Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Template-based prediction of protein function

Donald Petrey, T Scott Chen, Lei Deng, Jose Ignacio Garzon, Howook Hwang, Gorka Lasso, Hunjoong Lee, Antonina Silkov and Barry Honig

We discuss recent approaches for structure-based protein function annotation. We focus on template-based methods where the function of a query protein is deduced from that of a template for which both the structure and function are known. We describe the different ways of identifying a template. These are typically based on sequence analysis but new methods based on purely structural similarity are also being developed that allow function annotation based on structural relationships that cannot be recognized by sequence. The growing number of available structures of known function, improved homology modeling techniques and new developments in the use of structure allow template-based methods to be applied on a proteome-wide scale and in many different biological contexts. This progress significantly expands the range of applicability of structural information in function annotation to a level that previously was only achievable by sequence comparison.

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

Corresponding author: Petrey, Donald (dsp18@columbia.edu)

Current Opinion in Structural Biology 2015, **32**:33–38

This review comes from a themed issue on **Sequences and topology**

Edited by **Anna Panchenko** and **Madan Babu**

http://dx.doi.org/10.1016/j.sbi.2015.01.007

Introduction

It has been estimated that less than 1% of sequences in current sequence databases have an experimentally verified function [1] and, realistically, this situation is unlikely to change. Computational approaches offer the only viable solution to this problem. Numerous methods continue to be developed to infer protein function, most commonly based on sequence similarity, the presence of certain small sequence motifs, evolutionary history, and genomic location. Many of these methods are automatic and the best of them outperform simple orthology transfer, that is, annotation transfer based on the best PSIBLAST hit [2].

Three-dimensional structure information generally plays only a minor role in automated methods but of course is invaluable in the manual annotation of the function of individual proteins. The overall limited use of protein structural information is due in large part to the small number of protein structures available relative to the numbers of sequences. However, this situation is changing and homology modeling is currently making structural information available for large numbers of proteins [3]. Moreover, it has been shown that modeled proteins can be effectively used to annotate function [4–7].

Structure-based methods for function annotation can be based on the properties of the structure of a given protein itself, such as the presence of surface cavities, surface patches containing evolutionarily conserved or covarying sets of residues, or biophysical features such as electrostatic potentials [8]. Here we focus exclusively on so-called 'template'-based approaches, in which the function of a protein is assigned based primarily on its similarity to other proteins whose function is known. The wide applicability of such approaches is highlighted by the observation that, in general, there will be at least one, and usually several, proteins in structural databases that carries out a similar function using a mechanism similar to a query protein of interest [9–12]. This suggests that there are many new directions where protein structural information can be applied and, most significantly, used on a genome-wide scale [13•].

Templates are used in several ways in function annotation. Given a 'query' protein with unknown function, a database of templates is searched for structurally similar proteins based on different metrics such as global sequence or structural similarity or local similarity of protein substructures. Whether the query has a function similar to the template is then evaluated by looking for similarities and differences sequence, geometric or biophysical features after superposing the query and template structures. By similar function we typically mean similar interaction properties (e.g., 'these two proteins interact' or 'this protein binds molecules of a certain type at this location'), but methods are also being developed to predict more specific functions such as enzyme class. Below we discuss recent progress in template-based function annotation. Although many of the methods are not new, their combination, especially in the context of machine learning approaches, is a recent development that has significantly expanded the role of structure in protein function annotation.

Exploiting global structural similarity

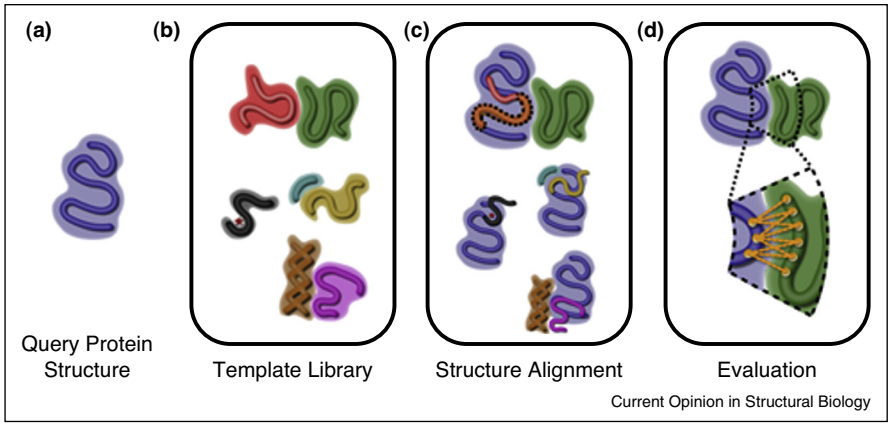
The general strategy involved in using templates to identify binding properties of a given query is to search a database of protein complexes to identify those where one member of the complex (the template) shares some global similarity (both close and remote) with the query (Figure 1). The query and interacting partner of the template are placed in the same coordinate system using the transformation that structurally aligns the query and template structures, at which point it is necessary to determine the likelihood that an interaction will occur. The interaction partner can correspond to another protein, a peptide, a nucleic acid or a small molecule.

Global similarity can be sequence or structure based. In sequence-based approaches if two query proteins are homologous to two other proteins that form a complex of known structure, the query proteins are first superimposed on their respective homologs in the complex. The likelihood of the query proteins forming a complex can be assessed using scoring schemes based on different factors such as overall sequence similarity, sequence similarity limited to the predicted interface [14] or sequence and structural similarity combined with biophysical properties of the predicted interface [15–17]. Interfacial residues in the query proteins are defined as those that align to interfacial residues in their respective templates. We use the term interaction model (Figure 1) to define the method used to score the putative interaction. This can range from an energetic analysis of the full three-dimensional structure to just a sequence analysis of interfacial residues.

In other structure-based approaches, templates are identified based solely on geometric similarity to the queries rather than on sequence similarity. Geometric similarity in principle enables a much broader coverage since there are many cases where structural and functional relationships are not detectable by sequence. The limited number of protein structures that have been determined experimentally limits the scope of this approach but homology modeling significantly expands the ability to exploit structural relationships. As an example, an experimental structure is available for at least one domain in about a quarter of the human genome but this number increases to about two thirds if homology models are used [13•]. The uncertain accuracy of many homology models may have limited their use in geometric alignments but we believe that they have in fact been underused. This is because, at the very least, they contain important information about a protein's fold which can in turn be used to identify proteins with similar structures that may be functionally related.

We recently developed a structure-based approach to predict whether two proteins interact which relies heavily on homology models to provide extensive structural coverage of genomes [13•]. If the two putative interaction partners have one or more domains whose structure is found in the PDB or homology model databases, these structures are used to identify geometrically similar proteins. If any pair of these forms a complex of known structure, this complex is used to create an interaction model of the two query proteins. The confidence score for a modeled interaction is based on overall structural similarity, the quality of the alignment in the interfacial region and the nature of the predicted interfacial residues. These scores are combined using Bayesian statistics. Although the use of homology models generates less

Figure 1



Function annotation using a template library. The structure of a query protein **(a)** is used to scan a library of templates with known function **(b)**. Templates can be proteins with various binding partners including other proteins (green), peptides (teal), RNA/DNA (brown) or small molecules (red star). For each complex in the library, the query, template and binding partner are placed in the same coordinate system by superposing the template and query based on global or local similarity (**(c)** dotted line). An interaction model is then created which defines the parameters used to determine whether the query has functional properties similar to the template. These can range from an estimate of the physical interaction energy derived from residues interactions (**(d)** yellow lines) in a 3-dimensional model of the interface, properties such as sequence conservation and covariation in the interface, or other features used as input to machine learning approaches.