BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, pages 1440–1448
doi:10.1093/bioinformatics/btl104

Sequence analysis

Phylogeny-independent detection of functional residues

Florencio Pazos*, Antonio Rausell and Alfonso Valencia

Protein Design Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3. Campus U. Autónoma, 28049 Cantoblanco, Madrid, Spain

Received on January 19, 2006; revised and accepted on March 16, 2006
Advance Access publication March 21, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Current projects for the massive characterization of proteomes are generating protein sequences and structures with unknown function. The difficulty of experimentally determining functionally important sites calls for the development of computational methods. The first techniques, based on the search for fully conserved positions in multiple sequence alignments (MSAs), were followed by methods for locating family-dependent conserved positions. These rely on the functional classification implicit in the alignment for locating these positions related with functional specificity. The next obvious step, still scarcely explored, is to detect these positions using a functional classification different from the one implicit in the sequence relationships between the proteins. Here, we present two new methods for locating functional positions which can incorporate an arbitrary external functional classification which may or may not coincide with the one implicit in the MSA. The *Xdet* method is able to use a functional classification with an associated hierarchy or similarity between functions to locate positions related to that classification. The *MCdet* method uses multivariate statistical analysis to locate positions responsible for each one of the functions within a multifunctional family.

Results: We applied the methods to different cases, illustrating scenarios where there is a disagreement between the functional and the phylogenetic relationships, and demonstrated their usefulness for the phylogeny-independent prediction of functional positions.

Availability: All computer programs and datasets used in this work are available from the authors for academic use.

Contact: pazos@cnb.uam.es

Supplementary information: Supplementary data are available at http://pdg.cnb.uam.es/pazos/Xdet_MCdet_Add/

INTRODUCTION

If the genomic era was characterized by the massive sequencing of complete genomes, the so-called 'post-genomic' era is being, may be, characterized by an unexpected lack of tools for obtaining relevant information from these raw sequences. Today, we know the complete sequences of hundreds of genomes from the three kingdoms, and 'environmental sequencing' (Venter *et al*., 2004) (the organism-independent sequencing of DNA repertories directly extracted from environmental samples) is boosting the number of available sequences. There is also an increasing number of proteins of known three-dimensional (3D) structures without associated functional information, in part owing to the Structural Genomics projects (Brenner, 2001).

Determining which residues in a protein are responsible for its function is very important in order to understand its molecular mechanism, to modify this function in our benefit (biotechnology) or to correct problems related with this function (e.g. pathologies). The experimental characterization of function and functional features (functional sites, etc.) is very expensive, time consuming and difficult to automate. This justifies the development of computational methods for predicting functional sites and other functional features for these uncharacterized sequences.

Some methods use previously known functional sites to derive sequence profiles (Mulder *et al*., 2003) or structural templates (Di Gennaro *et al*., 2001; Porter *et al*., 2004) in order to match new sequences/structures against them. Other techniques are able to detect functional sites without previous knowledge of them. Some of these methods are able to predict functional sites based on single sequences, like the method developed by Ofran and Rost for predicting protein interaction sites (Ofran and Rost, 2003). Others are based on single 3D structures. They look for structural features frequently associated with active sites and binding sites, like low-stability regions (Elcock, 2001) or special connectivity patterns extracted from residue–residue contact networks (Del Sol and O'Meara, 2004). Nevertheless, most of the methods predict functional positions based on multiple sequence or structure alignments of related proteins, and they work under the common assumption of conservation of functional residues during evolution.

The advantage of structural alignments is that they can relate remote homologs (Pazos and Sternberg, 2004), and their drawback is that they need 3D structures to work, which are more scarce than sequences.

Since sequences are still more abundant than structures, there is a plethora of methods for predicting functional sites from sequence alignments. The first information extracted from sequence alignments was related with fully conserved positions (Zuckerkandl and Pauling, 1965). Fully conserved positions are related with sites important for the function or the structure of the protein. Later, the concept of conservation was extended to family-dependent conservation: positions that are conserved within subfamilies being the aminoacid type different between different subfamilies. These

*To whom correspondence should be addressed.

family-dependent conserved positions have been related with functional specificity. That is, they are associated with the functional feature which distinguishes the functional subfamilies within the alignment, in contrast to fully conserved positions which are associated to the function which is common to all the proteins in the alignment. There are different approaches for detecting these positions: based on Principal Component Analysis (PCA) and neural network classifiers (Andrade *et al*., 1997; Casari *et al*., 1995), explicit phylogenetic trees (del Sol Mesa *et al*., 2003; Lichtarge *et al*., 1996), the detection of positions correlated with the phylogeny (del Sol Mesa *et al*., 2003; La *et al*., 2005) and others (Bickel *et al*., 2002; Livingstone and Barton, 1993). Conservation and family-dependent conservation are sometimes combined with structure information to restrict the predictions to the positions with the structural characteristics expected for a functional site (Aloy *et al*., 2001; Armon *et al*., 2001; Glaser *et al*., 2006; Kinoshita and Ota, 2005; Landgraf *et al*., 2001; Yu *et al*., 2005).

These methods for locating family-dependent conserved positions do not take a functional classification as input but they use the one implicit in the alignment. Hence, their assumption is that the functional classification of the proteins coincide with the sequence-based classification represented by the alignment. According to the accepted scenario of divergent evolution to function, this should be the situation in most of the cases. Nevertheless, one can imagine certain specific situations where there is a disagreement between the alignment-based classification and the functional classification of the proteins. Many functional and structural requirements 'push' together the evolution of a protein family, but only one phylogeny can be observed, which arises from a combination of all the different functional constraints. Hence, the specific divergence owing to a function we are interested in can be masked within this composite phylogeny. Another situation which could result in a function/phylogeny disagreement is when the alignment does not reflect the true phylogeny, e.g. in structural alignments linking distant proteins for which much of the sequence information relating the proteins has been lost (i.e. SH3 domains). There are not many methods which can incorporate an external functional classification. Mirny and Gelfand (2002) developed a method which uses information on orthology/paralogy to define the functional subfamilies in the MSA which can naturally use an external functional classification. Hannenhalli and Russell developed a method based on the comparison of subtype-specific sequence profiles which allows the user to impose an external functional classification (definition of the subtypes) (Hannenhalli and Russell, 2000). Although these interesting works probably constitute the first approaches for the phylogeny-independent detection of functional residues, these methods are in many senses exploratory and still have a number of intrinsic drawbacks. For example, they consider the functional classes as disjoint classes and do not have the possibility of incorporating relationships between them (i.e. functional distances or functional hierarchies). Moreover, it is unclear how many of the examples presented correspond to cases where the functional classification does not follow the phylogeny.

In this work we present two new supervised methods for detecting functional sites from multiple protein alignments which can incorporate an external functional classification instead of using the one implicit in the alignment. One of the methods (*Xdet*) can incorporate quantitative information on 'functional similarities' or hierarchical functional classifications in order to detect positions in the alignment related with that functional organization. The other method (*MCdet*) is based on a vectorial representation of the alignment on which Multiple Correspondence Analysis (MCA) is used to locate the residues which better follow the pattern of presence/absence of a given function. We tested the methods in different scenarios representing different degrees of coincidence between the functional and the phylogenetic classifications, and where that disagreement arises from different causes.

MATERIALS AND METHODS

Xdet method

This method is intended to locate positions in a multiple protein alignment which are related to the functional classification of the proteins, ideally when the functional classes can be related by a hierarchy, or distances between them can be defined. The idea is that, in these positions, a sharp amino acid change between two proteins would be related with a high functional difference between these proteins, and the other way around.

A schema of the *Xdet* method is shown in Figure 1. For each position in the alignment, a matrix quantifying the amino acid changes for all pairs of proteins is constructed based on a substitution matrix. In this matrix, a given entry represents the similarity between the residues of two proteins at that position. An equivalent matrix is constructed from an external explicit functional classification where each entry represents the 'functional similarity' between the corresponding proteins (for the functional feature we are interested in). These two matrices are compared with a Spearman rank-order correlation coeficient (Press *et al*., 1992). So, for a multiple alignment of N proteins of leght L, being A_{ijk} the similarity between the amino acids of proteins i and j at position k (see below), and F_{ij} the functional similarity between proteins i and j (see below), the score for position k is calculated as

$$r_k = \frac{\sum_{i,j} (A'_{ijk} - \bar{A'}) \cdot (F'_{ij} - \bar{F'})}{\sqrt{\sum_{i,j} (A'_{ijk} - \bar{A'})^2} \cdot \sqrt{\sum_{i,j} (F'_{ij} - \bar{F'})^2}},$$

where A' and F' are the ranked values of A and F respectively [ties being assigned midranks (Press *et al*., 1992)]. $\bar{A'}$, $\bar{F'}$ are the corresponding average values of these ranked matrices. Positions with >10% gaps are excluded from the calculations, and 0 is used as the similarity between any amino acid and a gap.

Positions with high r_k values are the ones for which similarities between amino acids are correlated with the functional similarities between the corresponding proteins, and hence are predicted as the ones related with functional specificity. P-values for these scores are obtained using a background distribution of random scores generated from 1000 random protein-function assignments using the same alignment.

As a measure of 'functional similarity' between proteins (F) very different metrics can be used, depending on the problem we are dealing with and the associated definition of 'function': chemical similarity between ligands, metrics for measuring similarities in hierarchical functional classifications like the ones implicit in *Gene Ontology* (Harris *et al*., 2004) or EC, functional hierarchies based on expert knowledge, similarity between enzyme functional parameters (K_{cat}, K_m, ...), and so on. In the basic case where the functional classification does not have associated (quantified) functional similarities, one can just use '1' and '0' for representing similarities between proteins belonging to the same or different functional classes respectively. We show examples of some of these different metrics of functional similarity in this work. Similarly, we can use different metrics of similarity between amino acids (A), either any of the available substitution matrices or the identity matrix (0,1), depending on whether we expect the function we are studying to be related with conservative or non-conservative changes.

MCdet method

This method is based on a simultaneous vectorial representations of sequences, residues and functions on related spaces. This allows to study

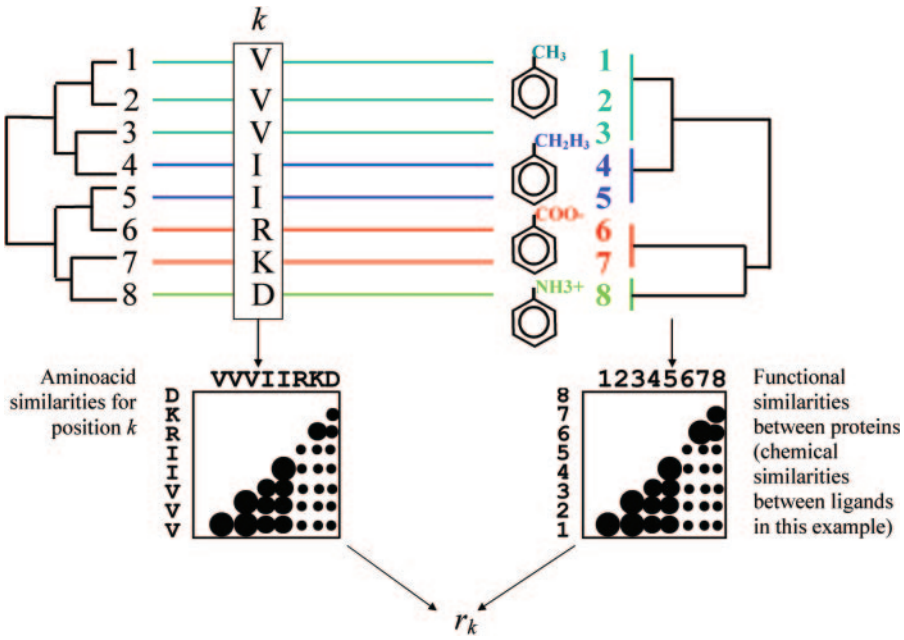

Fig. 1. Schema of the *Xdet* method. A sequence alignment of eight proteins is depicted. The implicit phylogeny based on the sequence relationships extracted from the alignment is shown on the left. The colors of the proteins represent a functional classification which is not reflected in the phylogeny. In this case, this function is the binding of an effector (small molecule) chemically slightly different for the different members of the family (right). A functional similarity can be defined between the proteins for this particular function (i.e. the chemical similarity between the effector they bind). These similarities could be represented in a tree-like structure or a hierarchy (rightmost tree). To asses whether a given position in the alignment is related with that particular functional hierarchy, a matrix containing all the amino acid changes occurring at that position is constructed and compared with an equivalent matrix containing the functional similarities between the proteins previously defined. The values of these two matrices are depicted here as circles with a radius proportional to the similarity value.

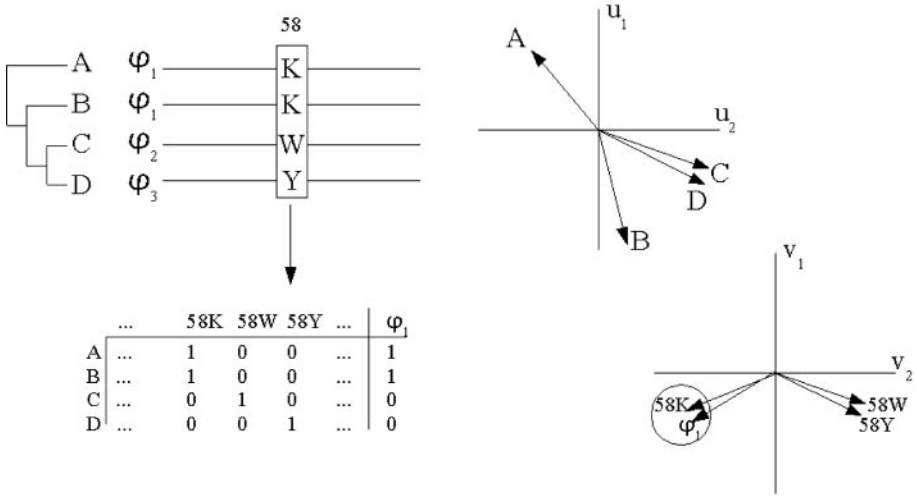

Fig. 2. Schema of the *MCdet* method. A MSA of four sequences (A, B, C and D) is depicted. The functions of the proteins (Φ1, Φ2 and Φ3) are not related with the alignment-based phylogeny (depicted on the left). MCA is applied to a binary representation of the sequences and the functions of the proteins (matrix at the bottom) generated as explained in Materials and Methods. This procedure leads to related spaces (defined by the eigenvectors of the MCA treatment $-u_i$ and v_i-) where the vectors of positions important for functional specificities lay close to the vectors of the functions they determine (right).

the relationships between these three sets using vector analysis techniques, MCA in this case.

MCA is based on correspondence analysis (Greenacre, 1984), a multivariate descriptive statistical technique that can be viewed as an equivalent to PCA when dealing with qualitative data, so that objective numerical values can be assigned to them (Greenacre, 1984; Lebart *et al*., 1984).

A schema of the *MCdet* method is shown in Figure 2. Given a multiple sequence alignment (MSA) of N sequences and L positions, a data matrix W of dimension $N \times Q$ (where $Q = 21L$) is constructed representing each position l in the alignment as a complete disjunctive category with 21 different modalities (representing the 20 amino acid types plus the gap) just coding the presence of a modality with '1'and its absence with '0'.

Columns in X without any '1' are removed for subsequent consistency without loss of generality, resulting in a matrix X of dimension $N \times P$, where $P < Q$. Given the data matrix X as defined above, with general term x_{ij}, let us define the following frequencies:

$$x_{nS} = \sum_p x_{np} \quad x_{Sp} = \sum_n x_{np} \quad x_{SS} = \sum_n \sum_p x_{np}$$
$$f_{ns} = x_{ns}/x_{ss} \quad f_{sp} = x_{sp}/x_{ss} \quad f_{np} = x_{np}/x_{ss}.$$

Let Y be the matrix with the general term $y_{np} = f_{np}/(f_{Sp}\sqrt{f_{nS}})$. Y is a transformed data matrix in which considering Euclidean distances between column vectors will be equivalent to considering χ^2 distances within the original data matrix X.

Let Z be the matrix with general term $z_{np} = f_{np}/\sqrt{(f_{nS} \cdot f_{Sp})}$ and Z' its transpose. The space generated by the eigenvectors of ZZ' provides a proper decomposition of the sequences-residues association between its sources of variation (Peña, 2002).

The next step is to project the columns of matrix Y into the space generated by the eigenvectors of matrix ZZ'. Let v_k be the k-th eigenvector associated with the k-th non-null eigenvalue λ_k of matrix ZZ' (excluding the trivial solution $\lambda = 1$). The coordinates of residue 'p' in factor 'k'of the space of sequences is

$$c_{pk} = \sum_n \frac{v_{kn} \cdot f_{np}}{f_{Sp}\sqrt{f_{nS}}}$$

Let $\Phi = (\phi_1, \phi_2, \ldots..., \phi_N)$ be a binary column vector $N \times 1$ coding for a subfunction Φ in such a way that its general term ϕ_n is '1' if sequence 'n' has function Φ, or '0' otherwise. In MCA, the vector Φ can be readily projected as a supplementary column into the space generated before, together with residues 'p', so that it will tend to be projected closer to those residues having the same presence/absence profile along the whole sequence population. We will predict those residues to be responsible of conferring that subfunction to a given sequence.

The coordinates of a supplementary column Φ in factor 'k' of the space of sequences is

$$c_{\Phi k} = \frac{1}{\sum_n \Phi} \sum_n (\Phi_n \cdot \frac{v_{kn}}{\sqrt{f_{nS}}}).$$

Therefore, the candidates that are responsible for the function Φ will be determined by those 'p' that minimize the Euclidean distance to Φ, i.e.

$$d(p, \Phi) = \sqrt{\sum_k c_{pk} \cdot c_{\Phi k}},$$

where k is the maximum number of no-null eigenvalues of ZZ'.

When calculating distances $d(p,\Phi)$, we consider in the analysis the whole set of eigenvectors k which accounts for an explained variance of 100% (note that for this particular application of MCA we are not interested in dimensionality reduction).

P-values are associated to these distances in the same way as described for *Xdet*.

Examples

We tried these two methods in different sets of aligned proteins for which we carefully checked that the functional classification is not implicit in the alignment. These examples illustrate different real scenarios where supervised methods should be applied. They cover different degrees of overlap between the phylogenetic and the functional classification, different definitions of function, and different ways of quantifying functional similarities.

Ras oncogene structural homologs We started from the structural alignment automatically generated by the *Dali* program (Holm and Sander, 1994) from the 3D structure of the *Ras* oncogene (PDB id: 1ctqA). This alignment contains proteins binding different ligands, including nucleotides (GTP, FMN, FAD, etc.), nucleosides, sugars, and so on. The alignment was filtered leaving only chains with a structural similarity with the master (1ctqA) higher than 6.0 (ZFSSP score), removing redundancy above 40% sequence identity, and removing structures without bound ligand. The final alignment contains 24 proteins binding different ligands. In this case the disagreement between the classification of the proteins implicit in this alignment and the functional classification (according to the ligand they bind) is mainly due to the fact that structural alignment (plus the redundancy cutoff imposed) is relating remote homologs at very high distances (Fig. 3A).

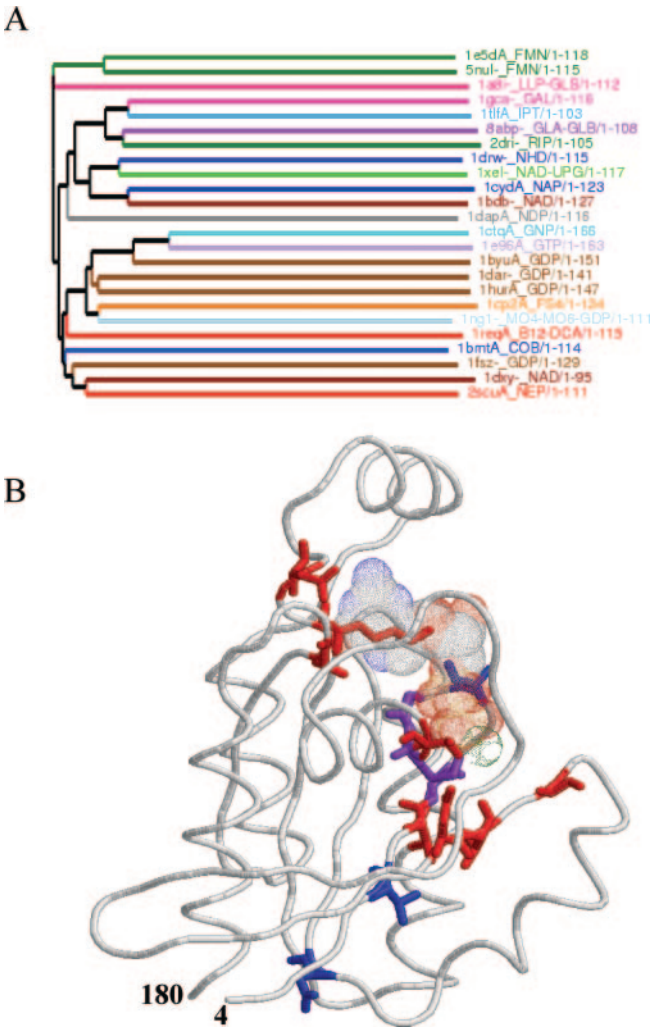

Fig. 3. Results for the *Ras* oncogene structural homologs. (**A**) Phylogenetic tree derived from the structure-based alignment, drawn with *Belvu* (E. Sonnhammer, unpublished data) which implements a neighbour-joining algorithm to calculate the layout of the tree from the alignment. Additional trees generated with Bayesian techniques are available for this and the forthcoming examples in Supplementary Material. Proteins are labelled with their PDB codes and the bound ligands (PDB nomenclature). The tree is colored according to the set of bound ligands. Proteins binding the same ligand could be colored different if the complete set of bound ligands is not exactly the same. (**B**) Predictions of the methods mapped on the structure of the human RhoA (PDB 1ftn), a GTP-binding protein. The bound GDP is shown in Van der Walls representation and colored by atom identity (CPK). The residues predicted by the methods are shown in sticks representation and colored blue for the *Xdet* method, red for *MCdet* and purple for the positions predicted by both methods. The figure was generated with Rasmol (Sayle and Milner-White, 1995).

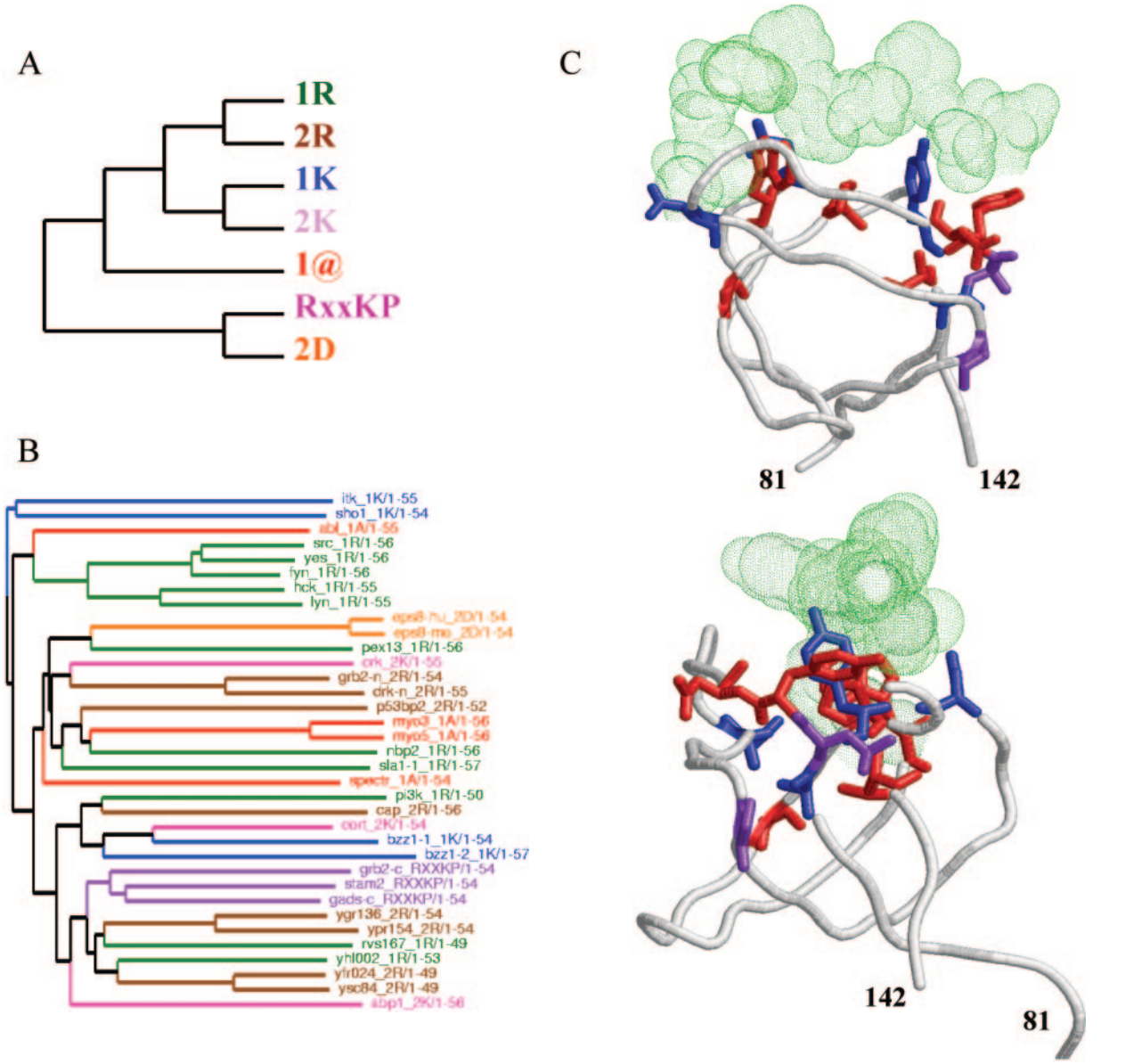

Fig. 4. Results for the SH3 domains. (**A**) Hierarchical functional classification of SH3 functional classes adapted from (Cesareni *et al*., 2002). (**B**) Tree obtained with *Belvu* from the alignment of SH3 domains described in Methods. Proteins are colored according to the functional class they belong to (panel A). (**C**) Two orthogonal views of the predictions of both methods mapped on the structure of Fyn Tyrosine Kinase bound to a synthetic peptide (PDB 1fyn). Color codes as in Figure 3.

As a measure of 'functional similarity' for the *Xdet* method we used the chemical similarity between the bound molecules measured as the Tanimoto coefficient (Holliday *et al*., 2002). We obtained the Tanimoto coefficients for all pairs of ligands from the SuperLigands server (http://bioinf.charite.de/superligands). For pairs where one of the proteins is binding more than one ligand, we took the pairs of ligands with highest similarity. Solvent and non-functional small molecules are excluded. All the guanine nucleotides (GTP, GDP, GNP) are considered as a single class ('GXP') since the binding properties of the proteins they are bound to are identical and the fact that they are binding a particular guanine nucleotide is due to factors extrinsic to the protein (crystallization with artificial non-hydrolysable nucleotides, etc). So, for example, two proteins binding GTP and GNP respectively are considered here as 100% functionally identical.

MCdet was used to predict the residues responsible for this GXP function (see Results) since that is the one with enough representatives in the alignment. Sufficient representatives of the target function is a crucial requirement for the reliability of the predictions of this method (see Discussion).

SH3 domains SH3 domains are peptide recognition modules which bind to some Prolin-rich motifs in proteins. Despite having a common evolutionary origin and a similar overall structure, they are very divergent in sequence. SH3 domains can be grouped in different functional classes depending on the peptide they bind (Cesareni *et al*., 2002). These domains are very divergent and a phylogenetic tree based on this alignment does not reflect the functional classification (Fig. 4A and B).

The functional similarities between the classes for feeding the *Xdet* method were obtained from a hierarchical functional classification of SH3 domains developed by experts (Cesareni *et al*., 2002) (Fig. 4A). In the context of the *Xdet* method, the functional distance between two classes is just the number of branches between these classes and the first internal node they converge in the functional tree. The functional similarity (sim) is obtained from the functional distance (dist) as sim = max_dist − dist, where max_dist is the maximum distance (4 in this case). For example, sim(1R,1R) = 4 (highest); sim(1R,2R) = 3, sim (1R,2D) = 0, etc.

The *MCdet* method was used to predict the residues associated with the 1R class.

Structural alignment of TIM-barrel hydrolases We started from the structural alignment automatically generated by the *Dali* server for the 1qumA structure and filter it leaving only ZFFSP $\geq$ 7.0 and %seq_id $\leq$ 30%. This structural alignment contains TIM-barrel structures, most of them enzymes belonging to different EC classes. We restricted the alignment to the hydrolases (EC: 3.2.1.*). We ended up with 20 sequences belonging to 10 subclasses of hydrolases (3.2.1.1; 3.2.1.2; 3.2.1.35; ...). This alignment contains long, unrealistic distances between proteins owing to remote homology. In this case we used binary functional similarities for the *Xdet* method: sim(*A*,*B*) = 1 if *A* and *B* belong to the same hydrolase subclass, and 0 otherwise. The *MCdet* method was used to predict the positions responsible for the 3.2.1.1 subclass (the one with enough representatives in the alignment).

Lactate/malate dehydrogenases This family of homologous enzymes encompasses two main functional subfamilies (EC: 1.1.1.27 and 1.1.1.37), acting on lactate and malate respectively.

We started from the Pfam (Bateman *et al*., 2004) alignment PF00056 ('lactate/malate dehydrogenase, NAD binding domain'). This domain covers residues 1–145 of the *Escherichia coli* malate dehydrogenase. From this Pfam alignment we removed redundancy >80% seq. id., ending up with an alignment of 46 proteins.

The phylogeny-function disagreement in this case arises because there is a group of malate dehydrogenases which is clearly more similar to the lactate dehydrogenases than to the rest of malate dehydrogenases (Fig. 6A).

As in the previous example, for the *Xdet* method we used a binary functional similarity (1 for pairs of proteins belonging to the same class and 0 for the rest).

RESULTS

Ras oncogene structural homologs

This example represents a case where a set of proteins is related by structural alignments. There is a disagreement between the classification of the proteins implicit in the alignment and the functional classification (Fig. 3A). It also illustrates the cases where the functional classification is based on the ligand bound to the protein and where 'functional similarities' between proteins can be quantified from the chemical similarities between these ligands.

Even for groups for which there is an overall good agreement between the functional and the phylogenetic classification (like GXP: GTP, GDP, GNP, etc) there are prominent exceptions, like the FtsZ cell division protein (PDB 1fsz), which is a GTPase far from the GXP group.

The residues responsible for functional specificity predicted by both methods for the structural neighbours of the Ras oncogene (see Methods) are shown in Figure 3B, mapped in the structure of a GTP-binding protein, the human RhoA (PDB 1ftn). The *MCdet* method was used in this case to predict the residues responsible for the 'guanine nucleotide binding' function (GXP). *Xdet* is designed to predict positions with a 'global' importance for conferring binding specificity, instead of being related with a particular ligand. The 11 residues closer to the vector representing the G × P function according to *MCdet* and the six residues with highest correlation with the matrix of functional distances according to the *Xdet* method are shown. Three residues are common to both methods. The predictions of both methods clearly cluster around the bound nucleotide (GDP in this example). *MCdet* predicted residues mostly close to the Guanine and the phosphate groups. All the *Xdet* predictions close to the ligand point to the phosphate groups, reflecting that this is the region conferring 'global' specificity for ligand binding in this family of structural neighbours (the region where the ligands are more different). Interestingly, the predictions of both methods extend a little bit beyond the last (β) phosphate of the bound GDP, in the region where the third phosphate (γ) should go in the GTP form of the molecule. There are some predictions far from the GDP for which we do not have an obvious explanation (possible false positives), like V9 and D78 for *Xdet*, or T60 and G62 for *MCdet*.

SH3 domains

This second example illustrates a case where relationships between proteins are based on remote homology. The functional classification is based on expert knowledge (G. Cesareni *et al*., 2002 and personal communication). We try to illustrate how such a complex 'ontology-based' classification can be used to quantify functional similarities. The intrinsic complexity of this functional classification (and its eventual drawbacks) maybe is leading to some disagreement between the functional and the phylogenetic classification, although it is clear that the remote homology is mainly responsible for it. SH3 domains are the prototypical case of remote homology were classical sequence-based methods are difficult to apply.

Figure 4B shows the phylogenetic tree generated with the neighbour-joining method implemented in *Belvu* (E. Sonnhammer, unpublished data) from the alignment of SH3 domains described in Methods. It can be seen that the phylogeny do not account for the different functional subtypes. This function/phylogeny disagreement is also present in more sophisticated Bayesian trees (Supplementary Material).

Figure 4C shows the predictions of both methods mapped on the structure of the SH3 domain of the Fyn Tyrosine Kinase bound to a synthetic peptide (PDB 1fyn). The *MCdet* method was used to predict the residues responsible for the '1R' functional specificity. The sets of residues predicted by the two methods clearly follow the bound peptide, and they are specially concentrated in its terminal ends. Binding specificity of SH3 ligands is known to reside mainly in the variable ends, the central part being more conserved (conserved Prolins) across all the ligands (Cesareni *et al*., 2002). The predictions extend a little far beyond the C-term of the peptide. Actually, the two residues in common between the predictions of both methods are there. This could indicate that this region is also important when binding natural substrates (longer proteins). Some of the predicted residues, like Y137 (highest score in *Xdet*) and L90, have been extensively described in the literature as important for determining specificity (Cesareni *et al*., 2002).

TIM-barrel hydrolases

The third example covers the other extreme of the sequence relations. In this case it cannot be discarded that some of the structural similarities are the consequence of a process of convergent

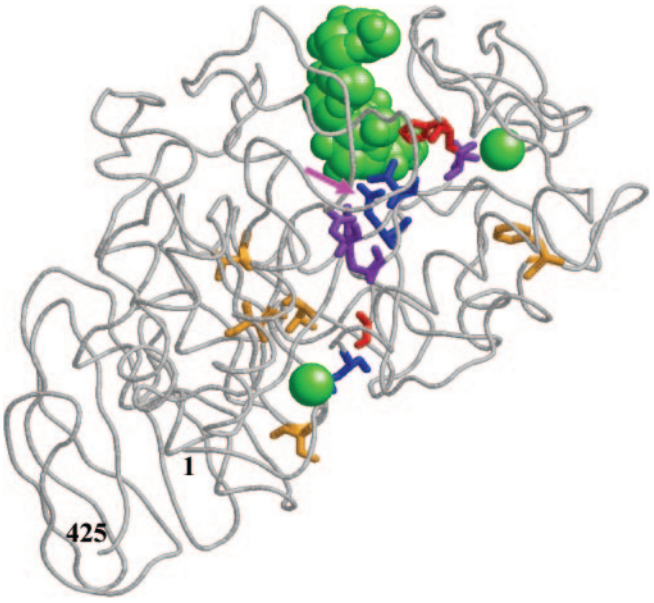

Fig. 5. Results for the TIM-barrel hydrolases. The predictions are mapped on the structure of the *B.subtilis* α-Amylase complexed with maltopentaose (PDB 1bag) (Fujimoto *et al*., 1998). The bound maltopentaose and the two calcium ions are colored green. The color code for the predicted positions is the same as for Figures 3 and 4. The pink arrow marks the E208Q mutant (Fujimoto *et al*., 1998). Additionally, the residues predicted by the MB-method (del Sol Mesa *et al*., 2003) based on the phylogeny implicit in the alignment are shown in sticks and colored yellow (except D97 which is also predicted by *Xdet* and *MCdet* and hence colored purple).

evolution. The disagreement between the alignment-based and the functional classification comes from these distant relationships. In this case, the functional classes are enzymatic activities differing in their substrate specificity (last EC number) that allows a simple quantification of the distances in a binary form.

The *Xdet* method was used to predict positions with a global importance for differentiating the various 3.2.1.* classes. The *MCdet* method was used to predict the positions specifically responsible for the 3.2.1.1 function, in contrast to the rest of 3.2.1.*.

Figure 5 shows the residues predicted by each method mapped on the structure of a *Bacillus subtilis* α-amylase (PDB 1bag) (Fujimoto *et al*., 1998). There are three residues in common between the five residues predicted by the *MCdet* method and the six predicted by *Xdet*. Of the 11 residues, 9 predicted by both methods cluster in the active site of the protein, indicated in this case by the presence of a bound maltopentaose. The second highest-scoring residue reported by *Xdet* (208) is mutated in the original paper reporting the 1bag 3D structure (E208Q) (Fujimoto *et al*., 1998) to demonstrate its involvement in the active site of this protein. Two residues (D171 and G172), independently predicted by *Xdet* and *MCdet* respectively, are relatively far from the active site but clustered together. Still, D171 is coordinating a calcium ion indicating a possible functional role in this protein, although such possibility is not described in the original publication (Fujimoto *et al*., 1998).

To illustrate the difference between these supervised methods and the ones which rely on the phylogeny represented by the alignment to detect positions responsible for functional specificity, we apply the MB-method (del Sol Mesa *et al*., 2003) to the same multiple alignment. That approach is methodologically similar to *Xdet* (see Discussion) but it uses the functional classification implicit in the alignment instead of an external one. As expected, the equivalent 6 residues predicted by that method with highest correlation values are not clustering around the active site (Fig. 5). Only one residue predicted by MB-method (D97) is also predicted by *Xdet* and *MCdet*. This illustrates the importance of using methods guided with the actual functional information (supervised) for cases where it is suspected that the alignment is representing unrealistic phylogenetic distances (alignments based on remote homology, etc.)

Lactate/malate dehydrogenases

For the last example we present a case where the proteins are closely related, the sequence relations are very clear and the alignment can be obtained with any common MSA program. As in the previous case, the functional classes are enzymatic activities differing in the substrate specificity (last EC number) and the functional distances between them are defined in a binary way. This family of proteins has already been used by Hannenhalli and Russell to test their method (Hannenhalli and Russell, 2000).

Figure 6A shows the phylogenetic tree obtained from the MSA of the lactate/malate dehydrogenases (LDH, MDH) described in Methods. It can be seen that there is a group of MDHs (let's call them MDH') which is closer to the LDHs than to the rest of MDHs. This function/phylogeny disagreement is also present in a Bayesian tree generated from the same Pfam alignment (Supplementary Material). An unsupervised method would be, to some extend, confused by this 'non-functional' phylogeny and would try to locate residues common to LDH + MDH' and different from MDH, which would obviously not reflect the malate versus lactate specificity.

Figure 6B shows the predictions of both methods for this alignment mapped on the structure of the *Aquaspirillium arcticum* malate dehydrogenase (PDB: 1b8u). For the *MCdet* method we show the intersection between the positions predicted to be specific for malate and the ones predicted to be specific for lactate. These should correspond to positions which tend to be conserved within the MDHs and within the LDHs but with the amino acid type being different in these two functional classes. It can be seen that most of the positions independently predicted by the two methods are relatively close to the active site of the protein marked by the bound NAD and oxalacetate. These predictions include two of the six positions detected by Hannenhalli and Russell with their supervised method (Hannenhalli and Russell, 2000): 95 and 99 in 1b8u (102 and 107 in their article). The other four positions predicted by Hannenhalli and Russell are outside the Pfam-based alignment used here, which only covers the N-terminal domain of this family. Position 95 is detected by both, *MCdet* and *Xdet* (2nd highest *Xdet* score). This position is R in MDH and Q in LDH (Fig. 6B) and there is plenty of experimental information demonstrating its importance on determining the lactate/malate specificity (Hannenhalli and Russell, 2000). In this example, the relationship between specificity-determining residues and closeness to the substrate is not very obvious since even the experimentally determined position 95 is not at contact distance with the substrate (Fig. 6B).

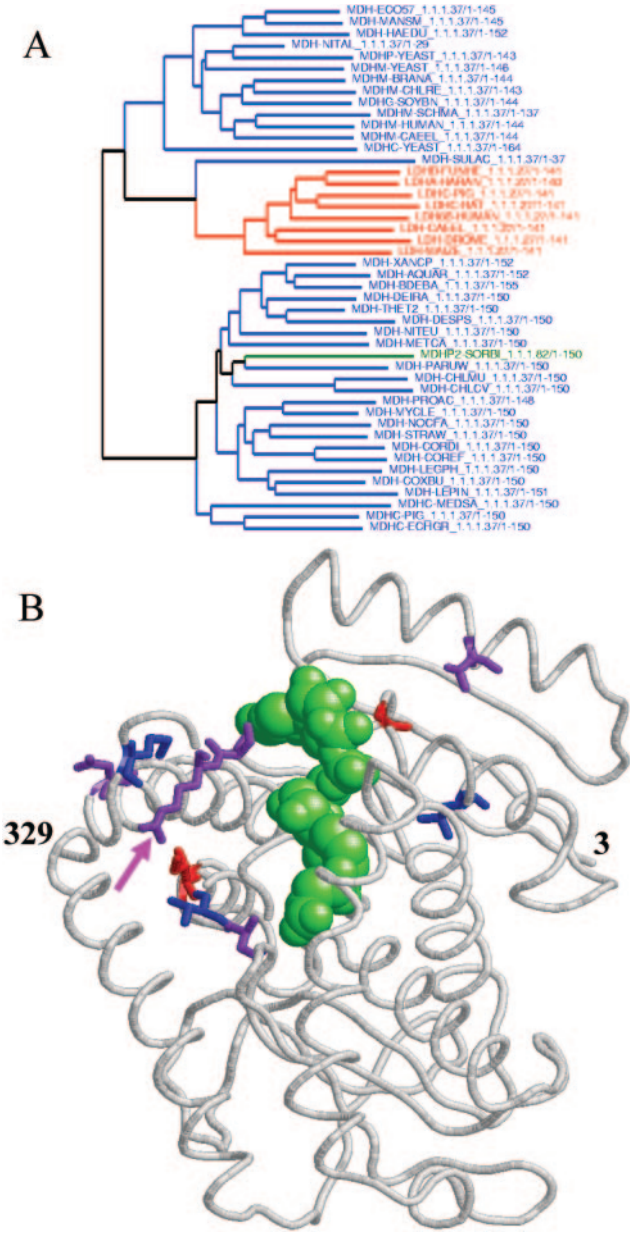

Fig. 6. Results for the Lactate/malate dehydrogenases. (**A**) Phylogenetic tree generated with *Belvu* from the MSA of the lactate/malate dehydrogenases described in Methods. MDHs are colored blue while LDHs are colored red. The green protein is a MDH which uses NADPH instead of NADH as cofactor. (**B**) Positions predicted by both methods mapped on the structure of the *Aquaspirillum Arcticum* malate dehydrogenase (PDB 1b8u). Color code as in Figures 3–5. The bound NAD is colored green. Position 95 (R in MDH; Q in LDH) is marked with a pink arrow.

DISCUSSION

In this work we present two new supervised methods for the detection of positions responsible for functional specificity from protein alignments. Since they are informed with an external functional classification, they are intended to be applied when this functional classification is not the one implicit in the alignment. We show four examples that cover the full range of sequence relations, from proximal to purely based on structural alignments. For each one of them we carefully checked that the functional classification is in disagreement with the phylogeny-based classification. These examples cover different degrees of overlap between the phylogenetic and the functional classification, different definitions of function, and different ways of quantifying functional similarities.

These two methods are complementary, between them and with existing approaches, since they have their own advantages and limitations. The advantage of *Xdet* with respect to *MCdet* and other existing supervised methods is that it is the first method that can naturally incorporate quantitative information on functional hierarchies and/or functional similarities, compared with the other methods which can only work with 'disjoin' functional classes not related by distances or hierarchies. This is important due to the complexity of the 'protein function' phenomenon which require complex hierarchical functional classifications and ontologies to be represented (Harris *et al*., 2004). Another advantage is that it does not require many examples of the different functional classes to work. Its main disadvantages are that it predicts residues with a 'global' importance for defining the classes (it is not designed to locate residues responsible for one of the classes), and that it is expected to work better when a rich functional classification (many classes related by a rich functional hierarchy) is available. The *Xdet* approach is methodologically similar to our previously described MB-method (del Sol Mesa *et al*., 2003) in which the matrix with the functional similarities was substituted by a matrix containing the percentages of sequence identity between the proteins. So, the assumption of the MB-method is that the functional classification is the one implicit in the alignment (reflected in the sequence similarities between the proteins). Hence, the *Xdet* method uses a similar technology with a different biological goal: to account for cases like the ones discussed in this paper, for which the functional classification is not implicit in the phylogeny.

The advantage of the *MCdet* method is that it can detect positions conserved within one subfamily but not within others (responsible for the specificity of that subfamily only), while other methods require the positions to be conserved within all subfamilies. On the other hand, Hannehalli and Russell's method has a better estimator for small samples (few sequences) based on Dirichlet mixtures (Hannenhalli and Russell, 2000), while *MCdet* is expected to work optimally in cases with enough sequences for the target function. The advantage of Mirny and Gelfand's method with respect to *MCdet* is its possibility to incorporate information on similarity between amino acids (Mirny and Gelfand, 2002).

These two supervised methods are complementary to the existing unsupervised ones in the range of applicability. For cases where a function/phylogeny disagreement is suspected for the function we are interested in, it does not make sense to apply unsupervised methods, as illustrated with the *Xdet* versus MB comparison for the hydrolase example (Results).

We will see an increasing number of cases for which certain functional features will not be reflected in the phylogeny as we know more sequences and structures. Nevertheless, it is still difficult to automatically find large sets of these examples, and their associated functional classes and annotated functional sites. This makes it difficult to test the methods presented here in datasets large enough to extract statistically meaningful cutoffs or confidence values, and to tune the parameters. We plan to work in that direction

in the future. We also plan to explore in the future the possibility of using the sets of predicted residues to assign protein to functional classes, as done in other works (Hannenhalli and Russell, 2000). For these cases where function and phylogeny do not correlate, the functional assignment cannot be done by the standard methods based on sequence similarity.

In most of the examples presented in this work, the disagreement between the sequence-based and the functional classification comes mostly from the fact that they are based on structural alignments which relate very distant proteins. We think that such cases will become quite frequent in the future as structural genomics projects continue to produce 3D structures which will allow establishing links between proteins which can not be related at the sequence level.

The two methods presented here complement the still limited andscape of approaches for the phylogeny-independent (supervised) detection of functional sites. Together with the existing unsupervised and supervised approaches, they can help in the interpretation of the incoming stream of sequences and structures in functional terms.

A careful inspection of the ''TIM-barrel hydrolases'' example shows that, spite there is an overall agreement between the functional and the phylogenetic groups, the specific distances between proteins are long and unrealistic (due to remote homology). This makes methods based on phylogenetic distances to fail, as shown when comparing the results of Xdet and MB-method (Results). This example illustrates the importance of using supervised methods not only when there is an overall function-phylogeny disagreement, but also when the details of the phylogeny (specific distances) are suspected to be wrong.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the members of the Protein Design Group, especially David de Juan for interesting discussion and comments, José M. Gonzalez for the data on the SH3 domains and José M. G. Izarzugaza for the help with Bayesian trees. F.P. and A.R. are the recipients of a 'Ramón y Cajal' contract and a FPI fellowship (BIO2004-00875) respectively, both from the Spanish Ministry for Education and Science. This work has been partially funded and funding for the Open Access publication charges was provided by the *GeneFun* (LSHG-CT-2004-503567) and *BioSapiens* (LSHC-CT-2003-505265) EU projects.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. *et al.* (2001) 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.*, **311**, 395–408.

Andrade,M.A. *et al.* (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.*, **76**, 441–450.

Armon,A. *et al.* (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.

Bateman,A. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Bickel,P.J. *et al.* (2002) Finding important sites in protein sequences. *Proc. Natl Acad. Sci. USA*, **99**, 14764–14771.

Brenner,S.E. (2001) A tour of structural genomics. *Nat. Rev. Genet.*, **2**, 801–809.

Casari,G. *et al.* (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.*, **2**, 171–178.

Cesareni,G. *et al.* (2002) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett.*, **513**, 38–44.

Del Sol,A. and O'Meara,P. (2004) Small-world network approach to identify key residues in protein–protein interaction. *Proteins*, **58**, 672–682.

del Sol Mesa,A. *et al.* (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, **326**, 1289–1302.

Di Gennaro,J.A. *et al.* (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.*, **134**, 232–245.

Elcock,A. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

Fujimoto,Z. *et al.* (1998) Crystal structure of a catalytic-site mutant alpha-amylase from *Bacillus subtilis* complexed with maltopentaose. *J. Mol. Biol.*, **277**, 393–407.

Glaser,F. *et al.* (2006) A method for localizing ligand binding pockets in protein structures. *Proteins*, **62**, 479–488.

Greenacre,M.J. (1984) *Theory and Application of Correspondence Analysis*. Academic Press, London.

Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, **303**, 61–76.

Harris,M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Holliday,J.D. *et al.* (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen*, **5**, 155–166.

Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.*, **22**, 3600–3609.

Kinoshita,K. and Ota,M. (2005) P-cats: prediction of catalytic residues in proteins from their tertiary structures. *Bioinformatics*, **21**, 3570–3571.

La,D. *et al.* (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins*, **58**, 309–320.

Landgraf,R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Lebart,L., Morineau,A. and Warwick,K.M. (1984) *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, 175, New York.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **6**, 645–756.

Mirny,L.A. and Gelfand,M.S. (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.*, **321**, 7–20.

Mulder,N.J. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.

Ofran,Y. and Rost,B. (2003) Predicted protein–protein interaction sites from local sequence information. *FEBS Lett.*, **544**, 236–239.

Pazos,F. and Sternberg,M.J.E. (2004) Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl Acad. Sci. USA*, **101**, 14754–14759.

Peña,D. (2002) *Analisis de Datos Multivariantes*. McGraw Hill, Madrid.

Porter,C.T. *et al.* (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Sayle,R. and Milner-White,E. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.

Venter,J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Yu,G.X. *et al.* (2005) *In silico* discovery of enzyme–substrate specificity-determining residue clusters. *J. Mol. Biol.*, **352**, 1105–1117.

Zuckerkandl,E. and Pauling,L. (1965) Evolutionary divergence and convergence in proteins. In Bryson,V. and Vogel,H.J. (eds), *Evolving Genes and Proteins*. Academic Press, New York, 97–166.