Automated protein function prediction—the genomic challenge

Iddo Friedberg

Received (in revised form): 25th January 2006

Abstract

Overwhelmed with genomic data, biologists are facing the first big post-genomic question—what do all genes do? First, not only is the volume of pure sequence and structure data growing, but its diversity is growing as well, leading to a disproportionate growth in the number of uncharacterized gene products. Consequently, established methods of gene and protein annotation, such as homology-based transfer, are annotating less data and in many cases are amplifying existing erroneous annotation. Second, there is a need for a functional annotation which is standardized and machine readable so that function prediction programs could be incorporated into larger workflows. This is problematic due to the subjective and contextual definition of protein function. Third, there is a need to assess the quality of function predictors. Again, the subjectivity of the term 'function' and the various aspects of biological function make this a challenging effort. This article briefly outlines the history of automated protein function prediction and surveys the latest innovations in all three topics.

Keywords: *computational function prediction; genomic annotation; gene ontology, phylogenomics*

INTRODUCTION

Arguably the main push that has moved bioinformatics from the wings of life sciences to center stage has been the growing influx of genomic information over the past decade. BLASTing a newly found sequence has been routine for over a decade now, and the memory of a time when sequence database searching was not an integral part of molecular biology is rapidly fading. With the exponential increase in the number of proteins being identified by sequence genomics projects and their molecular structure being determined by structure genomics projects, we are facing a sea of data from which actual information needs to be carefully distilled. Putting it simply, what do all the genes do? It is impossible to perform a functional assay for every uncharacterized gene in every genome. Moreover, it is impossible to keep up with the influx of data by manually curated annotation. Given this state of affairs, scientists have been turning to sophisticated computational methods for assistance in annotating the huge volume of sequence and structure data being produced.

This review will discuss the problems, recent solutions and future challenges for automated function prediction (AFP) in bioinformatics. The review begins with the challenge posed by merely defining protein function, the attempts to objectify this definition, and render functional annotations amenable to computational processing. Special emphasis is placed on ontologies as they are currently the dominant accepted solution to this problem. We continue to explore the techniques for function prediction: homology-based annotation transfer, phylogenomic methods, sequence patterns, structure similarity, structure patterns, genomic context, microarray data, and the latest aggregate methods. Finally, we discuss attempts at quality assessment of function prediction programs and the unique challenges it presents.

WHAT IS FUNCTION?

The definition of biological function is ambiguous, and the exact meaning of the term varies based on the context in which it is used [1★–3]. It is obvious

Iddo Friedberg, Burnham Institute for Medical Research, Program in Bioinformatics and Systems Biology, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA. E-mail: idoerg@burnham.org

Iddo Friedberg is a postdoctoral associate at the Burnham Institute in La Jolla, California. His research interests encompass a wide range of topics in structural bioinformatics that include function prediction from structures of uncharacterized proteins and the identification of 'structural signatures' which span different protein folds to better understand protein evolution, folding and functionality.

that the biological function of a protein has more than one aspect. Take for example a protein kinase; in the biochemical aspect, a kinase's function would be the phosphorylation of a hydroxyl group of a specific substrate. The scope of interest implied by this definition does not require any more than a 'disembodied' protein performing alone *in vitro*. However, proteins perform their function within an organism, and this has consequences ranging from the subcellular to the whole-organism level. In a physiological aspect, the same kinase may be part of a signaling pathway, where a protein both phosphorylates, and is phosphorylated by, interacting partners. A mutation in this kinase might cause a disease, so yet another aspect is a phenotypic or medical one. Therefore, when speaking of a protein's function, we must always specify the aspect or aspects of the functional description. When setting out to use a function prediction tool, let alone to develop one, we must keep in mind which functional aspect or aspects we are trying to predict, and use the appropriate vocabulary.

DESCRIBING FUNCTION

Having defined the functional aspect or aspects of interest, how should function be described in a computationally amenable way? Protein sequence and structure information are easily rendered machine legible. Protein sequences are represented as character strings that are suited for many tasks: pairwise and multiple alignments, motif finding, database searching, and a slew of other tasks aimed at extracting biological information from the sequence. The ability to express sequence information as a character string amenable for computational processing dovetails with algorithms able to analyse this information. As for structure, although the representation is more complex, the PDB [4] and mmCIF [5] file formats do so for most practical purposes. There is a limited, well-defined syntax involved in both cases.

In contrast to sequence and structure information, the annotation of a protein is written in human language, which conveys the subtleties and intricacies of its function as well as the experimental evidence which supports it, its research history, and other characteristics. As is accepted in human language, particularly in science where the vocabulary is invented and reinvented daily, many terms are synonymous. This synonymy is confusing to humans and even more so to machines. Other factors add to the problem, such as mixing up or omitting functional aspects of a protein. Finally, what constitutes functional information and what does not?

To make functional annotation open to computational processing, there is a need to convey it in a controlled and well-defined fashion. The need for a controlled vocabulary and well-defined relationships in describing function was first recognized by biochemists and took the form of the Enzyme Commission Classification, or EC (http://www.hem.qmul.ac.uk/iubmb/enzyme/). EC classifies reactions in a four-level hierarchy, and those are noted by a four-position identifier, going from the general 'Lyase' (4.-.-.-) in the first position through the more specific 'Nitrogen lyase' (4.3.-.-) on to 'Ammonia lyases' (4.3.1.-) to the ultimately specific 'Histidine-ammonia lyase' (4.3.1.3) in the fourth position. EC answers the requirements both for a controlled vocabulary and for a well-defined relationship between terms. However, there are other functions besides the enzymatic and other functional aspects besides the biochemical that are wanted in annotation. In 1993, the first genomic scheme to categorize gene products was suggested for *Escherichia coli* [6]. Other annotation schemes were suggested, mainly following the need to annotate budding genomic projects, and those are reviewed in [2, 7]. The common thread among those schemes is the establishment of a controlled vocabulary and in many cases a categorization that proceeds from the general to the specific.

The Gene Ontology (GO) [8] currently serves as the dominant approach for machine-legible functional annotation. GO is a framework consisting of controlled vocabularies describing three aspects of gene product function: **molecular function**, **biological process** and **cellular location.** The latter, although not a functional aspect *per se*, is deemed important for functional annotation since proteins do not operate in a vacuum (or rather in a saline solution) but within a well-defined context of the living cell. Each ontology is implemented as a directed acyclic graph (DAG) where terms are represented as nodes in the graph and are arranged from the general to the specific. The DAG arrangement means that each node may have more than a single parent—this is to describe functions that are involved in more than a single biological process, cellular compartment, or molecular function.

The GO Annotation (GOA) project's mission is to annotate genomes and various sequence and structure databases using GO terms [9]. When GO terms are assigned to a gene product, an evidence code stating how the annotation was obtained is assigned as well. In this manner, the reliability of the annotation is noted—is it based on experimental evidence that can be traced to an author (high reliability)? Or it is simply inferred by homology transfer annotation that has not been reviewed by a curator (low reliability)? GO has nine such evidence codes, and they are discussed in the GO site (http://www.geneontology.org/GO.evidence.shtml). Note that the evidence code is not a part of the ontology DAG, but rather is associated only upon assignment of a term, or terms, to a gene or gene product. There are other, more specialized ontologies that are used in more specific aspects of molecular biology or for other purposes, such as medically interesting annotation or genomic aspects unique to a given organism (http://www.obo.org). An example of GOA is provided in Figure 1.

By standardizing an annotation and defining the relationships between terms using a graph, annotations may be computationally processed [10]. Given a GO-annotated genome, a researcher can, for example, extract all the gene products involved in the synthesis of a particular metabolite. Within the scope of this discussion, GO provides a standard way for programs to output their function predictions. This is useful if those programs serve as components in more-extensive workflows, for example, a system predicting functions for all unknown ORFs in a genome and then extracting those that are predicted to be involved in a metabolic pathway of interest.

Naturally, the use of a controlled vocabulary comes at the expense of the detail and subtlety of natural language. However, GO is considered to be a good working compromise between the need to convey functional information with all its minutiae and the need to make it standardized and computer legible. GO terms and relationships are being updated constantly so that new functions, error corrections, and amended definitions can be included.

Automated function prediction by homology-based transfer

Having defined function and the means of describing function, we now turn to discussing function prediction programs. Bioinformatics' first and most-often-used class of tools have to do with searching sequence databases by using sequence similarity tools. As a matter of fact, the 1997 PSI-BLAST paper by Altschul *et al.* [11] has over 13 000 citations as of September 2005, showing how routinely sequence-based database searching is used. Chiefly, the reason that researchers BLAST protein sequences against databases is to learn about some aspect of their function. The researcher aims to answer this question by finding a significant sequence similarity to another protein that is already in the database and whose function was experimentally characterized. This is essentially the most widely used form of computational function prediction—annotation transfer by sequence similarity, also known as homology-based transfer. The biological rationale for homology-based transfer is that if two sequences have a high degree of similarity, then they have evolved from a common ancestor *and* they have similar, if not identical, functions. This statement may seem obvious to the point of being trivial; however, we shall see how a homology-based transfer may not be very reliable for functional annotation even in high-alignment identity percentages.

As databases grow in the number of sequences they hold, homology-based transfer begins to breakdown in three aspects. The first aspect is the observation that even with a high sequence similarity annotation transfer may be erroneous. Shah and Hunter [12, 13] have attempted to discriminate between enzymatic functions based on sequence alignments and have concluded that it is necessary to establish functionally significant subregions for discrimination purposes. Rost [14] has established that even at high sequence similarity rates, enzymatic function may not necessarily be conserved. Tian and Skolnick [15] have concluded that a 40% sequence identity is sufficient to establish catalytic mechanism similarity, but as for substrate similarity, a 60% identity or higher is necessary to determine that with acceptably low false-positive error rates. However, they have observed that information loss due to a high false-negative rate does occur at such high identity percentages. At the other end of the sequence identity scale, Galperin *et al.* [16] have shown that enzymes that are supposedly analogous due to undetectable sequence similarity are, in fact, homologous. Pawlowski *et al.* [17] have shown that a distant but significant sequence similarity correlates well with functional similarity, although a very high

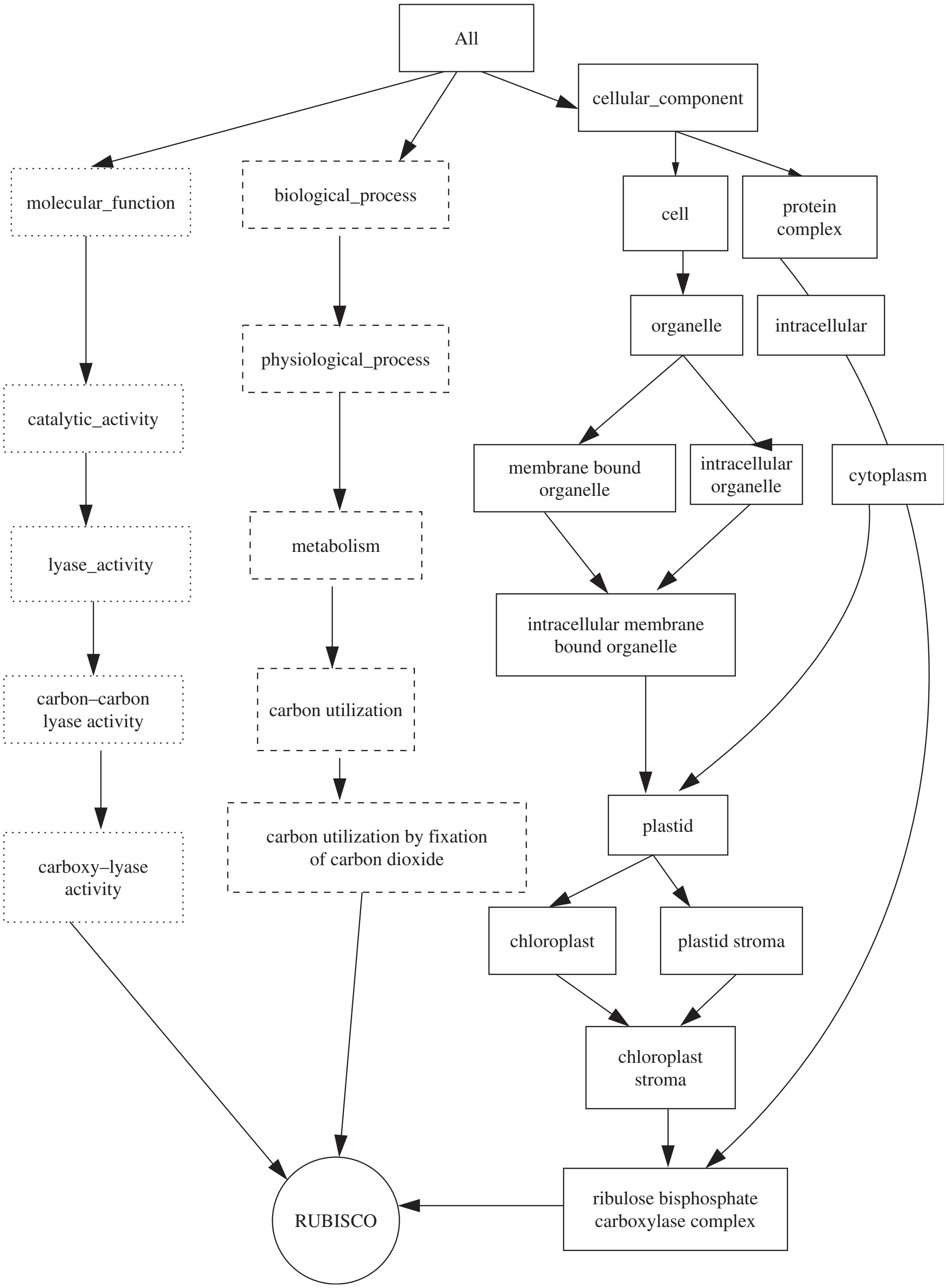

Figure 1: Gene Ontology. Gene Ontology is explained here via the annotation of Ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO), which catalyses the first step of carbon fixation in photosynthetic organisms. The molecular function, biological process, and cellular-component aspects are shown in dotted, dashed, and solid outlines, respectively. The associated terms proceed directionally, from the general to the specific. Note the use of multiple parents for some of the GO terms, for example, the 'intracellular membrane bound organelle' is a term denoting a membrane-bound organelle and an intracellular organelle. Figure is a modified GenNav image (http://mor.nlm.nih.gov/perl/gennav.pl).

sequence identity percentage is required for proteins to share all four EC numbers, and a very low (30% or less) sequence identity percentage is required to share the first [18]. Earlier research has shown that at 35% sequence identity, 60% of aligned enzymes share four EC numbers [19]. Clearly, sequence similarity is correlated to functional similarity, but exceptions are seen on both ends of the similarity scale. A related form of erroneous transfer is due to domain shuffling: the addition, deletion and redistribution of domains [20, 21]. Errors in annotations can be caused because database hits with significant e-values may occur, but the query and hit may have a different overall domain structure.

The second aspect of the breakdown of homology-based annotation transfer is that sequence-based tools are not sensitive enough to discover similarity between proteins, especially when not only the databases are growing in size, but the diversity of sequences is growing as well (Figure 2). In other words, we are not only collecting more sequences, we are collecting more new and different sequences. Homology-based transfer is even less effective since the number of clustered similar proteins for which we do not have a single annotated reference sequence is rapidly growing.

The third aspect is the propagation of erroneous annotations throughout the database. As more sequences enter the database, more are annotated by homology transfer, which increases the ability of errors to propagate and be amplified based on a single erroneous annotation [22, 23] Strangely, some of these errors can be the result of over-annotation by a misuse of the ontology method, for example, the misinterpretation of partial EC numbers such as '1.1.1.-' and consequently, the assignment of a wrong last EC number [24★]. An alarming estimate of the prevalence of incompatibility between different annotations of *M. genitalium* [25★★] (and by implication, errors) has prompted some of the error-control mechanisms discussed subsequently, such as the evidence codes in GO.

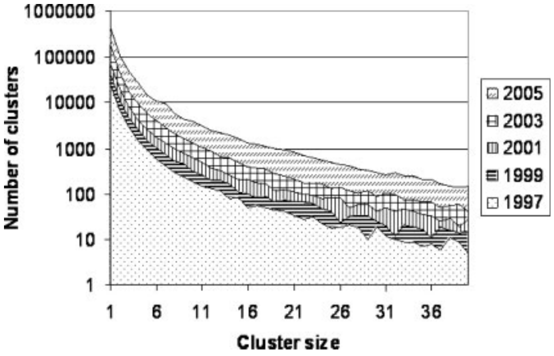

Figure 2: Increase of sequence diversity over time. GenPept sequences from the years 1997, 1999, 2001, 2003 and 2005 were clustered at 60% sequence identity, separately for each year. *X*-axis: cluster size, in number of sequences per cluster. *Y*-axis: $\log_{10}$ of number of clusters in a cluster size. As can be clearly seen, the diversity of sequences, represented by the number of clusters populated by few sequences, is growing. Sequences were clustered using CD-HIT [132].

Pfam [26] is arguably the database of choice for those seeking order within the protein sequence universe. Pfam 18 (2005) contains 7973 families composed of aligned sequences, out of which 1979 (25%) are annotated as having an unknown function. As we shall see, Pfam annotation is used by function prediction programs, either by directly querying Pfam or by using umbrella databases that include Pfam information such as InterPro [27]. SMART [28, 29], CDD [30], and PRODOM [31] are other databases consisting of multiple alignments of protein domains. All these databases have proteins arranged in homologous clusters, which, when possible, are annotated. These databases are often deferred to when producing homology-based annotation transfers. It should be emphasized that the use of these databases for homology transfer should be done with caution, as they annotate proteins on a domain level. A multi-domain query aligned to Pfam, for example, should be carefully checked for mis-annotations due to domain shuffling, as mentioned eariler. Also, the 'granularity' of these databases varies. For example, a single Pfam family may contain several proteins which perform the same enzymatic reaction on different substrates.

On a side note, adding yet more complexity to the fray, enter the meta-genomics projects: the sequencing of DNA extracted from the Azores seawater, South Pacific islands topsoil, and the subway vents of New York City have added a new level of diversity to our current body of knowledge [32, 33]. Venter and coworkers [35] have found that many gene families are much more diverse than previously thought. It has been proposed that for prokaryotic gene families, association with habitat rather than with taxonomic identity is much more informative [34]. Thankfully, the database curators did not include those sequences in the 'brand name' in databases, so as not to confuse and overload those databases with organism-dissociated sequence data. They are available, though, and harboring a trove of information on the diversity of protein families, their annotation is a massive task yet to be fully appreciated.

The programs reviewed in this section rely primarily, but not exclusively, on homology-based annotation transfer for function prediction. Mostly, programs that produce an ontological or keyword output were considered for this review. Naturally, there are many similarity-based search programs, which have been reviewed frequently in literature and will not be covered here.

PFP is a simple server that initially uses homology-based transfer for function prediction to establish function (http://dragon.bio.purdue.edu/pfp/pfp.html). It uses three iterations of PSI-BLAST against UniProt [35], and the results are cross-referenced against UniProt's GOA to produce a function prediction with a probability score. Additionally, Hawkins and Kihara have analysed the probabilistic associations in UniProt between different GO terms using function association matrices (FAMs). A FAM entry contains a score representing the strength of the association between any two GO terms per ontology, per database. This means that for every GO term, the probability of its being associated with another GO term can be looked up. Given a homology-based function prediction produced by PSI-BLAST, additional functions may be associated with a known degree of probability. Thus, sensitivity is increased by this association, but specificity is not compromised as the method uses a low number of PSI-BLAST iterations. To be clear, 'sensitivity' denotes the fraction of true-positive GO-term predictions out of all true positives, and 'specificity' addresses the fraction of true positives out of all predictions (true and false-positives). The low number of PSI-BLAST iterations is important, as it was shown that above a certain number of PSI-BLAST iterations, alignment accuracy drops remarkably, compromising the ability to perform an accurate homology transfer [36].

GOPET [37] uses multiple support vector machines (SVMs) to classify predictions into 'correct' and 'incorrect' categories. The features used for classification include sequence similarity measures, frequency of GO terms, quality of annotation of homologs, and annotation level within the GO hierarchy. Initially, each training sequence—assigned with known GO terms—is BLASTed against protein databases. The annotations of the retrieved sequences are compared with the standard of truth of the training annotations and are classified as either 'correct' or 'incorrect,' and the features listed previously are noted. Those features are then used to construct an SVM to differentiate between the correct and incorrect GOAs. Given a training sequence, the GOAs of the resulting sequences are noted, and their features are calculated and mapped into the feature space. Then, the correct/incorrect labels are assigned. Prediction accuracy is further enhanced using a voting approach to combine information from different classifiers. With more classifiers pointing to a given GO term, the prediction is considered more reliable. OntoBlast [38] and GOblet [39, 40] provide a GO-based output of BLAST, with BLAST's e-value used for scoring the results. GOtcha [41] offers a high-sensitivity search using BLAST—the probability for each GO term is normalized by its frequency in the genome, and a sub-DAG of the GO nodes is provided for each query with the term probabilities assigned.

PHYLOGENOMIC CONSIDERATIONS

The output of BLAST is a list of proteins ordered by increasing e-value, roughly corresponding to a descending similarity score between the query sequence and those found similar to it in the database. Intuitively, the top sequence would be taken as the basis for annotation transfer to the query sequence. This may not always be the correct thing to do. This is because the sequence most similar to the query is not necessarily the one identical in function. The reason is that when gene duplication occurs in an organism, the duplicate, termed a paralog, is 'free' to assume a new function, whereas in orthologs—homologs due to speciation—function is more likely preserved. However, due to database bias or unequal evolution rates, highly similar sequences may not be from orthologous proteins, not sharing the same function, and thus, promoting erroneous annotation transfers (Figure 3). This is where phylogenomic considerations come into play. The term phylogenomics, coined in 1998 by Jonathan Eisen [42★★], denotes the application of phylogenetic information to genomic studies. Phylogenomics states that the evolutionary history of putative homologs must also be considered when assigning function. In practical terms, annotation

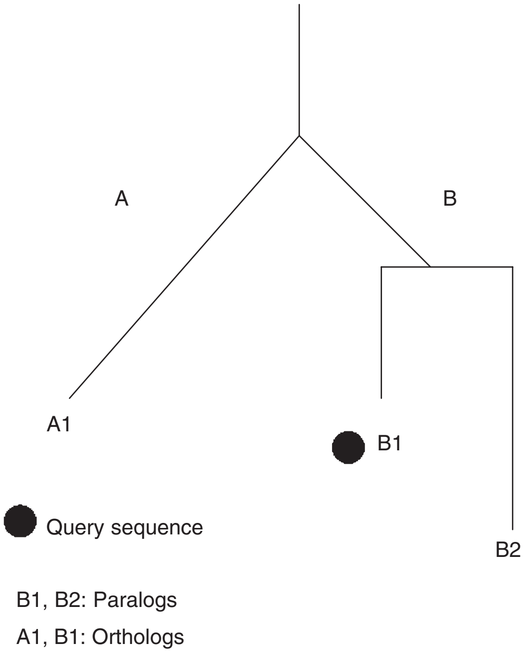

Figure 3: Similar as opposed to orthologous. Given two organisms A and B. Genes B1 and B2 are paralogous, as they share a common ancestor prior to a duplication event. A1 is orthologous to B1, B2. B1 is more similar to B2 than either B proteins are to A1. However, B1 and B2 have different functions, and A1 actually shares the same function with B1. Therefore, given B1 as a query sequence, annotation transfer by similarity would produce an error, as it will transfer function from B2. Given knowledge of the evolutionary tree, A1 should be used for functional annotation transfer.

transfer is performed from the closest ortholog, not from the most similar sequence. There are several methods based on this insight, known as phylogenomics-based methods. These methods are ideally used when BLAST produces two or more different annotations, and the way to discriminate and decide which one is correct is by considering the evolutionary history of the homologs in question. RIO [43] and SIFTER [44] check whether a protein has orthologs, by inferring gene duplications on a gene tree by comparing it with a species tree, thus sorting out the orthologous and paralogous events. Annotation transfer is performed from those with a good orthology score, not with a good similarity score. OrthoStrapper [45] uses the same principle, but the phylogenetic tree is inferred from many bootstrapped trees. A recent review of phylogenomic methods outlining a generic method to perform phylogenomic analyses was written by Kimmen Sjölander [46★★].

THE USE OF SEQUENCE PATTERNS

Proteins that share a common function but are otherwise diverse will usually share one or more common sequence or structure patterns necessary to maintain their structure and function. This is because proteins perform their functions using a relatively small part of their structure—of the 100–300 amino acids in a domain, only fewer than 10 may make up the ‘business ends’ of the protein that are the binding and catalytic sites. Therefore, to predict a protein’s function, in many cases there is no need to use annotation transfer from a homologous protein. All that is required is to identify a sequence- or structure-based signature or feature that can be associated with a function. This signature can either be located in one place along the sequence or be a ‘fingerprint’ composed of several such patterns. The canonical PROSITE [47] is a well-established database of such patterns, constructed by manual curation of multiple sequence alignments. The PROSITE language uses a regular-expression-like syntax consisting of the one-letter codes for amino acids. However, there are many different techniques for collating, compiling and recognizing sequence patterns. These include variations on regular expression syntax, multiple sequence alignment profiles, and hidden Markov models, to name a few. This is a well-established field in bioinformatics and is too broad to be included in this review. It is mentioned because sequence patterns are essential in the makeup of many recent function prediction servers. Also, the umbrella sequence databases, such as InterPro, provide functional information based on the presence of canonical sequence motifs.

Databases of sequence patterns usually contain cross-references to whole protein chains and their possible function. PROSITE is one, and PRINTS is another motif (or ‘fingerprints’) database [48]. Some databases integrate the information from others. One of the first was BLOCKS [49], which had its own motifs generated using the PROTOMAT program, but also integrated PRINTS, PROSITE, and Pfam information. Currently, one of the most comprehensive and heavily used resources is InterPro, which was mentioned earlier. InterPro contains multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins automatically generated from PRINTS. PRINTS is an annotation resource that includes ‘signatures’ (another synonym for sequence patterns) from PROSITE and whole-domain signatures from

PANTHER [50], Pfam and TIGRFAM. As we shall see, databases such as InterPro are often used for predicting protein structure by annotation transfer. Due to this 'cross talk' between the databases, it is crucial that the information be kept correct and up-to-date.

GENOMIC CONTEXT-BASED PREDICTIONS

One exciting class of methods for assigning function using sequence, but without using homology-based transfer, is the genomic context-based predictions. This technique has also been dubbed phylogenomic or phylogenetic profiling, but to avoid confusion with the phylogenomic methods described earlier, I shall use the term 'context-based'. The inter-genomic profile of a gene is a vector of bits, each bit representing the presence or absence of a homolog in a given genome. Two proteins with identical or nearly identical inter-genomic profiles are considered to have evolved together, and it is very likely that they are functionally associated [51★–54★★, 55]. Furthermore, in prokaryotic genomes, functionally related genes also tend to be co-located on the chromosome. These two observations have been translated into function-prediction algorithms, using co-evolution, chromosomal proximity, or both as functional predictors. Function is inferred by matching the inter-genomic profiles of the unknown protein to those which are known. The Phydbac2 server [56, 57] makes use of this trait for functional annotation. However, in addition, Phydbac2 uses chromosomal proximity and gene fusion analysis. In bacteria, having two genes co-located on a chromosome and both homology and co-location preserved in many genomes, indicates a functional association, and thus may be used for function prediction. Additionally, certain genes may fuse into two domains on the same gene, clearly showing a functional association. SNAP [58] takes this approach one step further and builds a graph of altering similarity and neighborhood relationships in completely sequenced bacterial genomes—that is, groups of genes are typified both by their co-location on genomes and by homology, although co-location need not necessarily imply actual genomic proximity. Genes with similar similarity–neighborhood graphs (*SN*-cycles) are considered to be functionally associated. In a more recent study, Barker and Pagel [59] have incorporated the phylogeny of 15 studied species into the absence/presence data of gene pairs to produce functional association predictions.

It should be noted that other contextual prediction methods exist, notably by inference from protein–protein interactions. However, this review surveys the genomic contextual methods. For comprehensive reviews on functional inference by interactions, see the recommended reading list at the end of this article.

STRUCTURE ALIGNMENT AND STRUCTURE PATTERNS

The structure of a protein is much more informative than the amino acid sequence alone. Knowing the structure allows us to explain the biochemical mechanism by which the protein implements its functionality. If the function is unknown, a three-dimensional (3D) structure-based similarity to other structures can reveal more about its function. If the 3D structure is of a known fold, then that protein may have a function similar or identical to other proteins inhabiting this fold. Furthermore, structure is much better preserved than sequence, so many proteins with little or no sequence similarity still have a structural similarity [60, 61★★]. There are a few folds that are inhabited by many proteins performing many different functions, but most folds we know of are associated with a single function, though this may be simply due to an uneven sampling of sequence space [62]. Also, those folds that are functionally diverse are in many cases within a given functional milieu—TIM barrels are usually enzymes that catalyse a reaction that has something to do with oligo- or polysaccharides. After aligning the novel protein with those from its fold, the functional transfer may be assessed for plausibility by examining the similarity of the catalytic site, which should have conserved amino acids and side-chain orientations. Programs that scan PDB for structural similarity given a query sequence include CE [63], DALI/FSSP [64, 65], FATCAT [66], VAST [67], FAST [68], Matras [69, 70], DaliLite [71], GRATH [72], and AnnoLite. A critical review of several structure alignment servers has been performed, with the conclusion that multiple methods should be used as no single method is completely accurate [73].

Of the methods mentioned here, Annolite was written explicitly as a structure alignment-based function-prediction tool. Annolite (http://www.salilab.org/~marcius/beta_dbali/?page=tools

Table 1: Function prediction by database scanning using structure alignment

Name	Description	URL	Reference
Global structure similarity search			
DALI	Search and optimal alignment using distance matrices	http://www.ebi.ac.uk/dali/	[64, 65]
VAST	Vector alignment	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	[67]
CE	Combinatorial extension	http://cl.sdsc.edu/	[63]
FATCAT	Flexible alignment of protein backbone	http://fatcat.burnham.org/	[60]
AnnoLite	Homology transfer from functionally annotated databases	http://salilab.org/DBAli/?page=tools&action=f.annolitechain	
Matras	Markov transition model of evolution	http://biunit.aist-nara.ac.jp/Matras/	[69]

&action=f_annolitechain) accepts a query structure and provides GOA along with a probability score. This is done by annotation transfer from similar structures.

However, even when confronted by a protein with a novel fold or with low similarity to a known fold, it is still possible to tease out functional information. This can be done by analysing structure patterns of the protein. The rationale for structure patterns is the same as for sequence-based patterns—identify unique markers associated with a function. However, structure patterns by nature are markedly different from sequence patterns. As we have seen, sequence patterns are short amino acid stretches, similar over different proteins, associated with a function. These can be represented as consecutive short regular expressions or profiles derived from multiple sequence alignments. Structure patterns are best described as identifiable spatial regions within the protein's structure, which leaves much more room for descriptive methods. Those range from identification of 3D shapes completely dissociated from the amino acids, to a string of characters representing amino acids and their physical environment. A typical prediction by the 3D motif method contains the following elements: an algorithm to generate a 3D motif library, a library of 3D patterns, and a search algorithm designed to scan that library. The rest of this section reviews existing methods. It should be noted that most of them fall short of being full AFP methods as they do not produce a functional prediction for the query protein. Rather, they show the mapping of putative functional sites, leaving the final compilation of results up to the user. Nevertheless, they are reviewed here due to their importance in providing functional annotation. A table reviewing some of these methods is maintained by Martin Jambon at (http://martin.jambon.free.fr/search-protein-3D-sites.html). The functional association of structure fragments is very strong and has been shown to transcend different folds [74, 75].

(Patterns In Non-homologous Tertiary Structures (PINTS) [76] allows the comparison of a protein structure against a database of patterns, or a PDB format pattern against a pattern database. PHUNCTIONER [77] extracts conserved residues and associates them with GO terms. Several more such databases and associated search algorithms exist, including the Catalytic Site Atlas [78★★], PDBFun [79], PDBSite and PDBSiteScan [80, 81], SuMo [82, 83], pvSOAR [84], SARIG [85], FEATURE [86, 87], RIGOR [88] and PatchFinder [89] (Table 2).

By default, structures are represented as a collection of thousands of 3D coordinates, which represent the atoms making up the protein. This representation is computationally expensive. To reduce this expense, many algorithms seek to represent the 3D structure more simply while preserving the necessary spatial and physicochemical information. For example, SuMo uses a chemical group representation for residues, arranged in triangles, and graph theory to superimpose them for database searching. FEATURE defines microenvironments in the protein structure as a series of concentric spheres, or alternatively as a 3D cubic lattice. Each spatial partition (cube or sphere) is then examined for physicochemical properties of the residues enclosed in it, and based on those, a feature vector is assigned to it. SeqFEATURE [87] further automates the process of site creation by enabling the creation of structure patterns from known sequence patterns. Theoretical Microscopic Titration Curves

Table 2: Structure motif finding methods

Name	Functional site representation	URL
SuMo	Triplets of functional groups	http://sumo-pbil.ibcp.fr/
PINTS	Selected atoms	http://www.russell.embl.de/pints/
PDBFun	Sidechain centroid	http://pdbfun.uniroma2.it/
PDBSite	Backbone centroid	http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsitescan/
SARIG		
RIGOR	C-alpha	http://portray.bmc.uu.se/cgi-bin/spasm/scripts/spasm.pl
pvSOAR	Side chain centroids of cavity amino acids	http://pvsoar.bioengr.uic.edu/
WEBFEATURE	Concentric spheres	http://feature.stanford.edu/webfeature/
THEMATICS	Predicted pH titration curves	No server associated yet.

(THEMATICS) [90, 91] uses computed pH titration curves of residues on a protein structure to locate active residues and differentiate between different proteins that may share similar structures but different functions.

Locating the functional amino acid residue positions can be done through evolutionary considerations. Evolutionary Trace (ET) ranks residue importance by correlating amino acid variations in a multiple sequence alignment with evolutionary divergences in a phylogenetic tree [92]. As such, it is reminiscent of (although predating) the phylogenomic methods. The obvious difference being that ET maps groups of residues rather than whole proteins. ET has been shown to correlate well with functional sites and as such serves as the basis for several methods for structure-based functional site detection [93–95].

A good structure feature-based prediction method will be able to recognize structural features in a given protein and integrate that information into a functional prediction. It is up to the algorithm designer to locate the important cues for functionality and use them to deliver a good prediction for which bits are important. It bears repeating that the programs listed in this section are not strictly function predictors as they do not produce a list of keywords/ontology terms or free text as their output, but rather are a stepping stone toward understanding a protein's function.

EXPRESSION MICROARRAY-BASED PREDICTIONS

'Birds of a feather flock together' (conversely, 'guilt by association' [96★★]) is the motto underlying this category of function-prediction methods. When gene-expression data are clustered, genes involved in the same pathways tend to cluster together. Eisen *et al.* [97★] proposed a metric based on correlation of expression intensity and then proceeded to show that genes involved in similar cellular functions are co-expressed. From their results they suggested that unknown genes, co-expressing with known genes, can be annotated by virtue of association. Other methods based on the same rationale have been proposed [98★]. It is clear that this method is useful for annotating the cellular pathway aspect of function rather than the molecular function itself. If we find an unknown gene in the cluster of genes responsible for cholesterol metabolism, we can safely infer that it too has to do with cholesterol metabolism. Pinpointing its exact biochemical function, however, is still problematic.

Expression microarray analysis for the purpose of function prediction will not be covered here since even a proper introduction to the subject is beyond the scope of this review. A review of ontological analysis of gene-expression data was recently published and is recommended reading for understanding the current scope of the tools that mediate an understanding of the biological phenomena underlying gene-expression data [99]. Many GO-based tools are available for this type of analysis from the GO site (http://geneontology.org/GO.tools.microarray.shtml).

OTHER PREDICTION METHODS

This section offers an overview of several 'canonical', and also recently described function-prediction servers and available software, mostly those that produce ontological output—exclusively or in addition to other functional output—in response to a structure- or sequence-based input.

GeneQuiz [100] was probably the first high-profile function-prediction server, having been the first available, and one of the best known by virtue of seniority. It used a variety of global sequence and sequence pattern search strategies (mainly from the BLAST family).

PropSearch [101] is probably the first method to use a purely feature-oriented approach to function prediction. ProtFun [102★★] is the current best-known player in this category. Sequence homology-based methods classify proteins within sequence space, and structure-based methods classify them within structure space. In contrast to those, a mixed-feature method such as PropSearch or ProtFun uses a series of predicted biophysical and localization signals as well as post-translational modification features to classify proteins. Among the attributes ProtFun uses are a number of positively or negatively charged residues, grand average of hydrophobicity [103], secondary structure prediction, low-complexity regions, and predicted glycosylation sites. ProtFun uses a neural network in which those features are used in calculating the probability a query sequence belongs in any one of twelve biological processes (e.g., 'energy metabolism' or 'replication and transcription' as outlined in [7]) as well as classifying it as an enzyme or non-enzyme and, if the former, assigning it a top-level EC classification. Although ProtFun uses a controlled vocabulary, the predictions provided are coarsely grained and are more suitable for a statistical overview of a large group of genes or a whole genome or for obtaining preliminary functional hints for later analysis.

Spearmint [104] generates rules for annotation based on SwissProt keywords rather than on GOA. The classifying attributes are taken from pattern databases such as PROSITE, Pfam and PRODOM, and an association between these and SwissProt keywords is determined and put in the form of a C4.5 decision tree [105]. Rulebase is similar to Spearmint in essence, only differing in that the rules are generated by manual curation rather than by an automated approach. However, this rule-based association is reliable as long as the SwissProt keyword annotation is reliable, which is often not the case; thus, both Rulebase and Spearmint may introduce erroneous annotations. For this reason, Xanthippe, a rule-based system aimed to detect and flag annotation errors, was constructed [106] to check inconsistencies in rule associations. (Xanthippe was Socrates's critical and shrewish wife).

Proteome Analyst (PA) uses BLAST and then assigns a set of attributes to the protein, based on SwissProt keywords drawn from the nearest homolog; it also enables the user to create custom classifiers [107, 108].

ProKnow [109] is a hybrid server utilizing information from sequence and structure (if available). Using multiple sequence alignment, multiple structure alignment, and Bayesian statistics, it outputs GO terms along with associated probability scores and strength of evidence (or 'clues') outlining how the predictions were made. ProFunc [110] is a mixed method that uses fold matching, residue conservation, surface cleft analysis, and functional 3D templates to identify both the protein's likely active site and its possible homologs in the PDB. Although ProKnow does not produce a prediction *per se*, it does collate and present information in a systematic way. JAFA is an aggregate server that queries other programs (currently GOtcha, PhydBac, GOblet, InterproScan and GOfigure) and presents the final results using GO in a concise, non-redundant fashion (http://jafa.burnham.org).

STRING [111★★, 112] integrates information from genomic proximity conserved along genomes, high-throughput experiments, co-expression conserved along genomes, phylogenetic profiling and literature mining. It predicts functional associations as well as direct interactions.

FSSA [113] is a method to discriminate between proteins that share the same overall fold but may have different functions. FSSA does so by locating positions that are highly conserved in multiple structural alignments. Then, it locates those that may be conserved to maintain function only, versus those that are conserved to maintain structure. It does so by calculating the conservation within a SCOP [114]-fold versus the conservation in a superfamily (more closely related proteins of the same or highly similar functions). The rationale is that structure-maintaining positions will be conserved in multiple alignments throughout the fold, whereas function-maintaining positions will be conserved only within a superfamily.

Assessment of automated function prediction

In the last section of this review, we will consider ways to assess how well AFP programs are performing. Annotation quality assessment is necessary to obtain an impression of the prediction quality of each server and of how well the field is performing in general. GASP, launched in 1999, was arguably the first genome-wide annotation quality assessment of its kind [115]. However, the participants noted that as no 'gold standard' for function exists, it is impossible to assess the quality of functional annotation. Another quality asessment was held as part of the BioCreAtIvE (Critical Assessment of Information Extraction for Biology) biological text-mining competition [116, 117]. Here, the organizers used GO-annotated genes with a good evidence code, and the function-prediction challenge (task 2, part 3 of BioCreAtIvE) was to assign a GO term using text mining. The CASP6 [118] assessment was not of AFP methods but rather of those that were handcrafted and submitted in addition to the main focus of CASP, namely that of protein structure prediction. Another assessment was conducted as part of the meeting on AFP that took place in 2005 (http://BioFunctionPrediction.org). The AFP assessment tested predictions for proteins whose functions were experimentally determined but were not yet public knowledge. As such, the AFP assessment had a 'gold standard,' and it was masked from the assessors.

For evaluating assessments, the measure used at the AFP 2005 meeting was *semantic similarity* as applied to gene ontologies [119★★]. Each GO term is assigned a frequency based on its frequency in the corpus of proteins and the cumulative frequency of its children nodes. Thus, the root node of the ontology (**molecular function**, **biological process**, or **cellular location**) always has a frequency of one. In order to measure the distance between any two nodes, the minimal subsuming node is found, and its frequency is translated into a similarity measure—the higher the frequency, the lesser the similarity. This reflects the observation that if two terms have a minimal subsumer relatively common in the corpus, that is, if they have a high frequency, then that parent is not a very informative term. For example, if 'enzyme activity' is the minimal subsumer of the two terms in question, then the terms are not very similar. If 'tyrosine kinase activity' a less-frequent term, is the minimal subsumer, then the terms would be more similar than in the former example. See Figure 4 for details. This similarity measure has been shown to be well correlated with sequence similarity [119★★]. Other similarity scores have been suggested for assessing functional similarity based on DAG levels [120] and on frequency in the corpus [121★, 122].

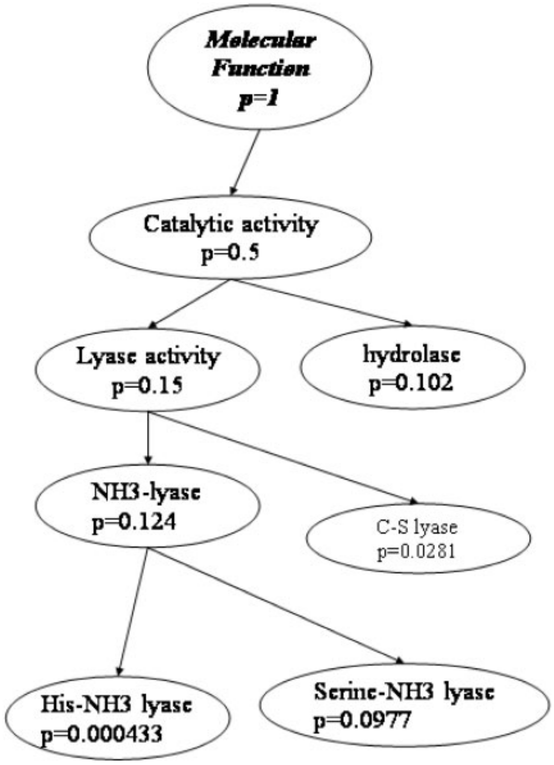

Figure 4: Semantic similarity between two GO terms. Using the method developed in [119**], each node is annotated with its frequency in UniProt. This includes the sum of the probabilities of all the nodes it subsumes, thus the root, 'molecular.function'; in this case, it has $P = 1$. Semantic similarity between any two nodes is inversely proportional to the probability of the minimal subsuming node of the two nodes in question. Thus, the semantic similarity distance between more specific terms, 'His-NH$_3$-lyase activity' and 'Ser-NH$_3$-lyase activity' (subsumed by 'NH3-lyase activity'), is shorter than 'NH$_3$-lyase activity' and 'C-S lyase activity' (subsumed by 'lyase activity'), even though the number of edges counted is the same. The actual equation used to calculate the distance is: $\text{sim}(c_1, c_2) = -\ln(P_{ms}(c_1, c_2))$. Where c_1 and c_2 are the two terms between which the distance is being measured, P_{ms} is the probability of the minimal subsumer of c_1 and c_2, and $\text{sim}(c_1, c_2)$ is the similarity score. See [119**] for a full description.

A completely different assessment scale was suggested by Ouzounis and Karp [123]—the transitive annotation-based scale (TABS), for assessing functional annotation. The difference between TABS and the methods described earlier is that TABS does not rely on an association between the functional terms, such as a tree or a DAG, for deriving a distance measure. TABS is an ordinal scale ranging from 0 (a complete agreement between the original annotation and the predicted annotation),

Table 3: Transitive annotation-based scale (TABS): a qualitative distance scale for the assessment of annotation reproducibility in genome projects

Score	Description	Comment
7	False positive	Original annotation predicts function without any supporting evidence.
6	Over-prediction	Original annotation predicts a specific biochemical function without sufficient supporting evidence.
5	Domain error	Original annotation overlooks different domain structure of query and reference proteins.
4	False negative	Original annotation does not provide predicted function although there is sufficient evidence to characterize the query protein.
3	Under-prediction	Original annotation predicts a nonspecific biochemical function although a more detailed prediction could have been made.
2	Undefined source	Original annotation contains undefined terms, non-homology-based predictions, and so on.
1	Typographical error	Original annotation contains typographical errors that may be propagated in the database.
0	Total agreement	Original annotation is correct, but annotations may be only semantically (but not computationally) identical.

Score, the score between two assignments; Description, a description of the potential disagreement between two projects; Comment, explanatory comments for ranking/scores. We consider scores of 0–3 as relatively benign compared with scores of 4–7 as the latter have a much more significant impact on genome sequence and database quality.

through 3 (under prediction), to 7 (false positive, see Table 3). TABS is used to locate 'problem spots' along a genome by comparing annotation schemes and seeing which locations receive the worst TABS scores.

However, there is great variance in the predictive scope of different methods [19, 25**]. For example, the phylogenomic-based methods are particularly suited when there are several choices for the function through homology-based annotation transfer, but the correct choice needs to be resolved [44, 46**]. This is because phylogenomic methods are able to make the distinction between similar proteins and directly orthologous ones and provide the correct sequence for orthology (and not merely similarity)-based transfer. In contrast, phylogenomic methods are not suited when a gene's function is completely unknown and cannot be found by homology-based transfer. Structural similarities (if a solved structure exists), phylogenetic profiling, or guilt-by-association microarray-based predictions might be more suitable. Additionally, some methods do not even propose to predict molecular function,but are concerned with pathway information, or subcellular location [1].

The main difference between the AFP2005 assessment and the CASP6 assessment was that in the case of CASP6 there was no 'functional gold standard' as the targets provided were used primarily for structure prediction (CASP's main interest) and had little or no knowledge regarding true function. As such, the CASP6 endeavor was geared toward setting standards for function prediction rather than assessing performance. Additionally, the predictions were generally not automatic but handcrafted as is the norm in CASP's main venue, assessment of structure predictions. Based on the discussions that took place in the CASP6 meeting, the conclusions of Soro and Tramontano [118] were as follows: (1) molecular function-prediction assessment should be limited to enzymes, where ambiguity is minimized and the framework provided by EC enables a good comparison and (2) making the prediction results useful for the experimental community is a top priority, and as such, a reliability ranking should be given to the predictions. BioCreAtIvE was focused on text-mining capabilities, and function prediction (using text-mining tools) was merely one in a large number of tasks.

DISCUSSION

There are several goals that function prediction needs to meet in the genomic era, the obvious one being improvement of annotation quality and genomic coverage—the proportion of genes and gene products in a genome which are annotated. Progress in function annotation will be measured chiefly by improvements upon these two traits. A less obvious but equally important goal is that of maintaining a good standard for information exchange. Programs dealing with sequences, sequence alignments, molecular structures and microarray data, all have one standard or a few standards for communicating information. This enables, for example, one program to read a set of sequences from a database, another to perform a multiple alignment, and yet a third to resolve conserved positions in the alignment. In contrast, with respect to function annotation, we are just beginning to set the standards for information exchange. Part of it is because—as it is with any young field defining itself—there are several conflicting suggestions being put forth, some of which will eventually propagate and be accepted. Nevertheless, a large part of the problem is due to the ambiguous and multifaceted character of biological function. This ambiguity makes it all the more difficult to set rigorous standards for prediction quality assessment without seeming unduly arbitrary and exclusive. Ontologies, chiefly GOs, have been gaining ground as the standard for information exchange about function, especially within the milieu of microarray data analysis. Even with the acceptance of GO, technical problems such as assessment of prediction quality are not yet satisfactorily resolved, as discussed in the 'Assessment of automated function prediction' section. A working information exchange standard is necessary in order to be able to 'plug in' function-prediction programs into the grander scheme of genomic analysis.

Function-related information is rarely complete, different functional aspects are slow to reveal themselves. Moreover, the scope of function annotation incompleteness is unknown. This is best exemplified in 'moonlighting proteins,' which perform several different functions in a context-dependent manner, and it can sometimes take years to discover that a protein having one function also possesses another [124★★]. This is in contrast to sequence and structure information in which, barring small errors, the information is completely known (sequence), or at the very least the scope of what is unknown is well demarcated and may sometimes be correctly filled in by predictive schemes, such as loop prediction in structures.

Distinct classes of function prediction are leading to the emergence of aggregate programs such as ProKnow, STRING and ProFunc, which use several data sources and/or algorithms to predict the function. Intuitively, aggregate methods are especially well suited for function prediction as they would need to use a wide set of features to predict different functional aspects. For example, while prediction of biochemical function can be well addressed by analysing sequence motifs, the physiological aspect might be better done by looking at the genomic context.

As for any predictor, feature selection is of utmost importance. The term 'feature selection' is used in machine learning to describe the selection of traits, or 'features' used in any classification problem. For example, when classifying schoolchildren into age groups, height is a good classifying feature. Studies assessing the best features for function prediction are only emerging now [125–128★]. Better feature-selection systems will not only enable us to create better predictors, but will also enable us to understand which chemical and physical elements in the protein are important for establishing function. Several good reviews about function prediction have been published recently, and those are recommended for additional information and other perspectives on the subject [1, 2, 21, 22, 129–131].

The expectations from AFP should be on a practical level, as by definition any controlled (or 'limited') vocabulary cannot capture the complex and multifaceted nature of function as well as natural language. Having entered the genomic era, the life sciences community faces a formidable task—to annotate the hundreds of genomes being sequenced and the structures being solved. AFP will play a pivotal role in this effort. Happily, computational biologists are gearing up to meet this challenge, as attested to by the many different methods surveyed in this article.

Key Points

- Few functional annotations are derived by experiments, and most functional annotations are automated. The exponential growth in sequence diversity means that the method of choice, homology transfer, is not performing as well as it used to.
- Nucleotide or amino acid sequence, sequence patterns, protein structure patterns, chromosomal location, phylogenetic information, expression data, molecular interaction data and gene co-evolution are all being used for function prediction.
- Different methods are better at predicting certain functional aspects. Combined approaches drawing on the strengths of different methods are currently emerging.
- Functional annotation can be confusing and ambiguous; ontologies are the tool of choice for standardizing annotation.

Acknowledgements

Many thanks to Steven Brenner, Cindy Cook, Alexey Eroshkin, Adam Godzik, Inbal Halperin, Martin Jambon and Sheila Podell for critical reading and sage advice. Thanks to Weizhong Li for providing an advance copy of a new version of CD-HIT, used in generating the data in Figure 2. This study was supported by the National Institutes of Health grant 1P01-GM63208.

References

★ Papers of particular interest

★★ Papers of extreme interest

1. ★ Rost B, Liu J, Nair R, *et al*. Automatic prediction of protein function. *Cell Mol Life Sci* 2003;**60**:2637–50.
2. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 2003; **36**:307–40.
3. Skolnick J, Fetrow JS. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *Trends Biotechnol* 2000;**18**:34–9.
4. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
5. Westbrook JD, Fitzgerald PM. The PDB format, mmCIF, and other data formats. *Methods Biochem Anal* 2003;**44**:161–79.
6. Riley M. Functions of the gene products of *Escherichia coli*. *Microbiol Rev* 1993;**57**:862–952.
7. Rison SC, Hodgman TC, Thornton JM. Comparison of functional annotation schemes for genomes. *Funct Integr Genomics* 2000;**1**:56–69.
8. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**:25–9.
9. Camon E, Barrell D, Lee V, *et al*. The Gene Ontology Annotation (GOA) Database – an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol* 2004;**4**:5–6.
10. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. *Genome Biology* 2005;**6**:R46.
11. Altschul S, Madden T, Schaffer A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 1997;**25**: 3389–402.
12. Shah I, Hunter L. Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**:276–83.
13. Shah I, Hunter L. Identification of divergent functions in homologous proteins by induction over conserved modules. *Proc Int Conf Intell Syst Mol Biol* 1998;**6**:157–64.
14. Rost B. Enzyme function less conserved than anticipated. *J Mol Biol* 2002;**318**:595–608.
15. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity?. *J Mol Biol* 2003;**333**:863–82.
16. Galperin MY, Walker DR, Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* 1998;**8**:779–90.
17. Pawlowski K, Jaroszewski L, Rychlewski L, Godzik A. Sensitive sequence comparison as protein function predictor. *Pac Symp Biocomput* 2000;42–53.
18. Valencia A. Automatic annotation of protein function. *Curr Opin Struct Biol* 2005;**15**:267–74.
19. Devos D, Valencia A. Practical limits of function prediction. *Proteins* 2000;**41**:98–107.
20. Doolittle RF, Bork P. Evolutionarily mobile modules in proteins. *Sci Am* 1993;**269**:50–6.
21. Doolittle RF. The multiplicity of domains in proteins. *Annu Rev Biochem* 1995;**64**:287–314.
22. Bork P. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res* 2000;**10**:398–400.
23. Gilks WR, Audit B, de Angelis D, *et al*. Percolation of annotation errors through hierarchically structured protein sequence databases. *Math Biosci* 2005;**193**:223–34.
24. ★ Green ML, Karp PD. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucl Acids Res* 2005;**33**:4035–9.
25. ★★ Brenner SE. Errors in genome annotation. *Trends Genet* 1999;**15**:132–33.
    An early comparison of the annotation of a genome by three different methods shows very little agreement between them.
26. Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;**32**:D138–41.
27. Mulder NJ, Apweiler R, Attwood TK, *et al*. InterPro, progress and status in 2005. *Nucl Acids Res* 2005;**33**:D201–5.
28. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *PNAS* 1998;**95**:5857–64.
29. Letunic I, Copley RR, Schmidt S, *et al*. SMART 4.0: towards genomic data integration. *Nucl Acids Res* 2004;**32**: D142–4.
30. Marchler-Bauer A, Panchenko AR, Shoemaker BA, *et al*. CDD: a database of conserved domain alignments with links to domain three-dimensional structure 10.1093/nar/30.1.281. *Nucl Acids Res* 2002;**30**:281–3.
31. Servant F, Bru C, Carrere S, *et al*. ProDom: automated clustering of homologous domains. *Brief Bioinform* 2002;**3**: 246–51.
32. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004;**304**:66–74.
33. Eyers L, George I, Schuler L, *et al*. Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol* 2004;**66**:123–30.
34. Rodriguez-Valera F. Environmental genomics, the big picture?. *FEMS Microbiol Lett* 2004;**231**:153–8.
35. Apweiler R, Bairoch A, Wu CH. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:D115–9.
36. Friedberg I, Kaplan T, Margalit H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci* 2000;**9**:2278–84.
37. Vinayagam A, Konig R, Moormann J, *et al*. Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics* 2004;**5**:116.
38. Zehetner G. OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res* 2003;**31**:3799–803.
39. Hennig S, Groth D, Lehrach H. Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Res* 2003;**31**:3712–5.
40. Groth D, Lehrach H, Hennig S. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. *Nucleic Acids Res* 2004;**32**:W313–7.
41. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* 2004;**5**:178.
42. ★★ Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 1998;**8**:163–7.
    Introduces phylogenomics to the study of functional analysis.

43. Zmasek C, Eddy S. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002;**3**:14.

44. Engelhardt BE, Jordan MI, Kathryn Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol* 2005;**1**:e45.

45. Storm CE, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002;**18**:92–99.

46. ** Sjölander K. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 2004;**20**:170–9.
An in-depth review of the phylogenomic methodology.

47. Hulo N, Sigrist CJA, Le Saux V, *et al*. Recent improvements to the PROSITE database. *Nucleic Acids Res* 2004;**32**: D134–7.

48. Attwood TK, Bradley P, Flower DR, *et al*. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 2003; **31**:400–2.

49. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* 2000;**28**:228–30.

50. Mi H, Lazareva-Ulitsky B, Loo R, *et al*. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005;**33**:D284–8.

51. * Gaasterland T, Ragan MA. Constructing multigenome views of whole microbial genomes. *Microb Comp Genomics* 1998;**3**:177–92.

52. Gaasterland T, Ragan MA. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics* 1998;**3**:199–217.

53. Ragan MA, Gaasterland T. Microbial genescapes: a prokaryotic view of the yeast genome. *Microb Comp Genomics* 1998;**3**:219–35.

54. ** Pellegrini M, Marcotte EM, Thompson MJ, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS* 1999;**96**:4285–8.
A seminal article on context based predictions.

55. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. *Nature* 2000; **405**:823–6.

56. Enault F, Suhre K, Poirot O, *et al*. Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res* 2003;**31**: 3720–2.

57. Enault F, Suhre K, Claverie J-M. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. *BMC Bioinformatics* 2005;**6**.

58. Kolesov G, Mewes H-W, Frishman D. SNAPping up functionally related genes based on context information: a colinearity-free approach. *J Mol Biol* 2001;**311**:639–56.

59. Barker D, Pagel M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* 2005;**1**:e3.

60. Brenner SE, Chothia C, Hubbard TJ, Murzin AG. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol* 1996;**266**:635–43.

61. ** Rost B. Protein structures sustain evolutionary drift. *Fold Des* 1997;**2**:S19–24.
Arguing that protein structure space is much more redundant than previously thought.

62. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 2001;**307**:1113–43.

63. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res* 2001;**29**:228–29.

64. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 1993;**233**:123–38.

65. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 1997;**25**: 231–34.

66. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 2003; **19(Suppl 2)**:II246–55.

67. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins* 1995;**23**:356–69.

68. Zhu J, Weng Z. FAST: a novel protein structure alignment algorithm. *Proteins* 2005;**58**:618–27.

69. Kawabata T. MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res* 2003;**31**:3367–69.

70. Kawabata T, Nishikawa K. Protein structure comparison using the markov transition model of evolution. *Proteins* 2000;**41**:108–22.

71. Holm L, Park J. DaliLite workbench for protein structure comparison. *Bioinformatics* 2000;**16**:566–67.

72. Harrison A, Pearl F, Sillitoe I, *et al*. Recognizing the fold of a protein structure. *Bioinformatics* 2003;**19**:1748–59.

73. Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. *Proteins* 2004;**54**:260–70.

74. Friedberg I, Godzik A. Connecting the protein structure universe by using sparse recurring fragments. *Structure (Camb)* 2005;**13**:1213– 24.

75. Friedberg I, Godzik A. Fragnostic: walking through protein structure space. *Nucleic Acids Res* 2005;**33**:W249–51.

76. Stark A, Russell RB. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res* 2003;**31**:3341–4.

77. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci USA* 2004;**101**:14754–9.

78. ** Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 2004;**32**: D129–33.
A database documenting enzyme active sites and catalytic residues in enzymes of 3D structure.

79. Ausiello G, Zanzoni A, Peluso D, *et al*. pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res* 2005;**33**(W):133–7.

80. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* 2005;**33**: D183–7.

81. Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucleic Acids Res* 2004;**32**: W549–54.

82. Jambon M, Andrieu O, Combet C. The SuMo server: 3D search for protein functional sites. *Bioinformatics* 2005.

83. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* 2003;**52**:137–45.

84. Binkowski TA, Freeman P, Liang J. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res* 2004;**32**: W555–58.

85. Amitai G, Shemesh A, Sitbon E, *et al*. Network analysis of protein structures identifies functional residues. *J Mol Biol* 2004;**344**:1135–46.

86. Wei L, Altman RB. Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac Symp Biocomput* 1998;497–508.

87. Wei L, Altman RB. Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *J Bioinform Comput Biol* 2003;**1**:119–38.

88. Kleywegt GJ. Recognition of spatial motifs in protein structures. *J Mol Biol* 1999;**285**:1887–97.

89. Nimrod G, Glaser F, Steinberg D, *et al*. *In silico* identification of functional regions in proteins. *Bioinformatics* 2005;**21(Suppl 1)**:i328–37.

90. Ko J, Murga LF, Wei Y, Ondrechen MJ. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics* 2005;**21(Suppl 1)**:i258–65.

91. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA* 2001;**98**:12473–8.

92. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;**257**:342–58.

93. Glaser F, Pupko T, Paz I, *et al*. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;**19**:163–4.

94. Berezin C, Glaser F, Rosenberg J, *et al*. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 2004;**20**:1322–4.

95. Landau M, Mayrose I, Rosenberg Y, *et al*. ConSurf 2005. *Nucleic Acids Res* 2005;**33**:W299–302.

96. ** Walker MG, Volkmuth W, Sprinzak E, *et al*. Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* 1999;**9**: 1198–1203.
Introduction of the now-ubiquitous Guilt by Asssociation (GBA) algorithm for analysis of gene expression arrays.

97. * Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *PNAS* 1998;**95**:14863–8.

98. * Brown MP, Grundy WN, Lin D, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;**97**: 262–7.

99. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005;**21**:3587–95.

100. Andrade MA, Brown NP, Leroy C, *et al*. Automated genome sequence analysis and annotation. *Bioinformatics* 1999;**15**:391–412.

101. Hobohm U, Sander C. A sequence property approach to searching protein databases. *J Mol Biol* 1995;**251**:390–9.

102. ** Jensen LJ, Gupta R, Blom N, *et al*. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* 2002;**319**:1257–65.
A feature based predictor categorizing proteins using 12 diferent functional criteria, as defined in Riley, 2003.

103. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**: 105–32.

104. Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 2001;**17**:920–6.

105. Quinlan JR. C4.5: Programs for Machine Learning. *Morgan Kaufmann* 1993.

106. Wieser D, Kretschmann E, Apweiler R. Filtering erroneous protein annotation. *Bioinformatics* 2004;**20**:i342–7.

107. Lu P, Szafron D, Greiner R, *et al*. PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Res* 2005;**33**: D147–53.

108. Szafron D, Lu P, Greiner R, *et al*. Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res* 2004;**32**:W365–71.

109. Pal D, Eisenberg D. Inference of protein function from protein structure. *Structure (Camb)* 2005;**13**:121– 30.

110. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 2005;**33**:W89–93.

111. ** von Mering C, Jensen LJ, Snel B, *et al*. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005;**33**:D433–7.
A comprehensive database of known and predicted protein protein interactions.

112. Korbel JO, Doerks T, Jensen LJ, *et al*. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biology* 2005;**3**:e134.

113. Wang K, Samudrala R. FSSA: a novel method for identifying functional signatures from structural alignments. *Bioinformatics* 2005;**21**:2969–77.

114. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; **247**:536–40.

115. Reese MG, Hartzell G, Harris NL, *et al*. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* 2000;**10**:483–501.

116. Blaschke C, Leon EA, Krallinger M, Valencia A. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 2005;**6(Suppl 1)**:S16.

117. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005; **6(Suppl 1)**:S1.

118. Soro S, Tramontano A. The prediction of protein function at CASP6. *Proteins* 2005;**61(Suppl 7)**:201–13.

119. ** Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* 2003;**19**:1275–83.
A distance measure between nodes on a Gene Ontology graph, based on semantic similarity.

120. Joslyn CA, Mniszewski SM, Fulmer A, Heaton G. The gene ontology categorizer. *Bioinformatics* 2004;**20**:i169–77.

121. ⋆ Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* 2003;**326**:1–9.

122. Shakhnovich BE. Improving the precision of the structure-function relationship by considering phylogenetic context. *PLoS Comput Biol* 2005;**1**:e9.

123. Ouzounis CA, Karp PD. The past, present and future of genome-wide re-annotation. *Genome Biol* 2002;**3**: COMMENT2001.

124. ⋆⋆ Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. *Trends Genet* 2003;**19**:415–7.
An additional challenge to associating function with proteins: proteins that have several distinct functions.

125. Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol* 2004;**11**:463–75.

126. Deng M, Zhang K, Mehta S, *et al*. Prediction of protein function using protein–protein interaction data. *J Comput Biol* 2003;**10**:947–60.

127. Al-Shahib A, Breitling R, Gilbert D. Franksum: new feature selection method for protein function prediction. *Int J Neural Syst* 2005;**15**:259–75.

128. ⋆ des Jardins M, Karp PD, Krummenacker M, *et al*. Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**:92–9.

129. Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* 2000;**18**:609–13.

130. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* 2005;**15**:275–84.

131. Ofran Y, Punta M, Schneider R, Rost B. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today* 2005;**10**:1475–82.

132. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;**17**:282–3.