Structure-based function prediction: approaches and applications

Pier Federico Gherardini and Manuela Helmer-Citterich
Advance Access publication date 3 July 2008

Abstract
The ever increasing number of protein structures determined by structural genomic projects has spurred much interest in the development of methods for structure-based function prediction. Existing methods can be roughly classified in two groups: some use a comparative approach looking for the presence of structural motifs possibly associated with a known biochemical function. Other methods try to identify functional patches on the surface of a protein using only its physicochemical characteristics. This review will cover both kinds of approaches to structure-based function prediction as well as their use in real-world cases. The main issues and limitations in using protein structure to predict function will also be discussed. These are mainly: the assessment of the statistical significance of structural similarities and the extent to which these methods depend on the accuracy and availability of structural data.

Keywords: bioinformatics; function prediction; protein structure; structural comparison; active sites; structural genomics

INTRODUCTION
Before the advent of structural genomics the main interest in solving the structure of a protein was to understand and better analyse the determinants of its function, which was already assigned after biochemical or genetic experiments. The increasing number of fully sequenced genomes together with the progress in homology modelling of protein structures shifted the interest on characterizing the largest number of different folds to have the best possible sampling of structure space. The rationale is that, as more and more structural folds are characterized, homology modelling of an increasing number of proteins should become possible and more reliable. In light of this goal, targets for structure determination are selected among proteins with very low sequence identity to proteins of known structure. As a consequence a large number of structures (over one-third of those solved by the Midwest Center for Structural Genomics [1] just to cite an example) belong to proteins of unknown function. This fact has enormously increased the interest in computational methods for structure-guided functional inference. Such methods have therefore already been included in numerous reviews about function prediction [2–5] and have also been the subject of dedicated papers [6–10].

The availability of structural information is generally believed to be a very strong aid in function prediction for two essential reasons:

1. Structural comparison methods are potentially able to identify very distant evolutionary relationships between proteins. Moreover, only structural data makes the identification of independently evolved functional sites possible [11–13].
2. Function depends on structure. Therefore, the structure of a protein directly reveals the mechanistic determinants of its function.

These distinctions immediately suggest the classification of functional annotation methods that has been used in this review (Figure 1). A first group of methods uses a comparative approach searching for common features between the query protein and some database of protein structures. Other methods analyse the physicochemical characteristics

Corresponding author. Pier Federico Gherardini, Centre for Molecular Bioinformatics, Department of Biology, University of Tor Vergata, Rome, Italy. E-mail: pier.federico.gherardini@uniroma2.it

Pier Federico Gherardini is doing his PhD in Manuela Helmer-Citterich’s lab. His main interests are the development and application of local structural comparison programs.

Manuela Helmer-Citterich is a Full Professor of Bioinformatics at the University of Tor Vergata in Rome.

© The Author 2008. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
of a protein surface to identify patches that have features (e.g. shape, electrostatic properties, etc.) characteristic of functional sites.

Whatever the approach all the methods discussed in this review can be used to infer the biochemical function of a protein, i.e. for example, whether it binds a particular ligand or catalyses a chemical reaction. As such they do not (at least directly) tell anything about its biological role, i.e. for example, whether it is involved in a certain biological pathway or has a role in the development of some disease. Such kind of predictions can sometimes be made once the biochemical function is known but falls outside the scope of these methods.

COMPARATIVE APPROACHES

Similar to sequence comparison methods, structural comparison algorithms can be classified as global or local. Global comparison algorithms, summarized in Table 1, are mainly used in protein structure classification and to identify evolutionary links between distant homologues. They can also be used for function prediction but one should be aware that the relationship between fold and function is extremely complex and numerous examples are known of folds hosting a great variety of functions [42]. It should indeed be noted that the function of a protein usually depends more on the identity and location of a few residues comprising the active site than on the overall fold. Therefore, the usefulness of global comparison methods is essentially indirect and lies in their capability of identifying remote homology relationships. In order to directly analyse and compare the residues effectively involved in protein function, local structural comparison methods have been developed.

Local structural comparison refers to the possibility of detecting a similar 3D arrangement of a small set of residues, possibly in the context of completely different protein structures. In applying such algorithms one can either:

1. compare two entire protein structures in search for local similarities, without any a priori assumption; or
2. use a pre-defined structural template to screen a structure. A template represents the spatial arrangement of the residues involved in some biochemical function and can be regarded as a 3D extension of the linear sequence motif concept.

Figure 1: An overview of structure-based function prediction methods as classified in this review. Comparative methods can be either global or local with the latter class including template-based methods. Methods based on structural calculations are based on the observation that the functional patches of a protein have characteristics that set them apart from the surface as a whole. Such distinctive features can be used alone or combined using machine learning methods.
### Table 1: Global structural comparison methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Structure representation</th>
<th>Search strategy</th>
<th>Webserver</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DaliLite</td>
<td>Cα-Cα distance matrix</td>
<td>Branch and bound</td>
<td><a href="http://www.ebi.ac.uk/DaliLite/">http://www.ebi.ac.uk/DaliLite/</a> <a href="http://www.ebi.ac.uk/msd-srv/ssm/">http://www.ebi.ac.uk/msd-srv/ssm/</a></td>
<td>[14]</td>
</tr>
<tr>
<td>SSM</td>
<td>Graph with nodes representing secondary structure elements and edges their spatial relationship</td>
<td>Subgraph isomorphism followed by Cα alignment in 3D</td>
<td><a href="http://www.ebi.ac.uk/DaliLite/">http://www.ebi.ac.uk/DaliLite/</a> <a href="http://www.ebi.ac.uk/msd-srv/ssm/">http://www.ebi.ac.uk/msd-srv/ssm/</a></td>
<td>[15]</td>
</tr>
<tr>
<td>GRATH</td>
<td>Graph with nodes representing secondary structure elements and edges their spatial relationship</td>
<td>Subgraph isomorphism</td>
<td>–</td>
<td>[16]</td>
</tr>
<tr>
<td>SSAP</td>
<td>Cα, Cβ. Several other structural features are used in scoring</td>
<td>Double dynamic programming</td>
<td><a href="http://www.cathdb.info/cgi-bin/cath/SSAPServer.pl">http://www.cathdb.info/cgi-bin/cath/SSAPServer.pl</a></td>
<td>[17]</td>
</tr>
<tr>
<td>CATHEDRAL</td>
<td>Combines GRATH and SSAP</td>
<td>Combines GRATH and SSAP</td>
<td><a href="http://www.cathdb.info/cgi-bin/cath/CathedralServer.pl">http://www.cathdb.info/cgi-bin/cath/CathedralServer.pl</a></td>
<td>[18]</td>
</tr>
<tr>
<td>VAST</td>
<td>Graph with nodes representing secondary structure elements and edges their spatial relationship</td>
<td>Subgraph isomorphism</td>
<td><a href="http://www.cathdb.info/cgi-bin/cath/CathedralServer.pl">http://www.cathdb.info/cgi-bin/cath/CathedralServer.pl</a></td>
<td>[19]</td>
</tr>
<tr>
<td>CE</td>
<td>Cα atoms</td>
<td>Extension of seed matches using a greedy heuristic; optimization of best alignments</td>
<td><a href="http://cl.sdsc.edu/ce.html">http://cl.sdsc.edu/ce.html</a></td>
<td>[20]</td>
</tr>
<tr>
<td>LSQMAN</td>
<td>User-defined atom types (typically Cα)</td>
<td>Alternates structural superposition and alignment to improve an initial transformation</td>
<td>Downloadable from: <a href="http://xray.bmc.uu.se/usf/s">http://xray.bmc.uu.se/usf/s</a></td>
<td>[21]</td>
</tr>
<tr>
<td>DEJAVU</td>
<td>Matrices of distances and angles between vectors of secondary structure elements</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://portray.bmc.uu.se/cgi-bin/dejavu/scripts/dejavu.pl">http://portray.bmc.uu.se/cgi-bin/dejavu/scripts/dejavu.pl</a></td>
<td>[22]</td>
</tr>
<tr>
<td>FATCAT</td>
<td>Cα atoms</td>
<td>Similar to CE but takes flexibility into account</td>
<td><a href="http://fatcat.burnham.org/">http://fatcat.burnham.org/</a></td>
<td>[25]</td>
</tr>
<tr>
<td>TOPS</td>
<td>Graph with nodes representing secondary structure elements and edges their spatial relationship</td>
<td>Branch and bound</td>
<td><a href="http://www.tops.leeds.ac.uk/">http://www.tops.leeds.ac.uk/</a></td>
<td>[26]</td>
</tr>
</tbody>
</table>

This table is intended as a quick overview, please refer to the original papers for details about the algorithms involved. See Novotny et al. [27] and Kolodny et al. [28] for a comprehensive evaluation of fold comparison methods.

### Table 2: Local structural comparison methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Structure representation</th>
<th>Search strategy</th>
<th>Webserver</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSAM</td>
<td>Vector from Cα to functional part of residue</td>
<td>Subgraph isomorphism</td>
<td>–</td>
<td>[29]</td>
</tr>
<tr>
<td>CavBase</td>
<td>Physicochemically labelled surface points</td>
<td>Subgraph isomorphism</td>
<td>–</td>
<td>[30]</td>
</tr>
<tr>
<td>eF-Site</td>
<td>Curvature and electrostatic potential of surface points</td>
<td>Subgraph isomorphism</td>
<td><a href="http://ef-site.hgc.jp/ef-seek/">http://ef-site.hgc.jp/ef-seek/</a></td>
<td>[31]</td>
</tr>
<tr>
<td>C-alpha Match</td>
<td>Coordinates of Cα atoms</td>
<td>Geometric hashing</td>
<td><a href="http://bioinfo3d.cs.tau.ac.il/calpha.match/">http://bioinfo3d.cs.tau.ac.il/calpha.match/</a></td>
<td>[32]</td>
</tr>
<tr>
<td>Prospect</td>
<td>Each residue is represented as a triangle using the N, Cα and C atoms</td>
<td>Geometric hashing</td>
<td>–</td>
<td>[33]</td>
</tr>
<tr>
<td>SiteEngine</td>
<td>Physicochemically labelled surface points</td>
<td>Geometric hashing</td>
<td><a href="http://bioinfo3d.cs.tau.ac.il/SiteEngine/">http://bioinfo3d.cs.tau.ac.il/SiteEngine/</a></td>
<td>[34]</td>
</tr>
<tr>
<td>PINTS</td>
<td>Cα, Cβ, functional atom of each residue</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://www.russell.embl-heidelberg.de/pints/">http://www.russell.embl-heidelberg.de/pints/</a></td>
<td>[36]</td>
</tr>
<tr>
<td>RIGOR/SPASM</td>
<td>Cα, geometric centroid of side chain</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://portray.bmc.uu.se/cgi-bin/spasm/scripts/spasm.pl">http://portray.bmc.uu.se/cgi-bin/spasm/scripts/spasm.pl</a></td>
<td>[37]</td>
</tr>
<tr>
<td>Query3d</td>
<td>Cα, geometric centroid of side chain Template expressed as a series of arbitrary constraints</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://pdbfun.uniroma2.it">http://pdbfun.uniroma2.it</a></td>
<td>[38]</td>
</tr>
<tr>
<td>JESS</td>
<td>Cα, geometric centroid of side chain</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://www.catalyst.scripps.edu/pdbsitescan.html">http://www.catalyst.scripps.edu/pdbsitescan.html</a></td>
<td>[39]</td>
</tr>
<tr>
<td>PDBSiteScan</td>
<td>Coordinates of N, Cα and C atoms</td>
<td>Recursive search, branch and bound</td>
<td><a href="http://wwwbioc">http://wwwbioc</a> recognize.com/pdbfun/</td>
<td>[40]</td>
</tr>
<tr>
<td>SuMo</td>
<td>Triangles of chemical groups</td>
<td>Graph-based heuristic</td>
<td><a href="http://sumo-pbil.ibcp.fr/">http://sumo-pbil.ibcp.fr/</a></td>
<td>[41]</td>
</tr>
</tbody>
</table>
The various methods available for local structure comparison, summarized in Table 2, differ essentially in two aspects: the way the protein structure is represented and the computational strategy that is used to search for similarities. The level of detail in the representation goes from very approximate, i.e. only the Cα atoms, to elaborate schemes that take into account the presence of different chemical groups along the amino acids side chains. We will not go into the details of the different representation schemes since constructing elaborate representations is usually quite easy in that one only has to decide which residues are allowed to match. Once the pairing rules are established the actual search strategy (see further) is usually independent of the way residues are represented. Increasing the level of detail is not necessarily advantageous. Torrance et al. [43] compared the performance of templates using only the Cα and Cβ with that of templates using three functional atoms to describe each residue. Cα/Cβ templates were found to consistently outperform the others. Different levels in the quality of the structures as well as alternative residues conformations, e.g. a bound/unbound transition in a ligand-binding site, can determine structural differences even in genuinely identical sites. Because of these differences a higher level, i.e. less detailed, description might be better for practical purposes.

The work by Kolodny and Linial [44] is one of the most interesting theoretical results about the approximability of the problem of structural comparison, and the effectiveness of various algorithmic approaches in solving it. They first discretized rotation and translation space and showed that its size depends polynomially on the lengths of the proteins, n, and on 1/ε for an approximation parameter ε. On the other hand the possible correspondences between two protein structures grow exponentially with the number of residues. Their algorithmic strategy is therefore to search exhaustively the transformation space (whose size is polynomial in n) and then choose the best solution according to some scoring function. Their theoretical algorithm runs in $O(n^{10}/\varepsilon^6)$ time. This strategy is possible because the structures reside in a three-dimensional Euclidean space. They were also able to show that, if one uses a representation of protein structure that does not take this fact into account e.g. distance matrices, the problem becomes significantly harder, and therefore such an algorithm either fails to find optimal solutions or is inefficient.

In terms of search strategy, three approaches prevail in practice: detection of sub-graph isomorphism [45], geometric hashing and recursive enumeration using a branch and bound strategy. It should be emphasized that graph methods are presented separate from those based on recursive enumeration only to follow the way these methods are reported in the literature. There is really no algorithmic difference since sub-graph isomorphism (at least as applied in the methods discussed) relies on clique detection algorithms (e.g. the Bron–Kerbosch algorithm [46]) that use a branch and bound strategy. These two approaches can therefore be mapped one onto the other. In general terms, the core of these methods is a recursive procedure that is used to extend initial candidate solutions. The extension stops when the algorithm determines that the current path cannot lead to solutions that are better than the current best one. In such case, the recursion goes back one level and the candidate is extended in another direction or another candidate is selected. The running time of these methods depends dramatically on how similar the proteins to be compared are. If the structures are very similar, then there will be a large number of seed matches to explore or, in graph theoretic terms, a very dense product graph to analyse. This fact, combined with other considerations about the interpretation of the results (see further), makes local structural comparison algorithm not suited to the comparison of two homologous protein structures.

Geometric hashing [47], first used for structure comparison by Fischer et al. [32] is a technique where the coordinates of a structure are expressed relative to several reference frames, for example, one for each set of three points of the protein. Since the points used as a reference belong to the structure itself this representation is invariant under both rotation and translation. For each frame the positions in which the other points end up are used as keys to a hash table. The value stored in the table is the reference frame itself. Once such representation has been calculated it is possible to compare two structures using a series of fast look-ups.

In general, local structural comparison methods can also be used to search for templates. Algorithms more specialized for templates, however, allow for a semantically more complex description. As an example they may allow different geometric
constraints or substitution rules for different portions of the template. The JESS [39] algorithm even gives the possibility of expressing the template as a set of high-level constraints of arbitrary nature. To overcome the difficulty that structural templates must be derived manually some authors have developed methods for the automatic discovery of structural motifs characterizing a protein family [48–50]. Polacco and Babbitt [48], for example, used a genetic algorithm to derive specific motifs that distinguish proteins belonging to a given enzymatic family from a background of unrelated structures. Their approach involves randomly modifying a set of structural motifs so as to maximize their discriminative power in successive rounds of structural comparisons. Oldfield [51] derived templates by constructing a hash table from inter-residue distance in order to count over-represented residue configurations. The program DRESPAT [50] uses a graph theoretic method to enumerate patterns recurring in a set of structures. The authors also empirically derived a function for assessing the significance of the patterns discovered.

**Significance assessment**

The problem of assessing the statistical significance of a local structural similarity is, at the time of this writing, largely unresolved. The biggest gap, in terms of statistical analysis, between sequence comparison and structure comparison is that in the latter case there is no universally accepted random model that can be used as a basis for significance assessment. The definition of ‘random’ is especially problematic in this case and much depends on what sort of information one wants to derive from the comparison of two protein structures. Proteins are subjected to strong constraints in order to achieve stability; this is reflected in the fact that the same scaffolding elements (helices, sheets and supersecondary structures, for example) are reused all across structure space. How should similarities between these elements be considered? A perfect match between two Greek-key motifs could be considered very interesting in the context of a fold comparison application, and would definitely stand out if the reference state (or background distribution) consists of points randomly positioned in space.

Alternatively, if one is using structural comparisons for functional annotation, such a match would probably be considered irrelevant since it only reflects the fact that proteins are composed of similar fragments. In other words, one would like to be able to separate structural similarities between residues that are due to the fact that they perform similar functions (e.g. active sites) from those that simply derive from the necessity to attain a stable conformation in solution, and therefore carry no functional information. RMSD alone is not of much help here because matches between secondary structure elements will in general tend to be better than matches between, for example, active sites. In this case, the choice of reference state should take into account scaffolding elements as opposed to randomly scattered points.

In order to solve this problem, different authors have either used empirical methods, i.e. using similarities between random pairs of structures to fit a distribution to be used as a background, or semi-empirical approaches. In the latter case, theoretical considerations are used to build a model that is parametrized by some terms that are eventually estimated by fitting to various runs of comparisons. Empirical approaches are also often used by template-based methods. Since templates are constructed from the analysis of a protein family one can usually derive template-specific RMSD thresholds in order to discriminate true positives from random similarities (e.g. see Torrance et al. [43], and Arakaki et al. [52]).

Betancourt and Skolnick [53] used structural comparisons between randomly selected protein fragments of different lengths to derive a similarity measure (‘relative RMSD’, RRMSD) which is dimensionless and independent of protein size. They first defined an aligned correlation coefficient (ACC) which is a measure, strictly related to RMSD that expresses the similarity of two chains after optimal superposition. They subsequently calculated the average and SDs of the ACC from continuous N-residues fragments of almost 1300 non-homologous structures chosen at random from the PDB [54]. The plot of this quantity as a function of fragment length defines two characteristic lengths approximately given by 4.7 and 37 residues. They argue that fragments of less than five residues have very restricted conformations: similarities of this length should therefore be considered not significant since they are simply the result of the general constraints imposed by protein folding. Fragments between 5 and 37 residues on the other hand have significant correlations generated by the recurring types of secondary structure elements. They then
defined a new similarity measure between two polypeptides, RRMSD, as the RMSD relative to the approximate average RMSD between two random protein fragments of equivalent size. An RRMSD value of 0.0 means that the two structures are identical, while a value of 1.0 means that they are as different as random structures on average. An important shortcoming of their work relative to the problem of local structural similarity is that they considered only fragments of residues contiguous in sequence while local comparison methods usually do not have this restriction.

Stark et al. [55] used a semi-empirical method to derive a formula for the statistical significance of a local structural match as a function of the number of residues involved, their RMSD, their abundance in the data set under analysis and the number of atoms used to represent each residue during the comparison. They used a series of geometric considerations to derive a crude formula for the expected number of matches with RMSD under a given threshold. If a search algorithm uses more than one atom to represent a residue, their model takes into account the mutual dependence of the atomic positions. The formula includes a number of parameters that have been estimated by selecting random patterns of residues and searching for them in a background database.

Since no definitive way exists to identify statistically significant matches, it is very important to integrate structural comparison methods with detailed functional annotations in order to steer the search towards functionally significant residues. When a match between two proteins involves residues for which functional annotations are available, one can more easily derive clues about the significance of a correspondence even in the absence of a rigorous statistical framework. Various tools have been developed to ease the functional annotation of protein structures. The E-MSD structure database [56] aggregates an enormous amount of functional information about protein structures, coming from a variety of sources. SPICE [57] is a graphical client for the DAS [58] system; it allows annotations from different laboratories hosting a DAS server to be mapped and displayed on the structure of a protein of interest. The pdbFun [59] webserver integrates residue-level functional annotations with the Query3d [38] local structural comparison method so that the residues to be used in a comparison run can be selected on the basis of functional information.

METHODS BASED ON STRUCTURAL CALCULATIONS

In general, all these methods are based on the observation that the functional patches of a protein have physicochemical characteristics that set them apart from the surface as a whole. Indeed, these peculiarities ultimately are the reason why these patches possess a function at all. The aim of these methods is usually to predict either the location of a ligand-binding site or that of an enzyme active site. This review is focused on purely structure-based methods; therefore, algorithms that also use sequence analysis (residue conservation, in particular) have been excluded. Table 3 provides a summary of publicly available programs.

Countless algorithms exist that employ the notion that functional sites are usually located in clefts on the protein surface [60]. This simple fact is used either directly to predict the location of functional sites, or as a first step to identify candidate residues before further scoring procedures are applied. Methods for identifying cavities in a protein surface include PASS [61], CASTp [62], LIGSITE [63], VOIDOO [64], SURFNET [65], APROPOS [66], CAVER [67] and PocketPicker [68]. Besides being located in clefts various authors have reported active site residues as being close to the centroid of the structure [69], having a destabilizing effect on the structure [70], interacting with a high number of residues of the same protein [71], having perturbed pKa values [72] and inducing peaks in the electrostatic potential around the protein [73]. All these observations have been used to develop methods aiming at the inference of active site location from structure. Electrostatic calculations have also been used to predict DNA binding sites. In particular, they have been combined with the analysis of the curvature of the molecular surface [74] and the detection of specific structural motifs [75].

The THEMATICS [72] program shows the power of computing chemical properties of the structure in order to predict active site location. This method starts with the observation that amino acids involved in catalysis usually have pKa values that differ from the standard values in solution. Therefore, a computational procedure is used to calculate the theoretical pKa of each amino acid side chain of a given protein structure. Cluster of residues with perturbed pKa values are assumed to identify the location of the active site. THEMATICS has recently been applied to a test set of 169 enzymes.
The authors distinguish between the prediction of all the residues listed as catalytic in the data set, and the prediction of a cluster of residues containing some of the catalytic residues, which they call site prediction. In the first case, THEMATICS has a recall rate of 50% and a precision of 18%. Active site predictions instead are more precise with a success rate of 85%. The authors also claim that the main improvement of THEMATICS with respect to the other methods tested is an increase in precision, i.e. their prediction are less spread out on the protein surface and more tailored towards the real active site. The high success rate for the prediction of catalytic sites shows that this method can be useful for functional annotation of proteins from structural genomic projects, at least in providing clues to the location of the active site.

Several methods exist that take into account a combination of structural features, such as hydrophobicity, surface curvature, electrostatic properties, etc. to infer the location of active sites. One of the earliest examples of this approach is the paper by Jones and Thornton [77]. They developed a method to predict protein–protein interaction sites taking into account six structural parameters, namely: solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area. They first performed a preliminary analysis [78] and verified that different types of protein–protein interfaces have different properties. This notion was used to construct three scoring functions, one for each category of interface, that are different linear combinations of the six parameters.

Keil et al. [79] used six physicochemical properties to describe protein surfaces and trained a neural network to classify surface points as not involved in binding, or forming complexes either with other proteins, DNA/RNA molecules or small ligands. Similarly Nayal and Honig [80] combined several surface descriptors to develop a method for predicting ligand binding sites. Each surface cavity is described using 408 physicochemical and structural features. These features were then used to train a classifier that distinguishes between those cavities that are likely to bind a ligand and those that are not. Interestingly, only 18 of these features proved to be statistically significant and these are mostly related to the size and shape of the cavity. Therefore, when dealing with small molecule binding cavities, size and shape seem to be more important than electrostatic interactions.

The ProMateus server [81] took the ‘combination of features’ approach even further. This server allows the user to propose new structural characteristics that may be useful in predicting protein–protein and protein–DNA interaction sites. Researchers can download a database and upload back the values of the new feature whose usefulness they want to

<table>
<thead>
<tr>
<th>Table 3: A summary of publicly available methods that predict functional sites by calculating physicochemical properties of the protein structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>ProMate</td>
</tr>
<tr>
<td>ProMateus</td>
</tr>
<tr>
<td>HotPatch</td>
</tr>
<tr>
<td>Q-SiteFinder</td>
</tr>
<tr>
<td>THEMATICS</td>
</tr>
<tr>
<td>PreDs</td>
</tr>
<tr>
<td>HTHQuery</td>
</tr>
</tbody>
</table>
explore. ProMateus will perform a series of statistical analyses and train a logistic regression model using the features already present together with the new one that is being proposed. A feature selection procedure will determine whether the new property is irrelevant, is relevant but overlaps with existing features or provides new information that effectively improves the prediction. This server is also an interesting experiment on the applicability of an open and community-driven research approach.

Several methods used to predict ligand binding sites calculate the interaction energy between the surface of the protein and a chemical probe. Cluster of regions with favourable interaction energy are then predicted to be ligand binding pockets. Q-SiteFinder [82] uses a methyl group to probe the structure. Conversely the method by Silberstein et al. [83] uses a variety of hydrophobic compounds to scan the structure and identifies a 'consensus' site that binds the highest number of probes. Ruppert et al. [84] developed a method that scans the surface with three molecular fragments (hydrophobic probe, hydrogen bond donor and acceptor). Clusters of points with high affinity for the probes are used to define the 'stickiest' regions of the surface. This representation of molecular surface can be used directly for small molecule docking, using the probes virtually bound to the binding pocket as anchors for the chemical groups of the ligand.

Since structural genomics is going to increase the number of sequences amenable to homology modelling it is interesting to investigate whether structure-based function prediction methods can be applied to models. For example, Szilagyi and Skolnick [85] developed a method to predict DNA binding sites and evaluated its performance as a function of the errors in the atomic coordinates. They used 10 easily computable features consisting in the proportion and spatial asymmetry of some amino acids and the dipole moment of the protein. These features were used to train a logistic regression model. Interestingly, their algorithm only needs the position of the Cα atoms and can therefore be applied to protein models and low-resolution structures. They subsequently generated a set of structural decoys with deviations up to 6 Å and evaluated their method on this set of incorrect structures. Since the properties used are quite coarse-grained there is a very small drop in performance when structures 6 Å away from the native are used. Methods using more specific properties of the structure may be more sensitive to small coordinate errors and therefore could be much less effective when applied to models.

**PRACTICAL APPLICATIONS**

Numerous cases have been reported where the structure of a protein provided essential clues for the discovery of its function (see Zhang and Kim [90] and Shin et al. [91] for a review). In most of the examples reported to date, the key to function prediction was the usage of fold comparison methods. Indeed such methods markedly increase the probability of finding a homologue from which function can be transferred to the protein of interest. However, this approach can often give only general indications; a detailed comparison of the active sites is necessary in order to make fine-grained distinctions, e.g. in ligand binding specificity or catalytic mechanism. For instance, in a revealing experiment, Shin et al. [92] determined the structure of a protein, which has a phosphatase domain belonging to a well-characterized protein family and a substrate binding domain whose fold was previously unknown. Local structural comparison methods were used to analyse the latter domain and infer specificity for carbohydrate molecules. Such prediction was then experimentally validated [93]. Besides such specific examples various authors have performed large-scale function prediction experiments [1, 94–97], some of which detailed further, that have definitely shown that incorporating structure-based methods in functional annotation pipelines provides fundamental insights.

In evaluating these methods, it is very important to distinguish what insights can be gained from the structure that was not already available by analysing the sequence. Watson et al. [1] recently directly compared the usefulness of structure- and sequence-based methods for function prediction. They used 282 non-redundant proteins solved by the Midwest Center for Structural Genomics, only 92 (33%) of which had a known function and were used as a benchmark. They used a variety of methods for function prediction, but the structure-based ones included only comparative approaches (both local and global). The results of their analysis show that, when sequence similarity is strong, sequence-based methods have the best performance. On the other hand they could be used only for 21% of the 67% of proteins of unknown function.
The performance of structure-based methods was evaluated using the area under the receiver operating characteristic curve, and global and local structural comparison methods showed a good performance of 0.83 and 0.70, respectively. Interestingly, a number of cases are described where template methods were fundamental to achieve a correct prediction. The authors also note that structure-based methods can be useful in restricting the options when sequence divergence is high and sequence-based methods suggest a wide range of possible functions. This work shows that structure-based functional inference can be useful in practice. Obviously, their data set is biased towards protein with low sequence identity with known proteins. In such a niche, structure-based methods have clear advantages. In a real-life scenario, their usefulness obviously depends on how often the more straightforward sequence-based methods cannot be applied.

Ferré et al. [96] used the SURFACE [98] database of surface patches annotated for their binding abilities and also by mapping PROSITE [99] and ELM [100] patterns on the structure. The Query3d structural comparison algorithm [38] was used to compare this compendium of functionally characterized patches with a set of 513 protein chains of unknown function and coming from structural genomics projects. The authors identified 534 matches and were thus able to suggest one or more molecular functions for 191 of these chains. Interestingly, a literature search revealed that 60% of the functional assignments were validated by experiments already performed, demonstrating the power of local structural comparison methods. Stark et al. [97] performed a similar analysis on a different data set of 157 structural genomics proteins. By using local structural comparison of functional sites they were able to increase the confidence of 17 functional assignments made by fold comparison. More interestingly they were able to suggest a function for 12 proteins with novel folds.

CONCLUDING REMARKS

Even though structures are generally believed to be more informative than sequences it is not completely clear whether structure comparison can outperform sequence comparison in the inference of protein function. One thing that must be taken into account is the sheer size of sequence information available. Often this is more than enough to compensate for the supposed lesser informativeness, so that using structure comparison methods does not add much to what has been discovered by sequence analysis alone. The ideal application of these methods is in inferring the function of a protein that has no close homologues of known function. In this sense, they are related to structural genomics. Alternatively, when several structures of a protein family are available, they can be useful in the fine-grained distinction of function specificity (e.g. ligand binding specificity) between homologous proteins.

This review has classified available methods as either using a comparative approach or using structural data to calculate some properties relevant to function. With respect to the first class of methods we believe that the two main areas needing improvement are the integration of functional annotations and the development of statistical models for significance assessment. As already noted above these two issues are somewhat complementary since a better integration of existing annotations would partly alleviate the problem of not having reliable statistical models. Methods performing structural calculations will become more and more useful as both computing power and our knowledge of what make active sites ‘special’ increase.

These limitations notwithstanding, structure-based function prediction as a methodology has already proven itself to be very useful, especially when dealing with proteins that do not have homologues of known function. Therefore, if the combined advances in structural genomics and modelling techniques will make it significantly easier to obtain the structure of a protein, we may expect structure-based methods to become standard tools in functional annotation pipelines.

Key Points

- Structure-based function prediction methods can be broadly classified as using a comparative approach or performing a computational analysis of structural properties.
- These methods are especially useful when the protein of interest does not have close homologues.
- Several concrete examples have already proven the usefulness of these methods.
- Structural genomics is going to increase the reliability and applicability of structure-based function prediction.

Funding

Funding to pay the Open Access publication charges for this article was provided by AIRC.
Acknowledgements
This review was supported by AIRC.

References
20. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–47.