) of fourth level EC numbers with a PDB structure, out of 6028 in total. By searching for close homologues, we are able to extend M-CSA coverage of PDB and UniProtKB to 51 993 structures and to over five million sequences, respectively, of which about 40% and 30% have a conserved active site.

INTRODUCTION

Enzymes are the macromolecules that catalyze the chemical reactions of life. The study of enzymes draws from the fields of biochemistry, genomics, protein structure, organic chemistry, computational chemistry, thermodynamics, and metabolomics, amongst others. As discussed below, current literature and biological databases with enzyme information mirror this diversity.

Enzymes are one of the most common products of the translation of genetic information. Protein sequence databases, most notably UniProtKB and its manually curated subset, Swiss-Prot, capture protein sequence data, including that for enzymes [1]. Sequence is but a small part of understanding how enzymes work, but due to the explosion of sequencing data (there are, at the moment, more than 89 million sequences in UniProtKB across 141 511 proteomes), it is an essential tool if one wants to extend current knowledge throughout the tree of life. Currently, 47.2% (262 310) of the 555 426 reviewed Swiss-Prot entries and 13.5% (12 050 143) of the complete UniProtKB dataset are annotated as enzymes, the latter figure primarily reflecting incomplete knowledge and annotation.

The three-dimensional structures of enzymes, as deposited in the wwPDB [2], are a much richer source of information for understanding function than sequence alone. At the moment, the wwPDB contains 133 397 structures, 45.9% (61 168) of which annotated as enzymes [3]. The number of available enzyme structures is not as vast as the number of sequences (and many sequences have more than one crystal structure), which is an important limitation. Without structure, elucidation of a specific mechanism is much more difficult, because the identification of catalytic and binding residues can only proceed through indirect means. Structure can also help elucidate mechanisms that have been debated for many years (e.g. lysozyme) [4]. Besides revealing functional mechanisms, structure also allows the identification of distant evolutionary links, since structural domains, like the ones defined in CATH [5] and SCOP [6], are more conserved than sequence [7,8]. The version of CATH (v4.1) used here contains 2737 homologous

superfamilies, 46% (1268) of these are annotated with at least one EC number and 845 are found in M-CSA.

The chemical reaction is the complete chemical transformation that the enzyme performs, and is often described with an Enzyme classification (EC) number (9). Currently, there are 6,028 reactions in the EC classification (10). In a strict sense, the EC classifies enzyme reactions, not enzymes, but each EC number is often linked to all the enzyme sequences that catalyse that reaction, even if the enzyme mechanisms differ (e.g. there are three cases of evolution for the chloroperoxidase reaction: EC 1.11.1.10). The primary EC database is hosted in ExplorEnz (11) and replicated by IntEnz (12) and ExPASy ENZYME (10). Other databases of enzyme reactions include more annotation than just the EC data. RHEA (13) describes reactions using ChEBI (14) identifiers to define reactants and products, and provides an RXN file for each reaction. The Enzyme and Reaction subsets of the KEGG database (15) define reactions using molecules from KEGG Compound, and these are also linked to their metabolic database, KEGG Pathway. Databases with functional information, such as BRENDA (16) and Sabio-RK (17), include kinetic data across enzymes of different organisms, mutants, and in different chemical conditions. Enzymes are also crucial components of metabolism databases. Examples of these include Meta-Cyc (18), which contains pathways for several organisms, ReconMap (19), which focus on human metabolism, and the already mentioned KEGG Pathway. In M-CSA, the overall reaction is primarily described as a list of reactants and products and is also assigned an EC number (to the most detailed level possible).

These databases tell us about enzyme function, and where this function fits in the overall scheme of cells, but they do not tell us how enzymes work. This is the unique contribution of the M-CSA (which stands for **M**echanism and **C**atalytic **S**ite **A**tlas) and its parent databases, MACiE (**M**echanism, **A**nnotation and **C**lassification **i**n **E**nzymes), and CSA (**C**atalytic **S**ite **A**tlas). Data about the reaction mechanisms of enzymes are found throughout the literature in several forms. Catalytic residues are usually identified by noting their conservation in related homologues, and by testing the lack of enzyme activity when these residues are mutated. Measuring enzyme activity under several biochemical conditions can also be insightful. For example, simply changing the pH may reveal the protonation states most conducive to catalysis. Protein structures, which can be found in the wwPDB, reveal the position of the active centre and the 3D disposition of catalytic residues. Structures that contain the substrate, the products or a similar molecule are yet more informative. Computational chemistry studies of enzyme mechanisms have been gaining traction over the last decade, propelled by new algorithms and increasing computational power. QM/MM models, in particular, which combine a QM description of the active center—usually using DFT methods—with a Molecular Mechanics description of the rest of the enzyme have the potential to provide a complete prediction of the mechanistic process in atomic detail (20,21).

M-CSA, as well as MACiE (22–24) and CSA (25–27), were created to capture and organize these and other kinds of mechanistic data available in the literature, and to make them available in a standardised and computer readable format for the community. Additionally, these datasets have been helpful to explore overall themes related to enzyme mechanisms such as the evolution of new chemical function and the roles of specific catalytic residues, cofactors, and metal ions in the chemistry of life (28–30). M-CSA includes annotation for the complete catalytic reaction and also for each step of that reaction, which includes the curly arrow description of the stepwise chemical reaction mechanism, the role of each catalytic residue and any cofactors, as well as the primary literature that supports such data. We also provide annotation for the associated protein sequences and structures, and when appropriate, we link to the databases mentioned above. M-CSA is complementary to other mechanism databases like the Structure-Function Linkage Database (SFLD) (31), which annotates diverse enzyme superfamilies (groups of enzymes that are evolutionarily related and perform a disparate set of overall chemical transformations utilising conserved chemical components), and EzCatDB (32), which uses a hierarchic classification of catalytic mechanisms (RLCP) to classify reaction mechanisms.

M-CSA represents a complete overhaul of the MACiE and CSA databases and websites, as well as an update of the content. We merged the two databases together due to their similarities and to avoid duplication of effort. As of September 2017, the updated database contains 423 manually curated entries with detailed reaction mechanisms and 538 manually curated entries where likely catalytic residues have been identified, but the complete mechanism is not known. We extend this annotation to over five million homologues sequences using the UniProtKB reference dataset, and to 51,993 homologous PDB structures. Below, we describe in more detail the features of the M-CSA website and the main differences with respect to the parent databases.

CONTENT UPDATE

Merging CSA and MACiE

MACiE (22–24) was a database of enzyme mechanisms that primarily annotated individual mechanism reaction steps with 2D curly arrow schemes, the roles of catalytic residues and cofactors, and a text description. Other annotation in MACiE included the overall reaction and links to PDB, CATH, UniProtKB and other databases. The CSA (Catalytic Site Atlas) (26,27), on the other hand, was a manually curated database of enzyme catalytic residues. Each entry in CSA contained a list of catalytic residues with their chemical functions annotated, together with the literature evidence, the overall reaction, and a reference PDB entry. Additionally, each entry included a list of homologous PDB structures which greatly increased the coverage of the dataset. The annotation of CSA overlapped with the annotation of MACiE, and there had been some efforts in the past to standardize the two databases using the EMO ontology and controlled vocabulary (27). The only annotation that MACiE lacked, in comparison with CSA, was the function of each catalytic residue at the overall reaction level and the PDB homologues associated with each entry. Apart from that, the CSA annotation was essentially a subset of the MACiE annotation.

The integration of the two databases accomplished here avoids the future duplication of effort and resources which, among other advantages, will make future updates easier. In M-CSA we have two kinds of entries, 'catalytic site' entries that represent the level of annotation associated with CSA entries, and 'detailed mechanism' entries that represent the level of annotation associated with MACiE entries. Both types of entries are now created and updated using the same web interface, and it is easy to promote 'catalytic site' entries to 'detailed mechanism' entries by just adding the missing mechanistic annotation.

While most of the data migration to the new database was automated, part of the merging process was performed manually. Besides the addition of the overall function field for each residue in MACiE, there was some duplication between CSA and MACiE that had to be removed. Duplication was identified initially at the EC, structure and sequence levels, although not all entries that share one of these are duplicates, because the same EC reaction can be catalysed by different catalytic sites/mechanism, and the same structure or sequence can catalyse more than one reaction. Manual curation was necessary to identify true duplicates, and to decide which fields to keep from each entry, if these were inconsistent. A list of current entries with the same EC or UniProtKB ID is given in the SI. During the data integration process, we found 303 CSA entries that were duplicates of existing MACiE entries and 127 CSA entries that were determined to be duplicates of other CSA entries.

Data migration and literature review

Over the years, new studies are published that invalidate, or complement, previous mechanism proposals. In the previous versions of MACiE we could only save a single mechanism proposal for each entry. We now store all the mechanism proposals we can find for each enzyme. We rate these alternatives from one to three stars where: one star denotes a proposed mechanism which has been disproved by more recent data; two stars denote a proposed mechanism that does not explain all the evidence and three stars denotes a proposed mechanism that is consistent with all existing evidence. Of the 423 detailed mechanism entries in the database, 50 entries have more than one mechanism proposal, and 10 have more than two.

Besides adding new mechanism proposals, and new literature references that support new and existing mechanisms, other changes to the curated data included: the update of text fields where necessary; the selection of more relevant PDB representatives or UniProtKB IDs as reference proteins; the addition of new reactions to the same entry (where the same mechanism at the same reaction centre works on different reactants); linking M-CSA reactions to KEGG and RHEA reactions, and lastly, selecting a primary literature reference for each mechanism proposal. In the old database, schemes of reaction were only available as RXN files, which did not include the electron arrows. Arrows were added on top of a static .gif image. In the new version of the website we use MarvinJS to draw the reaction schemes, which allow us to embed the curly arrows in the same file.

At the time of the release all entries coming from both the CSA and MACiE have been brought up to the same stan-
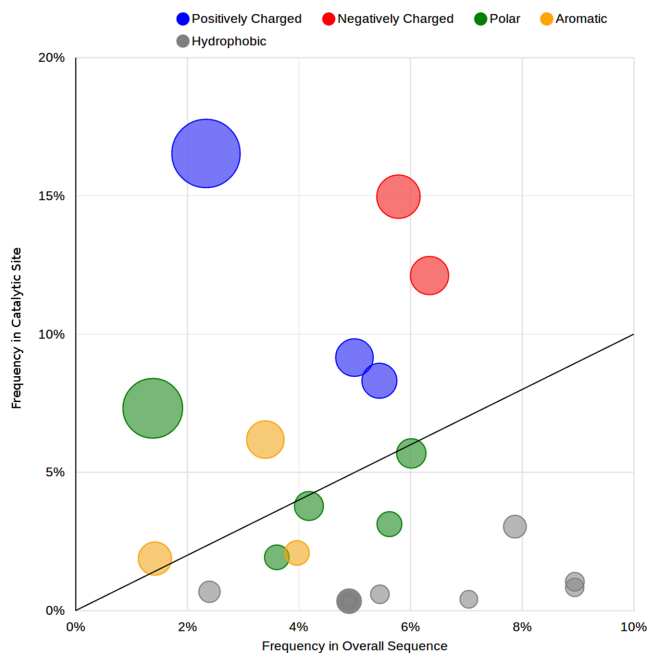

Figure 1. Frequencies and catalytic propensity of the 20 amino acids as they appear in current the M-CSA dataset.

dard, where the annotations for the two types of entries are equivalent, except for the lack of a mechanism in catalytic site entries. Additionally, a complete literature review of all the entries in M-CSA is in progress.

New statistics

The statistics page in the website provide real time analysis of the database. These plots are interactive and some of them link to the appropriate enzyme entries. One example is shown below (two more examples are given in the SI), but we encourage the reader to try others online. We plan to expand these in the future as we further study the M-CSA dataset.

The catalytic propensity of an amino acid captures how frequently an amino acid is involved in catalysis compared with random chance, i.e. the ratio between the percentage of that residue that is catalytic (in the whole M-CSA dataset) and the percentage of that residue in the protein sequences. Figure 1 plots the frequency of the 20 amino acids in the active site of all entries in MACiE against their frequency in the complete protein sequence. The catalytic propensity is indicated by the size of the circle (bigger means higher propensity) and the position in the plot (near the upper left corner means higher propensity). If the propensity is <1, then the propensity for that residue to be catalytic is less than expected by chance, and if it is >1, then the residue is more catalytic than might be expected. Colours indicate the type of amino acid. The distribution is clear: hydrophobic amino acids very rarely have a catalytic role, even though the four most frequent amino acids in these proteins are hydrophobic (alanine, leucine, glycine and valine). Charged amino acids, together with Cysteine, have the highest propensity. Histidine is seven times more likely to appear

as a catalytic amino acid than expected from a random distribution.

FEATURES OF THE NEW WEBSITE AND DATABASE

Home and entry pages

All the pages in the new website were designed from scratch. We try to present all the information captured in the database in a clear and intuitive manner. The user can navigate the website using a navigation bar that is always visible. Searching is also available from the top header in all pages.

The home page contains a brief description of M-CSA, additional navigation links, citation information, and a brief text description of the main statistics. The most prominent part of the homepage is a frame that summarizes the data of an example entry, to give new users an immediate feel for the information contained in M-CSA. This frame contains the identification of the example enzyme, some links to other databases, the description of the enzyme, and an image of one step of the mechanism. The user can click on the image to be taken to the respective entry page.

The entry page contains all the data associated with each enzyme mechanism in the database. A 3D Litemol window shows the reference PDB structure for this enzyme, where the 'show active centre' button can be used to zoom into the catalytic residues which are automatically highlighted, as well as any reactants and/or cofactors that happen to be in the structure. In the bottom part of the entry page, in the mechanism box, there is now a new table, similar to one available for the CSA, that shows the overall role of each catalytic residue in the reaction. Reaction schemes are shown on separate tabs, together with the role of the residues in each step. This page also links to the PDB and UniProt sequence homologues pages, where the user can browse the homologues and check the conservation of the catalytic residues in these sequences. Homologous sequences were found by running every M-CSA reference sequence and PDB sequence against the UniProfRef and PDB databases, respectively, using HMMER (33).

Browse, search and download pages

The browse page gives the user an overall view of the database. The default view shows a table with all the entries. This table can be sorted by any of UniProtKB, PDB, EC and CATH identifiers. The doughnut and sunburst charts, together with the residues and cofactors selectors, allow the user to query the database in graphical way. For example, Figure 2 shows how the user can query for all the entries with detailed mechanism, CATH 1.10.-.- (Orthogonal Bundle) and with at least a cysteine and a histidine as catalytic residues.

The statistics pages can also be used for browsing. Most of the plots have interactive elements that the user can click to retrieve relevant M-CSA entries or affect what is being shown on other plots. For example, in the Residues page, the user can click on a residue and then on a role, to be shown a table with all the entries where that residue has the selected role.

The search page contains a single search box and several buttons that can be toggled to select which tables to

search in the database. Most fields and tables in the database can be searched in this manner. The search functionality can also be accessed by the search box situated in the main header, which is visible across the website. In that case, the default search fields will be used (enzyme name, EC, PDB, and UniProtKB identifiers). The results of the search are presented in a table under the search box.

We provide flat files with the data in M-CSA in the downloads page. Currently, the more complete file provides the reference PDB and UniProtKB sequences for each entry, as well as the EC number for the reference reaction and a list of reactants and products. It also contains a list of all the catalytic residues and cofactors together with their annotated roles for each entry. We also provide flat files in the same format as the old CSA website, for users that have incorporated that in their workflow.

Curator pages

The curator pages can be used to edit the database, and make the changes immediately visible across the website. Remote submission for new entries is now possible, through these pages. Additional automation and data checks (examples in SI), in relation to previous databases facilitate curation and minimize the introduction of errors. The documentation pages describe how to add new entries to the database using the curator tools.

We use the ChemAxon MarvinJS plugin (www.chemaxon.com) to draw the chemical mechanistic steps and save these in the ChemAxon Marvin Document (MRV) file type. Besides the molecules' 2D coordinates, the MRV file type also saves the position of electron flow arrows, which is essential for our purposes.

The process of entering a new entry into the database is as follows. Curators are assigned a password protected account which allows them to create and edit their own entries. After the edition of a new entry is finished, the curator flags that entry as complete. A member of staff, which is a curator with more privileges, will then review that entry and flag it so it can be shown on the public part of website. At this point, the ownership of the entry is changed and the original curator cannot edit this entry anymore. Only staff curators can edit or delete public M-CSA entries.

Technology

The new website has been implemented using the Django Web Framework v1.10 (djangoproject.com) together with a PostgreSQL database (www.postgresql.org). The database schema is shown in the SI. Python is the only language used on the server side. Custom JavaScript and CSS, apart from the JS plugins and the EBI template, are used sparingly on the client side. We use LiteMol (34), a JavaScript plugin initially developed for PDBe (35), to show PDB structures and highlight the catalytic residues and substrate. In the curator pages, we use the ChemAxon MarvinJS plugin v17.15.0 (www.chemaxon.com) to draw chemical schemes of reaction and chemical compounds that are not available in ChEBI, although curators are encouraged to add compounds directly to ChEBI, instead, if they have access. Reaction Decoder (36), the atom mapping tool that is part of
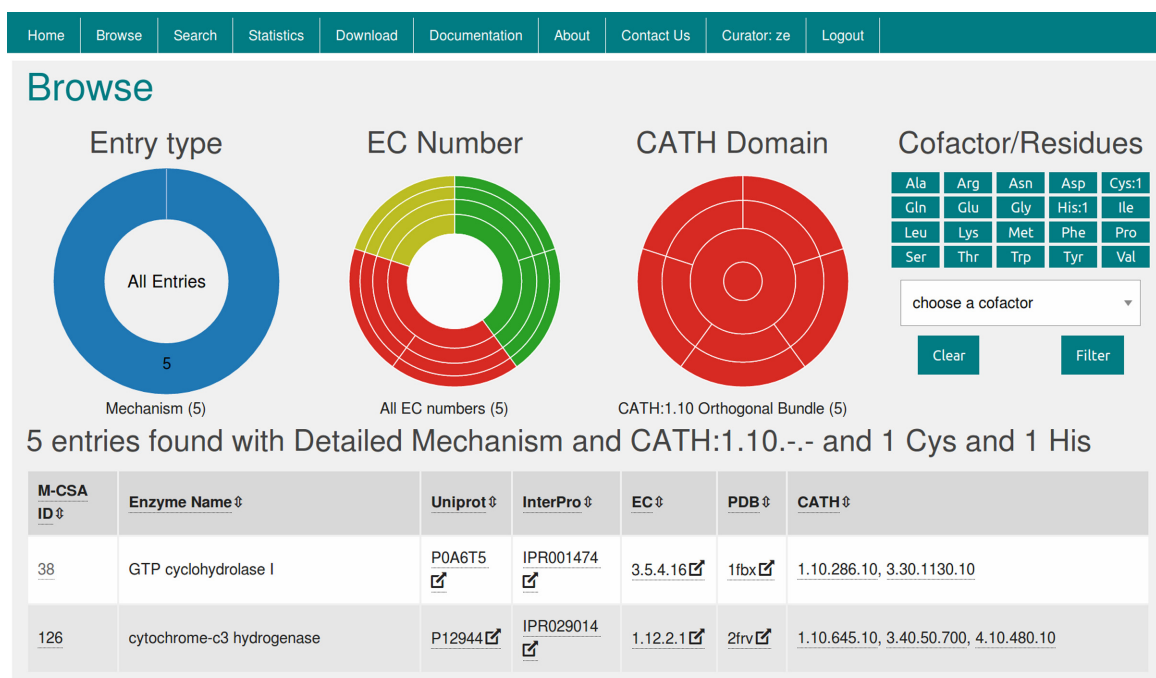
Figure 2. The browse page. Example of how to find all M-CSA entries with detailed mechanism, CATH:1.10.-.- and at least one His and one Cys in the catalytic site. The bottom three rows of the results table are omitted in the screenshot.

EC-Blast (37), is used to map RXN files, in order to calculate bond changes at the overall reaction level. All plots on the website, including the sunbursts plots in the browse page and all the plots in the statistics pages are rendered by NVD3 (nvd3.org), a D3.js (d3js.org) extension.

FUTURE DEVELOPMENTS

We hope to continue updating M-CSA in the future. In 2011, the last time these databases were updated, MACiE and the CSA covered over 70% of the available EC numbers with 3D-structures (a proxy for the number of cases with an available mechanism). Since 2011, the number of EC numbers with sequences and 3D structures has more than doubled (1196–2793). However, it cannot be assumed that because there is a structure, the mechanism is known. We do not currently have a good estimate of how many more mechanisms are available since 2011, although a conservative estimate would be double, in line with the increase of EC numbers with a 3D structure.

We are currently working on a 3D representation of the mechanism, for which we are building 3D models guided by the 2D schemes and some simple structural constraints. Ideally, we would like to use structures coming from computational chemistry studies, but these are not generally available. We encourage computational chemists to send us their models of enzyme mechanisms for inclusion in the database.

Since curation is now simpler and can be done through the website, we encourage enzymologists to submit new reaction mechanisms, especially for reactions not currently captured in M-CSA. In turn, and coupled with the rest of the analysis we want to perform, these users would be able to compare their proposed mechanism with others and see how it fits into the overall landscape of catalysis. For example, we could detect that a catalytic residue is doing a never-seen role, which could indicate that something is wrong with the mechanism or that the user has found some new chemistry.

We also want to use more protein structural tools to improve the database. At the moment, we are finding homologous proteins through sequence search, but structural alignments like those provided by CATH-tools (38), can uncover more distant relatives. Additionally, 3D template search using the catalytic residues (39,40) can help us find similar active sites in other structures, which may not be apparent from the one-dimensional sequence similarity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

REFERENCES

1. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

2. Burley,S.K., Berman,H.M., Kleywegt,G.J., Markley,J.L., Nakamura,H. and Velankar,S. (2017) Protein Data Bank (PDB): the single global macromolecular structure archive. In: *Protein Crystallography, Methods in Molecular Biology*. Humana Press, NY, pp. 627–641.

3. de Beer,T.A., Berka,K., Thornton,J.M. and Laskowski,R.A. (2014) PDBsum additions. *Nucleic Acids Res.*, **42**, D292–D296.

4. Kirby,A.J. (2001) The lysozyme mechanism sorted — after 50 years. *Nat. Struct. Mol. Biol.*, **8**, 737–739.

5. Sillitoe,I., Lewis,T.E., Cuff,A., Das,S., Ashford,P., Dawson,N.L., Furnham,N., Laskowski,R.A., Lee,D., Lees,J.G. *et al.* (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **43**, D376–D381.

6. Andreeva,A., Howorth,D., Chandonia,J.-M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.

7. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

8. Marsden,R.L., Ranea,J.A., Sillero,A., Redfern,O., Yeats,C., Maibaum,M., Lee,D., Addou,S., Reeves,G.A., Dallman,T.J. *et al.* (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Philos. Trans. R. Soc. B Biol. Sci.*, **361**, 425–440.

9. McDonald,A.G. and Tipton,K.F. (2014) Fifty-five years of enzyme classification: advances and difficulties. *FEBS J.*, **281**, 583–592.

10. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.

11. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.

12. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.

13. Morgat,A., Lombardot,T., Axelsen,K.B., Aimo,L., Niknejad,A., Hyka-Nouspikel,N., Coudert,E., Pozzato,M., Pagni,M., Moretti,S. *et al.* (2017) Updates in Rhea – an expert curated resource of biochemical reactions. *Nucleic Acids Res.*, **45**, D415–D418.

14. Hastings,J., De Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.

15. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

16. Placzek,S., Schomburg,I., Chang,A., Jeske,L., Ulbrich,M., Tillack,J. and Schomburg,D. (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.*, **45**, D380–D388.

17. Wittig,U., Kania,R., Golebiewski,M., Rey,M., Shi,L., Jong,L., Algaa,E., Weidemann,A., Sauer-Danzwith,H., Mir,S. *et al.* (2012) SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res.*, **40**, D790–D796.

18. Caspi,R., Billington,R., Ferrer,L., Foerster,H., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.

19. Noronha,A., Daníelsdóttir,A.D., Gawron,P., Jóhannsson,F., Jónsdóttir,S., Jarlsson,S., Gunnarsson,J.P., Brynjólfsson,S., Schneider,R., Thiele,I. *et al.* (2017) ReconMap: an interactive visualization of human metabolism. *Bioinformatics*, **33**, 605–607.

20. Chung,L.W., Sameera,W.M.C., Ramozzi,R., Page,A.J., Hatanaka,M., Petrova,G.P., Harris,T.V., Li,X., Ke,Z., Liu,F. *et al.* (2015) The ONIOM method and its applications. *Chem. Rev.*, **115**, 5678–5796.

21. Sousa,S.F., Ribeiro,A.J.M., Neves,R.P.P., Brás,N.F., Cerqueira,N.M.F.S.A., Fernandes,P.A. and Ramos,M.J. (2017) Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **7**, e1281.

22. Holliday,G.L., Bartlett,G.J., Almonacid,D.E., O'Boyle,N.M., Murray-Rust,P., Thornton,J.M. and Mitchell,J.B.O. (2005) MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**, 4315–4316.

23. Holliday,G.L., Almonacid,D.E., Bartlett,G.J., O'Boyle,N.M., Torrance,J.W., Murray-Rust,P., Mitchell,J.B.O. and Thornton,J.M. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res.*, **35**, D515–D520.

24. Holliday,G.L., Andreini,C., Fischer,J.D., Rahman,S.A., Almonacid,D.E., Williams,S.T. and Pearson,W.R. (2012) MACiE: exploring the diversity of biochemical reactions. *Nucleic Acids Res.*, **40**, D783–D789.

25. 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.

26. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

27. Furnham,N., Holliday,G.L., Beer,T.A.P. de, Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.

28. Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L. and Thornton,J.M. (2009) Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics*, **25**, 2088–2089.

29. Fischer,J.D., Holliday,G.L., Rahman,S.A. and Thornton,J.M. (2010) The structures and physicochemical properties of organic cofactors in biocatalysis. *J. Mol. Biol.*, **403**, 803–824.

30. Holliday,G.L., Mitchell,J.B.O. and Thornton,J.M. (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.*, **390**, 560–577.

31. Akiva,E., Brown,S., Almonacid,D.E., Barber,A.E., Custer,A.F., Hicks,M.A., Huang,C.C., Lauck,F., Mashiyama,S.T., Meng,E.C. *et al.* (2014) The Structure–Function Linkage Database. *Nucleic Acids Res.*, **42**, D521–D530.

32. Nagano,N., Nakayama,N., Ikeda,K., Fukuie,M., Yokota,K., Doi,T., Kato,T. and Tomii,K. (2015) EzCatDB: the enzyme reaction database, 2015 update. *Nucleic Acids Res.*, **43**, D453–D458.

33. Finn,R.D., Clements,J., Arndt,W., Miller,B.L., Wheeler,T.J., Schreiber,F., Bateman,A. and Eddy,S.R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.

34. Sehnal,D., Deshpande,M., Varekova,R.S., Mir,S., Berka,K., Midlik,A., Pravda,L., Velankar,S. and Koca,J. (2017) Real-time interactive visualisation of large macromolecular assemblies and molecular machines at atomic resolution. *Biophys. J.*, **112**, a178.

35. Velankar,S., van Ginkel,G., Alhroub,Y., Battle,G.M., Berrisford,J.M., Conroy,M.J., Dana,J.M., Gore,S.P., Gutmanas,A., Haslam,P. *et al.* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.

36. Rahman,S.A., Torrance,G., Baldacci,L., Cuesta,S.M., Fenninger,F., Gopal,N., Choudhary,S., May,J.W., Holliday,G.L., Steinbeck,C. *et al.* (2016) Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics*, **32**,2065–2066.

37. Rahman,S.A., Cuesta,S.M., Furnham,N., Holliday,G.L. and Thornton,J.M. (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods*, **11**, 171–174.

38. Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **208**, 1–22.

39. 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.

40. 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.