doi:10.1016/S0022-2836(02)01451-1 *J. Mol. Biol.* (2003) **326**, 1289–1302

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Automatic Methods for Predicting Functionally Important Residues

Antonio del Sol Mesa†, Florencio Pazos† and Alfonso Valencia*

Protein Design Group National Center for Biotechnology, Cantoblanco Madrid 28049, Spain

Sequence analysis is often the first guide for the prediction of residues in a protein family that may have functional significance. A few methods have been proposed which use the division of protein families into subfamilies in the search for those positions that could have some functional significance for the whole family, but at the same time which exhibit the specificity of each subfamily ("Tree-determinant residues"). However, there are still many unsolved questions like the best division of a protein family into subfamilies, or the accurate detection of sequence variation patterns characteristic of different subfamilies. Here we present a systematic study in a significant number of protein families, testing the statistical meaning of the Tree-determinant residues predicted by three different methods that represent the range of available approaches. The first method takes as a starting point a phylogenetic representation of a protein family and, following the principle of Relative Entropy from Information Theory, automatically searches for the optimal division of the family into subfamilies. The second method looks for positions whose mutational behavior is reminiscent of the mutational behavior of the full-length proteins, by directly comparing the corresponding distance matrices. The third method is an automation of the analysis of distribution of sequences and amino acid positions in the corresponding multidimensional spaces using a vector-based principal component analysis. These three methods have been tested on two non-redundant lists of protein families: one composed by proteins that bind a variety of ligand groups, and the other composed by proteins with annotated functionally relevant sites. In most cases, the residues predicted by the three methods show a clear tendency to be close to bound ligands of biological relevance and to those amino acids described as participants in key aspects of protein function. These three automatic methods provide a wide range of possibilities for biologists to analyze their families of interest, in a similar way to the one presented here for the family of proteins related with *ras-p21*.

Keywords: functional residue; Tree-determinant position; bioinformatics; protein structure; protein function

**Corresponding author*

Introduction

Several methods have been developed to predict the residues in a protein family that may be involved in a given biological activity. Good candidates for functionally important sites in a multiple sequence alignment (MSA) are the completely conserved positions. However, it would be interesting to explore other sequence patterns indicating possible functionally important sites in a protein family. Other positions subject to specific variation between protein families may reveal key aspects of the evolution of the functional specificity and

† These two authors contributed equally to this work.

Present addresses: A del Sol Mesa, Bioinformatics Unit, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK; F. Pazos, Alma Bioinformatica, Centro Empresarial Euronova, Ronda de Poniente, Tres Cantos, Madrid 28760, Spain.

Abbreviation used: MSA, multiple sequence alignment.

E-mail address of the corresponding author: valencia@cnb.uam.es

0022-2836/03/$ - see front matter

provide additional information about the composition of the binding sites.

Casari *et al*.[1] introduced the "SequenceSpace" analysis method to detect those residues with a tendency to be conserved within a subfamily of proteins, but which differ between subfamilies (Tree-determinant positions), and regarded them as a result of the evolutionary scenario in which conservation and specificity are present in a delicate balance.

Similarly, Livingstone & Barton[2] analyzed those positions with conservation patterns in one or more subfamilies, even if they were not conserved in every subfamily. They extended their observations to include a characterization of the physico-chemical properties of these positions.

Landgraf *et al*.[3] developed a method that uses sequence and structure information to predict functional patches in protein surfaces. The method compares a regional similarity matrix for the residues in a surface patch with the global similarity matrix of the alignment and detects both regions whose conservation pattern is similar to the one of the whole alignment and those that are different.

Lichtarge *et al*.[4,5] proposed an "Evolutionary Trace" procedure, which predicts active sites and functional interfaces in proteins with known structure. In order to generate this method, they manually determined clusters of protein families based on the correlation between sequence identity and functional characteristics. Then, those completely conserved residues in the entire family, and those invariant residues that change between subfamilies, are mapped onto the structure to give a three-dimensional functional map. This method was recently automated[5] and applied to the prediction of patches on proteins surfaces related with protein function in a large set of proteins. Armon *et al*.[6] and Pupko *et al*.[7] use a more accurate evaluation of the rate of evolution per site to take into account possible artefacts related with the non-equal representation of the species in the MSA.

There have been other interesting approaches, which attempt to classify sequences in functional specific subfamilies within a protein family, and to detect conservation patterns related to a certain functional activity of the protein family.[8–12] To compare the results of all of these approaches is beyond the scope of this paper.

Our group has been particularly involved in the use of this type of methodology for the prediction of specificity residues in proteins of the *ras* superfamily. In 1994, using an early version of SequenceSpace, we predicted the involvement of three separate sequence regions in the switch of specificity between *rab5* and *rab6*, a prediction that was successfully corroborated by the replacement of the predicted region and the assessment of the function of the chimeras in cellular systems.[13] More recently, we have used the same approach to predict two key positions for determining the differential affinity for external activating effectors of *ras* and *ral* proteins.[14] As in the previous case, the exchange of these two positions was sufficient to produce a switch of the corresponding specificities. This experience made us confident in the predictions generated by the expert use of SequenceSpace and related approaches.

However, there are still many questions about the nature of Tree-determinant positions. For example, one of the important points to study in greater depth is the most appropriate way of dividing a protein family into subfamilies in order to associate the Tree-determinants with sites, which are likely to be responsible for functional differences between these subfamilies. Within this context one could discuss different questions like: is there any optimal division of a protein family into subfamilies exhibiting more of these positions? Do they follow certain variation patterns among subfamilies?

These questions have an intrinsic biological interest, and may help to find the Tree-determinants in cases where we do not have enough sequences or there is no optimal divergence among sequences in a protein family alignment.

Although the approaches described partially answers some of these questions, one could think about the possibility of exploring automatically different specificity levels of division of a protein family into subfamilies in order to understand some of the questions addressed above. Moreover, most of the methods described are not fully automatic and thus unsuitable for testing in large data sets or for other large-scale applications, like the prediction of functionally important residues in a genomic context.

In pursuit of these goals, we implemented three fully automatic methods for the detection of Tree-determinant residues that represent the main approaches to this problem. Although the methods have as a common general purpose the search for the best Tree-determinants involved in the functional activity of the protein family, they are based on distinct concepts and thus deal with different aspects of the problem. We also considered the completely conserved positions that complement the predictions of the first three methods.

"The Level Entropy Method" (S-method) is based on the automatic search for different levels of a protein family splitting into subfamilies to search for an optimal reliable level according to the number of Tree-determinants involved in the function of the protein family. In order to do that, we first analyzed different cuts of a phylogenetic tree of a protein family and evaluates the relation between the stability of the "cut level" and the number of Tree-determinants, normalized by the amount of conserved positions in each subfamily. Using the concept of Relative Entropy from Information Theory,[15] we measure the distance between the distribution of Tree-determinants and the product of the distributions of conserved positions in each subfamily. The general model is related to the one developed by Hannenhalli & Russel,[16]

although the explicit implementation is different. We are aware that in some cases, due to the complexity in the function of certain protein families, more than one level of division could show Tree-determinants involved in different functional activities. However, as mentioned above, within this method we aim to find the most informative level regarding the number of Tree-determinants possibly involved in biological activity.

The "Mutational Behavior Method" (MB-method) searches for Tree-determinant positions in a multiple sequence alignment whose mutational behavior is similar to the mutational behavior of the whole family. We calculate a correlation coefficient between the position change matrix, representing the mutational behavior of the potential Tree-determinant position, and the protein change matrix, representing the mutational behavior of the whole family. This implementation is similar to the one recently published by Landgraf *et al*.[3]

Finally, the third method "SequenceSpace Automatization Method" (SS-method) deals with the automation of the human analysis of the classification results produced by SequenceSpace.[1] SequenceSpace detects functionally important residues from a multiple sequence alignment. It is based on a vector representation of the aligned proteins and residues followed by a principal component analysis that allows the selection of those axes in the sequencespace most populated by the proteins in the family and the characteristic residues of the different subfamilies. Although this method works more effectively than other approaches,[17] it has the disadvantage of requiring human inspection and manipulation of the results (*via* an interactive interface), which renders it unsuitable for statistical purposes, like the prediction of functional residues for large sets of proteins or complete genomes. The human intervention mainly consists of the search for protein clusters (protein subfamilies), the discrimination of those clusters according to biological expertise, the search for residue clusters, and the selection of the matching residue clusters and protein clusters. The clustering step presented here goes one step towards the automatic processing of the SequenceSpace results. Although automatic implementation of human expert knowledge is impossible at present, the method attempts to identify the residues with similar tendencies by identifying clear clusters in the multi-dimensional space using a straightforward geometrical criterion.

To check the efficacy of the three methods (combined with the completely conserved positions), we tested our predictions on two independent sequence-non-redundant lists of protein families whose representative proteins have a known structure (and only one chain) in the Protein Data Bank.[18] The first list of 191 protein families is composed of proteins binding various chemical groups, like prosthetic groups and ions (hereafter "heteroatoms"), which are potentially required for their biological activity. From the list of heteroatoms in the PDB files, we excluded those representing solvent molecules (water, heavy water and others). The second list with 112 protein families includes proteins with annotated functionally important sites ("SITE" records in PDB). The alignment for each protein family was taken from the HSSP database.[19]

Our aim was to examine the closeness between the predicted Tree-determinants (or conserved positions) and the heteroatoms or functionally important annotated sites to assess whether they could be involved in some biological activity relating the heteroatoms or other biological functions. We also examined the closeness between the Tree-determinants themselves to assess the degree to which the predicted Tree-determinants form clusters, i.e. whether they are located around a possible binding or active site. We pay special attention to the predicted Tree-determinant in physical contact with the heteroatoms, and also to the ones that coincide with the annotated functional sites.

Together with the automatic analysis of many protein structures we examined in more detail several other examples. In particular, we present the detailed results obtained with the three methods for the protein *ras-p21*, which was previously analyzed in the publications by Casari *et al*.[1] and Lichtarge *et al*.[4]

Results

Coverage of the three methods

The three methods can be used to find Tree-determinant residues at various reliability levels, corresponding to internal parameters that regulate the stability of the predictions (see Materials and Methods). Briefly, the S-method gives a number of Tree-determinants taking into account certain cut-off values: the average bootstrapping value of the level to be considered, the size of the jump in local maximum of Relative Entropy between different levels and the percentage of sequence conservation. The MB-method has an intrinsic score for every Tree-determinant indicating the degree of correlation between the mutational behavior of this position and the mutational behavior the whole family. The SS-method can work at different levels of confidence by selecting a minimal number of residue clusters in which a given alignment position has to be present.

In the framework of the constraints of the methods, it is always possible to have enough Tree-determinants for the MB-method, in most cases for the SS-method and in fewer cases for the S-method. In the S-method the constraint that reduces the number of possible predicted Tree-determinants is the requirement that the subfamilies should have enough sequences, and with

Table 1.

		S	MB	SS	CONS00
(a) Average, median and mode *z*-score values for distances between predicted residues and heteroatoms					
Common[a](61)	Average	−0.72	−0.73	−0.59	−0.95
	Median	−0.82	−0.75	−0.60	−1.11
	Mode	−1.20	−1.00	−0.75	−1.50
All[b]		(69)	(149)	(134)	(185)
	Average	−0.76	−0.63	−0.57	−0.90
	Median	−0.87	−0.73	−0.59	−0.96
	Mode	−1.20	−1.10	−1.10	−1.40
(b) Average, median and mode *z*-score values for distances between predicted residues and annotated PDB sites					
Common[a] (32)	Average	−1.02	−0.76	−0.80	−1.11
	Median	−0.92	−0.84	−0.88	−1.17
	Mode	−1.68	−1.10	−1.22	−1.47
All[b]		(34)	(87)	(87)	(108)
	Average	−1.03	−0.98	−0.97	−1.11
	Median	−1.04	−0.97	−0.97	−1.27
	Mode	−1.50	−1.50	−1.36	−1.47
(c) Average, median and mode *z*-score values for distances between pairs of Tree-determinats in the heteroatoms test set					
Common[a] (61)	Average	−0.39	−0.26	−0.23	−0.24
	Median	−0.53	−0.27	−0.22	−0.32
	Mode	−1.10	−0.50	−0.30	−0.60
All[b]		(69)	(149)	(134)	(185)
	Average	−0.40	−0.32	−0.28	−0.20
	Median	−0.56	−0.30	−0.29	−0.23
	Mode	−1.10	−0.80	−0.50	−1.20
(d) Average, median and mode *z*-score values for distances between pairs of Tree-determinats in the PDB sites test set					
Common[a] (32)	Average	−0.38	−0.27	−0.30	0.00
	Median	−0.41	−0.24	−0.27	−1.11
	Mode	−0.50	−0.56	−0.32	−0.40
All[b]		(34)	(87)	(87)	(108)
	Average	−0.28	−0.32	−0.26	−0.07
	Median	−0.46	−0.48	−0.33	−0.06
	Mode	−0.50	−1.00	−0.50	−0.45

[a] Averge, median and mode calculated on the list of proteins to which all the methods can be applied (with in brackets).
[b] Average, median and mode calculated on the individual list of proteins that fulfil the requirements for the application of each methods (with in brackets).

enough conservation in the different subfamilies (see Materials and Methods). However, we selected a large overlapping set of protein families for the three methods that allows a fair comparison among them (Table 1).

Tree-determinants and conserved positions are predictors of heteroatom contacting residues

The most general hypothesis is that Tree-determinant residues are responsible for substrate binding specificity, and so they are expected to be in contact with, or in close proximity to the bound compounds in protein binding sites (see Introduction). To test this hypothesis we analyzed whether Tree-determinant residues are close to bound heteroatoms in three-dimensional protein structures, assuming the unavoidable errors introduced by equating bound compounds (heteroatoms) and functional binding sites.

We initially selected a list of 191 sequence non-redundant protein families that bind heteroatoms (excluding solvent). Owing to the restrictions of the methods, each provided predictions for a different subset of protein families from the initial list.

The left column of Figure 1 shows the distributions of the *z*-scores of all the Tree-determinants in the list of proteins binding heteroatoms for each method. Clearly Tree-determinants tend to be closer to the heteroatoms than the rest of the residues and the distributions are shifted toward significantly shorter distances (negative *z*-scores). Completely conserved positions also show this tendency. The median and mode of these distributions are shown in the Figure. Almost half the proteins have their Tree-determinant residues closer to the heteroatoms with *z*-score values better than −1.0 and approximately the 3% of them have *z*-scores better than −2.0. The median and mode are selected as the appropriate Figures for these non-symmetrical distributions that are poorly described by the average value (see Discussion). Table 1(a) shows the median and mode values of the *z*-scores for a large set of protein families analyzed with each method, and for a common list of proteins predicted by the three methods and the completely conserved positions (61 protein families).

The SS-method is the one that produces slightly worst results. The other two methods show similar values, with slightly better predictions when there are enough observations for the application of the S-method. Therefore the last two methods (S- and MB-methods) are more useful to predict the closeness between Tree-determinants and heteroatoms. Completely conserved positions tend to be closer to heteroatoms, so within our context they are more likely to be involved in some functional activity related to the heteroatoms.

In terms of direct contact with bound heteroatoms, Tree-determinants are predicting a direct contact only in 13.5% of the cases for the S-method, in 16.3% for the MB-method and in 11.0% for the SS-method. Although that is better than the corresponding random predictions, they are clearly poor values. It is clear that the detected proximity does not imply direct contact.

It is important to remember that we are applying automatic methods to large collections of protein families where some of the bound heteroatoms may not represent the complete substrate and products in the binding sites, or the compounds may not be related with the function of the protein. This may reduce the reliability of the results since any empty binding site or any compound binding to an artificial site will be directly translated into erroneous predictions.

We combined the predictions generated by the three methods and the conserved positions (Table 2(a)). We consider separately the intersection between the three methods and the addition of the completely conserved positions to the predictions of each method. The intersection of the predictions of the Tree-determinant methods tends to improve

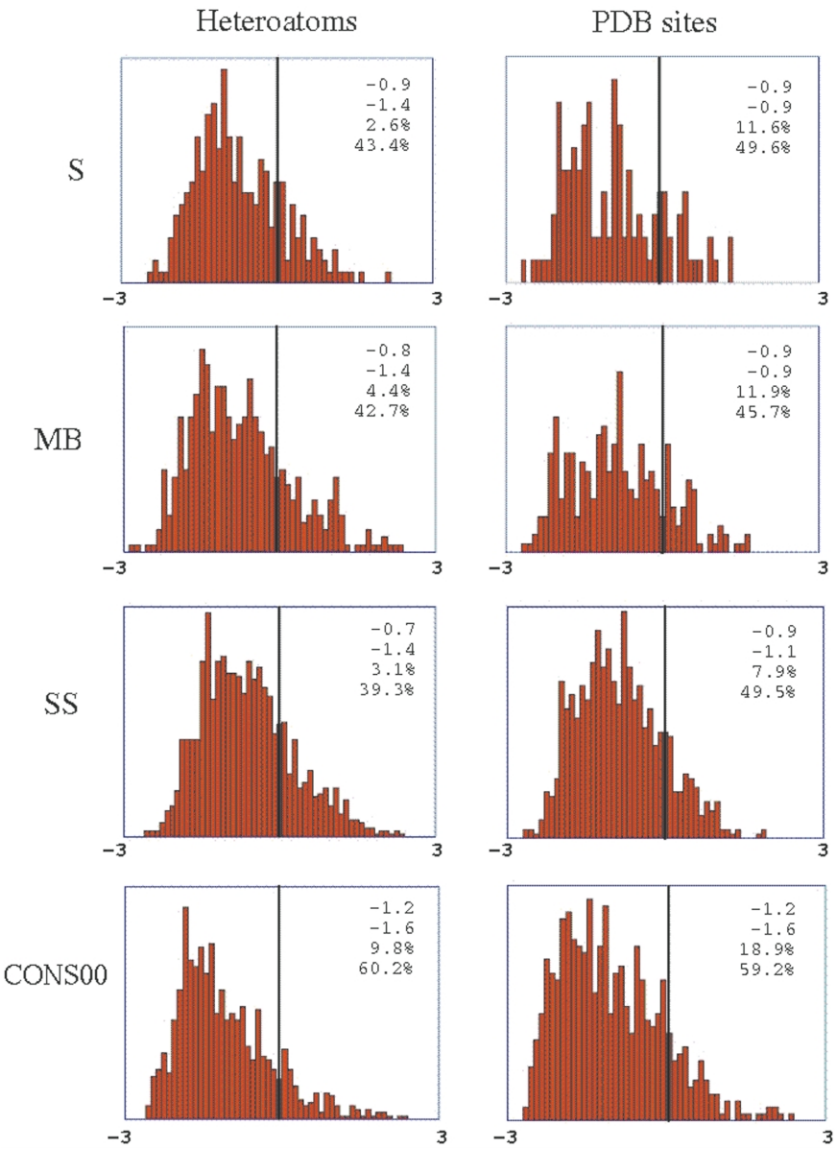

Figure 1. Distributions of the *z*-scores of distances for all Tree-determinants. The results include all the proteins to which it was possible to apply each one of the methods (not the common list). The evaluation has been carried out for the heteroatoms and the annotated sites test sets. The vertical lines mark *z*-score = 0 (random predictions). In the upper right part of each plot are shown the median of the distribution, the mode, the percentage of predicted residues with *z*-score better than −2.0 and the percentage with *z*-score better than −1.0, respectively.

the results by lowering the *z*-scores in most cases. This indicates that the residues predicted by two of the methods or by the three of them (fewer cases) form a more significant set of functionally important residues according to the criteria we are following here. This also implies that the three methods capture relatively different sets of functional residues.

The addition of the conserved positions to the predictions of the Tree-determinant methods tends to improve the results of the methods by decreasing the *z*-scores, and tends to worsen the results for the completely conserved positions. Results that support the known relation between conserved positions and binding sites.

Figure 2(a) shows the predictions of the three methods and the completely conserved positions for phthalate dioxygenase reductase (PDB code: 2pia), which is a prototypical iron–sulfur flavoprotein.[20] The heteroatoms, in this case, are an iron–sulfur (Fe_2S_2) cluster and an FMN prosthetic group. About one third of the total number of predicted Tree-determinants and conserved positions coincide with the annotated functionally important sites belonging to the iron–sulfur binding loop (residues 271–280) and, hence, they are binding the heteroatoms. All the Tree-determinants predicted by the three methods are clearly grouped around the iron–sulfur cluster whereas the conserved residues are distributed around the iron–sulfur and FMN groups.

Tree-determinants and conserved residues are predictors of positions involved in different types of biological function

The three methods were also tested on a list of 112 non-redundant protein families with annotations for functionally important sites (SITE records in PDB). We again have a set of proteins predicted by each method and a set common to the three of them and the completely conserved positions.

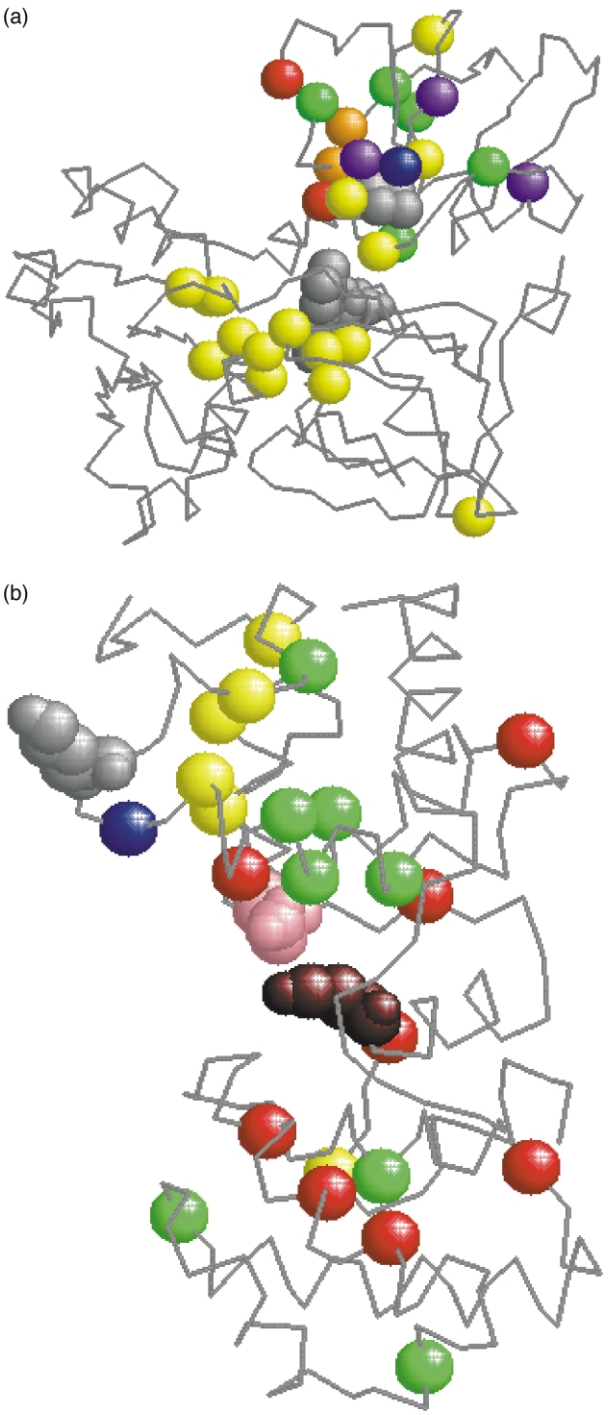

Figure 2. (a) Predicted Tree-determinant and conserved residues for the protein phthalate dioxygenase reductase. Phthalate dioxygenase reductase (PDB code 2pia) is an iron–sulfur flavoprotein used to illustrate a typical case of prediction of Tree-determinant residues contacting heteroatoms. The heteroatoms of this protein (iron–sulfur cluster—top in the Figure and FMN—bottom) are in grey and full-atom spacefill representation. For the predicted Tree-determinants and conserved residues only the Cα is shown. The completely conserved residues are drawn in yellow. The residues predicted by the MB-method are in red, the ones predicted by the S-method in green and the ones predicted by these two methods in orange. The residue predicted by the SS and S methods is in blue. The positions predicted by the three methods are in purple. (b) Prediction of three Tree-determinants and the completely conserved positions for endonuclease III. Endonuclease III (PDB code 2abk) has two annotated SITES in PDB: K120/D138 involved in DNA glycosylase and lyase activities; and K191, involved in DNA binding. The residues of these annotated sites are in grey, pink and brown (see below) and full-atom representation. For the predicted Tree-determinants and conserved residues only the Cα is shown. The completely conserved positions are in yellow. The predictions of the MB-method, the S-method and the SS-method are represented in red, blue and green respectively. The residue in pink is a completely conserved position that is annotated as a functional site and the one in brown is predicted by the MB-method and also annotated as SITE.

Table 2.

	CONS00 (union)	CONS10 (intersection)	SS	MB
(a) Combination of the three Tree-determinant prediction methods and conserved positions for the heteroatoms test set				
SS	−0.96	−1.58		
	129	72		
MB	−0.71	−2.26	−1.56	
	143	23	48	
S	−1.32	−1.39	1.64	−1.78
	68	39	42	17
(b) Combination of the three Tree-determinant prediction methods and conserved positions for the PDB site test set				
SS	−1.54	−1.85		
	83	47		
MB	−1.35	−2.03	−1.22	
	78	13	34	
S	−1.28	−0.72	−1.13	−1.04
	33	19	22	13

The Table show the results of the intersection of the residues predicted by each pair of methods and the addition of the completely conserved positions to the predictions of each method. The results are in the form of a matrix in which each entry corresponds to the intersection of two methods. For each intersection, the mode and the number of common proteins are shown. For the intersection with conserved residues, the positions with conservation >90% (HSSP VAR < 10) were used because there is no intersection between completely conserved residues (conservation = 100%; HSSP VAR = 0) with Tree-determinants.

Although the annotations for the functionally important sites may be incomplete, this approach complements the previous one (heteroatom binding, Results) in the absence of a definitive criterion of functionality. Therefore, we aimed to assess the functional significance of the predicted Tree-determinants and conserved amino acids regarding their distances to the annotated positions.

We carried out a *z*-score analysis similar to that performed previously, but instead of the distances from the predicted positions to the heteroatoms, we considered the distances from the predicted positions to the residues annotated as SITE in PDB. In some cases, certain predicted positions coincide with the annotated positions, but in most

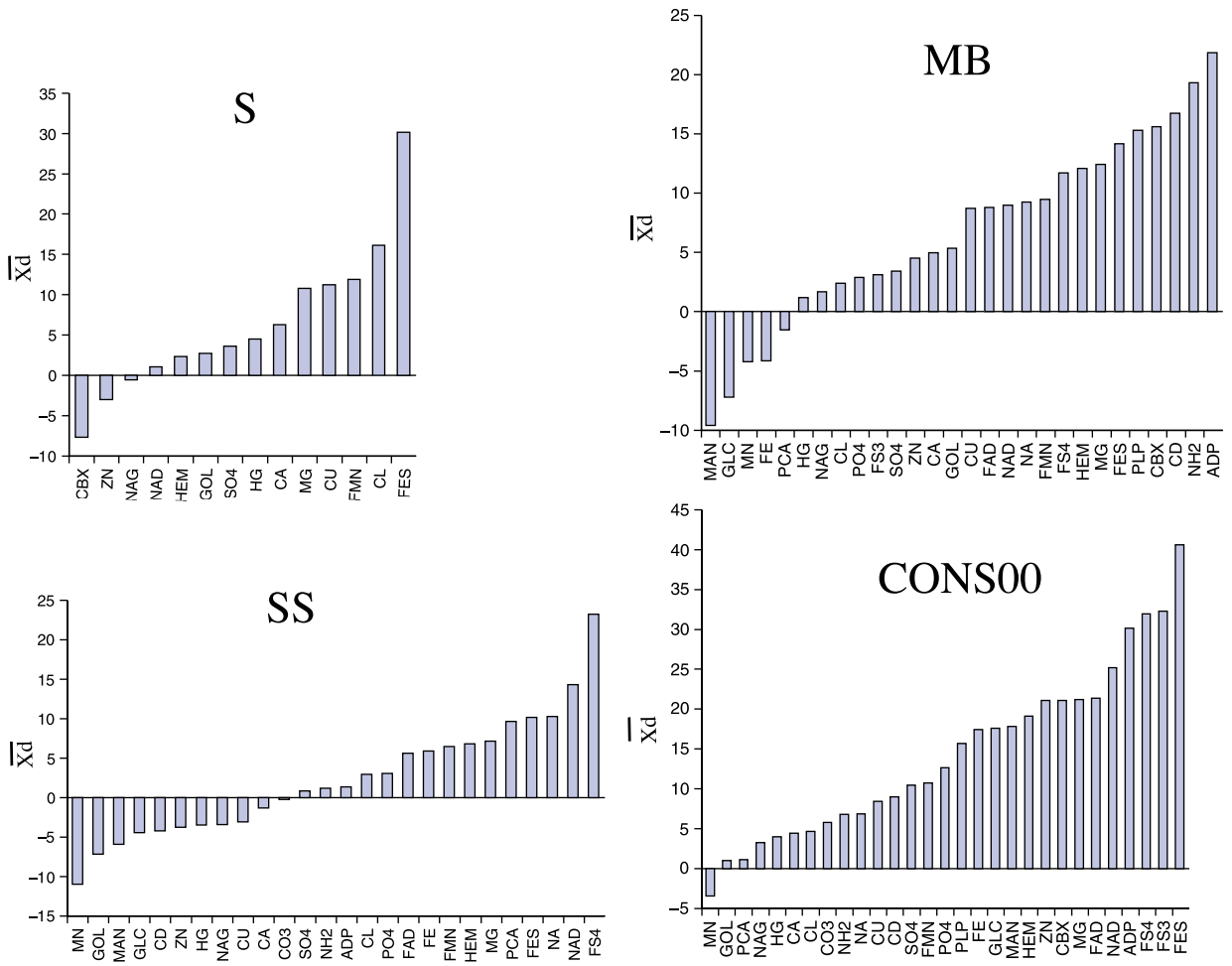

Figure 3. Average *Xd* values for proteins with various types of heteroatoms. The results are presented separately for the MB-method, S-method, SS-method and completely conserved positions. The analysis was performed on the longest list of proteins to which each method could be applied. The names of the heteroatoms are in PDB nomenclature.

cases the predicted positions co-localize with the annotated positions (right column in Figure 1). These tendencies are similar to the ones observed for the case of the heteroamons (Figure 1, left column).

In general, the results are encouraging. Unfortunately, the common list of proteins with alignments suitable for the three methods and the conserved positions is short (32 families), which compromises the statistical reliability of results based on this list.

In Table 2(b) (combination of the methods) we see a similar situation as in the case of the heteroatoms. The intersection of the Tree-determinant methods improves the results when we consider the distance from the predicted Tree-determinants to the annotated positions. The addition of the completely conserved positions to the predictions of the Tree-determinant methods improves the results of the previous methods at the expenses of reducing the accuracy of the predictions based on the conserved positions alone.

Figure 2(b) shows the predictions of the three Tree-determinant prediction methods and the completely conserved positions with the example of endonuclease III (2abk),[21] which has three functionally important residues annotated as SITE records in PDB (positions 120, 138 and 191). One of the completely conserved positions coincides with one of the annotated sites (138) and one of the residues predicted by the MB-method also coincides with one of the sites (120). The SS and MB methods have predictions close to the PDB sites, but some of them are also far apart (in Figure 2(b), see for example positions 51, 55, 81, 65 and 104). Even if we consider them as erroneous predictions, they may be involved in other functions not associated with the annotated SITEs. The residue predicted by the S-method (189) is close to one of the annotated sites (191). All the conserved residues but one are quite close to the annotated sites.

Distance between predicted positions

In both cases, the predictions for the heteroatom list and the PDB SITE list, we also considered the distances between the predicted positions

(Tree-determinant and conserved positions) to test whether they show a tendency to cluster in the three-dimensional structure.

For all the methods and the two test sets (heteroatoms and SITEs), the *z*-score values reveal only a slight tendency to cluster (Table 1(c) and (d)). This may be due to the presence of: (i) more than one binding region in the protein, involving more than one heteroatom, (ii) conformational changes in protein structures upon protein interaction that will separate residues that form part of a defined binding site, and (iii) large binding sites where the predicted residues are far apart.

Dependence of the predictions on the type of heteroatom

The heteroatoms include a broad range of chemical compounds crystallized with protein macromolecules. The average predictive power for all the proteins classified by type of heteroatom is shown in Figure 3 for the MB, S and SS methods, and for the completely conserved positions. The *Xd* average values (See Materials and Methods) for each type of heteroatom are calculated for each method based on its own protein list. The S-method is the one with the fewest observations, owing to the restrictive conditions imposed for its application.

The results for all the methods lead to a similar conclusion: large compounds (e.g. nucleotides and heme groups) are better predicted than small ones (ions). This result is not an artefact of the *Xd* calculation, since this parameter corrects by the size of the binding site and that of the protein. These results may be due to the fact that in certain cases small ions can be present as a result of the crystallization of the protein, and therefore functional residues are not expected to be associated with these ions. On the other hand, sugars, even if they are large, are not well predicted, which is not surprising since in many cases they are not part of binding sites but part of protein post-translational modifications (e.g. glycosylation) that cover the protein surface without a "localized" functional significance.

If we exclude from the calculations those ions and heteroatoms that are not directly related with biological binding activity, the *z*-score (mode) values become better than the ones obtained for all proteins (Table 1(a)): S-Method: −1.62; MB: −1.63; SS: −0.46 and CONS0: −1.59.

A relevant biological example

As described in Introduction, we used the manual version of SequenceSpace for the prediction of specificity regions in proteins of the *ras* superfamily. Other authors have also used this, or very related proteins such as α subunits of G proteins, to test their predictions.[1,4,22] Therefore, we analyzed the automatic predictions of Tree-determinants in this protein family (Figure 4).

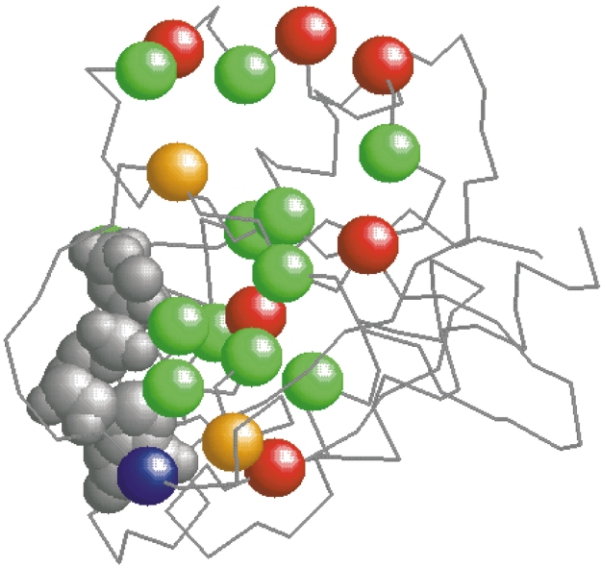

Figure 4. Functional residues predicted by the three methods for the *ras* protein (PDB code 5p21). The GTP group is in full-atom representation and coloured grey. For the predicted Tree-determinants only the Cα is shown. The residues predicted by the S, MB and SS methods are in blue, red and green respectively. The residues predicted by both, the S and MB methods, are in orange.

The S-method, with a bootstrap cut-off of 70% (see Materials and Methods), predicts as Tree-determinant position 28, which is a key component of the G1 region involved in the binding to the GTP-Mg cofactor.

The MB-method, with a correlation value cut-off of 0.6, predicts residue 37 in first position (highest correlation value), which is a well-known residue related with the change of specificity between *ras* and *ral* proteins,[14] and positions 22, 54, 65, 70, 73, 81 and 144, all of which are around the GTP binding site and the binding site for various *ras* effectors (switch I and II regions). From this set of residues, positions 37, 22, 54, 81 and 144 are clearly detectable by direct analysis of the SequenceSpace output by human experts.

The SS-method again includes residue 37, eight residues that are commonly detected by manual analysis of SequenceSpace (12, 18, 20, 22, 40, 56, 68 and 75) and another five residues that are particular to this method (9, 17, 64, 82, and 130). As with the other methods most of these residues are around the GTP binding site and/or the binding region for different *ras* effectors.

The conserved positions in all the known sequences (not shown in the Figure) belong to the GTP binding site, and the connection with various effectors (switch regions). Conserved residues also form part of the structural core of the protein that is substantially different of the binding site.

It is important to realise that the large part of the molecule with a clear functional activity related with the specific binding to other proteins (effectors, inhibitors and activators[14]) is well predicted

by all three methods. Unfortunately, the information about this protein–protein binding site is not present in the PDB records nor is it related with a heteroatom, therefore the automatic evaluation scheme used here will count these correct predictions of functional sites as errors of the methods. This situation could be common since few proteins have been co-crystallized with their effectors.

Discussion

The completely conserved positions in multiple sequence alignments are traditionally considered as likely candidates for functionally important sites.[23] Indeed their persistence during long periods of time has been interpreted as consequence of strong evolutionary pressure. We explore here other sequence patterns indicative of positions involved in functional activity of the whole protein family and determine the molecular basis of the functional specificity of their corresponding subfamilies.

Indeed, the difference between subfamily-specific residues (Tree-determinants) and conserved positions can be regarded as a product of the limited number of members known for a given family. In this sense, a conserved position may become a Tree-determinant as more sequences are added to the family. That is, the more sequences there are in a multiple alignment, the more likely we would be to find real Tree-determinants and conserved positions, since the definition of the subfamilies would become clearer.

Tree-determinant positions encapsulate the concept of evolutionary importance by their conservation in subfamilies and the concept of functional specificity by their difference between subfamilies. As mentioned in Introduction, Tree-determinants has been analyzed by several authors and, in particular examples, they have been shown to be functionally important positions for various protein families.[1–6,9,10,12]

In a few relevant cases these predictions have been used as a guide for the experimental switch of specificity between members of protein subfamilies: *rab5* and *rab6*,[13] *ras* and *ral*[14] and Gα,[22] These cases have required the expert use of the Tree-determinant detection methods. Despite the interesting work performed in key examples, the statistical assessment of the role of Tree-determinants as part of specific functionally important sites has not been tested systematically.

Here, we propose three different automatic methods to locate Tree-determinant positions that can serve as predictors for functionally important sites. Although these methods do not incorporate the biological knowledge of the user, they may be used for automatic predictions of functionally important sites in long lists of protein families, in systematic studies, and coupled to the pipeline of genome sequencing.

Various attempts have been made to construct comprehensive databases of annotated protein functional sites[8] and/or protein complexes. But there is a general lack of publicly available functional residue databases, which hinders the direct testing of methods for the prediction of functional sites.

For the current study, we tested three independent methods on two non-redundant lists of single chain protein families, one in which the binding sites are identified by the presence of various heteroatoms (191 protein families), and another in which functional sites have been manually annotated in the corresponding protein structures (PDB SITE records; 112 protein families).

For the first list, we assumed that heteroatoms bound to protein structures play some functional role. Considering as functional sites related to heteroatoms those that are close or bind them directly, we measured the distance from the predicted residues to the heteroatoms, correcting properly by the sizes of the protein and the binding site (see Materials and Methods). Regarding the second list, we took into account the distance from the predicted residues to the annotated positions.

We also considered the distances between the predicted residues to test whether they form clusters in the three-dimensional structures. That could indicate that these residues belong to an active site not related either with heteroatoms or with PDB sites, for example to protein–protein interaction surfaces.[24]

Despite the difficulties associated with binding site definition (binding of effectors or to cofactors not present in the protein crystal structures), the results clearly show a tendency of automatically detected Tree-determinants to be part of binding sites (proximity and to a lesser extent direct contact). But the Tree-determinants form clear clusters only in a few cases.

The distribution of the z-scores (Figure 1) are clearly skewed, with a long tail on the bad score regions (high z-score) that includes the poor predictions in many cases related with the insufficiency of the annotation of binding sites, described below. To assess the significance of the results we have used the standard median and mode estimators, since the mean is not an appropriate descriptor of asymmetrical distributions. In almost half of the cases the z-score values are better than -1, corresponding to mode values ranging from -0.9 to -1.4 and medians from -0.7 to -1.4 (Table 1 and Figure 1).

The relative low tendency of some proteins (tails of the distributions in Figure 1) to have predicted Tree-determinants close to heteroatoms or PDB sites, is not surprising since the large set selected for this experiment contains a number of imprecisions, i.e. functional sites not completely described in the PDB records, sites not completely conserved by the bound heteroatoms, and incomplete experimental data on the bound heteroatoms. These imprecisions, which are difficult to solve for the

full collection of proteins, should not prevent us for considering the potential for correct predictions and significant results shown for the majority of the proteins. This problem is illustrated by the case of the heteroatoms that cannot be predicted given their lack of biological significance. In fact, even the completely conserved residues are not significantly close to all the defined binding sites. Moreover, the results with sites described in PDB files are slightly better than the ones obtained for the sites described by their interaction with heteroatoms, since the quality of the annotations is slightly better for the manually annotated PDB files than for the sites described only by their proximity to bound heteroatoms.

We illustrated our results with concrete examples with known biological data, a heteroatom binding protein (iron–sulfur flavoprotein) and a endonuclease with PDB SITE annotation. In these examples, the predicted Tree-determinants and the conserved positions are close to the heteroatoms and to the PDB annotated sites, respectively.

We also followed in detail the predictions for the *ras* superfamily of GTPases, which has been used extensively as an example in most of the previous publications,[1,4,10] and carefully analyzed in several experimental approaches to the switch of specificity.[13,14] In this case, it is clear that the three methods, plus the conserved positions, point to the main interaction sites, where the *ras* proteins bind to specific effectors involved in the regulation of the GTP hydrolysis and exchange and also to the GTP binding site.

As a general tendency, the intersection of the predicted residues by two of the methods or by the three of them improves the results (decreases the values of the *z*-scores) regarding the distances from the predicted residues to the heteroatoms (or PDB sites). The addition of the conserved positions to the prediction methods improves the results in terms of distances from the selected positions to the heteroatoms (or PDB sites), decreasing the *z*-score values, reflecting the predictive power of the conserved positions.

We also analyzed the results of the methods in terms of the type of heteroatoms involved in the predictions (see Results) and noticed a common tendency of all methods to predict Tree-determinants associated with large heteroatoms (except for sugars) rather than with small ones (ions). This tendency was obvious after correcting the distances from the predicted positions to the heteroatoms by the size of the proteins and of the binding sites (*Xd* values). Taking into account that in many cases the protein crystals include some ions which are not functionally important for the proteins, and in other cases sugars are involved in some post-translational modifications (like glycosylation) without localized functional meaning, we could explain that the best predictions are associated not with ions or sugars, but with other bulky heteroatoms which appear in the protein crystal and are part of the functional mechanism of various proteins and therefore related to the functional specificity of the protein family.

It is interesting that the accuracy of the completely conserved positions as predictors of functionally important sites also depends on the type of heteroatom, in a way similar to the predictions of the Tree-determinant methods (see Results). With these results in hand we propose the use of Tree-determinant detection methods to predict functional sites, in particular those associated with biochemical functions (like the ones that require the presence of large cofactors).

In the future, we aim to use our automatic methods to predict binding sites in large collections of known structures (structural genomics) or protein sequences (complete genomes). At the same time, these methods can be used to explore the nature of the Tree-determinants involved in certain functional activities of different protein families, for example, to explore a possible correlation between heteroatom type and subfamily specificity.

Materials and Methods

The Level Entropy Method (S-method)

Phylogeny of a protein family

The main idea of this method is to explore automatically different possibilities of Tree-determinants, in the sense of positions that are conserved (85% conservation) within a subfamily, and differ between subfamilies, regardless of the physico-chemical characteristics of the residues. Since the method aims to analyze different divisions of the protein family in subfamilies, it incorporates a phylogenetic representation of the protein family and an appropriate random model.

We use as a starting point a phylogenetic tree of the protein family generated by the ClustalW program[25] from the HSSP[19] multiple sequence alignment of that family. Then, we analyze different levels of division of the tree, starting from the first division of the whole protein family that ClustalW produces and going in the direction of more specialized divisions. In this way, each level defines a specific division of the family into subfamilies to be analyzed. Figure 5 shows the first four levels of a tree, considering several possibilities of Tree-determinants. The fact that ClustalW generates and unrooted tree does not affect our goal, since we are using the tree just as a reference for grouping the proteins of the whole family in subfamilies.

As the main objective of this analysis is to study the Tree-determinants at each level, we regarded subfamilies as branches with more than two sequences. We pay attention only to the first half of the number of levels, since we aim to relate Tree-determinants with functional important positions for the whole family, and the more we advance in the division of subfamilies, the more likely we are to lose sequences on the way, according to the subfamily definition we imposed above.

Finally, we focused on the stable levels of the tree, particularly the first stable level. As a measure of stability we use the average of the bootstrapping values belonging to each branch of the division level, considering a level stable if the average of bootstrapping values exceed 75%.

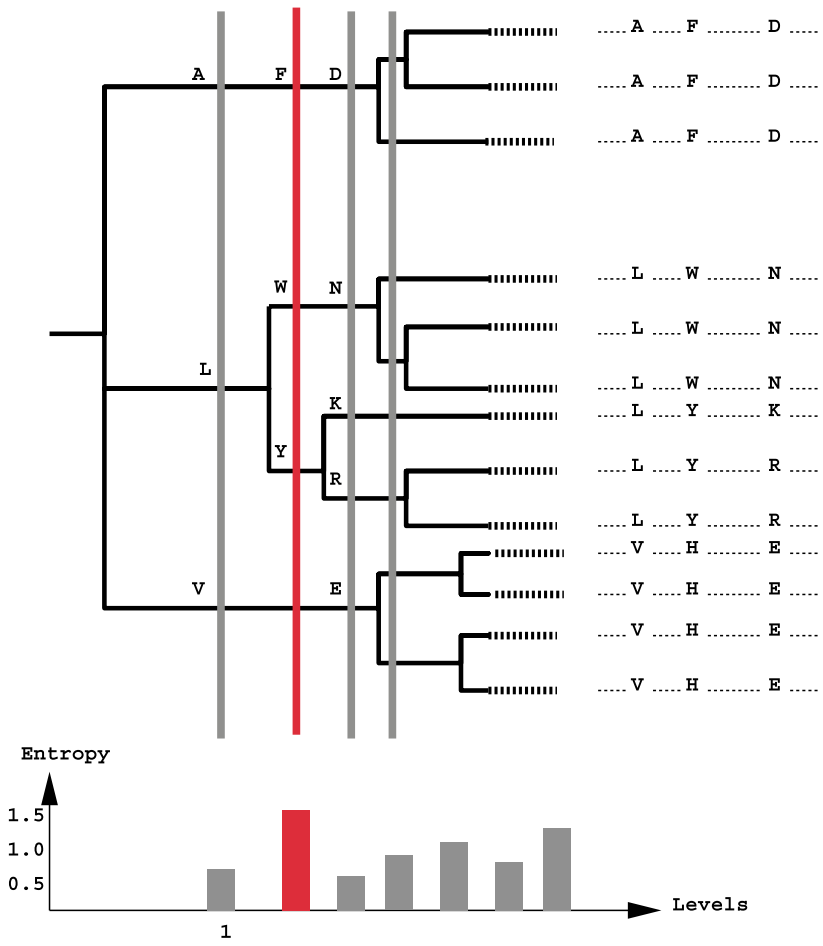

Figure 5. Schema of the Level Entropy Method (S-Method). The Figure represents different divisions of a protein family into subfamilies obtained by cutting the phylogenetic tree at different levels. The Tree-determinant positions at each division level are shown. The selected level (red) exhibits the local maximum of the Relative Entropy.

Entropy analysis

Once we have the division of the protein family in subfamilies at each level, we calculate the Tree-determinants at those level, requiring 85% or more conservation in each subfamily. We use the HSSP[19] multiple sequence alignment.

However, as subfamilies are further divided, the number of conserved positions in each subfamily tends to increase, and the probability of finding a Tree-determinant by chance also tends to increase. For this reason, we introduce the principle of Relative Entropy in order to normalize the number of Tree-determinants by the number of conserved positions in each subfamily. Relative Entropy (also known as the Kullback–Leibler distance in Information Theory)[15] could be thought of as a distance between two probability distributions. When one of the distributions is the joint probability distribution of n random variables and the other is the product of the probability distributions of each of the n variables, the Relative Entropy becomes an entropy measurement called Mutual Information, and it gives us an idea about how independent the n random variables are. In our problem, we introduce Relative Entropy (or in this case Mutual Information) in order to consider the distance of the "probability distribution" of Tree-determinants at a certain level from the product of probability distributions of conserved positions in each subfamily of that level, where instead of probability distributions we used frequencies.

Suppose we have a level with n subfamilies, then Relative Entropy is defined as

$$H(n) = \sum_{x_1,x_2,\ldots,x_n} p(x_1,x_2,\ldots,x_n)\log\frac{p(x_1,x_2,\ldots,x_n)}{\prod_{i=1}^{n} p(x_i)}x \qquad (1)$$

where the sums are taken over the 20 aminoacids, and $p(x_1,x_2,\ldots,x_n)$ stands for the frequency of obtaining by chance a Tree-determinant where in the ith subfamily one finds the residue $x_{ip(xi)}$ are the frequencies of obtaining by chance in the ith subfamily a conserved position with the residue x_i.

In other words, Level Relative Entropy considers implicitly the number of Tree-determinants of each type, corrected by the number of conserved positions that define that type of Tree-determinant at each subfamily.

Of all the stable levels (bootstrap $\geq$75%), we select the first that represents a jump or a local maximum of Relative Entropy. That will be our optimal level within the context of this analysis. Certain protein families do not have optimal levels. That is why it is possible to obtain results only for part of the list of non-redundant protein families (see Results).

Finally, note that we consider not only those Tree-determinants with the conserved residues different in each subfamily, but we also include conserved residues common to two or more subfamilies.

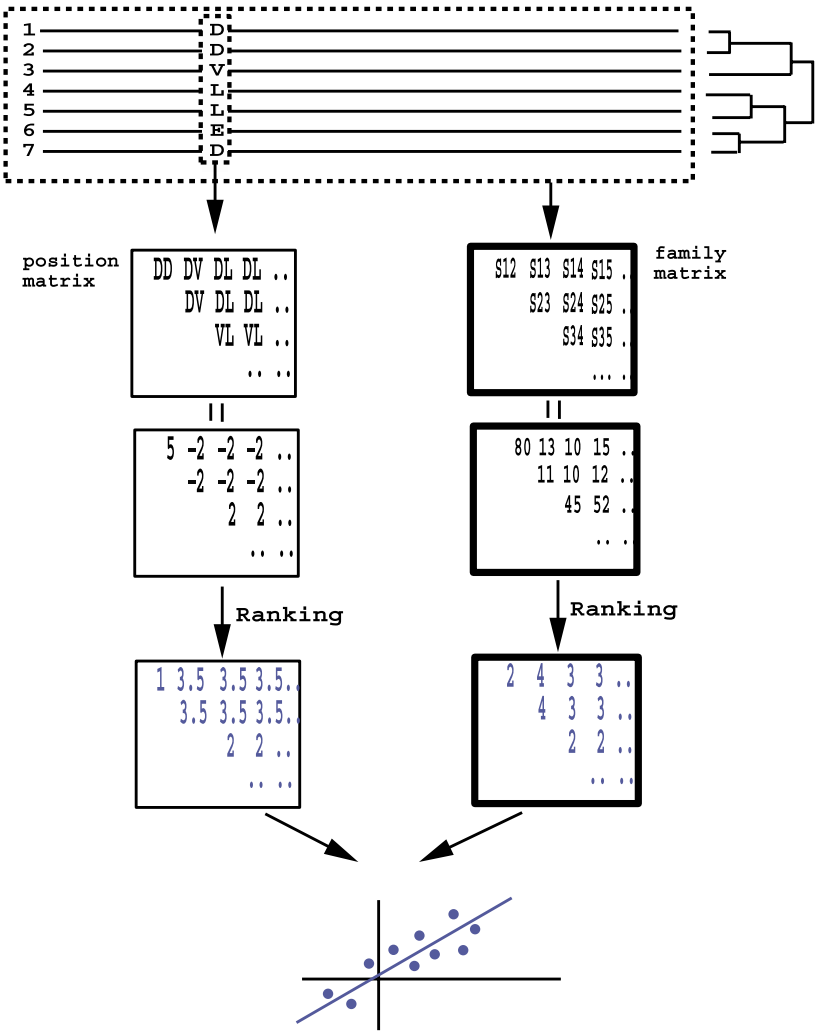

Figure 6. Schema of the Mutational Behavior Method (MB-method). The similarity between the mutational behavior of a position and the mutational behavior of the whole family is evaluated for each position in the multiple sequence alignment. The mutational behavior of the position is represented by a matrix containing the similarities for all pairs of residues occupying that position. The mutational behavior of the whole family is represented by a matrix containing the overall similarities for all pairs of proteins in the family. The similarity between the mutational behavior of the position and the mutational behavior of the family is evaluated as the similarity between these two matrices using a rank-correlation criterion.

The Mutational Behavior Method (MB-method)

This method takes into account that, given the definition of Tree-determinants given in Introduction, the variation pattern of such positions ought to reflect the variation pattern of the entire alignment (Figure 6). So, the "mutational behavior" of the Tree-determinant positions would be similar to the mutational behavior of the whole family.

One the one hand, the mutational behavior of the family is represented by a matrix whose elements are the homologies between each protein pair. On the other hand, the mutational behavior of each individual alignment position is represented by a matrix whose elements are the homologies between the pairs of residues at that position (Figure 6). These homologies are the substitution values taken from the McLachlan substitution matrix.[26] To test whether these two sets of data represent a common pattern of variation, we use the Spearman Rank-Order Correlation Coefficient defined as follows:[27]

$$r = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}} \qquad (2)$$

where R_i and S_i are the rank order values of the matrix elements belonging to the protein change matrix and position change matrix respectively (Figure 6). $\bar{R}$ and $\bar{S}$ are their average values.

A high value of this correlation coefficient for a given position means that the variation pattern of that position resembles the variation pattern of the whole family and therefore, that position is a good candidate for a Tree-determinant within the context of the present method.

For this study, we took the 10% of residues with highest correlation value as the predicted Tree-determinants.

The method does not use any explicit representation of the family phylogenetic tree but the phylogenetic information is contained in the protein homology matrix (distances between sequences). Therefore, this method is independent of any tree-construction algorithm. The method does not have as input specific divisions of the protein family into subfamilies, but those divisions are implicitly taken into account in the protein change matrix.

The idea was somehow inspired by our own independent way of calculating "correlated mutations".[28] Indeed, using two position matrices instead of a position matrix against a protein matrix, we calculate a correlated mutation between these two positions. Other authors have developed approaches to compare the variation of the full-length proteins with that of a given segment of those proteins in order to search for regions whose variability or conservation differs from that of the whole family.[29] More recently Landgraf *et al.*[3] developed a

similar method where a global similarity matrix is compared with a regional (surface patch) similarity matrix (see Introduction).

The SequenceSpace automation method (SS-method)

The final SequenceSpace output for residues is a list of the coordinates respect to the principal axes (the first six principal axes by default) for every amino acid type at each position in the alignment.[1]

We cluster this multidimensional cloud of points based on the distances between them to form groups of residues in similar positions of the multidimensional space. The algorithm calculates an optimal number of clusters and elements that correspond to them. The algorithm starts by selecting a random point as hub of the first cluster and the most distant point in the multidimensional space as the hub of the second cluster. Then, the points are assigned to clusters of the closest hub. If there is any point more distant of its hub than the average distance between hubs, a third cluster is created with the point farer apart from its hub as new hub. This process is iterated until no additional points match that condition on the distance to the hubs.

Once the clusters are determined, the one containing the very variable residues, that is usually the largest, is excluded since these residues contain less information about the structure of the family. Within the remaining clusters, the algorithm looks for positions represented by different amino acid types in different clusters (residues conserved and different between subfamilies). Finally, we select among them the most informative positions by picking the ones with a distance from the origin of coordinates higher than half of the maximum distance from any point to the origin.

Completely conserved positions

Although the methods described above deal with Tree-determinants, we also analyzed the completely conserved positions (defined as having zero variability in the HSSP files). The comparison between Tree-determinants and completely conserved positions as predictors of functionally important sites is discussed in the text.

z-Score analysis

To measure the significance of the closeness between Tree-determinants and heteroatom(s) (or PDB annotated site(s)) in a given protein, we consider the *z*-scores of the distances of each Tree-determinants to that heteroatom (or PDB annotated site) respect to the distributions of distances from all the protein residues to the heteroatom (or PDB site). To measure how significantly close the Tree-determinants are among themselves we do the same *z*-score calculation but for the distances between Tree-determinants with respect to the distribution of distances among all pairs of residues in the protein. The *z*-score is defined as follows:

$$z = \frac{r - \bar{r}}{\sigma} \tag{3}$$

where, in the first case, r is the distance of each Tree-determinant to the closest heteroatom (or PDB site), $\bar{r}$ is the average value of the distances of all the protein residues respect to the same heteroatom (or PDB site), and σ is the corresponding standard deviation. In the second case (evaluation of closeness between predicted Tree-determinants) we have the same meaning for all the symbols except for the fact that we are considering now distances among the Tree-determinants respect to the distances of all pairs of residues in the protein.

In both cases, we calculate the average value of the *z*-scores for all Tree-determinants in one protein, and the average of these average values in a set of proteins, to have single values of accuracy for one protein and for a set of proteins respectively (see Table 1). We also calculate the median and mode of the *z*-score values.

Xd analysis

In the previous section, we introduced a *z*-score analysis to evaluate how significantly close each of the predicted residues were to the heteroatom, to the PDB annotated site or to other predicted residue. However, if instead of single values of distance between a predicted residue and a heteroatom (PDB site or other residue), we want to assess the significance of the distances of a large set of predicted residues (sub-population) respect to all distances (population), we use the concept of *Xd* introduced in a previous publication.[24]

$$Xd = \sum_{i=1}^{i=n} \frac{P_{ic} - P_{ia}}{d_i xn} \tag{4}$$

where n is the number of distance bins in the distributions (there are 15 equally distributed bins from 4 Å to 60 Å); d_i is the upper limit for each bin (corrected to 60). P_{ic} is the percentage of predicted residues with distance to the heteroatom (PDB site, or other residue) between d_i and d_{i-1}, and P_{ia} is the same percentage for all residues in the protein. Defined in this way, $Xd > 0$ indicates positive cases for which the population of predicted residues is shifted to smaller distances with respect to the population of all residues.

Once the *Xd* value is calculated for the predictions of each method for each protein in the list (depending on the method), we calculate the average value of *Xd* over those proteins that bind the same heteroatom and discuss these data obtained for each type of heteroatom (see Results).

Acknowledgements

We would like to acknowledge interesting discussions in issues related with the prediction of binding sites with P. Aloy (EMBL), G. Casari (Cellzome), C. Ouzounis (EMBL-EBI), B. Rost (U. Columbia) and C. Sander (MIT). The continuous support and the fruitful exchanges with members of the protein design group have been particularly productive, particularly the inside provided by J. A. Garcia-Ranea into the predictions of the *ras* binding sites. A. del Sol developed the S-method and F. Pazos the MB and SS-methods. Both of them carried out the analysis of the results. A. Valencia originated the idea and contributed to analysis of the results. The three of them work together in the preparation of the manuscript. This work was supported in part by a grant from the CICYT (MCyT, Spain) by EC grants QLK3-CT-

1999-00875 and QLRT-2000-01663, and A. del Sol by a fellowship from the DIO-CNB department.

References

1. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178.
2. Livingstone, C. D. & Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **6**, 645–756.
3. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
4. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An Evolutionary Trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
5. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
6. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.
7. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.
8. Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. & Godzik, A. (1999). From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104–1115.
9. Dorit, R. L. & Ayala, F. J. (1995). ADH Evolution and the phylogenetic footprint. *J. Mol. Evol.* **40**, 658–662.
10. Andrade, M. A., Casari, G., Sander, C. & Valencia, A. (1997). Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* **76**, 441–450.
11. Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293.
12. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
13. Stenmark, H., Valencia, A., Martinez, O., Ulrich, O., Goud, B. & Zerial, M. (1994). Distinct structural elements of *rab5* define its functional specificity. *EMBO J.* **13**, 575–583.
14. Bauer, B., Mirey, G., Vetter, I. R., Garcia-Ranea, J. A., Valencia, A., Wittinghofer, A. *et al.* (1999). Effector recognition by the small GTP-binding proteins Ras and Ral. *J. Biol. Chem.* **274**, 17763–17770.
15. Shannon, C. & Weaver, W. (1963). *Mathematical Theory of Communication*, University of Illinois press, Champaign, IL.
16. Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.
17. Pazos, F., Sanchez-Pulido, L., García-Ranea, J. A., Andrade, M. A., Atrian, S. & Valencia, A. (1997). Comparative analysis of different methods for the detection of specificity regions in protein families. In *Biocomputing and Emergent Computation* (Lundh, D., Olsson, B. & Narayanan, A., eds), pp. 132–145, World Scientific, Singapore.
18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
19. Sander, C., Schneider, R. & The, H. S. S. P. (1993). data base of protein structure-sequence alignments. *Nucl. Acids Res.* **21**, 3105–3109.
20. Correll, C. C., Batie, C. J., Ballou, D. P. & Ludwig, M. L. (1992). Phthalate dioxygenase reductase: a modular structure for electron transfer from pyridine nucleotides to [2Fe–2S]. *Science*, **258**, 1604–1610.
21. Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P. & Tainer, J. A. (1995). Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J.* **14**, 4108–4120.
22. Lichtarge, O., Bourne, H. & Cohen, F. E. (1996). Evolutionary conserved Gαβγ binding surfaces support a model of the G protein–receptor complex. *Proc. Natl Acad. Sci. USA*, **93**, 7507–7511.
23. Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes And Proteins* (Bryson, V. & Vogel, H. J., eds), pp. 97–166, Academic Press, New York.
24. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.* **271**(4), 511–523.
25. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**, 189–191.
26. Mclachlan, A. D. (1971). Test for comparing related aminoacid sequences. *J. Mol. Biol.* **61**, 409–424.
27. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edit., Cambridge University Press, Cambridge.
28. Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.* **18**, 309–317.
29. Dopazo, J. (1997). A new index to find regions showing an unexpected variability or conservation in sequence alignments. *Comput. Appl. Biosci.* **13**(3), 313–317.

Edited by F. E. Cohen

(Received 4 December 2002; accepted 11 December 2002)