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

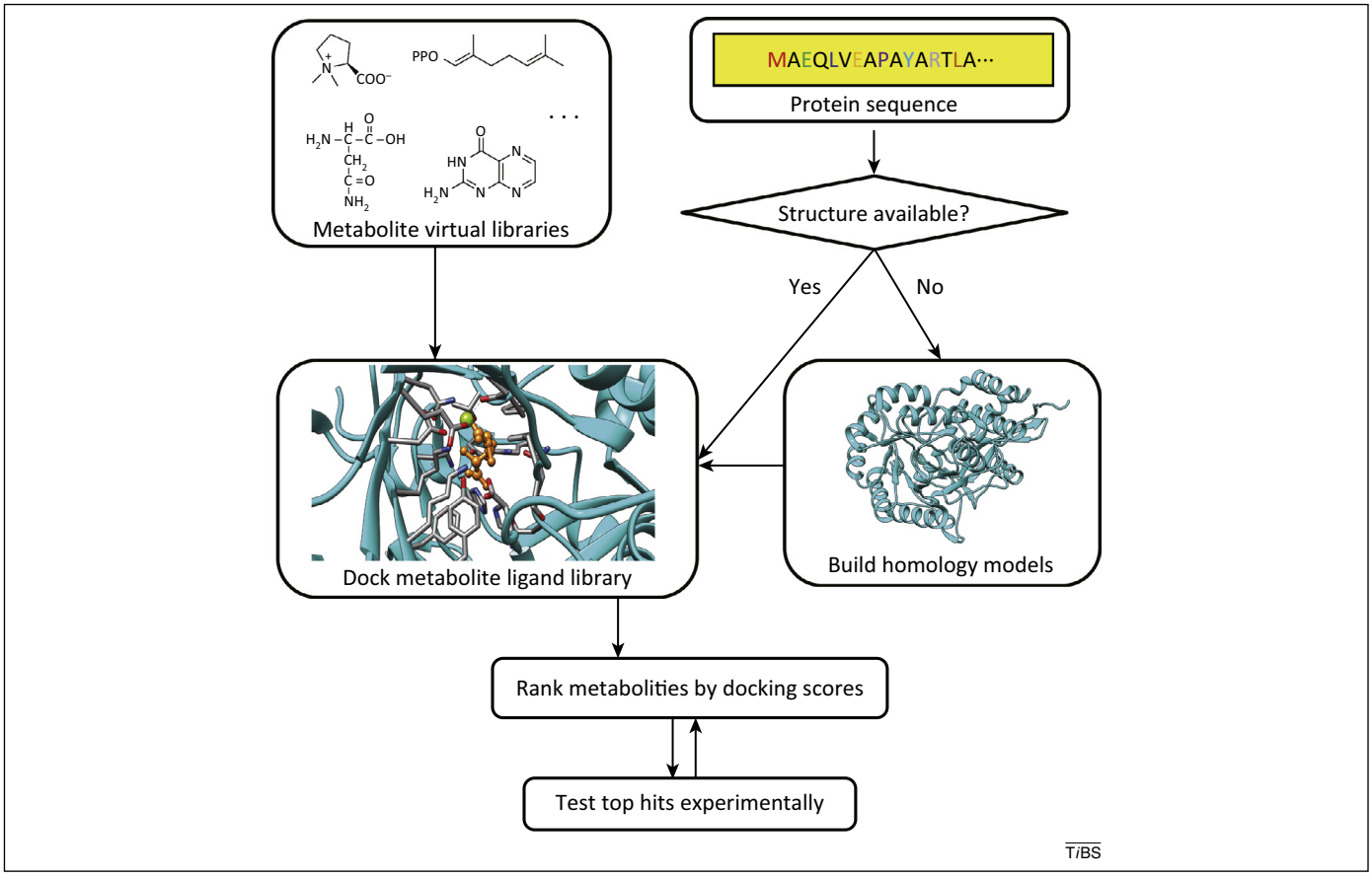

Figure 1. Structure-based virtual metabolite docking protocol for enzyme activity prediction. When no structure has been experimentally determined for a protein sequence, a model can be built using a variety of comparative modeling methods, but only when the structure of a homologous protein is available 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].

Even with these methodological improvements, there are numerous caveats to this approach, both fundamental and practical. A fundamental limitation is that docking methods can, at best, predict binding interactions, which is necessary but not sufficient for a ligand to be the substrate of an enzyme. In practice, experimental testing of top hits from metabolite docking frequently reveals many false positives, including weak substrates with very poor k_{cat} (but reasonable K_M); that is, metabolites that bind to the enzyme but are not efficiently turned over [27].

An important practical limitation of metabolite docking is that existing databases of metabolites are incomplete. A second practical limitation is that the structures used for docking must have ordered active sites, including any metal ions. However, it is possible to predict relatively small conformational changes associated with ligand binding, especially in side chains [37].

Another limitation for molecular mechanics-based scoring functions is that the electronic structures of transition states cannot be accurately described. In principle, combining quantum mechanics and molecular mechanics methods (QM/MM) can provide more accurate analysis of the mechanisms and specificities of enzymes. A proof-of-concept study has shown that such an approach may become practical for studying certain challenging aspects of enzyme specificity, compared with the more common use of quantum mechanical methods to investigate reaction mechanisms [38]. In the future, this type of approach may be particularly important when studying enzymes with intermediates that are radicals [e.g., P450 enzymes and radical *S*-adenosylmethionine (SAM) enzymes]. However, such calculations are currently prohibitively expensive to be used on a large scale.

Despite these limitations, metabolite docking has been shown to be useful in practice for generating testable hypotheses about function, which have proven to be correct in many cases. Herman *et al.* [30,36] and Fan *et al.* [28,29,39] docked the high-energy intermediates of metabolites and successfully predicted deaminase activity in several functionally uncharacterized enzymes of the amidohydrolase superfamily. Favia *et al.* [22] examined the ability of docking to identify cognate substrates of enzymes belonging to the short chain dehydrogenases/reductases superfamily. In several of these studies, subsequently determined co-crystal structures with metabolites confirmed the binding mode predicted by docking [23,24,27,32,33].

From sequence to function using homology models

Structural information can also be leveraged to help infer enzymatic function for proteins lacking structures. Although homology modeling remains imperfect [40], models have been successfully used to infer aspects of function in a many cases, including models of proteins based on the structures of proteins with which they have relatively low sequence identity (30% or lower); examples involving metabolite docking are discussed below [27,28,33]. The leverage of a single structure can be large; on average, each new structure determined by structural genomics efforts could be used to create models for hundreds or thousands of homologous sequences [41]. Pre-computed homology models can be obtained from databases such as SwissModel [42] and ModBase [43], which contain models for millions of protein sequences.

One of the simplest approaches to infer aspects of enzyme function, when no structure is available, is to identify putative active site residues in protein sequences by sequence alignment to proteins with solved structures. Changes in critical active site residues can suggest changes in the enzymatic reaction (e.g., changes in catalytic amino acids) or specificity. Constructing homology models can provide additional information about the predicted 3D arrangement of active site residues. Catalytic and other critical active site residues are frequently well conserved across homologs, facilitating accurate sequence alignment and, hence, the accuracy of the models, for regions surrounding the active site; nonetheless, allowing some degree of receptor flexibility in the docking protocol can be helpful to address small errors in, for example, side chain positioning [24,33,37].

Homology models have been used to predict accurately the substrate specificity of enzymes in the enolase (Figure 2B,C) and isoprenoid synthase (Figure 2D) superfamilies [24,25,27,33]. In each case, a structure of the enzyme was subsequently determined that confirmed the predicted binding mode, and *in vitro* enzymology confirmed that the ligands were proficient substrates. The examples in Figure 2C,D are taken from studies in which predictions were made for dozens of enzymes [27,33], using homology models constructed based on template structures with sequence identities as low as 25%. That is, it is straightforward to automate the process of creating multiple homology models, all based on a particular template structure, for a series of homologous proteins in a multiple-sequence alignment and then dock against all of them [28].

Finally, certain X-ray crystal structures can be used to help identify small molecule ligands in a complementary fashion. Almo and coworkers have estimated that 3–5% of all structures determined by the New York Structural Genomics Research Consortium contain organic ligands from the expression host that survived purification [44]. Unassigned electron density, at sufficient resolution, can be sufficient in some cases to infer the nature of the substrate, although determining the mass of the metabolite by mass spectroscopy provides a useful constraint. This type of detective work led Almo and coworkers to discover a novel metabolite, carboxy-*S*-adenosyl-L-methionine, and a pathway that uses it to modify RNA [44]. In cases where the identity of the ligand remains ambiguous, metabolite docking may provide a useful way of identifying ligands that match the electron density and are predicted to have favorable binding interactions [45,46].

Although the number of protein structures is smaller than the number of protein sequences inferred from genome sequencing, and will undoubtedly remain so, a variety of complementary approaches has emerged to utilize these structures to make inferences concerning enzymatic function. Currently, experimental testing remains essential, but the computational approaches can help guide the design of experiments, and focus attention on enzymes likely to have novel or unexpected activities. In favorable

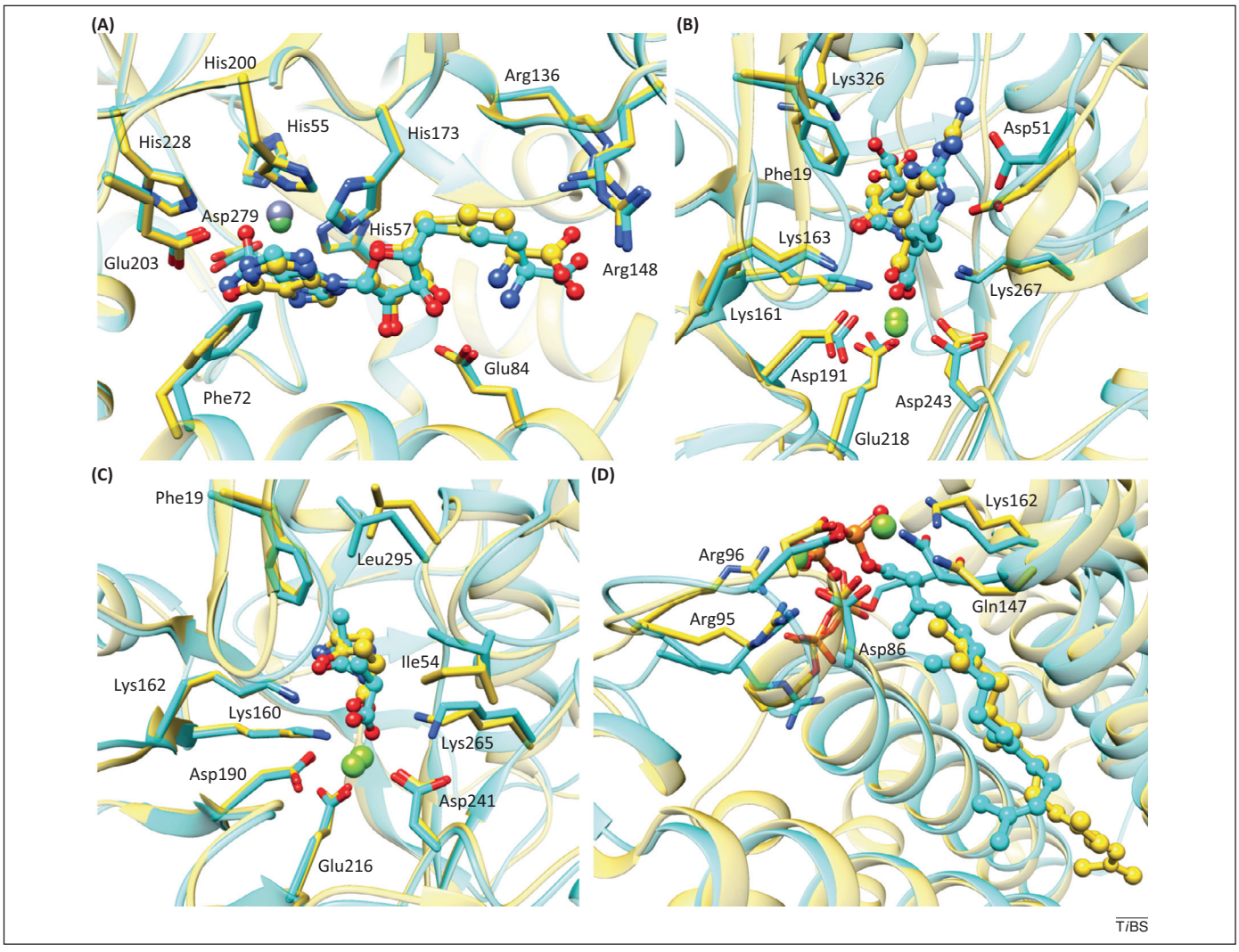

Figure 2. Predicted binding poses are in good agreement with subsequently determined experimental structures. Predicted ligand binding mode (cyan) superimposed with the X-ray crystal structure (gold) of: **(A)** *S*-adenosylhomocysteine deaminase (PDB: 2PLM); **(B)** *N*-succinyl-L-Arg racemase (PDB: 2P8C); **(C)** D-Ala-D-Ala epimerase (PDB: 3Q4D), and **(D)** a polyprenyl synthase (PDB: 4FP4). In (B–D), the docking predictions were made using homology models based on crystal structures with 35%, 39%, and 29% sequence identity, respectively.

cases, homology modeling can be used to extend the use of structure-based methods to large numbers of proteins lacking experimental structures. A major challenge is automating the metabolite docking methods, which remain technically complex; the Metabolite Docker web resource (http://metabolite.docking.org/) [47], and its application to metabolite docking, represents important progress in this direction.

Structural information in the context of pathways

As we have shown, a single structure (or model) of an enzyme can be used to make testable predictions concerning its potential substrate(s). However, *in vitro* activity does not, by itself, necessarily imply *in vivo* biochemical function. When enzymes can be placed into pathways or networks, additional information is available for predicting both *in vitro* and *in vivo* biochemical function.

In prokaryotes and certain eukaryotes, enzymes involved in pathways are frequently located in close proximity on the genome. In some cases, functionally related proteins also appear in certain organisms as gene fusions. A family of genome context analysis techniques takes advantage of these observations to infer functional relations among genes, even when they do not share sequence similarity. These techniques have been exploited by databases such as Metacyc [48], MicrobesOnline [49], STRING [50], SEED [51], and IMG [52]. Although genome proximity is not a useful source of information for most eukaryotes, other types of experiment, such as interactome mapping by mass spectroscopy or other methods [53,54], can be used in an analogous manner; that is, to develop hypotheses concerning proteins that have related functions.

Structural genomics efforts have added a structural perspective to biochemical pathways in certain organisms. The Joint Center for Structural Genomics has determined the structures of over 100 enzymes in the central metabolism of *Thermotoga martima*, and created homology models for hundreds of others [55]. In less well-studied organisms, it would be rare to find entire pathways for which each enzyme has been structurally characterized, but as in the case of *T. maritime*, it is frequently possible to

create models for multiple enzymes in a putative pathway. In this context, metabolite docking can be expanded to pathway docking; that is, metabolite docking against multiple structures or models of proteins hypothesized to participate in a metabolic pathway or network [26,32]. In addition to increasing potentially the *in vivo* relevance of the results, docking metabolites to multiple binding sites in the same pathway can also increase the reliability of *in silico* predictions of substrate specificity because the pathway intermediates are chemically similar even if the proteins involved are structurally unrelated. Put simply, the product of one enzyme is the substrate for another enzyme, and comparing the metabolite docking results can help to refine hypotheses concerning the individual protein functions as well as the overall pathway.

Pathway docking was first introduced by Kalyanaraman and Jacobson to 'predict' retrospectively the intermediates in the glycolysis pathway in *Escherichia coli* [26]. In this proof-of-concept study, a large and diverse *in silico* metabolite library derived from Kyoto Encyclopedia of Genes and Genomes (KEGG) was docked against structures and homology models of ten enzymes in the glycolysis pathway. The ranks of the 'correct' substrates were all within the top 1% of the hit list, and in six out of ten cases, cognate substrates were ranked within the top 0.3%, that is, among the top approximately 50 ligands.

Zhao *et al.* performed a prospective application of the pathway docking method, which led to the discovery of new enzymes in the hydroxyproline betaine/proline betaine metabolism pathways (Figure 3) [31,32]. The initial focus was an uncharacterized member of the enolase superfamily, HpbD, the *apo* structure of which was determined in a structural genomics effort. The genome contexts are similar for HpbD and its putative orthologs in approximately 20 organisms, suggesting a conserved pathway, and homology models could be created for many of these (Figure 3). Metabolite docking against the structure and several homology models suggested that the pathway involved catabolism of amino acid derivatives, especially *N*-modified proline derivatives. A model of a periplasmic binding protein encoded by a gene located close to HpbD was particularly informative and suggested that the binding site contained a cation-π cage comprising three Trp side chains (Figure 3); docking results strongly suggested that the cation would be a quaternary amine, specifically a betaine (*N*-trimethylated amino acid). The combined results led to the prediction of catabolic pathways for proline betaine and *trans*-4R-hydroxyproline betaine (both are important osmolytes in marine organisms), with HpbD performing inversion of stereochemistry at the Cα position [31,32]. Subsequent *in vitro* enzyme assays and *in vivo* metabolomics experiments confirmed

(A)

Name	TrEMBL annotation	Closest PDB homolog
HpbA	2,4-dienoyl-CoA reductase	28% 1PS9
HpbB	Putative aromatic-ring hydroxylating dioxygenase, α-subunit	62% 3N0Q
HpbC	Possible dioxygenase reductase subunit	25% 2PIA
HpbD	Mandelate recemase/muconate lactonizing enzyme	100% 2PMQ
HpbE	Oxidoreductase, NAD-binding site	-
HpbF	Aldehyde dehydrogenase	50% 4LIH
HpbG	Dihydrodipicolinate synthetase	33% 3NA8
HpbH	ABC glycine betaine/L-proline transporter, ATPase	39% 1Z47
HpbI	ABC glycine betaine/L-proline transporter, permease	44% 3L6G
HpbJ	ABC glycine betaine/L-proline transporter, substrate binding protein	48% 1R9L
HpbR	Transcriptional regulator, LysR family	28% 3KOS

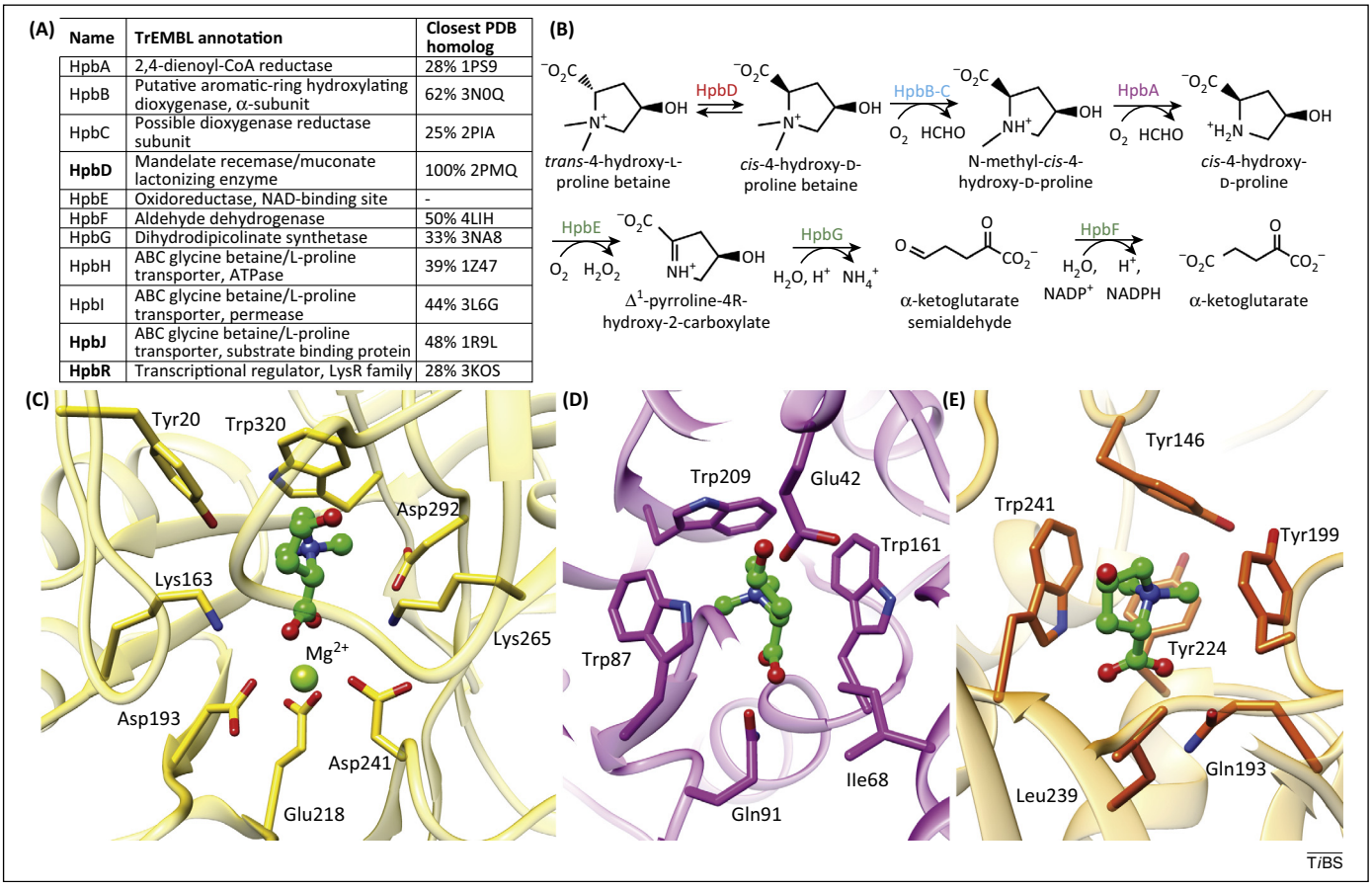

Figure 3. Structure-guided discovery of new enzymes in a novel hydroxyproline betaine metabolism pathway. **(A)** shows the name, TrEMBL annotation, and most similar homolog in the Protein Data Bank for each protein in the pathway. The automated TrEMBL annotations are incorrect or imprecise for all proteins in the pathway. However, there is rich structural information that can be used for modeling and docking, as shown in the closest PDB homolog column. The pathway is shown in **(B)**. **(C–E)** show the binding site and/or active site of the three proteins [HpbD, HpbJ, and HpbR, shown in bold in (A)] in the pathway, respectively, along with the docking-predicted binding mode for the ligand *trans*-4-hydroxy-L-proline betaine (ball-and-stick, green color). Both HpbJ and HpbR have a predicted cation-π cage, known for binding quaternary amines. In HpbD, two catalytic residues (Lys163 and Lys265) replace aromatic residues, leaving Trp320 as the key aromatic residue forming a cation-π interaction with the substrate.

these predictions and elucidated aspects of the regulation of these pathways.

Challenges and opportunities

No single computational or experimental approach alone is likely to 'solve' the problem of predicting or determining the functions of the millions of currently uncharacterized enzymes, especially for the most challenging goal of identifying novel enzymatic activities and biochemical pathways. However, the combination of sequence-based (bioinformatics) and structure-based computational methods, together with high-throughput protein expression, enzyme assays, crystallography, metabolomics, phenotyping, and potentially many other approaches, can provide powerful approaches to generate and evaluate hypotheses. A major challenge and opportunity is the development of methods to optimally combine these disparate types of computational and experimental data to make functional inferences. Even in the context of pathway docking, functional inferences have thus far been made with the aid of human knowledge and intuition, but certain aspects of the data integration can certainly be automated and systematized. The scope of the potential applications of these integrated approaches is vast, and we highlight a few opportunities here.

Biosynthetic pathways for natural products

Natural products, such as polyketides, nonribosomal peptides, isoprenoids, alkaloids, and ribosomally synthesized and post-translationally modified peptides, are structurally diverse secondary metabolites, many of which have biological activity and are used in modern medicine (erythromycin, vancomycin, taxol, morphine, duramycin, etc.). The biochemical pathways that create these natural products represent a challenge for function prediction because the chemical space is enormous; that is, the number of possible intermediates and end products of pathways in secondary metabolism is almost limitless. Moreover, the experimental characterization of the structures of these secondary metabolites is often challenging due to frequently complex ring structures and stereochemistry. For these reasons, the elucidation of the biosynthetic pathways of these high-value secondary metabolites remains challenging, even when the genome of the producing organism has been sequenced; for instance, only a small fraction of the tens of thousands of known alkaloids have their biosynthetic pathways fully elucidated [56,57].

One area of rapid progress has been the prediction of templated biosynthetic pathways for polyketides and nonribosomal peptides, due to the modular nature of the biosynthetic enzymes and their frequent occurrence in large gene clusters or operons. Sequence–structure–function relations have been well characterized for certain classes of enzyme in these pathways, such as polyketide synthases and nonribosomal peptide synthetases [58–60]. This knowledge has been harnessed in efforts to achieve combinatorial biosynthesis of novel polyketides and nonribosomal peptides [61–63]. However, elucidating the biosynthetic pathway of nontemplated natural products, such as isoprenoids and alkaloids, remains nontrivial.

Isoprenoid biosynthesis pathways present both opportunities and challenges with respect to function prediction [64,65]. In the biosynthesis of isoprenoids, isoprene units (C_5) are assembled by polyprenyl transferases to give long-chain terpenes such as geranyl pyrophosphate (C_{10}), farnesyl pyrophosphate (C_{15}), geranylgeranyl pyrophosphate (C_{20}), and squalene (C_{30}), which can then be converted into diverse carbon skeletons by terpenoid synthases (also called terpene cyclases), which are sometimes further modified by other enzymes, such as SAM-dependent methyl transferases. A paradigmatic isoprenoid pathway, the biosynthesis of cholesterol, is illustrated in Figure 4; the crystal structures of key enzymes in the pathway have been solved, including farnesyl pyrophosphate synthase [gold; Protein Data Bank (PDB): 1RQI], squalene synthase (light blue; PDB: 3WEG), and oxidosqualene-lanosterol cyclase (magenta; PDB 1W6K).

It is relatively straightforward to leverage structural information to predict the product specificities of the polyprenyl synthases. Product chain length has been shown to be determined primarily by the size of the cavity, and Wallrapp *et al.* [33] have shown that it is possible to predict chain-length specificity for sequences lacking structures through a combination of homology modeling and docking. By contrast, predicting the product specificity of isoprenoid synthases is challenging, because the number of possible products is enormous, and the enzymes must bind and stabilize several carbocations and transition states leading to a given product [66]. Despite these challenges, the potential impact of elucidating the sequence–structure–function relations of isoprenoid synthases is high, given the importance of these enzymes in the biosynthesis of complex, bioactive natural products and drugs.

Domains of unknown function

A high-value subset of functionally uncharacterized proteins is 'domains of unknown function' (DUFs). As the name suggests, no function is known for any member of a DUF protein family; thus, annotating even a single member of a DUF can have a large impact, by defining (in the case of enzymes) aspects of the biochemical capabilities. In Pfam 27.0, 26% (3885 out of 14 831) Pfam families are DUFs, with 'unknown function' or 'uncharacterized protein' in their descriptions [67]. Structures are available in the PDB for proteins in 379 DUF families (as of 2013) [68].

The potential impact of the systematic, structure-guided study of DUFs is suggested by the recent work of Bastard *et al.* [69], who determined that the DUF849 Pfam family contains β-keto acid cleavage enzymes of diverse substrate specificity. In this work, 14 novel *in vitro* enzymatic activities of the DUF849 Pfam family have been revealed through an integrated strategy, combining bioinformatics analysis to cluster the protein sequences and structural analysis using both crystal structures and homology models. The structural analysis was primarily qualitative (e.g., whether the substrate is neutral, positively, or negatively charged) but also supported by metabolite docking. High-throughput enzymatic screening confirmed many of the predictions and resulted in discovery of *in vitro* activities for 80 enzymes, including several novel functions; remarkably rapid progress for a protein family that was, until recently, entirely uncharacterized.

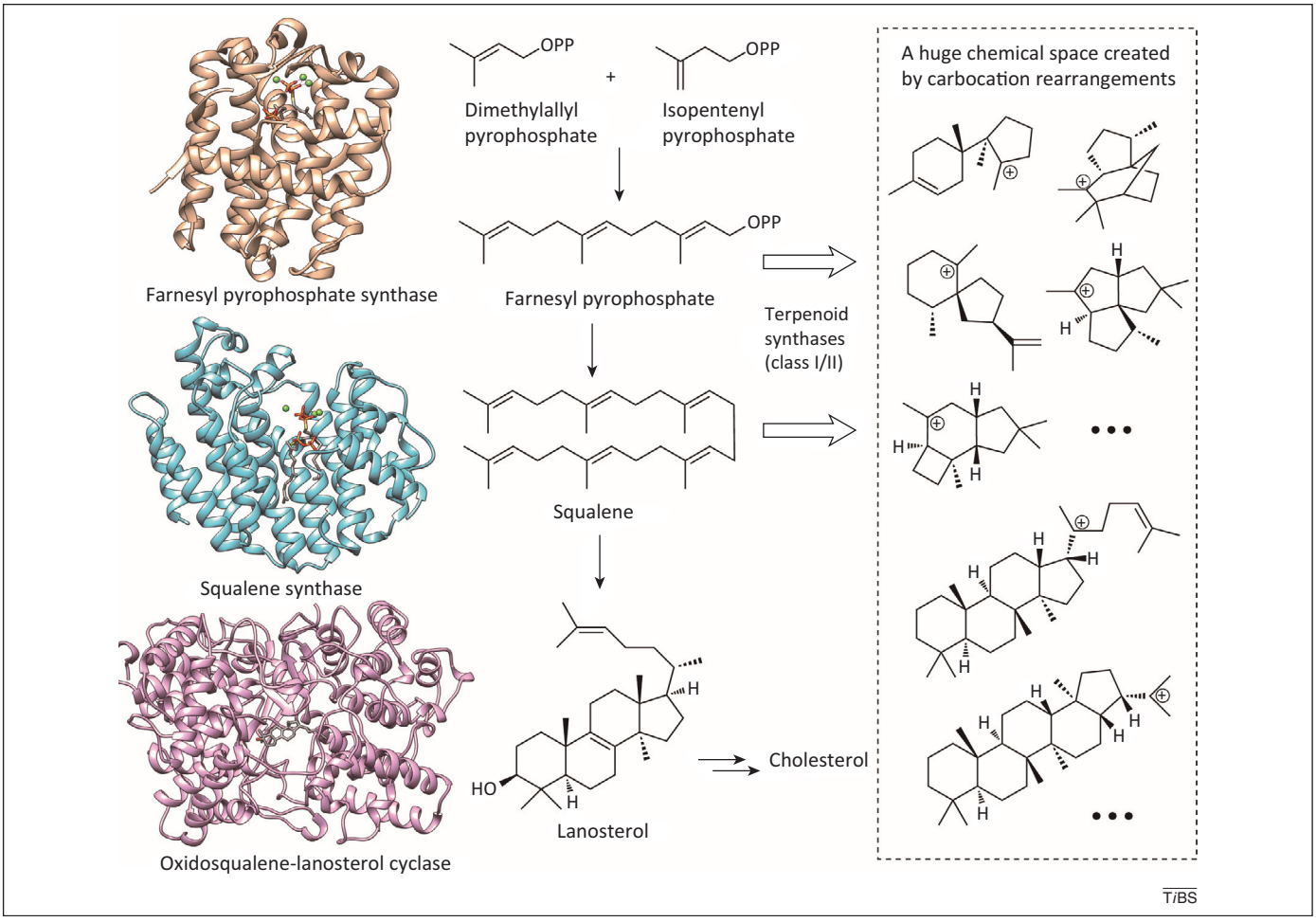

Figure 4. The biosynthesis of cholesterol: a paradigmatic isoprenoid pathway. Crystal structures of key enzymes in the pathway have been solved, including farnesyl pyrophosphate synthase [gold; Protein Data Bank (PDB): 1RQI], squalene synthase (light blue; PDB: 3WEG), and oxidosqualene-lanosterol cyclase (magenta; PDB 1W6K). These crystal structures provide opportunities to predict functions of related enzymes of the isoprenoid synthase superfamily. However, function prediction for the terpenoid synthases (also called terpene cyclases) is challenging due to the huge product chemical space created by carbocation rearrangements.

Missing links in metabolism: orphan enzyme activities and dead-end metabolites

In addition to the many functionally uncharacterized enzymes, there are also many enzyme activities that have been identified but are not associated with any protein sequence. In fact, despite considerable efforts over the past few years [70–76], 20% (1042 out of 5294 [77], as of February 2014) of enzyme commission (EC) numbers are not associated with sequence data in any of the three major enzyme databases (Metacyc [48], ExPasy [78], and Brenda [79]) and, thus, are described as orphan ECs. Other terms, such as 'orphan metabolic activities' and 'orphan enzymes', have also been used to describe the phenomenon. The original publication dates for orphan ECs ranges from the 1950s to today, with a mean of 1977 [72,73]. Many orphan ECs have biologically important roles, and could be an unexplored reservoir of new drug targets [72,80].

Our incomplete understanding of metabolism is also reflected by 'dead-end' metabolites. Metabolites in biochemical networks are generally linked to at least two enzymes; that is, each metabolite is both the product of one biochemical reaction and the substrate of another. Dead-end metabolites are those that currently can only be linked to one enzyme in an organism, and these can be readily identified by methods of automated metabolic network reconstruction [81]. For example, Mackie *et al.* recently identified 127 potential dead-end metabolites in *E. coli* K-12 [82].

The number of orphan enzyme activities and dead-end metabolites will naturally decrease as new enzyme functions are discovered. However, the ability to identify holes in our understanding of metabolism in specific species suggests new structure-based approaches. Instead of the current approach where a candidate enzyme is studied for functional clues, one could dock substrates (or intermediates) corresponding to orphan enzyme reactions and dead-end metabolites to structures or models of many uncharacterized enzymes within the relevant organism(s).

Although enzyme function can be predicted from protein sequence or, as emphasized in this review, protein structure, the combination of these approaches with high-throughput experimental methods of studying metabolism and methods to interrogate computationally the metabolic networks of entire organisms is likely to be even more powerful. Integrated experimental and computational methods have great promise to fill systematically holes in our understanding of both primary and secondary metabolism.

Box 1. Outstanding questions

- When the binding site of an enzyme is unknown and cannot be inferred from homologous proteins, can we predict the site using sequence- and/or structure-based methods? Can enzymes be readily identified from sequence or structure, compared with proteins that lack catalytic function?
- How complete are existing *in silico* databases of metabolites, for specific organisms (e.g., *Escherichia coli* or humans) and for life on Earth in general? Are there entirely new classes of secondary metabolite that have not yet been discovered?
- How can we define the functions of an enzyme when it catalyzes multiple reactions? What is the best way to predict functions of such enzymes?
- How can information from high-throughput metabolomics, protein interaction, and phenotyping experiments be optimally combined with sequence and structural information to infer enzyme activities and pathways or networks?
- Among the approximately 50 million protein sequences identified from genome sequences thus far, how many enzyme activities exist? What fraction of enzymes has multiple activities *in vitro* and *in vivo*?

Concluding remarks

In the sequence–structure–function paradigm, inferring function from structure has proven challenging, and many approaches to function prediction have not utilized structural information at all. In the case of enzymes, there has recently been rapid progress in experimental and computational approaches to inferring aspects of enzymatic activity from structure. Numerous challenges remain (Box 1), including the limitations of existing algorithms for metabolite docking and homology modeling, incomplete *in silico* databases of metabolites, and incomplete structural coverage of putative enzyme families, despite the advances made by high-throughput protein expression and structural biology (structural genomics). Nonetheless, structure-guided approaches have shown promise, particularly for the most challenging goal of identifying novel metabolites, enzyme activities, and biochemical pathways. As in drug discovery, where structural information is now routinely used to guide design, we believe that enzyme structures will prove to be an essential component of strategies for enzyme function prediction, not in isolation, but rather integrated with many other experimental and computational methods.

Acknowledgments

This work was part of the Enzyme Function Initiative supported by the National Institutes of Health Grant U54 GM093342. We thank John Gerlt for helpful discussions. We also thank Johannes Hermann and Frank Wallrapp for kindly sending us docked poses for Figure 2A and Figure 2D. M.P.J. is a consultant to Schrodinger LLC, which developed and distributes some of the software used in studies cited here.

References

1 ExPASy (2014) *UniProtKB/Swiss-Prot Protein Knowledgebase Release 2014_04 Statistics*, ExPASy
2 UniProt (2014) *UniProtKB/TrEMBL protein database release 2014_04 statistics*, UniProt
3 Friedberg, I. (2006) Automated protein function prediction: the genomic challenge. *Brief. Bioinform.* 7, 225–242
4 Schnoes, A.M. *et al.* (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* 5, e1000605
5 Seffernick, J.L. *et al.* (2001) Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. *J. Bacteriol.* 183, 2405–2410
6 Patti, G.J. *et al.* (2012) Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13, 263–269
7 Wagner, E.M. (2013) Monitoring gene expression: quantitative real-time rt-PCR. *Methods Mol. Biol.* 1027, 19–45
8 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
9 Wu, A.R. *et al.* (2014) Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46
10 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
11 Meier, M. *et al.* (2013) Proteome-wide protein interaction measurements of bacterial proteins of unknown function. *Proc. Natl. Acad. Sci. U.S.A.* 110, 477–482
12 Fuchs, H. *et al.* (2011) Mouse phenotyping. *Methods* 53, 120–135
13 Bassel, G.W. *et al.* (2012) Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *Plant Cell* 24, 3859–3875
14 Kufareva, I. *et al.* (2012) Compound activity prediction using models of binding pockets or ligand properties in 3D. *Curr. Top. Med. Chem.* 12, 1869–1882
15 Nilmeier, J.P. *et al.* (2013) Rapid catalytic template searching as an enzyme function prediction procedure. *PLoS ONE* 8, e62535
16 Yang, Y. *et al.* (2014) Understanding a substrate's product regioselectivity in a family of enzymes: a case study of acetaminophen binding in cytochrome P450s. *PLoS ONE* 9, e87058
17 Amin, S.R. *et al.* (2013) Prediction and experimental validation of enzyme substrate specificity in protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4195–E4202
18 Carbonell, P. and Faulon, J.L. (2010) Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* 26, 2012–2019
19 Meng, E.C. *et al.* (1992) Automated docking with grid-based energy evaluation. *J. Comput. Chem.* 13, 505–524
20 Wang, R.X. *et al.* (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* 46, 2287–2303
21 Kalyanaraman, C. *et al.* (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry* 44, 2059–2071
22 Favia, A.D. *et al.* (2008) Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. *J. Mol. Biol.* 375, 855–874
23 Xiang, D.F. *et al.* (2009) Functional annotation and three-dimensional structure of Dr0930 from *Deinococcus radiodurans*, a close relative of phosphotriesterase in the amidohydrolase superfamily. *Biochemistry* 48, 2237–2247
24 Song, L. *et al.* (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat. Chem. Biol.* 3, 486–491
25 Kalyanaraman, C. *et al.* (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. *Structure* 16, 1668–1677
26 Kalyanaraman, C. and Jacobson, M.P. (2010) Studying enzyme-substrate specificity in silico: a case study of the *Escherichia coli* glycolysis pathway. *Biochemistry* 49, 4003–4005
27 Lukk, T. *et al.* (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. *Proc. Natl. Acad. Sci. U.S.A.* 109, 4122–4127
28 Fan, H. *et al.* (2013) Assignment of pterin deaminase activity to an enzyme of unknown function guided by homology modeling and docking. *J. Am. Chem. Soc.* 135, 795–803
29 Hitchcock, D.S. *et al.* (2013) Structure-guided discovery of new deaminase enzymes. *J. Am. Chem. Soc.* 135, 13927–13933
30 Hermann, J.C. *et al.* (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature* 448, 775–779
31 Kumar, R. *et al.* (2014) Prediction and biochemical demonstration of a catabolic pathway for the osmoprotectant proline betaine. *MBio* 5, e00933–13
32 Zhao, S.W. *et al.* (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* 502, 698–702
33 Wallrapp, F.H. *et al.* (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1196–E1202
34 Rakus, J.F. *et al.* (2009) Computation-facilitated assignment of the function in the enolase superfamily: a regiochemically distinct galactarate dehydratase from *Oceanobacillus iheyensis*. *Biochemistry* 48, 11546–11558

35 Jacobson, M.P. *et al.* (2002) Force field validation using protein side chain prediction. *J. Phys. Chem. B* 106, 11673–11680
36 Hermann, J.C. *et al.* (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc.* 128, 15882–15891
37 Sherman, W. *et al.* (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* 49, 534–553
38 Tian, B.X. *et al.* (2013) Predicting enzyme-substrate specificity with QM/MM methods: a case study of the stereospecificity of D-glucarate dehydratase. *Biochemistry* 52, 5511–5513
39 Kamat, S.S. *et al.* (2011) Enzymatic deamination of the epigenetic base N-6-methyladenine. *J. Am. Chem. Soc.* 133, 2080–2083
40 Moult, J. *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP): round x. *Proteins: Struct. Funct. Bioinform.* 82, 1–6
41 Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science* 294, 93–96
42 Biasini, M. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gku340
43 Pieper, U. *et al.* (2014) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 42, D336–D346
44 Kim, J. *et al.* (2013) Structure-guided discovery of the metabolite carboxy-SAM that modulates tRNA function. *Nature* 498, 123–126
45 Binkowski, T.A. *et al.* (2010) Assisted assignment of ligands corresponding to unknown electron density. *J. Struct. Funct. Genomics* 11, 21–30
46 Lasker, K. *et al.* (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins: Struct. Funct. Bioinform.* 78, 3205–3211
47 Irwin, J.J. *et al.* (2009) Automated docking screens: a feasibility study. *J. Med. Chem.* 52, 5712–5720
48 Caspi, R. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 40, D742–D753
49 Dehal, P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.* 38, D396–D400
50 Franceschini, A. *et al.* (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815
51 Aziz, R.K. *et al.* (2012) SEED Servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* 7, e48053
52 Markowitz, V.M. *et al.* (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122
53 Babu, M. *et al.* (2012) Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* 489, 585–589
54 Havugimana, P.C. *et al.* (2012) A census of human soluble protein complexes. *Cell* 150, 1068–1081
55 Zhang, Y. *et al.* (2009) Three-dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science* 325, 1544–1549
56 Zotchev, S.B. (2013) Alkaloids from marine bacteria. *Adv. Bot. Res.* 68, 301–333
57 Ziegler, J. and Facchini, P.J. (2008) Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* 59, 735–769
58 Walsh, C.T. and Fischbach, M.A. (2010) Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.* 132, 2469–2493
59 Keatinge-Clay, A.T. (2012) The structures of type I polyketide synthases. *Nat. Prod. Rep.* 29, 1050–1073
60 Rottig, M. *et al.* (2011) NRPSpredictor2: a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* 39, W362–W367
61 Go, M.K. *et al.* (2012) Establishing a toolkit for precursor-directed polyketide biosynthesis: exploring substrate promiscuities of acid-CoA ligases. *Biochemistry* 51, 4568–4579
62 Williams, G.J. (2013) Engineering polyketide synthases and nonribosomal peptide synthetases. *Curr. Opin. Struct. Biol.* 23, 603–612
63 Wong, F.T. and Khosla, C. (2012) Combinatorial biosynthesis of polyketides: a perspective. *Curr. Opin. Chem. Biol.* 16, 117–123
64 Sacchettini, J.C. and Poulter, C.D. (1997) Biochemistry: creating isoprenoid diversity. *Science* 277, 1788–1789
65 Christianson, D.W. (2007) Roots of biosynthetic diversity. *Science* 316, 60–61
66 Tantillo, D.J. (2011) Biosynthesis via carbocations: theoretical studies on terpene formation. *Nat. Prod. Rep.* 28, 1035–1053
67 Punta, M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301
68 Goodacre, N.F. *et al.* (2013) Protein domains of unknown function are essential in bacteria. *MBio* 5, e00744–13
69 Bastard, K. *et al.* (2014) Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.* 10, 42–49
70 Yamada, T. *et al.* (2012) Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Mol. Syst. Biol.* 8, 581
71 Smith, A.A.T. *et al.* (2012) The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput. Biol.* 8, e1002540
72 Pouliot, Y. and Karp, P.D. (2007) A survey of orphan enzyme activities. *BMC Bioinformatics* 8, 244
73 Chen, L.F. and Vitkup, D. (2007) Distribution of orphan metabolic activities. *Trends Biotechnol.* 25, 343–348
74 Ramkissoon, K.R. *et al.* (2013) Rapid identification of sequences for orphan enzymes to power accurate protein annotation. *PLoS ONE* 8, e84508
75 Watschinger, K. and Werner, E.R. (2013) Orphan enzymes in ether lipid metabolism. *Biochimie* 95, 59–65
76 Lespinet, O. and Labedan, B. (2006) ORENZA: a web resource for studying ORphan ENZyme activities. *BMC Bioinformatics* 7, 436
77 ExplorEnz (2014) *Enzyme Database Statistics*, ExplorEnz
78 Artimo, P. *et al.* (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603
79 Schomburg, I. *et al.* (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* 41, D764–D772
80 Lespinet, O. and Labedan, B. (2006) Orphan enzymes could be an unexplored reservoir of new drug targets. *Drug Discov. Today* 11, 300–305
81 Karp, P.D. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* 11, 40–79
82 Mackie, A. *et al.* (2013) Dead end metabolites: defining the known unknowns of the *E. coli* metabolic network. *PLoS ONE* 8, e75210