gram MULTISUP,[32] which performs weighted multiple superposition of the structures.

Availability

The SSAP package can be obtained by sending a request to orengo@bsm.bioc.ucl.ac.uk. Programs are distributed as binary files compiled for execution on a Silicon Graphics machine running UNIX.

[32] T. P. Flores, personal communication.
[33] T. P. Flores, D. S. Moss, and J. M. Thornton, *Protein Eng.* **7,** 31 (1993).

[37] Understanding Protein Structure: Using Scop for Fold Interpretation

By STEVEN E. BRENNER, CYRUS CHOTHIA, TIM J. P. HUBBARD, and ALEXEY G. MURZIN

Introduction

The structure of a protein can elucidate its function, in both general and specific terms, and its evolutionary history. Extracting this information, however, requires a knowledge of the structure and its relationships with other proteins. These two aspects are not independent, for an understanding of the structure of a single protein requires a general knowledge of the folds that proteins adopt, while an understanding of relationships requires detailed information about the structures of many proteins.

Fortunately, this complex problem with its intertwined requirements is not insurmountable, for two reasons. First, protein structures can be fundamentally understood in ways that most of their sequences cannot. The comprehensibility of protein structures derives from the relatively few secondary structure elements in a given domain and the fact that the arrangement of these elements is greatly restricted by physics and probably by evolution. Second, resources are now available to aid recognition of the relationships between protein structures. The structural classification of proteins (scop) database hierarchically organizes proteins according to their structures and evolutionary origin.[1] As such, it forms a resource that allows researchers to learn about the nature of protein folds, to focus their investi-

[1] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.* **247,** 536 (1995).

Copyright © 1996 by Academic Press, Inc.
All rights of reproduction in any form reserved.

gation, and to rely on expert-defined relationships when seeking new ones. In addition, automated methods can compare and recognize structures that are similar by some criteria and therefore provide a method for associating new experiments with the scop hierarchy.

Scop Hierarchy

An explanation of the classification of folds in scop provides a tutorial in the understanding of protein structures, and vice versa. Thus, when we discuss the way proteins are organized in scop, we intend to indicate simultaneously how one can examine any protein structure. Because the whole of scop is an extended collection of many thousand classifications, relatively few examples are provided here. Likewise, scop contains hundreds of references, and therefore the descriptions of the classification will note only highlights.

To illustrate the scop hierarchy, we have chosen proteins that bind NAD or NADP (Table I), in particular those whose NAD(P)-binding domains share the so-called Rossmann-fold (Table II). The discovery of this shared fold in 1975[2] raised the question that is fundamental to our understanding of protein structure and evolution: What are the origins of structural similarity in proteins whose sequences show no significant sequence similarity? The many structures for the NAD(P)-binding proteins which have been determined since that time provide us with the clear examples of divergent and convergent evolution highlighted here.

The scop database is organized on a number of hierarchical levels, with the principal ones being family, superfamily, fold, and class. Within the hierarchy, the unit of categorization is the protein domain, rather than whole proteins, since protein domains are typically the units of protein evolution, structure, and function. Thus, different regions of a single protein may appear in multiple places in the scop hierarchy under different folds or, in the case of repeated domains, several times under the same fold.

In scop, families contain protein domains that share a clear common evolutionary origin, as evidenced by sequence identity or extremely similar function and structure. Superfamilies consist of families whose proteins share very common structure and function, and therefore there are compelling reasons to believe that the different families (with low interfamily sequence identities) are evolutionary related. Folds consist of one more superfamilies that share a common core structure (i.e., same secondary structure elements in the same arrangement with the same topological

[2] M. G. Rossmann, A. Liljas, C.-I. Brändén, and L. J. Banaszak, *in* "The Enzymes" (P. D. Boyer, eds.), 3rd ed., Vol. 11, p. 61. Academic Press, New York, 1975.

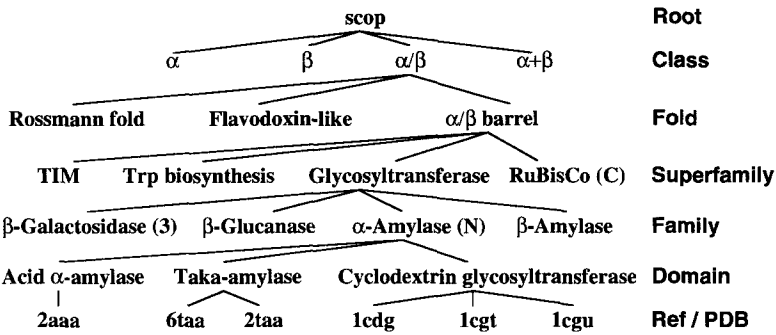

FIG. 1. Region of scop hierarchy. All the major levels, including class, fold, superfamily, and family, are shown. Also shown are individual proteins and, at the lowest level, either the Protein Data Bank (PDB) coordinate identifier or a literature reference.

connections). Frequently, proteins clustered together at this level will have considerable elaboration on the shared fold topology. Finally, folds are grouped into one of four classes depending on the type and organization of the secondary structure elements: all-α, all-β, α/β, and $\alpha+\beta$. In addition, there are several other classes for proteins that are very atypical and therefore difficult to classify. Figure 1 depicts a region of the scop hierarchy, with examples of items at all levels.

The following discussion describes how to classify a protein structure at each of the different levels given above, and to make use of that information.

Class

Before a protein structure can be properly classified, it needs to be divided into domains. The basic idea of a domain is simple: it is a region of the protein that has its own hydrophobic core and has relatively little interaction with the rest of the protein, so that it is essentially structurally independent. In practice, however, identification of domains is not a trivial task and can frequently be done correctly only by using evolutionary information to see, for example, how domains have been "shuffled" in different proteins. Typically domains are colinear in sequence, which aids their identification, but occasionally one domain will have another "inserted" into it, or two homologous domains will intertwine by swapping some topologically equivalent parts of their chains. Because of the problem of identifying domains on the basis of a single protein structure, it is usually best to iteratively refine domain definitions: first make a tentative set of assignments and, as understanding of the structure grows, refine as necessary. An application of this approach to the domain definition in NAD(P)-binding proteins is discussed in following sections.

Fortunately, the first real step in interpreting a domain structure, namely, placing it in the appropriate class, is usually a straightforward task. It should be readily apparent whether a domain consists exclusively of α helices, β sheets, or some mixture thereof. One caveat is that only the core of the domain should be considered: it is possible for an all-β protein to have very small adornments of α or 3_{10} helix, whereas so-called all-α structures may actually have several regions of 3_{10} helix, and in rare cases, small β sheet outside the α-helical core.

Domains with a mixture of helix and sheet structures are divided into two classes, α/β (alpha and beta), and $\alpha+\beta$ (alpha plus beta). The α/β domains consist principally of a single β sheet, with α helices joining the C terminus of one strand to the N terminus of the next. Commonly these proteins are divided into two subclasses: in one subclass, the β sheet is wrapped to form a barrel surrounded by α helices; in the other, the central sheet is more planar and is flanked on either side by helices. Domains that have the α and β units largely separated in sequence fall into the $\alpha+\beta$ class. Because the strands in these structure do not have the intervening helices, they are typically joined by hairpins, leading to antiparallel sheets much as are found in all-β class folds. However, $\alpha+\beta$ structures may have one, and often a small cluster, of helices packing tightly and integrally against the sheet.

When the four classes were originally defined on the basis of a handful of structures,[3] the distinction between the α/β and $\alpha+\beta$ classes was very clear. Since then, the picture has become more cloudy, and although most mixed-structure domains can be clearly placed in one class or the other, some domains defy easy classification. In the event that a protein with mixed secondary structure does not clearly fall into one of these two classes, it is best to hold this decision in abeyance and proceed to the more important task of identifying other structures with the same fold.

In addition to the four classes of globular protein structures, scop contains a few other classes. Most important is the multidomain class. Proteins here have multiple domains that would ordinarily be placed in different classes. However, the different domains of these proteins have never been seen independently of each other, so accurate determination of their boundaries is not possible and perhaps not meaningful or useful. Another important class contains the many small proteins having structures stabilized by disulfide bridges or by metal ligands rather than by hydrophobic cores. Membrane proteins frequently have unique structures because of their unusual environment, and therefore they also have their own class. The scop

[3] M. Levitt and C. Chothia, *Nature* (*London*) **261,** 552 (1976).

database also contains classes for the short peptides, theoretical models, and nonproteins (such as nucleic acids) in the Protein Data Bank.

Fold

Identification of the fold of a protein is one of the most difficult stages of classification, so much so that papers about new structures often fail to report structural similarity with other proteins. The problem is in part due to the fact that there are more than just a handful of different folds used in nature; about 50 are currently known for each of the four classes of globular proteins. The best way to characterize a fold is to look first at the major architectural features, and then identify more subtle characteristics, as described below for NAD(P)-binding proteins.

In addition, there is a shortcut that will aid identification of the fold of a large number of proteins. This is because there are about a dozen folds, such as the β/α barrel, that occur very frequently (as identified by the number of superfamilies they contain), so first comparing a structure of interest with each of these will often be expedient. For example, at least three such folds—the β/α barrel (currently contains 11 superfamilies), the SH3-like β barrel (6 superfamilies), and the ferredoxin-like $\alpha+\beta$ fold (16 superfamilies)—contain NAD(P)-binding proteins (Table I).

So far no all-α protein or domain that binds NAD or NADP has been found, but in the structure of catalase, currently classified as multidomain, the NADP molecule binds between an all-α and a β-barrel domain. The other known NAD(P)-binding structures are distributed between the three other major classes: all-β, $\alpha+\beta$, and α/β. Apart from the ADP-ribosylation toxins, which utilize NAD as substrate, these proteins are oxidoreductases, which use NAD or NADP as cofactor. The different folds and classes into which these enzymes fall clearly indicate multiple origins of their similar function. Nevertheless, their catalytic sites can be very similar. A particularly striking example of functional convergence can be seen in the active-site structures of NAD(P) oxidoreductases of the β/α-barrel fold and a family of the Rossmann-fold domains, which have their substrate-binding cavities and the catalytically essential tyrosine and lysine residues similarly located relative to the cofactor. There is also convergence of a different kind. The NADP-binding domain of ferredoxin reductase and related enzymes is topologically similar to the Rossmann-fold domain, and this similarity extends to the locations of the coenzyme-binding sites. However, because the protein structures do not superimpose well and the binding modes are very different, these proteins cannot be classified as having the same fold or superfamily (see below). Topological similarities of this kind are on intermediate level between class and fold, and, in the current version of scop, they

are silently indicated by listing folds with similar topologies together on the class page. This approach is also used to segregate different architectural motifs, like two-sheet sandwiches and single-sheet barrels in the all-β class. Future versions of scop will include the necessary additional levels of classification to make such distinctions explicit.

Superfamilies

Protein structures classified in the same superfamily are probably related evolutionarily, and therefore they must share a common fold and usually perform similar functions. If the functional relationship is sufficiently strong, for example, the conserved interaction with substrate or cofactor molecules, the shared fold can be relatively small, provided it includes the active site. This is in contrast with classification on the fold level, which ordinarily requires greater structural similarity.

We have already mentioned that NAD(P)-binding domains are classified in scop in several distinct superfamilies. The largest of these superfamilies includes all the original members of the Rossman-fold, hence its full name, the NAD(P)-binding Rossmann-fold domain (Tables I and II). All members of this superfamily bind the cofactor molecule in the same way; that is, they will be positioned identically when the whole protein structures are superimposed. This superfamily also includes a domain of succinyl-CoA synthetase that binds a different cofactor, coenzyme A (CoA), but shows good structural similarity and recognizes the common part of CoA

TABLE I
NAD(P)-BINDING PROTEINS IN SCOP DATABASE

Class	Fold[a]	Superfamily
All-β	SH3-like	R67 dihydrofolate reductase
α/β	β/α barrel	NAD(P) oxidoreductases
	—	FAD (also NAD)-binding motif
	—	NAD(P)-binding Rossmann-fold domain[b]
	—	Ferredoxin reductase-like, C-terminal domain
	—	Dihydrofolate reductases
	—	Isocitrate and isopropylmalate dehydrogenase
$\alpha+\beta$	Ferredoxin-like	HMG-CoA reductase, N-terminal domain
	—	ADP-ribosylation toxins
Multidomain	—	Heme-linked catalases

[a] If there is only one superfamily in the fold, a — is shown.
[b] The families of this superfamily are shown in Table II.

TABLE II
SUPERFAMILY OF NAD(P)-BINDING ROSSMANN-FOLD DOMAINS[a]

Family and protein	Specific features
Tyrosine-dependent oxidoreductases Short-chain dehydrogenases Dihydropteridin reductase UDP-galactose epimerase Enoyl-ACP reductase	Extensive structural similarity; coenzyme-binding and catalytic site are in one domain; rare left-handed $\beta\alpha\beta$ unit in extension of superfamily fold
Lactate and malate dehydrogenases Lactate dehydrogenase Malate dehydrogenase	Sequence similarity, extensive structural similarity; C-terminal catalytic domain has an unusual $\alpha+\beta$ fold
Alcohol dehydrogenase Alcohol dehydrogenase Glucose dehydrogenase Quinone reductase	Extensive structural similarity; N-terminal catalytic domain has a GroES-like all-β fold
Formate dehydrogenase Formate dehydrogenase D-Glycerate dehydrogenase Phosphoglycerate dehydrogenase	Extensive structural similarity; catalytic domain is formed by N- and C-terminal regions and has common flavodoxin-like α/β fold
Glyceraldehyde-3-phosphate dehydrogenase Glyceraldehyde-3-phosphate dehydrogenase Glucose-6-phosphate dehydrogenase Dihydrodipicolinate reductase	Common $\alpha+\beta$ fold in the catalytic domain, inserted in coenzyme domain in same topological location
6-Phosphogluconate and acyl-CoA dehydrogenases 6-Phosphogluconate dehydrogenase Acyl-CoA dehydrogenase	Superfamily fold is similarly extended; common all-α fold in the catalytic C-terminal domain

[a] The superfamily of NAD(P)-binding Rossmann-fold domains consists of a number of families. In addition to the fold and the cofactor-binding mode common to all members of the superfamily, there are some family-specific similarities in either sequence, structure, or domain organization.

and NADP molecules in very similar way. The scop classification differs from a traditional classification of the nucleotide-binding Rossmann-fold which contains all proteins that show topological (but not close structural) similarity and bind nucleotides, often in a very different way.

The NAD(P)-binding Rossmann-fold domain is well defined by its shared fold and the coenzyme-binding site. It is usually conjoined with another domain that contains the catalytic site of the whole enzyme. The catalytic domain may precede or follow the coenzyme domain, or it may interrupt or be interrupted by this domain. The protein folds of catalytic domains in different enzymes may be very different and fall in all four major classes and in a number of different folds (Table II).

Families

Most scop families cluster together homologous proteins with high sequence similarity. The structures of these proteins are also very similar (e.g., lactate and malate dehydrogenases). However, in extraordinary cases, extensive structural similarity and strong functional relationships can define families in the case of low sequence similarity, as in the case of the tyrosine-dependent oxidoreductases.[4] This family seems to be the largest family of NAD(P)-dependent enzymes, as evidenced by sequence similarity among its members. It could be further divided into a number of subfamilies, according to the extent of their sequence similarity.

A small number of scop families currently embrace the relationships which are above the standard family definition but below the superfamily level. It can be suggested that proteins that have similar domain organization and share a common fold in the catalytic domain, like dihydrodipicolinate reductase and the glyceraldehyde-3-phosphate and glycose-6-phosphate dehydrogenases, are likely to be even closer related than those sharing the common fold in their coenzyme domain only. Such similarities currently lead to coenzyme domains being assigned to the same family and the catalytic domains to the same family or superfamily, depending on the extent of structural similarity of the catalytic domains.

With future releases of the scop database, additional levels of classification around the family level will be introduced so that each will have a unique, well-defined meaning.

Role of Automated Systems

Computational approaches are now beginning to play an important role in the understanding of protein structures and can be fruitfully used in conjunction with the scop classification. At present there are two particularly valuable types of programs, both described in detail elsewhere in this volume.

Sequence comparison is a simple and reliable way of learning about the structural and evolutionary relationships of a protein. Two proteins with 40% sequence identity with each other will have very similar structures, and if a sequence has 30% identity to a protein of known structure, then an outline of its configuration can also be deduced. If there is significant similarity between a sequence and a protein in scop, then that sequence can be put into the appropriate family, which then defines its superfamily, fold, and class.

[4] L. Holm, C. Sander, and A. Murzin, *Nat. Struct. Biol.* **1**, 146 (1994).

The major limitation of sequence comparison is that it fails to identify many of the structural relationships in scop either because the sequence relationship has become too weak (for evolutionarily related proteins) or never existed (for evolutionarily unrelated proteins with similar folds). Structure–structure comparison programs use various methods to recognize similar arrangements of atomic coordinates and thus identify domains of similar structure. Although these methods lack complete accuracy, they can be used to suggest a shared fold between a protein of interest and others in scop. Manual inspection must then be used to verify the choice of fold and to select an appropriate superfamily. The selection of superfamily is the most challenging and most scientifically rewarding step of protein classification, for it ascribes a biological interpretation to chemical and physical data. For this reason, the assignment of all proteins of known structure to evolutionarily related superfamilies is perhaps the single most powerful and important feature of the scop database.

Resources

The scop database can be accessed on the World Wide Web, at the URL http://scop.mrc-lmb.cam.ac.uk/scop/. For improved access, mirrors of scop are available worldwide, and their addresses can be found from the above location.

[38] Detecting Structural Similarities: A User's Guide

By MARK BASIL SWINDELLS

Introduction

If I were to tell you that I had just discovered a new similarity between two structures, which consisted of a single equivalenced α helix, you would probably conclude that my future contributions to the field were likely to be limited. However, if I were to tell you that I had detected an unanticipated similarity in which 700 residues could be superimposed and that intriguing functional similarities were retained, you would, I would hope, be impressed. In reality, most similarities revealed by a typical database search fall between these two extremes and as a result it is frequently difficult to know whether they merely reflect the features of proteins per se, in which helices and strands are packed in a compact manner, or whether they have some additional functional or even evolutionary meaning.

Copyright © 1996 by Academic Press, Inc.
All rights of reproduction in any form reserved.