PROSITE: A documented database using patterns and profiles as motif descriptors

Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch and Philipp Bucher

Date received (in revised form): 31st May 2002

Amos Bairoch and Philipp Bucher
are group leaders at the Swiss Institute of Bioinformatics (SIB), whose mission is to promote research, development of software tools and databases as well as to provide education, training and service activities within the field of bioinformatics. Amos Bairoch has developed the SWISS-PROT protein and the PROSITE motif databases, whereas Philipp Bucher has developed the generalised profiles used in PROSITE.

Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Laurent Falquet and Marco Pagni
are researchers working at the SIB.

Alexandre Gattiker
is a PhD student doing a thesis in the field of bioinformatics.

Keywords: *pattern, regular expression, profile, weight matrix, database*

Abstract

Among the various databases dedicated to the identification of protein families and domains, PROSITE is the first one created and has continuously evolved since. PROSITE currently consists of a large collection of biologically meaningful motifs that are described as patterns or profiles, and linked to documentation briefly describing the protein family or domain they are designed to detect. The close relationship of PROSITE with the SWISS-PROT protein database allows the evaluation of the sensitivity and specificity of the PROSITE motifs and their periodic reviewing. In return, PROSITE is used to help annotate SWISS-PROT entries. The main characteristics and the techniques of family and domain identification used by PROSITE are reviewed in this paper.

INTRODUCTION

PROSITE is an annotated collection of motif descriptors dedicated to the identification of protein families and domains. The motif descriptors used in PROSITE are either patterns or profiles, which are derived from multiple alignments of homologous sequences. This gives to these motif descriptors the notable advantage of identifying distant relationships between sequences that would have passed unnoticed based solely on pairwise sequence alignment. Patterns and profiles have both their own strengths and weaknesses, which define their area of optimum application.

The core of the PROSITE database is composed of two text files:

- PROSITE.DAT is a computer-readable file that contains all the information necessary to programs that make use of PROSITE to scan sequence(s) for the occurrence of patterns or profiles. This file includes, for each of the entry described, statistics on the number of hits obtained while scanning the SWISS-PROT protein database[1] for a pattern or profile. Cross-references to the corresponding SWISS-PROT entries as well as to matched sequences from the PDB 3D-structure database[2] are also provided.
- PROSITE.DOC contains textual information that fully documents each pattern or profile.

Release 17.18 of PROSITE (August 4, 2002) contains 1147 documentation entries that describe 1567 different motif descriptors. In addition to these entries, a collection of 152 pre-release profiles (see below) is also available.[3]

PROSITE PATTERNS

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by pairwise sequence alignment. However, relationships can be revealed by the occurrence in its sequence of a particular cluster of residue types, which is variously known as a pattern, motif, signature or fingerprint. These motifs, typically around 10 to 20 amino acids in length, arise

Christian J. A. Sigrist,
Swiss Institute of Bioinformatics (SIB),
CMU, University of Geneva,
1 rue Michel Servet,
CH-1211 Geneva 4,
Switzerland

Tel: +41 22 702 58 68
Fax: +41 22 702 58 58
E-mail: christian.sigrist@isb-sib.ch

Patterns are regular expressions matching short sequence motifs usually of biological meaning

because specific residues and regions thought or proved to be important to the biological function of a group of proteins are conserved in both structure and sequence during evolution. These biologically significant regions or residues are generally:

- Enzyme catalytic sites (Figure 1a).
- Prostethic group attachment sites (heme, pyridoxal-phosphate, biotin, etc.).
- Amino acids involved in binding a metal ion.
- Cysteines involved in disulphide bonds.
- Regions involved in binding a molecule (ADP/ATP, GDP/GTP, calcium, DNA, etc.) or another protein.

As the sequence of biologically meaningful motifs is evolutionarily conserved, a multiple alignment of them can be reduced to a consensus expression called a *regular expression* or *pattern*. Each position of such a pattern can be occupied by any residue from a specified set of acceptable residues, and in addition can be repeated a variable number of times within a specified range. At strictly conserved positions only one particular amino acid is accepted, whereas at other positions several amino acids with similar physicochemical properties can be accepted. It is also possible to define which amino acid(s) is(are) incompatible with a given position, and conserved residues can be separated by gaps of variable lengths. Finally, the pattern syntax provides features to anchor a pattern either at the beginning or at the end of a sequence (Figure 1b). The complete syntax of a PROSITE pattern is available at http://www.expasy.org/tools/scanprosite/scanprosite-doc.html.

Patterns are qualitative motif descriptors

A regular expression is qualitative; it either does match or does not. There is no threshold above which we consider the match as statistically significant. However, it is possible to evaluate the accuracy of PROSITE patterns thanks to the statistics on the number of hits obtained while scanning the SWISS-PROT database[1] or by scanning randomised databases (see below).

Finally, it should be noticed that some

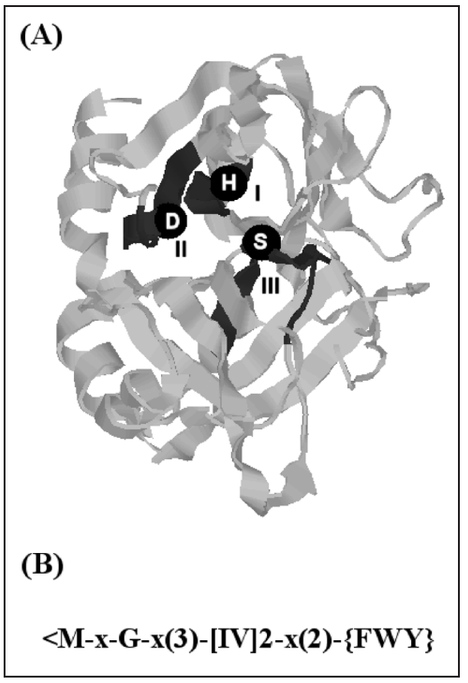

Figure 1: (a) The ScanProsite tool allows the visualisation of PROSITE motifs on 3D structures. The structure shown (1A46) is that of the serine proteases, trypsin domain (PS50240) of the human prothrombin (P00734). The active signatures are shown in dark and the residues involved in the catalytic mechanism as black balls: I, corresponds to the histidine (H) active site pattern (PS00134); II, to a user defined pattern centred on the aspartate (D) active site; and III, to the serine (S) active site pattern (PS00135). The close proximity between the residues involved in the catalytic mechanism is clearly visible. (b) A hypothetical pattern restricted at the N-terminal of a sequence (<) and translated as Met-any residue-Gly-any residue-any residue-any residue-[Ile or Val]-[Ile or Val]-any residue-any residue-{any residue but Phe or Trp or Tyr}. If any mismatch occurs at one of these positions, the pattern will not match

families or domains are defined not just by one pattern but by the co-occurrence of two or more patterns of low specificity. The presence of just one of these patterns is not sufficient to assign a protein to a particular family and/or domain. However, the simultaneous occurrence of linked patterns gives good confidence that the matched protein belongs to the set being considered (see for example the serine protease signatures PS00134 and PS00135, Figure 1a).

The advantages of patterns are their easy intelligibility for the user and the fact that patterns are directed against the most conserved residues. As these residues often are the more relevant for the biological function of the protein family or domain, further research can concentrate on them. Another advantage of patterns is that the scan of a protein database with patterns can be performed in reasonable time on any computer.

PROSITE PROFILES

Profiles are more sensitive than patterns

Although patterns largely proved their usefulness, they also have intrinsic limitations in identifying distant homologues as they do not accept any mismatch. Typical examples of important functional domains that contain only a few very well-conserved sequence positions are the globin, the immunoglobulin, and the SH2 and SH3 domains. The enhanced sensitivity of generalised profiles (or weight matrices) allows the detection of such poorly conserved domains or families. Another advantage of profiles over patterns is that they characterise protein domains over their entire length, not just the most conserved parts of it. This advantage is currently used to define automatically the limits of particular domains in SWISS-PROT entries in order to improve the consistency of the annotation.

Profiles usually correspond to protein domains

Profiles are quantitative motif descriptors

The increased discriminatory power of profiles is due to intrinsic capabilities of the profile descriptor as well as to the sophistication of the profile construction methods. Profiles are quantitative motif descriptors providing numerical weights for each possible match or mismatch between a sequence residue and a profile position. A mismatch at a highly conserved position can thus be accepted provided that the rest of the sequence displays a sufficiently high level of similarity. The automatic procedure used for deriving profiles from multiple alignments is capable of assigning appropriate weights to residues that have not yet been observed at a given alignment position, making for this purpose use of prior knowledge about amino acid substitutability contained in a substitution matrix. In contrast, the procedure by which patterns are usually developed does not allow rational guesses as to which not yet observed residues might be observed in the future.

Despite their obvious advantages, profiles are not superior to patterns for all purposes. In fact the two types of descriptors have complementary qualities. Patterns confined to small regions with high sequence similarity are often powerful predictors of protein functions such as enzymatic activities. Profiles covering complete domains are more suitable for predicting protein structural properties. Hence, a profile (PS50240) is able to detect the structural relationship of the non-enzymatic haptoglobin with the trypsin family of serine proteases (Figure 1a), even though the positions corresponding to the proteolytic active site residues of the proteases are occupied by different amino acids in haptoglobin and, as a consequence, no longer detected by the corresponding patterns (PS00134 and PS00135).[4]

The generalised profiles[5,6] used in PROSITE are an extension of the sequence profiles introduced by Gribskov and coworkers.[7] They are sequence-like linear structures consisting of alternating match and insert positions. A match position corresponds to a domain position, which is typically occupied by a single amino acid. It provides weights for each residue type occupying this position plus a deletion extension penalty. Insert positions contain weights for insertions

relative to the domain model defined by the sequence of match positions. In addition they provide parameters for the opening and closing of a deletion gap, as well as for the initiation and termination of a partial alignment to the profile.

The numerical weights of a profile serve to define a quality score for a profile-sequence alignment. A sequence region that can be aligned to a profile with a score higher than a threshold score is considered a match. Searching for multiple occurrences of a particular domain within the same sequence requires the execution of a dynamic programming algorithm that finds a maximal set of high-scoring profile-sequence alignments above a threshold score. Different alignment modes, such as global or local, are defined by profile-intrinsic parameters.[6] The profile format used in PROSITE comprises fields for so-called accessory parameters which define the search method to be used for a particular domain. They allow specification of appropriate cut-off values, different score normalisation modes, and instructions as to how to treat partly overlapping matches.

Construction of PROSITE profiles

Generalised profiles were not exclusively designed for characterising protein domains and can in principle be generated by many different methods. Most of the current profiles in PROSITE were generated by a standard automatic procedure implemented in the program *pfmake* of the PFTOOLS package. This method is based on Gribskov's original method[8] with modifications published later.[9,10] The development of a profile for a protein domain logically involves several steps as shown in Figure 2.

Patterns and profiles are built from multiple sequence alignments

The first and perhaps most critical part is the generation of a good multiple alignment of domains extracted from complete sequences. Whenever possible, we try to use correct alignments based on available 3D structures. The next step consists of attributing weights to individual sequences of the multiple alignment in order to eliminate bias due to over-representation of subfamilies. The program *pfw* from the PFTOOLS package computes Voronoi weights.[11] Several other methods have been proposed for this purpose (see Durbin *et al.*[12]) and apparently perform about equally well. The weighted multiple alignment is then converted into a so-called 'unscaled profile' (see below) with the aid of the program *pfmake*. This process involves as an intermediate step, the generation of a frequency profile, which is in fact a data structure equivalent to a hidden Markov model (HMM).[13,14] For this purpose, each column of the multiple alignment is mapped to either a match or an insert position of the profile, according to the number of gap characters it contains. By default, columns containing less than 50 per cent gap characters are kept as match positions, the others are assigned to insert positions (see supplemental data available online). The last step in the profile construction process consists of converting the frequency profile (HMM) into a searchable scoring profile equivalent to a profile-HMM.[15] The amino acid frequencies of the match positions are transformed into weights using Gribskov's original formula:

$$\left[M_{ij} = \sum_{j'} s_{jj'} f_{ij} \right]$$

where M_{ij} is the match weight for residue j at profile position i, $s_{jj'}$ the substitution score for residue pair jj' according to the substitution matrix, and f_{ij} the weighted frequency of residue j at profile position i. The substitution scores are usually taken from a PAM[16] or BLOSUM[17] matrix. The resulting numbers represent the weighted average of the substitution scores of the residue in the query sequence compared to the residues observed in the corresponding column of the multiple alignment. According to tests performed by two groups,[9,10] the BLOSUM45 matrix produces profiles

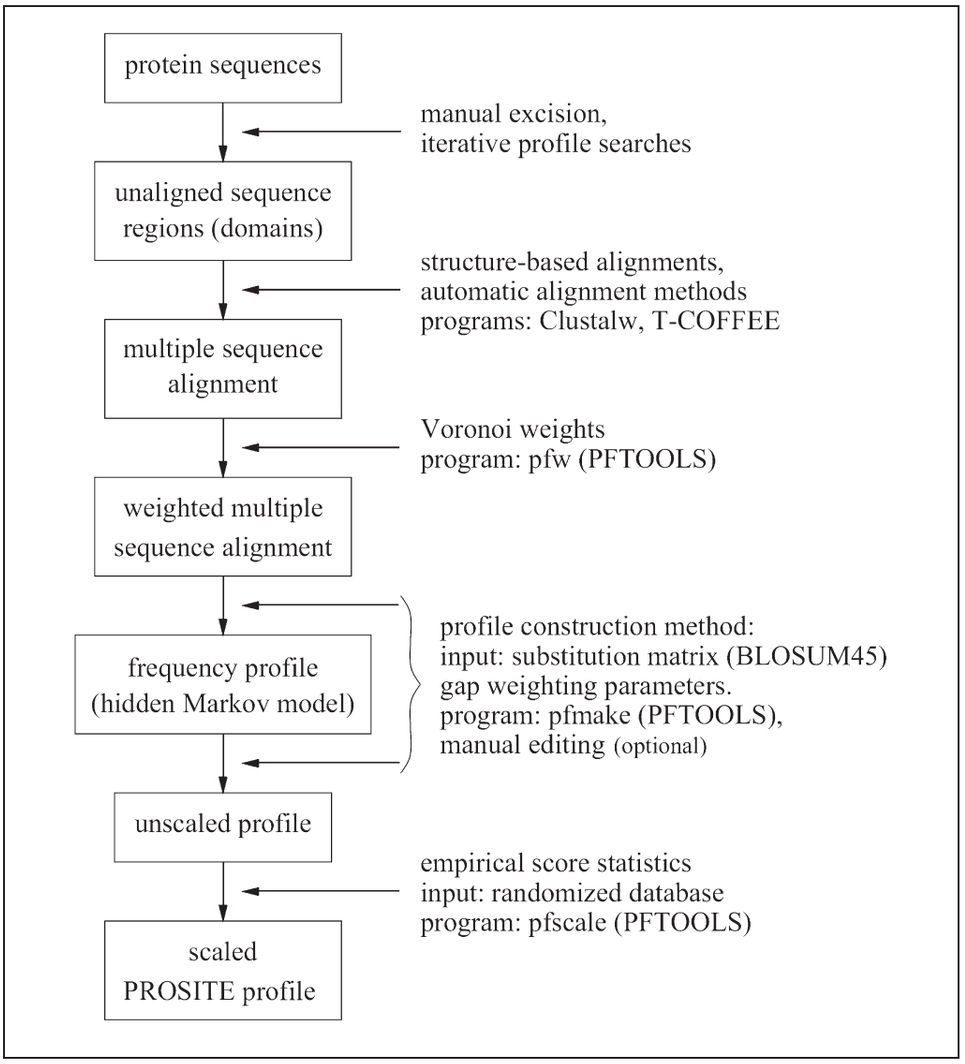

Figure 2: Construction of a PROSITE profile entry. The individual steps of the process are described in the text

The calibration step adjusts the score to a scale common to all predictors

with high sensitivity and selectivity. The so-called transition scores, ie the weights applied to insertions and deletions, are also calculated by a recipe inspired by Gribskov's method. Briefly, the penalty for a gap in the sequence or in the profile is very high at profile positions where no gaps have been observed before. However, it is reduced at positions where gaps do occur in the corresponding multiple alignment in a manner that depends only on the lengths of the gaps observed but not on their frequency. The precise method for scoring gaps used by the program *pfmake* is explained in the supplemental data available on-line.

Profile calibration

The method described above leads to an 'unscaled' profile, which assigns a so-called raw score to a potential match. Raw scores are based on arbitrary units and do not lend themselves to a useful biological interpretation. However, the generalised profile format used in PROSITE allows the definition of various mathematical functions to convert raw scores into more sensible normalised scores. When a large protein database is searched for domains with a profile, one is interested in the question whether a match with a given score is likely to occur by chance or not. To estimate this

probability the raw scores in PROSITE profiles are converted into so-called '$\log_{10}$ per residue *E*-values', allowing the computation of the number of expected matches with equal or higher score in a database of a given size. For instance a match with a normalised score of 9.0 or higher is expected to occur about once in a database of one billion residues. The normalised score is related to the raw score by a linear function whose parameters are equivalent to *K* and λ of the BLAST score statistics.[18] In the jargon of PROSITE, a profile containing a normalisation function that defines $\log_{10}$ per residue *E*-values is called a 'scaled' profile.

An empirical method is used for scaling PROSITE profiles. In the first step of the calibration process, a randomised protein database is searched with the unscaled profile in order to collect high-scoring matches. The cumulative distribution of the 2000 highest scores is then fitted to an extreme value distribution in order to estimate the parameters of the normalisation function. A more detailed description of the numerical recipe can be found at http://hits.isb-sib.ch/doc/motif_score.shtml (see also Hofmann and Bucher[19]). Three types of randomised databases are commonly used for this purpose, each one preserving different properties of real protein sequences such as their length distribution, composition, subfamily relationships and internal repetitiveness, making them more or less suited for a particular type of profile:

Generalised profiles are roughly equivalent to profile-HMMs

- reversed: created by taking the reverse sequence of each individual entry;
- window20: created by local shuffling of each individual sequence entry using a window width of 20 residues;
- db_global: created by global shuffling of each individual sequence entry.

The reversed database is used in most cases but it is not very appropriate for profiles containing regularly spaced repeated residues such as cysteines in zinc fingers or hydrophobic amino acids in helix loop helix domains, because these features are conserved in the reversed database. One of the other two databases may be used in such cases.

A normalised score of 8.5 is typically defined as the default cut-off value in PROSITE profiles. However, in some instances this threshold is not appropriate. There are profiles producing statistically significant matches to members of structurally related protein families, for which a higher cut-off value is indicated. Conversely, for short structural repeats it is sometimes necessary to choose a lower cut-off value to reach satisfactory sensitivity. Such decisions are always based on the match lists for SWISS-PROT presented as quality control information in each entry (see below).

The PROSITE profile format allows specification of multiple cut-off levels. The default level zero is used for the classification of matches as true and false positives and negatives, respectively. Usually, a second low cut-off level with a normalised threshold score of 6.5 is defined for weak matches, which must be interpreted with caution. Nevertheless, they can be very useful for gene discovery and the detection of remote homologues. Additional cut-off levels may be added in order to distinguish subfamilies of proteins or domains or to improve the detection of repeats (see below).

Generalised profiles *vs*. profile-HMMs

Generalised profiles and profile-HMMs are two widely used methods to model protein domains. The generalised profiles used in PROSITE[5] are an extension of the profiles first described by Gribskov *et al.*,[7] while profile-HMMs are a particular case of a class of the HMM probabilistic models.[13,14] Although the two models result from a different historical background, their equivalence has been demonstrated.[6]

A simple introduction to the HMM probabilistic models can be found in

Eddy.[15,20] The important point of profile-HMMs is that they are finite models that describes a probability distribution over an infinite number of possible sequences. A great advantage of profile-HMMs on generalised profiles is that they are formally built on the probability theory (reviewed by Rabiner[21]). The counterpart is that this theory restricts the flexibility of the models because the sum of the probability distribution over all modelled sequences must by definition equal 1. In other words, in a profile-HMM the probability of one sequence cannot be increased without decreasing the probability of another sequence, which makes the manual editing of the model very difficult.

Generalised profiles do not suffer from this restriction, making them very flexible and manipulable in a text editor. For example, it is possible to modify scores in a generalised profile to avoid or force a specific residue or family of residues in a specific position, which may be a way to discriminate two subfamilies of sequence motifs. Modification of the constraints on initiation and termination profile scores provides an easy way to change the behaviour of the model (global, local, semiglobal) and its anchoring to the sequence (no-, left-, right-anchoring).[6]

Although generalised profiles and profile-HMMs are both very effective in detecting motifs in distantly related sequences,[22,23] generalised profiles may be more interesting for the bioinformatician who wants to modify or add some capacities to the models without changing the base algorithms used for the searches.

QUALITY CONTROL OF PROSITE THROUGH SWISS-PROT

The PROSITE and SWISS-PROT databases reciprocally check their quality

The accuracy of PROSITE patterns and profiles can be evaluated thanks to the intimate connection of PROSITE with the SWISS-PROT knowledge base.[1] Each time a new motif descriptor is added to PROSITE, it is used to scan SWISS-PROT in order to attribute one of the following match statuses to the SWISS-PROT entries concerned by the motif:

- True positive: a protein belonging to the set being considered and matched by the motif.
- False positive: a protein that does not belong to the set being considered but picked up by the motif.
- False negative: a protein belonging to the set being considered but not detected by the motif.
- Unknown: a protein that could belong to the set being considered and matched by the motif.
- Partial: a protein belonging to the set being considered but not detected by the motif (possibly because its sequence is incomplete and the region, which should be detected by the motif, missing).

Reciprocally, every new protein entering SWISS-PROT is checked for the occurrence of PROSITE patterns and profiles and a match status for the relevant PROSITE entries is assessed. At every new PROSITE release, these SWISS-PROT match statuses are used to establish statistics for most motifs. These statistics allow the user to evaluate the ability of a motif to detect all or most of the sequences it is designed to describe (sensitivity) as well as its ability to give as few false positive results as possible (specificity). In addition, this process allows motifs to be permanently improved to give a better fit to the increasing number of proteins in SWISS-PROT.

Skipping short and degenerate motifs

Some PROSITE entries are too short or degenerate to have a biological meaning by their own as they are found in the majority of known protein sequences. These motifs, some of which predict post-translational modification sites (eg

N-glycosylation; PS00001), produce matches that are only indicative of a possible function. Independent biological evidence must be considered to confirm the appropriateness of these matches. There is no matchlist and hence statistics provided with these motifs as well as with compositional profiles, which do not characterise biologically defined objects but are directed against sequence regions enriched in a particular amino acid. As these profiles are only defined statistically, it is not possible to speak of true or false matches to these profiles, neither is it possible to assign a false negative status to a sequence.

Some pre-release profiles are not yet in PROSITE, but still available for users

PROSITE motifs belonging to these classes are tagged with the */SKIP-FLAG = TRUE* qualifier in their CC lines (for a definition of all the PROSITE lines see http://www.expasy.org/prosite/prosuser.html.

PROSITE DOCUMENTATION

Each PROSITE motif descriptor is linked to a document providing biological knowledge

Each PROSITE motif is linked to a corresponding documentation describing the protein family or domain it detects. The documentation contains a brief description of what is known about this particular protein family or domain: origin of its name, taxonomic occurrence, domain architecture of the proteins, function, 3D structure, main characteristics of the sequence and some references. Recently, for families or domains whose structure is known, a direct link to a representative PDB entry is provided in the documentation, in order to make the description of the 3D structure more comprehensible. All the information providing biological knowledge about a protein family or domain should also be used as an additional quality control for patterns and profiles. If the user has some information about its sequence that makes no sense with the description of the motif detected, the match should be considered with caution.

The documentation also contains direct information about the motif descriptors. Hence, for patterns the amino acid residues involved in the catalytic mechanism, metal ion or substrate binding or post-translational modifications are indicated. For profiles, residues are given if they cover the entire domain or protein.

Finally, the sensitivity and specificity of the motif are also indicated, as well as an expert to contact if necessary.

PRE-RELEASE PROFILES AND DOCUMENTATIONS

In addition to PROSITE profiles, there is a collection of pre-release profiles and of their corresponding preliminary documentations (QDOC) that is not integrated in PROSITE but available with InterPro. There can be several reasons why a profile and its documentation are considered as preliminary and not integrated in PROSITE. First, some profiles can be considered as under development and will need to be seriously redefined before their eventual integration in PROSITE. The second class represents profiles and documentation whose quality still needs some improvement in order to reach the PROSITE standards. Finally, the last ones can already be considered as PROSITE profiles. They are just waiting for their integration, which takes some time because of the intimate connection of PROSITE with SWISS-PROT: an integration of a profile not only concerns PROSITE but also implies more or less important changes in SWISS-PROT.

FUTURE DEVELOPMENTS

Usually it is difficult to detect all repeated copies of a domain in a protein because the most degenerate repeats are generally missed. We plan to improve the detection efficiency of repeats by introducing a new threshold for repeats. Once a repeat above the normal threshold is detected, this new lower threshold would be used to detect the additional degenerate copies of the same repeat. By this way most, if not all, copies of a repeat should be detected.

We have already constructed some

profiles using structural information and now want to study if this approach improves the sensitivity and specificity of profiles by restricting the position of insertions or deletions to particular position having only little effects on the 3D structure of the protein/domain-like loops.

Another project is to improve the automated annotation of SWISS-PROT thanks to PROSITE. Some information contained in PROSITE could help SWISS-PROT annotators in the annotation of domain or catalytic site features. Such an approach would ensure the consistency of SWISS-PROT.

The PROSITE database and programs to use it are available online or for download

HOW TO OBTAIN A LOCAL COPY OF PROSITE

A list of servers which distribute PROSITE has been recently published,[3] but please note that though PROSITE is free for academic users, the documentation entries are under copyright regulations. To obtain a licence, commercial users should e-mail the Swiss Institute of Bioinformatics: license@isb-sib.ch.

HOW TO MAKE USE OF PROSITE

Computer programs

We provide programs that have been specifically developed to help use PROSITE for both patterns and profiles searches:

- *ps_scan*[24], a program used to scan one or several PROSITE motifs against one or several protein sequences. ps_scan is available from ftp://ftp.expasy.org/databases/prosite/tools/.

- *PFTOOLS*, programs used to construct profiles or scan a sequence or a sequence library against a profile or a profile library. PFTOOLS are available from ftp://ftp.expasy.org/databases/prosite/tools/or ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools/.

Interactive Web access to PROSITE

To browse the PROSITE documentation and motif entries, users should go to http://www.expasy.org/prosite/. Web access to PROSITE allows users to benefit from the latest PROSITE updates and from hyperlinks connecting a PROSITE entry to other relevant sources of information. In addition, it has recently been made possible for the user to display the match list of a PROSITE motif as a multiple alignment available in different formats.

To scan a sequence for PROSITE motifs, one can make use of the following tools.

ScanProsite

ScanProsite[24] allows either to scan a protein sequence – from SWISS-PROT or provided by the user – for the occurrence of PROSITE motifs or to scan the SWISS-PROT, TrEMBL and/or PDB databases for the occurrence of a pattern that can originate from PROSITE or be provided by the user. ScanProsite also allows the user to visualise the position of a PROSITE motif or of his own pattern on the 3D structure (if known) of the matched proteins (Figure 1a). Recently, we added the possibility for the user to evaluate the specificity of a pattern by using it to scan a randomised version of the current SWISS-PROT database. The URL for ScanProsite is http://www.expasy.org/tools/scanprosite.

ProfileScan

ProfileScan allows a protein sequence – from SWISS-PROT or provided by the user – to be scanned for the occurrence of profiles stored in PROSITE and in the pre-release collection. The new URL for ProfileScan is http://hits.isb-sib.ch/cgi-bin/PFSCAN.

Acknowledgements

PROSITE is supported by grant no. 3100–63879.00 from the Swiss National Science Foundation.

References

1. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28(1), pp. 45–48.
2. Berman, H. M., Westbrook, J., Feng, Z. *et al.* (2000), 'The Protein Data Bank', *Nucleic Acids Res.*, Vol. 28(1), pp. 235–242.
3. Falquet, L., Pagni, M., Bucher, P. *et al.* (2002), 'The PROSITE database, its status in 2002', *Nucleic Acids Res.*, Vol. 30(1), pp. 235–238.
4. Kurosky, A., Barnett, D. R., Lee, T.-H. *et al.* (1980), 'Covalent structure of human haptoglobin: a serine protease homolog', *Proc. Natl Acad. Sci. USA*, Vol. 77(6), pp. 3388–3392.
5. Bucher, P. and Bairoch, A. (1994), 'A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation', in 'Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 53–61.
6. Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996), 'A flexible motif search technique based on generalized profiles', *Comput. Chem.*, Vol. 20(1), pp. 3–23.
7. Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987), 'Profile analysis: detection of distantly related proteins', *Proc. Natl Acad. Sci. USA*, Vol. 84(13), pp. 4355–4358.
8. Gribskov, M., Lüthy, R. and Eisenberg, D. (1990), 'Profile analysis', *Methods Enzymol.*, Vol. 183, pp. 146–159.
9. Lüthy, R., Xenarios, I. and Bucher, P. (1994), 'Improving the sensitivity of the sequence profile method', *Protein Sci.*, Vol. 3(1), pp. 139–146.
10. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'Improved sensitivity of profile searches through the use of sequence weights and gap excision', *Comput. Appl. Biosci.*, Vol. 10(1), pp. 19–29.
11. Sibbald, P. R. and Argos, P. (1990), 'Weighting aligned protein or nucleic acid sequences to correct for unequal representation', *J. Mol. Biol.*, Vol. 216(4), pp. 813–818.
12. Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. (1998), 'Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press, Cambridge.
13. Krogh, A., Brown, M., Mian, I. S. *et al.* (1994), 'Hidden Markov models in computational biology. Applications to protein modeling', *J. Mol. Biol.*, Vol. 235(5), pp. 1501–1531.
14. Baldi, P., Chauvin, Y., Hunkapiller, T. and McClure, M. A. (1994), 'Hidden Markov models of biological primary sequence information', *Proc. Natl Acad. Sci. USA*, Vol. 91(3), pp. 1059–1063.
15. Eddy, S. R. (1996), 'Hidden Markov models', *Curr. Opin. Struct. Biol.*, Vol. 6(3), pp. 361–365.
16. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978), 'A model of evolutionary change in proteins', in Dayhoff, M. O., Ed., 'Atlas of Protein Sequence and Structure', Vol. 5, National Biomedical Research Foundation, Washington, DC, pp. 345–352.
17. Henikoff, S. and Henikoff, J. G. (1992), 'Amino acid substitution matrices from protein blocks', *Proc. Natl Acad. Sci. USA*, Vol. 89(22), pp. 10915–10919.
18. Karlin, S. and Altschul, S. F. (1990), 'Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes', *Proc. Natl Acad. Sci. USA*, Vol. 87(6), pp. 2264–2268.
19. Hofmann, K. and Bucher, P. (1995), 'The FHA domain: A putative nuclear signalling domain found in protein kinases and transcription factors', *Trends Biochem. Sci.*, Vol. 20(9), pp. 347–349.
20. Eddy, S. R. (1998), 'Profile hidden Markov models', *Bioinformatics*, Vol. 14(9), pp. 755–763.
21. Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proc. IEEE*, Vol. 77, pp. 257–286.
22. Hofmann, K. (2000), 'Sensitive protein comparisons with profiles and hidden Markov models', *Brief. Bioinform.*, Vol. 1(2), pp. 167–178.
23. Karplus, K., Barrett, C. and Hughey, R. (1998), 'Hidden Markov models for detecting remote protein homologies', *Bioinformatics*, Vol. 14(10), pp. 846–856.
24. Gattiker, A., Gasteiger, E. and Bairoch, A. (2002), 'ScanProsite: a reference implementation of a PROSITE scanning tool', *Applied Bioinform.*, Vol. 1(2), pp. 51–52.