REPRESENTING STRUCTURE-FUNCTION RELATIONSHIPS IN MECHANISTICALLY DIVERSE ENZYME SUPERFAMILIES

SCOTT C.-H. PEGG[†], SHOSHANA BROWN[†], SUNIL OJHA[†], CONRAD C. HUANG[*], THOMAS E. FERRIN[*†], PATRICIA C. BABBITT[†*]

[†]Dept. of Biopharmaceutical Sciences and [*]Dept. of Pharmaceutical Chemistry
University of California, San Francisco, 94143

The prediction of protein function from structure or sequence data remains a problem best addressed by leveraging information available from previously determined structure-function relationships. In the case of enzymes, the study of mechanistically diverse superfamilies can provide a rich source of structure-function information useful in functional determination and enzyme engineering. To access these relationships using a computational resource, several issues must be addressed regarding the representation of enzyme function, the organization of structure-function relationships in the superfamily context, the handling of misannotations, and reliability of classifications and evidence. We discuss here our approaches to solving these problems in the development of a Structure-Function Linkage Database (SFLD) (online at http://sfld.rbvi.ucsf.edu).

1. Introduction

The solution of a protein's three-dimensional structure often does not immediately lead to the determination of its function.[1] Typically, we take the natural step of leveraging the information gained from experimental determinations of function by asking the question, "Is this structure and active site one that I've seen before, and if so, what does it do and how?" As the number of both solved protein structures and experimental determinations of function increase (with the former growing much faster than the latter), there is a growing need for computational methods for storing and searching representations of protein function in a way that correlates specific aspects of function with sequence and structural features. Ideally, such representations of function should go beyond simple identification of conserved and/or functionally validated residues in sequence alignments.

In the case of enzymes, the study of mechanistically diverse superfamilies—sets of homologous enzymes which, while often sharing very little sequence similarity to each other and often catalyzing different overall reactions with a variety of substrates and products, share the same fold and conserve a specific partial reaction (or some other aspect of mechanism) enabled by a conserved set of residues[2, 3]—allows us to leverage structure-function information at multiple levels. At the highest level, we can infer only a partial mechanistic step and the associated functionally important residues that are common across all members of the superfamily. At the lowest, most detailed level, we can determine the specific function of a single enzyme, including its mechanism as performed by specific active-site residues and co-factors. Often, however, because of the sophisticated level of chemical intuition required to identify the partial reaction(s) associated with conserved structural characteristics in such diverse proteins, only researchers who are intimately familiar with a given enzyme superfamily can take advantage of the structure-function information it contains. A resource that allows other investigators to utilize this information represents a valuable tool.

The terrain of mechanistically diverse enzyme superfamilies presents a number of obstacles that must be surmounted, some common to any database linking structural and functional information, others unique to mechanistically diverse enzyme superfamilies. The former include handling multi-functional enzymes, representing function in a computationally accessible format, and dealing with potential inaccuracies in annotation. Unique to analysis of mechanistically diverse enzyme superfamilies is the need to capture chemical function both in terms of the overall chemical reactions performed, but also at the level of the common partial reaction (or common chemical capability) associated with all of the different overall reactions represented in a superfamily. Partial reactions are captured in the Structure-Function Linkage Database (SFLD) by way of a partial reaction table in the relational database schema (see Fig. 4 below). Providing this information is especially important for the SFLD because it is only these partial reactions that correlate with active site similarities across all diverse members of a given superfamily. Identifying these partial reactions and linking them to structure provides a powerful tool for difficult problems in functional inference and protein engineering.[3, 4]

In this paper we discuss several major issues involved in the development of a computational resource for the storage and leverage of structure-function data in mechanistically diverse enzyme superfamilies and our specific approaches to handling these issues in the SFLD. Currently, five different superfamilies, representing over 3,800 sequences, are available in the database, with substantial expansion planned over the next year. Several of these superfamilies are used as examples here. Access to the SFLD is provided by a world-wide web based graphical user interface which accommodates many search and browse capabilities linking sequences, structures, and representations of the associated chemical transformations. A more detailed description of the content, uses, and the scientific principles motivating development of the SFLD will be presented elsewhere (manuscript in prepraration).

2. Representing Structure-Function Relationships in Mechanistically Diverse Enzyme Superfamilies

2.1. *Organization of Enzyme Superfamilies*

Within a mechanistically diverse enzyme superfamily, the elements of sequence and structure that deliver catalytic function are conserved to varying degrees. While all members of a given superfamily will possess the sequence and structural elements relating to the ability to perform the conserved mechanistic step (e.g., partial reaction) which helps define the superfamily, other subsets of the superfamily will possess a superset of conserved elements relating to other aspects of catalytic function. To clarify the distinction between these conserved elements, we have organized the enzyme superfamilies of the SFLD into a hierarchy of groupings. Figure 1 illustrates this hierarchy using the enolase superfamily[5] as an example.

At the top level of the hierarchy, enzymes are classified into the same superfamily if they appear to be evolutionarily related (based on sequence and structural information) and to share a common chemical capability (in the case of the enolase superfamily example, abstraction of a proton alpha to a carboxylic acid). The subgroup classification at the middle level of the hierarchy is superfamily-specific, and is defined by SFLD curators. In the enolase superfamily, enzymes are divided into subgroups based on active site residue motifs. At the next level of the hierarchy are families of enzymes each of whose members catalyze the same overall reaction. At the bottom of the hierarchy is a single enzyme, referred to as an enzyme functional domain (EFD). (See section 2.3 below for an example of how EFDs are defined.) According to the SFLD schema, an EFD need not be classified into a family to be classified into a subgroup or superfamily. Thus, if the full catalytic function of an EFD cannot be reliably determined, it may still be placed into a higher-level category in the hierarchy.

Because the hierarchical organization of EFDs into superfamilies, subgroups and families is based on functional as well as evolutionary criteria, functional classification of new sequences and structures is facilitated. For example, if an uncharacterized enzyme can be placed within a superfamily, the reaction catalyzed by the enzyme can be expected to utilize the chemical capability common to the superfamily. The overall reaction catalyzed by the enzyme may then be inferred based on additional information, such as operon context, or by further classifying the enzyme into a subgroup or family.

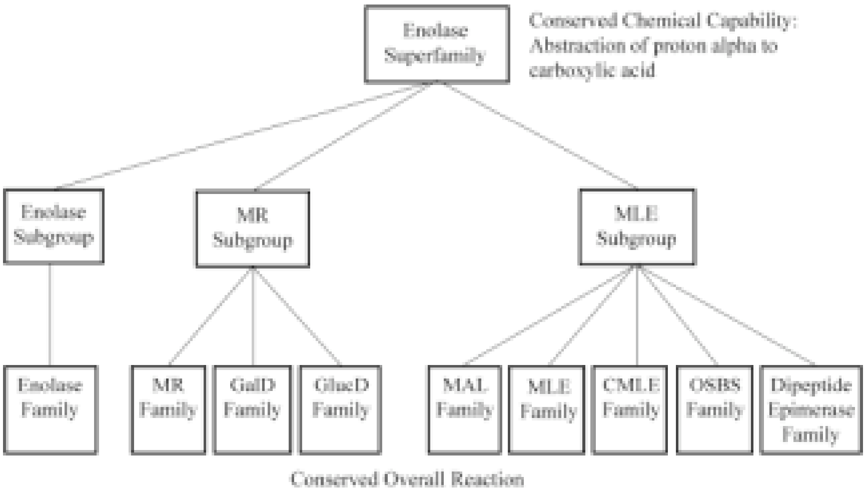


Figure 1. Hierarchical classification of EFDs in the enolase superfamily, based on sequence, structure and function. At the top level of the hierarchy, EFDs are classified into the same superfamily if they appear to be evolutionarily related based on sequence and structural information and to share a common chemical capability. The subgroup classification at the middle level of the hierarchy is superfamily-specific, and is defined by SFLD curators. At the bottom level of the hierarchy, enzymes in the same family are thought to catalyze the same overall reaction. Abbreviations used in this example: MR: mandelate racemase, GalD: galactonate dehydratase, GlucD, glucarate dehydratase, MAL: b-methylaspartate ammonia-lyase, MLE: muconate cycloisomerase, CMLE: chloromuconate cycloisomerase, OSBS: o-succinylbenzoate synthase.

2.2. *Representation of Catalyzed Reactions*

In order for the reactions catalyzed by the enzymes in mechanistically diverse superfamilies to be rapidly searched and compared to each other, they must be stored in a computationally accessible format that allows for not just comparisons of overall and partial reactions, but comparisons of substrate and product substructure. This issue has been addressed in the field of small molecule synthetic chemistry, where the standard has become the use of SMILES/SMARTS strings.[6] SMILES/SMARTS provides the type of functionality required to link enzyme chemistry to the sequence and structure information provided in the SFLD, as these strings of ASCII characters represent the chemical structures of participants in a reaction, including chirality. We have adopted this format for the SFLD, allowing users to search the overall and partial reactions using both reactions and substructures as queries. Figure 2 gives an example of some SMILES/SMARTS representations and queries.

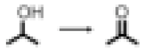
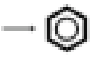
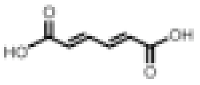
SMARTS Query	Meaning	Schematic
C([OD1])>>C(=O)	Reaction contains the conversion of an alcohol to a ketone	
>>c1ccccc1	Product of the reaction contains a benzene ring	
C(=O)([OD1])C=CC=CC(=O)[OD1]	Either the substrate or the product of the reaction contains mandelate	

Figure 2. Examples of possible SMARTS queries and their chemical meanings.

While very flexible, and a good solution in the field of synthetic organic chemistry, SMILES/SMARTS does not provide a comprehensive solution for the study of enzyme chemistry, however. Extension of these representations, currently underway in our laboratory, will also be required. These include representation of the chemical contributions of active site residues as well as metals and complex cofactors.

2.3. *Enzymes with Multiple Functions*

A major hurdle in representing structure-function information is the issue of multi-functional enzymes—single protein sequences that catalyze multiple chemical reactions. Multi-functional enzymes can be viewed as one of three types. The first type consists of enzymes with multiple fused, but fundamentally independent domains, each with a separate active site. For example, phosphoribosylanthranilate isomerase (PRAI) and indoleglycerolphosphate synthase (IGPS), which catalyze two consecutive reactions in the tryptophan biosynthetic pathway, occur as a single protein chain in *E. coli*. Although the *E. coli* protein catalyzes both reactions, the n-terminal domain is responsible for the IGPS reaction, while the c-terminal domain is responsible for the PRAI reaction.[7-10] Furthermore, PRAI and IGPS occur as two physically separate protein chains in other organisms, such as *T. maritima*.

The SFLD accommodates this type of multifunctional protein by storing enzyme information at the level of the enzyme functional domain (EFD). An EFD is an enzyme, or part of an enzyme, that is capable of catalyzing a chemical reaction on its own. Thus, the *E. coli* IGPS-PRAI protein would be divided into two separate EFDs, one corresponding to the n-terminal domain of the protein, and one corresponding to the c-terminal domain of the protein.

The second type of multifunctional enzyme represents proteins that are capable of catalyzing an adventitious secondary reaction in the same active site responsible for catalyzing its primary reaction. One example, illustrated in Fig. 3, is the enolase superfamily enzyme, o-succinylbenzoate synthase (OSBS), from *Amycolaptosis sp*. In addition to the biologically relevant OSBS reaction, this enzyme also catalyzes the industrially important n-acylamino acid racemase (NAAAR) reaction.[11] The enzyme utilizes the same active site to catalyze both of these very different overall reactions. Catalytic promiscuity has been noted in other enzymes,[12, 13] but because it is difficult to determine possible secondary reactions based on the primary reaction of an enzyme, the overall incidence of this type of promiscuity is unknown. Many enzymes have also been noted to exhibit a third, and much more common form of catalytic promiscuity, e.g. turning over a variety of substrates related to their primary substrate.[12, 14, 15]

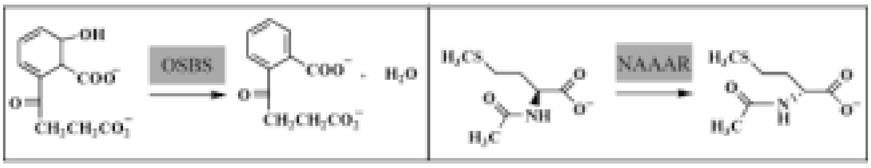


Figure 3. The O-succinylbenzoate synthase (OSBS) enzyme in *Amycolaptosis* sp. catalyzes both the OSBS reaction and the N-Acylamino Acid Racemase (NAAAR) reaction using the same active site.

The SFLD accommodates these latter two types of multifunctional enzymes by providing a many-to-many relationship between the EFD and Reaction tables—a single EFD can have an arbitrary number of Reaction entries, and can turn over multiple substrates. When known, the SFLD schema marks a single canonical reaction for each EFD to indicate which of the multiple reactions catalyzed has the greatest biological relevance. Figure 4 shows a simplified version of the SFLD schema. Note the inclusion of a partial reaction table, which, as discussed above, is key to the ability to correlate conserved structural elements/residues to the chemical functions conserved at the superfamily level.

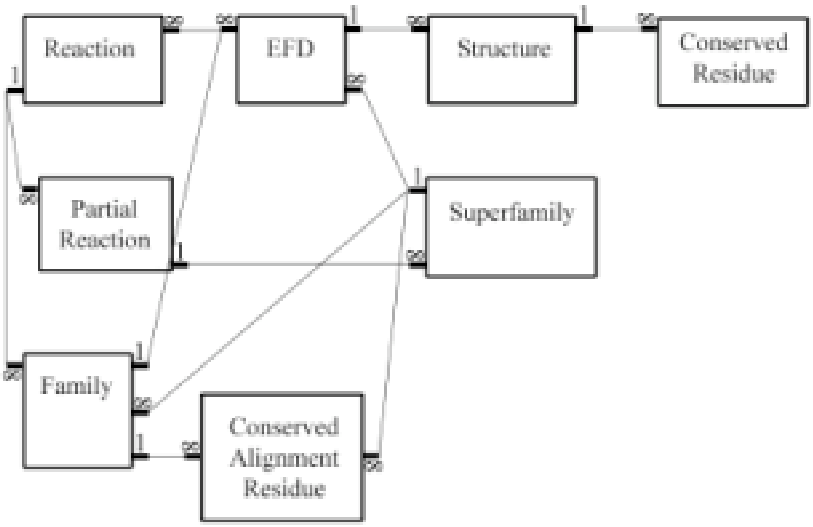


Figure 4. A simplified version of the SFLD schema.

2.4. *Functional Annotation and Misannotation*

The sheer size of many of the most commonly used sequence databases and sequencing projects requires the use of automated methods for assigning protein function to newly sequenced open reading frames (ORFs), leading to significant levels of misaannotation.[16-18] These typically sequence based methods face an especially difficult challenge when dealing with mechanistically diverse enzyme superfamilies.[2] This is because a superfamily member of unknown function can show a statistically significant level of sequence similarity to other members of the superfamily which, although they share a common mechanistic step, do not perform the same overall reaction. If, in the set of statistically significant matches, the closest characterized sequence represents an enzyme whose function is from a different family than the unknown sequence of interest, that sequence will often mistakenly be assigned the function of the characterized sequence.

The use of explicitly linked structure-function data helps us address the problems of annotation and misannotation. The SFLD allows users to place sequences of unknown function which appear to belong to an enzyme superfamily into a multiple sequence alignment of the superfamily, the subgroup, or family, thus providing an easily accessible basis for functional assignment. The superfamily level alignment includes information about the positions and residues of the superfamily that deliver catalytic function, allowing users to quickly evaluate whether the new sequence possesses the catalytic machinery required to perform partial reactions common to all members of the superfamily. Placement of these sequences in multiple alignments at the subgroup or family level aids in determining the likelihood that a given ORF has been accurately annotated with regard to the identity of its substrate and overall chemical reaction. This simple analysis, while not foolproof, has proven useful in evaluating the accuracy of annotations within a superfamily. Our lab was able to determine, for example, that 8 of the 30 sequences annotated in Genbank[19] as muconate cycloisomerases, while certainly members of the enolase superfamily, lack catalytic residues required to perform the specific annotated function.[20] To further enrich these capabilities, work is underway to link these alignments with the Chimera visualization software,[21] allowing users to view relevant three-dimensional structures simultaneously with related multiple sequence alignments.

Of course, multiple sequence alignments do not always provide enough information for a biologist to determine the precise function, e.g., the family level classification, of a new member of an enzyme superfamily. Often, only a subset of the functional and conserved residues of a family or superfamily can be identified in the new sequence. This only allows us to predict accurately that this enzyme will include the mechanistic step conserved throughout the superfamily, but not the substrate or the product. Alternatively, new sequences may perform new reactions for which the functionally important residues have not been identified (or cannot be inferred by alignment to known families). The SFLD classifies such sequences as within the proper superfamily, or in some cases within a subgroup labeled "unknown function", but not within a family. Information about the metabolic pathway, and when available, operon context, can also aid in the determination of an enzyme's primary catalytic role.[22-24] We are in the process of integrating these types of information into the SFLD.

2.5. *Guidance for Protein Engineering*

Protein engineering, whether to generate new functions or to improve on old ones, requires choosing a template protein for use as a scaffold. The kind of information captured in the SFLD can be used to guide this choice. Because mechanistically diverse enzyme superfamilies have been used by nature to evolve many different enzymatic reactions, it follows that superfamily members could be useful templates in the lab to re-engineer new and different enzymes as well. For example, it has been shown that two members of the enolase superfamily can be reengineered, via a single point mutation, to perform the very different reaction of a third member.[4] The key to success in this experiment was the recognition, provided by the superfamily context, of the common partial reaction all three members share. In effect, the active site of each of the superfamily members is already pre-organized to perform the proton abstraction step required for any other member, simplifying the re-engineering problem. The SFLD exploits these principles by allowing users to identify superfamily scaffold proteins potentially capable of performing a fundamental partial reaction required to generate entirely new chemical reactions.

2.6. *Grading the Reliability of Functional Information*

As mentioned above, the SFLD can facilitate the functional classification of an uncharacterized protein by placing that protein into the appropriate superfamily, subgroup, or family. The reliability of such a classification depends greatly on the quality of the classifications within the SFLD itself. For example, if an uncharacterized protein X closely resembles proteins in the adenosine deaminase family within the amidohydrolase superfamily, one might want to determine whether the closest relatives of protein X have been experimentally determined to perform the adenosine deaminase reaction or whether their family classification was made based merely on sequence similarity to other experimentally characterized members of the family. If protein X is closely related to experimentally characterized family members, that strengthens the argument for assigning protein X the adenosine deaminase function.

The SFLD uses evidence codes to indicate the type and reliability of functional information. SFLD evidence codes are based on those developed by the Gene Ontology consortium,[25] but have been modified to fit the requirements of representing structure-function information. Where applicable, evidence codes are paired with the literature references upon which they are based. The assignment of a particular EFD to a family, for example, comes with an evidence code and literature references deemed relevant by the curator to this assignment. This is similar to approaches adopted by other resources such as SwissProt.[26] Some examples of evidence codes that might be used for family assignments, ordered roughly in order of reliability, are:

> *IES (Inferred from Experiment and Sequence)*: Used when family membership is assigned based on an experimental assay that shows that the EFD in question catalyzes the canonical family reaction, and there is clear sequence and/or structural similarity to existing family members.

> *ISS (Inferred from Sequence or Structural similarity)*: Used when family assignment is based on overall sequence or structural similarity, reviewed for accuracy by a human curator, to existing family members.

IEA (Inferred from Electronic Annotation): Used when family assignment is based on overall sequence or structural similarity to existing family members but has not been reviewed for accuracy by a human curator.

The evidence codes used in the SFLD and their definitions can be found at http://sfld.rbvi.ucsf.edu/ecodes.html.

2.7. *Metadata*

Whenever we try to place the boundaries of classification upon biological systems, we inevitably are confronted with cases that appear to stretch the rules. An enzyme, for example, may have as its biologically relevant function the catalysis of multiple reactions. The humulene synthase enzyme from *Abies grandis*, for instance, is known to catalyze reactions leading to at least 52 distinct products, only a fraction of which are of known biological importance.[27] The presence of such information, while often useful to users, is difficult to predict prior to curation. To handle these cases, nearly all of the SFLD tables contain a "metadata" field in which the curators of a family or superfamily can enter textual information.

2.8. *Methods of Searching the SFLD*

The main purpose of the SFLD is to facilitate the leveraging of structure-function data. Thus, it is of the highest importance that users be able to access the data via methods most informative from their own scientific perspectives. For example, a protein engineer looking to design an enzyme to perform a particular reaction might want to search the SFLD for enzymes catalyzing similar reactions or underlying partial reactions. Such searches can be performed by entering a SMARTS query, or by sketching chemical structures using a Java applet on the SFLD search page. Alternatively, users can also search the reactions by Enzyme Commission number[28] or simply browse a list of all the reactions in the database.

Those interested primarily in the determining the function of an uncharacterized protein can query the SFLD using its sequence. This sequence is matched to pre-generated hidden Markov models[29] representing the superfamilies, subgroups, and families in the SFLD. The resulting matches (with their scores and expectation values) are displayed, along with a hyperlink for each match leading to a dynamically generated alignment of the query sequence to the multiple sequence alignment used to construct the hidden Markov model. The alignments produced highlight the conserved residues that participate in enzymatic catalysis, and provide links to literature references of the experiments through which the structure-function relationship was determined. Users can also view a query sequence in the context of the superfamily/subgroup/family in the form of a dendrogram generated using ClustalW's neighbor joining algorithm[30] and can view relevant structures associated with the multiple alignments.

Users can also browse the SFLD in multiple ways. Lists of all superfamilies and EFDs within any hierarchical level of a superfamily are available, as well as lists of all reactions, and structures. Users can easily navigate the SFLD hierarchy of superfamily, subgroup, family, and enzyme functional domain levels.

When a three-dimensional structure is available, a link is provided allowing users to open and view the structure in Chimera[21] with a single mouse-click. Methods of searching the database using three-dimensional coordinates and new representations of enzyme function are currently being developed.

3. Conclusion

Developing a resource to represent structure-function relationships that can be leveraged for biological discovery in the context of mechanistically diverse enzyme superfamilies has required us to address many of the issues involved in making any biological database, including dealing with multi-functional enzymes and grading the reliability of data. It has also presented some unique, domain-specific challenges in terms of data organization and representation, such as the implementation of a structure-function knowledge hierarchy that reflects the patterns of conservation in enzyme superfamilies, and the representation of enzyme function itself. The SFLD is a first attempt at addressing some of these challenges, and provides a computational resource for those investigating enzyme structure-function relationships for applications that range from determination of the function of a new protein to providing guidance for engineering a new function into an existing enzyme scaffold.

Acknowledgments

We thank John Gerlt for expert advice on the enolase and crotonase superfamilies, and Kinkead Reiling for his advice on the terpene cyclase superfamily. This work was supported by NIH R01-GM60595 and NSF DBI-0234768 to PCB, and NIH P41-RR01081 to TEF.

References

1. Watson JD, et al, *IUBMB Life* 55, 249 (2003).
2. Gerlt JA, Babbitt PC, *Genome Biol* 1, REVIEWS0005 (2000).
3. Gerlt JA, Babbitt PC, *Annu Rev Biochem* 70, 209 (2001).
4. Schmidt DM, et al, *Biochemistry* 42, 8387 (2003).
5. Babbitt PC, et al, *Biochemistry* 35, 16489 (1996).
6. Weininger D, *J. Chem. Inf. Comp. Sci.* 28, 31 (1988).
7. Wilmanns M, Priestle JP, Niermann T, Jansonius JN, *J Mol Biol* 223, 477 (1992).
8. Yanofsky C, Horn V, Bonner M, Stasiowski S, *Genetics* 69, 409 (1971).
9. Kirschner K, Szadkowski H, Henschen A, Lottspeich F, *J Mol Biol* 143, 395 (1980).
10. Cohn W, Kirschner K, Paul C, *Biochemistry* 18, 5953 (1979).
11. Palmer DR, et al, *Biochemistry* 38, 4252 (1999).
12. Copley SD, *Curr Opin Chem Biol* 7, 265 (2003).
13. O'Brien PJ, Herschlag D, *Chem Biol* 6, R91 (1999).
14. Jensen RA, *Annu Rev Microbiol* 30, 409 (1976).
15. Miller BG, Raines RT, *Biochemistry* 43, 6387 (2004).
16. Bork P, Koonin EV, *Nat Genet* 18, 313 (1998).
17. Brenner SE, *Trends Genet* 15, 132 (1999).
18. Abascal F, Valencia A, *Proteins* 53, 683 (2003).
19. Benson DA, et al, *Nucleic Acids Res* 32 Database issue, D23 (2004).
20. Dodevski I, Pegg, S. C.-H., Babbitt, P. C. (unpublished)
21. Pettersen EF, et al, *J Comput Chem* 25, 1605 (2004).
22. Green ML, Karp PD, *BMC Bioinformatics* 5, 76 (2004).
23. Osterman A, Overbeek R, *Curr Opin Chem Biol* 7, 238 (2003).
24. Reed JL, Vo TD, Schilling CH, Palsson BO, *Genome Biol* 4, R54 (2003).
25. Ashburner M, et al, *Nat Genet* 25, 25 (2000).
26. Boeckmann B, et al, *Nucleic Acids Res* 31, 365 (2003).
27. Steele CL, Crock J, Bohlmann J, Croteau R, *J Biol Chem* 273, 2078 (1998).
28. Webb EC, NC-IUBMB. *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. New York, NY: Academic Press (1992).
29. Eddy SR, *Bioinformatics* 14, 755 (1998).
30. Thompson JD, Higgins DG, Gibson TJ, *Nucleic Acids Res* 22, 4673 (1994).