D776–D782 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 17 October 2011*
doi:10.1093/nar/gkr852

FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies

Nicholas Furnham[1,*], Ian Sillitoe[2], Gemma L. Holliday[1], Alison L. Cuff[2], Syed A. Rahman[1], Roman A. Laskowski[1], Christine A. Orengo[2] and Janet M. Thornton[1]

[1]European Bioinformatics Institute, European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and [2]Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

Received August 10, 2011; Accepted September 24, 2011

ABSTRACT

FunTree is a new resource that brings together sequence, structure, phylogenetic, chemical and mechanistic information for structurally defined enzyme superfamilies. Gathering together this range of data into a single resource allows the investigation of how novel enzyme functions have evolved within a structurally defined superfamily as well as providing a means to analyse trends across many superfamilies. This is done not only within the context of an enzyme's sequence and structure but also the relationships of their reactions. Developed in tandem with the CATH database, it currently comprises 276 superfamilies covering ~1800 (70%) of sequence assigned enzyme reactions. Central to the resource are phylogenetic trees generated from structurally informed multiple sequence alignments using both domain structural alignments supplemented with domain sequences and whole sequence alignments based on commonality of multi-domain architectures. These trees are decorated with functional annotations such as metabolite similarity as well as annotations from manually curated resources such the catalytic site atlas and MACiE for enzyme mechanisms. The resource is freely available through a web interface: www.ebi.ac.uk/thorton-srv/databases/FunTree.

INTRODUCTION

The majority of chemical reactions known to occur in biology appear to have been created by the modulation of an existing reaction through the evolution of the enzyme responsible. To begin to understand in detail how enzymes have evolved new functions requires the combination of protein 3D structure, sequence, phylogenetic, chemical and mechanistic data. This combination of information is crucial given the continual flood of data from structural genomic projects, since insights into the evolution of enzyme function provide one of the best routes for predicting functions of uncharacterized enzymes (1). Current resources either provide details on just a subsection of this combination of data or advance extensive detailed analysis on a relatively small number of enzyme superfamilies (2–5).

In order to address this challenge, we have developed a resource that brings together manually curated data from the CATH (6) classification of domains from protein structures, sequences from UniProtKB (7) and CATH-Gene3D (8), as well as functional and chemical information from a variety of sources including the manually curated MACiE (9) and Catalytic Site Atlas (CSA) (10) databases. The data are presented through phylogenetic analysis and is combined with the examination of relationships between metabolites obtained by exploiting tools for comparing small molecules.

THE FUNTREE PIPELINE

Protein domains, structurally defined by CATH, that are identified as having an enzyme function are selected using the MACiE database. This identifies, through careful manual annotation, the location of the residues involved in the enzyme mechanism. FunTree processes the superfamilies of domains that have the active site residues located within the domain. The workflow by which data are collected, processed and presented is shown in Figure 1. Recent studies have highlighted the problems of relying on

*To whom correspondence should be addressed. Tel: +44 1223 464631; Fax: +44 (0)1223 494 496; Email: nickf@ebi.ac.uk

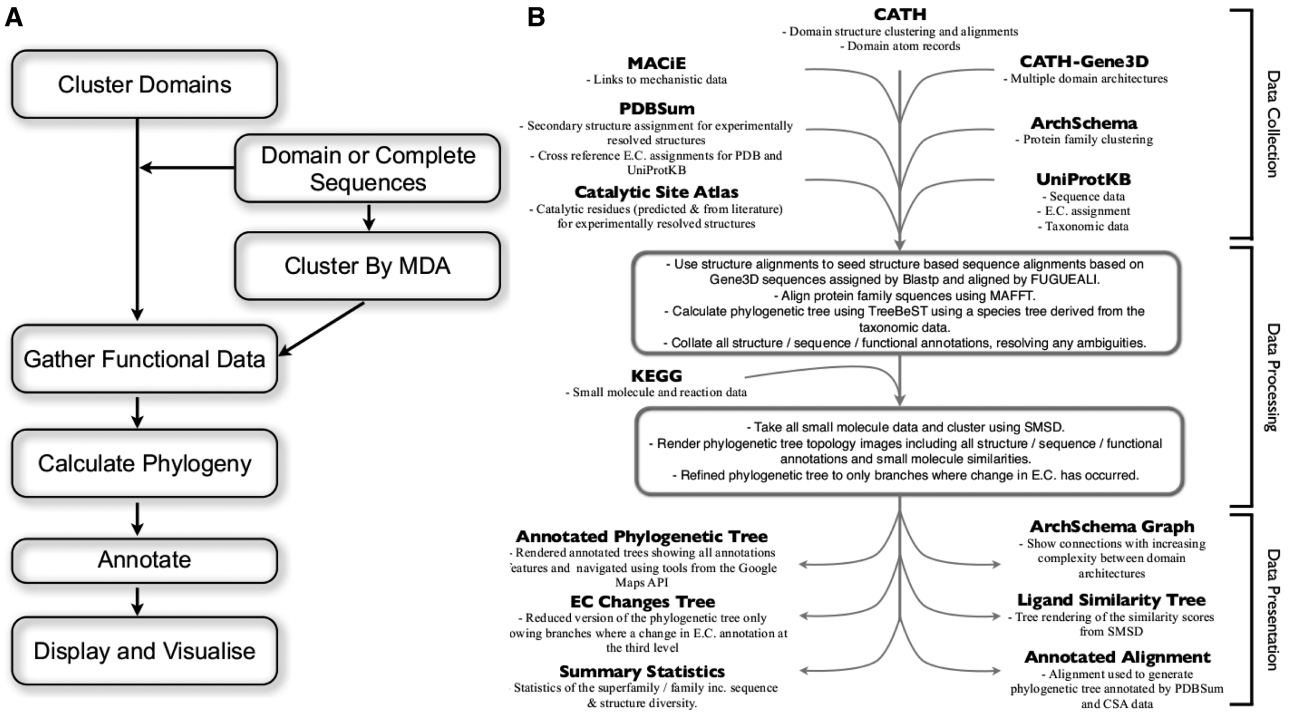

Figure 1. The FunTree pipeline. (**A**) An overview of the workflow for collecting and processing sequence, structure and functional information in FunTree. (**B**) A detailed schematic representation of the various steps in data collection, processing and visualization in FunTree.

functional annotations, especially those generated by automated methods (11,12), thus FunTree only uses sequences with functional annotations from the reviewed section of UniProtKB or where a functional annotation is made on deposition of structural data.

Changes to an enzyme's function can arise from modifications of a single domain or from a change to the combination of domains making up in the complete protein sequence. To capture both factors, we generate two types of cluster based on either the superfamily domain or the complete protein sequence:

Structurally similar groups

Protein domain superfamilies can show considerable sequence and structural diversity outside of the common structural core. This makes it difficult to effectively superimpose all domains within some superfamilies. Thus, we grouped non-redundant domains with <35% sequence identity to all other members of the cluster, whose structures could be aligned by CORA (13) and superimposed using the McLachlan algorithm (14) as implemented in the program Profit (Martin,A.C.R. and Porter,C.T.; http://www.bioinf.org.uk/software/profit/) with a root mean squared deviation of <9 Å, to generate multiple structure alignments. These clusters are described as structurally similar groups (SSG), and are then subsequently populated with sequence relatives. These are collected from CATH-Gene3D (a resource which contains sequences for all known and predicted domains in 1867 genomes) and are assigned to one of the SSGs using BLASTp (15) to scan against the sequences of known structural domains. BLASTp is used because it is very fast and can scan through vast numbers of sequences in CATH-Gene3D. Once assigned the sequence is aligned to the profile of the structurally informed sequence alignment using FUGUALI [part of the FUGUE (16) software]. The resulting robust structurally informed sequence alignments are used to undertake the phylogenetic analysis.

MDAs

A protein can be made up of one or more domains that may be contributing to overall function (17). For each sequencein FunTree, the multi-domain architecture (MDA) is assigned by considering the order of known or predicted structural domains mapped to the sequence. Domain structure assignments are taken from CATH-Gene3D by initially scanning the sequence against Markov models built from CATH domains. Regions of sequences that are unassigned and are large enough to be considered as a domain are checked against the PFam database (18) and if a non-overlapping PFam domain is found, it is included in the MDA, Subsequently, we group together proteins within a superfamily that share the same domains in the same order along the sequence: i.e. the same MDA. Grouping of MDAs is carried out by ArchSchema (19), which also visualizes the relationships between MDAs as a directed graph. For each superfamily, entire protein sequences which share the same MDA are aligned using MAFFT (20). Alignments generated are used to perform the phylogenetic analysis of the MDA clusters.

Phylogenetic analysis. We perform phylogenetic analysis on both the SSG and the MDA alignments. However, some enzyme superfamilies can be very large with

thousands to tens of thousands of sequences, which makes both aligning all the sequences and conducting the phylogenetic analysis difficult. In order to overcome this, sequences are first filtered by taxonomic lineage and uniqueness of function. This removes sequences sharing the same genus level and having the same function (i.e. E.C. number) and, for simplicity taking the first occurrence as the single representative. If there are still many thousands of sequences left, a stricter filter is applied at the kingdom level. In both cases, however, if a sequence has a function annotated that has not been previously seen for that taxonomic rank, then the sequence is included.

For both the SSG and MDA alignments, phylogenetic trees are generated using the TreeBest software [as described in the methods for compiling the TreeFam database (21)]. The method uses species relationships to guide the tree building, thus a taxonomic tree for those sequences in the alignment is generated using the species relationships as defined by the NCBI taxonomic database (22).

By systematically traversing the tree, it can be simplified by collapsing nodes whose branches have a commonality in their annotation. For the purposes of this study, we define commonality at the subsubclass (third level) of the four-level E.C. classification, which broadly can act as a proxy for a change in general chemistry. Thus, nodes in the pruned tree correspond to different reaction chemistries. This collapsed version of the tree is also generated and presented.

Metabolite analysis. Functional data, in the form of E.C. classifications (23), are collected from either annotations from the reviewed section of UniProtKB or if present from the annotations made in the deposition of the protein structure. These identifiers are then used, *via* KEGG (24), to collect the reaction performed and the small molecules used by the enzyme. All the small molecules within the superfamily and within each SSG/MDA groups are compared with each other using the Small Molecule Subgraph Detector (SMSD) toolkit (25) to generate an all-by-all comparison matrix for each case. The metabolites are clustered using PVCLUST (26), implemented in the R statistical package and the results are rendered as a similarity tree using software developed in-house.

Data presentation and navigation

FunTree data are presented through a publicly accessible website—http://www.ebi.ac.uk/thornton-srv/databases/FunTree, for the superfamily, structurally similar groups (SSGs) and MDA groups. The website can be searched by superfamily or small molecule names and synonyms as well as specific superfamily, sequence, structure, E.C. or small molecule identifiers. In addition, the data can be browsed by superfamily, E.C. code, structure or metabolites. SSG and MDA groups provide different views of the superfamily data—i.e. in terms of structural similarity and similarity of domain composition, respectively. An SSG may contain domain relatives in different MDAs, and conversely a given MDA may be present in one or more SSGs. To navigate between the various groups and show how they relate, FunTree displays a simple bifurcated graph with the two branches representing the division between SSGs and MDAs. Clicking on a particular SSG branch highlights the MDA branches that contain members of the SSG; conversely, clicking on an MDA branch highlights the SSG branches that have members belonging to this MDA (Figure 2).

On the top level page describing the superfamily, the following data are presented: a summary of statistics such as the sequence diversity as measured by ScoreCons (27) and the average SSAP (28) scores of the domain structures; a similarity tree of the small molecules; an ArchSchema graph of the MDA and a representation of the E.C. hierarchy (an 'E.C. wheel') showing which E.C. numbers are present in the superfamily. At the SSG or MDA level, the page shows the 'grouping-specific' general statistics, a similarity tree of the small molecules, an E.C wheel, a phylogenetic tree, the annotated alignment used to build the phylogenetic tree and a collapsed version the phylogenetic tree. The collapsed tree shows nodes with changes at the third level of the E.C. classification, which, as mentioned above, broadly act as a proxy for a change in general chemistry. Each leaf of the tree also lists the full E.C. numbers the collapsed node represents.

As the phylogenetic trees can be large and contain many annotations, the tree is rendered as a series of images at various zoom levels that can be navigated using the GoogleMaps API. The GMMap Image Cutter (R. Milton 2008, http://www.casa.ucl.ac.uk/software/googlemapimagecutter.asp) is used to generate the image tiles that are used by the GoogleMaps API to display the tree. Embedded in a web page, the tree can be navigated using the tools familiar to anyone who has used Google Maps navigation tools. Thus, the tree can be scrutinized within a web page by being dragged, panned and zoomed using the navigation tools or click-and-drag mouse motions, as well as allowing for overlays to show hyperlinks and additional notes when a mouse hovers over a specific part of the map. Also provided is an in-page thumbnail overview, which tracks the movements in the main image. This aids navigation when the image is zoomed in.

Each leaf of the tree is annotated with links to sequence, structure and mechanism data if known. In addition, the E.C. numbers are annotated and coloured according to their similarity at the third level of the E.C. hierarchy. The small molecules involved in each E.C. reaction are represented by coloured boxes, where the colour shows the similarity relationship based on the SMSD scores. The more similar the molecules, the closer their colours are according to the colours of the rainbow. The complete reaction is also annotated as an image, appearing when the mouse is hovered over the annotation. Finally, the domain architecture of the complete sequence is depicted as a series of coloured bars, with each unique domain in the MDA given a unique colour. At the nodes in the tree the bootstrap values are displayed and a link to a JMol (29) view of the superimposition of any structures present in the clade rooted at the node. The structures are shown

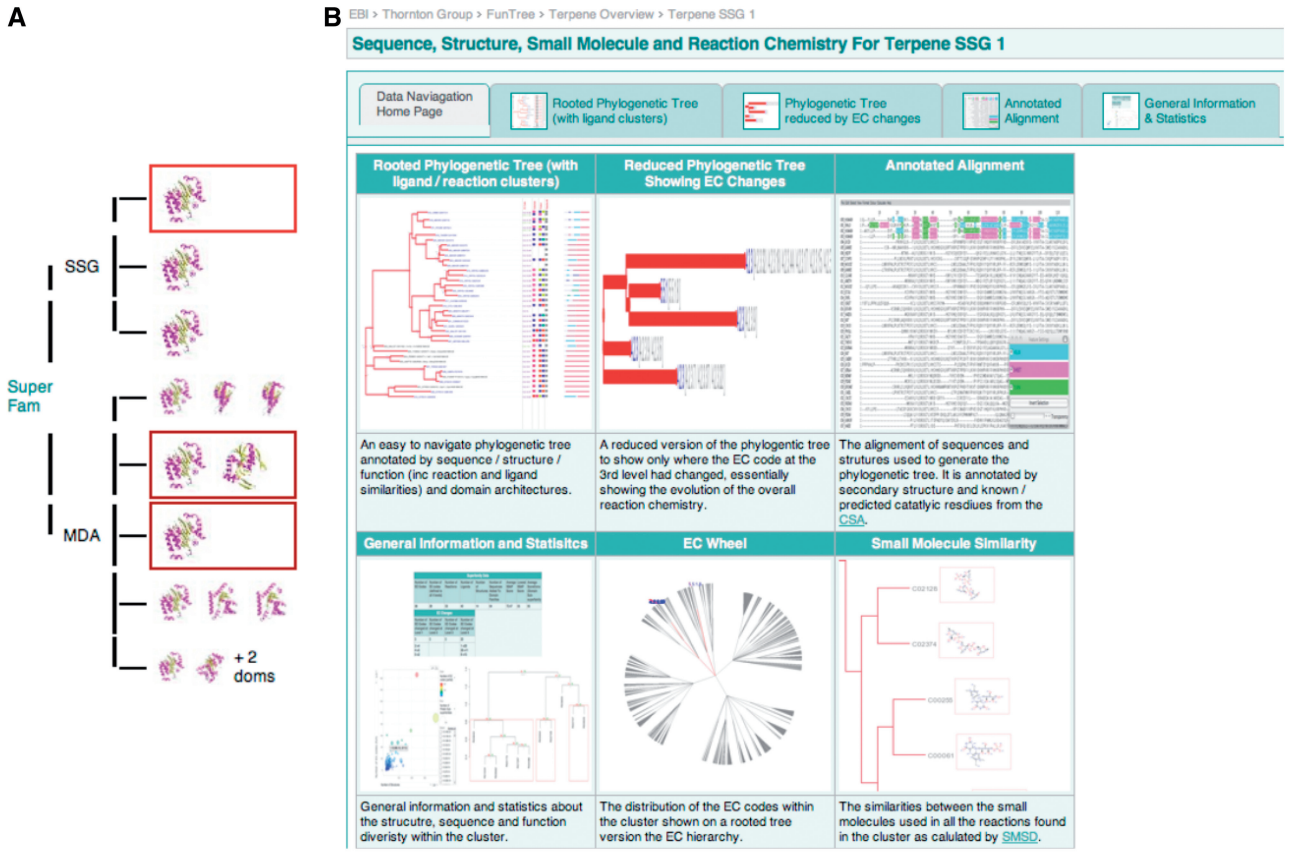

Figure 2. Navigating FunTree web pages. (**A**) A screenshot of the bifurcated tree developed to aid navigation between the two groupings within each structurally defined superfamily. At the root is a link to the superfamily level of data, while each branch shows the SSG/MDA groups. For each group, the domain composition is shown as a thumbnail of the domain structure. As some MDA groups are made up of a large number of domains only the first three domains are displayed as thumbnails, with the number of extra domains given numerically. When navigating data within a group, that group is highlighted in the navigation tree as red, with the corresponding SSG/MDA found within the group highlighted in dark red. (**B**) Shows a screenshot of the top data navigation page for the SSG/MDA group and is similar to the page displayed for the superfamily. Six representations of the data are available to view, each can be accessed either via the thumbnail on the data navigation page or via the tabs at the top of the page. A breadcrumb trail is also provided.

as protein cartoons, coloured based on the colours assigned to the E.C. code in the tree and the active site residues are highlighted as space filled atoms coloured red. The active site information is derived from the CSA (Figure 3).

In the collapsed phylogenetic tree, the third level E.C. code representing the branch is highlighted and all the full E.C. numbers are listed. In addition, the sequence alignment the phylogenetic tree is based on is shown in Jalview (30). If any sequence has a known structure, the secondary structure assignments provided by PDBsum (31) are annotated along with catalytic site residues as defined by the CSA. Distinction is made between the catalytic residues identified in the CSA by curation from the literature, and those inferred on the basis of sequence comparison.

All types of tree images are processed and rendered, along with data collection, processing and integration, using software developed in-house for FunTree. All software is written in Python making particular use of the BioPython (32), Pycluster and PIL libraries. Associated data relating to the trees and superfamilies are stored in a MySQL database.

Overview of 276 superfamilies

The FunTree pipeline has been applied to 276 CATH superfamilies. These superfamilies represent over 2 million sequences from UniProtKB and nearly 3 million domain sequences (32% of CATH-Gene3D sequences) as defined by CATH-Gene3D. All four CATH classes and 60% of all CATH architectures are present. Though these 276 superfamilies represent only 11% of CATH homologous superfamilies they include some of the largest superfamilies, so that 48% of structurally characterized domains classified by CATH are present. In total, FunTree captures 2167 E.C. numbers (71% of E.C. numbers assigned to sequences) of which 1817 are fully classified and 1360 represent chemically balanced reactions with 1589 unique metabolites.

The largest number of SSGs and MDAs are found in the P-loop containing nucleotide triphosphate hydrolases with 27 SSGs and 687 unique MDAs. The top 10% of superfamilies in FunTree ordered by either the number of SSGs or MDAs account for ~50% of sequences of all sequences represented by the 276 superfamilies. This top 10% have with a mean number SSGs of 5 and

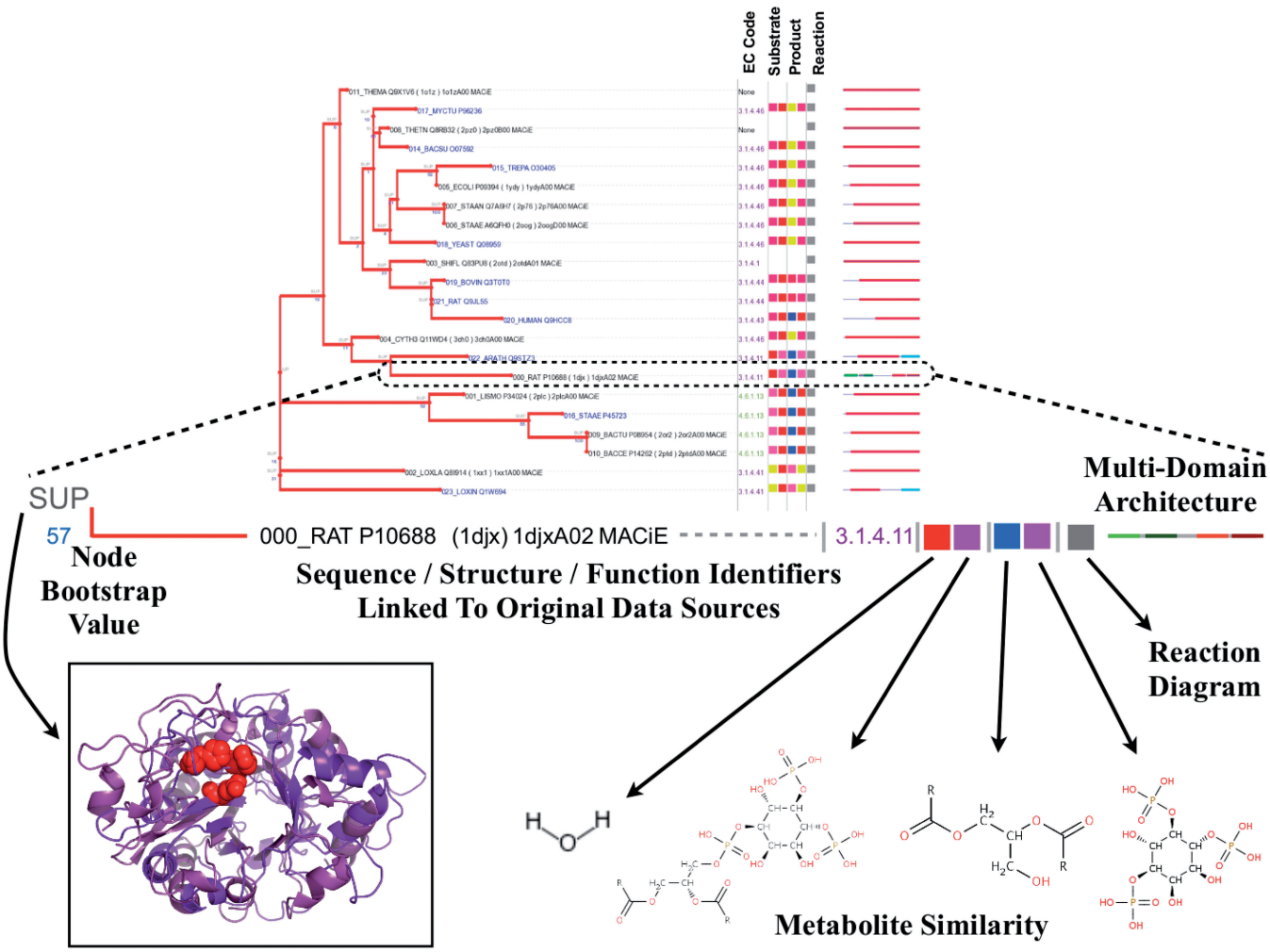

Figure 3. An example of the phylogenetic tree as visualized in FunTree. A single branch of the tree is highlighted to show the range of information held for each branch. Branches with annotation in black have structural information associated with it, while those in blue have just sequence information. The tree image is imbedded in the web page using the GoogleMaps API to navigate around the tree. Annotation on each branch is hyperlinked to the underlying data source. Relationships between metabolite data is shown as coloured boxes, with the colouring based on a rainbow scale with similar metabolites having similar colours. At each node in the tree, a bootstrap value is provided in blue as well as a link to a JalView window showing the superimposition of any structures in the clade rooted at the node. The structures are coloured by the same colours given to the E.C. numbers in the tree and well as having any catalytic residue information for the CSA highlighted as red filled spheres.

MDAs of 113. The rest of the superfamilies only have an average of one structurally similar group and only 7 different domain architectures (Figure 4). This accords with previous observations (6).

The purpose of this resource is to explore the evolution of functional catalytic diversity. The distribution of the number of associated functions for each superfamily, as defined by the E.C., shows that some exceptional superfamilies have many different enzyme functions [the NAD(P)-binding Rossmann-like domain has the most with 223 unique E.C. numbers], while 49 others have only one. The top 10% of superfamilies by number of sequences in FunTree account for 849 unique functions as defined by E.C. number, with an average of 35 E.C. numbers per superfamily. The rest have on average only 6 E.C. numbers per superfamily.

Of the 276 superfamilies, about two-thirds (177) show some or all of their functional diversity at the fourth serial number level of the E.C. classification, which indicate changes in substrate specificity. The promiscuity of a superfamily can be gauged by analysing the diversity shown by multiple differences in the serial number (note that in some reactions, as defined by the E.C. number, the substrates include an 'R' group which indicates a variable moiety and provides another level of substrate diversity). Of these 177 superfamilies, 150 have more than 50% of their E.C. diversity coming from changes in this level, the rest coming from changes at the higher levels. However, nearly an equal number of superfamilies (176 superfamilies) include at least one member where the diversity is occurring at the third level E.C. or above, which can act as proxy for a change in chemistry. Thirty-nine of these general chemistry diverse superfamilies show that all the diversity is occurring through changes at the third level or above and none are occurring at the serial number (fourth level) of the E.C. In our data set, there are 67 superfamilies

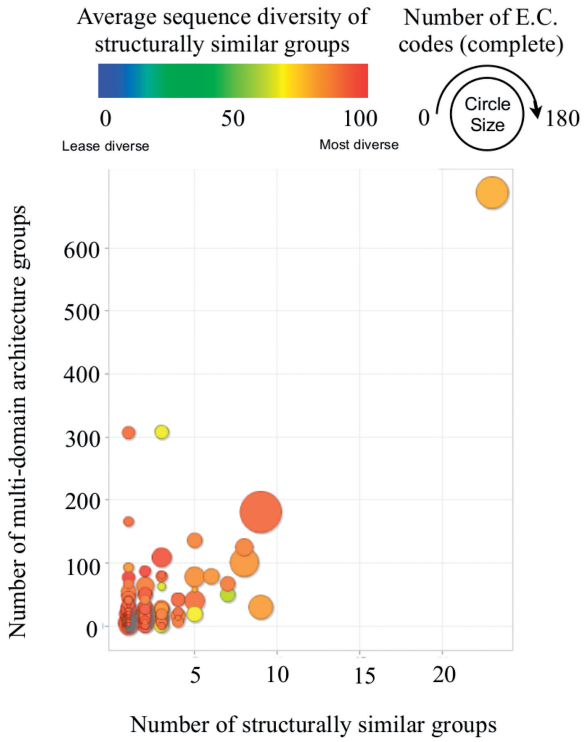

Figure 4. Overview of superfamily data. For each of the 276 superfamilies catalogued in FunTree, the number of structurally similar groups and the number of multi-domain architectures are shown, with most superfamilies having a few SSGs and MDAs while some exceptional families have many SSGs and MDAs. The sequence diversity of structurally similar groups is highly diverse, as indicated by the circle colour intensity. Most superfamilies are also functionally diverse, as indicated by the size of the circle.

(~25%), where the single domain carries out 80% or more of the enzyme functions found in the superfamily.

CONCLUSIONS

It is difficult to combine structural, sequence, phylogenetic, functional and chemical data together effectively for a large number of superfamilies, thus we had to develop a complex pipeline. Bringing together this range of data into a single resource allows the investigation of the evolution of novel enzyme functions within structurally defined superfamilies. It has permitted not only the exploration of specific enzyme superfamilies but provides a means to analyse trends across many superfamilies. Exploring individual families has reinforced the observation that enzyme evolution is incredibly complex, with many different routes being taken to obtain different reactions, mechanisms and specificities within a superfamily.

In practice, the FunTree resource allows a number of questions to be addressed. For example, for a given superfamily the catalytic diversity (by E.C. number) can be gauged as well as the range and diversity of known substrates and products. Furthermore, FunTree can provide the evolutionary progression in terms of function of the superfamily. In the future, we envisage that it would be possible to place new sequences into FunTree, allowing a user to see how it positions in 'functional' space. FunTree also allows other more general questions to be addressed for all superfamilies, such as which E.C. numbers are 'related' in terms of evolution and what are the common structural paradigms of enzyme evolution that underlie functional evolution.

We will continue to update the resource in parallel with the CATH/CATH-Gene3D update process to include new sequences, structures and functions as they become available. As new tools become available to analyse similarities between enzyme reactions based on metabolite substructure similarity and bond order changes, these will be introduced as another similarity measure appended to the branch annotation. Using such tools should remove some of the problems in comparing E.C. codes.

By beginning to gather, catalogue and classify the emergence of catalytic reactions, users can analyse shifts in functionality across and within enzyme superfamilies and may help in designing new enzymes as well as aid in function prediction.

FUNDING

Wellcome Trust (Grant No. 081989/Z/07/A to N.F. and I.S.); Biotechnology and Biological Sciences Research Council to ALC; European Molecular Biology Laboratory (to G.L.H. and S.A.R.); and in part by US Department of Energy Contract (DE-AC02-06CH11357 to R.A.L.) as part of the Midwest Center for Structural Genomics. Funding for open access charge: EMBL and Wellcome Trust Grant.

Conflict of interest statement. None declared.

REFERENCES

1. Rentzsch,R. and Orengo,C.A. (2009) Protein function prediction–the power of multiplicity. *Trends Biotechnol.*, **27**, 210–219.
2. Bartlett,G.J., Borkakoti,N. and Thornton,J.M. (2003) Catalysing new reactions during evolution: economy of residues and mechanism. *J. Mol. Biol.*, **331**, 829–860.
3. Brown,S.D., Gerlt,J.A., Seffernick,J.L. and Babbitt,P.C. (2006) A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.*, **7**, R8.
4. Glasner,M.E., Gerlt,J.A. and Babbitt,P.C. (2006) Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.*, **10**, 492–497.
5. Todd,A.E., Orengo,C.A. and Thornton,J.M. (1999) Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.*, **3**, 548–556.
6. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
7. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
8. Lees,J., Yeats,C., Redfern,O., Clegg,A. and Orengo,C. Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res.*, **38**, D296–D300.
9. Holliday,G.L., Almonacid,D.E., Bartlett,G.J., O'Boyle,N.M., Torrance,J.W., Murray-Rust,P., Mitchell,J.B. and Thornton,J.M. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res.*, **35**, D515–D520.

10. 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.
11. Furnham,N., Garavelli,J.S., Apweiler,R. and Thornton,J.M. (2009) Missing in action: enzyme functional annotations in biological databases. *Nat. Chem. Biol.*, **5**, 521–525.
12. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
13. Orengo,C.A. (1999) CORA–topological fingerprints for protein structural families. *Protein Sci.*, **8**, 699–715.
14. McLachlan,A. (1982) Rapid comparison of protein structures. *Acta Crystallog. Sect. A*, **38**, 871–873.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
17. Bashton,M. and Chothia,C. (2007) The generation of new protein functions by the combination of domains. *Structure*, **15**, 85–99.
18. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
19. Tamuri,A.U. and Laskowski,R.A. (2010) ArchSchema: a tool for interactive graphing of related Pfam domain architectures. *Bioinformatics*, **26**, 1260–1261.
20. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
21. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
22. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
23. IUBMB. (1992) IUBMB (1992) Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.
24. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
25. Rahman,S., Bashton,M., Holliday,G., Schrader,R. and Thornton,J. (2009) Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminformat.*, **1**, 12.
26. Suzuki,R. and Shimodaira,H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.
27. Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.
28. Orengo,C.A., Taylor,W.R. and Russell,F.D. (1996) *Methods in Enzymology*, Vol. 266. Academic Press, San Diego, USA, pp. 617–635.
29. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/ (5 October 2011, date last accessed).
30. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
31. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.
32. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.