The Rough Guide to In Silico Function Prediction, or How To Use Sequence and Structure Information To Predict Protein Function

Marco Punta[1,2,3], **Yanay Ofran**[4]*

[1] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York, United States of America, [2] Columbia University Center for Computational Biology and Bioinformatics (C2B2), New York, New York, United States of America, [3] Northeast Structural Genomics Consortium (NESG), Columbia University, New York, New York, United States of America, [4] The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

A Tutorial in PLoS Computational Biology

Introduction

Choosing the right function prediction tools. The vast majority of known proteins have not yet been characterized experimentally, and there is very little that is known about their function. New unannotated sequences are added to the databases at a pace that far exceeds the one in which they are annotated in the lab. Computational biology offers tools that can provide insight into the function of proteins based on their sequence, their structure, their evolutionary history, and their association with other proteins. In this contribution, we attempt to provide a framework that will enable biologists and computational biologists to decide which *type* of computational tool is appropriate for the analysis of their protein of interest, and what kind of insights into its function these tools can provide. In particular, we describe computational methods for predicting protein function directly from sequence or structure, focusing mainly on methods for predicting molecular function. We do not discuss methods that rely on sources of information that are beyond the protein itself, such as genomic context [1], protein–protein interaction networks [2], or membership in biochemical pathways [3]. When choosing a tool for function prediction, one would typically want to identify the best performing tool. However, a quantitative comparison of different tools is a tricky task. While most developers report their own assessment of their tool, in most cases there are no standard datasets and generally agreed-upon measures and criteria for benchmarking function prediction methods. In the absence of independent benchmarks, comparing the figures reported by the developers is almost always comparing oranges and apples (for discussion of this problem see [4]). Therefore, we refrain from reporting numerical assessments of specific methods. For those cases in which independent assessment of performance is available, we refer the reader to the original publications. Finally, we discuss only methods that are either accessible as Web servers or freely available for download (relevant Web links can be found in Table S1).

What is protein function? The first problem we face when dealing with protein function is well-illustrated by the title of a 1998 article by Schubert et al. [5], "The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase." What is the function of the protein that is described in this paper? The authors report the solution of the crystal structure of CbiF, which is an enzyme implicated in the biosynthesis of vitamin B12 (cobalamin). More specifically, CbiF transfers a methyl group from an S-adenosyl-L-methionine molecule to a precursor of vitamin B12 (cobalt-precorrin-4). Vitamin B12 is a compound that "helps maintain healthy nerve cells and red blood cells, and is also needed to make DNA" [6]. Its deficiency is related to anemia, as well as to several neurological and psychiatric symptoms [7]. As we see, CbiF function comes in different flavors: molecular/enzymatic (methyltransferase), metabolic (cobalamin biosynthesis—directly—and DNA biosynthesis—indirectly), and physiological (maintenance of healthy nerve and red blood cells, through B12), along with possible consequences related to their malfunctioning. There are, obviously, numerous ways to describe each of these aspects of the protein function. Enzymatic function, for example, may be characterized through: reaction (methylation), substrate (cobalt-precorrin-4), or ligand (S-adenosyl-L-methionine).

Classifying and predicting. Since protein function has many facets, its prediction has different meaning for different people. It may mean the prediction of the cellular process in which the protein is involved, or the nitty-gritty of its enzymatic activity, or rather its physiological role. Therefore, when attempting to predict protein function one should first define clearly the kind of function she or he wants to predict. When predicting function automatically on a large scale, this problem is intensified by the need to standardize and quantitatively assess the similarity of functions between proteins. While defining sequence and structural similarity may be easy, there is no a priori straightforward measure we can use to put a number on the similarity of functions

Citation: Punta M, Ofran Y (2008) The Rough Guide to In Silico Function Prediction, or How To Use Sequence and Structure Information To Predict Protein Function. PLoS Comput Biol 4(10): e1000160. doi:10.1371/journal.pcbi.1000160

Published October 31, 2008

Editor: Fran Lewitter, Whitehead Institute, United States of America

Funding: MP was supported by grant U54-GM75026-01 from the National Institutes of Health (NIH).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yanay@ofranlab.org

between two proteins. Prediction methods could not be developed, or rigorously assessed, without such measure. Several large-scale projects attempted to respond to this challenge by building classification systems or ontologies of biological functions (see [8,9] for review). One such enterprise was launched as early as 1955 by the International Congress of Biochemistry, which created the Enzyme Commission to come up with a nomenclature for enzymes. In this numerical classification, each enzymatic function could be described by a set of four numbers (which, together, are dubbed EC number). Each of these four numbers represents specific description of the enzyme and its activity. For instance, when comparing carboxylesterase (3.1.1.1) and isochorismatase (3.3.2.1), one can tell that they share the basic enzymatic activity of a hydrolase (all hydrolases have 3 as the first number), but they act on different types of bonds: hydrolases with 3.1.-.- act on an ester bond and those with 3.3.-.- act on an ether bond. This system is infinitely expandable to include any new enzyme, but it does not cover functions that are not enzymatic. The Gene Ontology (GO) project provides a controlled vocabulary to describe the function of any gene product in any organism. It developed three structured controlled vocabularies to cope with the multifaceted nature of the biological function. For each gene product, GO can provide a number for its cellular component, the biological process in which it is involved, and its specific molecular function. Various algorithms have been proposed to assign a score for the similarity between numbers within each of these three ontologies [10,11]. Thus, GO has become the standard for assessing the performance of function prediction methods.

Function Annotation Transfer from Sequence

Homology useful but different from "same function". The most widely used approach for function prediction is homology transfer. Given an unannotated protein, this approach suggests searching for an annotated homolog and using the experimentally verified function of the latter to infer the function of the former. However, this procedure should be implemented with caution. Homology is often confused with similarity of function. In reality, homology between two proteins simply means that they have a common evolutionary origin. Whether or not they have since retained similarity in any of their properties is something that needs to be checked in each individual case. An important distinction in this context is between orthologous and parologous sequences: orthologs are genes that originated from a common ancestor through a speciation event, while paralogs are the results of duplication events within the same genome. In general, function tends to be more conserved in orthologs than in paralogs [12]. So, when attempting to predict the function of an unannotated protein based on its homology to an annotated one, one should search for orthologs rather than paralogs (Figure 1A). Although several databases have been created to help identify orthologous genes (e.g., COGs [13] and InParanoid [14]), "proven orthologs are as rare in the literature as diamonds in bare rock" [12]. Orthologs, additionally, may also diverge functionally, sometimes more than corresponding paralogs [12]. Finally, there exist functional similarities between proteins that are not reflected in homology. These facts underline the difficulty of the task of transferring function from a homologous template.

In practice, the most common way to infer homology is by detecting sequence similarity (note, however, that remote relationships will generally be missed by sequence similarity approaches; see the section about structure below). Popular sequence alignment methods include PSI-BLAST [15], HMMER [16], and SAM [17]. When investigating the function of a protein, we ought to align its sequence against a database of annotated proteins, such as SWISS-PROT [18], in order to find its homologs of known function. The question we need to address is how two homologous proteins relate functionally. As we mentioned previously, several studies have shown that homology (both orthology and paralogy) does not guarantee conservation of function (Table 1). Indeed, relatively small differences in sequence can sometimes cause quite radical changes in functional properties, such as a change of enzymatic action, or even a loss or acquisition of the enzymatic activity itself. It is also apparent that there is no sequence similarity threshold that guarantees that two proteins share the same function (see references in Table 1). Thus, although higher sequence similarity increases confidence in function annotation transfer, there is no threshold that can be considered safe. An extreme case is represented by the so-called "moonlighting proteins" or proteins that perform multiple and, at times, significantly different functions [19,20]. For example, η-crystallin is a protein that plays a structural role in the eye lens of several species, while working as an enzyme in other tissues. Homologs of these proteins may retain only some of the original functions [21]. As a consequence, function annotation transfer may result in erroneous or incomplete assignments (Figure 1B).

The multi-domain nature of many proteins can also be the cause of annotation transfer errors (Figure 1C). In fact, in databases storing entire sequences (such as SWISS-PROT [18]), functional annotation of a protein may refer to any of its domains. If the query protein (i.e., the protein whose function we wish to predict) does not align to that specific domain, annotation transfer is totally unjustified and will very likely result in a misannotation. While a number of databases and tools attempt to split proteins into domains based on sequence (Pfam [16], PRODOM [22], SMART [23]), the most reliable way to identify protein domains is by using, when possible, structural knowledge (SCOP [24], CATH [25]).

Some of these problems can be mitigated by the use of phylogenomic inference that frames sequence evolutionary relationship into a phylogenetic context as described in [26].

To complicate matters further, bear in mind that databases contain incorrect annotations, mostly caused by erroneous automatic annotation transfer by homology [27] (Figure 1D). Thus, always check the source of the annotation before you use it.

In conclusion, homology between two proteins does not guarantee that they have the same function, not even when sequence similarity is very high (including 100% sequence identity) (Table 2). Bottom line: when annotating function, you won't get too far with the classic 25%–30% sequence identity that is so powerful for structure prediction. On the positive side, the higher the sequence similarity the better the chance that homologous proteins in fact share functional features (Tables 1 and 2). As we have seen, correct transfer of functional annotation from a protein to its homolog depends on whether the two proteins are orthologs or paralogs, on the level of sequence similarity, on the type of annotation we want to transfer (for example, prediction of subcellular localization typically requires lower sequence identity than prediction for enzymatic function [28]), and on the specific domain aligned. No sequence similarity threshold is safe for blind annotation transfer.

Sequence signatures predict functional traits. In some cases, a

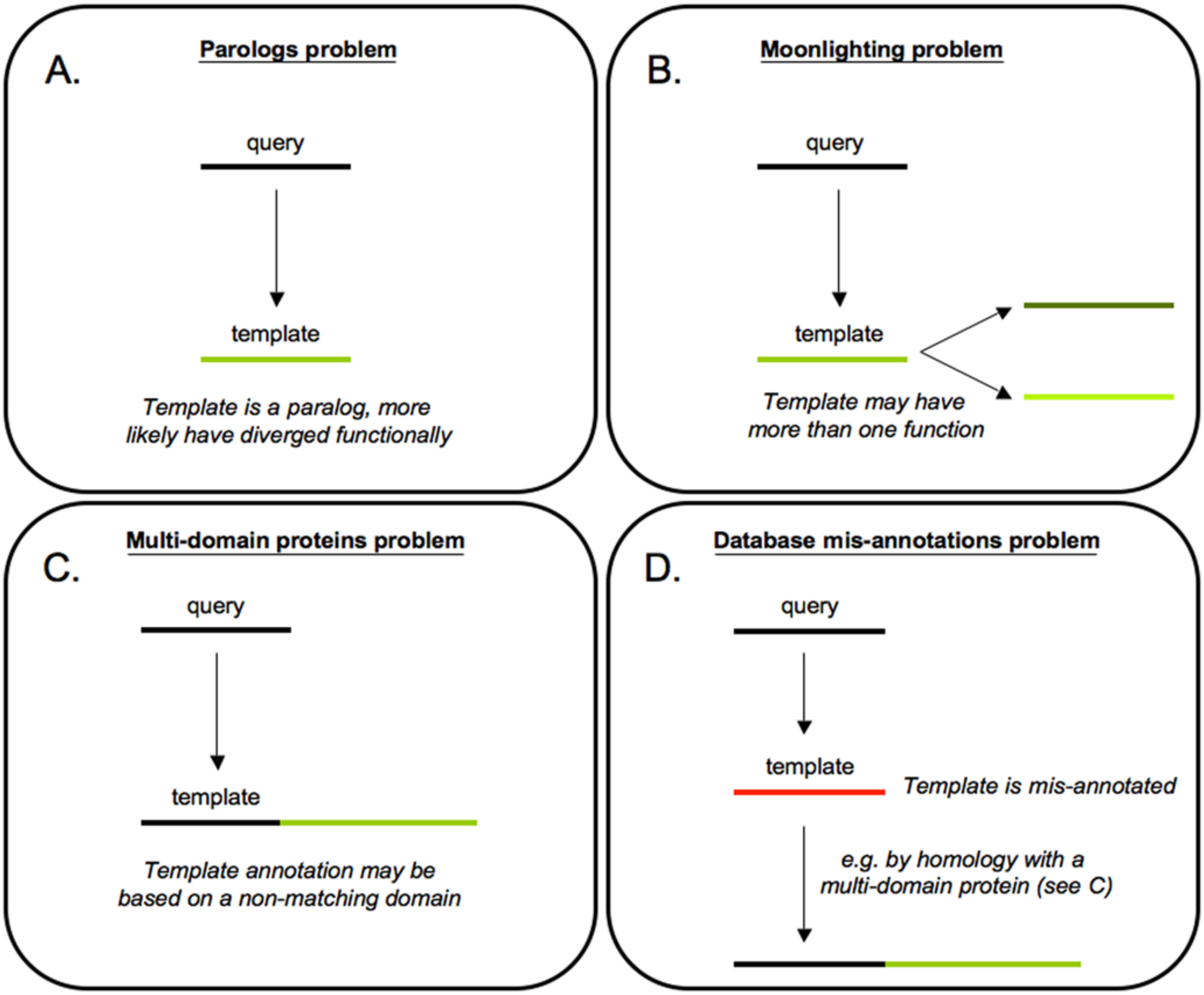
A.
Parologs problem
query
template
Template is a paralog, more likely have diverged functionally
B.
Moonlighting problem
query
template
Template may have more than one function
C.
Multi-domain proteins problem
query
template
Template annotation may be based on a non-matching domain
D.
Database mis-annotations problem
query
template
Template is mis-annotated
e.g. by homology with a multi-domain protein (see C)

Figure 1. Homology based annotation transfer: Problems. (A) Paralogy problem: Paralogs are more likely to diverge functionally with respect to orthologs. If our putative template is a paralog, the probability that the query has similar function decreases. (B) Moonlighting problem: If the template performs multiple functions, the query could have retained only some of them (and vice-versa, if the query were a moonlighting protein, using a non-moonlighting template would result in an incomplete annotation of the query). (C) Multi-domain proteins problem: If the template is annotated based on the function of a domain that is not aligned to the query, annotation transfer is not possible. (D) Database mis-annotations problem: Database entries may have been mis-annotated; the risk is especially high if annotation was performed automatically via homology transfer.
doi:10.1371/journal.pcbi.1000160.g001

relatively small sequence signature may suffice to conserve the function of a protein even if the rest of the protein has changed considerably during the course of evolution. Alternatively, non-homologous proteins could acquire the same functional motif independently (convergent evolution). Thus, two proteins that would not find each other in a sequence search may still have common sequence signatures that could surrender their functional relatedness. Clearly, if two proteins have some level of overall sequence similarity and also share a common motif, the confidence of annotation transfer increases. Several computational tools are dedicated to the identification of functional motifs (e.g., PRINT-S [29], BLOCKS [30], PROSITE [31], InterPro [32], and ELM [33]). They usually offer a large library of sequence motifs that have been collected either manually by experts, or automatically by pattern-searching algorithms, or by a combination of the two. When a query sequence is submitted to these tools, it is compared to all known motifs in search of a match. Finding one of these well-characterized motifs in a newly discovered sequence could offer some insights into its function.

More generally, residues that are crucial for the function of the protein can often be identified through the use of multiple sequence alignments that highlight conservation patterns in protein families (see [34] and [35] for more detailed discussion of these methods). This approach is possible, of course, when multiple homologs of the protein of interest are available. Importantly, even when the function of specific conserved residues within the protein family is not known, multiple sequence alignments point to regions that may be of interest for experimental functional characterization (e.g., by means of site directed mutagenesis). Multiple sequence alignments are also relevant as input to methods that map sequence conservation on the protein surface (see below).

Table 1. Do's and Don'ts of annotation transfer by homology.

Functional property to be conserved	Sequence identity	Conservation rate	Reference
Non-enzyme	50%	98%*	[88]
All 4 EC numbers	70%**	90%	[89]
All 4 EC numbers	40%**	70%	[89]
First 3 EC numbers	50%**	90%	[89]
First 3 EC numbers	30%**	70%	[89]
All 4 EC numbers	50%	30%	[90]
First 3 EC numbers	25%	70%	[91]
SWISS-PROT keywords	40%	70%	[92]
Subcellular localization (11 classes)	70%	90%	[93]

*98% of non enzymes that have at least one enzyme homolog.
**Global identity, defined in [89].
Note: different estimates for the same functional aspects reflect the different methods, procedures, and datasets used to assess sequence similarity by the various groups.

doi:10.1371/journal.pcbi.1000160.t001

Function Annotation Transfer from Structure

Structure better than sequence alone. Proteins live and function in 3D, and therefore structural information is very helpful for predicting function. The need for tools to predict function from structure is intensified by the success of the structural genomics enterprises that deposit hundreds of new experimentally solved structures of proteins with unknown function [36]. Structural information, however, does not have to come directly from the protein of interest but can also be derived from a homologous protein via modeling [37]. Unfortunately, as with sequence, two proteins having the same overall structural architecture, and even conserved functional residues [38], can have unrelated functions. Additionally, two proteins can perform the same function while having radically different structures [39]. Still, structure may help function prediction in several ways. Structural similarity between two proteins may reveal their common evolutionary origin even in the absence of significant sequence similarity, possibly suggesting similar function (Figure 2A). Or, it may indicate evolutionary convergence caused by common functional constraints. Prokaryotic virulence effectors offer some remarkable examples of functional convergence. Some of these proteins, in order to be able to tamper with the biological processes of the host, have adapted to mimic host proteins. This is achieved by either mimicking their overall architecture or, more often, their local structural features [40,41]. Numerous methods have been developed to perform structural comparisons, using the Protein Data Bank [42] or structure classification databases (SCOP [24], CATH [25]) as a source. Among the most used structural alignment methods are SSM [43], FATCAT [44], DALI [45], and CATHEDRAL [46] (see [47] for a comparison of the performance of several methods). In general, it is suggested to use more than one method since different methods may capture different valid matches. Most programs provide a PDB-type output file for the two aligned proteins that can be uploaded to one of the many available structure visualization programs (e.g., VMD [48], AstexViewer 2.0 [49]). When evaluating the functional implications of a match, we need to consider how functionally promiscuous a given structural architecture is (i.e., whether or not it is known to relate to many functions [50]), and we have to check the conservation of functional residues. Functional residues may not be

Table 2. Do's and Don'ts of annotation transfer by homology.

			Yes	No
Homology	=	Same function		√
Orthology	=	Same function		√
Paralogy	=	Same function		√
Orthology	=	>Probability of same function	√	
Paralogy	=	<Probability of same function	√	
Same sequence	=	Same function		√
Sequence similarity>threshold	=	Same function		√
Homology+conservation of functional residues	=	Same function		√
Similar structure	=	Similar function		√
>Sequence similarity	=	>Probability of same function	√	
>Structure similarity	=	>Probability of same function	√	

doi:10.1371/journal.pcbi.1000160.t002

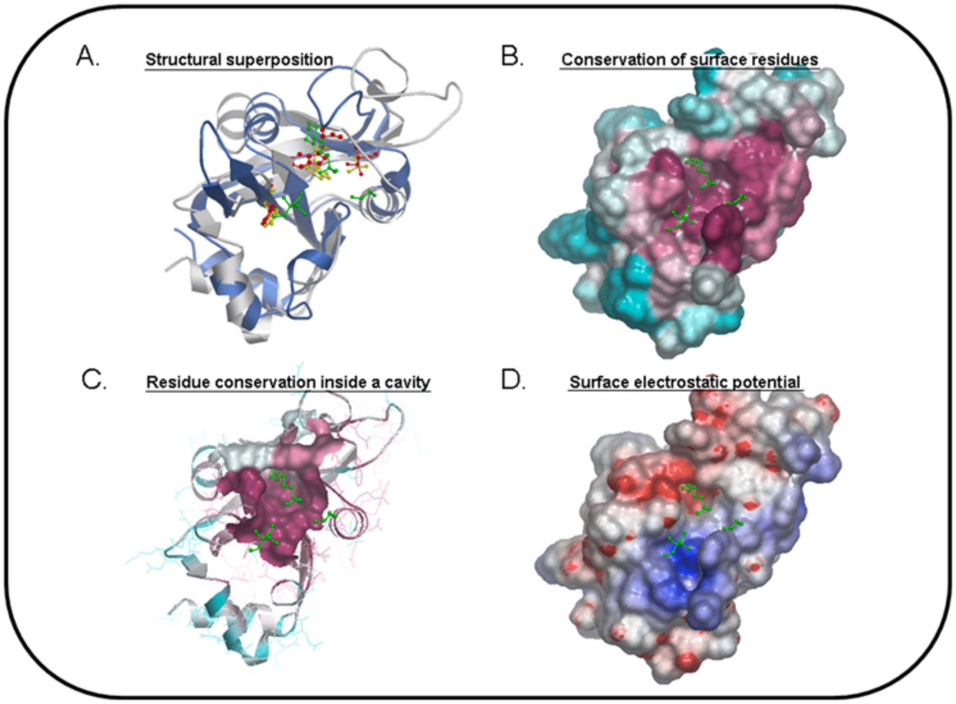
A.
Structural superposition
B.
Conservation of surface residues
C.
Residue conservation inside a cavity
D.
Surface electrostatic potential

Figure 2. Using structure to predict function. The protein represented here is PDBid: 2eve. All figures are derived from the Northeast Structural Genomics Consortium structure gallery (http://nmr.cabm.rutgers.edu:9090/gallery/jsp/Gallery.jsp). AstexViewer 2.0 [49] is used for visualization. (A) Superposition of 2eve structure (gray) and of the structure of a homolog (blue, PDBid: 2ar1), using Skan [59]. 2eve hosts three co-crystallized small non-functional ligands (green; ball and stick). Three structurally aligned residues of 2eve and 2ar1 are also shown (red and yellow; ball and stick). (B) Surface residue conservation: Conserved residues (mauve) versus variable residues (cyan). Conservation is calculated as follows: homologs of 2eve are collected using three iterations of PSI-BLAST [15] retaining all homologs with E-value$<10-3$ and reducing redundancy at 80% sequence identity with CD-HIT [85]. Then, a multiple sequence alignment is created using CLUSTALW [86]. Finally, the multiple sequence alignment is used as input to ConSurf [54], which uses it to calculate residue conservation. (C) Residue conservation within the protein largest cavity (as defined by SCREEN [87]). (D) 2eve surface electrostatic potential (using GRASP2 [59]) (positive in blue, negative in red).
doi:10.1371/journal.pcbi.1000160.g002

perfectly conserved in proteins of similar function. In fact, specific residues may be responsible for different ligand or substrate binding affinities or for different reaction rates in enzymes. However, disruption of the 3D core of an active site in an overall conserved structural architecture should be a serious concern [51]. Catalytic Site Atlas [52] and MACiE [53] are databases where you can find detailed information about functional residues and their specific role in enzymes.

Even in the absence of a structurally related protein, structure may provide important functional information by highlighting properties of the protein's accessible surface that may relate to function. These include residue conservation (Consurf [54], siteFiNDER|3D [55], TRACE [56], Figure 2B), cavities (CASTp [57], Q-SiteFinder [58], Figure 2C), and electrostatic patches (GRASP2 [59], Figure 2D). In general, structural knowledge, although not a panacea for all problems, is an extremely powerful tool for computational function prediction.

Structural motifs reveal binding sites. The idea is similar to sequence motifs: functional aspects may be defined by local structural signatures. Residues found in functional signatures may be not be adjacent in sequence; however, they do tend to cluster in the 3D structure, forming binding sites for ions, small molecules, DNA, RNA, or other proteins. There are databases and tools for searching such structurally defined motifs in a structure of interest (JESS [60], RIGOR [61], PAR-3D [62], PINTS [63], and PDBSiteScan [64]). As usual, the effectiveness of such methods depends on the specific function being predicted and on the desired level of detail of the prediction.

De Novo Function Prediction Using Sequence and Structure

De novo predictions push the limit. What can we do when the protein whose function we want to predict has no significant similarity to any annotated protein? Several approaches have been suggested to predict protein function de novo. That is, using sequence or structure information without relying on similarity to a specific protein but rather on the "generic" properties that are common to proteins of the same function. Indeed, proteins of the same function have to adapt to similar constraints (e.g., pH, properties of a ligand, structural flexibility), which will be reflected in their sequence and structural features. De novo methods are generally based on machine learning algorithms that are able to capture significant non-trivial correlations between features and functions. These methods are usually less accurate than annotation transfer but enjoy higher coverage, eventually protruding into experimentally yet unexplored regions of the sequence space and allowing annotation of entire genomes. Hereafter, we report on some of the most successful de novo methods.

Functional residues. Residues that have similar function in different proteins are likely to possess similar physicochemical characteristics. For example, residues that bind DNA share common structural and physicochemical features in most DNA-binding proteins (e.g., secondary structures, geometries, solvent accessibility, charge, hydrophobicity). Once these features are characterized and quantified, it may be possible to search for residues that possess them, thus predicting their function. There are several methods for the prediction of DNA binding residues from sequence (e.g., DISIS [65] and bindN [66]) or structure (e.g., Patchfinder+ [67]). Another example is represented by residues that bind metals. The number and type of residues binding to a given metal may considerably differ from protein to protein. For this reason, known sequence metal binding motifs are useful but cover only a small fraction of all binding sites [68]. Recently, de novo methods have been developed that specialize in predicting metal binding sites from sequence (MetalDetector [69]) and from structure (MetSite [70] and CHED [71]), the latter exploiting successfully the tight clustering of metal binding residues in 3D.

Subcellular localization. Knowing the subcellular localization of a protein helps to narrow down the number of functions the protein can perform and can be very relevant for its experimental characterization [72]. Subcellular localization can be predicted from homology and motifs, with the aforementioned limitations. De novo methods, instead, exploit the known correlation between amino acid composition and localization [73]. LOCtree [74], BaCelLo [75], TARGETp [76], Protein Prowler [77], and the PSORT suite of programs [78]—some combining de novo, homology, and motifs—are among the best methods available.

Programs that predict function combining different sources of information. Another, more ambitious, approach is to integrate various aspects of proteins and to try to associate them with specific GO numbers. Since protein function is a multifaceted notion, its comprehensive prediction requires data from many sources. Thus, these methods attempt to integrate all sorts of information that pertain to function such as structure, sequence information, physicochemical features, and even protein interaction data. Such an approach is taken, for example, by ProtFun [79], which combines 14 different sequence-based prediction methods such as prediction of glycolization sites, number of negative and positive residues, predicted transmembrane helices, predicted subcellular localization, and other features, and integrates them to yield a GO term. ProKnow [80] relies predominantly on structural features that are associated with specific functions as well as on sequence motifs and interaction data. Similarly, ProFunc [81] uses structure and sequence motifs, combined with identification of active and binding sites and integrates them with interaction data and knowledge of genomic sequences to yield a comprehensive prediction of function.

Several more de novo methods that are relevant for function exist, including predictors of coil-coiled regions [82], natively unstructured regions [83], and post-translational modifications [84].

Supporting Information

Table S1 Publicly available tools.
Found at: doi:10.1371/journal.pcbi.1000160.s001 (0.18 MB DOC)

References

1. Gabaldon T, Huynen MA (2004) Prediction of protein function and pathways in the genome era. Cell Mol Life Sci 61: 930–944.
2. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43. doi:10.1371/journal.pcbi.0030043.
3. Gianchandani EP, Brautigan DL, Papin JA (2006) Systems analyses characterize integrated functions of biochemical networks. Trends Biochem Sci 31: 284–291.
4. Godzik A, Jambon M, Friedberg I (2007) Computational protein function prediction: Are we making progress? Cell Mol Life Sci 64: 2505–2511.
5. Schubert HL, Wilson KS, Raux E, Woodcock SC, Warren MJ (1998) The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase. Nat Struct Biol 5: 585–592.
6. MedlinePlus (2005) Medline Plus. Available: http://www.nlm.nih.gov/medlineplus/. Accessed 23 July 2008.
7. Reynolds E (2006) Vitamin B12, folic acid, and the nervous system. Lancet Neurol 5: 949–960.
8. Thomas PD, Mi H, Lewis S (2007) Ontology annotation: Mapping genomic regions to biological function. Curr Opin Chem Biol 11: 4–11.
9. Bard JB, Rhee SY (2004) Ontologies in biology: Design, applications and future challenges. Nat Rev Genet 5: 213–222.
10. Lee SG, Hur JU, Kim YS (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. Bioinformatics 20: 381–388.
11. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23: 1274–1281.
12. Theissen G (2002) Secret life of genes. Nature 415: 741.
13. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.
14. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041–1052.
15. Altschul S, Madden T, Shaffer A, Zhang J, Zhang Z, et al. (1997) Gapped Blast and PSI-Blast: A new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
16. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, et al. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27: 260–262.
17. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, et al. (2005) SAM-T04: What is new in protein-structure prediction for CASP6. Proteins 61 (Supplement 7): 135–142.
18. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190–D195.
19. Jeffery CJ (2004) Molecular mechanisms for multitasking: Recent crystal structures of moonlighting proteins. Curr Opin Struct Biol 14: 663–668.
20. Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24: 8–11.
21. Bateman OA, Purkiss AG, van Montfort R, Slingsby C, Graham C, et al. (2003) Crystal structure of eta-crystallin: Adaptation of a class 1 aldehyde dehydrogenase for a new role in the eye lens. Biochemistry 42: 4349–4356.
22. Corpet F, Gouzy F, Kahn D (1998) The ProDom database of protein domain families. Nucleic Acids Res 26: 323–326.
23. Ponting CP, Schultz J, Milpetz F, Bork P (1999) SMART: Identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 27: 229–232.
24. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C (1999) SCOP: A Structural Classification of Proteins database. Nucleic Acids Res 27: 254–256.
25. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM (1997) CATH—A hierarchic classification of protein domain structures. Structures 5: 1093–1108.
26. Brown D, Sjolander K (2006) Functional classification using phylogenomic inference. PLoS Comput Biol 2: e77. doi:10.1371/journal.pcbi.0020077.
27. Linial M (2003) How incorrect annotations evolve—The case of short ORFs. Trends Biotechnol 21: 298–300.
28. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y (2003) Automatic prediction of protein function. Cell Mol Life Sci 60: 2637–2650.
29. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, et al. (1999) PRINTS prepares for the new millennium. Nucleic Acids Res 27: 220–225.
30. Henikoff JG, Henikoff S (1996) Blocks database and its applications. Methods Enzymol 266: 88–104.
31. Hofmann K, Bucher P, Falquet L, Bairoch A (1999) The PROSITE database, its status in 1999. Nucleic Acids Res 27: 215–219.
32. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro—An integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150.
33. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31: 3625–3630.
34. Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3: e123. doi:10.1371/journal.pcbi.0030123.
35. Wallace IM, Blackshields G, Higgins DG (2005) Multiple sequence alignments. Curr Opin Struct Biol 15: 261–266.
36. Shapiro L, Harris T (2000) Finding function through structural genomics. Curr Opin Biotechnol 11: 31–35.
37. Petrey D, Honig B (2005) Protein structure prediction: Inroads to biology. Mol Cell 20: 811–819.
38. Bartlett GJ, Borkakoti N, Thornton JM (2003) Catalysing new reactions during evolution: Economy of residues and mechanism. J Mol Biol 331: 829–860.

39. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36: 307–340.

40. Desveaux D, Singer AU, Dangl JL (2006) Type III effector proteins: Doppelgangers of bacterial virulence. Curr Opin Plant Biol 9: 376–382.

41. Stebbins CE, Galan JE (2001) Structural mimicry in bacterial virulence. Nature 412: 701–705.

42. Berman HM, Westbrook J, Feng Z, Gillliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.

43. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60: 2256–2268.

44. Ye Y, Godzik A (2004) FATCAT: A Web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res 32: W582–W585.

45. Holm L, Sander C (1996) DALI/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 25: 231–234.

46. Redfern OC, Harrison A, Dallman T, Pearl FM, Orengo CA (2007) CATHEDRAL: A fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput Biol 3: e232. doi:10.1371/journal.pbio.0020232.

47. Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. J Mol Biol 346: 1173–1188.

48. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graph 14: 33–38, 27–38.

49. Hartshorn MJ (2002) AstexViewer: A visualisation aid for structure-based drug design. J Comput Aided Mol Des 16: 871–881.

50. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA (2000) From structure to function: Approaches and limitations. Nat Struct Biol 7 (Supplement): 991–994.

51. Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 347: 565–581.

52. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129–D133.

53. Holliday GL, Almonacid DE, Bartlett GJ, O'Boyle NM, Torrance JW, et al. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): Novel tools for searching catalytic mechanisms. Nucleic Acids Res 35: D515–D520.

54. Armon A, Graur D, Ben-Tal N (2001) ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307: 447–463.

55. Innis CA (2007) siteFiNDER|3D: A Web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res 35: W489–W494.

56. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342–358.

57. Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res 31: 3352–3355.

58. Laurie AT, Jackson RM (2005) Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21: 1908–1916.

59. Petrey D, Honig B (2003) GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol 374: 492–509.

60. Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. Bioinformatics 19: 1644–1649.

61. Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285: 1887–1897.

62. Goyal K, Mohanty D, Mande SC (2007) PAR-3D: A server to predict protein active site residues. Nucleic Acids Res 35: W503–W505.

63. Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. Nucleic Acids Res 31: 3341–3344.

64. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2004) PDBSiteScan: A program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. Nucleic Acids Res 32: W549–W554.

65. Ofran Y, Mysore V, Rost B (2007) Prediction of DNA-binding residues from sequence. Bioinformatics 23: i347–i353.

66. Wang L, Brown SJ (2006) BindN: A Web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34: W243–W248.

67. Shazman S, Celniker G, Haber O, Glaser F, Mandel-Gutfreund Y (2007) Patch Finder Plus (PFplus): A Web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res 35: W526–W530.

68. Passerini A, Punta M, Ceroni A, Rost B, Frasconi P (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins 65: 305–316.

69. Lippi M, Passerini A, Punta M, Rost B, Frasconi P (2008) MetalDetector: A Web server for predicting metal binding sites and disulfide bridges in proteins from sequence. Bioinformatics; in press.

70. Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, et al. (2004) Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342: 307–320.

71. Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M (2008) Prediction of transition metal-binding sites from apo protein structures. Proteins 70: 208–217.

72. Nair R, Rost B (2007) Predicting proteins subcellular localization using intelligent systems. In: Leon D, Markel S, eds. In Silico Technology in Drug Target Identification and Validation. Boca Raton (Florida): CRC Press. pp 261–284.

73. Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 238: 54–61.

74. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 348: 85–100.

75. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) BaCelLo: A balanced subcellular localization predictor. Bioinformatics 22: e408–e416.

76. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953–971.

77. Hawkins J, Boden M (2006) Detecting and sorting targeting peptides with neural networks and support vector machines. J Bioinform Comput Biol 4: 1–18.

78. Nakai K, Horton P (1999) PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–36.

79. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, et al. (2002) Prediction of human protein function from post-translational modifications and localization features. J Mol Biol 319: 1257–1265.

80. Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13: 121–130.

81. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: A server for predicting protein function from 3D structure. Nucleic Acids Res 33: W89–W93.

82. Gruber M, Soding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140–145.

83. Ferron F, Longhi S, Canard B, Karlin D (2006) A practical overview of protein disorder prediction methods. Proteins 65: 1–14.

84. Zhou F, Xue Y, Yao X, Xu Y (2006) A general user interface for prediction servers of proteins' post-translational modification sites. Nat Protoc 1: 1318–1321.

85. Jaroszewski L, Li W, Godzik A (2002) In search for more accurate alignments in the twilight zone. Protein Sci 11: 1702–1713.

86. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4690.

87. Nayal M, Honig B (2006) On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins 63: 892–906.

88. Todd AE, Orengo CA, Thornton JM (2002) Sequence and structural differences between enzyme and nonenzyme homologs. Structure 10: 1435–1451.

89. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333: 863–882.

90. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318: 595–608.

91. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297: 233–249.

92. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.

93. Nair R, Rost B (2002) Sequence conserved for subcellular localization. Protein Sci 11: 2836–2847.