Exploring enzyme evolution from changes in sequence, structure, and function

Jonathan D Tyzack[1], Nicholas Furnham[2], Ian Sillitoe[3], Christine M Orengo[3], Janet M Thornton[1]

[1]EMBL-EBI, Wellcome Genome Campus, CB10 1SD, United Kingdom

[2]London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

[3]UInstitute of Structural and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Running Head

FunTree: exploring enzyme evolution

Corresponding authors:

Jonathan D. Tyzack	tyzack@ebi.ac.uk
Nicholas Furnham	Nick.Furnham@lshtm.ac.uk

Abstract

The goal of our research is to increase our understanding of how biology works at the molecular level, with a particular focus on how enzymes evolve their functions through adaptations to generate new specificities and mechanisms. FunTree [1] is a resource that brings together sequence, structure, phylogenetic, chemical and mechanistic information for 2,340 CATH superfamilies [2] (which all contain at least one enzyme) to allow evolution to be investigated within a structurally defined superfamily.

We will give an overview of FunTree's use of sequence and structural alignments to cluster proteins within a superfamily into structurally similar groups (SSGs) and generate phylogenetic trees augmented by ancestral character estimations (ACE). This core information is supplemented with new measures of functional similarity [3] to compare enzyme reactions based on overall bond changes, reaction centres (the local environment atoms involved in the reaction) and the structural similarities of the metabolites involved in the reaction. These trees are also decorated with taxonomic, Enzyme Commission (EC) code and GO annotations, forming the basis of a comprehensive web interface that can be found at http://www.funtree.info. In this chapter, we will discuss the various analyses and supporting computational tools in more detail, describing the steps required to extract information.

Key Words

FunTree, enzyme evolution, CATH, EC--Blast, phylogenetic tree

Introduction

FunTree is a resource for exploring the evolution of protein function through relationships in sequence, structure, phylogeny and function. It catalogues 2,340 CATH superfamilies with over 400,000 representative sequences (selected to cover taxonomic lineage and function), over 70,000 structural domains, and 2,358 EC (Enzyme Commission) codes.

FunTree can be used to place structures and sequences in the context of their structural and functional evolution, allowing the investigation of how novel enzyme functions have evolved within a structurally similar group (SSG). This can also be helpful to identify new but currently unobserved reactions and substrates for known enzymes, as well as possible reactions for enzyme sequences/structures of unknown function.

Often CATH superfamilies can be structurally highly diverse hindering the confident atomic superimposition of all structures in the superfamily. A key step in the generation of FunTree is the sub--clustering of each superfamily into distinct SSGs, where all inter structure SIMAX scores are less than 9 Angstroms (where SIMAX is the RMSD between two domains multiplied by the number of residues in the larger domain divided by the number of aligned residues).

The core output of FunTree is a phylogenetic tree for each SSG (discussed in 2.1), calculated from structure--guided sequence alignments using a novel agglomerative clustering technique. The resulting alignments are provided to TreeBest [4], along with a species tree derived from species relationships in the NCBI Taxonomic definitions, to generate a maximum likelihood based phylogenetic tree.

The phylogenetic trees are decorated with information such as EC code, GO annotations, and multi domain architecture (MDA), and augmented with various ancillary analyses describing the diversity in areas such as enzyme chemistry and taxa distribution. These will all be described in more detail in the Methods section, however it is important to note that some annotations such as GO and EC are assigned to entire gene products rather than the individual structural domains included in the SSGs. Most functions can be ascribed to a single domain but many are a product of domain combinations or multiple gene products. Thus, as FunTree is a domain centric resource, some annotations might be relevant at the protein rather than domain level. The FunTree pipeline describing the various steps in collecting, processing and presenting the data is shown in Fig. 1.

The trees generated in FunTree can become difficult to navigate due to their size and mixed media content. To facilitate easy navigation a web interface has been constructed using the javascript D3 libraries [5] to provide intuitive and user--friendly functionality (e.g. using the mouse wheel to zoom and dragging images to pan) and interaction with the trees (e.g. collapsing and expanding nodes by clicking).

Methods

FunTree can be browsed by CATH superfamily or searched by CATH superfamily, UniProtKB accession, EC code or by entering a text string for a fuzzy search. Overview statistics and high--level results are produced at the CATH superfamily level, with the phylogenetic trees and lower level results produced at the structurally similar group (SSG) level. These are discussed in more detail in the remainder of this section.

1. CATH superfamily results gateway

This page is the gateway for results at the CATH superfamily level for the selected domain (Fig. 2), where each thumbnail provides a link to a detailed analysis of the selected results. The SSGs within the superfamily are shown in Clusters with a link to lower level results for that SSG.

1.1. Domain architectures

This page shows an interactive force directed graph generated by ArchSchema [6] of the multi domain architectures (MDAs) associated with the current search, with the current domain shown at the centre connected to increasingly more complicated architectures.

a. The coloured graph nodes represent the different MDAs and can be dragged to reorganize the graph. Hovering over the graph nodes shows the following information for that MDA:

    I. Number of sequences

    II. Number of structures

    III. List of EC codes (annotated by UniProtKB [7])

    IV. List of structures

b. The colored domain bars show the domain composition, where hovering over the bar reveals the domain name and clicking opens the webpage for that CATH superfamily.

1.2. Overview Stats

This page contains a dynamic, interactive plot allowing various properties of CATH superfamilies to be plotted on 2 axes (Fig. 3). The different properties that can be plotted on either a linear or log scale and also used to scale and colour the data points are:

a. Alphabetical order (x--axis only)

b. Average conservation score for each position in the alignment (ScoreCons) [8] for SSGs

c. Number of multi--domain architectures (MDAs)

d. Number of full Enzyme Commission (EC) codes

e. Number of partial EC codes

f. Number of sequences

g. Number of structures

h. Number of structurally similar groups (SSGs)

1.3. EC Wheel

This page shows the EC hierarchy as an unrooted tree with EC codes within the superfamily labelled outside the wheel.

a. Nodes/leaves for class, sub--class, sub--subclass and numerical identifier are highlighted for the enzymes found in the superfamily.

1.4. EC--Blast

This page shows the EC classification rendered as a circular rooted tree

a. Leaves represent EC code and are coloured by primary EC class. EC codes that are found in the superfamily are pushed out of the circle and are coloured blue.

b. Links between EC codes found in the superfamily and their 10 most similar functions as calculated by EC--Blast are highlighted in blue, tracing the path through the tree between them.

c. Hovering over an EC code highlights in red connections to the top 10 most similar reactions, which are also listed on the right of the page.

1.5. CATH

This is a link to the CATH page [2] for that superfamily containing further information on structure and function.

2. Structurally Similar Group (SSG) results gateway

This page is the gateway for results at the SSG level for the selected domain (Fig. 4). Each thumbnail provides a link to a detailed analysis of the selected results.

2.1. FunTree: Rooted Phylogenetic Tree

This page contains a rooted phylogenetic tree for the SSG selected, with annotations and links embedded in the nodes and leaves (Fig. 5).

a. Navigation is implemented using the mouse wheel to zoom; dragging the image to pan; clicking on a node to collapse/expand that node; clicking on text for links to data sources; and hovering over text/images for more information.

b. At each node to the tree a confidence score can be found. This is the confidence bootstrap score provided by TreeBest for bifurcation at the node. Please note that as these trees are automatically generated some of the bifurcations might have low confidence scores and should be considered with caution.

c. The annotations at the end of each leaf are as follows:

   I. The first number/text section is the node name (internal to FunTree) made up of a reference and the taxonomic code

   II. If the leaf represents an enzyme, the next three circles show the similarity between reactions in the EC code on a bond change, reaction centre and sub--structure basis respectively. Colouring is based on the degree of similarity as calculated by EC--Blast.

   III. Primary EC code, containing a hyperlink to the IntEnz database

   IV. UniProtKB identifier, containing a link to the UniProtKB record

   V. If the sequence represents a known structural superfamily then the PDB (linked to PDBe entry [9]) and CATH domain (linked to the CATH superfamily page) are shown.

VI. The MDA of the protein at each leaf is depicted showing the domains as uniquely coloured bars along a line, the position and length of which are proportional to the total sequence. Hovering over each bar shows the CATH superfamily and clicking navigates to the CATH superfamily page.

2.2. Taxa Distribution

Shows the distribution of taxonomic classes within the SSG tree.

a. Hovering over the band reveals the taxonomic lineage (shown top left) as well as the percentage of sequences in the tree that belong to that group

2.3. Ancestral Character Estimation (ACE) Tree:

This is a circular representation of the phylogenetic tree based on SSG alignments showing likelihoods of functions at ancestral nodes.

a. Hovering over a node shows the EC code/function with the maximum likelihood for that node.
b. Hovering over leaves of the tree shows the contribution to function annotation from each internal node in the lineage.

2.4. Reaction Clustering

This page shows a tree representing the similarities between reactions based on bond changes calculated by EC--Blast, where the clustering is made using the PVClust [10] methods as implemented in R (Fig. 6).

a. The tree can be zoomed using the mouse wheel or moved/panned by dragging the image.

b. Each leaf shows a schematic of the reaction with colour coding highlighting the atoms that are involved in the reaction.

2.5. GO Clustering

A tree representing the similarities between GO annotations using a semantic similarity score.

a. The tree can be zoomed using the mouse wheel or moved/panned by dragging the image.

2.6. Ligand Clustering

This page shows a similarity tree of all the small molecules found in all the reactions in the SSG. The similarities are calculated using SMSD [11] and the clustering is made using the PVClust methods as implemented in R.

a. The tree can be zoomed using the mouse wheel or moved/panned by dragging the image.
b. By hovering over the leaves of the tree the reaction is displayed, and the other ligands in the reaction are highlighted.

2.7. EC Wheel

The functionality is as described in 1.3 but for data at the SSG level.

2.8. Annotated Alignment

This page shows the multiple sequence alignment generated with the BioJS [12] module (Fig. 7) that was used to build the phylogenetic tree. The sequences in the alignment are annotated by secondary structure where available and catalytic residues as catalogued the M--CSA [13] (bright red if from the curated M--CSA, light red if from the predicted M--CSA).

a. The alignments can be scrolled vertically (to show more sequences) and horizontally (to show different parts of the sequence)
b. The sequences can be selected, ordered and filtered by the various data fields including sequence identity
c. Other formatting options include editing the colour scheme and hiding/showing visual elements such as labels and headers
d. There is also functionality to import and export data for external analysis

2.9. Overview Stats

The functionality is as described in 1.2 but for data at the SSG level.

3. Examples of the Application of FunTree

As FunTree holds data across many domain superfamilies, it is possible to use FunTree to make large--scale general observations about how enzymes have evolved their function [14]. These observations can be made at the domain and residue level, exploring how function is modulated via the addition/removal of domains within a multi--domain architecture or adaptations of the catalytic/binding pocket. This allows analyses to be prepared comparing the number and types of evolutionary steps observed within domain superfamilies [15].

Furthermore, detailed analysis within a single superfamily or for a specific enzyme can be undertaken. An example of this is the evolution of functionality within the Phosphatidylinositol--Phosphodiesterase Superfamily (CATH 3.20.20.190), which is summarised briefly here but discussed more comprehensively in reference [16]. This superfamily shows relatively high structural conservation, presenting just one structurally similar group, but the phylogenetic tree generated within FunTree reveals three clades (see Fig. 8). Clades C1 and C3 show hydrolase activity (EC: 3.1.4) using a metal co--factor, whereas Clade 2 exhibits a transition to lyase activity (EC: 4.6.1). The structure--informed sequence alignment reveals that none of the three metal chelating residues are conserved in Clade 2, so that a metal is no longer bound, resulting in the cyclic intermediate leaving the active site prior to hydrolysis and giving the change from hydrolase to lyase functionality. The mechanistic changes that give rise to this change in functionality can be explored further using the Mechanism and Catalytic Site Atlas (M--CSA [13], formerly called MACiE [17] and CSA [18]).

FunTree is an important resource providing a comprehensive analysis of the evolution of enzyme functionality within structurally similar sub--divisions of CATH superfamilies. Not only will this improve our understanding of the link between enzyme structure and function, but, coupled with FunTree's various supporting analyses such as structural alignments and measures of molecular similarity, offers potential to inform de--novo enzyme design, annotate sequences/structures of unknown function and propose novel indications for drugs.

References

1. Sillitoe I, Furnham N (2016) FunTree: advances in a resource for exploring and

contextualising protein function evolution. Nucleic Acids Res 44:D317–D323. doi: 10.1093/nar/gkv1274

2. Sillitoe I, Lewis TE, Cuff A, et al (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381. doi: 10.1093/nar/gku947

3. Rahman SA, Cuesta SM, Furnham N, et al (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods 11:171–174. doi: 10.1038/nmeth.2803

4. Ruan J, Li H, Chen Z, et al (2007) TreeFam: 2008 Update. Nucleic Acids Res 36:D735–D740. doi: 10.1093/nar/gkm1005

5. Bostock M (2017) https://d3js.org. https://d3js.org.

6. Tamuri AU, Laskowski RA (2010) ArchSchema: a tool for interactive graphing of related Pfam domain architectures. Bioinformatics 26:1260–1261. doi: 10.1093/bioinformatics/btq119

7. Uniprot Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37:D169–D174. doi: 10.1093/nar/gkn664

8. Valdar WSJ (2002) Scoring residue conservation. Proteins Struct Funct Genet 48:227–241. doi: 10.1002/prot.10146

9. Gutmanas A, Alhroub Y, Battle GM, et al (2014) PDBe: Protein Data Bank in Europe. Nucleic Acids Res 42:D285–D291. doi: 10.1093/nar/gkt1180

10. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22:1540–1542. doi: 10.1093/bioinformatics/btl117

11. Rahman S, Bashton M, Holliday GL, et al (2009) Small Molecule Subgraph Detector (SMSD) toolkit. J Cheminform 1:12. doi: 10.1186/1758-2946-1-12

12. Yachdav G, Goldberg T, Wilzbach S, et al (2015) Anatomy of BioJS, an open source community for the life sciences. Elife. doi: 10.7554/eLife.07009

13. Ribeiro AJM, Holliday GL, Furnham N, et al (2017) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. submitted

14. Furnham N, Dawson NL, Rahman SA, et al (2016) Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies. J Mol Biol 428:253–267. doi: 10.1016/j.jmb.2015.11.010

15. Tyzack JD, Furnham N, Sillitoe I, et al (2017) Understanding enzyme function evolution from a computational perspective. Curr Opin Struct Biol 47:131–139. doi: 10.1016/j.sbi.2017.08.003

16. Furnham N, Sillitoe I, Holliday GL, et al (2012) Exploring the Evolution of Novel Enzyme Functions within Structurally Defined Protein Superfamilies. PLoS Comput Biol 8:e1002403. doi: 10.1371/journal.pcbi.1002403

17. Holliday GL, Bartlett GJ, Almonacid DE, et al (2005) MACiE: a database of enzyme reaction mechanisms. Bioinformatics 21:4315–4316. doi: 10.1093/bioinformatics/bti693

18. Furnham N, Holliday GL, de Beer TAP, et al (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42:D485–D489. doi: 10.1093/nar/gkt1243

Fig. 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.

Fig. 2: Superfamily gateway. CATH superfamily results for CATH 3.20.20.120 Enolase. Each thumbnail provides a link to a detailed analysis of the selected results. The SSGs within the superfamily are shown in Clusters with a link to lower level results for that SSG.

Fig. 3: CATH superfamily Overview Stats. The plot shows the number of sequences on the y--axis against the number of structures on the x--axis with colour representing the number of EC codes and size representing the number of partial EC codes.

Fig. 4: SSG (structurally similar group) Gateway: SSG results for CATH 3.20.20.120 SSG1. Each thumbnail provides a link to a detailed analysis of the selected results. The SSGs within the superfamily are shown in Clusters with a link to lower level results for that SSG.

Fig. 5: Rooted Phylogenetic Tree for SSG1 in CATH 3.20.20.120 Enolase. Each node contains a score that measures the confidence in the bifurcation. Each leaf contains labels for reaction similarity represented as green circles, EC code/function (where available), UniProtKB sequence, representative PDB domain (where available) and a domain bar representing the multi--domain architecture (MDA). See 2.1 for further details.

Fig. 6: Reaction Clustering for SSG1 in CATH 3.2.2.120 Enolase. A tree representing the similarities between reactions based on bond changes calculated by EC--Blast, where the clustering is made using the PVClust methods as implemented in R

Fig. 7: Annotated alignment for Phosphatidylinositol (PI) phosphodiesterase

Fig. 8: Summary of Phylogenetic, Functional, Metabolite and Multi--Domain Architectures for the Phosphatidylinositol--phosphodiesterase Superfamily (3.20.20.190) [16]. This shows a diagrammatic representation of the FunTree phylogenetic tree with associated functional data and multi--domain architectures. Domain 3.20.20.190 performs all molecular functionality and is represented in green in the Multi--Domain Architecture analysis. Three major clades (C1–C3) are highlighted. Within the first group a number of functional sub--groups can be observed, with differences in function defined by changes in substrate or product formed.