EDITORIAL

New computational approaches to understanding molecular protein function

Jacquelyn S. Fetrow[1]*, **Patricia C. Babbitt**[2]

1 Office of the President, Albright College, Reading, Pennsylvania, United States of America, 2 Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America

* jfetrow@albright.edu

OPEN ACCESS

Citation: Fetrow JS, Babbitt PC (2018) New computational approaches to understanding molecular protein function. PLoS Comput Biol 14 (4): e1005756. https://doi.org/10.1371/journal.pcbi.1005756

Published: April 5, 2018

Funding: The authors received no specific funding for this article.

Competing interests: The authors have declared that no competing interests exist.

Defining function

Function is like beauty—its definition lies in the eye of the beholder or, in this case, the researcher. At the broadest level, we define organismal function—the function that the protein plays in the overall organism. This function can be observed by understanding the impact on the organism of deletion or mutation of the protein. Physiological function is the function the protein plays in pathways, such as metabolic or signaling pathways. Another level of function is the cellular level. Approaches to understanding cellular function attempt to identify a protein's interaction partners and its location within the cell. Biochemical or molecular function is another level of function, one that identifies the molecular functional details of a functional site, including reaction mechanism, substrate binding, and molecular details of the binding of regulatory molecules. The International Union of Pure and Applied Chemistry (IUPAC) Enzyme Classification (EC) system was an early approach to identifying molecular function. The Gene Ontology (GO) system of classifying function recognizes ways of defining function, using distinct cellular components, molecular function, and biological process hierarchies [1]. The emphasis in the collection of research articles that comprise this Focus Feature is on understanding the different molecular functions that can exist within a protein superfamily and the features that define those differences.

Brief history of protein function identification/prediction

With the ongoing genome sequencing projects, the number of predicted proteins continues to increase exponentially. Understanding the biological role of these protein sequences requires an understanding of their functions, a goal that has resulted in many experimental and computational methods to evaluate gene product function. Large-scale experiments have aided in identification of protein-protein interactions, nonprotein binding partners, and expression levels, each of which define some manner of cellular function of the gene product. Other large-scale projects, such as the Enzyme Function Initiative (EFI, [2]) have focused on molecular or biochemical function.

Computational methods also abound. Our goal here is not to provide a thorough review but rather to provide a brief overview of new work in the field, with illustrative examples. The most commonly used approach is function annotation transfer, exemplified by common application of BLAST and Position-Specific Iterative BLAST (PSI-BLAST) [3,4]. In its simplest form, pairwise sequence similarity between 2 proteins is determined and, if the proteins are similar enough (i.e., if the similarity between them is deemed significant), function annotation is transferred from one protein to the other. There are many other more sophisticated approaches that can be applied for annotation transfer (e.g., ClustalOmega [5]) that take advantage of multiple alignments or alignment of a query protein to a probabilistic profile, as well as a plethora of other approaches, all of which are useful in various ways [6]. On the other hand, inappropriate application of annotation transfer has resulted in substantial annotation errors because the annotation transferred is not substantiated by available evidence [7]. Most often, misannotations arise from transfer of a more detailed molecular function than is warranted.

Other approaches to understanding molecular function involve motifs, either sequence motifs exemplified by PRINTS [8] and PROSITE [9] or structural motifs exemplified by Fuzzy Functional Forms [10] and Patterns in Non-homologous Tertiary Structures (PINTS) [11], or methods that use structure to inform sequence, such as Catalytic Site Atlas [12] and Enzyme Function Inference by a Combined Approach (EFICAz) [13]. All of these methods identify patterns of amino acids that correlate with a given function. To predict a functional site, a sequence or structure of unknown function is evaluated for how well it matches each motif. Many methods for evaluation of the match between motif and sequence (or structure) have been developed.

Finally, several groups have developed methods to cluster proteins into functionally relevant groups. These approaches—exemplified by Genome Modelling and Model Annotation (GEMMA) [14], Subfamily Classification in Phylogenomics (SCI-PHY) [15], Active Sites Modeling and Clustering (ASMC), [16], and Two-Level Iterative Clustering Process (TuLIP) [17]—cluster proteins using sequence, structure, evolutionary relationships, genomic context, or other metrics and correlate those clusters with some level of protein function. The goal of these approaches is not to predict function per se but rather to cluster proteins into functionally relevant groups, often using network methods [18,19], thus informing function annotation from an assumption that detailed molecular functional information can be transferred between members of the cluster. This approach has thus far most often applied to the molecular function of enzyme superfamilies.

Why molecular function?

The emphasis of this Focus Feature is molecular function, particularly for enzymes. A primary goal of molecular function analysis is to identify and understand the role of mechanistic or functional determinants, patterns of amino acids that distinguish one functional group from another. These are typically the amino acids that enable substrate binding and catalytic mechanism. Such features distinguish one molecular functional family from another. Their identification has long been the subject of expert analysis in the pharmaceutical and structure-based drug discovery programs, as these features are used to design the most specific inhibitors for a given functional site.

Understanding molecular functional determinants would also provide a better understanding of evolution. Contemporary protein superfamilies are the result of numerous genetic events, including gene duplications and horizontal transfers. The relationship between molecular function and evolution has long been debated [20]. Indeed, an underlying assumption in many annotation transfer approaches such as those mentioned above is that molecular function can be transferred from one family member to another. As has been clearly demonstrated [7,21–23], this assumption does not always hold, especially at detailed levels of molecular function. The ability to computationally identify molecular function at the most detailed level would provide the opportunity to compare trees and branches of evolutionary pathways, as has been illustrated by the analysis of the carbohydrate kinases [24].

Hierarchy in molecular function

Hierarchy is inherent in molecular function. Early on, this hierarchy was recognized in the IUPAC EC system, in which each enzyme was assigned a 4-digit EC number, W.X.Y.Z. In the

EC system, wherein W represents 1 of 6 main types of chemical reactions (classes) to which the enzyme belongs, X indicates a more detailed level of reaction, i.e., subclass, Y indicates the sub-subclass (typically reaction specificity), and Z is the serial number of the enzyme in its sub-subclass and describes substrate specificity, including pertinent cofactors.

Hierarchy is also represented in the GO system of molecular function classification [1]. As an example from the paper in this Focus Feature series by the Orengo group [25], the GO molecular function ontology term GO:0008800 represents "beta-lactamase activity." This function is further subdivided into GO:0033250 "penicillinase activity" and GO:0033251 "cephalosporinase activity," representing a hierarchy of molecular function in this superfamily from less specific to more specific.

The Structure-Function Linkage Database (SFLD) also defines a molecular functional hierarchy, in this case as superfamily, subgroup, and family (Fig 1) [26]. Superfamilies represent broad groups of proteins, such as the enolases or glutathione transferases, which are homologous (descended from a common ancestor) and which share a common reaction step or other chemical capability. Members of an SFLD subgroup share some, but not necessarily all, reaction steps, while family members share all or almost all steps of a reaction mechanism and perform the same molecular function. Some superfamilies within the SFLD database have been utilized as a "gold standard" for validation of approaches to predicting molecular function [7,14,27].

A key feature of annotation transfer methods is that the level of functional hierarchy must correlate with the similarity method and mathematical and statistical comparison that is being done [28]. An illustrative comparison between similarity network analysis [19] and the SFLD hierarchy is shown in Fig 1. In the original ASMC publication, by de Melo-Minardi and coworkers (contributors to this Focus Feature), for instance, the level of the functional hierarchy

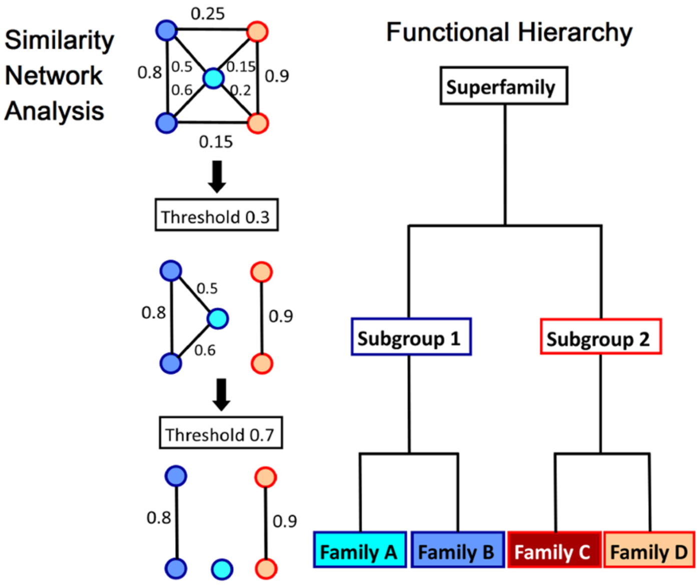


Fig 1. An illustration of the concept of molecular functional hierarchy and its correlation with network analysis. Similarity network analysis (left) uses edge thresholds to identify clusters that are progressively more similar to each other. As an example of molecular functional hierarchy, the Structure-Function Linkage Database (SFLD) hierarchy is shown on the right. Ideally, network clustering would capture the biologically relevant functional boundaries that would correlate with defined level of functional hierarchy, such as those defined by SFLD or the Gene Ontology (GO) hierarchies. (Note: the figure is illustrative and is not meant to suggest that an edge threshold of 0.3 correlates with the subgroup level of the SFLD hierarchy. One challenge in this field, illustrated by some of the papers in this Focus Feature, is which edge metric and threshold correlate with which levels of functional hierarchy).

https://doi.org/10.1371/journal.pcbi.1005756.g001

is called out explicitly for each protein superfamily studied [16]. For Multi-level Iterative Sequence Searching Technique (MISST), the topic of one of the papers in this Focus Feature series [29], the aim is to identify functional families using SFLD curated families as the gold standard definition of proteins that share all or almost all catalytic reaction steps. Defining the level of molecular function (Fig 1) being targeted by a given method is essential to interpretation of the results.

Focus feature on molecular function prediction

The goal of this Focus Feature is to present some recent work on determining molecular function and the details of features that distinguish one molecular function from another in a superfamily. Three research groups take the approach of clustering proteins into functionally relevant groups. The fourth research group describes a computational approach to enumerating potential reaction products in a protein superfamily, i.e., an approach to identifying what each of those functionally relevant clusters within a superfamily might do.

The contribution by Orengo and her colleagues describes their most recent work on clustering the beta lactamase protein superfamily [25]. Building on their previously published FunHMMer [30] work, this analysis distinguishes the A, C, and D classes of serine beta lactamases. Filtering these sequences using CD-HIT [31] at a 60% identity threshold divides Class A into 151 clusters. Nine of these correlate with 15 known types of the Class A beta lactamases. One hundred and forty-two are newly identified. Orengo and her colleagues then tackle the difficult problem of identification of the residues that confer antibiotic resistance and differentiate these clusters. As described in the Focus Feature contribution, an active site structure profile (ASSP) is used to identify key functional positions near the active site, and a second shell parsimony approach (SSPA) to identify additional mutations and drive more detailed functionally relevant clustering—to identify types and variants within each class. Such approaches are broadly useful for identifying mechanistic or functional determinants (such as those that confer antibiotic resistance).

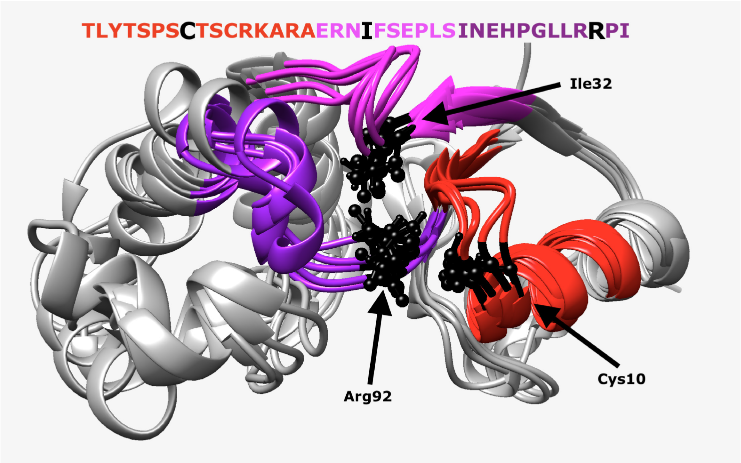
The contribution of de Melo-Minardi and her colleagues [32] describes the development of a new method for functionally relevant clustering that evolved from their published ASMC approach [16] to identifying functionally relevant clusters in the PFAM protein families database [33] superfamilies. Comparative models are built for each member of a PFAM superfamily that is at least 30% sequence identical to another superfamily member of known structure. Genetic programming (GP) is used to explore the optimal combinations of evidence regarding function (sequence, structural, and genomic context and active site information) that give the best separation of clusters. Spectral clustering divides the family into a specified number of clusters; mutual information identifies the optimal number of clusters. This method, called a GP (for genetic programming) approach by the authors, is applied to a variety of superfamilies, including nucleotidyl cyclases, kinases, serine proteases, enolases, and crotonases, with good success.

A third method, MISST [29], for clustering proteins in functionally relevant ways is also part of this Focus Feature series. This method implements an approach called Deacon Active Site Profiler (DASP), [34]. DASP utilizes the concept of active site profiling [35] to identify sequences that share functional site features in common. This method was built on a key observation that isofunctional clusters, groups of sequences that share most details at their molecular sites, self-identify in DASP searches [17]. This observation is built into an iterative search process, MISST, that both aggregates sequences and identifies when a cluster should be subdivided because it is composed of more than one isofunctional group. In this contribution, the approach is applied to the peroxiredoxin superfamily. Notably, this approach does not start with all members of a superfamily, as do the first 2 methods, but rather aggregates sequences from the GenBank database into functionally relevant clusters within a superfamily.

Notably, the above methods are not function prediction methods because, once functionally relevant clusters are identified, the key question remains to be answered: What do the proteins in each cluster actually do? That is, what is the reaction mechanism of the proteins in each cluster, and what are the substrates utilized and products produced by each cluster? Jacobson and his colleagues present an approach to answering that question, which builds on previous work on the triterpenoid synthases that utilized docking to identify potential substrates and functional families [36]. Their contribution to this Focus Feature series describes the use of structural modeling, virtual reactions, and energy calculations to systematically enumerate plausible terpenoid carbocations, thus exploring chemical space of the products of the monoterpenoid synthases [37]. Five of the 74 identified skeletons are found in natural products. Others can be connected to substrates by small carbocation rearrangements already known to occur. The authors hypothesize that these skeletons may represent currently unidentified natural products of one or more of the terpene synthases for which detailed molecular function is uncharacterized. The enumeration algorithm they describe, iGEN, could be used to create carbocations in the active site of an enzyme, which may allow both the prediction of novel terpenoid skeletons and matching them to a terpenoid synthase of unknown function. Indeed, a proof of concept of this approach has been published [38].

Commonalities in approach: Functional site signatures

A common feature of several of the clustering approaches in this Focus Feature is the definition of a "functional site signature." At its simplest, a signature consists of the "mechanistic determinants" or "specificity-determining positions" or "functional determinants"—only those residues that are actively involved in substrate binding or catalytic mechanism (as illustrated for an arsenate reductase, Fig 2). Godzik and colleagues called these residues "signature positions" in their work on the carbohydrate kinases [24]. A more encompassing definition of an active site signature was defined by Cammer and colleagues [35] and includes all residues



```
Active Site Signature:

  1Z3E   TLYTSPSCTSCRKARAERNIFSEPLSINEHPGLLRRPI

Active Site Profile:

  1Z3E   TLYTSPSCTSCRKARAERNIFSEPLSINEHPGLLRRPI
  3L78   TLFLSPSCTSCRKARAEHNIMTSPLSISKNPSLLRRPI
  3FZ4   TFYEYPKCSTCRRAKAAIDIKKNPPALASDGXLIKRPL
  2M46   KFYQYKNCTTCKKAAKPIDIVQHTPTLSSDGMLVKRPL
  3GKX   LFLQYPACSTCQKAKKNRLIVDDNPTLATNGKLVKRPL
          :     *::*::*     *     ::  .  *::**:
```

Fig 2. **Specificity determining or "signature" positions are amino acids directly involved in an enzyme's activity.** Three such residues are shown as black side chains (left) for 5 arsenate reductase enzymes. Residues within close structural space are shown as colored fragments. The sequences of those colored fragments create the active site signature (right), a residue sequence originally defined by Cammer and colleagues [35]. Signatures can be aligned to create a profile, which allows direct comparison of residues in and near the active site. This active site profile concept is used by 3 of the approaches for functionally relevant clustering of protein superfamilies included in this Focus Feature. (The authors gratefully acknowledge Mikaela Rosen for creating these figures).

https://doi.org/10.1371/journal.pcbi.1005756.g002

within 10 Å of the centers of geometry of 3 key specificity-determining residues. Similarly, Orengo's ASSP algorithm identifies all residues within 8 Å of the key active site serine in the beta lactamases to build the ASSP.

Each of the 3 clustering approaches use the concept of active site profiling, including a multiple sequence alignment that represents the active site composition for each protein in the family. This alignment is termed an active site profile (ASP) by Fetrow and colleagues [35] and an ASSP by Orengo and colleagues [25]. In the Orengo method, ASSP is used to identify functionally important residues. In de Melo-Minardi's contribution, these residues are used to optimize the separation of clusters [32]. In MISST, the profiles are used by DASP to search the sequence databases to identify protein sequences that contain fragments that are statistically similar to those in the query profile [29]. No matter how such profiles are used, the definition of such functionally relevant signatures is valuable in informing our understanding of the mechanistic differences between functionally relevant clusters.

Challenges and a view to the future

A key challenge for clustering proteins in functionally relevant ways is the identification of the number of clusters that are optimal for defining all distinct molecular functions within a protein superfamily—that is, the number of clusters that captures biologically relevant functional boundaries that correlate with the level of molecular functional hierarchy being pursued (Fig 1). The previously published ASMC method utilized the full hierarchy of clusters, with manual pruning of branches [16], while de Melo-Minardi's current contribution solves this problem using mutual information to identify an appropriate number of clusters [32]. FunHMMer identified many families in the beta lactamases, and Orengo and her colleagues used sequence identity and active site residue analysis to distinguish and/or combine these into functionally relevant clusters in the hierarchy. MISST is an aggregative, rather than divisive, clustering approach, starting with a small, representative set of proteins, aggregating new members, and identifying the point at which a cluster contains more than one isofunctional cluster. Thus, it only has to identify when a cluster is "complete," rather than at what level the clustering correlates with molecular function.

Another key challenge for methods aimed at molecular function identification is validation. Isofunctional clusters, defined as a family by SFLD, have been defined and experimentally validated for only a small number of families. For example, the enolase superfamily and its functionally relevant families have been studied extensively [39–41], and these functionally relevant groups (along with those of other SFLD superfamilies) have been used for validation as well [27]. More recently, the peroxiredoxins [42], FGGY carbohydrate kinases [24], and matrix metalloproteins [43] have also been described at a detailed molecular functional level and could serve as additional validation standards.

Clearly, our understanding of molecular function continues to evolve. Each of the methods described in this Focus Feature builds on previous work, and each continues to aid our understanding of molecular function and relationships between functionally relevant clusters within a protein superfamily. For the future, much new work seeks to improve automated and "agnostic" approaches for functional inference, while other communities, such as Critical Assessment of Function Annotation (CAFA) [6], have developed improved approaches for evaluation of such methods. And while the "messiness" of biology continues to challenge general solutions for identifying functional boundaries, new approaches such as those described in the Focus Feature offer new directions for exploration. The need for continued progress remains clear, however, as the volume of sequencing will continue to increasingly outpace experimental validation.

The Protein Molecular Function Prediction Focus Feature consists of the following 4 research articles:

Novel computational protocols for functionally classifying and characterising serine beta-lactamases. David Lee, Sayoni Das, Natalie L. Dawson, Dragana Dobrijevic, John Ward, Christine Orengo. Published 22 Jun 2016. *PLOS Computational Biology*. http://dx.doi.org/10.1371/journal.pcbi.1004926

Isofunctional protein subfamily detection using data integration and spectral clustering. Elisa Boari de Lima, Wagner Meira Júnior, Raquel Cardoso de De Melo-Minardi. Published 27 Jun 2016. *PLOS Computational Biology*. http://dx.doi.org/10.1371/journal.pcbi.1005001

Defining the product chemical space of monoterpenoid synthases. Boxue Tian, C. Dale Poulter, Matthew P. Jacobson. Published 12 Aug 2016. *PLOS Computational Biology*. http://dx.doi.org/10.1371/journal.pcbi.1005053

An atlas of peroxiredoxins created using an active site profile-based approach to functionally relevant clustering of proteins. Angela F. Harper, Janelle B. Leuthaeuser, Patricia C. Babbitt, John H. Morris, Thomas E. Ferrin, Leslie B. Poole, Jacquelyn S. Fetrow. Published 10 Feb 2017. *PLOS Computational Biology*. http://dx.doi.org/10.1371/journal.pcbi.1005284

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25: 25–9. https://doi.org/10.1038/75556 PMID: 10802651
2. Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, et al. The Enzyme Function Initiative. Biochemistry. 2011; 50: 9950–62. https://doi.org/10.1021/bi201312u PMID: 21999478
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Shang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25: 3389–3402. PMID: 9254694
5. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol Clifton NJ. 2014; 1079: 105–116. https://doi.org/10.1007/978-1-62703-646-7_6 PMID: 24170397
6. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016; 17: 184. https://doi.org/10.1186/s13059-016-1037-6 PMID: 27604469
7. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009; 5: e1000605. https://doi.org/10.1371/journal.pcbi.1000605 PMID: 20011109
8. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, et al. The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. Database J Biol Databases Curation. 2012; 2012: bas019. https://doi.org/10.1093/database/bas019 PMID: 22508994
9. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010; 38: D161–166. https://doi.org/10.1093/nar/gkp885 PMID: 19858104
10. Fetrow JS, Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol. 1998; 281: 949–68. https://doi.org/10.1006/jmbi.1998.1993 PMID: 9719646

11. Stark A, Russell RB. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. Nucleic Acids Res. 2003; 31: 3341–4. PMID: 12824322

12. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 2004; 32: D129–33. https://doi.org/10.1093/nar/gkh028 PMID: 14681376

13. Kumar N, Skolnick J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. Bioinforma Oxf Engl. 2012; 28: 2687–2688. https://doi.org/10.1093/bioinformatics/bts510 PMID: 22923291

14. Lee DA, Rentzsch R, Orengo C. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. 2010; 38: 720–37. https://doi.org/10.1093/nar/gkp1049 PMID: 19923231

15. Brown DP, Krishnamurthy N, Sjölander K. Automated protein subfamily identification and classification. PLoS Comput Biol. 2007; 3: e160. https://doi.org/10.1371/journal.pcbi.0030160 PMID: 17708678

16. de Melo-Minardi RC, Bastard K, Artiguenave F. Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinforma Oxf Engl. 2010; 26: 3075–3082. https://doi.org/10.1093/bioinformatics/btq595 PMID: 20980272

17. Knutson ST, Westwood BM, Leuthaeuser JB, Turner B, Nguyendac D, Shea G, et al. An approach to functionally relevant clustering of the protein universe: Active site profile-based clustering of protein structures and sequences. Protein Sci Publ Protein Soc. 2017; https://doi.org/10.1002/pro.3112 PMID: 28054422

18. Enright AJ, Ouzounis CA. BioLayout—an automatic graph layout algorithm for similarity visualization. Bioinforma Oxf Engl. 2001; 17: 853–854.

19. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS ONE. 2009; 4: e4345. https://doi.org/10.1371/journal.pone.0004345 PMID: 19190775

20. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity determining residues. Genome Biol. 2002; 3: PREPRINT0002. PMID: 11897020

21. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333: 863–82. PMID: 14568541

22. Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318: 595–608. https://doi.org/10.1016/S0022-2836(02)00016-5 PMID: 12051862

23. Addou S, Rentzsch R, Lee D, Orengo CA. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol. 2009; 387: 416–430. https://doi.org/10.1016/j.jmb.2008.12.045 PMID: 19135455

24. Zhang Y, Zagnitko O, Rodionova I, Osterman A, Godzik A. The FGGY carbohydrate kinase family: insights into the evolution of functional specificities. PLoS Comput Biol. 2011; 7: e1002318. https://doi.org/10.1371/journal.pcbi.1002318 PMID: 22215998

25. Lee D, Das S, Dawson NL, Dobrijevic D, Ward J, Orengo C. Novel Computational Protocols for Functionally Classifying and Characterising Serine Beta-Lactamases. PLoS Comput Biol. 2016; 12: e1004926. https://doi.org/10.1371/journal.pcbi.1004926 PMID: 27332861

26. Akiva E, Brown S, Almonacid DE, Barber AE, Custer AF, Hicks MA, et al. The Structure-Function Linkage Database. Nucleic Acids Res. 2014; 42: D521–530. https://doi.org/10.1093/nar/gkt1130 PMID: 24271399

27. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC. A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biol. 2006; 7: R8. https://doi.org/10.1186/gb-2006-7-1-r8 PMID: 16507141

28. Leuthaeuser JB, Knutson ST, Kumar K, Babbitt PC, Fetrow JS. Comparison of topological clustering within protein networks using edge metrics that evaluate full sequence, full structure, and active site microenvironment similarity. Protein Sci. 2015; 24: 1423–1439. https://doi.org/10.1002/pro.2724 PMID: 26073648

29. Harper AF, Leuthaeuser JB, Babbitt PC, Morris JH, Ferrin TE, Poole LB, et al. An Atlas of Peroxiredoxins Created Using an Active Site Profile-Based Approach to Functionally Relevant Clustering of Proteins. PLoS Comput Biol. 2017; 13: e1005284. https://doi.org/10.1371/journal.pcbi.1005284 PMID: 28187133

30. Das S, Sillitoe I, Lee D, Lees JG, Dawson NL, Ward J, et al. CATH FunFHMMer web server: protein functional annotations using functional family assignments. Nucleic Acids Res. 2015; 43: W148–153. https://doi.org/10.1093/nar/gkv488 PMID: 25964299

31. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma Oxf Engl. 2012; 28: 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

32. Boari de Lima E, Meira W, Melo-Minardi RC de. Isofunctional Protein Subfamily Detection Using Data Integration and Spectral Clustering. PLoS Comput Biol. 2016; 12: e1005001. https://doi.org/10.1371/journal.pcbi.1005001 PMID: 27348631

33. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016; 44: D279–285. https://doi.org/10.1093/nar/gkv1344 PMID: 26673716

34. Huff RG, Bayram E, Tan H, Knutson ST, Knaggs MH, Richon AB, et al. Chemical and structural diversity in cyclooxygenase protein active sites. Chem Biodivers. 2005; 2: 1533–1552. https://doi.org/10.1002/cbdv.200590125 PMID: 17191953

35. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, et al. Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol. 2003; 334: 387–401. PMID: 14623182

36. Tian B-X, Wallrapp FH, Holiday GL, Chow J-Y, Babbitt PC, Poulter CD, et al. Predicting the functions and specificity of triterpenoid synthases: a mechanism-based multi-intermediate docking approach. PLoS Comput Biol. 2014; 10: e1003874. https://doi.org/10.1371/journal.pcbi.1003874 PMID: 25299649

37. Tian B, Poulter CD, Jacobson MP. Defining the Product Chemical Space of Monoterpenoid Synthases. PLoS Comput Biol. 2016; 12: e1005053. https://doi.org/10.1371/journal.pcbi.1005053 PMID: 27517297

38. Chow J-Y, Tian B-X, Ramamoorthy G, Hillerich BS, Seidel RD, Almo SC, et al. Computational-guided discovery and characterization of a sesquiterpene synthase from Streptomyces clavuligerus. Proc Natl Acad Sci U S A. 2015; 112: 5661–5666. https://doi.org/10.1073/pnas.1505127112 PMID: 25901324

39. Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, et al. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. Biochemistry (Mosc). 1996; 35: 16489–501.

40. Gerlt JA, Babbitt PC, Jacobson MP, Almo SC. Divergent evolution in enolase superfamily: strategies for assigning functions. J Biol Chem. 2012; 287: 29–34. https://doi.org/10.1074/jbc.R111.240945 PMID: 22069326

41. Gerlt JA, Babbitt PC, Rayment I. Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. Arch Biochem Biophys. 2005; 433: 59–70. https://doi.org/10.1016/j.abb.2004.07.034 PMID: 15581566

42. Nelson KJ, Knutson ST, Soito L, Klomsiri C, Poole LB, Fetrow JS. Analysis of the peroxiredoxin family: using active-site structure and sequence information for global classification and residue analysis. Proteins. 2011; 79: 947–64. https://doi.org/10.1002/prot.22936 PMID: 21287625

43. Ratnikov BI, Cieplak P, Gramatikoff K, Pierce J, Eroshkin A, Igarashi Y, et al. Basis for substrate recognition and distinction by matrix metalloproteinases. Proc Natl Acad Sci U S A. 2014; 111: E4148–4155. https://doi.org/10.1073/pnas.1406134111 PMID: 25246591