BMC Bioinformatics

Methodology article

A structural study for the optimisation of functional motifs encoded in protein sequences

Allegra Via* and Manuela Helmer-Citterich

Address: Centre for Molecular Bioinformatics, Dept. of Biology, University of Rome Tor Vergata, Rome (Italy)

Email: Allegra Via* - allegra@cbm.bio.uniroma2.it; Manuela Helmer-Citterich - citterich@uniroma2.it

* Corresponding author

Published: 30 April 2004

BMC Bioinformatics 2004, **5**:50

Received: 05 February 2004

Accepted: 30 April 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/50

Abstract

Background: A large number of PROSITE patterns select false positives and/or miss known true positives. It is possible that – at least in some cases – the weak specificity and/or sensitivity of a pattern is due to the fact that one, or maybe more, functional and/or structural key residues are not represented in the pattern. Multiple sequence alignments are commonly used to build functional sequence patterns. If residues structurally conserved in proteins sharing a function cannot be aligned in a multiple sequence alignment, they are likely to be missed in a standard pattern construction procedure.

Results: Here we present a new procedure aimed at improving the sensitivity and/ or specificity of poorly-performing patterns. The procedure can be summarised as follows: 1. residues structurally conserved in different proteins, that are true positives for a pattern, are identified by means of a computational technique and by visual inspection. 2. the *sequence* positions of the structurally conserved residues falling outside the pattern are used to build *extended* sequence patterns. 3. the *extended* patterns are optimised on the SWISS-PROT database for their sensitivity and specificity. The method was applied to eight PROSITE patterns. Whenever structurally conserved residues are found in the surface region close to the pattern (seven out of eight cases), the addition of information inferred from structural analysis is shown to improve pattern selectivity and in some cases selectivity and sensitivity as well. In some of the cases considered the procedure allowed the identification of functionally interesting residues, whose biological role is also discussed.

Conclusion: Our method can be applied to any type of functional motif or pattern (not only PROSITE ones) which is not able to select all and only the true positive hits and for which at least two true positive structures are available. The computational technique for the identification of structurally conserved residues is already available on request and will be soon accessible on our web server. The procedure is intended for the use of pattern database curators and of scientists interested in a specific protein family for which no specific or selective patterns are yet available.

Background

One major challenge in the post-genomic era is the assignment of function to the enormous number of ORFs derived from newly sequenced genomes [1]. The comparison with databases of protein sequences or families of aligned proteins does not always provide biologically

useful annotation to hitherto uncharacterised protein sequences [2]. Protein function usually imposes tight constraints on the evolution of specific regions of protein structure; residues directly or indirectly involved in a function are often clustered in a short sequence motif (signature, pattern or fingerprint) that is conserved across the different proteins sharing that function. When a motif encoding a specific function matches the sequence of all the proteins sharing the function and no other sequences, its presence in a newly determined sequence can be used to associate that function to the corresponding protein. Many methods have been developed to identify sequence patterns [3-8]. Most of them start from multiple sequence alignments of homologous sequences and aim at identifying conserved regions potentially important for the biology of the aligned proteins. However, structures are more conserved than sequences; in addition, key functional residues always occupy defined positions in the three dimensional space [9]. In some cases, though, such residues are dispersed along the sequence and are difficult to align in a multiple sequence alignment. This observation, together with the increased availability of protein three-dimensional structures, has led to the development of algorithms for the identification, search and comparison of structural motifs. These algorithms can be used to access protein structure databases [10-19]. Many of these techniques allow the identification and comparison of structurally conserved clusters of residues independently on their order and proximity in the sequence. The derived patterns, however, are three-dimensional patterns and cannot be applied to proteins of unknown structure: this imposes a strict limit on large-scale inference of biological functions in the context of proteomics. Many functional motifs obtained from literature and from multiple sequence alignments are collected in the PROSITE database [20] in the form of deterministic patterns or profiles. Many of them only match all the known true positives (i.e. they do not have false negatives or false positives). However, a large number of PROSITE patterns (referred thereafter as "leaky" patterns) select false positives and/or do not select all the proteins known to belong to the family or to share the function associated to the pattern. In other words, they have low sensitivity (ability to detect true positives) and/or low selectivity (ability to detect only true positives). A procedure developed for increasing the sensitivity and specificity of a PROSITE motif would be extremely helpful for protein functional annotation. To this end, we hypothesized that – at least in some cases – the weak sensitivity and/or specificity of a pattern might be due to the absence, in the pattern, of some functionally and/or structurally important residues, that have been missed because they are not at conserved positions in the primary structure vis-à-vis the motif core. Kasuya and Thornton [21] and Jonassen *et al*. [22] show that structural information improves the ability of a PROSITE pattern to discriminate true from false positive matches. This happens because the structural requirements for the function add constraints to the protein sequence. Jonassen *et al*. search the Protein Data Bank (PDB) [23] sequences with 'softer' variants of each PROSITE pattern and then use structure constraints to reject false positives (i.e. they reject matches whose structure cannot be superimposed onto the three-dimensional fragment corresponding to the PROSITE pattern). Functional sites usually are among the best-conserved parts of a protein structure [24]. Amino acids found in the same spatial positions even in distantly-related protein structures that share a function are likely to be involved in that function. We describe here a procedure that allows the identification of structurally conserved residues for the construction of new sequence motifs. To this aim, we superimpose protein structures that are true positives associated to a PROSITE 'leaky' pattern and identify conserved amino acids that might be structurally and/or functionally relevant. The identification of such conserved residues relies on both a computational and visual analysis of protein surface regions likely to be involved in protein function. Such putative functional surface regions are defined as the solvent exposed residues located in a region close to the set of residues belonging to a PROSITE pattern. The sequence positions of structurally conserved residues NOT belonging to a PROSITE pattern are added to the PROSITE original pattern positions to derive a new *extended* pattern (figure 1). New *extended* patterns are then tested. When necessary they are made less stringent and then optimised, for false and negative matches, by scanning the SWISS-PROT database [25]. The procedure is applied to eight 'leaky' patterns chosen as test cases from the PROSITE database. In seven out of the eight PROSITE patterns analysed, the new *extended* patterns display a greater discriminating power than the original ones.

Results

Criteria for the procedure application

The procedure can be applied to PROSITE patterns, or to any other functional motif, matching at least two true positive sequences for which the structure is known. Links between PROSITE pattern entries and true positive structures can be established as explained in Methods. To avoid biases in the procedure for identifying structurally conserved residues, only matches on non-redundant true positive structures are to be considered. To test the method, a non-redundant set of PDB (Protein Data Bank) [23] chains was compiled as described in Methods. Other criteria to establish non-redundancy can be adopted, however.

The procedure

The true positive structures of a "leaky" PROSITE pattern are used to build a multiple 3D alignment by superimpos-

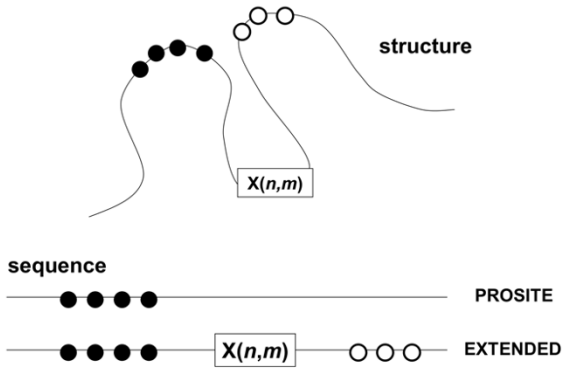


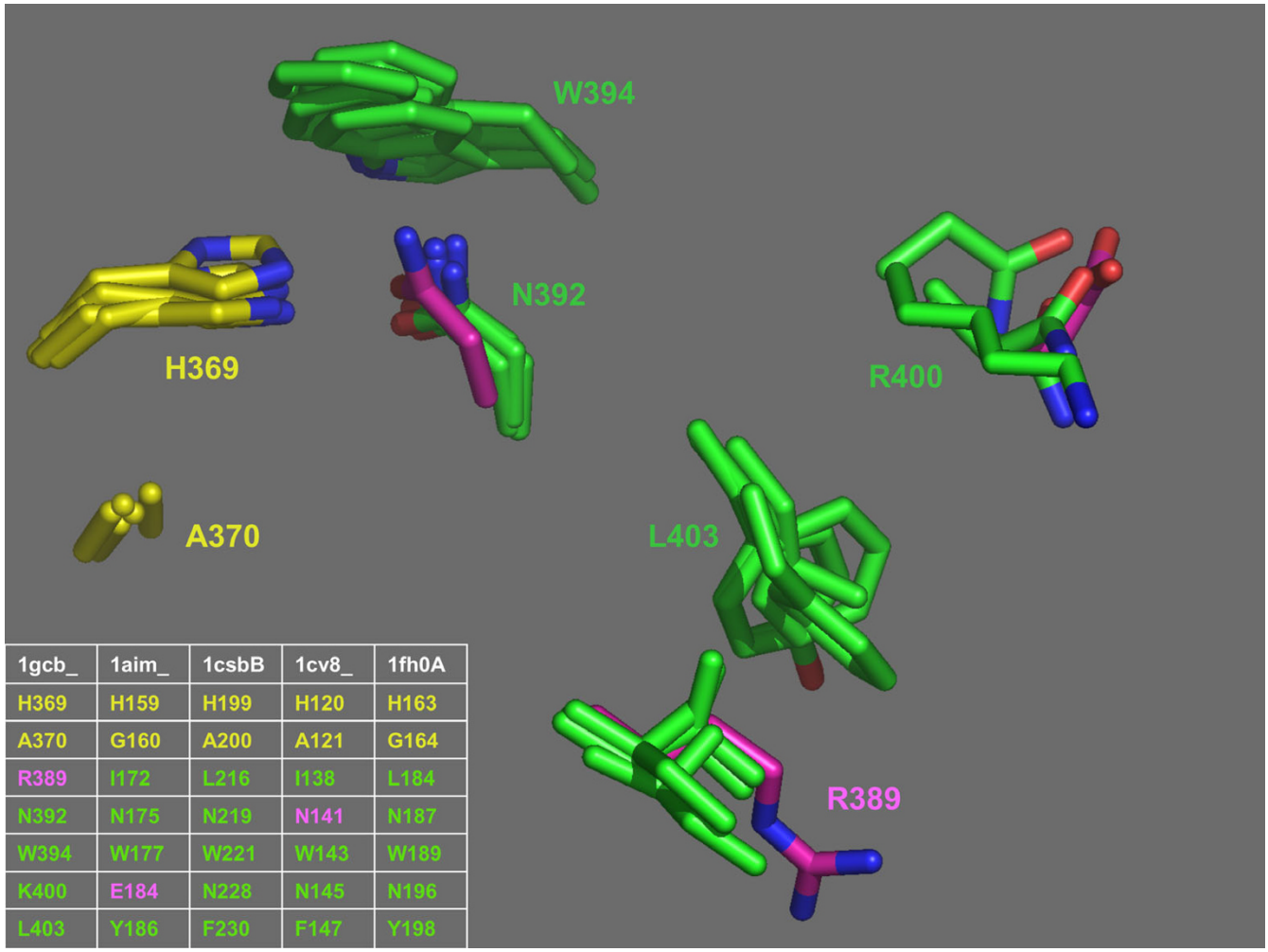
Figure 1
Functional residues localised in space and dispersed along the sequence. Functional residues localised in space can be dispersed along the sequence. Residues belonging to a PROSITE pattern are represented as black-filled circles. Residues represented in white are spatially close to the PROSITE residues but often positioned at variable distances from them in the different structures analysed. Under certain circumstances, the sequence positions (in white) of the structurally conserved residues can be added to the PROSITE pattern positions (in black) in order to obtain a new *extended* sequence pattern with an increased ability of discriminating between true and false positives, than the PROSITE original one.

	1kmmC	1aszA	1e24A	1pysA	1qf6A
1.50 R	R311	R531	R480	R321	R520
1.50 G	G308	G528	G477	G318	G516
1.50 F	F125	F338	F274	F216	F379
1.50 R	R113	R325	R262	R204	R363
1.40 E	E310	E530	D479	E320	E519
1.38 L	L15	L547	L496	F335	F265
1.38 W	R113	F324	R262	R204	R363
1.36 F	Y14	L547	L496	F335	F265
1.28 F	Y122	M335	F274	F216	F379
1.20 R	v	R531	R480	R321	R520
1.20 R	S278	R333	R420	R321	W478
1.20 F	v	F548	F497	F336	F265
1.20 R	R311	v	R480	R321	R520
1.20 R	R311	R531	v	R321	R520
1.20 E	E310	E530	v	E320	E519
1.20 G	G308	G528	G477	G318	v

Figure 2
Example of a Heavy Elements Table (HET). The Heavy Elements Table (HET) obtained by applying the 3D profile method to the set of structurally aligned TP structures corresponding to the PROSITE AA_TRNA_LIGASE_II_2 pattern. Only the first representative lines of the HET are reported. Capital letters indicate the residue type in one-letter code while the number corresponds to the residue number in the PDB file. The 'v' symbol indicates a lacking residue in a cell. Residues belonging to the AA_TRNA_LIGASE_II_2 PROSITE pattern are underlined. The number at the beginning of each line is the profile score of the corresponding cell and the letter represents the *consensus* amino acid in that cell.

ing the pattern's conserved residues. A computational analysis [18] of the surface area in the proximity of the structurally aligned fragments is then performed, aimed at identifying a number of residues conserved in all the superimposed structures. These conserved residues (called Heavy Elements) are placed in a table called HET (Heavy Elements Table, see Methods and figure 2). Subsequent visual inspection of the structural alignment is used to identify one or more conserved residues missed by the computational procedure and/or to discard one or more residues erroneously identified as conserved (see figure 5). The final list of conserved residues is inserted in a table called R-HET (Refined Heavy Elements Table, see Methods, figure 3 and figure 4). In principle, a R-HET contains both residues which belong to the PROSITE pattern analysed and residues which do not. Such residues are added to the PROSITE 'leaky' pattern positions, giving rise to a 'rough' *extended* pattern. A 'rough' *extended* pattern is composed of $(n + m)$ positions, where n represents the number of the PROSITE pattern positions and m is the number of structurally conserved residues, which do not belong to the PROSITE pattern. A 'rough' pattern is then tested on the SWISS-PROT sequence database: the positions NOT belonging to the PROSITE moiety of the 'rough' *extended* pattern are gradually 'softened' (see Methods) in order to obtain an *extended* pattern, matching all the true positives and less false positives in the SWISS-PROT database than the PROSITE original pattern. Also the PROSITE positions of the *extended* pattern are then 'softened', with the aim of detecting previously missed true positives. As a measure of pattern performance, the number of true positives (TP), false positives (FP) and false negatives (FN) on the SWISS-PROT database are used. To compare the performance of the PROSITE with the different *extended* patterns, sensitivity (S_n), selectivity (S_l), specificity (S_p) and correlation (C) values of each pattern are calculated.

Application of the procedure to a test set of eight PROSITE patterns
The procedure was applied to eight PROSITE patterns (see additional file 1 and additional file 2), fetching at least three true positive structures in the non-redundant PDB dataset, and characterised by a low level of specificity and/or sensitivity (see Methods).



1gcb_	1aim_	1csbB	1cv8_	1fh0A
H369	H159	H199	H120	H163
A370	G160	A200	A121	G164
R389	I172	L216	I138	L184
N392	N175	N219	N141	N187
W394	W177	W221	W143	W189
K400	E184	N228	N145	N196
L403	Y186	F230	F147	Y198

Figure 3
Residues conserved in space in a test case and the corresponding R-HET. The R-HET corresponding to the THIOL_PROTEASE_HIS pattern (table on the left) and the 3D superimposition of the corresponding residues. The side chains of the residues belonging to the THIOL_PROTEASE_HIS PROSITE pattern are represented in yellow. Other colours are used to represent the side chains of the residues not belonging to the PROSITE signature, whose 3D position is conserved across the different structures. In the case of the G160 and G164 residues, corresponding to A370, the Cα is represented (as a sphere) instead of the side chain. Residues identified by the computational procedure are in green. Residues added to the R-HET by visual inspection are in magenta. For the figure, residues belonging to the R-HET only, were structurally re-aligned.

1) In six out of the eight examined PROSITE patterns (AA_TRNA_LIGASE_II_1, AA_TRNA_LIGASE_II_2, ASP_PROTEASE, EGF_1, LIPOCALIN and RRM_RNP_1), two different *extended* patterns with a similar correlation (C) have been obtained (*extended 1* and *extended 2*): the first favouring a higher sensitivity (S_n) and the second privileging selectivity (S_l) and specificity (S_p) (see table 1 also for the definition of S_n, S_l, S_p and C). For example, the LIPOCALIN PROSITE pattern matches 70 true positive, 82 false positive and 35 false negative sequences on the SWISS-PROT database (release 40.7) (see additional data file 2). The LIPOCALIN *extended 1* pattern, on the same SWISS-PROT release, also has 70 true positives and 35 false negatives. Interestingly, however, the number of false positives is much lower (FP = 11) (see additional data file 3). In this case, the sensitivity of the *extended 1* pattern is unchanged (S_n = 0,667) with respect to the PROSITE pattern sensitivity, whereas the selectivity of the *extended 1* pattern (S_l = 0.864) is higher than the PROSITE pattern selectivity (S_l = 0.461) (see table 1). The *extended 2* pattern matches more true positive sequences (TP = 81) than the PROSITE one, less false positives (FP = 12) and less false

negatives (FN = 24). In the literature many efforts are made in building a single optimised functional motif corresponding to a specific function, which is the result of a deal between sensitivity and specificity. For function inference, however, being given the opportunity to choose the more specific or sensitive functional motif would be useful: a more specific motif would provide more reliable function assignments (though missing some) whereas a more sensitive one would detect a higher number of potential true positives (even if with a lower degree of reliability).

The optimisation and testing procedure resulted in three different outcomes:

2) In one case (THIOL_PROTEASE_HIS) an *extended* pattern with selectivity and specificity equal to 1.0 was obtained (table 1). The number of false positive matches of such *extended* pattern on both the SWISS-PROT releases used is zero. The correlation of such a pattern increases from C = 0.596 (PROSITE) to C = 0.886 (*extended*).

3) In one case (CYTOCHROME_C) no new *extended* pattern was identified with a better correlation than the PROSITE original one.

See additional file 4 for a detailed description and analysis of the test cases under study.

aligned struct / cell number	1kmmC	1aszA	1e24A	1pysA	1qf6A
Cell 1	I108	I130	I257	P199	F358
Cell 2	F112	[F324]	F261	F203	[H362]
Cell 3	R113	R325	R262	R204	R363
Cell 4	R121	H334	H270	H212	R375
Cell 5	Y122	M335	[N271]	[Q213]	[V376]
Cell 6	F125	F338	F274	F216	F379
Cell 7	A306	G526	G475	G316	A513
Cell 8	G308	G528	G477	G318	G516
Cell 9	E310	E530	D479	E320	E519
Cell 10	R311	R531	R480	R321	R520

Figure 4
Example of a Refined Heavy Elements Table (R-HET). The Refined Heavy Elements Table (R-HET) corresponding to the AA_TRNA_LIGASE_II_2 pattern. The residues manually added to each cell of the multiple alignment grid after visual inspection are reported in square brackets. The residues belonging to the sequence PROSITE pattern are underlined. Residues are indicated in one-letter code and the number corresponds to the residue number in the PDB file.

Biological role of the structurally conserved residues in the test cases analysed

By analysing the SWISS-PROT and PDB annotations, the positions added to the PROSITE AA_TRNA_LIGASE_II_1, AA_TRNA_LIGASE_II_2, ASP_PROTEASE, EGF_1, LIPOCALIN and RRM_RNP_1 patterns seem to represent weak structural constraints rather than playing a clear functional role. In the THIOL_PROTEASE_HIS *extended* pattern, two of the new positions added to the PROSITE pattern are occupied by residues known to be involved in the catalytic activity of the eukaryotic thiol proteases [26]. Proteins belonging to the thiol protease family (e.g. calpains, cathepsins L, B or K, papain) share a low sequence identity except in the vicinity of the active site [26]. Also, the global structures display several significant differences whereas the structural region around the active site is well conserved across different proteins of the family. The structural analysis, focused on the surface area nearby the histidine residue of the catalytic triad (Cys-His-Asn), allowed us to identify, among the best conserved residues, the asparagine of the catalytic triad and a tryptophan, which is positioned +2 along the sequence with respect to the Asn (Figure 3). In the case of the active form of calpains, Hosfield *et al*. [27] showed that such conserved Trp residue has a weak interaction with the His residue of the catalytic triad and helps maintaining the His orientation. Moreover, mutation of the Trp to Tyr reduced the activity of calpain to 5% of the wild-type value. Hosfield *et al*. results are in agreement with the role of such conserved Trp residue in other cysteine proteases [28,29]. The Cys residue of the catalytic triad belongs to a structural domain which is different from that of the His and Asn residues. Among the *extended* patterns defined in this work, the THIOL_PROTEASE_HIS *extended* pattern seems to be the only one containing positions occupied by residues known to play a functional role, besides the PROSITE pattern positions, and it is the only one that matches NO false positives.

Discussion

In this work we developed a procedure aimed at increasing the specificity and/or sensitivity of functional sequence patterns – such as PROSITE ones -, matching at least two true positive protein sequences of known structure. The surface region around the three-dimensional fragment corresponding to a PROSITE pattern is analysed. Residues conserved in different proteins sharing the same pattern are identified by means of the 3D profile procedure [18] and a further visual analysis of the structures. Visual inspection is complementary to the automatic procedure and enables us to identify surface conserved residues that would be otherwise missed and/or to discard amino acids erroneously considered structurally aligned by the computational procedure. Structurally conserved residues, which are co-linear in the corresponding

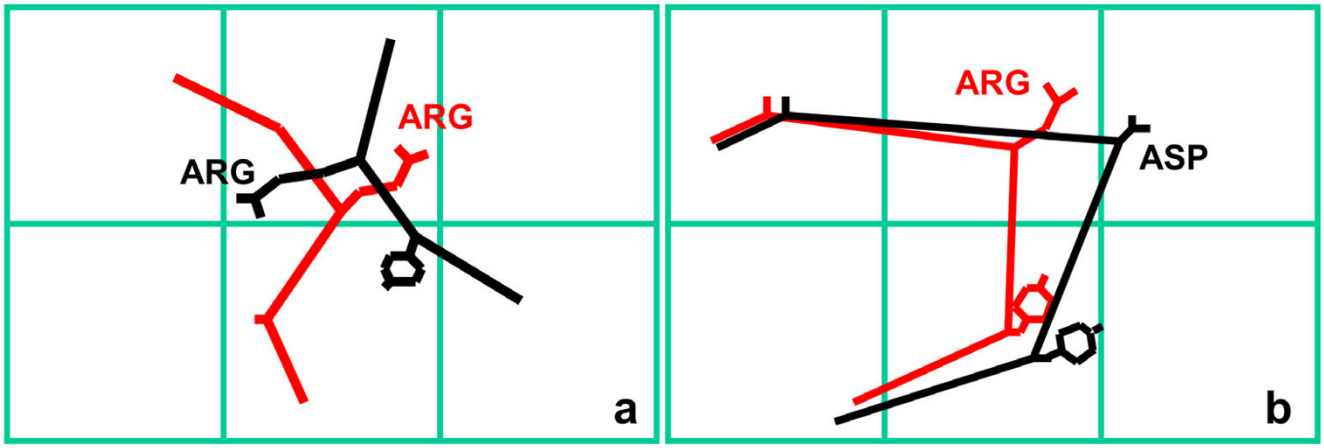


Figure 5

Example of residues discarded (a) or included (b) in a R-HET after visual analysis. a. Two identical residues (ARG) belonging to two aligned proteins (protein A and B), which are clearly not correctly structurally aligned, though they fall in the same cell of the HET. The 3D multiple alignment grid is represented in green as a 2D grid. Structure segments of protein A and B are reported in red and black, respectively. This pair of residues is discarded after visual analysis. b. Two different residues (ARG and ASP) belonging to two aligned proteins (protein A and B), which clearly correspond though they do not fall in the same cell of the HET. This pair of residues is included in the R-HET after visual inspection.

Table 1: Sensitivity, selectivity, specificity and correlation of PROSITE test cases and the corresponding *extended* patterns.

AC[a]	type[a]	S_n[b]	S_l[b]	S_p[b]	C[b]
AA_TRNA_LIGASE_II_1	prosite	0.648	0.802	0.9994	0.720
	extended 1	0.648	0.908	0.9997	0.767
	extended 2	0.719	0.813	0.9994	0.764
AA_TRNA_LIGASE_II_2	prosite	0.531	0.464	0.9977	0.494
	extended 1	0.528	0.617	0.9988	0.569
	extended 2	0.622	0.528	0.9979	0.572
ASP_PROTEASE	prosite	0.974	0.836	0.9996	0.902
	extended 1	0.974	0.921	0.9998	0.947
	extended 2	0.984	0.900	0.9998	0.941
EGF_1	prosite	0.679	0.792	0.9993	0.733
	extended 1	0.679	0.936	0.9998	0.796
	extended 2	0.750	0.864	0.9995	0.805
LIPOCALIN	prosite	0.667	0.461	0.9992	0.554
	extended 1	0.667	0.864	0.9999	0.759
	extended 2	0.771	0.871	0.9999	0.820
RRM_RNP_1	prosite	0.582	0.537	0.9985	0.558
	extended 1	0.582	0.719	0.9993	0.646
	extended 2	0.629	0.614	0.9988	0.620
THIOL_PROTEASE_HIS	prosite	0.785	0.459	0.9986	0.596
	extended 1	0.785	1.000	1.0000	0.886

[a] Pattern accession number (AC) in the PROSITE database and the type of pattern (type): PROSITE, *extended* 1 or *extended* 2[b] Sensitivity S_n (defined as $S_n = TP/(TP + FN)$), selectivity S_l ($S_l = TP/(TP + FP)$), specificity S_p ($S_p = TN/(TN + FP)$)and correlation C ($C = [TP \times TN - FP \times FN]/ [(TP + FP) \times (FP + TN) \times (TN + FN) \times (FN + TP)]^{1/2}$) of the patterns on the SWISS-PROT release 40.8. TN is the number of true negatives, which is calculated as the total number of sequences on the SWISS-PROT release 40.8 (= 101659) less the sum of true positive, false negative and partial sequences.

sequences, are used to develop new *extended* patterns each containing, as a subset, its PROSITE original pattern. The positions in sequence of the structurally conserved residues may precede and/or follow the PROSITE pattern. Structurally conserved residues added to PROSITE patterns may be located even in very distant positions along the protein sequences (see, for example, the *consensus* signatures shown in the table reported in the additional file 3). Indeed, functionally important residues, which are generally localised in space, may be dispersed along the sequence (figure 1). For this reason, the *extended* patterns cannot be easily identified in multiple sequence alignments. In many cases, PROSITE pattern true positive sequences do not all belong to the same protein family. Therefore sequence alignments such as those provided by Pfam [30] would not allow the identification of the conserved positions which we detected using the computational and visual analysis of structurally aligned proteins. Our method requires only that a set of proteins shares the function associated to a PROSITE 'leaky' pattern, independently of the degree of similarity displayed by their sequences.

Extended patterns, obtained by introducing sequence constraints derived from structural data, are expected to display a better correlation than the corresponding PROSITE patterns. They can be used for data mining not only in structure databases, but also in protein sequence databases. In this respect, the procedure described here is a powerful tool for sequence analysis and function inference in the context of proteomics. The method relies on both automatic and manual contributions: the computational approach allows us to carefully and exhaustively explore a functional surface region of a particular set of proteins sharing a function, thus ensuring the detection of ALL the potential conserved residues. Visual inspection of the multiple alignment of structures, on the other hand, guarantees accuracy of analysis and cuts out ambiguous or even wrong assignments. It would be interesting to explore the possibility of making the procedure entirely automated (without losing too much accuracy) and using it to systematically analyse the complete PROSITE database in order to develop *extended* patterns for the vast majority of PROSITE 'leaky' patterns. In the eight test cases analysed, the procedure proved to be successful in all but one case. The CYTOCHROME_C represents a particular case where, in the protein surface regions nearby the PROSITE pattern, no extra structurally conserved residues are found, neither by means of the computational method nor through a visual analysis of the structures. This is not expected to be a frequent situation, especially in the case of 'leaky' patterns. All the CYTOCHROME_C *extended* patterns constructed using the surface conserved residues belonging to some subsets of structures have a worse performance than the original CYTOCHROME_C PROSITE pattern. More generally, the method presented in this work is likely to fail with very short functional patterns, matching structures with a very different local fold in the region of the 3D PROSITE fragment. The method is effective when residues playing a direct or indirect role in the biological function of a protein family are structurally well conserved in the proteins sharing that function and irregularly dispersed in the corresponding sequences.

Interestingly, in seven out of the eight patterns examined, a correlation between the functional relevance of the conserved residues that were added to the PROSITE 'leaky' pattern and the performance of the corresponding *extended* pattern, can be observed. The AA_TRNA_LIGASE_II_1, AA_TRNA_LIGASE_II_2, ASP_PROTEASE, EGF_1, LIPOCALIN and RRM_RNP_1 *extended 1* and *2* patterns display a better performance than the original PROSITE patterns. In some cases the improvement is remarkable, yet the patterns still match false positives. Could this be due to the absence of some other residue/s with a crucial functional role in the pattern representation? In the case of the THIOL_PROTEASE_HIS *extended* pattern, which matches NO false positives, we found that two of the structurally conserved residues added to the PROSITE 'leaky' pattern are directly involved in the protease function.

Conclusions

The 'traditional' construction of a functional motif or pattern typically involves the identification of the residues conserved in a multiple sequence alignment of related proteins. Such residues are the 'clues' for developing a new pattern. The basic condition for building a pattern is the existence of a set of related proteins. In many cases a functional pattern thus obtained does not match all and only the sequences of proteins sharing the function associated to the pattern. The procedure described in this work can be applied to any type of poorly-performing functional pattern for which at least two three-dimensional true positives exist. It is based on the identification of structurally conserved residues in the surface area close to the three-dimensional fragment corresponding to the 'leaky' pattern. For the identification of such residues we use a procedure, which is available on request and will soon be accessible on our web server [31]. The procedure consists of an automated structural superimposition plus a protein surface analysis, which allows the identification of the best-conserved residues in the examined region. In cases for which only one true positive structure is available, the procedure returns those residues close to the pattern under study and which are exposed to the solvent. The structurally conserved residues represent a sort of 'clue' for *extending* the original pattern. The structural conservation can be refined and optimised by visual inspection. The degree of refinement will generally depend on the pattern

curator needs. Once the structurally-conserved residues have been identified, the construction of a 'rough' pattern is straightforward, as is its test on a sequence database, following the steps described in Results. The application of the entire procedure to a set of 'leaky' patterns shows that in seven out of eight cases pattern selectivity was improved, as was, in some cases, selectivity and sensitivity.

Methods

Databases

Sequence patterns used to test our procedure were obtained from the PROSITE database, release 17.01 [20], containing 1501 entries, 1331 of which are patterns. Profile and rule entries are not considered in this work.

A non-redundant set of PDB (Protein Data Bank) [23] chains (referred in the following as nr-PDB) was compiled from the nrpdb.032002 (released on March 2002) NCBI file [32] by choosing the set of non-redundancy groups identified by the BLAST p-value 10e-40 for PDB chains sequence similarity. Such a cut-off divides the PDB chains in 4321 groups. In each group, chain structures are scored on the basis of the structure quality. A representative structure is then selected. 126 out of such 4321 groups contain low-resolution structures only, and have been therefore discarded. 1028 groups have a NMR structure as a group representative. In such cases, the best ranking X-ray structure was taken as a group representative. A library of 4195 sequences was derived from the nr-PDB chains and used for pattern matching.

The *extended* patterns were built and optimised on the SWISS-PROT database [25] release 40.8 (101,659 sequences). To verify the pattern performance on a set of protein sequences different with respect to the 'training' set, the final *extended* patterns were subsequently tested on the SWISS-PROT database release 41.0 (122,564 sequences).

Links between PROSITE pattern entries and true positive nr-PDB structures

The PROSITE.DAT file [33] provides, for each pattern, a list of links to SWISS-PROT entries labelled with T (true positive, TP), F (false positive, FP) or N (false negative, FN) and links to the corresponding true positive PDB structures. Such links, however, are too sparse and not all redundancy groups are therein represented. Therefore, links were established between PROSITE pattern entries and nr-PDB structures as follows. Each PROSITE pattern *P* was searched in the nr-PDB sequences. Links from each matching nr-PDB chain to the corresponding SWISS-PROT entries were used when available, yet not all the PDB files contain links to SWISS-PROT entries. The EMBOSS [34] program *water* for pairwise sequence alignment was used to associate nr-PDB to SWISSPROT sequences for sequence identity higher than 90%. A nr-PDB chain matching a PROSITE pattern *P* is assumed to be a true positive for *P* if displaying a link to a SWISS-PROT entry annotated as true positive for *P*-

Selection of the PROSITE 'leaky' patterns used to test the procedure

In the current work, the number of TP, FP, FN and partial sequences for each PROSITE pattern was extracted from the PROSITE.DAT file. PROSITE patterns were then sorted according to the percentage of FP and FN. Then a set of six PROSITE patterns, for which at least three true positive structures exist in the nr-PDB, was selected among the 'leakiest' ones (see additional data file 1). The true positives, false negatives and partial sequences, belonging to this subset, were checked through the SWISS-PROT annotations. AA_TRNA_LIGASE_II_1, AA_TRNA_LIGASE_II_2, EGF_1, LIPOCALIN, RRM_RNP_1, THIOL_PROTEASE_HIS display >18% false positives and miss >12% true positives in the SWISSPROT database. Instead of enriching the testing set with some more 'leaky' pattern chosen at random, we considered two examples of biologically interesting patterns: the CYTOCHROME_C (63% false positive matches and 0.6% false negatives) and the ASP_PROTEASE (16.3% false positive matches and 2.2% false negatives). The former was chosen since it is an interesting biological well-known pathological example of a sequence pattern. True positives are found in a wide range of proteins, with different folds and, in some cases, even a different local structure [21]. The latter is a pattern describing a widely-studied family of proteins whose structures show a very conserved and spatially localised active site. Such functional/structural information is partially lost in the corresponding PROSITE pattern. One of the most interesting 'leaky' PROSITE patterns is the ATP_GTP_A pattern (known as "p-loop"). Due to the huge number of false positive hits, this pattern is not annotated for true positives, false positives and false negatives in the SWISS-PROT database. The p-loop and many local structural features of its true positive proteins have already been analysed in Via *et al*. [35].

Retrieval of the structures associated to functional motifs

A structure fragment corresponding to a sequence pattern *P* is defined as the set of residues starting with the residue in the first position of *P* and ending with the residue in the last position of *P*. In the additional data file 2, for each PROSITE pattern of the sample, the list of true positive structures in the nr-PDB is reported. The CYTOCHROME_C PROSITE pattern matches 49 structures in the nr-PDB database. For this specific pattern, however, we decided to use a less redundant PDB (BLAST p-value 10e-7) and considered only 28 non-redundant

chains (between brackets in the sixth column of the table reported in the additional file 2).

In three cases, namely the ASP_PROTEASE, the RRM_RNP_1 and the CYTOCHROME_C, there are structures whose sequence matches the pattern more than once. More precisely, each one of the 1bxoA, 1zap_ and 2rmpA sequences contains two matches of the ASP_PROTEASE PROSITE pattern, the 2up1A chain contains two matches of the RRM_RNP_1 PROSITE pattern; 1eb7A, 1etpA, 1fcdC, 1wad_ and 2cthA chains contain two matches each of the CYTOCHROME_C PROSITE pattern, which is also found three times in 1czj_, 1hh5A, 1qjdA, 2cy3_ and 3caoA structures, four times in 1bvb_, 1prcC and 1qdbA and up to eight and nine times in 1fgjA and 19hcA, respectively.

Identification of key residues in the surface area nearby the PROSITE residues

The first step of the procedure consists of identifying key residues in the surface area nearby the PROSITE residues. To this aim, a structure fragment corresponding to a sequence pattern P must be defined. Here, such a structure fragment is defined as the set of residues starting with the residue in the first position of P and ending with the last position of P. A protein structure may contain even more than one structure fragment corresponding to a single pattern P. Then the true positive structures for a pattern P are superimposed onto the residues belonging to the pattern. To superimpose any pair of structures on the pattern residues, the correspondence between residue pairs has to be defined unambiguously. When superimposing structural fragments corresponding to PROSITE patterns, it is possible either to consider only residues matching non-wildcard positions or to include also residues matching fixed-length wildcards [21,22]. Lin *et al*. [36] performed a conformational analysis of long spacers of fixed length in PROSITE patterns and found that, for the majority of the cases analysed, the entire backbone of a long spacer is structurally well conserved. In this study, we consider PROSITE patterns with spacers of both fixed and variable length. Thus we decided to superimpose only residues matching non-wildcard positions. The multiple structural alignment is performed onto the residue pseudo-atoms. Pseudo-atom co-ordinates of each residue are calculated as the average of the residues side-chain atom co-ordinates. When superimposing two or more structures, one of them has to be taken as 'target' (the target structure establishes the reference system) and the other(s) as 'probe(s)' subjected to a rigid body rotation and translation with respect to the target structure. The 'best target' or 'master' 3D structural fragment corresponding to a PROSITE pattern is selected as the fragment with the lowest average pair-wise RMSD to all other fragments in a group of structures (a group consists of the nr-PDB true positives of that PROSITE pattern). In this procedure, structures matching a pattern more than once are not included in the first round of superimposition. The 3D fragments corresponding to multiple matches are superimposed, one at a time, onto the master fragment previously identified and the RMSD is calculated. Only the one with the lowest RMSD is retained while the others are discarded. Once the structural superimposition has been done, the 3D profile method [18] is used to identify conserved residues in the functional region of each group of superimposed structures. The method starts from a multiple superimposition of n protein structures, transforms each one of them into a surface structure (by retaining only exposed residues) and places each structure into a 3D grid, centred around the 3D fragment corresponding to the PROSITE pattern. Then the n grids are merged into a unique 3D multiple alignment grid. The kth cell of the 3D multiple alignment grid contains a set of residues, each one coming from the kth cell of each single protein template grid. The sequence profile [37] associated to the residues located in the same cell is calculated for each cell of the 3D multiple alignment grid, which we call 3D profile. The procedure also generates a table with the 100 cells displaying the highest scoring profile values of the 3D multiple alignment grid. Since the most conserved elements are called 'heavy' elements [18], this table is referred as to Heavy Elements Table (HET). For each PROSITE pattern, a HET can be produced displaying residues belonging to the linear pattern but also sometimes highlighting other interesting features, as will be discussed below. An example of HET is reported in Figure 2.

A web server to obtain a HET starting from a user-defined pattern plus one or more true positive structures will be soon available.

Construction of an extended *pattern*

The procedure for the construction of an *extended* pattern can be divided into two steps. The first step consists of the extraction of structural information from HETs to obtain a first 'rough' *extended* pattern. The second concerns more specifically a procedure for gradually 'softening' and testing the 'rough' *extended* pattern obtained in the first step on a sequence database.

Analysis of a 3D profile heavy elements table and visual inspection of the structurally conserved residues

A Heavy Elements Table (HET) is a collection of structurally conserved residues in the functional/binding region across different structures of proteins sharing a biological property. Some of these 'heavy' elements, even if falling in the same cell of the 3D multiple alignment grid, can be structurally and/or functionally unrelated. Here, we want to focus only on conserved residues that are co-linear in the corresponding sequences. For example, consider a res-

idue located near the N-terminus of a protein A and a residue positioned close to the C-terminus of a protein B superimposed onto A. Such residues are not taken into account in the construction of the *extended* sequence pattern, albeit they are localised in the same spatial position of the two superimposed structures A and B. Thus, among the 'heavy' elements, only those displaying a good structural superimposition – comprising the same spatial orientation – and preserving the same order in the corresponding protein sequence, are selected. To this end, a visual inspection to check the position and orientation of the HET residues in the structural alignment can be performed using a molecular graphic software, such as Swiss-PdbViewer [38] or PyMol [39]. For each pattern, the visual analysis allows the construction of a new HET (named Refined Heavy Elements Table, or R-HET), which contains a strictly conserved subset of the original HET cells (figure 3 and 4).

In more detail, the visual analysis of the structural alignment makes it possible to check whether residues belonging to the same cell in the HET are indeed structurally well aligned or not. The 3D profile method uses, besides the quality of the superimposition, the chemical and physical similarity of the residues falling into the same cell to score the cell itself. Two superimposed residues displaying reversed side chain directions are likely to encode different functions in the corresponding structures, although both of them fall in the same three-dimensional cell (figure 5a). The visual analysis makes it possible to identify and discard such erroneously-aligned residues. On the other hand two residues, from two superimposed structures, may be actually aligned, even though their physicochemical properties differ (e.g. a lysine and an aspartic acid) and/or the residues are not sufficiently close in space to fall in the same cell of the grid (figure 5b). The visual analysis highlights such situations. Residues absent in a cell of the HET, which are revealed to be aligned instead by visual inspection, are added to the corresponding cell of the R-HET (see figure 4).

In general, the visual inspection allows the exclusion of residues from structurally-aligned proteins erroneously selected by the automated 3D profile method while taking into account aligned residues which the method failed to detect. We want to emphasise that the use of a visual inspection is subjective and depends on the degree of accuracy that is required in the pattern analysis. A HET can be used as a starting point for the construction of an *extended* pattern as it is, above all when a more automated procedure is preferred.

Construction of a 'rough' extended pattern

Once a HET or a R-HET is generated, the sequence intervals between structurally conserved residues are determined in each sequence as shown in the following for the R-HET described in figure 4.

I108-x(3)-F112-R113-x(7)-R121-Y122-x(2)-F125-x(180)-PROSITE

I320-x(3)-F324-R325-x(8)-H334-M335-x(2)-F338-x(187)-PROSITE

I257-x(3)-F261-R262-x(7)-H270-N271-x(2)-F274-x(200)-PROSITE (1)

P199-x(3)-F203-R204-x(7)-H212-Q213-x(2)-F216-x(100)-PROSITE

F358-x(3)-H362-R363-x(11)-R375-V376-x(2)-F379-x(135)-PROSITE

where 'PROSITE' indicates the PROSITE pattern positions, namely, in the AA_TRNA_LIGASE_II_2 case:

PROSITE = [GSTALVF] - [^DENQHRKP] - [GSTA] - [LIVMF] - [DE] -R- [LIVMF] -X(1) - [LIVMSTAG] - [LIVMFY]

The first rough extended pattern is then deduced from (1):

[FIP] -x(3) - [FH] -R-x(7, 11) - [RH] - [MNVQY] -x(2)-F-x(100, 200) - PROSITE (2)

The rough extended pattern includes, in each position, all the residues found in a column of the corresponding R-HET and considers, for the intervals between fixed positions, a variable range of residues according to the minimum and maximum length of the intervals between columns of (1).

Testing an extended pattern on the SWISS-PROT database

The last step of the entire procedure consists of scanning a sequence database (e.g. SWISS-PROT) with an *extended* pattern. This step is similar to the one adopted by PROSITE in developing a new pattern starting from a 'core' pattern (see the PROSITE user manual at [40]).

A 'rough' extended pattern is initially searched on the set of true positive sequences of the SWISS-PROT database. If the rough extended pattern does not select all the known true positive sequences, the pattern is gradually generalized, working left-to-right as follows:

(a) in each position the match set of identities and ambiguous positions is extended by including residues similar to the one/s already present in that position. After the inclusion of each residue, the pattern is searched on the

SWISS-PROT database and a trade-off between sensitivity and specificity is carried out. If the number of true positive matching sequences grows by n units while the number of false positives grows by m > n units, the residue is removed and another residue is substituted and tested.

(b) if the pattern resulting from step (a) is not satisfactory, then the number of wildcard positions is increased. After increasing the number of wildcard positions by a unit, the pattern is searched on the SWISS-PROT database and a trade-off between sensitivity and specificity is performed. The pattern with the increased number of wildcard positions is accepted only if the number of true positive matching sequences grows by n units whereas the number of false positives (defined as not-true positives) grows by m < n units.

(c) when adding a new residue to a pattern position, some other residues could turn out to be superfluous. In order to minimise the number of constraints, residues whose absence did not affect the pattern sensitivity and specificity were discarded.

Since the "PROSITE cores" of the *extended* patterns are not modified by this procedure, the number of false negative sequences is not expected to decrease. Consequently, the PROSITE region of each extended pattern is 'softened' as described in (a) and (b), with the aim of obtaining supplementary *extended* patterns (*extended 2*) matching some true positives missed by the PROSITE and the first *extended* (*extended 1*) patterns. Tests on the SWISS-PROT database and trade-offs between sensitivity and specificity were performed for the *extended 2* patterns as well. To confirm the findings, the final *extended 1* and *extended 2* patterns can also be tested on a different SWISS-PROT release. In this case, further false positive, false negative and partial sequences can be identified through the SWISS-PROT annotation.

Authors' contributions

AV and MHC conceived the research and authored the manuscript. AV developed and tested the procedure. MHC supervised the work.

Additional material

Additional File 1

The PROSITE patterns examined to test the procedure file1.pdf is a pdf (adobe) file. It contains a table listing the PROSITE patterns used to test the methodology. For each pattern, the PROSITE identification number, the consensus signature and a brief description are provided.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-50-S1.pdf]

Additional File 2

Test cases number of matching sequences, TP, FP, FN, partial sequences and nr-PDB true positives file2.pdf is a pdf (adobe) file. It contains a table listing, for each PROSITE pattern used to test the procedure, the number of true positives, false positives, false negative and partial sequences on the SWISS-PROT database. The true positives on the non-redundant PDB database are also provided.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-50-S2.pdf]

Additional File 3

Consensus and other data of the extended *patterns corresponding to the PROSITE test cases analysed file3.pdf is a pdf (adobe) file. It displays a table containing* extended *patterns data. The signature consensus is provided together with the number of true positives, false positives, false negatives and partial sequences on the SWISS-PROT sequence database.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-50-S3.pdf]

Additional File 4

Detailed description of the analysed test cases Description.pdf is a pdf (adobe) file, which contains a detailed description of the PROSITE patterns used as test cases. Data (true positives, false positives etc.) obtained with the extended *patterns are accurately analysed and discussed.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-50-S4.pdf]

Acknowledgements

We thank Gianni Cesareni for helpful discussions and for critically reviewing the manuscript and Armin Lahm for useful suggestions. We gratefully acknowledge the support of Telethon, AIRC, GENEFUN and EU grant QLRI-CT-2000-00127. AV is supported by FIRB "Bioinformatica per la Genomica e la Proteomica".

References

1. Shapiro L, Harris T: **Finding function through structural genomics.** *Curr Opin Biotechnol* 2000, **11:**31-35.
2. Gutteridge A, Bartlett GJ, Thornton JM: **Using a neural network and spatial clustering to predict the location of active sites in enzymes.** *J Mol Biol* 2003, **330:**719-734.
3. Smith HO, Annau TM, Chandrasegaran S: **Finding sequence motifs in groups of functionally related proteins.** *Proc Natl Acad Sci USA* 1990, **87:**826-830.
4. Smith RF, Smith TF: **Automatic generation of primary sequence patterns from sets of related protein sequences.** *Proc Natl Acad Sci USA* 1990, **87:**118-122.
5. Neuwald AF, Green P: **Detecting patterns in protein sequences.** *J Mol Biol* 1994, **239:**698-712.
6. Jonassen I, Collins JF, Higgins DG: **Finding flexible patterns in unaligned protein sequences.** *Protein Sci* 1995, **4:**1587-1595.
7. Nevill-Manning CG, Wu TD, Brutlag DL: **Highly specific protein sequence motifs for genome analysis.** *Proc Natl Acad Sci USA* 1998, **95:**5865-5871.
8. Brazma A, Jonassen I, Eidhammer I, Gilbert D: **Approaches to the automatic discovery of patterns in biosequences.** *J Comput Biol* 1998, **5:**279-305.
9. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273:**595-603.
10. Nussinov R, Wolfson HJ: **Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques.** *Proc Natl Acad Sci USA* 1991, **88:**10495-10499.
11. Fisher D, Bachar O, Nussinov R, Wolfson HJ: **An efficient automated computer vision based technique for detection of**

three dimensional structural motifs in proteins. *J Biomol Struct Dyn* 1992, **9:**769-789.
12. Kleywegt GJ: **Recognition of spatial motifs in protein structures.** *J Mol Biol* 1999, **285:**1887-1897.
13. Orengo CA: **A review of methods for protein structure comparison. Patterns in Protein Sequence and Structure.** In *Springer series in Biophysics Volume 7*. Edited by: Taylor WR. Heidelberg: Springer-Verlag; 1992:159-188.
14. Vriend G, Sander C: **Detection of common three-dimensional substructures in proteins.** *PROTEINS* 1991, **11:**52-58.
15. Orengo CA, Taylor WR: **A local alignment method for protein structure motifs.** *J Mol Biol* 1993, **233:**488-497.
16. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P: **A graph-theoretic approach to the identification of three-dimensional patterns od amino acid side-chains in protein structures.** *J Mol Biol* 1994, **243:**327-344.
17. Russell RB: **Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution.** *J Mol Biol* 1998, **279:**1211-1227.
18. de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M: **Three-dimensional profiles: a new tool to identify protein surface similarities.** *J Mol Biol* 1998, **284:**1211-1221.
19. Jonassen I, Eidhammer I, Taylor WR: **Discovery of local packing motifs in protein structures.** *PROTEINS: Structures, Function, and Genetics* 1999, **34:**206-219.
20. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
21. Kasuya A, Thornton JM: **Three-dimensional structure analysis of PROSITE patterns.** *J Mol Biol* 1999, **286:**1673-1691.
22. Jonassen I, Eidhammer I, Grindhaug SH, Taylor WR: **Searching the protein structure databank with weak sequence patterns and structural constraints.** *J Mol Biol* 2000, **304:**599-619.
23. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C: **The Protein Data Bank.** *Acta Crystallogr D Biol Crystallogr* 2002, **D58:**899-907.
24. Irving A, Whisstock JC, Lesk AM: **Protein structural alignments and functional genomics.** *Proteins* 2001, **42:**378-382.
25. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
26. Berti PJ, Storer AC: **Alignment/phylogeny of the papain superfamily of cysteine proteases.** *J Mol Biol* 1995, **246:**273-283.
27. Hosfield CM, Elce JS, Davies PL, Jia Z: **Crystal structure of calpain reveals the structural basis for Ca^{2+}-dependent protease activity and a novel mode of enzyme activation.** *EMBO J* 1999, **18:**6880-6889.
28. Arthur JSC, Guthier S, Elce JS: **Active site residues in m-calpain: identification by site-directed mutagenesis.** *FEBS Lett* 1995, **368:**397-400.
29. Bromme D, Bonneau PR, Purisima E, Lachance P, Hajnik S, Thomas DY, Storer AC: **Contribution to activity of histidine-aromatic, amide-aromatic, and aromatic-aromatic interactions in the extended catalytic site of cysteine proteinases.** *Biochemistry* 1996, **35:**3970-3979.
30. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam Protein Families Database.** *Nucleic Acids Res* 2002, **30:**276-280.
31. **Centro di Bioinformatica Molecolare** [http://cbm.bio.uniroma2.it]
32. **Non-Redundant PDB Chain Set** [http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html]
33. **ExPASy molecular biology anonymous FTP server of the Swiss Institute of Bioinformatics (SIB)** [ftp://us.expasy.org/databases/prosite/release]
34. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends in Genetics* 2000, **16:**276-277.
35. Via A, Ferrè F, Brannetti B, Valencia A, Helmer-Citterich M: **Three-dimensional view of the surface motif associated with the p-loop structure: *cis* and *trans* cases of convergent evolution.** *J Mol Biol* 2000, **303:**455-465.
36. Lin KY, Wright J, Lim C: **Conformational analysis of long spacers in PROSITE patterns.** *J Mol Biol* 2000, **299:**537-548.
37. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: Detection of distantly related proteins.** *Proc Natl Acad Sci USA* 1987, **84:**4355-4358.
38. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb-Viewer: An environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
39. **PyMOL Home Page** [http://pymol.sourceforge.net/]
40. **The PROSITE database of protein families and domains. User Manual** [http://us.expasy.org/prosite/prosuser.html#meth13]