Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Is protein classification necessary? Toward alternative approaches to function annotation

Donald Petrey and Barry Honig

The current nonredundant protein sequence database contains over seven million entries and the number of individual functional domains is significantly larger than this value. The vast quantity of data associated with these proteins poses enormous challenges to any attempt at function annotation. Classification of proteins into sequence and structural groups has been widely used as an approach to simplifying the problem. In this article we question such strategies. We describe how the multifunctionality and structural diversity of even closely related proteins confounds efforts to assign function on the basis of overall sequence or structural similarity. Rather, we suggest that strategies that avoid classification may offer a more robust approach to protein function annotation.

Address
Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, United States

Corresponding author: Honig, Barry (bh6@columbia.edu)

Current Opinion in Structural Biology 2009, **19**:363–368

This review comes from a themed issue on
Sequences and Topology
Edited by Anna Tramontano and Adam Godzik

Available online 5th March 2009

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2009.02.001

Introduction

Protein classification schemes codify various relationships between groups of proteins that are often based on global sequence or structural similarities. However, the complex nature of evolutionary relationships between proteins raises questions about the possibility of generating a reliable classification [1]. The function of a protein is not easy to define and difficulties in describing it can occur at all levels of a classification hierarchy, even when an unambiguous sequence relationship is evident. The problem is exemplified by the so-called 'moonlighting' proteins [2,3], which can play multiple roles in the cell. Such proteins have been able to acquire new functions with only minimal changes in either sequence or structure. Conversely, it has become increasingly apparent that proteins can undergo significant changes in sequence and/or structure while still maintaining the same or a similar function [4•,5•]. The picture that is emerging of protein sequence/structure/function space is one of multiple and complex relationships that in many ways defy classification.

One example of the problem involves the organization of proteins into discrete categories based on their 'fold'. Structural alignments have revealed numerous geometric relationships between local fragments of proteins that have been classified, globally, as belonging to different folds. Such observations have led to the view that protein structure space is continuous rather than discrete ([6–8,9•], D Petrey, M Fischer, B Honig, **Structural and functional relationships between proteins with different global topologies suggest a dynamic approach to function annotation**. *Proc Natl Acad Sci U S A*, unpublished data). The importance of this issue is amplified by observations that proteins can have different global topologies, and hence be assigned different fold classifications, but can still share a biological function [10•]. Indeed that evolutionary relationships might exist between proteins that have been classified differently can be inferred from the very existence of so many different folds and topologies (over 1000 topologies in CATH and folds in SCOP). As has been recently pointed out, it is unlikely that each of these folds appeared independently but, rather they probably evolved from a smaller, less diverse set of ancestral proteins [11•]. This in turn suggests that there are numerous functional relationships between proteins with seemingly unrelated structure. The potential existence of such relationships implies that there is a richness of diverse information in structural databases that has yet to be uncovered. This article provides specific examples and suggests a general strategy for mining this information.

Mechanisms for the evolution of structurally diverse proteins with common functions

A conceptual problem that arises in understanding the origins of structural diversity is that most mutations are neutral in the sense that they do not usually cause structural changes while those that do would generally be expected to result in disordered structures (e.g. see [12]) and hence would be expected to also abrogate function. However, evidence has recently accumulated implying that the standard mechanisms of evolution, ranging from point mutations, to large deletions, insertions, or rearrangements of secondary structure elements (SSEs) can result in structural diversity while maintaining function. That small changes in sequence can lead to large changes in structure has been known for some time

and there have been recent reports of this phenomenon using both designed [13] and naturally occurring proteins [14,15]. At the extreme are 'chameleon' sequences which can adopt multiple conformations depending on factors such as oligomeric state [16] or temperature [17].

The conservation of the sequence and structure of a functional site combined with structural diversity in other regions provides a simple mechanism to conserve function but not global structure. The α/β and all-α ferredoxins offer an apparent example of this mechanism. Despite a structural similarity that only involves a pair of helices surrounding the functional Fe–S cluster, an evolutionary relationship between these two groups of proteins [18] is suggested both by their sequence similarity and by the existence of both types of ferredoxins as individual domains in the protein dihydropyrimidine dehydrogenase.

Insertion, deletion, and rearrangement of individual domains are other mechanisms for the evolution of novel functions [19,20]. A related mechanism can occur within domains, for example the rearrangement of structural fragments that consist of a number of SSEs that play a functional role [4•,5•,10•] and would also account for proteins with different global topologies and related functions. A possible objection to the idea is that significant changes within a domain would be expected to be highly destabilizing. However, recent experimental evidence suggests that this is not an issue. Graziano *et al.* [21] created a library of DNA elements representing random SSEs from proteins with known structure. Members of the library were randomly combined, producing several stably folded proteins one of which was highly homologous to the protein aspartate racemase from *Polaromonas* sp. In another example, individual members of the family of cobaltochelatases still preserve function even after significant deletions, insertions, duplications, and substitutions of multiple SSEs [22]. Finally, Peisajovich *et al.* [23] proposed and experimentally verified a mechanism involving circular permutation in a DNA methyltransferase which preserves function even after significant rearrangement.

Analysis of protein structures has also suggested evolutionary mechanisms by which different structures can share a common function. For example, it has been shown that proteins with different topologies can evolve from a conserved structural core [4•,5•,10•,24•]. Moreover, analyses of individual functional families [5•] and of homologous superfamilies in CATH [4•] have shown that structural changes associated with the evolutionary mechanisms summarized above are in fact quite common and often occur while still preserving overall function. An important example is provided by the evolution of heteromeric protein–protein interactions via the duplication of genes that are involved in homodimeric interactions [25]. It was found that in the evolution of heteromeric proteins, the general location of the interface seen in homodimers was conserved. This suggests that the prediction of protein–protein interactions sites may benefit from the exploitation of apparently remote structural relationships.

Classification can obscure functional relationships

The observation of functional relationships between proteins that have been classified as structurally unrelated provides some of the most striking consequences of the evolutionary mechanisms summarized in the previous section. A recent example is the identification of an evolutionary path relating the P22 and phage λ Cro proteins. Both proteins are transcription factors and both contain a DNA-binding helix–turn–helix motif but they have very different global topologies; P22 Cro is an all-α protein and λ Cro is an α/β protein. The two proteins have a low level of sequence identity (25%) and could not be unambiguously related based on sequence alone. However, a relationship has been established through the identification of a series of proteins with sequence similarities (>40%) using transitive sequence searches [26•,27•]. Classifications that would place the two Cro proteins in different categories would clearly obscure the relationship between them. Related studies have been reported by Lupas and coworkers [28,29] who demonstrated an evolutionary connection between proteins in different folds. As in the Cro protein example, they identified functional relationships that involve only a few SSEs, in one case a small βαβ fragment containing conserved residues [24•].

We have recently discussed other cases where common structural fragments in proteins with different overall topologies play a similar functional role ([9•], D Petrey, M Fischer, B Honig, **Structural and functional relationships between proteins with different global topologies suggest a dynamic approach to function annotation**. *Proc Natl Acad Sci U S A*, unpublished data). For example, Figure 1 represents the SSEs of three sugar-binding proteins. Each is classified as a different fold in SCOP and a different architecture in CATH (a 'jelly roll', a 'β-prism', and a 'β-propeller'). Nevertheless, they all contain a common substructure consisting of a four-stranded and three-stranded sheet with identical connectivity. Moreover, the overall locations of sugar-binding sites on the surface of this substructure are also conserved. An evolutionary relationship between these proteins is also suggested by their more general biological function shown in Figure 1B. Each protein plays a similar role in distinct but related pathways: the jelly-roll facilitates viral entry into bacterial cells and the β-propeller facilitates bacterial entry into eukaryotic cells. Moreover, both of these activities are mediated through interactions with sugar-modified proteins on the cell surface. Although the