Threading your way to protein function

Steven M Firestine, Andrew E Nixon and Stephen J Benkovic

It is still very difficult to determine the function of a protein from its sequence. One potential solution to the problem combines the concept of enzyme superfamilies with modern methods of protein structure prediction. Active-site templates can be used as search tools to identify new members of the superfamilies.

Address: The Pennsylvania State University, Department of Chemistry, 152 Davey Laboratory, University Park, PA 16802, USA.

Chemistry & Biology October 1996, **3**: 779–783

 ISSN 1074-5521

The ability to relate a protein's sequence to its structure and function is vitally important in biochemistry. This type of information is required to design intelligent site-directed mutagenesis strategies, to determine the evolutionary relationships between enzymes and to engineer novel proteins. However, perhaps the most important reason for relating sequence to function is the determination of the function of unknown but sequenced genes. Before we can benefit from the information being generated by the various genome sequencing efforts, we must be able to identify the function of any gene of interest. It has been estimated that as much as 60 % of human sequences will have no relation to currently sequenced genes. Thus, it is important to develop methods for determining gene function which do not rely on finding long stretches of homology. Researchers are currently developing genetic approaches to decipher gene function, but these methods are costly and labor intensive. A more efficacious approach would be to devise new computational methods that allow protein function to be determined from sequence data alone.

Recent advances in computing power have greatly simplified the investigation and analysis of a newly cloned gene. Sequence analysis of a new gene is typically accomplished using sequence alignment programs, such as BLAST, which search genetic databases to determine homology to existing genes [1]. Identification of homologous genes immediately suggests a possible function for the protein product; however, in cases where little or no similarity to known genes exists, sequence analysis must progress to other programs that attempt to ferret out other important pieces of information from the sequence. These programs (for example PROMOTIF) are aimed at identifying any of the conserved structural motifs that are found in proteins of a specific function, for example the β–α–β unit that binds ADP [2,3]. The list of motifs is growing quite rapidly as the number of solved structures increases and the computational methods to identify motifs are improved [4].

Once these search possibilities are exhausted, determining information about the protein of interest becomes a matter of *in vitro* rather than *in silico* biochemistry. To extend the computational methods used in sequence analysis, additional approaches are required. Here, we examine the recent developments in defining enzyme reaction families and outline the methods designed to solve the sequence–structure problem. We propose that it should be possible to define a structural fingerprint for enzymes that catalyze a specific chemical reaction. A number of groups are pursuing this goal, using the serine protease family of enzymes as a model system [5–8]. This work has verified that active-site architecture is distinct for any class of enzymes and that it is possible to determine a protein's function simply by comparing a very small region (the active site) with an existing database.

Enzyme superfamilies: structure and function

Enzymes have evolved over millions of years to catalyze reactions with high specificity and high rates of reaction. During this evolution, the structure of the enzyme is bound by nothing but overall stability and the limitations imposed by the nature of the transition state of the reaction that they catalyze. Given these constraints and the relatively limited set of protein folds, it seems probable that there will be only a limited set of structural and spatial arrangements for any given active site [9]. One consequence of this hypothesis is that enzymes that catalyze similar reactions must have similar active-site structures, and thus these enzymes can be classified as belonging to a particular superfamily. By superfamily we mean those enzymes that possess similar active sites and similar reaction mechanisms. Such enzyme superfamilies are well known; for example, the serine proteases and lipases use a common catalytic motif that hydrolyzes carboxylic acid derivatives using an active-site serine as a nucleophile [10,11]. Although these enzymes have identical active sites, they have distinct overall topologies.

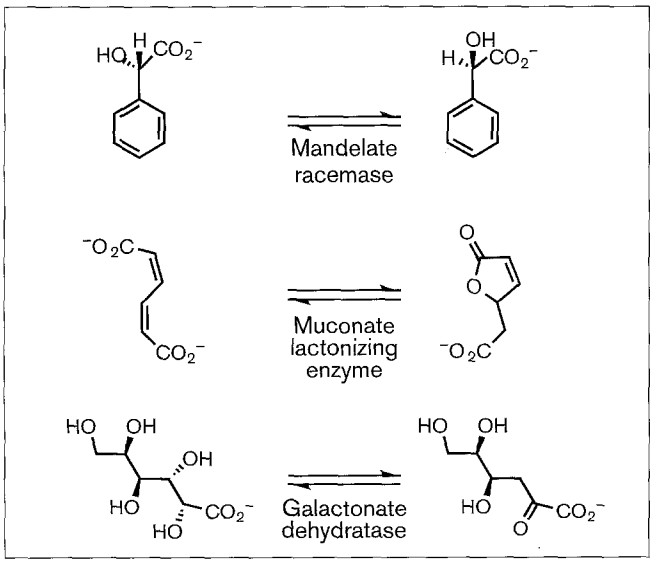
Another recently described example of an enzyme superfamily is the nucleotidyltransferases. Recent structural investigations of DNA polymerase β indicate that this enzyme is similar to kanamycin nucleotidyltransferase rather than the larger and more processive DNA and RNA polymerases [12]. This surprising conclusion suggests that DNA polymerase β and kanamycin nucleotidyltransferase have similar structures, because they both catalyze the same type of reaction. This example also indicates that the catalytic machinery is the most important conserved feature in enzyme evolution, as the substrate specificity of these enzymes is quite different.

The power of the concept of enzyme superfamilies for determining protein function was demonstrated in 1995 by Gerlt and colleagues, who identified the enzyme galactonate dehydratase (GalD), based upon sequence similarities to mandelate racemase (MR) (Fig. 1) [13]. The connection between GalD and MR, and an understanding of the mechanism of MR, led to a hypothesis regarding the catalytic activity of GalD. Cloning and analysis of GalD firmly established this enzymes as a member of the superfamily that catalyzes the abstraction of an α-proton from carboxylic acids [13]. This work demonstrated that it is possible to identify enzymes that catalyze similar reactions using the concept of superfamilies.

Structure–structure comparisons

For GalD and MR, traditional sequence analysis was important in determining the function of GalD and its membership in the superfamily. For other superfamilies, structural comparisons are necessary to determine membership in the family and evolutionary relationships between family members. Identification of the biotin carboxylase superfamily is a recent success in structural comparison (Fig. 2) [14,15]. Members of this superfamily catalyze the formation of an amide-like bond, with ATP activating the carboxylic acid. These enzymes all have a non-classical ATP-binding fold, classified as an ATP-grasp fold. The residues responsible for ATP binding in the four enzymes of this superfamily (biotin carboxylase, D-Ala:D-Ala ligase, glutathione synthetase, and succinyl CoA synthetase) have a low degree of sequence similarity. The implication from the similar structures and reaction mechanisms of these enzymes is that nature has found a particular spatial arrangement of the amino acids in these proteins that is optimal for the type of reaction that the proteins catalyze.

Figure 1



Reactions catalyzed by mandelate racemase, muconate lactonizing enzyme and galactonate dehydratase.

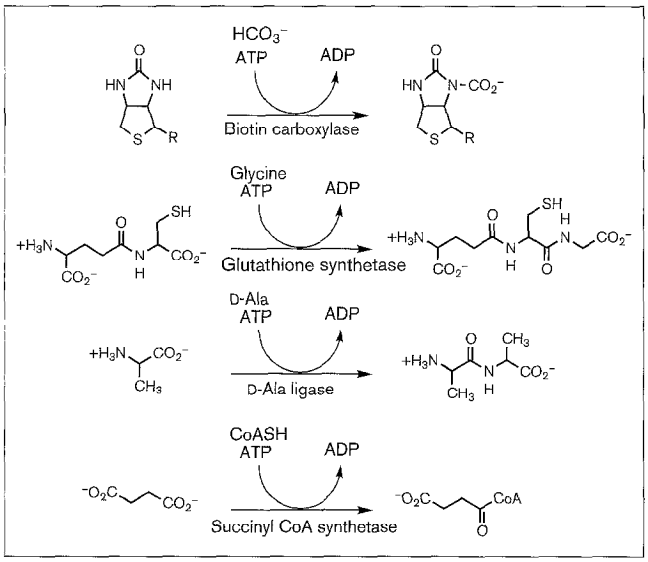
The elucidation of the biotin carboxylase superfamily was accomplished using a three-dimensional search algorithm that looked for similar shapes in whole proteins [14]. Recent structural comparisons have shown, however, that the active sites of enzymes in superfamilies are nearly identical, suggesting that overall structural comparisons may not be necessary. Thus, instead of complete three-dimensional comparisons, perhaps only structural templates of the active sites would be necessary for structure–structure comparisons.

This approach was described recently by Wallace *et al.* [5], using the serine protease family. This enzyme family is a good model system as the catalytic mechanism (hydrolysis of the peptide bond through the concerted action of a catalytic triad of Asp, His, and Ser) and the structures of the enzymes in this family have been well characterized [10]. In addition, there are numerous structures of serine proteases complexed with various ligands, so a template that can act as a generalized active site can be developed and used to test the technique [6,7]. A structural template specific for the Ser-His-Asp catalytic triad was derived from the structures of serine proteases and lipases contained in the protein structure database (PDB). The template was generated in the following way. First, all of the occurrences of both catalytic and noncatalytic Ser-His-Asp residues were extracted. Noncatalytic triads were then filtered out based upon RMS differences to the parent triad of α-lytic protease, leaving only the well-defined catalytic triads. Inspection of the templates created revealed that only the relative positions of the Ser and Asp sidechain atoms were important in governing the conformation of the catalytic triad. Thus, the positions of two sidechain oxygen atoms were taken as the initial trial three-dimensional template, which was used to screen a representative data set of structures in the PDB [16]. All known Ser-His-Asp catalytic triads were identified in this screen, confirming that a template-based search of three-dimensional structures, where the template is based upon active-site geometry alone, can identify enzymes that catalyze similar reactions.

Sequence–structure comparisons

The concept of enzyme superfamilies derives from the idea that the structure of the active site is defined by the type of chemical reaction that it catalyzes. In principle, if one could determine the structure of the active site from the sequence alone, then the function of the protein could be determined. Although this implies that the problem of protein folding must first be solved, we believe that if current computational methods were applied, active site template structures for each of the superfamilies could be derived from sequence data alone. The key elements of

Figure 2



Reactions catalyzed by biotin carboxylase, D-Ala:D-Ala ligase, glutathione synthetase and succinyl CoA synthetase.

this idea, which we have termed template-based functional determination, are the design and use of the template.

The template designed by Wallace *et al.* [5] focused on the relative positions of only two atoms in the catalytic triad, and as such represents the minimal structure that could be used to define an active site. Such a template is suitable for structure–structure comparisons, because most of the important information (the active-site geometry) is already contained within the crystal structure. When little is known about the three-dimensional structure of a protein, however, more information must be incorporated into the template to minimize the number of fortuitous matches to unrelated sequences. This can be achieved by including not only conserved residues but also conserved elements of secondary structure in the catalytic center. Thus, sequence matches can only be made by locating a particular residue within the active site on a defined structural element.

There are two complementary methods for determining if a particular sequence could adopt the conformation of the template, threading and inverse folding. Both methods rely upon the fact that many proteins have similar folds despite the absence of sequence homology. Threading is the alignment of an input sequence with a known three-dimensional structure; the method involves calculating the probability that a given residue will occur in a particular three-dimensional environment. A recent critical assessment of this technique, in which several groups attempted to predict the overall fold of newly solved structures, resulted in varying degrees of success [17]. In template-based searching, sequences would be threaded onto the active-site template rather than the whole structure. The threading would identify a number of proteins that are likely to contain an active site similar to the template and thus probably catalyze a reaction that is similar to that catalyzed by the parent enzyme.

Inverse protein folding addresses the problem from an alternative perspective. It seeks to determine all possible sequences that could fold into a given structure. This approach can be seen as an extension of threading, where a library of random sequences are threaded onto a target structure. The library of sequences generated by applying inverse folding algorithms can then be used as a search template for standard sequence–sequence comparisons. Inverse protein folding has been used successfully by Godzik [18] to identify members of the plastocyanin family. This technique is attractive because it effectively reduces the problem of whether a given sequence could adopt a particular conformation to that of a sequence comparison between the input sequence and a library containing all possible sequences that could fold into a structure that resembles the active site.

Theoretical steps necessary for template-based functional determinations

Before a template can be developed for an enzyme superfamily, important features of the active site need to be determined. These features can be identified by a combination of both structural analysis and traditional biochemical methods, such as site-directed mutagenesis. To illustrate these principles, we have chosen mandelate racemase as an example. Mandelate racemase catalyzes the interconversion of (*R*)- and (*S*)-mandelate (Fig. 1). This enzyme is a member of the recently described superfamily responsible for the abstraction of protons that are alpha to a carboxylic acid, and provides a representative model to outline the steps for template-based functional determination [13].

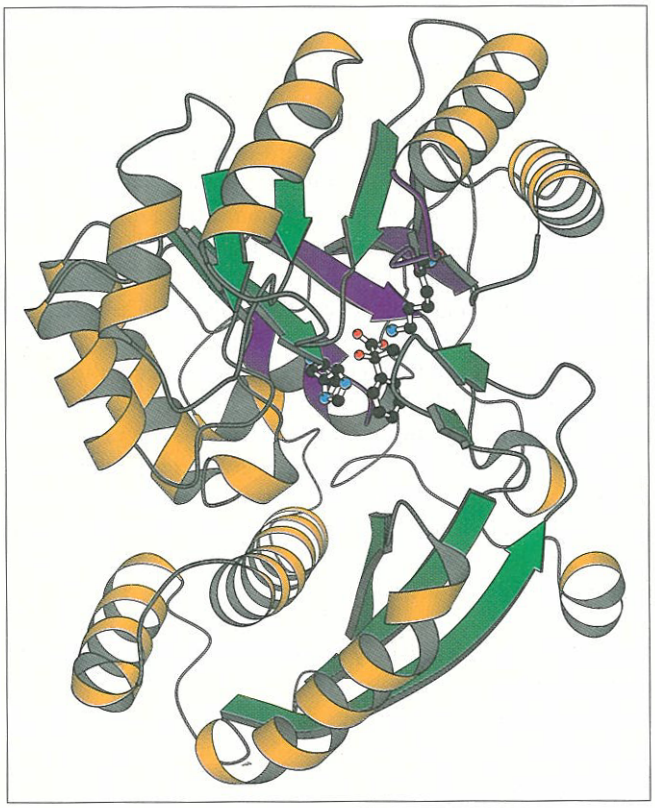
Step 1. Identification of the residues and secondary structures of the active site

The crystal structure (Fig. 3) of mandelate racemase shows the location of the active site, based on the presence of a covalent inhibitor, (*R*)-α-phenylglycidate [19]. Using this structure, the secondary structure of the active site can be identified (Fig. 3, blue color). Further inspection also reveals several amino acids responsible for either substrate binding or catalysis. The function of these amino acids has been investigated by site-directed mutagenesis, and they have been found to be important for catalysis [20,21]. In addition, similar amino acids have also been shown to be important for catalysis of muconate lactonizing enzyme, another member of the superfamily [22].

Step 2. Developing a template for the active site

The active site of mandelate racemase (Fig. 4) represents the staring point for the development of the template for

Figure 3



Structure of mandelate racemase. Shown in blue are the structural elements surrounding the active site as defined by the presence of the inhibitor α-phenylglycidate. Important amino acids are shown in detail as ball and stick representations. The figure was produced from the PDB file by the program MOLSCRIPT [23].

the superfamily. The template that is constructed must incorporate the features of this active site as well as those from the other members of the family. For example, mandelate racemase uses both lysine and histidine for catalysis, whereas muconate lactonizing enzyme requires only lysine. These differences are noted in the template along with any other conserved amino acids. Equally important is the secondary and tertiary structure of the active site. Thus, not only are the conserved amino acids noted, but also the type of secondary structure where these residues are located.

A simple inspection of the active site outlined in Fig. 4 indicates that it is composed of discontinuous pieces of secondary structure. This arises because the different geometrical elements of the active site come from different parts of the protein. This problem is circumvented by computationally connecting these pieces via loops of undefined length. The variable length allows for the presence of additional domains within the protein. Although the absence of any length restriction on these connecting loops may suggest that almost any protein can be made to adapt to the template, we feel that the likelihood that any given protein will possess both the required secondary structure and conserved amino acids is small.

Step 3. Using the template as a search tool

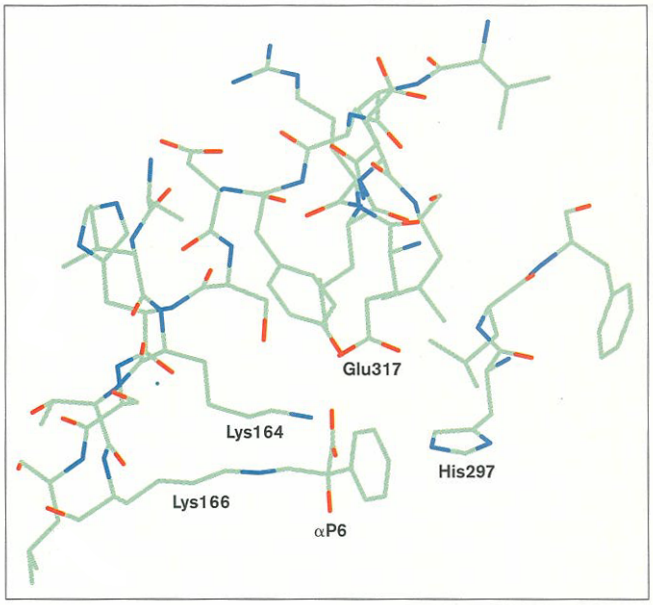
Once the template is formulated, either threading or inverse folding is used to determine which polypeptide sequences adopt the conformation of the active site. The proteins identified in this way may thus have the same catalytic function as the enzymes from which the template is derived.

Conclusions

We have suggested a way by which large numbers of sequences could be screened for potential function. This approach focuses on structural templates designed from the crystal structures of members of enzyme superfamilies. These templates would then be used in template-based sequence–structural determinations to investigate whether a particular sequence has the potential to adopt the structural features of the active site.

Currently, there is no single program that can accomplish the method outlined in this paper. However, many of the programs and algorithms needed as the basis for the

Figure 4



The active site of mandelate racemase, extracted from the complete structure. Important catalytic residues are indicated along with the position of the covalent inhibitor. Although the active site magnesium is important in catalysis, we feel that the ion is not required for the template to be functional, and therefore the magnesium is not shown. The following amino acids constitute the structure shown: 134–142, 162–166, 297–299, and 317–322. The figure was prepared using the program Quanta (Molecular Simulations).

method have already been developed. These programs need to be modified to increase speed and accuracy, as well as to handle discontinuous sequences such as those that would arise from the template. A database of active-site templates for the known superfamilies will also need to be constructed, a task which is not trivial. Once these difficulties have been overcome, template-based functional determinations should become a major tool in sequence analysis.

Acknowledgements

The authors would like to thank the Damon Runyon–Walter Winchell Foundation (S.M.F.) and the National Institutes of Health (GM13306, A.E.N. and S.J.B.) for funding.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
2. Hutchinson, E.G. & Thornton, J.M. (1996). PROMOTIF – A program to identify and analyse structural motifs in proteins. *Protein Sci.* **5**, 212–220.
3. Wiernega, R.K., Terpstra, P. & Hol, W.G.J. (1986). Prediction of the occurrences of the ADP-binding βαβ-fold in proteins, using amino acid sequence fingerprint. *J. Mol. Biol.* **187**, 101–107.
4. Bork, P. & Koonin, E.V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366–376.
5. Wallace, A.C., Laskowski, R.A. & Thornton, J.M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipase. *Prot. Sci.* **5**, 1001–1013.
6. Barth, A., Frost, K., Wahab, M., Brandt, W., Schlader, H.D. & Franke, R. (1994). Classification of serine proteases derived from steric comparisons of their active site geometry, part II: Ser, His, Asp arrangements in proteolytic and non-proteolytic proteins. *Drug Des. Dis.* **12**, 89–111.
7. Barth, A., Wahab, M., Brandt, W. & Frost, K. (1993). Classification of serine proteases derived from steric comparisons of their active sites. *Drug Des. Dis.* **10**, 297–317.
8. Fischer, D., Wolfson, H., Lin, S.L. & Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: Potential implications to evolution and to protein folding. *Prot. Sci.* **3**, 769–778.
9. Orengo, C.A., Flores, T.P., Taylor, W.P. & Thornton, J.M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
10. Blow, D.M. (1990). More of the catalytic triad. *Nature* **221**, 337–340.
11. Brady, L., *et al.*, & Menge, U. (1990). A model for interfacial activation in lipases from the structure of a fungal lipase–inhibitor complex. *Nature* **351**, 767–770.
12. Holm, L. & Sander, C. (1995). DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* **20**, 345–347.
13. Babbitt, P.C., *et al.*, & Gerlt, J.A. (1995). A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids. *Science* **267**, 1159–1161.
14. Artymiuk, P.J., Poirette, A.R., Rice, D.W. & Willet, P. (1996). Biotin carboxylase comes into the fold. *Nat. Struct. Biol.* **3**, 128–132.
15. Wolodko, W.T., Fraser, M.E., James, M.N.G. & Bridger, W.A. (1994). The crystal structure of succinyl-CoA synthetase from *Escherichia coli* at 2.5 Å resolution. *J. Biol. Chem.* **269**, 10883–10890.
16. Bernstein, F.C., *et al.*, & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
17. Lemer, C.M.-R., Rooman, M.J. & Wodak, S.J. (1995). Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins* **23**, 337–355.
18. Godzik, A. (1995). In search of the ideal protein sequence. *Protein Eng.* **8**, 409–416.
19. Neidhart, D.J., *et al.*, & Gerlt, J.A. (1991). Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2.5-Å resolution: Identification of the active site and possible catalytic residues. *Biochemistry* **30**, 9264–9273.
20. Kallarakal, A.T., *et al.*, & Kenyon, G.L. (1995). Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the K166R mutant. *Biochemistry* **34**, 2788–2797.
21. Mitra, B., *et al.*, & Kenyon, G.L. (1995). Mechanism of the reaction catalyzed by mandelate racemase: importance of electrophilic catalysis by glutamic acid 317. *Biochemistry* **34**, 2777–2787.
22. Neidhart, D.J., Kenyon, G.L., Gerlt, J.A. & Petsko, G.A. (1990). Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature* **347**, 692–694.
23. Kraulis, P.J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–960.