Predicting protein function from sequence and structure

David Lee, Oliver Redfern and Christine Orengo

Abstract | While the number of sequenced genomes continues to grow, experimentally verified functional annotation of whole genomes remains patchy. Structural genomics projects are yielding many protein structures that have unknown function. Nevertheless, subsequent experimental investigation is costly and time-consuming, which makes computational methods for predicting protein function very attractive. There is an increasing number of noteworthy methods for predicting protein function from sequence and structural data alone, many of which are readily available to cell biologists who are aware of the strengths and pitfalls of each available technique.

There are now >600 completely sequenced genomes of cellular organisms[1], contributing to more than five million unique protein sequences in the publicly accessible databases[2,3]. Experimental determination of the functions of all these proteins would be a hugely time-consuming and costly task and, in most instances, has not been carried out. Currently, approximately 20%, 7%, 10% and 1% of annotated proteins in the *Homo sapiens, Mus musculus, Drosophila melanogaster* and *Caenorhabditis elegans* genomes, respectively, have been experimentally characterized (traceable author source (TAS) annotations in Gene Ontology (GO))[4]. However, as the volume of data has increased, so too have the number and sophistication of computational methods for predicting function[5–7]. Knowledge of the three-dimensional (3D) structure of a protein can also provide a crucial insight into its mode of action, but currently the structures of <1% of sequences have been experimentally solved[8].

Protein function can be thought of on different interdependent levels and may be divided into three major categories: molecular function, biological process and cellular component (BOX 1). Molecular function describes activity on the molecular level, such as catalysis, whereas biological process describes broader functions that are carried out by assemblies of molecular functions, such as a particular metabolic pathway. Cellular component describes the compartment or compartments of a cell in which the protein performs its function. Computational methods exist to predict all of these aspects of function. Furthermore, most biological processes are carried out by groups of interacting proteins and these interactions can also be predicted *in silico*. Ideally, laboratory data should be integrated with theoretical approaches where possible, but the scope of this review is to focus on what can be achieved by exploiting sequence and structural data using computational means alone.

The most common and generally more accessible approach to function prediction is 'inheritance through homology' — that is, the knowledge that proteins with similar sequences frequently carry out similar functions. However, with the recent increase in the number of complete genome sequences, the possibility of establishing orthology has also increased. As discussed later in this review, this greatly improves the reliability of function transfer, although the coverage provided by identifiable orthologues tends to be small compared with that achieved by homologue detection. However, many of the incorrect annotations found in databases today are a consequence of the overly liberal application of inheritance through homology and this is compounded by the fact that the source of these annotations is often not given. Estimates of the error rate for the annotation of complete genomes vary from <5% to >40% depending on the types of function[9,10]. The development of computational protein-function prediction would be greatly assisted by establishing higher-quality benchmark datasets against which to test the methods[11]. In practice, the annotations assigned to enzymes are more amenable to computational analysis and a corresponding bias is seen in the literature.

Here, we introduce some basic concepts that are important to function prediction and provide a guide to methods that are commonly used for sequence-based function prediction and structure-based function prediction. Further information for many of the programs included in this article can be found in Supplementary information S1 (table). This list is not intended to be

Orthologue
A homologue that is found in separate species and has been separated by speciation rather than by a gene duplication event.

Homologue
Protein sequences are homologous if they have descended, usually with divergence, from a common ancestral sequence.

Biomolecular Structure and Modelling Group, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK.
e-mails:
dlee@biochem.ucl.ac.uk; ollie@biochem.ucl.ac.uk; orengo@biochem.ucl.ac.uk
doi:10.1038/nrm2281
Published online
24 October 2007

Box 1 | Computational approaches to protein-function prediction

Protein function can be divided into three broad areas: molecular function, biological process and cellular component. Molecular function describes activity at the molecular level, such as catalysis, which is commonly predicted through methods that identify homologues or orthologues. Biological process describes broader functions that are carried out by assemblies of molecular functions, such as a particular metabolic pathway. Genomic inference methods can identify the direct physical protein–protein interactions and indirect functional associations that are found in biological processes. Cellular component describes the compartment(s) of a cell in which the protein performs its function. This component can be predicted through methods that predict signal sequences, residue composition, membrane association or post-translational modifications (BOX 2). Within these areas there are broad categories of computational methods, all of which ultimately depend on experimental data. In this review we focus on predicting molecular function using homology, orthology and motifs.

comprehensive, but instead focuses on those methods that are widely used, of high quality, and easily accessible at the time of writing this review.

Homology, orthology and paralogy

When considering the field of protein-function prediction, it is vital to consider the concepts of homology, orthology and paralogy with respect to evolutionary relationships between proteins (reviewed by Fitch[12]). Protein sequences are homologous if they have descended, usually with divergence, from a common ancestral sequence. Homologues can be further divided into orthologues and paralogues. Orthologues are found in different species and have been separated by a speciation event, rather than by gene duplication. Paralogues are the product of gene duplication within a species, but because gene duplication can occur before speciation, paralogues can also exist in different species. The additional terms outparalogue and inparalogue refer to paralogues that arise before and after speciation, respectively.

These concepts are relevant to function prediction because orthologues are likely to occupy the same or a similar functional niche in different species. Conversely, paralogues — although they possibly still maintain considerable sequence similarity to their parents — are free to evolve new functions. Ideally, orthology and paralogy are determined by examining the closest common ancestor of any two species being compared. However, even if this is possible, determining orthology can still be complicated. For example, an orthologous gene could be lost from the genome while its paralogue is retained. Furthermore, members of multigene families within one genome can also exhibit functional overlap or even redundancy between their members. Various approaches used to recognize orthologues are described further below.

Paralogue
A homologue that is the product of a gene duplication event within a species.

Sources of annotation

When developing new methods for predicting protein function or assessing functional similarities, it is advantageous for the functional descriptors to be easily computer-readable. Until recently, most annotations were accumulated in an *ad hoc* fashion, usually as free text containing a whole spectrum of terminology and synonyms. Although natural language processing and automatic extraction of information from biological literature continues to improve[13], the most significant step forward has been in the structuring and standardization of annotations.

There are many excellent sources of functional annotations (for example, COG, Gene Ontology (GO), ENZYME, Swiss-Prot, FunCat, KEGG, MetaCyc and Reactome; see also Supplementary information S1 (table)). One of the more comprehensive sources is the GO project[4], which provides three structured vocabularies (ontologies) to describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Such vocabularies restrict descriptions to specific terms with uniform spelling. The controlled vocabularies are also structured so that they can be queried at different levels: for example, GO can be used to find all the gene products in the mouse genome that are involved in signal transduction, or to focus on only the receptor tyrosine kinases. Similarities between the sets of annotation terms assigned to separate proteins can be automatically quantified using a 'semantic similarity measure' (Lord *et al.*[14] and Schlicker *et al.*[15]), which accounts for the information content of different annotations. A useful resource for annotating enzymes is Enzyme Commission (EC) numbers. They comprise a hierarchical set of four numbers: the first number refers to the enzyme class ('1', for example, refers to oxidoreductases); the second number refers to the type of bond or group that is acted on (for example, '4' denotes an enzyme that acts on a peptide bond); and the next two levels give progressively more specific details of the catalysed reaction and its substrates. Therefore, two different enzymes that catalyse the same reaction would be annotated with the same EC number. A comparison of functional annotation schemes is made by Rison *et al.*[16].

Sequence-based function prediction

Unfortunately, there is no perfect protocol that can guarantee prediction of the correct function of a protein from its sequence, but in FIG. 1 we propose a suggested workflow that shows how some of the methods described in this review might be applied sequentially or in parallel to maximize the functional information that can be predicted from sequence.

As a first step, a simple, common approach to predict the function of a given protein sequence is to use the gateways provided by the National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI). These resources guide the user to a range of curated data, including protein and domain family information, functional sites and function prediction methods. Searching with a protein accession code, gene name or similar term will generate a list of links to these resources, each of which provides user instructions. Furthermore, tools are being actively developed that aim to integrate many of these diverse resources seamlessly (for example, InterPro[17] from EMBL–EBI).

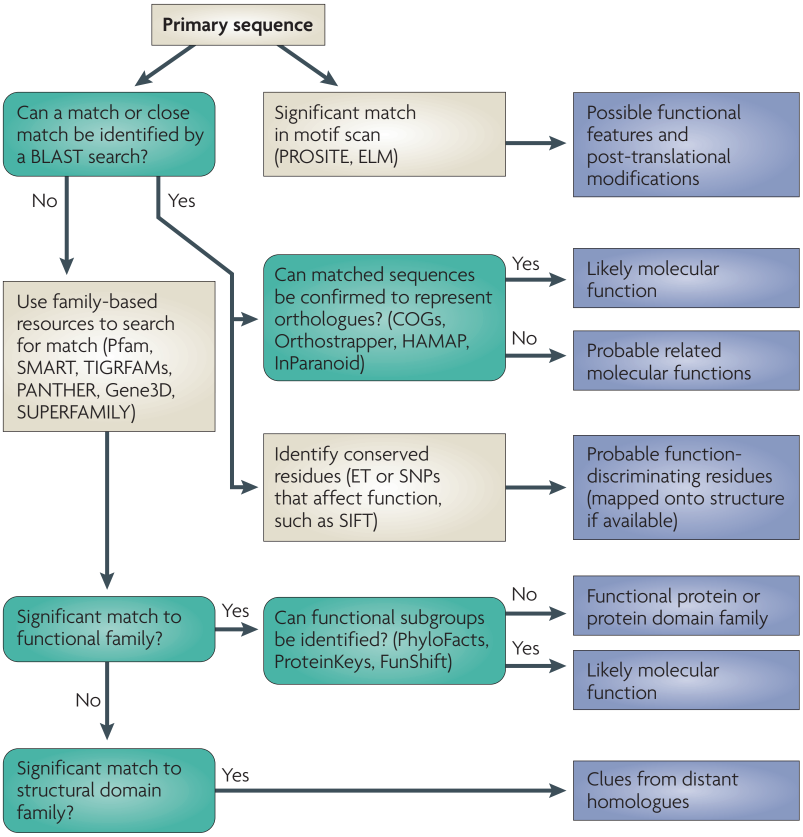
Primary sequence
Can a match or close match be identified by a BLAST search?
No
Yes
Significant match in motif scan (PROSITE, ELM)
Possible functional features and post-translational modifications
Use family-based resources to search for match (Pfam, SMART, TIGRFAMs, PANTHER, Gene3D, SUPERFAMILY)
Can matched sequences be confirmed to represent orthologues? (COGs, Orthostrapper, HAMAP, InParanoid)
Yes
Likely molecular function
No
Probable related molecular functions
Identify conserved residues (ET or SNPs that affect function, such as SIFT)
Probable function-discriminating residues (mapped onto structure if available)
Significant match to functional family?
Yes
Can functional subgroups be identified? (PhyloFacts, ProteinKeys, FunShift)
No
Functional protein or protein domain family
Yes
Likely molecular function
No
Significant match to structural domain family?
Yes
Clues from distant homologues

Figure 1| **Flow chart suggesting a possible strategy for molecular function prediction from a protein sequence and some possible outcomes.** There is no perfect workflow that can guarantee prediction of the correct function of a protein from its sequence, but we propose a simplified workflow to demonstrate the use of some commonly used methods that are described in this review. For clarity, the workflow is presented as sequential steps that are intended to give a logical overview of what to try next if one method has failed, and the depth of prediction that could be achieved. No single method is guaranteed always to give the best prediction so, at each step, various methods should be tried and the results compared. If a BLAST search is used as a first step and close relatives are identified, it may be possible to identify an annotated orthologue and, thus, give a confident prediction of the likely molecular function. Finding several close relatives may make it possible to identify function-discriminating residues (for example, using Evolutionary Trace (ET)) or single nucleotide polymorphisms (SNPs) that affect function (for example, using SIFT). If the BLAST matches are more distant relatives or no matches are found, it may be possible to identify a function using family-based resources. If a functional family can be identified, it may then be possible to obtain a more specific prediction of protein function by identifying the subgroup that the sequence belongs to. Finally, if no significant match is found to a functional family, then clues to the function may be gained by using structural domain families that identify distant homologues. Motif scans can provide predictions of a different nature from the other more global methods of sequence comparison presented here, and can provide an additional starting point to carry out a parallel analysis. The methods used in this workflow mainly address the prediction of molecular function. Other types of method are described in the main text sections 'Predicting interactions and associations' and 'Non-homology-based prediction'.

If an accession code or text search fails to locate the query protein at NCBI or EMBL–EBI, its sequence can be submitted to the Basic Local Alignment Search Tool (BLAST; either BLAST EBI or BLAST NCBI)[18]. A BLAST search compares the query sequence with protein sequences from various databases and if an exact match is not found, the search usually identifies a similar sequence from which it may be possible to inherit annotations. However, these resources provide few guidelines on when annotations can safely be inherited from other proteins. Automatic methods that perform a statistical analysis of the GO annotations associated with the matches can sometimes help to improve the accuracy of annotation transfer[19], as can data-mining approaches that also exploit the associations between GO terms[20].

There have been many studies aimed at establishing sequence similarity measures for safely transferring function between related proteins (discussed below). However, genes evolve at different rates owing to both uneven selection pressure on their functions and the inherent mutation rate of different species, which means that it is difficult to establish a similarity measure that is reliable in all cases. Rodents, for example, accumulate point mutations more rapidly than apes[21], and the evolutionary rates of proteins in different gene families[22] may vary by several orders of magnitude. Therefore, it is often best to exhaust the possibility of establishing orthology using resources such as HAMAP[23], which contains manually defined orthologous families for prokaryotes (FIG. 1). In practice, however, orthology itself is usually inferred by using sequence methods such as those used in COG[24] and its eukaryotic extension KOG, or those in InParanoid[25] (for eukaryotic orthologues) and Orthostrapper[26]. COG extends the classic bidirectional best-match approach — in which two genes form a pair by each being the best match to the other — by looking for two genes from different genomes that have the highest level of identity both to each other and to a single gene from a third genome. This is taken to be a strong indication that they are orthologues.

However, establishing orthology is not straightforward and provides limited coverage. Over the past ten years, many new family-based resources have emerged that group together protein sequences or individual protein domains into putative evolutionary families from across many different sequenced genomes. Family resources make it easier to gauge the reliability of functional inheritance through homology by organizing the matches to putative homologues in a more informative manner than that obtained by a BLAST search against an unstructured sequence database. Salient family characteristics can be readily identified and spurious matches can be more easily filtered out. In addition, multiple consistent matches can improve confidence in the results, and family resources allow more distant matches to be identified. These approaches and other rapidly developing areas of bioinformatics, such as using genomic inference to predict protein interactions and non-homology-based methods, are briefly reviewed below.

Family-based resources. Family-based resources group together either whole multidomain sequences or individual protein domains into putative evolutionary families. There are now many resources available (for example, the Munich Information centre for Protein Sequences (MIPS), Pfam, TIGRFAMs, ProtoNet, SYSTERS, ProDom, PANTHER, PRINTS, SMART, PhyloFacts, SCOP, SUPERFAMILY, CATH and Gene3D;

see also Supplementary information S1 (table) for more details). For some of these (such as MIPS[27] and Pfam[28]), the classification of homologues with related functions into families is accompanied by considerable manual validation and biological descriptions of the families are also provided. The InterPro[29] server has links to many of these resources.

Although general protein family resources such as Pfam can often give greater coverage than those that concentrate on orthologues, it is vital to remember that in many of the most highly populated families, function can diverge considerably between paralogues. What level of sequence similarity within a protein family provides a safe threshold for inheriting functional annotation? In an early study of CATH enzyme superfamilies[30,31], a large proportion of homologous relatives (>90%) were observed to have some similarity in the chemistry of their reactions, and in only 25% of superfamilies were some distant relatives found that showed completely different catalytic actions. Therefore, in many enzyme families, identifying relatives can often provide some useful clues on shared functional characteristics. Extension of these analyses to non-enzyme families shows similar trends, although caution is certainly necessary[25,26].

Various analyses have suggested that for functional transferability, 40% pairwise sequence identity can be used as a confident threshold to transfer the first three digits of an EC number, but to transfer all four digits of an EC number with at least 90% accuracy, >60% sequence identity is needed[32,33]. Lower thresholds can be used (30% sequence identity) for domain relatives that share similar multidomain contexts. Furthermore, because gene families evolve at different rates, family-specific thresholds are safer and lead to higher levels of functional annotation in many genomes (for example, a fivefold increase in GO annotations in *D. melanogaster*).

Another approach to building protein family resources is sequence clustering, which can automatically place all sequences into groups based on some measure of similarity; however, the clusters tend not to be manually validated. Generally, the stricter the criteria for clustering, the smaller the clusters and the more likely the proteins are to be functionally related. However, some sophisticated clustering methods are being developed that automatically identify the number of clusters that are required to segregate functions optimally[34].

Many family-based resources are now attempting to overcome the problem of functional diversity between relatives by identifying subgroups or subfamilies with more specific functions within the family; for example, dopamine and histamine receptors might be classified as two subgroups or subfamilies of the G-protein-coupled receptor family. In some resources (such as PANTHER[35]), this involves considerable expert curation[35], whereas others (for example, PhyloFacts[36]) exploit mainly computational strategies that are based on highly specific sequence profiles or hidden Markov models (HMMs)[36]. The success of a resource can often be increased by identifying specific residues that discriminate between functions, as described below.

Function-discriminating residues. In some protein families, ligand- and protein-binding sites can be predicted using phylogenetic analysis and assessment of 'tree-determinant residues'[37], because the functionally active residues are those that are most likely to have been conserved over evolution. Some of the most successful approaches[38] use a phylogenetic tree to rank the residues by evolutionary importance and then map this ranking onto a structure if one is available. The highest-ranked residues often cluster together in 3D space and can be used to identify functional sites (FIG. 2). Various servers exist to perform these analyses[39–41], and some approaches[42] are able to distinguish between residues that are conserved for functional reasons and those that have remained owing to structural constraints.

Other implementations[43] combine both sequence similarity and species information to distinguish between orthologues and paralogues. The accuracy appears to improve by purging those sequences with errors (for example, sequencing errors or incomplete fragments of sequences) that can degrade functional site identification[44]. Alternatively, external functional classifications can be overlaid on the phylogenetic tree[45] to overcome the constraints that are imposed by using a single tree to identify a functional property that is likely to arise from multiple factors. Moreover, analyses of function-discriminating residues (FDRs) can be exploited to predict whether a single nucleotide polymorphism (SNP) will affect protein function[46]. Several of these servers are listed in the Supplementary information S1 (table).

Other methods for identifying FDRs in protein families exploit entropy-based approaches[47–49] (for example, ProteinKeys; see also Supplementary information S1 (table)). In these methods, it is the diversity of amino acids at each position in a multiple alignment that is assessed, rather than the likelihood of specific mutations. Compositional differences between subfamilies can thus be scored without having to impose conservation. For example, the FunShift database[49] documents functional subgroups within Pfam protein domain families using an entropy-based approach. This identifies sites that are conserved between two subfamilies and those sites that have different evolutionary rates in the two subfamilies. EFICAz[50] recognizes FDRs in enzyme families by an iterative procedure that progressively clusters related sequences into functional subfamilies by combining information from pairwise sequence comparison, recognition of FDRs in Pfam enzyme families and recognition of multiple PROSITE patterns of high specificity to infer enzyme function. FIGURE 1 shows how the recognition of FDRs using methods such as ProteinKeys and FunShift can be applied to identify functional subgroups within a family.

Methods that seek FDRs do not necessarily require family resources or phylogenetic trees, but can be applied to any group of putative relatives in which a multiple sequence alignment or profile can be constructed. ClustalW[51] is one of the most commonly used methods for doing this, although new approaches show great promise[52–54]. Analysis of residue conservation can then

Phylogenetic tree
Shows the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. Each node that has descendants represents the most recent common ancestor of those descendants, with edge lengths sometimes corresponding to time estimates.

be applied to these alignments or the profiles derived from them. For example, filtering of PSI-BLAST profiles for patterns of catalytic site residues in the Catalytic Site Atlas[55] has been shown to give more specific enzymatic function annotations[56]. Similarly, PROSITE has a new section (ProRule) with manually created rules that increase discriminatory power by providing additional information about amino acids that are functionally and/or structurally crucial.

Sequence motifs. Globular protein domain databases have been reviewed above but there are often large segments of multidomain proteins that are disordered and do not intrinsically fold into a regular tertiary structure. Sometimes these are just linkers between globular domains but often they contain functional sites such as protein interaction sites, cellular localization signals, post-translational modification sites or cleavage sites. Patterns and regular expressions have been developed to describe and identify these short motifs, which often consist of only a few residues. Some motifs are also found within globular domains and can suggest a function in cases where profile methods have failed. A general problem with motifs, however, is that short sequence matches typically have low statistical significance and the false-positive rate can be high. Some specialized motif resources (for example, PROSITE, PRINTS (also a family-based resource), ELM) are summarized in Supplementary information S1 (table). FIGURE 1 shows that motif scans could be performed in parallel with the other prediction methods described above.

Predicting interactions and associations

Molecular biology has recently entered a new era in which revolutionary new experimental techniques also reveal the activity of proteins in space and time and as interacting components in complexes, pathways and networks. Various computational approaches exist for predicting protein interactions using protein sequences[57] or structures[58]. We now briefly review some promising new prediction methods that have been developed to address these aspects of protein function. BOX 2 outlines the important contribution that expression data is also now making to protein-function prediction.

Inheritance through homology. Protein interactions can be predicted through inheritance from proteins with known interactions, derived from various experimental approaches. MIPS[59] contains a manually curated yeast protein-interaction dataset and is often regarded as the gold standard of protein-interaction databases. Other accessible databases include DIP, IntAct, MINT, BIND, STRING, SCOPPI, SNAPPI-DB, iPfam, PSIMAP, PIBASE and 3did (see Supplementary information S1 (table) for descriptions). STRING[60] uses the COG database to automatically transfer associations to orthologous protein pairs in other organisms. Reconstruction of whole metabolic pathways and networks is also supported by several resources such as KEGG, MetaCyc, IMG and PUMA2. TRANSPATH[61] specializes in signalling pathways and Reactome[62] and the Human Protein Reference Database[63] specialize in the human genome. Protein interactions can also be predicted *de novo* by computational means and the main approaches are described below.

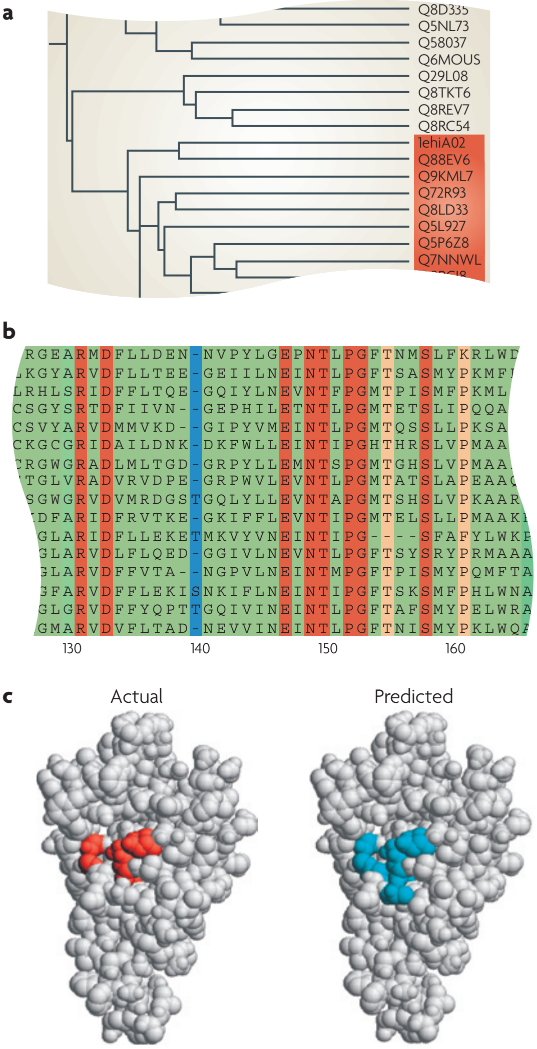
a
Q8D335
Q5NL73
Q58037
Q6MOUS
Q29L08
Q8TKT6
Q8REV7
Q8RC54
1ehiA02
Q88EV6
Q9KML7
Q72R93
Q8LD33
Q5L927
Q5P6Z8
Q7NNWL
b
RGEARMDFLLDEN-NVPYLGEPNTLPGFTNMSLFKRLWD
LKGYARVDFLLTEE-GEIILNEINTLPGFTSASMYPKMF
LRHLSRIDFFLTQE-GQIYLNEVNTFPGMTPISMFPKML
CSGYSRTDFIIVN--GEPHILETNTLPGMTETSLIPQQA
CSVYARVDMMVKD--GIPYVMEINTLPGMTQSSLLPKSA
CKGCGRIDAILDNK-DKFWLLEINTIPGHTHRSLVPMAA
CRGWGRADLMLTGD-GRPYLLEMNTSPGMTGHSLVPMAA
TGLVRADVRVDPE-GRPWVLEVNTLPGMTATSLAPEAA
SGWGRVDVMRDGSTGQLYLLEVNTAPGMTSHSLVPKAAR
DFARIDFRVTKE-GKIFFLEVNTIPGMTELSLLPMAAK
GLARIDFLLEKETMKVYVNEINTIPG----SFAFYLWKP
GLARVDLFLQED-GGIVLNEVNTLPGFTSYSRYPRMAAA
GLARVDFFVTA--NGPVLNEINTMPGFTPISMYPQMFTA
GFARVDFFLEKISNKIFLNEINTIPGFTSKSMFPHLWNA
GLGRVDFFYQPTTGQIVINEINTLPGFTAFSMYPELWRA
GMARVDVFLTAD-NEVVINEINTLPGFTNISMYPKLWQ
130
140
150
160
c
Actual
Predicted

Figure 2 | **The evolutionary trace (ET) method for identifying specificity residues. a** | A phylogenetic tree of related sequences is constructed for the protein of interest and related proteins. Functional subgroups tend to form distinct branches of the tree (red). **b** | A sub-alignment is constructed of sequences from those proteins in the same functional subgroup (those in red in part **a**). The level of conservation of columns in the alignment is calculated (in the alignment excerpt shown, colours represent a sliding scale in which the most conserved residues are red, those with intermediate conservation are green and the least conserved are blue) and conserved positions are mapped onto the structure or sequence of the unknown protein. **c** | A comparison of the actual substrate-binding residues (left panel) and predicted residues (right panel) when the ET method was applied to the ATP-grasp domain of a D-Ala–D-lactate ligase (Protein Data Bank entry 1EHI). It can be seen that the ET method achieves a close prediction of the substrate-binding site.

Box 2 | Combining functional genomics data with computational methods

Computational methods for predicting protein interactions and functional associations (see the main text section on genomic inference of protein function 'Predicting interactions and associations') can be validated using gene expression data, on the basis that proteins that interact should be expressed in the same cell types or under similar conditions. A considerable amount of the available expression data has been annotated and made publicly available in resources such as ArrayExpress[107]. A functional association between proteins predicted by phylogenetic profiling, for example, could be further investigated by searching ArrayExpress using protein accession codes or other search terms.

By contrast, it is possible to start with the experimental expression data and use computational analyses to validate groups of genes being expressed under similar conditions or with similar expression profiles. As part of this approach, it can be advantageous to integrate and compare 'local' private expression data with related public data. A collaboration between 'wet' and 'dry' scientists in the ENFIN[108] network is attempting to validate experimental data (such as gene expression and proteomics data) by combining many different computational analyses. Further experimental validation is then performed on genes that have been identified with high confidence by multiple prediction methods.

The most common source of expression data is from microarray experiments but these often sacrifice specificity for scale, yielding large quantities of relatively low-quality data[109]. Thus, sophisticated computational methods are necessary to achieve an accurate functional interpretation of these large-scale datasets[109–111]. A recent example can be seen in the computational prediction of cancer-gene function[112], which demonstrates the common approach of using statistical methods to generate a ranked list of overexpressed genes or lists of co-expressed genes in a profile, and then identifying enriched functional categories using Gene Ontology. These data can then be integrated with online protein–protein interaction data. The 'Chipping Forecast' published by *Nature Genetics*[113] also provides a periodic guide to the analysis of expression data.

Proteomics data can be similarly integrated with online computational methods. A recent example is the assignment of function to nucleolar proteins identified by mass spectrometry in an effort to characterize the protein complexes that constitute the human nucleolus[114].

Gene neighbour methods. Gene neighbour methods[64,65] use the organization of prokaryotic genomes in which interacting genes are often located next to each other in operons and are hence co-transcribed. This can be extended to eukaryotes, in which interacting co-regulated genes are sometimes found to cluster in the genome[66]. A detailed analysis of genomic context using expression data is presented by Korbel *et al.*[67].

The Rosetta Stone method. The Rosetta Stone method[68] is based on the observation that in some organisms, interacting proteins are encoded by separate genes, whereas in other organisms, their orthologues are fused into a single polypeptide chain. An example is the Trp synthetase-α and -β subunits, which are fused in fungi but separate in bacteria[69]. It is possible that the fusion of interacting proteins is sometimes subject to selection pressure because it effectively increases their relative concentration and removes the requirement for their co-regulation. More recent approaches have included statistical measures[70] to detect these 'gene fusion' events and have focused on all homologues of fused and non-fused proteins to improve the predictive coverage, rather than restricting the analysis to orthologues.

Phylogenetic profiling. Phylogenetic profiling[71] is based on the hypothesis that during evolution, functionally associated genes are likely to be inherited or eliminated in a co-dependent manner. Creating presence–absence profiles is now a common way of identifying these gene associations but crucially depends upon correctly determining orthology, which is a non-trivial task. As a result, the methods tend to be more successful for prokaryotes, in which it is more straightforward to identify orthologues. Recent approaches exploit domains rather than entire proteins[72] and one method[73] uses the number of occurrences of domain predictions for CATH structural families to overcome the problem of orthology assignment, thus finding functional relationships in eukaryotes that are undetectable by the conventional presence–absence profile comparisons.

Tree similarity. Methods to detect similarities in phylogenetic trees[74] use sequence comparison to reveal the co-evolution of interacting non-homologous protein families. Interacting proteins often co-evolve by accumulating correlated mutations and this can sometimes be seen in the correlation between the distance matrices that are used to construct trees for the families of the two proteins[74]. Recent improvements involve attempts to remove any background similarity between the trees of two protein families caused by speciation by subtracting a rescaled 16S rRNA phylogenetic tree, which is considered to be the canonical tree of life[75].

Some groups have attempted to combine these various approaches to predict protein interactions[76]. It is often found that the predictive power of a combined approach is greater than that of the components used individually.

Non-homology-based prediction

Recent analyses of genomes[77] have identified many singleton sequences for which homology cannot be used to infer function. In these cases it may be possible to apply non-homology-based methods that make use of subcellular localization and other protein features such as membrane association and post-translational modifications. Proteins that do not share significant global sequence similarity but perform similar or related functions might be expected to share some common features because they must share the same cellular machinery for modification and sorting, and operate in similar environments.

TIM barrel
Consists of eight α-helices and eight parallel β-strands that alternate along the peptide backbone. The structure is named after triose phosphate isomerase, a conserved glycolytic enzyme.

Superposition
After equivalent residues in two protein structures have been determined, the coordinates of one protein can be transformed onto the other.

Rossmann fold
Composed of three or more parallel β-strands linked by two α-helices and is found in proteins that bind nucleotides, such as the NAD and FMN co-factors.

There are various methods for the computational prediction of protein subcellular localization in prokaryotes[78] and eukaryotes[79]. The ProtFun[80] method predicts protein function from post-translational modifications and protein-sorting signals as well as from other, simpler aspects of protein sequence such as length, isoelectric point and amino-acid composition. Another approach[81] uses DNA microarray data from cell-cycle-regulated genes to show that protein features such as phosphorylation, glycosylation, subcellular location and instability or degradation can also be used to distinguish these genes. Lobley *et al.*[82] have recently demonstrated the use of patterns of native disorder in proteins to infer function.

Structure-based function prediction

In the remainder of this review, we explore the relationship between structure and function. Although protein structure is more conserved than sequence[83], knowledge of the specific fold adopted by a given protein does not directly imply a function. For example, there are 27 different homologous superfamilies that adopt the TIM barrel fold alone, covering over 60 different EC classifications[84]. Clearly, simply identifying this fold in a novel structure would do little to predict its function reliably.

The use of structural similarity in function prediction also poses additional problems that arise from artefacts of the crystallization procedure. A range of cognate and non-cognate ligands are used to stabilize the protein structure and facilitate the formation of crystals. In some cases, any conformational change that occurs during substrate binding can cause significant changes to the overall structure. Therefore, even structures of the same protein might exhibit significant structural differences when superposed. However, structural data can be used to detect proteins with similar function whose sequences have diverged beyond a level of similarity that can be reliably detected using sequence comparison methods. Generally, approaches to predict function from structure rely on trying to find globally similar structures and then, if no match is found, to focus on any structural similarities between known or predicted functional sites. An outline of an approach to predicting function from structural data is summarized in FIG. 3.

Predicting function by protein-fold comparison. Two proteins that exhibit high structural similarity along the entire length of their amino-acid sequence are likely to have the same or a similar function. Several popular methods for aligning and quantifying this relationship are available as web services (for example, DALI[85], CE[86], SSAP[87], STRUCTAL[88] and CATHEDRAL; see Supplementary information S1 (table)). When assessing the significance of the similarity between two structures, it is important to take into account both the quality of superposition and the number of residues in the alignment. Many common motifs, such as β-meanders, are observed within a range of diverse folds and, hence, detecting this motif is unlikely to suggest a key functional similarity. One advantage of structural methods is that they usually produce superior alignments to BLAST[18] and other sequence-profile methods when identity dips below 40%.

How similar must two proteins be at the structural level to have similar functions? An analysis of the CATH database[84] revealed that although most domains that share the same fold are associated with a single function, a small number of 'superfolds' (such as the ubiquitous Rossmann fold) can be associated with upwards of 50 different functions. Furthermore, these superfolds are the most common folds and account for >50% of domain sequences with predicted structures.

In highly variable superfamilies — those that exhibit significant structural divergence — different functions can evolve through the insertion of secondary structure elements[89]. Although these might originate from disparate regions of the primary sequence, they tend to co-locate in the structure to produce a larger structural motif or surface feature that modifies the geometry of the active site or promotes different protein–protein interactions. The ATP-grasp superfamily, shown in FIG. 4, is a good example of this mechanism of functional change, where ATP binding is conserved but insertion of secondary

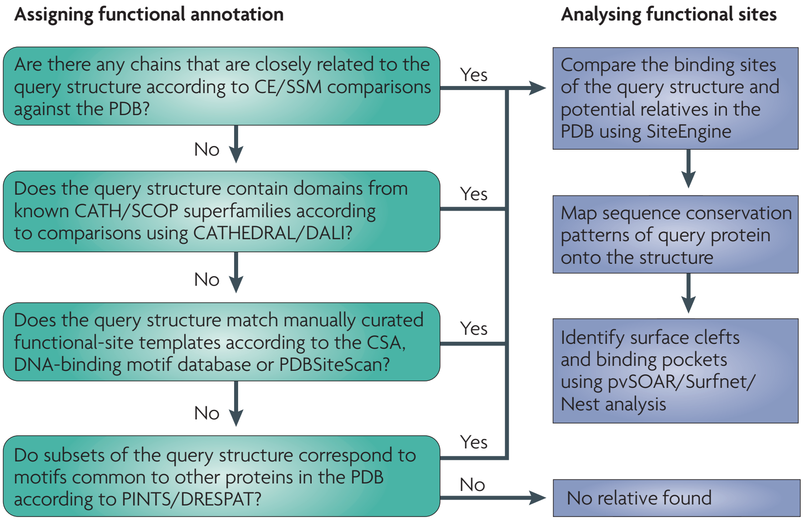
Assigning functional annotation
Analysing functional sites
Are there any chains that are closely related to the query structure according to CE/SSM comparisons against the PDB?
Yes
No
Does the query structure contain domains from known CATH/SCOP superfamilies according to comparisons using CATHEDRAL/DALI?
Yes
No
Does the query structure match manually curated functional-site templates according to the CSA, DNA-binding motif database or PDBSiteScan?
Yes
No
Do subsets of the query structure correspond to motifs common to other proteins in the PDB according to PINTS/DRESPAT?
Yes
No
Compare the binding sites of the query structure and potential relatives in the PDB using SiteEngine
Map sequence conservation patterns of query protein onto the structure
Identify surface clefts and binding pockets using pvSOAR/Surfnet/ Nest analysis
No relative found

Figure 3 | **Flow chart suggesting a possible strategy for function prediction from a protein structure and some possible outcomes.** We propose a simplified workflow to demonstrate the use of some commonly used methods to predict function from structure. As with sequence data, the first step is often to look for globally similar structures using fast comparison methods that are designed to operate at the whole-protein-chain level (for example, SSM and CE) from which to infer function. If no close match can be found, the query protein might contain previously observed structural domains of known function, which may be identified from protein family databases (for example, SCOP and CATH) using structure comparison algorithms such as CATHEDRAL and DALI. Putative relatives can often be verified by comparison to known functional motifs that have been manually curated from the literature and can be identified using PDBSiteScan or the Catalytic Site Atlas (CSA). If similar curated sites are not available, then automatic structural motifs, which are predicted to correlate with functional similarities, can be identified using PINTS or DRESPAT. Once the function of a protein has been inferred by one of these techniques, putative binding sites can be compared with a library of known sites, such as those implemented in pvSOAR or Surfnet and by comparing additional physicochemical properties such as charge and hydrophobicity using programs such as SiteEngine or Nest analysis. Patterns of sequence conservation, either from Evolutionary Trace analysis or calculated from a multiple sequence alignment, can also be mapped onto the query and matched structures to see if their predicted functional sites co-locate. PDB, Protein Data Bank.

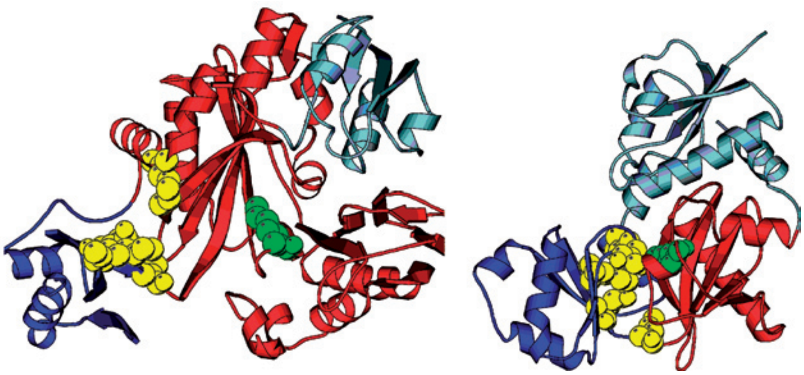

Figure 4 | **Change of protein function in the ATP-grasp superfamily by insertion of secondary structure elements.** The structure on the right shows D-Ala–D-Ala ligase and the structure on the left shows biotin carboxylase (box-like geometry), both of which contain an ATP-grasp domain (red) with two additional domains — a small domain (dark blue) and the B-domain (light blue). The yellow residues are involved in ATP binding and the green residues represent substrate-binding residues. Despite the noticeable difference in size between the ATP-grasp domains, the location of the active site appears to be conserved in this superfamily. However, the insertion of a significant number of secondary structures in the ATP-grasp domain of biotin carboxylase brings about a change to its substrate specificity. Hence, it appears that although these two proteins have conserved their function during evolution with respect to ATP binding, they bind different substrates and have different molecular functions. Therefore, caution is always required when inferring function from structurally similar family members.

structure elements permits the two proteins to bind different substrates. As a rule of thumb, most superfamilies with a high level of structural similarity also exhibit high functional similarity, but the data are more sparse[90].

Full alignment of two protein structures is not necessary to produce a similarity measure that proves useful for functional annotation[91]. A recent approach scored the similarity of two proteins by simply comparing their internal residue contacts — that is, residues that co-locate within 8–10 Å in the structure — and detected additional similarites over global alignment methods.

Predicting function using local 3D templates. To retain a specific function through evolution, the local environment of a functional site must be preserved, even if other portions of the fold have become altered. Indeed, enzymatic catalysis is performed by a limited set of residues that comprise the active site, and the specificity of DNA-binding proteins is often conferred by relatively small regions of positive charge on the surface of the protein structure. Using whole-fold comparison to assign function is limited by the fact that small changes in a binding or active site can cause a divergence of function. As a consequence, there are several methods that focus on comparing smaller structural motifs associated with a specific function.

The Catalytic Site Atlas[55] held at the EBI is a database of protein structures, the catalytic residues of which (up to six per protein) have been manually annotated from the literature. Annotations have been carefully expanded to include close relatives using conservative PSI-BLAST[18] profiles. Structural templates are constructed from the catalytic residues of the proteins in the database, and a fast search algorithm[92] is used to compare these to structures of unknown function so that the EC number can potentially be transferred. However, recognizing the correct relative can be challenging. First, catalytic residues can frequently move relative to one another on substrate binding, causing their geometry to vary between structures with and without bound ligands. Second, the probability of matching small structural templates at random is high, which creates difficulties in distinguishing between true and false matches. One attempt to address this problem compares the local environments around known or predicted catalytic residues and the corresponding residues in the matched protein[93]. It exploits the idea that the environment around the active site often exhibits higher sequence similarity than is evident across a global alignment of the query and matched structures.

Related methods use similar knowledge-based approaches[94,95] to compare functional information (SITE records) contained in the Protein Data Bank (PDB) structure files (FIG. 3). However, there are no specific rules imposed by the PDB as to what should be contained in the SITE records. Consequently, these records can contain various data (such as information about disulphide bridges, residues that are implicated in binding biologically irrelevant ligands, mutated residues, catalytic residues and so on) and may not always be important for comparisons of function.

Because it is time-consuming to manually design structural templates for a particular function, several groups have endeavoured to derive these automatically using novel algorithms. Some detect common structural motifs through pairwise comparison of side-chain patterns in diverse members of protein families[96]. These motifs can then be scanned to see whether they are present in novel structures. Other algorithms[97], while also seeking common side-chain patterns within superfamilies, make no assumptions about the location or nature of these motifs. Hydrophobic residues are often excluded when constructing templates because these are more likely to be buried in the protein core. For many methods that seek small structural motifs, distinguishing between genuine similarities and background is hampered by high rates of false positives.

A recent method[98] uses a novel approach. Rather than identifying 3D templates for structurally conserved regions in protein families, random sequence-conserved residues in known enzyme structures are selected to build motif templates. Interestingly, the best templates generally contain known functional residues, although there are also a few additional positions that have no known functional role but might afford a structural scaffold for catalytic or binding residues.

Comparing local structural features. Analysis of the surface of the protein, as well as pockets such as the active-site cleft, can often yield information on potential protein–protein interactions and small-molecule binding. One of the key reasons that enzymes catalyse reactions so effectively is that they are able to isolate their substrates in binding pockets or clefts, creating a unique chemical environment. Indeed, the active site is usually found in one of the two largest surface clefts[99]. In a similar fashion

Superfamily
A group of evolutionarily related proteins that often have the same overall domain structure, but may have diverged beyond recognition at the sequence level.

Structural template
Many methods of predicting function from structure involve listing specific residues and expected inter-atom distances in a template file, which can then be compared against other structures.

SITE record
Part of a Protein Data Bank file containing details of which residues are relevant to the protein function (for example, those involved in substrate binding).

De novo sequence method
A method that does not rely upon homology between sequences for transferring functional annotations but rather on the recognition of features such as residue composition and subcellular localization signals.

Meta-server
In the context of this review, a meta-server is a gateway to a well-benchmarked set of prediction methods.

to the template searching discussed above, binding sites in unannotated proteins can be compared against a library of known sites, such as those implemented in pvSOAR[100]. Related approaches attempt to improve performance by including comparisons of the physico-chemical properties of the amino acids in the binding site[101]. The conservation of charge and hydrophilicity is often useful for picking out genuine functional homologues (using programs such as SiteEngine; FIG. 5). Binding sites with similar physico-chemical properties in comparable 3D conformations can be used to identify similar enzymatic functions. In a similar vein, the electrostatic surface of functional site (eF-Site) database[102] provides information about electrostatic-potential surfaces that can be used to identify similar patterns of charge in binding and interaction sites.

Molecular interactions in the cell — either between protein surfaces or proteins and their ligands — rely on electrostatic contacts between charged or polar residues. It is possible to apply a molecular cartography approach to reduce protein surfaces to a spherical map[103]. By comparing mainly charged and hydrophobic residues, the similarity of two protein maps can be used to identify functional subgroups within protein families, for example, to distinguish between monomeric and tetrameric haemoglobin subunits. Because some active-site residues have perturbed titration curves, a different approach is to identify active-site residues by using theoretical microscopic titration curves to model the electrostatics of the protein and predict pK_a values of ionizable groups[104].

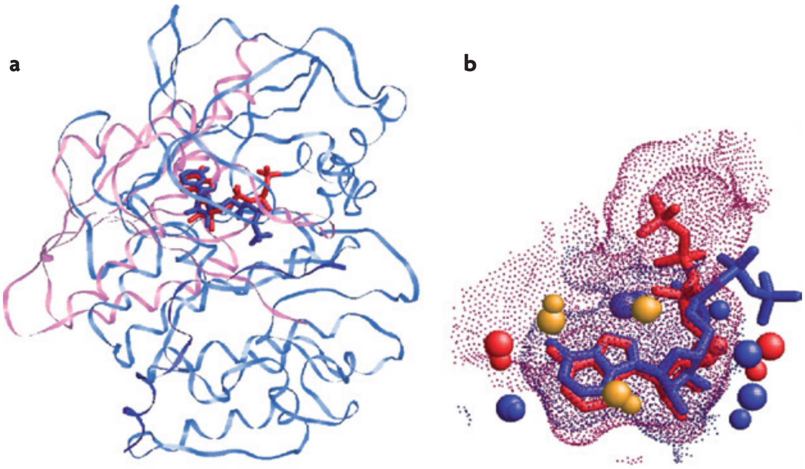

Figure 5 | **Using surface features and physico-chemical properties to recognize similarities between binding sites.** Analysis of the protein surface as well as pockets such as the active-site cleft can often yield information about binding sites. Using SiteEngine, binding sites can be compared between proteins of unknown function and a library of binding sites from proteins of known function. **a** | The proteins of a cAMP-dependent protein kinase (blue) (Protein Data Bank (PDB) entry 1ATP) and a structure of unknown function MJ0577 (pink) (PDB entry 1MJH) are superimposed based on the similarity of their binding sites. The ATP molecules from 1ATP are coloured blue and those from 1MJH are coloured red. **b** | A closer view of the active sites of cAMP-dependent protein kinase and protein MJ0577 and the surface properties used to detect the ATP-binding site. The surfaces of the active sites are represented as small dots and are coloured red for 1MJH and blue for 1ATP. SiteEngine detects that the known ATP-binding site of cAMP-dependent protein kinase has similar physico-chemical properties to that of MJ0577 and, hence, ATP-binding properties can be predicted for this structure of unknown function. Figure reproduced with permission from REF. 115 © (2004) Elsevier.

Servers for function prediction. For the convenience of the general user, the ProFunc[105] and ProKnow[106] servers can be used that extract both structural and sequence data from a query structure to carry out functional annotation using several of the methods described above. ProFunc combines BLAST and HMM searches with 3D template-based and surface-cleft analysis. ProKnow extends this approach by providing a probability model for GO annotations of the protein in question.

Conclusions and perspectives

There is considerable activity today in the field of computational protein-function prediction, and although 'inheritance through homology' remains the most common and easily accessible approach, *de novo* sequence methods have been developed recently that do not rely on the inheritance of annotation through homology. The body of protein-function annotations that most prediction methods depend upon is becoming increasingly computer-readable and is being organized in ways that enhance the scope of predictions.

The scarcity of experimentally solved protein structures means that most function prediction is carried out by comparing protein sequences, and the recent substantial growth in complete genome sequences is making these methods more powerful. Family-based resources that exploit profiles and sequence clustering — and in many cases involve significant manual curation — can be extremely valuable in providing information on the variation in functional properties across a family. This makes it far easier to assess the accuracy of transferring functions between particular relatives. Many methods are also available to identify the function-discriminating residues in proteins and these are being used to divide families into more specific subfamilies.

With the advent of the structural genomics initiatives, an increasing number of protein structures are being experimentally determined while their function is still unknown. In these cases, function can sometimes be predicted by using the structure rather than the sequence of the protein. By analogy with sequence comparison, global comparisons can be made using fold-comparison methods, usually by identifying the individual structural domains in a protein, and local comparisons can be made using structural templates from the active site of enzymes. Other features that can be used for function prediction when a structure is available include conserved surface patches, clefts and electrostatic potential.

Where is the field of computational protein-function prediction heading and what are the important matters that need to be addressed? Initiatives to integrate more experimental data and to validate the performance of prediction methods will increase their value to biologists. In general, it is currently best to seek and compare the results of several prediction methods, and meta-servers simplify this by providing easy access to a range of the best-performing methods. Meta-servers could be improved by developing better and more stable software environments to provide easier and more powerful interfaces and to support workflows. Confidence in the results returned by these servers could be improved by integrating third-party validation processes in addition to those provided by the developers of the server. In some cases, more balance between the sequence-based and structure-based methods would be desirable because the sequence-based methods are still among the most mature and reliable methods for inferring protein function.

Structural genomics projects are enriching the data that computational methods rely on by increasing the diversity of protein sequences for which the structure has been determined. Structural genomics groups are also planning to select targets to maximize the number of protein functional families for which at least one structure has been determined. Greater collaboration between different fields within the biological sciences will enhance this selection process and, hopefully, will ultimately lead to better tools for function prediction.

1. Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N. C. The Genomes On Line Database (GOLD) v2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332–D334 (2006).
2. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
3. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **34**, D16–D20 (2006).
4. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
One of the best and most comprehensive attempts to standardize and organize the annotation of protein function.
5. Whisstock, J. C. & Lesk, A. M. Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340 (2003).
A thorough and fairly recent review of the whole field of protein-function prediction from sequence and structure.
6. Bork, P. *et al.* Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725 (1998).
7. Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **15**, 275–284 (2005).
8. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
9. Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
10. Devos, D. & Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431 (2001).
11. Godzik, A., Jambon, M. & Friedberg, I. Computational protein function prediction: are we making progress? *Cell Mol. Life Sci.* **64**, 2505–2511 (2007).
12. Fitch, W. M. Homology: a personal view on some of the problems. *Trends Genet.* **16**, 227–231 (2000).
An interesting discussion of some important concepts in the field of protein-function prediction.
13. Krallinger, M. & Valencia, A. Text-mining and information-retrieval services for molecular biology. *Genome Biol.* **6**, 224 (2005).
14. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275–1283 (2003).
15. Schlicker, A., Domingues, F. S., Rahnenfuhrer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**, 302 (2006).
16. Rison, S. C., Hodgman, T. C. & Thornton, J. M. Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics* **1**, 56–69 (2000).
17. Mulder, N. J. *et al.* New developments in the InterPro database. *Nucleic Acids Res.* **35**, D224–D228 (2007).
18. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
19. Martin, D. M., Berriman, M. & Barton, G. J. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* **5**, 178 (2004).
20. Hawkins, T., Luban, S. & Kihara, D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci.* **15**, 1550–1556 (2006).
This method performed well in the CASP7 function-prediction category.
21. Blair, H. S. & Kumar, S. Genomic clocks and evolutionary timescales. *Trends Genet.* **19**, 200–206 (2003).
22. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. USA* **102**, 5483–5488 (2005).
23. Gattiker, A. *et al.* Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.* **27**, 49–58 (2003).
24. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
25. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
26. Storm, C. E. & Sonnhammer, E. L. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**, 92–99 (2002).
27. Mewes, H. W. *et al.* MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **34**, D169–D172 (2006).
28. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
29. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
30. Pearl, F. *et al.* The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* **33**, D247–D251 (2005).
31. Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
This paper examines the sequence–structure–function paradigm through an analysis of enzymes within superfamilies in the CATH database. It gives several examples of the different ways in which sequence and structure can change over evolution to produce new functions.
32. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* **333**, 863–882 (2003).
33. Rost, B. Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608 (2002).
34. Marttinen, P., Corander, J., Toronen, P. & Holm, L. Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics* **22**, 2466–2474 (2006).
35. Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
36. Krishnamurthy, N., Brown, D. P., Kirshner, D. & Sjolander, K. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol.* **7**, R83 (2006).
37. del Sol, M. A., Pazos, F. & Valencia, A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302 (2003).
38. Yao, H. *et al.* An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261 (2003).
39. Joachimiak, M. P. & Cohen, F. E. JEvTrace: refinement and variations of the evolutionary trace in JAVA. *Genome Biol.* **3**, RESEARCH0077 (2002).
40. Morgan, D. H., Kristensen, D. M., Mittelman, D. & Lichtarge, O. ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics* **22**, 2049–2050 (2006).
41. La, D. & Livesay, D. R. MINER: software for phylogenetic motif identification. *Nucleic Acids Res.* **33**, W267–W270 (2005).
42. Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504 (2004).
43. Engelhardt, B. E., Jordan, M. I., Muratore, K. E. & Brenner, S. E. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.* **1**, e45 (2005).
44. Yao, H., Mihalek, I. & Lichtarge, O. Rank information: a structure-independent measure of evolutionary trace quality that improves identification of protein functional sites. *Proteins* **65**, 111–123 (2006).
45. Pazos, F., Rausell, A. & Valencia, A. Phylogeny-independent detection of functional residues. *Bioinformatics* **22**, 1440–1448 (2006).
46. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
47. Valdar, W. S. Scoring residue conservation. *Proteins* **48**, 227–241 (2002).
48. Pirovano, W., Feenstra, K. A. & Heringa, J. Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.* **34**, 6540–6548 (2006).
49. Abhiman, S. & Sonnhammer, E. L. FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.* **33**, D197–D200 (2005).
50. Tian, W., Arakaki, A. K. & Skolnick, J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.* **32**, 6226–6239 (2004).
51. Thompson, J. D., Higgins, D. G. & Gibson, T. J. 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**, 4673–4680 (1994).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
53. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
54. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
55. Porter, C. T., Bartlett, G. J. & Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32**, D129–D133 (2004).
56. George, R. A. *et al.* Effective function annotation through catalytic residue conservation. *Proc. Natl. Acad. Sci. USA* **102**, 12299–12304 (2005).
57. Shoemaker, B. A. & Panchenko, A. R. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol.* **3**, e43 (2007).
An accessible introduction to computational methods for predicting protein-interaction partners.
58. Aloy, P. & Russell, R. B. Structural systems biology: modelling protein interactions. *Nature Rev. Mol. Cell Biol.* **7**, 188–197 (2006).
59. Guldener, U. *et al.* MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34**, D436–D441 (2006).

60. von Mering, C. *et al.* STRING 7 — recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358–D362 (2007).
A good example of a state-of-the-art protein-interaction database.

61. Krull, M. *et al.* TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **34**, D546–D551 (2006).

62. Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39 (2007).

63. Mishra, G. R. *et al.* Human protein reference database — 2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).

64. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).

65. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901 (1999).

66. Teichmann, S. A. & Babu, M. M. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* **20**, 407–410 (2002).

67. Korbel, J. O., Jensen, L. J., von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnol.* **22**, 911–917 (2004).

68. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).

69. Burns, D. M., Horn, V., Paluh, J. & Yanofsky, C. Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase α and β chains. *J. Biol. Chem.* **265**, 2060–2069 (1990).

70. Marcotte, C. J. & Marcotte, E. M. Predicting functional linkages from gene fusions with confidence. *Appl. Bioinformatics.* **1**, 93–100 (2002).

71. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).

72. Pagel, P., Wong, P. & Frishman, D. A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.* **344**, 1331–1346 (2004).

73. Ranea, J. A. G., Yeats, C., Grant, A. & Orengo, C. A. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D "Phylo-Tuner" method applied to eukaryotic genomes. *PLoS Comput. Biol.* (in the press).

74. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614 (2001).

75. Pazos, F., Ranea, J. A., Juan, D. & Sternberg, M. J. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–1015 (2005).

76. Qi, Y., Bar-Joseph, Z. & Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* **63**, 490–500 (2006).

77. Lee, D., Grant, A., Marsden, R. L. & Orengo, C. Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins* **59**, 603–615 (2005).

78. Gardy, J. L. & Brinkman, F. S. Methods for predicting bacterial protein subcellular localization. *Nature Rev. Microbiol.* **4**, 741–751 (2006).

79. Donnes, P. & Hoglund, A. Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics* **2**, 209–215 (2004).

80. Jensen, L. J. *et al.* Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265 (2002).

81. de Lichtenberg, U., Jensen, T. S., Jensen, L. J. & Brunak, S. Protein feature based identification of cell cycle regulated proteins in yeast. *J. Mol. Biol.* **329**, 663–674 (2003).

82. Lobley, A., Swindells, M. B., Orengo, C. A. & Jones, D. T. Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol.* **3**, e162 (2007).

83. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).

84. Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res.* **35**, D291–D297 (2007).

85. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138 (1993).

86. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747 (1998).

87. Taylor, W. R. & Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **208**, 1–22 (1989).

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

89. Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A. & Orengo, C. A. Structural diversity of domain superfamilies in the CATH database. *J. Mol. Biol.* **360**, 725–741 (2006).

90. Orengo, C. A., Sillitoe, I., Reeves, G. & Pearl, F. M. Review: what can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134**, 145–165 (2001).

91. Lisewski, A. M. & Lichtarge, O. Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Res.* **34**, e152 (2006).

92. Barker, J. A. & Thornton, J. M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**, 1644–1649 (2003).

93. Laskowski, R. A., Watson, J. D. & Thornton, J. M. Protein function prediction using local 3D templates. *J. Mol. Biol.* **351**, 614–626 (2005).

94. Ivanisenko, V. A. *et al.* PDBSiteScan: a tool for search for the best-matching superposition in the database PDBSite. *Third International Conference on Bioinformatics of Genome Regulation and Structure* **3**, 149–152 (2002).
Description of the PDBSiteScan server, which allows the user to compare a query protein structure against known functional sites in solved structures in the PDB.

95. Golovin, A., Dimitropoulos, D., Oldfield, T., Rachedi, A. & Henrick, K. MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins* **58**, 190–199 (2005).

96. Stark, A. & Russell, R. B. Annotation in three dimensions. PINTS: Patterns In Non-homologous Tertiary Structures. *Nucleic Acids Res.* **31**, 3341–3344 (2003).

97. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N. & Sarawagi, S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.* **326**, 955–978 (2003).

98. Polacco, B. J. & Babbitt, P. C. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics* **22**, 723–730 (2006).

99. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452 (1996).

100. Binkowski, T. A., Joachimiak, A. & Liang, J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.* **14**, 2972–2981 (2005).

101. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. *Nucleic Acids Res.* **33**, W337–W341 (2005).

102. Kinoshita, K. & Nakamura, H. eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics* **20**, 1329–1330 (2004).

103. Pawlowski, K. & Godzik, A. Surface map comparison: studying function diversity of homologous proteins. *J. Mol. Biol.* **309**, 793–806 (2001).

104. Ko, J., Murga, L. F., Wei, Y. & Ondrechen, M. J. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics* **21** (Suppl. 1), i258–i265 (2005).

105. Laskowski, R. A., Watson, J. D. & Thornton, J. M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33**, W89–W93 (2005).
Description of the ProFunc server, which combines sequence and structure comparison methods to predict protein function from a given structure.

106. Pal, D. & Eisenberg, D. Inference of protein function from protein structure. *Structure* **13**, 121–130 (2005).
Description of the ProKnow server, which, like ProFunc, aims to combine a range of homology-detection methods for a given structure to predict function. Gene Ontology terms from matched proteins are combined using a statistical framework to provide the user with a combined significance score for each predicted function.

107. Parkinson, H. *et al.* ArrayExpress — a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* **35**, D747–D750 (2007).

108. Kahlem, P. & Birney, E. Dry work in a wet world: computation in systems biology. *Mol. Syst. Biol.* **2**, 40 (2006).

109. Breitling, R., Amtmann, A. & Herzyk, P. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* **5**, 34 (2004).

110. Breslin, T., Eden, P. & Krogh, M. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics* **5**, 193 (2004).

111. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).

112. Hu, P., Bader, G., Wigle, D. A. & Emili, A. Computational prediction of cancer-gene function. *Nature Rev. Cancer* **7**, 23–34 (2007).

113. Editorial. A decade of genome-wide biology. *Nature Genetics* **37**, S3 (2005).

114. Hinsby, A. M. *et al.* A wiring of the human nucleolus. *Mol. Cell* **22**, 285–295 (2006).

115. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **339**, 607–633 (2004).

Acknowledgements

We would particularly like to acknowledge E. Sideris for help with the figures in this manuscript.

DATABASES

Protein Data Bank: http://www.pdb.org/pdb/home/home.do
1ATP | 1EHI | 1MJH

FURTHER INFORMATION

Christine Orengo's homepage:
http://www.biochem.ucl.ac.uk/research/orengo/orengo.htm
Programs: 3did | BIND | BLAST EBI | BLAST NCBI | Catalytic Site Atlas | CATH | CATHEDRAL | CE | ClustalW | COG | DALI | DIP | DRESPAT | EFICAz | eF-Site | ELM | ENZYME | Evolutionary Trace | FunCat | FunShift | Gene3D | GO | HAMAP | Human Protein Reference Database | IMG | InParanoid | IntAct | InterPro | iPfam | KEGG | MetaCyc | MIPS | MINT | Nest | Orthostrapper | PANTHER | PDBSiteScan | Pfam | PhyloFacts | PIBASE | PINTS | PRINTS | ProDom | ProFunc | ProKnow | PROSITE | ProteinKeys | ProtFun | ProtoNet | PSIMAP | PUMA2 | pvSOAR | Reactome | SCOP | SCOPPI | SIFT | SiteEngine | SMART | SNAPPI-DB | SSAP | SSM | STRING |STRUCTAL | SUPERFAMILY | Surfnet | Swiss-Prot | SYSTERS | TIGRFAMs| TRANSPATH

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF