

Computational Prediction of Functionally Important Regions in Proteins

Florencio Pazos^{*1} and Jung-Wook Bang²

¹Protein Design Group, National Center for Biotechnology (CNB-CSIC), Madrid, Spain

²Department of Computing, Imperial College, London, UK

Abstract: Current projects for the massive characterization of proteomes are generating protein sequences and, to less extent, three dimensional structures with unknown function. Experimentally determining functional features of a protein is expensive, time consuming and difficult to automate. There is therefore a demand for computational methods for predicting protein functional features, which can be coupled to the pipelines of genome sequencing and structure determination. This review focuses on current *in-silico* methods for predicting regions in proteins with some functional importance (catalytic sites, binding sites, protein interaction regions, etc.) using sequence and/or three-dimensional structure information.

Keywords: Functional site, active site, binding site, protein function, protein interaction.

INTRODUCTION

The main characteristic of the so-called “post-genomic” era is the need for methods and tools for analyzing the huge amount of data (mainly in the form of biological polymer sequences) produced in the genomic era. An extreme reductionist approach to Biology led to the idea that most (if not all) biological knowledge could be obtained from the sequences of biomolecules. Nevertheless, the massive determination of genomic sequences faced with a lack of methods for obtaining biological information from these sequences. Genome sequencing projects continue to produce the genetic repertoire of many organisms, leading to more than 240 fully sequenced genomes available today. “Environmental sequencing”, the organism-independent sequencing of whole ecosystems [1], is a recent phenomenon which is boosting the amount of available sequences. To a lesser extent, structural genomics projects [2] are also producing an increasing number of protein three-dimensional (3D) structures for which functional information is not available (more than 500 right now). In the past, the determination of a protein 3D structure was a latter step after an intensive characterization of the protein (including functional characterization). The 3D structure was basically used to map the previously-known functional features, to give them an “spatial” interpretation, and to design ways of inhibiting or changing the function. But this paradigm is changing now with the massive determination of protein 3D structures previous to any knowledge of their function. Nevertheless, the rate of structure determination is still very low compared with sequence determination, which is reflected in the very different orders of magnitude of the number of known 3D structures ($\sim 10^4$) compared with the number of known sequences ($\sim 10^6$).

All these factors are leading to an increasing number of sequences and structures for which functional information is

not available. Determining which residues in a protein are responsible for its function is very important in order to understand the molecular mechanism of this function, to modify it in our benefit (biotechnology) or to correct problems related with this function (e.g. pathologies due to mutations). The experimental characterization of function and functional features (functional sites, etc.) is very expensive, time consuming, and difficult to automate. Interestingly, some attempts to characterize at least very basic functional features (like dispensability or phenotypic effect) are being carried out in a genome-wide fashion [3,4]. This difficulty in the experimental determination of functional features is pushing the development of computational techniques for this purpose.

This review tries to summarize the landscape of current methods and programs for the prediction of functionally important regions in proteins from sequence and/or 3D structure information. It is difficult to give a comprehensive and accurate definition of “functionally important residue” in the same way it is difficult to define “protein function”, although everyone seems to have intuitive ideas of what these concepts are. A possible definition is that functional residues are those required for the protein to perform its molecular function or biological role, in the sense that these residues can not be freely changed (except to some compatible aminoacids) without directly affecting the function of the protein. “Directly” means that the effect of the mutation in the function should be direct and not due to an effect in the 3D structure (i.e. a mutation in the structural core which avoid the protein to fold and hence to perform its function). Functionally important regions are considered here in a broad sense, including active/catalytic sites, protein binding sites, small ligand binding sites, nucleic acids binding sites, etc. Nevertheless, some important but very specific functional features have been deliberately skipped, such as disulphide bridges or glycosylation sites. Methods for predicting the molecular or cellular function of whole protein chains (“annotation”) are covered in other excellent reviews [5-8]. The review consists of two main sections which describe the sequence-based methods and the structure

*Address correspondence to this author at the Protein Design Group, National Center for Biotechnology (CNB-CSIC), Campus Universidad Autónoma, Cantoblanco, 28049 Madrid, Spain; Tel: +34.915854669; Fax: +34.91.5854506; E-mail: pazos@cnb.uam.es

Table 1. Some Public Web Servers for Locating Functionally Important Sites in Proteins

Server name	Description	URL	Ref.
TraceSuite II	Implementation of the Evolutionary Trace algorithm. Prediction of functional residues from sequence	http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html	[29]
Consurf server	Consurf method. Prediction of functional residues from sequence and 3D structure	http://consurf.tau.ac.il/	[61]
Conseq server	Conseq method. Prediction of functional residues from sequence	http://conseq.bioinfo.tau.ac.il/	[16]
Treedet server	Implementation of various methods including MTreedet and S-method. Prediction of functional residues from sequence	http://pdg.cnb.uam.es/treedet/	[28]
ISPRED	Prediction of protein interaction sites from sequence and 3D structure	http://gpcr.biocomp.unibo.it/cgi/predictors/pp/pred_pp.cgi	[65]
PROCAT server	Location of possible enzyme active sites in 3D structures by matching against a database of templates	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html	[71]
PINTS server	Scanning of 3D structures against a database of 3D motifs or a 3D motif against 3D structures	http://www.russell.embl.de/pints/	[73]
Disopred	Prediction of disordered regions in proteins from sequence	http://bioinf.cs.ucl.ac.uk/disopred/	[56]
InterPro	Centralized repository of protein information. It contains information on protein functional sites (if available) and information on possible domains and motifs	http://www.ebi.ac.uk/interpro/	[21]

The first column contains the name of the server. The second column is a short description including the methods implemented on it and the type of input it requires (sequence and/or 3D structure). See the text for a full description of the methods. The third column contains the URL to access the server, and the last column the bibliographic references to the server and/or the methods implemented.

based-methods respectively. This distinction is useful from the user point of view, since one type of methods or the other can be chosen depending on the available information for the target protein. We also include a list of web servers which implement some of the methods for the prediction of functionally important sites discussed in the text, indicating the type of input (sequence and/or structure) they require (Table 1).

PREDICTION OF FUNCTIONALLY IMPORTANT REGIONS FROM SEQUENCE

The number of known sequences is orders of magnitude higher than the number of known 3D structures. This justifies the development of methods for the prediction of protein important regions based on sequence-information alone, despite being less accurate than the ones which take into account structural information as well, which are discussed in the next section. Some methods take a single sequence as input, although most of them use evolutionary information in the form of multiple sequence alignments (MSAs). We will use the following nomenclature herein: aminoacid: each one of the 20 natural aminoacids; residue: an aminoacid within a protein; position: a column within a multiple sequence alignment relating the equivalent residues of the aligned proteins.

Conserved Regions

Homologous proteins, those sharing a common ancestor, can be grouped together and their sequences aligned to compare equivalent residues. These MSAs are rich sources of structural and functional information [9] (Fig. 1a) since they show the aminoacid changes allowed by the evolution at each position due to structural or functional requirements. Fully conserved positions in MSAs are interpreted as important residues for the structure and function of the

protein, since no changes have been allowed on those positions during the evolutionary process. These positions were the first indicators of functionality [10] and they are related with all types of functional sites: active sites, ligand-binding and protein-protein interaction sites [11], nucleic acid binding sites, etc. Not all the conserved positions are related to function but many are conserved due to structural requirements (forming the structural core of the protein). These can be distinguished, to some extent, by the aminoacid which is conserved: some aminoacids tend to have structural roles when conserved (e.g. Trp, Leu, Gly, Cys) while others tend to be part of binding and active sites (mainly apolar aminoacids, or specific types e.g. Asp, Ser, Cys, His) [12,13]. Although the problem of locating conserved positions could look trivial at first sight, there is not a unique method, indeed there are many different approaches which produce different results [14]. For example, a position with Arg's, Lys's and His's would be reported as variable by a method based on entropy, whereas a method which takes into account conservative substitutions would report it as quite conserved. Some methods incorporate complex models for evaluating the evolutionary conservation of positions taking into account the phylogeny of the sequences in the MSA [15-17], so as to avoid artifacts due to the peculiarities or uneven distribution of sequences in the alignment (i.e. highly similar sequences resulting in most of the positions being conserved).

A functional motif could be composed not only by fully conserved positions but also by others showing distinct aminoacid distributions. Sequence profiles extracted from MSAs [18] are able to capture not only sequence conservation but other aminoacid distributions within positions. In the simplest form of a profile, a position is encoded by a vector of 20 components representing the fraction of each one of the 20 aminoacids (fully conserved

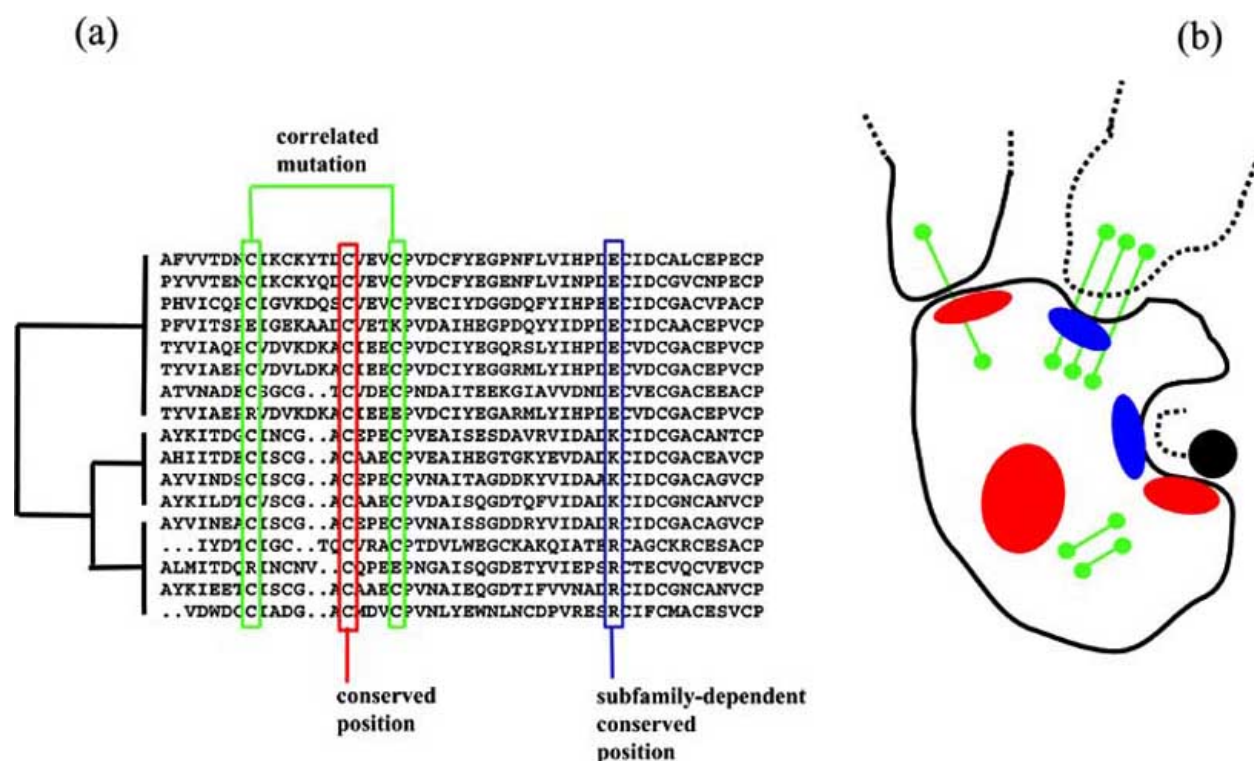


Fig. (1). Information extracted from multiple sequence alignments (MSAs) related with protein structure and function. a) fully conserved, family-dependent conserved positions and pairs of positions showing a correlated behavior are exemplified in a MSA. b) An ideal model illustrating the relationships between these positions and functional and structural features. Conserved positions (red) are in the structural core of the protein and in the active sites. Family dependent conserved positions (blue) are also present in the active site conferring specificity. For example, in a family of proteins binding the same class of substrates but with slight differences (i.e. different length of a hydrocarbon chain), conserved positions would bind the conserved part of the substrate, whereas family-dependent conserved positions would bind the variable part. These two types of positions also map in protein interaction sites, depending on whether the whole family interacts with the same partner or each subgroup specifically interact with different partners. Pairs of correlated positions (green) have been related to residue closeness. In the case of inter-protein correlations, these pairs are many times pointing to the interaction surface but not directly on it.

positions would have “1.0” in the corresponding position and “0.0” in the rest.) Several databases of motifs and domains have been developed aligning related proteins and extracting the evolutionary conserved parts or motifs, either automatically or assisted by expert knowledge. The two main exponents are *Prosite* [19], which contains short sequence motifs related to function, and *Pfam* [20], which encloses the sequence characteristics of protein families in Hidden Markov Models (HMMs). Both systems can be used to predict the function and to locate the functional sites for a new sequence by matching it against the libraries of *Prosite* motifs or *Pfam* HMMs. These and other similar methods are now available in the centralized *InterPro* resource [21].

Family-Dependent Conserved Regions

Another kind of position shows a more subtle pattern of conservation. The position is clearly conserved but the aminoacid type is different within different subgroups of proteins in the MSA (Fig. 1a). These subgroups can be defined according to different criteria (phylogenetically, phenotypically, functionally, etc.). The fact that these positions are conserved is indicative of their functional

importance, whereas the fact that the aminoacid type is different for different subfamilies indicates that this importance is “subfamily-specific”, that is, these positions are important for the feature used for defining the subfamilies. If subfamilies are defined according to functional criteria, these positions are related with functional specificity. A general model for illustrating the relationship between fully conserved and subfamily-specific positions is shown in Fig. 1b. Conserved positions are present in structural cores (due to structural reasons) and active sites. Family-specific positions are also in active sites close to conserved positions (i.e. conferring specificity for substrates with slightly different characteristics) and in other parts of the protein related with specificity, like protein-protein interaction sites (reflecting the interaction with different partners). Since the aminoacid type is different in the different subfamilies (some times remarkably different, i.e. Arg vs. Asp), programs for the automatic detection of conservation fail to detect these positions in many cases. It is also important to take into account that these positions are not fully captured in profiles (previous section) since these only consider the fraction of each one of the 20 residues in

the position, regardless of the protein each residue belongs to. (In other words, in the simplest form of profiles, shuffling the residues within a position would produce identical profiles.) That is why a new set of techniques have been developed to deal with this problem. One of the first programs for the detection of these family-specific conserved positions was *SequenceSpace* [22]. This elegant program is based on a vectorial representation of the MSA in a high dimensional space ($20 \times L$ dimensions, being L the length of the MSA). A reduction of this space by principal component analysis (PCA) produces a low dimensional space preserving most of the information of the original one in which proteins with similar sequences cluster together in similar regions of the space. A similar vectorial treatment for the individual residues (instead of full-length proteins) produces an equivalent space where the residues with most of the information for explaining the separation of the subfamilies go to the same regions of the space where these subfamilies are. This approach has been successfully applied for locating regions related with functional specificity and differential recognition of interacting partners in many protein families [23-26]. In some cases, experimental mutations of the positions detected by this approach as responsible of specificity produced changes in the interaction specificity [27]. Since the original *SequenceSpace* approach requires a manual step in which the user interactively examines the protein and residue spaces in order to locate these family-specific positions, several improvements emerged trying to mechanize this process by automatically clustering proteins and residues in the corresponding spaces and looking for the equivalences [28]. Another manner of locating these positions is the *Evolutionary Trace* (ET) methodology. The basic ET method [29] is based on successive hierarchical partitions of a MSA in different sets of subfamilies following the corresponding phylogenetic tree from the root to the leaves. A "rank" is assigned to each position which represents the partition in which it becomes conserved within subfamilies. A fully conserved positions would have rank=1 since it is conserved in the root partition of the tree (only one subfamily). A position conserved within the two main subfamilies of the tree (second partition) would have rank=2 and so on. In the last partition of the tree (one "subfamily" per protein) every position is conserved (lowest rank). Positions with highest ranks are predicted as functionally important residues, and it has been demonstrated for many protein families that they are related with functional/binding sites [30-33]. The basic ET method, which represents conservation in a qualitative way, has been recently improved incorporating entropy as a quantitative measure of conservation in the scoring and ranking of the positions [34]. Another method, developed by Hannehalli & Russell allows the user to impose a subfamily definition (i.e. according with functional criteria), instead of using the one implicit in the phylogeny of the MSA [35]. The *S-method* [28] explores successive partitions of a phylogenetic tree from the root to the leaves in a way similar to ET, but in this case only one partition is used to obtain the positions conserved within the resulting subgroups. As mentioned before, the probability of finding this kind of positions by chance increases as we divide the tree closer to the leaves. The concept of Relative Entropy is used in the *S-method* to normalize the number of family-specific conserved positions

by the number of conserved positions in each subfamily. The partition which produces the highest relative entropy is taken as the "best" one, and the corresponding family-dependent conserved positions reported. The *MTreedet* method [28] looks for positions in MSAs whose mutational behavior resembles the one of the whole alignment, since this is the expected conduct for positions responsible for subfamily separation (Fig. 1a). The mutational behavior of a position is represented by a matrix containing the similarities for all pairs of aminoacids at that position, whereas the mutational behavior of the whole alignment is represented by an equivalent matrix containing the overall similarities for the corresponding pairs of proteins. These matrices are compared with a correlation criteria which produces a score for every position in the MSA. Positions with highest scores are taken as the predicted functional sites. (Fig. 2) shows examples of these correlation patterns for a functional and a "non-functional" position in ferredoxin. Other approaches are also based in this idea that functionally important regions preserve the overall phylogeny of the whole family [36,37]. There are many other approaches for handling this problem of locating subfamily-specific positions [38-42].

We can also include in this section of sequence-based approaches variations of methods for predicting functional sites using sequence and structure information (next section) in which structural features are substituted by predicted structural features. For example, the *ConSeq* server uses sequence conservation and predicted sequence accessibility to locate functional sites [16].

The methods described in this section are not specific in the type of functional sites they detect (protein binding, DNA-binding, catalytic site, ...) since the assumption behind (conservation due to importance) equally applies to all of them.

Protein Interaction Sites

Beside the general methods described above for the sequence-based prediction of functionally important residues, there are some specific methods for locating protein-protein binding sites. Ofrañ and Rost have developed a neural network able to distinguish between interacting and non-interacting residues which takes as input a single sequence [43]. "Correlated mutations" are another type of signal that can be extracted from MSAs related with binding sites (Fig. 1). Correlated changes observed in MSAs can be interpreted as compensatory mutations (a mutation in a position is compensated by another mutation in a position spatially close to the first one). A weak relationship between this mutational behavior and actual spatial closeness has been found [44,45]. The relationship between compensatory changes and protein interactions has been experimentally studied [46]. There is a relationship between inter-protein correlated mutations and interacting surfaces, which allows to use these correlations as constraints to select between different docking solutions, or to predict interaction regions from sequence information alone [47].

Disordered Regions

An increasingly important concept, which breaks the paradigm of molecular biology stating that 3D structure determines function, are the intrinsically unstructured

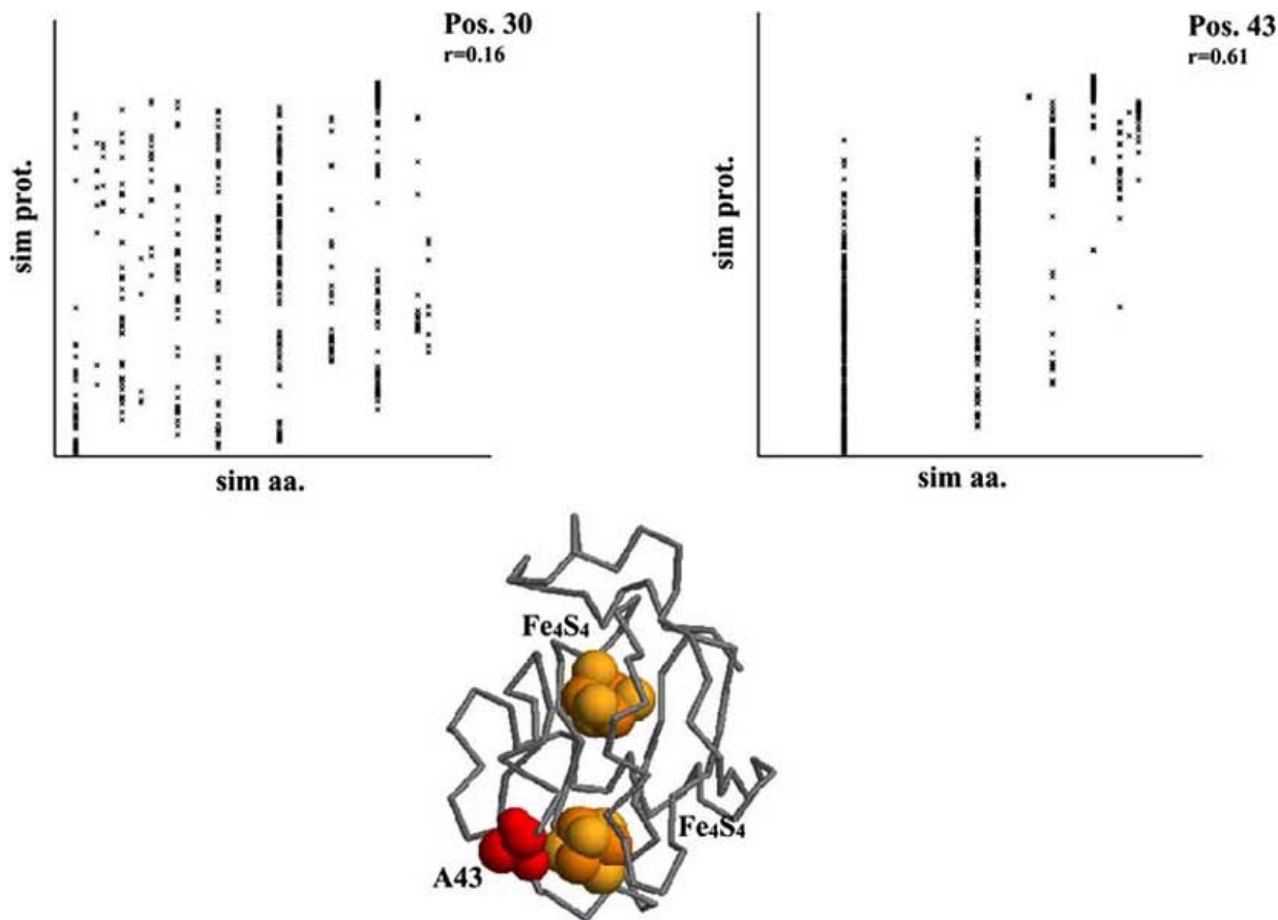


Fig. (2). Examples of the mutational behavior of a functional and a “non-functional” position in ferredoxin. Many methods are based on the idea that functional regions resemble the phylogeny of the whole family. An example of the application of the *MTreedet* method to ferredoxin is shown. Correlations between amino acid similarities in a position (X-axis) and the corresponding similarities in the whole family (Y-axis) are shown for two positions in ferredoxin: a functional position (43) involved in the binding of the iron-sulfur centers of the protein (which functions as an electron transporter), and a “non-functional” position (30). Position 43 shows a high correlation (the highest one) with the overall phylogeny of the family (high similarities between residues tend to be related with high similarities between the corresponding proteins and the other way around) which is reflected in the correlation coefficient. Other positions with high correlation values are also binding the Fe/S clusters (data not shown).

proteins and protein segments [48-51]. These unstructured polypeptides (lacking defined 3D structure in the native functional form) are experimentally detected as missing regions in X-ray crystal structures, highly mobile regions in NMR, or as certain characteristics of CD spectra indicative of secondary structure absence. Against the paradigm that structure determines function, what determines the function of these proteins is actually their lack of structure. The functional importance of these regions arises from their frequent involvement in protein-protein interactions. In many cases these unstructured segments become structured after binding, shifting the binding equilibrium to the bound form. This type of entropy-driven interaction has the unique specificity/affinity characteristics required for some biological complexes [52].

Although intrinsically unstructured regions have been related with low complexity segments in the primary sequence, [53,54] specific predictors based on neural networks or support vector machines (SVM) have been developed to locate these regions [55,56].

PREDICTION OF FUNCTIONALLY IMPORTANT REGIONS FROM STRUCTURE

When the 3D structure of the protein is available, we can add to the battery of methods described in the previous section, the ones which use structural information to predict functional features.

Several studies demonstrated that simple sequence (i.e. amino acid type) or structural (i.e. planarity) features are not sufficient to distinguish active or binding sites from the rest

of the surface in general [57-60]. Nevertheless, there are some signatures specific for certain types of complexes: for example, permanent homodimers tend to have planar interfaces rich in hydrophobic residues [57].

Combination of Sequence and Structural Information

The simplest methods which use structural information are just based on mapping the predictions of the sequence-based methods described in the previous section on the 3D structure, in order to see whether the predicted positions have the structural characteristics expected for a functional site (clustered or accessible to the solvent). Ben-Tal's group developed systems which look for clusters of conserved residues in the surface of the protein taking into account the distribution of sequences in the MSA for calculating conservation, as discussed in the previous section [15,61]. The method developed by Aloy *et al.* searches for conserved apolar residues which cluster in the surface of the protein [62]. The method successively removes distant sequences in the MSA until the set of conserved residues forms a cluster. This method is able to locate functional regions and binding sites which can be used as constraints for protein-protein and protein-DNA physical docking. Landgraf *et al.* method looks for regions of the protein following the same phylogeny as the whole family [36], a concept common in many methods described in the previous section. In this case, 3D information is used to define those regions (as sets of residues close in the surface of the protein). Some methods are also based on mapping sequence conservation on 3D structures but using a simplified alphabet which reflects the functional groups of the aminoacids, instead of considering the 20 aminoacids [63]. Sequence information has also been used in combination with stability profiles derived from 3D structures to locate enzyme active sites [64].

For the special case of predicting protein-protein interaction surfaces there are methods which use 3D and sequence information. 3D structure is used to define "surface patches" (groups of neighbor residues in the surface), whereas sequence information (MSAs) is used to extract the evolutionary characteristics of the members of a patch, generally in the form of sequence profiles. These patches are then used as input to machine learning techniques, either neural networks [65,66], or support vector machines [67], which are trained to classify the patches as either interacting or non-interacting.

Based on Structurally Related Proteins

There are other methods which are primarily based on structural information (generally extracting information from sets of structurally related proteins) to detect functionally important regions. The advantage of using structural alignments, instead of MSAs, is that they are able to capture distant relationships and hence discard artifacts due to the relatively evolutionary closeness of sequences in a MSA (i.e. conservation not due to functional requirements but to a short divergence time).

Following a nomenclature taken from protein structure prediction, these methods for function prediction from 3D structure can be classified as similarity-based or ab-initio methods, depending on whether they look for already-known or unknown functional sites respectively. The first class of

methods use databases of 3D templates (a set of residues in a 3D layout) derived from known functional/active sites and look for matches of these 3D templates in protein structures, whereas the second class can locate functional sites not described before. Skolnick's group developed a 3D representation of active sites called Fuzzy Functional Form (FFF) [68]. 3D structures can be scanned against these FFFs in order to locate functional sites. The methodologies developed in Thornton's group for retrieving and matching 3D templates of active sites allowed the creation the PROCAT database of enzyme active sites [69-71] and associated tools for predicting active sites trained on this database [72]. The PINTS database includes different sets of 3D profiles and a tool to scan a protein structure against them (or the other way around, a 3D profile against 3D structures) providing a robust statistical significance of the eventual matches [73]. The second class of methods does not rely on already-known functional profiles. Pazos & Sternberg developed a method for predicting 3D profiles associated with a given function based on structural alignments of proteins sharing that function [74]. We can also include in this class methods for the automatic unsupervised localization of sets of residues with a similar 3D layout within a set of protein structures (structurally related or not) [75], since these methods can be used to locate functional sites not described before, as these sites are expected to be structurally similar even in unrelated proteins (e.g. the catalytic triad of proteases).

There are also methods for locating and comparing protein surface motifs. The concept of "surface motif" is the 2D equivalent of the sequence motif in 1D [76,77]. It has been shown for some cases that proteins with different sequences and structures converge to similar surface patches [77] reflecting similar functions or binding properties.

Based on Single Structures

Related with what has been discussed in the previous section about disordered regions, it is known that in many cases binding and active sites are energetically unstable. This is due to many factors, including the need to put together energetically unfavorable combinations of residues because of functional requirements (for example two positive residues in a catalytic site), or to maintain unstable interacting regions that, being stabilized upon binding, favor the bound form in the binding equilibrium. Luque and Freire related the uneven distribution of the stabilization energy in proteins with binding sites and cooperativity effects [78]. Relationships between functionally important residues and the ones which negatively contribute to the electrostatic stabilization energy have also been found [79].

Theoretical microscopic titration curves have been also used to locate active sites in enzymes from structural information alone [80]. Concepts taken from Graph Theory have also been applied to the detection of functional and binding sites. In this case, a protein structure is represented as an undirected graph where the residues are the nodes and the residue-residue contacts the edges. Within this graph, residues (nodes) with special connectivity characteristics have been located and related with functional and binding sites [81,82]. Predictors based on neural networks have been trained to predict metal binding sites [83], which are

frequently related with functional sites. Many other works tried to related other structural features with functionality, like certain backbone conformations related with anion binding [84], characteristics of hydrogen-bonds [85] or surface clefts [86].

In the special case of predicting the complex (and implicitly the interfaces involved in the interaction) between two interacting proteins of known 3D structures there is a plethora of physical docking techniques. A discussion on that extensive field is outside the scope of this review. The interested readers can refer to excellent revisions on this subject [87-89].

DISCUSSION

The need to obtain functional information for the vast stream of new sequences and structures pushed the development of a plethora of methods for detecting functional sites from sequence or structure information, most of them available as web servers (Table 1) or standalone computer programs. The final user of these programs would like to know which method is better, in order to select among the many available ones. No exhaustive comparison of methods for predicting functional regions has been done so far, unlike protein structure prediction, a field where worldwide initiatives like *CASP* [90] or *EVA* [91] have been running for years to compare the different protein structure prediction methods. Two problems for comparing these methodologies are that they are very different in nature and they have been trained and tested in different datasets. But the main problem, maybe an unsolvable one, is the lack of a clear definition of “functional site”. Whereas in protein structure prediction “protein structure” is a univocal concept, and hence there is a clear gold-standard to compare against, in functional site prediction there is not such a standard. We think that working in a good definition of functional site and a good database for training, testing and comparing the methods is more urgent than developing new methods or improving the existing ones.

Depending on the information available for the target protein (sequence alone, 3D structure, sequence homologs, ...) different sets of methods can be used. Moreover, related with the problem discussed above, the user has to consider which kind of functionality he/she wants to predict (active sites, binding sites, allosteric sites...) and use the most appropriated method, since some of these techniques are specific for a certain type of functional feature. Table 1 contains a list of some of the methods described which are accessible as web servers, indicating the type of input they require.

In spite of these problems associated with functional site prediction, these methods are clearly demanded since it is totally impossible to experimentally obtain functional data for the overwhelming stream of new sequences and structures. *In-silico* methods can restrict the number of *wet* experiments, for example providing a set of candidate residues to mutate (instead of blindly trying all residues in the protein). A clever combination of both, experimental and *in-silico*, approaches will help us to interpret the genetic information in functional terms, which is the final goal of the so-called post-genomic era.

ACKNOWLEDGEMENTS

We want to thank the members of the Protein Design Group (CNB-CSIC, Madrid) and the Structural Bioinformatics Group (Imperial College, London), specially Prof. Alfonso Valencia, David Juan and Prof. Michael J. E. Sternberg, for interesting discussions and support. F.P. is the recipient of a “Ramón y Cajal” contract from the Spanish Ministry for Science and Technology. J-W.B. is supported by a UK Department of Trade and Industry Beacon Award QCB/C/012/00003.

ABBREVIATIONS

3D	=	Three-dimensional structure
MSA	=	Multiple sequence alignment
NMR	=	Nuclear magnetic resonance
CD	=	Circular dichroism

REFERENCES

- [1] Venter JC, Remington K, Heidelberg JF, *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **2004**; 304: 66-74.
- [2] Sali A. 100,000 protein structures for the biologist. *Nat Struct Biol* **1998**; 5: 1029-1032.
- [3] Winzler EA, Shoemaker DD, Astromoff A, *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **1999**; 285: 901-906.
- [4] Sönnichsen B, Koski LB, Walsh A, *et al.* Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **2005**; 434: 462-469.
- [5] Norin M, Sundström M. Structural proteomics: developments in structure-to-function predictions. *Trends Biotech* **2002**; 20: 79-84.
- [6] Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo C. From structure to function: approaches and limitations. *Nat Struct Biol* **2000**; 7: 991-994.
- [7] Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofra Y. Automatic prediction of protein function. *Cell Mol Life Sci* **2003**; 60: 2637-2650.
- [8] Valencia A. Automatic annotation of protein function. *Curr Opin Struct Biol* **2005**; 15: 267-274.
- [9] Devos D, Merino E, Pazos F, Valencia A. Multiple sequence alignments information in structure and function prediction. In: Press, I. (ed.), *Artificial Intelligence and Heuristic Methods for Bioinformatics* **2002**; 83-94.
- [10] Zuckerkandl E, Pauling L. Evolutionary Divergence and Convergence in Proteins. In: Bryson V, Vogel HJ (Eds), *Evolving Genes And Proteins*. Academic Press, New York, **1965**; 97-166.
- [11] Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **2001**; 42: 108-124.
- [12] Ouzounis C, Perez-Iratxeta C, Sander C, Valencia A. Are binding residues conserved? *Pacific Symposium on Biocomputing* **1998**; 3: 399-410.
- [13] Villar HO, Kauvar LM. Amino-acid preferences at protein binding sites. *FEBS Lett* **1994**; 349: 125-130.
- [14] Valdar WS. Scoring residue conservation. *Proteins* **2002**; 48: 227-241.
- [15] 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.
- [16] Berezin C, Glaser F, Rosenberg J, *et al.* ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **2004**; 20: 1322-1324.
- [17] Mayrose I, Mitchell A, Pupko T. Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account. *J Mol Evol* **2005**; 60: 345-353.
- [18] Gribskov M, Luethy R, Eisenberg D. Profile analysis. *Meth Enzymol* **1990**; 183: 146-159.
- [19] Bairoch A. PROSITE: A dictionary of sites and patterns in proteins. *Nucl Acids Res* **1992**; 20: 2013-2018.

- [20] Bateman A, Coin L, Durbin R, *et al.* The Pfam protein families database. *Nucl Acids Res* **2004**; 32: D138-141.
- [21] Mulder NJ, Apweiler R, Attwood TK, *et al.* The InterPro Database, 2003 brings increased coverage and new features. *Nucl Acids Res* **2003**; 31: 315-318.
- [22] Casari G, Sander C., Valencia, A. A method to predict functional residues in proteins. *Nat Struct Biol* **1995**; 2: 171-178.
- [23] Pazos F, Sanchez-Pulido L, García-Ranea JA, Andrade MA, Atrian S, Valencia A. Comparative analysis of different methods for the detection of specificity regions in protein families. In: Lundh, D., Olsson B, Narayanan A (Eds), *Biocomputing and Emergent Computation*. World Scientific, Singapore, New Jersey, London, Hong Kong, 1997; 132-145.
- [24] Atrian S, Sanchez-Pulido L, González-Duarte R, Valencia A. Shaping of *Drosophila* Alcohol Dehydrogenase through evolution. Relationship with enzyme functionality. *J Mol Evol* **1997**; 47: 211-221.
- [25] Azuma Y, Renault L, Garcia-Ranea JA, Valencia A, Nishimoto T, Wittinghofer A. Model of the ran-RCC1 interaction using biochemical and docking experiments. *J Mol Biol* **1999**; 289: 1119-1130.
- [26] Morillas M, Gomez-Puertas P, Bentebibel A, *et al.* Identification of conserved amino acid residues in rat liver carnitine palmitoyltransferase I critical for malonyl-CoA inhibition. Mutation of methionine 593 abolishes malonyl-CoA inhibition. *J Biol Chem* **2003**; 278: 9058-9063. Epub 2002 Dec 9023.
- [27] Bauer B, Mirey G, Vetter IR, *et al.* Effector recognition by the small GTP-binding proteins Ras and Ral. *J Biol Chem* **1999**; 274: 17763-17770.
- [28] del Sol Mesa A, Pazos F, Valencia A. Automatic Methods for Predicting Functionally Important Residues. *J Mol Biol* **2003**; 326: 1289-1302.
- [29] Lichtarge O, Bourne HR, Cohen FE. An Evolutionary Trace method defines binding surfaces common to protein families. *J Mol Biol* **1996**; 257: 342-358.
- [30] Lichtarge O, Bourne H, Cohen FE. Evolutionary conserved G $\alpha\beta$ binding surfaces support a model of the G protein-receptor complex. *Proc Natl Acad Sci USA* **1996**; 93: 7507-7511.
- [31] Lichtarge O, Yamamoto KR, Cohen FE. Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J Mol Biol* **1997**; 274: 325-337.
- [32] Onrust R, Herzmark P, Chi P, *et al.* Receptor and $\beta\gamma$ binding sites in the alpha subunit of the retinal G protein transducin. *Science* **1997**; 275: 381-384.
- [33] Sowa ME, He W, Wensel TG, Lichtarge O. A regulator of G protein signaling interaction surface linked to effector specificity. *Proc Natl Acad Sci USA* **2000**; 97: 1483-1488.
- [34] Mihalek I, Res I, Lichtarge O. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* **2004**; 336: 1265-1282.
- [35] Hannehalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* **2000**; 303: 61-76.
- [36] 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.
- [37] La D, Sutch B, Livesay DR. Predicting protein functional sites with phylogenetic motifs. *Proteins* **2005**; 58: 309-320.
- [38] Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci* **1993**; 6: 645-756.
- [39] Andrade MA, Casari G, Sander C, Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol Cybern* **1997**; 76: 441-450.
- [40] Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol* **2002**; 321: 7-20.
- [41] Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN. Finding important sites in protein sequences. *Proc Natl Acad Sci USA* **2002**; 99: 14764-14771.
- [42] Caffrey DR, O'Neill LA, Shields DC. A method to predict residues conferring functional differences between related proteins: application to MAP kinase pathways. *Protein Sci* **2000**; 9: 655-670.
- [43] Ofra Y, Rost B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett* **2003**; 544: 236-239.
- [44] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins* **1994**; 18: 309-317.
- [45] Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* **1997**; 2: S25-S32.
- [46] Mateu MG, Fersht AR. Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization. *Proc Natl Acad Sci USA* **1999**; 96: 3595-3599.
- [47] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* **1997**; 271: 511-523.
- [48] Dunker KA, Lawson JD, Brown CJ, *et al.* Intrinsically disordered protein. *J Mol Graphics Modell* **2001**; 19: 26-59.
- [49] Bracken C, Iakoucheva LM, Romero PR, Dunker AK. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol* **2004**; 14: 570-576.
- [50] Uversky VN. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci* **2002**; 11: 739-756.
- [51] Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* **2002**; 27: 527-533.
- [52] Dunker AK, Garner E, Guilliot S, *et al.* Protein disorder and the evolution of molecular recognition: theory, predictions, and observations. *Pacific Symp Biocomputing* **1998**; 3: 473-484.
- [53] Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. *Meth Enzymol* **1996**; 266: 554-571.
- [54] Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* **2001**; 42: 38-48.
- [55] Iakoucheva LM, Dunker AK. Order, disorder, and flexibility: prediction from protein sequence. *Structure (Camb)* **2003**; 11: 1316-1317.
- [56] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **2004**; 337: 635-645.
- [57] Jones S, Thornton JM. Analysis of Protein-Protein Interaction Sites using Surface Patches. *J Mol Biol* **1997**; 272: 121-132.
- [58] Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol* **1999**; 285: 2177-2198.
- [59] Ofra Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol* **2003**; 325: 377-387.
- [60] Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins* **2002**; 47: 334-343.
- [61] 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.
- [62] Aloy P, Querol E, Aviles FX, Sternberg MJE. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* **2001**; 311: 395-408.
- [63] Innis CA, Anand AP, Sowdhamini R. Prediction of functional sites in proteins using conserved functional group analysis. *J Mol Biol* **2004**; 337: 1053-1068.
- [64] Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J Mol Biol* **2003**; 327: 1053-1064.
- [65] Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* **2002**; 269: 1356-1361.
- [66] Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **2001**; 44: 336-343.
- [67] Koike A, Takagi T. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel* **2004**; 17: 165-173.
- [68] Di Gennaro JA, Siew N, Hoffman BT, *et al.* Enhanced functional annotation of protein sequences via the use of structural descriptors. *J Struct Biol* **2001**; 134: 232-245.
- [69] Wallace AC, Borkakoti N, Thornton JM. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **1997**; 6: 2308-2323.

- [70] Bartlett G, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **2002**; 324: 105-121.
- [71] Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl Acids Res* **2004**; 32: D129-133.
- [72] Gutteridge A, Bartlett G, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* **2003**; 330: 719-734.
- [73] Stark A, Russell RB. Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucl Acids Res* **2003**; 31: 3341-3344.
- [74] Pazos F, Sternberg MJE. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci USA* **2004**; 101: 14754-14759.
- [75] Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol* **2003**; 326: 955-978.
- [76] de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M. Three-dimensional profiles: a new tool to identify protein surface similarities. *J Mol Biol* **1998**; 284: 1211-1221.
- [77] Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. *J Mol Biol* **2004**; 339: 607-633.
- [78] Luque I, Freire E. Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins* **2000**; S4: 63-71.
- [79] Elcock A. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol* **2001**; 312: 885-896.
- [80] Ondrechen MJ, Clifton JG, Ringe D. THEMATIC: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA* **2001**; 98: 12473-12478.
- [81] Amitai G, Shemesh A, Sitbon E, *et al.* Network analysis of protein structures identifies functional residues. *J Mol Biol* **2004**; 344: 1135-1146.
- [82] Del Sol A, O'Meara P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins* **2004**; 58: 672-682.
- [83] Sodhi JS, Bryson K, McGuffin LJ, Ward JJ, Wernisch L, Jones DT. Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* **2004**; 342: 307-320.
- [84] Watson JD, Milner-White EJ. A novel main-chain anion binding site in proteins: the nest. A particular combination of phi, psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often as functionally important regions. *J Mol Biol* **2002**; 315: 171-182.
- [85] Fernández A, Scheraga HA. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc Natl Acad Sci USA* **2003**; 100: 113-118.
- [86] Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. protein clefts in molecular recognition and function. *Protein Science* **1996**; 5: 2438-2452.
- [87] Smith GR, Sternberg MJE. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol* **2002**; 12: 28-35.
- [88] Janin J, Seraphin B. Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol* **2003**; 13: 383-388.
- [89] Russell RB, Alber F, Aloy P, *et al.* A structural perspective on protein-protein interactions. *Curr Opin Struct Biol* **2004**; 14: 313-324.
- [90] Moulton J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins* **2003**; 53: 334-339.
- [91] Koh IYY, Eyrich VA, Marti-Renom MA, *et al.* EVA: evaluation of protein structure prediction servers. *Nucl Acids Res* **2003**; 31: 3311-3315.