Review

CellPress

Leveraging structure for enzyme function prediction: methods, opportunities, and challenges

Matthew P. Jacobson[1,2], Chakrapani Kalyanaraman[1,2], Suwen Zhao[1,2], and Boxue Tian[1,2]

[1] Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, CA 94158, USA
[2] California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA 94158, USA

The rapid growth of the number of protein sequences that can be inferred from sequenced genomes presents challenges for function assignment, because only a small fraction (currently <1%) has been experimentally characterized. Bioinformatics tools are commonly used to predict functions of uncharacterized proteins. Recently, there has been significant progress in using protein structures as an additional source of information to infer aspects of enzyme function, which is the focus of this review. Successful application of these approaches has led to the identification of novel metabolites, enzyme activities, and biochemical pathways. We discuss opportunities to elucidate systematically protein domains of unknown function, orphan enzyme activities, dead-end metabolites, and pathways in secondary metabolism.

The challenge of protein function assignment

The rapid advances in genome-sequencing technology have created enormous opportunities and challenges for defining the functional significance of encoded proteins. Although the number of genome sequences continues to grow rapidly, experimentally verified functional annotations lag well behind and are growing at a slower pace. As of May 2014, the UniProtKB (TrEMBL and Swiss-Prot) database contained 56 010 222 sequences, but only 545 388 sequences (~1%) are listed in Swiss-Prot, the manually annotated and reviewed portion of UniProtKB [1,2], where experimental information about function is reported. High-throughput bioinformatics methods are clearly needed to bridge this gap, but many significant challenges remain for reliably predicting the functions of proteins using the most common approaches, which are based primarily on transferring the relatively small number of experimentally determined functions to large collections of proteins based on sequence similarity. The rates of misannotation in the major repositories of protein sequence information, such as GenBank and TrEMBL, are unknown but estimated to be large [3,4].

One fundamental challenge is that there is no universal criterion sufficient to determine when a pair of proteins are likely to have the same or different functions; even if two proteins are highly homologous to one another and have similar structures, a change of only a few residues in the active site can change the functional specificity [5]. A second fundamental challenge is that annotation transfer, by definition, cannot identify new, uncharacterized protein functions. These challenges have motivated the development of diverse approaches to protein functional characterization and prediction. Such approaches use additional types of information beyond protein sequence, such as high-throughput metabolomics [6], RNA profiling [7–9], proteomics [10,11], and phenotyping experiments [12], and orthogonal types of bioinformatics information, such as genome organization (operons and gene clusters; domain fusions) and metabolic systems analysis [13].

Glossary

Homology modeling: a computational technique that builds an atomic model of a target protein using its sequence and an experimental 3D structure of a homologous protein (called the 'template'). The quality of a homology model depends on the accuracy of the sequence alignment between target and template, which varies (loosely) with the sequence identity (roughly speaking, pairwise identity higher than 40% is ideal, and lower than 25% is poor).
Ligand docking: a computational technique that predicts and ranks the binding poses of small molecule ligands to receptors (e.g., proteins). Docking usually comprises a sampling method that generates possible binding poses of a ligand in a binding site, and a scoring function that ranks these poses. Most scoring functions are empirical, and give only a crude estimate of the binding free energy of a ligand.
Secondary metabolism: biochemical pathways to produce organic molecules (i.e., secondary metabolites) that are not absolutely required for the survival of the organism. There are five particularly prevalent classes of secondary metabolite: isoprenoids, alkaloids, polyketides, nonribosomal peptides, and ribosomally synthesized and post-translationally modified peptides. Secondary metabolites are often restricted to a narrow set of species and have important ecological roles for the organisms that produce them. Many secondary metabolites are bioactive (antibacterial, anticancer, antifungal, antiviral, antioxidant, anti-inflammatory, antiparasitic, antimalaria, cytotoxic, etc.) and have been used as drugs and drug leads.
Structural genomics: an effort to determine the 3D, atomic-level structure of every protein encoded by a genome through a combination of high-throughput experimental and modeling approaches. The determination of a protein structure though a structural genomics effort often precedes knowledge of its function, motivating the development of methods to infer function from structure.

Corresponding author: Jacobson, M.P. (matt.jacobson@ucsf.edu).
Keywords: enzyme function prediction; protein structures; homology modeling; docking; metabolic pathways.

0968-0004/
 http://dx.doi.org/10.1016/j.tibs.2014.05.006

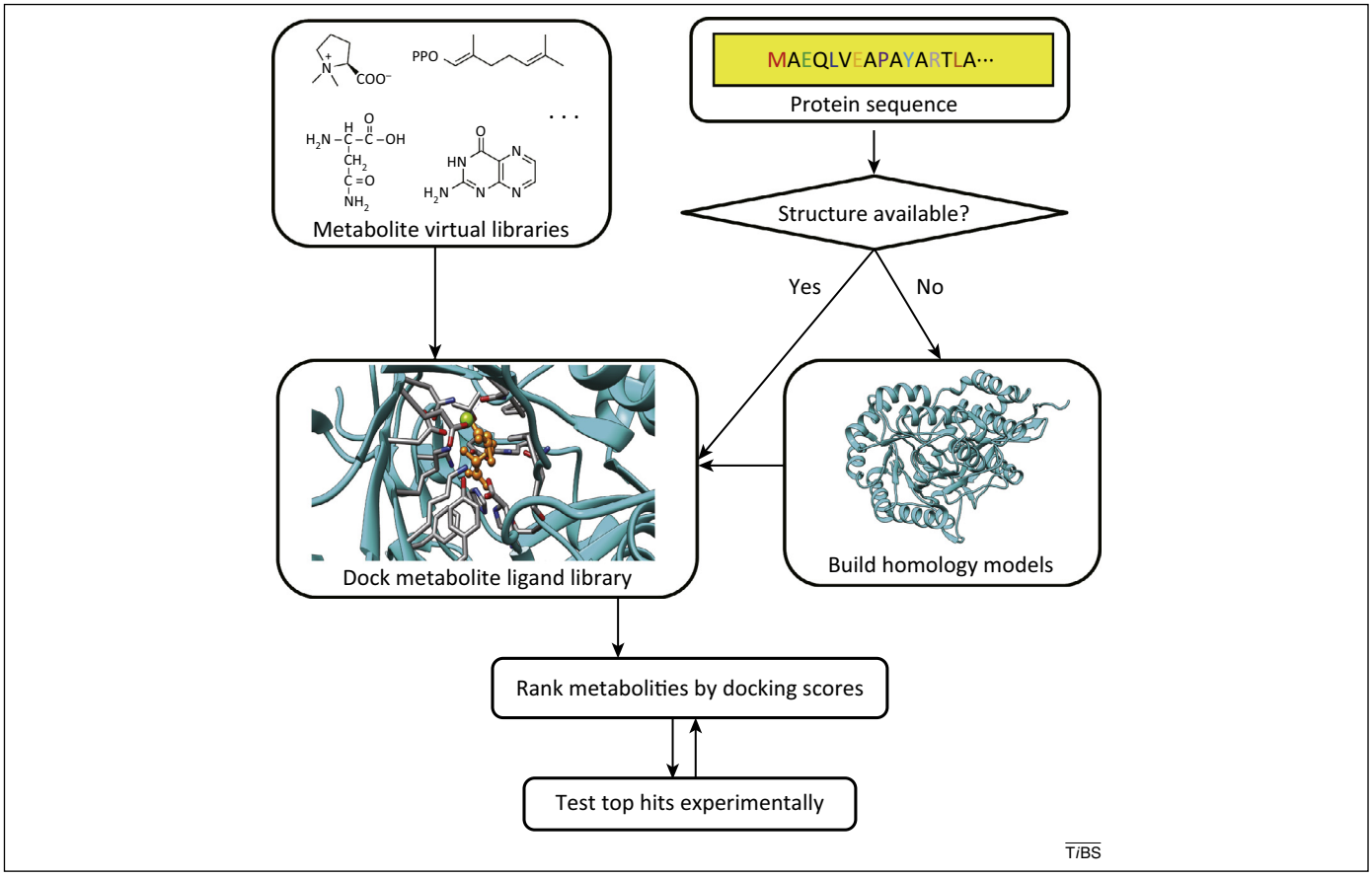

Figure 1. Structure-based virtual metabolite docking protocol for enzyme activity prediction. When no structure has been experimentally determined for a protein sequence, a model can be built using a variety of comparative modeling methods, but only when the structure of a homologous protein is available that has approximately 30% of greater sequence identity to the protein of interest. Whether using a structure of a model, it is critical that active site metal ions and cofactors are present, and that catalytic residues are positioned appropriate for catalysis. Virtual metabolites libraries can be constructed and 'docked' against the putative active sites of structures or models using computational tools more commonly used in structure-based drug design (e.g., Glide or DOCK). The docking scoring functions can be used to rank the ligands according to their estimated relative binding affinities. Top-scoring metabolites are typically inspected for plausibility (Is the predicted binding mode compatible with catalysis? Is the metabolite likely to be present in the relevant organism?), and then selected for experimental testing (*in vitro* enzymology). Protocols similar to that shown here have been used in retrospective and prospective studies [22–25,27–33,36,39].

In this review, we focus on the use of protein structure, in conjunction with other types of information, to aid function assignment, including the determination of novel functions and pathways. Structural information has been used to help elucidate many aspects of function, including protein–protein interactions (e.g., scaffolding) and regulation, but our focus here is biochemical function; that is, the determination of enzymatic activities *in vitro* and *in vivo*.

Using structure to infer small molecule binding

From structure to function

Structural genomics (see Glossary) efforts have generated a large number of structures for proteins with uncertain function. In the case of enzymes, these structures can be used to make inferences about function, either qualitatively, through inspection by an expert, or in more quantitative and automated ways. One class of methods generates functional hypotheses based on physicochemical similarity of the putative active site to the active sites of structurally and functionally characterized enzymes [14–18]. A second class of methods exploits computational tools developed primarily for computer-aided drug design to predict the substrates, products, or intermediates of an enzyme. Specifically, the strategy comprises docking an *in silico* metabolite library against an enzyme active site and experimentally testing the top-ranking metabolites to determine *in vitro* biochemical activity (Figure 1). Two excellent reviews are available describing the algorithms used in docking programs and their limitations [19,20], including their highly approximate treatment of key forces driving binding, such as electrostatics, solvation, and entropy losses. Although such algorithms have been extensively benchmarked and demonstrated their practical utility for computer-aided drug design, significant effort was required to test docking for enzyme-substrate recognition, resulting in various modifications to improve performance in this application [21–34]. Many metabolites are more highly charged than typical drug-like molecules; one successful approach for metabolite docking uses molecular mechanics-based scoring functions that treat electrostatics and solvation in a more realistic (and computationally expensive) [21,35]. Shoichet and coworkers introduced the concept of docking 'high energy intermediates' rather than substrates or products of enzymes, and demonstrated that this approach improved the ability to predict the binding mode of metabolites, and the ability to distinguish true substrates from false positives [30,36].