Predicting protein function from sequence and structural data
James D Watson, Roman A Laskowski and Janet M Thornton

When a protein's function cannot be experimentally determined, it can often be inferred from sequence similarity. Should this process fail, analysis of the protein structure can provide functional clues or confirm tentative functional assignments inferred from the sequence. Many structure-based approaches exist (e.g. fold similarity, three-dimensional templates), but as no single method can be expected to be successful in all cases, a more prudent approach involves combining multiple methods. Several automated servers that integrate evidence from multiple sources have been released this year and particular improvements have been seen with methods utilizing the Gene Ontology functional annotation schema.

Introduction
A measure of the success of the various structural genomics initiatives worldwide is that they now account for around a fifth of the weekly Protein Data Bank (PDB) releases and consist largely of proteins with low homology to those already in the PDB. This has been brought about by improvements in target selection, the automation of protein production and structure determination, and the use of high-throughput methodology to tackle many proteins in parallel. The price of this success, however, has been that a significant proportion — over one-third — of structural genomics structures are of proteins of unknown function, annotated merely as 'hypothetical proteins'. Although these provide a valuable contribution to our knowledge of protein structure, their worth can be significantly enhanced by knowing the biological roles that they play.

In the absence of experimental data, the function of a protein is usually inferred on the basis of its sequence similarity to a protein of known function. The more similar the sequence, the more similar the function is likely to be, although there are cases of proteins having 100% sequence identity yet performing different roles according to where they are expressed [1]. The problem with structural genomics targets is that many of them do not have any close sequence similarity to proteins of known function. In recent years, therefore, much effort has gone into the development of methods for deriving functional clues directly from the three-dimensional structure once it has been solved. These methods are the main subject of this review, although methods for predicting function from other sources will also be mentioned.

Definition of function
Many methods for predicting function from structure have been published, as have several reviews [2–4]. One problem with reviewing these methods concerns assessing their usefulness and rate of success. A large part of this stems from the difficulty of defining what is meant by ‘function’. Function can be described at many levels, ranging from biochemical function to biological processes and pathways, all the way up to the organism level [5]. Consequently, proteins are annotated with different degrees of functional specificity: for example, ‘ubiquitin-like domain’, ‘signalling protein’, ‘predicted serine hydrolase’, ‘probable eukaryotic D-amino acid tRNA deacetylase’ and so on. Judging the ‘accuracy’ of any such assignment is difficult.

Several schemes for classifying protein function have been developed over the years, most famously the Enzyme Commission (EC) numbering scheme for enzymes and the EcoCyc scheme for Escherichia coli proteins. A relatively recent development has been the Gene Ontology (GO) schema [6,7], which is open source, offers a controlled vocabulary and, most importantly, is a machine-readable ontology. This has made it one of the most commonly used ontologies in bioinformatics and provides a ready means of assessing the worth of any function prediction scheme.

Sequence-based methods
Prediction of a protein’s function from its structure is usually undertaken when sequence-based methods have failed (Table 1 provides a description of some of the sequence-based methods currently available). Typically, however, the sequence-based methods used are simple BLAST [8] or FASTA runs, which merely perform direct sequence-sequence comparisons. These have been largely superseded by more powerful and sensitive profile/
pattern-based methods, and these can provide important clues to function, even before the structure is obtained.

The most common type of profile is the hidden Markov model (HMM) and several methods exist for creating them (e.g. SAM-T and HMMER). The profiles are constructed from the sequences of whole protein families, where the family can be defined in terms of three-dimensional structure, as in SUPERFAMILY, or in terms of function, as in Pfam [9]. The PSI-BLAST program constructs profiles based on sequences matched by an initial BLAST search and uses the profiles to iteratively pull in more and more remote homologues, eventually bringing in more distant matches than is possible from a standard BLAST search.

The primary advantage of the profile methods is that they provide greater sensitivity compared to simple sequence-sequence comparison because the profiles implicitly contain information on both which residues within the family are well conserved and which are the most variable. Further sensitivity can be obtained using profile-profile comparisons (e.g. COUCH, COMPASS and PROF_SIM). A new method called HHsearch [10] goes one stage further and performs HMM-HMM comparisons. This is claimed to be faster and more sensitive than other methods, being able to detect 2–4 times more homologues than PSI-BLAST and 1.4–1.9 times more homologues than profile-profile methods when the HMMs include predicted secondary structure. Furthermore, the resultant sequence alignments are of higher quality across family, superfamily and fold levels.

### Structure-based methods

When sequence-based methods fail, functional clues need to be garnered from the protein’s three-dimensional structure. Methods for predicting function from structure can be classified according to the level of protein structure and specificity at which they operate, ranging from analysis of the protein’s overall fold to the identification of highly specific three-dimensional clusters of functional residues.

#### Fold matching

Proteins sharing similar functions often have similar folds, a result of descent from a common ancestral protein. Sometimes, however, the function of one or both proteins may alter during evolution while their folds remain largely unchanged, so in these cases the same fold may give rise to two functions. In the case of the so-called ‘super-folds’, the functional divergence can occur across a whole fold family, the best known example being the TIM-barrel fold, which supports over 60 different functions. So, finding a fold match between the target protein and one in the PDB does not always provide a reliable prediction of the protein’s function, but can usually suggest possible function types (e.g. enzyme, DNA-binding protein, etc.).
Hence, finding a fold match is invariably the first stop for structure-based function prediction. Several methods exist for fold searching, the best known being DALI [11]; other well-known methods include SSM [12] and GRATH [13], which use graph theory, VAST [14], which uses vector alignment of secondary structures, and CE [15], which employs combinatorial extension. Table 2 provides further details. For a review evaluating the fold comparison servers, see Novotny et al. [16]. Although the methods generally agree, they can sometimes give different results for the same comparison. Therefore, the best strategy is to use two or three different methods and compare the results.

New approaches have been developed that can improve the sensitivity or cut down the search time. One such algorithm, FAST [17], uses a directionality-based scoring scheme to align structures at the residue-residue level rather than by secondary structure. It is claimed to have higher sensitivity than DaliLite and CE, and to be faster than other residue-residue comparison methods. Another new approach uses deterministic annealing to achieve high accuracy of alignment [18], whereas a third uses sequence threading to identify sequences in pathogenic organisms that are expected to match the three-dimensional structure of proteins in the host organism [19]. The last approach is aimed at identifying potential virulence factors, but also has the potential to be used in generic protein comparisons.

Surface clefts and binding pockets

Below the level of the fold come various other aspects of a protein’s three-dimensional structure that may be associated with specific functions. The surface of the protein, particularly its clefts and pockets, can hold important clues to function. Cofactors, substrates and regulatory elements tend to bind in clefts on the surface or in a region between separate interacting protein chains. For the majority of enzymes, the active/catalytic site is found in one of the two largest clefts on the surface. Therefore, analysis of the clefts, including comparison with the clefts of proteins of known function, can potentially provide fairly strong indicators of what the protein might do.

One problem arises when a protein’s binding site undergoes a significant conformational change upon substrate binding; the ligand-free form of the protein will have a very different site that may be difficult to even recognize as the binding site, let alone allow prediction of what might bind there. Another issue is that of allosteric control, whereby remote pockets on the surface act as binding sites for molecules that regulate the activity of the protein through conformational change. In such cases, the major cleft may not be the most important site. For a review of methods to identify potential ligand-binding sites, see Campbell et al. [20].

Once located, the binding sites can be compared against databases of clefts and pockets. The pvSOAR [21] web server allows searches to be performed against the CASTp [22] database or against non-redundant PDB data sets. The server has a wide range of searches available. Full structures and individual pockets can be uploaded onto the server or existing PDB releases can be selected for analysis. Additional useful features allow particular pockets to be examined or, if a particular residue is of interest, surfaces containing this residue can be selected. Searching is rapid and statistically significant similar pockets are returned to the user for further inspection.

Another database of surface regions is SURFACE [23], although this can currently only be used for manual inspection of its contents rather than for searches against uploaded PDB files. The eF-Site [24] database provides electrostatic potential surfaces that can be used to identify similar patterns of charge in binding sites and on larger surfaces. A different approach, relying on residue conservation, is called conserved functional group (CFG) analysis [25]. This uses a simplified definition of chemical groups, builds a multiple sequence alignment, identifies CFG clusters and maps these onto the structure to find active sites. SiteEngine [26] also uses a simplified chemical classification to match binding sites. It represents residues as pseudo-centres [27] of a given physiochemical property and uses geometrical hashing to match binding sites according to these physiochemical properties rather than amino acid identities.

A novel use of the chemical and electrostatic properties of amino acids to specifically hunt for catalytic residues is THEMATICS (theoretical microscopic titration curves). The procedure models the electrostatics of the system and predicted pK_a values of ionisable groups. Each residue type has its mean net charge predicted as a function of pH. Curves are compared manually and active site residues are predicted on the basis of perturbed theoretical titration curves [28].

Residue template methods

The function of certain types of proteins is effected by a small number of residues found in a localized region of the three-dimensional structure. In enzymes, for example, the enzyme’s catalytic function will be performed by a small number of catalytic residues located in the active site. For DNA-binding proteins, a handful of residues on the surface may be responsible for the specificity of binding to a particular sequence of DNA. And so on. Often, the specific arrangement and conformation of the residues are crucial to the performance of the function and remain strongly conserved over evolutionary time, even as the remainder of the protein’s sequence and structure undergoes major changes.

Such highly conserved conformations underpin three-dimensional residue template methods (for a review,
Some approaches place the onus on the user to define patterns of residues to be searched for. The ASSAM program [36] and RIGOR/SPASM [37] programs search protein structures for occurrences of user-defined patterns of residues or residue properties. The MSDsite database allows users to define patterns for search against the PDB.

The PINTS (Patterns In Non-homologous Tertiary Structures) program is not a template method as such, but it exploits the same basic idea that similar arrangements of residues in two proteins may imply similar functions. The program detects the largest common three-dimensional arrangement of residues between any two structures, or between a query structure and all those in the PDB [38].

The Phunctioner [39**] program uses GO annotation as a core element of its assignment and validation. Proteins with the same GO annotation are extracted from the FSSP database to create position-specific substitution matrices (PSSMs). These three-dimensional profiles can then be used to scan a new protein for matches to particular functions. One benefit of the method is that it generates profiles for GO terms for which the functional residues are not known. These could form the basis of new structural templates of the type discussed above. Unfortunately, the method is not yet publicly available.

**DNA-binding proteins**

Proteins that bind DNA use a limited repertoire of structural motifs for binding the DNA. The most common of these is the helix-turn-helix (HTH) motif. However, many proteins that do not bind DNA also contain this motif, so the problem is one of correctly identifying the DNA binders and discarding the others. As the protein surfaces that bind DNA tend to be the most positively charged surfaces of the structure, the use of electrostatic potentials can significantly improve the differentiation between true and false positive hits to the HTH motif [40]. DNA-binding surfaces can also be located by eF-Site, already mentioned above, and other methods for identifying the protein–DNA interaction surface [41]. Some servers are listed in Table 2.

**Phylogenetic relationships**

The active sites of proteins can also be predicted using phylogenetic analysis and assessment of ‘tree-determinant residues’ [42], as the functionally important regions of the structure are those most likely to have been strongly conserved over evolutionary time [43,44]. The most common method is the evolutionary trace (ET) method [45], which uses a phylogenetic tree to rank the residues in a protein by evolutionary importance and map this onto the structure. The highest-ranked residues often cluster together and can be used to identify functional sites. Many servers exist to perform ET and a Java implementation of the method called JevTrace [46] is available for downloading. The method of Chelliah et al. [47] aims to distinguish between residues conserved for functional reasons and those conserved through structural constraints. The method achieves this distinction through the use of environment-specific substitution matrices when calculating conservation scores.

In some protein families, a subgroup can evolve a new function or substrate specificity. These so-called functional shifts are documented in the FunShift database [48], but are of greater importance for mutagenesis studies than for functional annotation.

**Machine learning techniques**

All the methods discussed so far rely on finding structural similarity, whether at the overall fold level or in the
<table>
<thead>
<tr>
<th>Method</th>
<th>Program/ server</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fold similarity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DALI</td>
<td><a href="http://www.ebi.ac.uk/dali/">http://www.ebi.ac.uk/dali/</a></td>
<td>An algorithm for optimal pairwise alignment of protein structures using $\alpha$ carbon residue-residue distance matrices.</td>
<td></td>
</tr>
<tr>
<td>GRATH</td>
<td><a href="http://www.biochem.ucl.ac.uk/cgi-bin/cath/Grath.pl">http://www.biochem.ucl.ac.uk/cgi-bin/cath/Grath.pl</a></td>
<td>-Graph theory approach to compare a protein structure against a representative library of structures in CATH.</td>
<td></td>
</tr>
<tr>
<td>FAST</td>
<td><a href="http://biowulf.bu.edu/FAST/">http://biowulf.bu.edu/FAST/</a></td>
<td>-A new, rapid algorithm for aligning three-dimensional structures of proteins. The executable can be downloaded for standalone use.</td>
<td></td>
</tr>
<tr>
<td><strong>Pockets and clefts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pvSOAR</td>
<td><a href="http://pvsoar.bioengr.uic.edu/">http://pvsoar.bioengr.uic.edu/</a></td>
<td>-Web server to detect surface similarities in proteins. Searches against the CASTp database or non-redundant PDB sets. Can also compare pockets against one another.</td>
<td></td>
</tr>
<tr>
<td>CASTp</td>
<td><a href="http://cast.engr.uic.edu/cast/">http://cast.engr.uic.edu/cast/</a></td>
<td>-A database providing identification and measurements of surface-accessible pockets, as well as interior inaccessible cavities, for proteins and other molecules.</td>
<td></td>
</tr>
<tr>
<td>SURFNET</td>
<td><a href="http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html">http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html</a></td>
<td>-A program to generate surfaces and void regions between surfaces from coordinate data supplied in PDB format. Can output in a variety of formats.</td>
<td></td>
</tr>
<tr>
<td>SURFACE</td>
<td><a href="http://cbm.bio.uniroma2.it/surface/">http://cbm.bio.uniroma2.it/surface/</a></td>
<td>-A database of protein surface patches annotated with sequence- and structure-derived information about function or interaction abilities.</td>
<td></td>
</tr>
<tr>
<td><strong>Active sites and templates</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSDsite</td>
<td><a href="http://www.ebi.ac.uk/msd-srv/mdsite/index.jsp">http://www.ebi.ac.uk/msd-srv/mdsite/index.jsp</a></td>
<td>-Active site searches based on ligand or active site information. Can also search an uploaded file. Searches can be restricted by author, keywords, experiment, resolution, etc.</td>
<td></td>
</tr>
<tr>
<td>PROCAT</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html">http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html</a></td>
<td>-A database of three-dimensional enzyme active site templates.</td>
<td></td>
</tr>
<tr>
<td>CSA</td>
<td><a href="http://www.ebi.ac.uk/thornton-srv/databases/CSA/">http://www.ebi.ac.uk/thornton-srv/databases/CSA/</a></td>
<td>-A database documenting enzyme active sites and catalytic residues in enzyme three-dimensional structures.</td>
<td></td>
</tr>
<tr>
<td>PINTS</td>
<td><a href="http://www.russell.embl.de/pints/">http://www.russell.embl.de/pints/</a></td>
<td>-Allows comparison of a protein structure against a database of patterns, a structural pattern against a database of protein structures or two proteins directly. Hits are evaluated using rmsd and E-value.</td>
<td></td>
</tr>
<tr>
<td>SPASM/ RIGOR</td>
<td><a href="http://usf.wehi.edu.au/usf/spasm.html">http://usf.wehi.edu.au/usf/spasm.html</a></td>
<td>-SPASM takes a user-defined motif of mainchain and/or sidechains, and compares it against a database of structures. RIGOR searches a database of pre-defined motifs to see if any of them occur in the submitted protein structure.</td>
<td></td>
</tr>
<tr>
<td>SuMo</td>
<td><a href="http://sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome">http://sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome</a></td>
<td>-Uses graph theory to match triangles of chemical groups rather than amino acids.</td>
<td></td>
</tr>
<tr>
<td>PDBSiteScan</td>
<td><a href="http://www.mgs.bionet.nsc.ru/mgs/gnw/pdbsiteScan/">http://www.mgs.bionet.nsc.ru/mgs/gnw/pdbsiteScan/</a></td>
<td>-Searches the PDBSite database (a database of functional sites extracted from the SITE records in all PDB files).</td>
<td></td>
</tr>
<tr>
<td>DRESPAT</td>
<td>Program available for download on request to <a href="mailto:ashish@it.itb.ac.in">ashish@it.itb.ac.in</a></td>
<td>-Graph theory approach to identify functional sites in proteins.</td>
<td></td>
</tr>
<tr>
<td>SARIG</td>
<td><a href="http://bioinfo2.weizmann.ac.il/~pietro/SARIG/V3/index.html">http://bioinfo2.weizmann.ac.il/~pietro/SARIG/V3/index.html</a></td>
<td>-The protein structure is represented as a residue interaction graph (RIG). Network analysis is then performed to identify active site residues.</td>
<td></td>
</tr>
<tr>
<td><strong>Phylogenetic analysis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ET</td>
<td><a href="http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html">http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html</a></td>
<td>-One of several ET servers.</td>
<td></td>
</tr>
<tr>
<td>ET</td>
<td><a href="http://www.cmpharm.ucsf.edu/~marcin/JEvTrace/">http://www.cmpharm.ucsf.edu/~marcin/JEvTrace/</a></td>
<td>-A JAVA implementation of the ET method is available for download.</td>
<td></td>
</tr>
<tr>
<td>ConSurf</td>
<td><a href="http://consurf.tau.ac.il/">http://consurf.tau.ac.il/</a></td>
<td>-Identifies functional regions in proteins by mapping phylogenetic information onto the surface. An advantage of the method is that it can be run solely from ATOM coordinates.</td>
<td></td>
</tr>
<tr>
<td>FunShift</td>
<td><a href="http://funshift.cgb.ki.se/">http://funshift.cgb.ki.se/</a></td>
<td>-Provides functional shift (divergence) analysis among the subfamilies of a protein domain family.</td>
<td></td>
</tr>
<tr>
<td><strong>Protein–DNA interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PreDS server</td>
<td><a href="http://pre-s-protein.osaka-u.ac.jp/~peds">http://pre-s-protein.osaka-u.ac.jp/~peds</a></td>
<td>-Uses protein surface and electrostatic potential to detect double stranded DNA binding sites.</td>
<td></td>
</tr>
</tbody>
</table>
data-mining approaches \cite{49} and machine learning techniques \cite{32} to recognize function directly from the three-dimensional structure. These include statistical methods, capable of recognizing function from genomic context, experimental data, microarray data and text mining PubMed. A database and some analysis tools for protein interaction data. All interactions are derived from literature curation or direct user submissions, and are freely available. The mammalian protein–protein interaction (PPI) database. A collection of manually curated high-quality PPI data collected from the scientific literature.

Many work has gone into the development of techniques (support vector machines and neural networks) to predict catalytic residues \cite{50}, protein function \cite{54,55}. As GO has a well-defined hierarchy, computer learning techniques should be able to spot patterns in the data that are not obvious. An example is the AnaGram \cite{56} server, which finds correlations between amino acid patterns, known as protomotifs, and functional annotations from the original SWISS-PROT entries, supplemented by information from PubMed. Although these methods are ideally suited for spotting subtle patterns not yet described, the major problem associated with machine learning and automated techniques is that they tend to work very well on the data set used to train the predictor, but are generally less successful on unseen data. Bearing this in mind, many groups report success rates in the region of 60–80\% for their predictors.

Combining methods

None of the methods described so far in this review are likely to be successful in all cases and there will still be difficult proteins about which no method is able to say very much at all. So, a sensible strategy is to use as many different methods as possible, incorporating data from multiple sources, to increase the chances of obtaining a function prediction for any given protein.

The ProFunc \cite{4} server (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) is one such attempt to combine multiple methods and multiple data sources. It has been developed in collaboration with the Midwest Center for Structural Genomics (MCSG), but access to it is freely granted to all. The methods currently incorporated in the server range from standard sequence searches (BLAST

<table>
<thead>
<tr>
<th>Table 2 (Continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Function prediction servers</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Another server that combines several different methods is ProKnow [59**] (http://www.doe-mbi.ucla.edu/Services/ProKnow). This extracts a variety of features from the submitted structure, including fold (taken from user-submitted DALI results), sequence motifs, structural motifs (RIGOR) and functional links (taken from the DIP database). Clues from each of these methods are used to assign function, with the significance of each result being weighted using Bayes’ theorem. The most likely overall function is returned to the user in terms of GO annotation, the confidence of which can be assessed using a combination of the number of sources providing the functional clues and the GO evidence codes (ranked by significance from high-confidence manually curated assignments to less confident electronic assignments). The authors suggest that, overall, 70% of assignments are correctly inferred.

From single structures to complexes and networks
Proteins do not function in a vacuum; they interact with other proteins in different locations under strict regulation by the cell. A protein’s function can be described at the biochemical level (e.g. protein X catalyses the transfer of a phosphate group from A to B) or within the context of a pathway (e.g. the protein kinase cascade) or as part of a signalling response to external conditions. The full complexity of interactions in the cell is overwhelming and data come from numerous experimental sources, such as yeast two-hybrid experiments, knockout studies and gene expression microarrays. Not all of these experimental techniques provide similar answers, and each will have associated biases and variation in the reliability of results. Additionally, theoretical predictions, such as co-occurrence of genes across genomes and mapping of GO terms [60–62] to protein–protein interaction and expression data, can also indicate potential functions or regulatory roles. For a review of the methods used for the prediction of protein function and pathways, see Gabaldon and Huynen [63].

An excellent resource that extracts data from these different sources is the STRING database at http://string.embl.de [64**]. It accepts a known gene name or a protein sequence, and returns what functional information is available from the different experimental methods. One of the most useful aspects of the database is the use of PubMed to find associations of terms in the literature. All of the evidence is weighted and scored by the confidence of the information obtained, thus providing a clear way to determine the significance of any predictions.

Conclusions and future directions
Recent years have seen the development of a variety of techniques for predicting protein function from three-dimensional structure, inspired primarily by the needs of structural genomics projects. Structure can provide clues to function in many cases, even if powerful sequence methods have failed to provide a conclusive functional assignment. Some targets are still hard to tackle by any method and certain methods are appropriate to only a specific protein type (e.g. enzyme, DNA-binding protein). The most sensible strategy, therefore, is to subject the target to a battery of different prediction methods. Web servers such as ProFunc and ProKnow are being developed to do just that.

Should several methods concur, there is greater confidence in their prediction. Specifically, the structure can greatly increase the confidence of any tentative assignments, such as from sequence matches to distant homologues. For example, if the protein is thought to be a serine protease and the structure contains the residues of the catalytic triad arranged in exactly the correct conformation, the confidence in the initial prediction is greatly increased. Ideally, all such predictions should be experimentally confirmed, as this is the only way to be absolutely sure of the protein’s function.

As more and more methods and data become available, housed on separate servers or in different programs, so it becomes more difficult to gather and piece together the different nuggets of information for a given protein or set of proteins. The reliance on web services and workflows is thus likely to increase. These are easy-to-use techniques for integrating disparate resources to automatically compile analyses from multiple sources. In Taverna [65], a workflow can be drawn up on screen using drag and drop methods, and then initiated. A paper by Stevens et al. [66] illustrates the use of such distributed web services to investigate the human disease Williams–Beuren syndrome. A wide variety of tools were used to automatically fill in and annotate gaps in a 1.5 Mb region of chromosome 7 deleted in sufferers of the syndrome. Eventually, such fully automated procedures for functional assignment could be bolted onto functional genomics pipelines, although of course any fully automated procedure is prone to the propagation of errors in databases and the pipeline needs to be extended to incorporate experimental confirmation of any functional assignment made.

versus UniProt, and BLAST versus PDB), through gene neighbour analysis and sequence motif scans (InterProScan and SUPERFAMILY), to wholly structure-based methods, including fold matching (SSM), mapping of residue conservation onto the protein surface, surface cleft analysis (SURFNET [57]), structural motifs (DNA-binding HTH motifs and ‘nests’ [58]) and three-dimensional residue templates. The last include enzyme active site templates from the CSA, automatically generated ligand- and DNA-binding templates, and a novel ‘reverse’ template method, wherein templates are generated from the target structure and scanned against representative structures in the PDB.
Update

A recently published novel method allows the comparison of protein binding pockets and shows potential for ligand prediction [67]. The method uses spherical harmonic expansion coefficients to describe the three-dimensional shape of a protein’s binding pocket. Pocket similarity can then be rapidly calculated on the basis of the distance between coefficients. The descriptors can be used to describe entire proteins, or even ligands and small molecules; it is this latter use that shows potential for the prediction of ligand binding to hypothetical proteins. Limitations and difficulties of the method are apparent, but the initial results are encouraging and, more importantly, show a move away from homology-based approaches towards more directly predictive tools.

Another new method for the prediction of protein–ligand binding sites is Q-SiteFinder [68] (http://www.bioinformatics.leeds.ac.uk/qsitefinder). This method locates energetically favourable binding sites using a methyl probe, clustering favourable probes and then ranking by total interaction energy. The novelty of this approach lies in its ranking: the most energetically favourable cluster is ranked first irrespective of the geometric size of the cluster. The authors suggest that this gives a greater success level, with at least one of the top three predictions correct for 90% of the proteins tested.

In an expansion of the CSA mentioned previously, a new web server (Catalytic Site Search) allows users to query the three-dimensional template library using the JESS algorithm. Users can submit queries based on PDB code or by uploading a protein structure. In their paper [69], the authors also perform an analysis of homologous template families to explore the evolution of active sites. The results show that templates based on protein backbone positions are more discriminating than those based on sidechain atoms. The web server is available at http://www.ebi.ac.uk/thornton-srv/databases/CSS and is also available for download.

Acknowledgements

Work was supported by the National Institutes of Health (grant GM 62414), by the US DoE (contract W-31-109-Eng-38) and also as part of the BioSapiens project. The BioSapiens project is funded by the European Commission within its FP6 program, under the thematic area ‘Life sciences, genomics and biotechnology for health’, contract number LIHG-CT-2003-503265.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


16. Novotny M, Madsen D, Kleywegt GJ: Evaluation of protein fold comparison servers. Proteins 2004, 54:260-270. The authors perform a wide-ranging evaluation of 11 publicly available fold comparison servers. They use the CATH database as a reference for their tests. The results show that no one server provides 100% accuracy and therefore multiple methods should be used to assess similarities to known structures.


21. Binkowski TA, Freeman P, Liao J: pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 2004, 32:W555-W558. The pvSOAR web server identifies similar surface regions among proteins, and takes advantage of the CASTp database of pockets and cavities. The authors describe how the server can be used to predict the function of hypothetical proteins, illustrating this with the E. coli BioH protein (PDB code 1m33) as an example.


Predicting protein function

Watson, Laskowski and Thornton


This new method describes the conservation of a protein's surface using chemical groups rather than the amino acids. A multiple sequence alignment is used to identify conserved functional group clusters, the size of which is determined by the number of proteins contributing to it. These are mapped onto the surface to identify active sites.


SiteEngine uses modified pseudo-centres and geometric hashing to compare surfaces with the aim of identifying conserved chemistry in similar pockets, which might indicate similar function.


This method distinguishes residues conserved for functional reasons from those that are highly conserved because they are constrained by the structure. By comparing the observed sequence conservation with the predicted conservation (based on amino acid type and environmental constraints), the authors construct environment-specific substitution tables for use in identifying functionally conserved residues.


The ProKnow server uses several structural methods for function identification using GO terminology. The method’s most important feature is that it attempts to weight the evidence using Bayes' theorem for each of the functional predictions, and provides details to the user concerning the confidence of each method and the confidence of each GO term.
assignment. This allows annotations to be ranked by significance. An overall success rate of 70% correct prediction is reported.


The STRING database takes data from protein-protein interaction experiments, microarray expression data, genome organization and co-occurrence to identify functional associations. One of the key features of the database is the use of text processing of PubMed to identify potential interacting partners from the literature.


Another protein–ligand binding site prediction method is described. The approach uses a methyl probe and energetic calculations to rank sites by favourable interaction energy rather than by pocket size. The method is available at http://www.bioinformatics.leeds.ac.uk/qsitefinder, and allows users to investigate any PDB code or upload a structure of their own for prediction.


The authors present a library of catalytic site structural templates based on information from the scientific literature. In an extension of previous work, a new web server is released that allows users to search the CSA using the JESS algorithm. The user can investigate a specific PDB code or submit a three-dimensional protein structure for analysis.