# Ligand binding: functional site location, similarity and docking

Stephen J Campbell, Nicola D Gold, Richard M Jackson* and David R Westhead†

Computational methods for the detection and characterisation of protein ligand-binding sites have increasingly become an area of interest now that large amounts of protein structural information are becoming available prior to any knowledge of protein function. There have been particularly interesting recent developments in the following areas: first, functional site detection, whereby protein evolutionary information has been used to locate binding sites on the protein surface; second, functional site similarity, whereby structural similarity and three-dimensional templates can be used to compare and classify and potentially locate new binding sites; and third, ligand docking, which is being used to find and validate functional sites, in addition to having more conventional uses in small-molecule lead discovery.

**Addresses**
School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK
*e-mail: jackson@bmb.leeds.ac.uk
†e-mail: westhead@bmb.leeds.ac.uk

**Current Opinion in Structural Biology** 2003, **13**:389–395

This review comes from a themed issue on
Sequences and topology
Edited by Mark Gerstein and Janet M Thornton

0959-440X/03/$ – see front matter

**DOI 10.1016/S0959-440X(03)00075-7**

**Abbreviations**

| | |
|---|---|
| **MPC** | mutually persistently conserved |
| **QSAR** | quantitative structure–activity relationship |
| **SH2** | Src homology 2 |

## Introduction

It is well known that protein function is intimately related to three-dimensional structure and high-throughput structural genomics projects are now starting to increase the structural information available for genome sequences. It seems likely, however, that many of these structures will be relatively poorly characterised in terms of biological or biochemical function. Ligand binding, the subject of this review, is a key aspect of protein function, mediating the ability of proteins to recognise their natural ligands for transport, signal transduction or catalysis, and also our ability to modulate function through the discovery of drugs. Facing the possibility of large amounts of relatively uncharacterised protein structure data, the field has focused on computational means of function prediction. This review covers improved methods for the identification of potential functional sites in new protein structures and new methods for the discovery of similarity in functional sites. It also covers the significant progress that has been made in the prediction of protein–ligand interactions (docking).

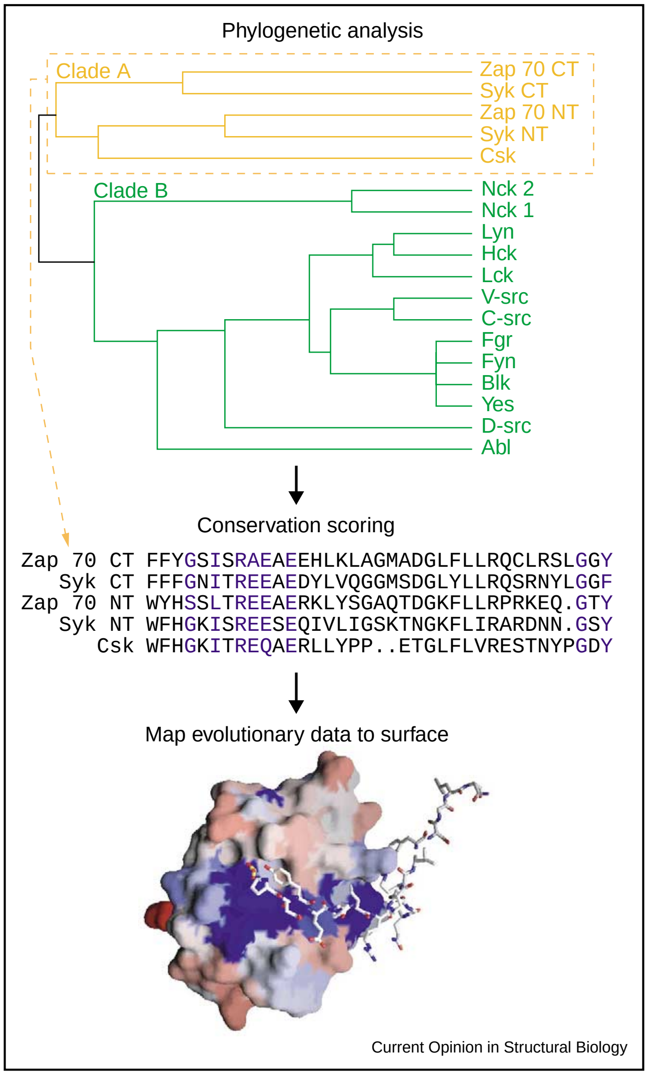
## Functional site location

A fundamental question to ask when considering an uncharacterised protein structure is the location of ligand-binding or active sites. In this section, we review methods for finding such sites in the absence of information about the ligand. If the ligand is already known, then alternative approaches are possible, including the exploitation of similarity to known binding sites and docking (see below).

Enzyme active sites commonly occur in large and deep clefts on the protein surface [1], and the need for significant favourable interactions between ligand and protein usually means that other small-molecule ligands also bind in surface depressions. Work related to the location of such clefts and depressions has been reviewed recently [2]. This review focuses on recent developments in functional site identification. Two major approaches have emerged: methods based on evolutionary information and methods that employ other criteria. The evolutionary method is potentially powerful because of the abundance of structural and sequence-related data for many protein families. Such methods [3–5] exploit the observation that important functional sites in proteins usually display a high level of conservation (see [6••] for a recent review). More recent methods are discussed below, some of which have been developed for interactions in protein–protein or other macromolecular systems, but are equally applicable to the discovery of small-molecule binding sites.

A common analysis method has been the mapping of evolutionary data to the three-dimensional surface of representative molecules [7–9]. Campbell and Jackson [10] divided the Src homology 2 (SH2) family into groups on the basis of binding site residue similarity. Subsequent conservation data are mapped to representative domains (see Figure 1) to investigate diversity between these groups within the ligand-binding region. Two other groups developed this technique and demonstrated the importance of taking into account the divergence levels expected with a particular phylogenetic depth, rather than simply considering the conservation against variability of residue positions. Pupko *et al.* [11••] describe a method to distinguish between residues that are conserved because of

Figure 1



```
Zap 70 CT FFYGSISRAEAEEHLKLAGMADGLFLLRQCLRSLGGY
   Syk CT FFFGNITREEAEDYLVQGGMSDGLYLLRQSRNYLGGF
Zap 70 NT WYHSSLTREEAERKLYSGAQTDGKFLLRPRKEQ.GTY
   Syk NT WFHGKISREESEQIVLIGSKTNGKFLIRARDNN.GSY
      Csk WFHGKITREQAERLLYPP..ETGLFLVRESTNYPGDY
```

Current Opinion in Structural Biology

Surface mapping of phylogenetic information. In this example, phylogenetic analysis has been used to divide a group of SH2 domains into two clades of similar sequences. The sequences from clade A have then been scored for conservation using a conservation scoring algorithm [9]. Conservation scores are produced for each residue position in the alignment and these values are mapped to the three-dimensional surface of the representative syk C-terminal SH2 domain structure. Red surface areas indicate low conservation, blue regions high conservation and white represents intermediate conservation. This method demonstrates the effective identification of the conserved phosphotyrosyl peptide binding site, in blue. (Based on a study by Campbell and Jackson [10].)

functional importance and those that appear to be conserved because of shortness of divergence time. They describe the shortcomings of the maximum parsimony approach, whereby conservation scores are assigned without taking branch lengths into account. The group found that functionally important regions often correspond to surface patches of slowly evolving residues and identified interaction regions in the Src SH2 domain. Blouin *et al.* [12] also use the maximum likelihood rate of evolution to qualify the level of evolutionary constraint on residue positions within a site. The technique allows the detection of functional divergence between subtrees.

In related work, Friedberg and Margalit [13•] propose a method for the identification of mutually persistently conserved (MPC) positions within pairs of sequence dissimilar and structurally similar protein families. MPC positions, which are positions that are conserved over long evolutionary times in both families, are found to correspond to key structural features, including catalytic residues and those that stabilise active sites. An interesting contrast is provided by the work of Kunin *et al.* [14], who compare multiple alignments of protein families that might or might not be structurally similar. Their cyclical relations consistency analysis is able to identify functional sequence motifs shared by proteins with different folds; for example, there are phosphate-binding regions that are shared by some Rossmann and TIM barrel folds.

Finally, two groups [15,16] have investigated the theory of co-evolution in protein–protein interactions. This might be applicable to protein–peptide (ligand) interactions. Goh and Cohen [15] developed a method for identifying binding partners for uncharacterised proteins and new binding partners for previously characterised proteins, whereas Bickel *et al.* [16] developed a statistical method whereby important sites are identified either by residue identity within and outside functional subfamilies, or by strong covariance between a pair of motif sites.

Although evolution is a powerful tool in many cases, there will always be a need for alternative approaches in cases where conservation does not indicate a well-defined site. Two groups have produced interesting and related approaches to the identification of key functional residues using electrostatic (Poisson–Boltzmann) calculations [17,18••]. Ondrechen *et al.* [17] observe that the theoretical titration curves of ionisable residues involved in catalysis are often anomalous, showing unexpected 'flat' regions and shifts in p$K_a$. Shifts in p$K_a$ are related to stability effects and Elcock [18••] employs the hypothesis that functionally important residues are often a thermodynamic disadvantage to a protein structure. Several experimental studies have shown that such residues can be mutated to give a more stable structure, and Elcock shows that calculation of the electrostatic component of the free energy also enables the identification of these destabilising and functionally important residues. It is likely that these approaches will be valuable in identifying potential functional sites, particularly if coupled with other evidence, for instance, from studies such as that of Bartlett *et al.* [19•], who showed that six amino acid residues (histidine, cysteine, aspartic acid, glutamic acid, arginine and lysine) account for >70% of all catalytic residues, and that they tend to occupy less flexible structural locations.

## Functional site similarity

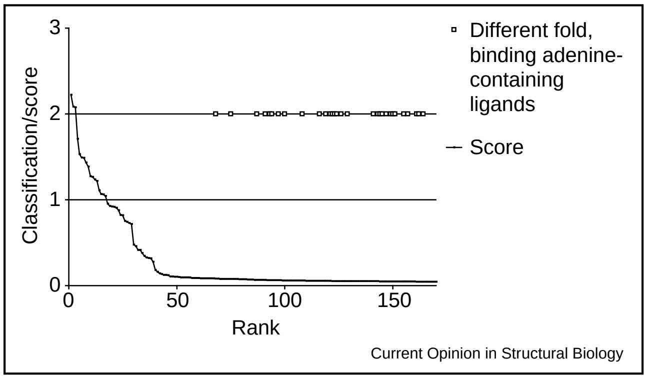
Prediction by similarity, for example predicting function using similarity at the sequence level, is a very strong theme in genome annotation, and recent years have seen much discussion of the precise nature of the relationship between similarity at the sequence, structural and functional levels [20,21]. Here we focus on similarity at the level of ligand-binding or functional sites. There are two motivations for this type of study. First, when proteins with different folds share aspects of function (e.g. ligand binding or catalysis), this can be reflected in site similarities independent of evolutionary homology. Second, even within evolutionary families, there are often functional differences between proteins that may be major (e.g. whether they are enzymes or nonenzymes) or minor (e.g. the specificity of different enzymes). Nagano *et al.* [22] describe the many functions associated with the ubiquitous TIM barrel fold. Functional differences might be more obvious from structural comparison of functional sites than they would be from comparison of sequences or overall tertiary structure. It is unsurprising, therefore, that databases of ligand-binding sites are starting to emerge that can, as in the case of LigBase [23], contain mappings of active site residues to structural alignments of protein family members.

There are two distinguishable approaches to functional site similarity. The first involves the creation of three-dimensional templates reflecting particular functions; an example is the continuing work of one group on three-dimensional 'fuzzy functional forms' to predict disulfide oxidoreductase activity, an approach that can be useful even with predicted structures [24]. Similar three-dimensional templates can also define the recognition properties of ligand-binding sites. Ligands containing adenine moieties have been the subject of considerable interest in the literature [25,26,27•,28••], probably because of the relatively rigid adenine ring structure and the ubiquitous nature of adenine in many key biological molecules (ATP, ADP, NAD, FAD, etc.) bound by nonhomologous proteins with different folds. Zhao *et al.* [28••] describe novel three-dimensional templates that define the consensus interaction energetics of the binding site volume in a diverse set of known adenine-binding sites. These templates out-perform traditional docking methods and energy calculations in distinguishing adenylate and guanylate recognition sites.

The above approach — describing the energetic properties of the binding site volume — is interesting, and contrasts with earlier work in which the focus has been more strongly on residues and chemical groups flanking the binding pocket and interacting with the ligand. Catalysis often involves universally conserved residues, but ligand binding often does not (there are many ways of binding the same ligand), so templates of this nature would seem to be a promising future direction for binding site analysis. Nevertheless, there have also been interesting developments in describing interactions with key protein atoms. Following a large volume of earlier work, which they review in detail, Rantanen *et al.* [27•] introduce a Bayesian approach using Gaussian mixtures to describe the distributions for particular protein atoms around ligand fragments. They demonstrate that their approach can narrow down the search for the positions and types of protein atoms interacting with particular ligand fragments, and consider that the main application of their methods is in the prediction of the nature of the ligand-binding pocket in the absence of protein structure information. However, these methods would be equally applicable to the analysis of potential sites in new and uncharacterised protein structures.

The second approach to functional site similarity does not involve the creation of three-dimensional templates for particular ligands or activities. Methods have been developed by Schmitt *et al.* [29••] and Kinoshita *et al.* [30] that provide similarity searches over functional site databases. These use related clique detection algorithms to find similarity, but employ different representations of the site. Schmitt *et al.* use chemical groups able to make hydrogen bonds and/or aliphatic interactions with ligands, whereas Kinoshita *et al.* represent the sites as surfaces with electrostatic and hydrophobic characteristics. Both methods are able to detect functional similarity in the absence of homology or fold similarity. As an example, Figure 2 shows some results from a similar approach based on differing site descriptions that was developed in our own laboratory (N Gold, D Westhead, unpublished data). Here, the query site binds the cofactor NAD; similarity is found to other NAD-binding sites and then,

**Figure 2**



The 160 highest-scoring results from a site similarity search based on an NAD-binding site in alcohol dehydrogenase. The highest-ranking hits come from the same fold and superfamily, but at lower ranks sites from other folds binding related adenine-containing ligands start to emerge. All the top 160 hits bind a molecule containing an adenine moiety.

at lower ranks, to binding sites for adenine-containing ligands in proteins with unrelated folds. Similarity searches of this nature are usually more demanding in terms of computational time than matching to three-dimensional templates, but they might provide useful information for sites not covered by templates and could lead to the discovery of new and interesting relationships and templates.

## Predicting ligand interactions by molecular docking

The increasing availability of protein three-dimensional structures coupled with continuing advances in docking and scoring methods have established docking as an important tool for small-molecule lead discovery. Two reviews on the subject have recently been published, covering many of the docking methods [31] and their importance for lead generation in structure-based drug design [32]. The number of applications to lead discovery has increased dramatically in recent years; however, these studies will not be covered here. Although there are several well-established docking methods that continue to undergo further development [32,33], new approaches to protein–ligand docking continue to be proposed that focus on efficient search criteria and models for solvation and protein flexibility. We also discuss scoring functions and proposed data sets for use in their validation.

### Docking methods

In most cases, a functional site such as an enzyme active site is already known and the practitioner either performs docking of a known substrate/inhibitor or an *in silico* screening of a ligand database. However, with the advent of structural genomics and the prospect of large numbers of structures with unknown functional properties, the ability to perform docking to validate functional sites will become important. This *ab initio* prediction problem has received little attention to date and new applications in this area are of particular interest. Hetenyi and van der Spoel [34] apply the well-known AutoDock program to try to reproduce experimental structures of several protein–peptide complexes without prior knowledge of the binding site. Their results are very encouraging, given that the whole protein constitutes a large search space. However, computational times (tens of hours) per ligand are a serious limiting factor for multiple ligands. Glick *et al.* [35•] address this problem by using a multiscale ligand representation in which a ligand is first docked using a very small number of feature points. These represent a more complex ensemble of rigid ligand conformations. Results for finding binding sites and a final close root mean square deviation solution are encouraging for the seven ligands featured and can be achieved at reasonable computational cost (minutes) per ligand.

One objective of docking programs is to target the biologically active conformations quickly, therefore limiting the relatively large amount of search space. In the program Q-fit [36•], a probabilistic method ranks receptor binding modes so that those with the lowest energy are sampled first in the docking procedure. Limits can be placed on the search depth, therefore restricting sampling to low energy conformation space. Alternatively, search algorithms such as the ‘Mining Minima’ optimiser combine several methodologies; this algorithm has undergone recent enhancement [37]. EUDOC [38] uses a more conventional systematic search method followed by focused finer searches. It was applied to a large ligand test set with good results.

The importance of protein flexibility is increasingly being recognised as fundamental to the wider applicability of docking methods [39]. Fradera *et al.* [40] analyse ligand-induced changes in protein binding sites. Using a fixed receptor structure can impose considerable limitations if the protein undergoes an induced fit on ligand binding, as is often the case. However, more sophisticated protein models come at the expense of increased computer time and, on a practical level, the results are often worse [32]. Docking with full protein flexibility is currently not feasible for a large number of ligands and therefore some level of approximation must be introduced. Kua *et al.* [41] apply ligand docking to different static conformations of the protein acetylcholinesterase taken from snapshots of a molecular dynamics trajectory. The docking energies correlate well with experimental binding affinities for a series of substrate and inhibitor analogues. This method would be computationally expensive for a large ligand data set. Alternatively, methods that combine several structures simultaneously (taken from experiments, molecular modelling or simulation) to produce an ensemble representation have recently been developed. In the program FlexE [42], discrete alternative conformations are explicitly taken into account. These can be combinatorially joined to create new protein structures. Osterberg *et al.* [33] generate a single representative grid of interaction energies. Using 21 HIV-1 protease structures as a test case, they show that the type of flexibility can pose problems. However, an energy-weighted average of the grids performs well for redocking of most of the ligands.

Given the increasing availability of three-dimensional structures arising from models built by comparative (homology) modelling, there is a need for docking methods to be able to handle lower-quality structural information. Often these models are not sufficiently accurate to apply conventional docking methods. Schafferhans and Klebe [43•] have developed a method to do this that integrates a ligand three-dimensional quantitative structure–activity relationship (QSAR) model and a model-built representation of the protein binding site using soft potentials.

Distance restraints can also help to constrain the docking search space. Hindle *et al.* [44•] have developed a docking methodology that allows incorporation of pharmacophore-type constraints. These guide the docking procedure to generate solutions that obey these constraints, allowing pharmacophore-based or experimental filters to be used. In [45], intraligand exchange transfer nuclear Overhauser effect data are used to restrict the conformation of a peptide ligand in docking. This method provides intramolecular proton distances for the peptide but no information on intermolecular contacts, and therefore docking of the protein–peptide complex is a useful additional tool.

### Scoring functions

The development of scoring functions continues to be a subject of considerable study within the docking community [46]. Clearly, the ability to develop a universally applicable function would greatly enhance the docking methodology. The functions can be broadly classified according to whether they are empirical or knowledge-based in origin; functions in the former category are based on a parameterised force field model (e.g. including van der Waals, electrostatic, entropy and solvation terms), whereas functions in the latter category are based on a statistical analysis of observer contacts in the structural database.

Wang *et al.* [47] report the development of a new consensus empirical scoring function for binding affinity prediction. In a particularly interesting development [48••], the knowledge-based potential DrugScore is tailored to a particular protein by using known ligand-binding affinities to reparameterise the interaction grid used in scoring. This adaptive method has considerable future potential. Increasingly, drug targets will already have ligand binding data available and increased amounts of data will facilitate better predictive models.

Both empirical and knowledge-based functions have their advantages and limitations, and several comparative studies have been reported recently. Stahl and Rarey [49] compare the empirical FlexX and PLP score functions and the knowledge-based PMF and DrugScore functions. Also, Perez and Ortiz [50] compare the AMBER force field against the knowledge-based PMF function, and Sotriffer *et al.* [51] compare Autodock's largely AMBER-based score function against their scoring function, DrugScore. The idea of using a consensus score from different methods has been investigated [52] and found to outperform the individual docking methods for predicting the top-ranked cluster.

### Validating methods

Comparison between different docking methods and scoring functions is difficult. Only general conclusions can be made when comparing two different studies. Standardising data sets and computer hardware (for run time comparison) would greatly help in making comparisons; however, this is clearly difficult to achieve without a coordinated effort. Recently, a large validation data set has been proposed [53]. This consists of 305 complexes with protonation states assigned by manual inspection. Care has been taken to remove unsuitable entries. Also, Roche *et al.* [54] have proposed a ligand–protein database that combines structural data with experimental binding data. In addition, they have generated sets of 'decoys' that could prove useful for testing new scoring functions.

## Conclusions

We have reviewed recent computational advances in understanding ligand binding. Protein functional site detection and characterisation is a major goal of structural bioinformatics and its application to other protein–biomolecule interactions is reviewed elsewhere in this issue. Several powerful techniques have emerged in recent years that will hopefully be of use both in detecting functional sites in new structures solved before any knowledge of function and in aiding the rational design of new pharmaceuticals. Mapping evolutionary data to the three-dimensional surface of proteins has emerged as a promising technique for many protein families; however, it is no panacea for all functional interactions. Similarly, the evidence is that structural (or feature) similarity in the binding sites of proteins will yield meaningful information only in certain cases; nevertheless, this could prove critical in the detection of new functional sites and is invaluable in understanding ligand recognition, including issues of cross-reactivity and toxicity. The possibility that protein docking methods can also be used for site detection and characterisation is an important new application of an older technique. Given the likely future growth in protein structure information, approaches such as these will become increasingly important tools for the functional characterisation of ligand-binding sites and for structure-based drug design.

## References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM: **Protein clefts in molecular recognition and function**. *Protein Sci* 1996, **5**:2438-2452.

2. Sotriffer C, Klebe G: **Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design**. *Farmaco* 2002, **57**:243-251.

3. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins**. *Nat Struct Biol* 1995, **2**:171-178.

4. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments**. *J Mol Biol* 2000, **303**:61-76.

5. Landgraf R, Xenarios I, Eisenberg D: **Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins**. *J Mol Biol* 2001, **307**:1487-1502.

6. Lichtarge O, Sowa ME: **Evolutionary predictions of binding surfaces and interactions**. *Curr Opin Struct Biol* 2002, **12**:21-27.
•• A thorough review of evolutionary predictions of binding surfaces and interactions, covering early and recent methods in the field. The section

on evolutionary predictions in the present review follows on from the scope of this article.

7. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families**. *J Mol Biol* 1996, **257**:342-358.

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

9. Valdar WS, Thornton JM: **Protein–protein interfaces: analysis of amino acid conservation in homodimers**. *Proteins* 2001, **42**:108-124.

10. Campbell SJ, Jackson RM: **Diversity in the SH2 domain family phosphotyrosyl peptide binding site**. *Protein Eng* 2003, **16**:217-227.

11. •• Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N: **Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues**. *Bioinformatics* 2002, **18**:S71-S77.

The authors address the issue of branching distance in the mapping of phylogenetic information to three-dimensional surfaces of proteins. Rate4Site, a tool developed by the group, uses the maximum likelihood method to compute amino acid replacement probabilities for each branch in a phylogenetic tree. The rate of evolution is then mapped onto the surface of a representative structure.

12. Blouin C, Boucher Y, Roger AJ: **Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information**. *Nucleic Acids Res* 2003, **31**:790-797.

13. • Friedberg I, Margalit H: **Persistently conserved positions in structurally similar, sequence dissimilar proteins: roles in preserving protein fold and function**. *Protein Sci* 2002, **11**:350-360.

Pairs of protein families similar in structure and dissimilar in sequence are studied. Within each family, persistently conserved positions are defined as residue positions remaining conserved over several PSI-BLAST iterations (or long evolutionary times). MPC positions are those structurally aligned positions that are persistently conserved in both families. MPC positions are shown to occur frequently in key structural and functional locations, such as active sites, areas important for active site stability and the termini of helices.

14. Kunin V, Chan B, Sitbon E, Lithwick G, Pietrokovski S: **Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs**. *J Mol Biol* 2001, **307**:939-949.

15. Goh CS, Cohen FE: **Co-evolutionary analysis reveals insights into protein–protein interactions**. *J Mol Biol* 2002, **324**:177-192.

16. Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN: **Inaugural article: finding important sites in protein sequences**. *Proc Natl Acad Sci USA* 2002, **99**:14764-14771.

17. Ondrechen MJ, Clifton JG, Ringe D: **THEMATICS: a simple computational predictor of enzyme function from structure**. *Proc Natl Acad Sci USA* 2001, **98**:12473-12478.

18. •• Elcock AH: **Prediction of functionally important residues based solely on the computed energetics of protein structure**. *J Mol Biol* 2001, **312**:885-896.

The author presents a method for identifying functional residues on the basis of electrostatics. It is applied to three uncharacterised structures from a structural genomics initiative, and identifies charged residues in areas of the protein surface also showing other characteristics of functional sites, including exposed hydrophobic residues and, in one case, a cluster of four cysteines that could potentially bind a metal ion. In a larger set of proteins, residues showing strong evolutionary conservation are found among either the most stabilising residues in the structure (and are therefore probably important for structural stability) or the most destabilising (and are therefore probably important for functional reasons).

19. • Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites**. *J Mol Biol* 2002, **324**:105-121.

A thorough study of the properties of catalytic residues in protein structures. Six amino acid residues (histidine, cysteine, aspartic acid, glutamic acid, arginine and lysine) are shown to account for 70% of all catalytic residues. These tend to be strongly conserved, have low crystallographic B-factors (indicating low structural flexibility) and make correspondingly high numbers of hydrogen bonds. Interestingly, glycine accounts for a large proportion of residues involved in catalysis using mainchain atoms.

20. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**:1113-1143.

21. Devos D, Valencia A: **Practical limits of function prediction**. *Proteins* 2000, **41**:98-107.

22. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions**. *J Mol Biol* 2002, **321**:741-765.

23. Stuart AC, Ilyin VA, Sali A: **LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures**. *Bioinformatics* 2002, **18**:200-201.

24. Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, Fetrow JS: **Enhanced functional annotation of protein sequences via the use of structural descriptors**. *J Struct Biol* 2001, **134**:232-245.

25. Denessiouk KA, Johnson MS: **When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families**. *Proteins* 2000, **38**:310-326.

26. Denessiouk KA, Rantanen VV, Johnson MS: **Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins**. *Proteins* 2001, **44**:282-291.

27. • Rantanen VV, Denessiouk KA, Gyllenberg M, Koski T, Johnson MS: **A fragment library based on Gaussian mixtures predicting favorable molecular interactions**. *J Mol Biol* 2001, **313**:197-214.

A new approach to the distribution of protein atom types around defined ligand fragments using a Bayesian approach with Gaussian mixture models. Maximum likelihood parameters of the models are estimated using an expectation maximisation algorithm. The method gives predictions of which protein atoms interact with particular ligand fragments (example protein atoms would be 'mainchain carbonyl oxygen' or 'sidechain hydroxyl oxygen). A detailed analysis of the errors made in predictions shows that the method can effectively narrow down the set of possible interactions for ligand functional groups.

28. •• Zhao S, Morris GM, Olson AJ, Goodsell DS: **Recognition templates for predicting adenylate-binding sites in proteins**. *J Mol Biol* 2001, **314**:1245-1255.

An approach to the development of three-dimensional templates for binding site recognition. In contrast to other approaches involving the residues and chemical groups flanking the site, these templates describe the energetic properties of the ligand-binding volume by employing grid-based affinity potentials. Consensus features of these potentials are extracted from diverse structures binding the same ligand. The templates are better able to distinguish adenylate- and guanylate-binding sites than traditional energy calculation or ligand docking.

29. •• Schmitt S, Kuhn D, Klebe G: **A new method to detect related function among proteins independent of sequence and fold homology**. *J Mol Biol* 2002, **323**:387-406.

A database of three-dimensional cavities, extracted from the Protein Data Bank, is employed. These are described in terms of the physico-chemical characteristics (e.g. hydrogen-bond donor/acceptor, whether aliphatic) of the chemical groups flanking the cavity. A clique-detection-based similarity search algorithm is shown to be able to detect and rank database entries with similar function (binding the same or similar ligands, or with similar catalytic mechanisms) for several example query cavities, independently of fold homology. The method can also be used as an idea generator for *de novo* drug design.

30. Kinoshita K, Furui J, Nakamura H: **Identification of protein functions from a molecular surface database, eF-site**. *J Struct Funct Genomics* 2001, **2**:9-22.

31. Taylor RD, Jewsbury PJ, Essex JW: **A review of protein–small-molecule docking methods**. *J Comput Aided Mol Des* 2002, **16**:151-166.

32. Abagyan R, Totrov M: **High-throughput docking for lead generation**. *Curr Opin Chem Biol* 2001, **5**:375-382.

33. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS: **Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock**. *Proteins* 2002, **46**:34-40.

34. Hetenyi C, van der Spoel D: **Efficient docking of peptides to proteins without prior knowledge of the binding site**. *Protein Sci* 2002, **11**:1729-1737.

35. Glick M, Grant GH, Richards WG: **Docking of flexible molecules**
• **using multiscale ligand representations**. *J Med Chem* 2002, **45**:4639-4646.

The number of feature points are progressively increased during the simulation and different conformations clustered to remove redundancy at a given level of feature points. This produces a hierarchy of models for the ligand, which are clustered and then pruned at each successive level, ultimately producing an all-atom ligand representation.

36. Jackson RM: **Q-fit: a probabilistic method for docking**
• **molecular fragments by sampling low energy conformational space**. *J Comput Aided Mol Des* 2002, **16**:43-57.

The receptor binding site is 'probed' before docking. Receptor binding modes are defined in three dimensions in terms of a triplet of receptor probes that potentially correspond to atom triplets in the ligand. The ranking of these triplets by energy allows an ordered search of binding modes, so that those with the lowest energy are sampled first.

37. Kairys V, Gilson MK: **Enhanced docking with the mining minima optimizer: acceleration and side-chain flexibility**. *J Comput Chem* 2002, **23**:1656-1670.

38. Pang YP, Perola E, Xu K, Prendergast FG: **EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases**. *J Comput Chem* 2001, **22**:1750-1771.

39. Ma B, Shatsky M, Wolfson HJ, Nussinov R: **Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations**. *Protein Sci* 2002, **11**:184-197.

40. Fradera X, De La Cruz X, Silva CH, Gelpi JL, Luque FJ, Orozco M: **Ligand-induced changes in the binding sites of proteins**. *Bioinformatics* 2002, **18**:939-948.

41. Kua J, Zhang Y, McCammon JA: **Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach**. *J Am Chem Soc* 2002, **124**:8260-8267.

42. Claussen H, Buning C, Rarey M, Lengauer T: **FlexE: efficient molecular docking considering protein structure variations**. *J Mol Biol* 2001, **308**:377-395.

43. Schafferhans A, Klebe G: **Docking ligands onto binding site**
• **representations derived from proteins built by homology modelling**. *J Mol Biol* 2001, **307**:407-427.

Ligands are aligned relative to each other to create a three-dimensional QSAR model, which is then aligned with the model protein binding site. Soft (Gaussian) functions are used for both representations and then the alignment optimises the overlap between them. The aim is to ultimately use this information to constrain the model-built structure in the modelling step via feedback between the protein and the QSAR models.

44. Hindle SA, Rarey M, Buning C, Lengaue T: **Flexible docking under**
• **pharmacophore-type constraints**. *J Comput Aided Mol Des* 2002, **16**:129-149.

The method applies a series of look-ahead checks to see if a currently generated solution obeys the pharmacophore constraints and removes those that do not. This can speed up calculation times.

45. Zabell AP, Post CB: **Docking multiple conformations of a flexible ligand into a protein binding site using NMR restraints**. *Proteins* 2002, **46**:295-307.

46. Gohlke H, Klebe G: **Statistical potentials and scoring functions applied to protein–ligand binding**. *Curr Opin Struct Biol* 2001, **11**:231-235.

47. Wang R, Lai L, Wang S: **Further development and validation of empirical scoring functions for structure-based binding affinity prediction**. *J Comput Aided Mol Des* 2002, **16**:11-26.

48. Gohlke H, Klebe G: **DrugScore meets CoMFA: adaptation of**
•• **fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein**. *J Med Chem* 2002, **45**:4153-4170.

Experimental protein–ligand binding affinities are used as a training set to adapt the DrugScore function to a particular protein. This tailor-made scoring function improves the predictive power for affinity prediction in the two proteins studied.

49. Stahl M, Rarey M: **Detailed analysis of scoring functions for virtual screening**. *J Med Chem* 2001, **44**:1035-1042.

50. Perez C, Ortiz AR: **Evaluation of docking functions for protein–ligand docking**. *J Med Chem* 2001, **44**:3768-3785.

51. Sotriffer CA, Gohlke H, Klebe G: **Docking into knowledge-based potential fields: a comparative evaluation of DrugScore**. *J Med Chem* 2002, **45**:1967-1970.

52. Paul N, Rognan D: **ConsDock: A new program for the consensus analysis of protein–ligand interactions**. *Proteins* 2002, **47**:521-533.

53. Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R: **A new test set for validating predictions of protein–ligand interaction**. *Proteins* 2002, **49**:457-471.

54. Roche O, Kiyama R, Brooks CL III: **Ligand–protein database: linking protein–ligand complex structures to binding data**. *J Med Chem* 2001, **44**:3592-3598.