PROGRESS IN BIOLOGICAL SCIENCES

iProsite: an improved prosite database achieved by replacing ambiguous positions with more informative representations

Mohammad-Hadi Foroughmand-Araabi[1], Bahram Goliaei[1, *], Mehdi Sadeghi[2]

[1]Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
[2]National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran 14155-6346, Iran.

Received: December 15, 2012; Accepted: March 1, 2013

Abstract

PROSITE database contains a set of entries corresponding to protein families, which are used to identify the family of a protein from its sequence. Although patterns and profiles are developed to be very selective, each may have false positive or negative hits. Considering false positives as items that reduce the selectiveness of a pattern, then, the more selective pattern we have, a more accuracy in protein family detection we will get. In this paper, we have provided a method for improving the PROSITE patterns by reconstructing them in a manner that they not only still match to true positive hits, but also match to less false positive hits. From 973 PROSITE patterns, 283 have been improved by our method. We have applied the provided method on the PROSITE database and the improved resulting database is available at http://cbp.ut.ac.ir/iPROSITE.

Keywords: Protein sequence pattern, Protein database, Protein family, PROSITE

Introduction

To determine the function of a protein sequence, usually, the first step is to align it to a sequence database. Achieving a significant alignment indicates that aligned sequences perform similar functions. This approach is useful in the case that a significant alignment could be found. However more commonly, we will find a set of partial matches to diverse proteins, which may not help in detecting the role of the new sequence. Considering this difficulty, an alternative method is to search the new sequence in *pattern* databases, in contrast to *sequence* databases. That is because pattern databases selectively specify the function of a protein sequence.

Pattern databases are developed by grouping related proteins, in function or structure. Then, to each group (family) a pattern is assigned based on sequence similarities between proteins of that group, which helps to distinguish members of this group from others. These patterns represent parts of protein sequences which are responsible for the function of the members of that group (1).

Different approaches for grouping proteins, and representing patterns, have given rise to different pattern databases, such as PROSITE (2), PRINTS (3), Pfam (4), and InterPro (5). Among various available protein pattern databases, one of the most popular is the PROSITE database. The PROSITE database contains protein sequence sites that involve in protein functions such as: enzyme catalytic sites, prosthetic group attachment sites, binding a metal ion or ligand, disulphide bonds, and etc. Also, each PROSITE entry contains extensive information on nomenclature, function, sequence features, and important literature references (2).

PROSITE database represents its patterns in two different formats, namely, patterns and profiles. In this paper, we have focused on PROSITE pattern entries. PROSITE patterns are suitable for representing small conserved regions. On the other hand, PROSITE profiles are good for representation of a whole protein domain. The residues which are specified by PROSITE patterns are often more relevant for the biological function (2), and thus, we have chosen PROSITE *pattern*

* Corresponding author: goliaei@ibb.ut.ac.ir
Tel.: +98-21-664-98672

entries in this study.

PROSITE patterns are used to detect function of a protein sequence, thus, having more accurate entries leads to better tools for analyzing protein sequences. Consequently, more false positives for a pattern lead to more wrong function assignment to an unknown sequence. In this paper, we aim to reduce the number of false positives for PROSITE patterns.

False positives for a PROSITE pattern could be found in information attached to entries. Technically speaking, PROSITE pattern entries contain information about the accuracy of the provided pattern. List of proteins which are related to the patterns are provided, for each entry. These proteins are categorized into five categories:

- True positive hits: Proteins which belong to the family and also match to the pattern.
- False positives: Proteins which does not belong to the family but falsely match to the provided pattern.
- False negatives: Proteins which belong to the family but do not match to the provided pattern.
- Unknown proteins: Proteins which match to the provided pattern, but, there are not enough confidence for them to be considered as family members, i.e. according to the biological literature this protein is not known to be a member of this family.
- Partial proteins: Proteins which belong to the family, but, since their sequence are partially achieved, they do not match to the pattern.

For each pattern, in addition to the number of *true* and *false positives*, the numbers of *true* and *false positive sequences* are defined. That is because a pattern may match a sequence at more than one position. As the number of *true positives* and *false positives*, all the matching places will be considered, even if they lie on one sequence. In contrast, in counting the number of *true positive sequences* or *false positive sequences*, only the numbers of such sequences (and not such matching places) are considered.

Our paper is the first work that reduces the number of false positive of PROSITE patterns without inclusion of any additional information. Some previous works added the secondary (6) and tertiary structure (7-9) of proteins to the pattern to obtain more selective patterns. In these works, obtained patterns has two parts, the first part, which is similar to PROSITE patterns, describes the sequence of the protein, while the second part describes the secondary or tertiary structure of the protein. In contrast, we contribute a computational approach that produces sequential patterns. The new database which is obtained from this approach is called iPROSITE.

Materials and methods

As the core of our study, we have considered *pattern* entries of PROSITE database. A pattern entry is defined as a sequence of acceptable amino acids. An acceptable amino acid set in a sequence may be represented as one of the following forms:

- One character which is the standard International Union for Pure and Applied Chemistry (IUPAC) one-letter code for an amino acid.
- A set of acceptable amino acids which is surrounded by square brackets “[]”.
- A set of non-acceptable amino acids which is surrounded by braces“{}”.
- Character “x” which stands for any possible amino acid.

For example, the pattern with ID PS00118 in the PROSITE database is represented as “C-C-{P}-x-H-{LGY}-x-C.” This pattern represents a sequence of eight consecutive positions. In this pattern, the first and second positions contain a “C” amino acid. The third position contains an amino acid which is not “P”. The forth position may contain any amino acid. Other positions follow a similar rule. In this study, we have not considered patterns with unknown amino acids, such as those patterns that contain the character “U”.

We have reconstructed pattern positions based on their true positive hits. In order to reconstruct pattern positions, we have extracted protein sequences from the Uniprot database (10). Then, we have aligned PROSITE patterns with these sequences, and found the matching subsequences. Then, we have tried to modify PROSITE to obtain patterns, with exactly the same sets of true positive hits, and reduced numbers of false positives. Particularly, for each pattern position, we have selected the amino acids which appear in the corresponding position of true positive hits.

Table 1. Alignment of PROSITE pattern PS01088 with its positive hits and false positives.

					Pattern & Matc hes									
Type of hit Pattern	**Protein**	**[LIVM](2) x**			**R**	**L [DE]**		**x(4)**				**R**	**L**	**E**
Positive hits	CAP1 BOVIN	L	V	E	R	L	E	R	V	V	G	Z	L	E
	CAP1 HUMAN	L	V	E	R	L	E	R	V	V	G	R	L	E
	CAP1 MACFA	L	V	E	R	L	E	R	A	V	G	R	L	E
	CAP1 MOUSE	L	V	E	R	L	E	R	A	V	G	R	L	E
	CAP1 PONAB	L	V	E	R	L	E	R	A	V	G	R	L	E
	CAP1 RAT	L	V	E	R	L	E	R	A	V	G	R	L	E
	CAP2 HUMAN	L	V	E	R	L	E	R	A	V	S	R	L	E
	CAP2 MOUSE	L	M	E	R	L	E	R	A	V	I	R	L	E
	CAP2 PONAB	L	V	E	R	L	E	R	A	V	S	R	L	E
	CAP2 RAT	L	M	Q	R	L	E	F	A	V	S	R	L	E
	CAP DICDI	L	L	K	R	L	D	Q	A	T	T	R	L	E
	CAP HYDVD	L	V	S	R	L	E	A	V	T	N	R	L	E
	CAP SCHPO	I	L	K	R	L	E	A	A	T	S	R	L	E
	CAP YEAST	L	L	K	R	L	E	E	A	T	A	R	L	E
False positives	CLMP1 CRYNB	L	V	Q	R	L	D	V	E	S	A	R	L	E
	CLMP1 CRYNJ	L	V	Q	R	L	D	V	E	S	A	R	L	E

Finally, we have created the improved version of the pattern by concatenating these newly obtained sets of acceptable amino acids.

In order to describe the fundamental essence of the proposed method, we discuss it by an example. Consider the PROSITE pattern with ID PS01088, which is "[LIVM](2)-x-R-L-[DE]-x(4)-R-L-E". This pattern contains 16 true, and 2 false positive hits. The alignment of this pattern with true and false positive hits are represented in Table. 1. Considering 9th position of this alignment (i.e. the position of the third "x"of "x(4)"), true positive hits in this position only contain amino acids "V" and "T", while, false positive hits contain "S". Thus, we propose to change the description of this position from "x" to "[VT]". After this modification, proteins CLMP1_CRYNB and CLMP1_CRYNJ are no longer false positive hits for this pattern. Note that, we apply this method to every position of every pattern, thus, we may change the acceptable amino acid set for more than one position.

In some cases, the above approach may produce biologically meaningless sets of amino acids for a position. In order to resolve this issue, we consider the physicochemical properties of amino acids that form an acceptable set. Since this property is also taken into account in the development of PROSITE patterns, we have restricted our patterns to use amino acid sets that appear in PROSITE patterns. To do this, after reconstruction of patterns, we have replaced the amino acid sets with the smallest amino acid set which has the following properties:

1- It is a superset of the obtained amino acid sets.

It is already represented in at least k PROSITE patterns, for a specific number k.

2- It is a subset of the acceptable amino acid set in the old pattern.

3- Let *S* be the obtained set of amino acids, *O* be the original set of amino acids, and A be the family of sets of amino acids that appear in PROSITE patterns at least *k* times. If *O* is a member of *A*, then, *O* is an appropriate choice for acceptable amino acid set. Otherwise, we have to choose a set *X* in *A* which is a superset of *S* and a subset of *O*. If there are more than one set with this property, we choose the smallest set. If the smallest set is not unique, we choose the one which appears more in the PROSITE database. In the case of equal number of appearances of these sets, we choose one randomly. Moreover, there may be no set with this property, that happens when no members of *A* is superset of *S* and subset of *O*, at the same time. In this case, we do not change the acceptable amino acid set for this position, i.e. we choose the set *O*.

We can easily show that our method does not produce any new false positive or false negatives. Since the acceptable amino acid set in improved patterns are supersets of obtained amino acids from true positive hits, then obviously, the new pattern matches to all true positive hits of the original pattern. Also, since we only replaced acceptable amino acid sets with their subsets, no protein might be added to false positive hits.

Results and Discussion

The PROSITE database is obtained by a semi-manual method. In contrast, we have proposed a computational method to improve the patterns of PROSITE database. This computational approach leads to reduction of number of false positive hits for 283 PROSITE patterns. Indeed, PROSITE database contains 973 patterns with fixed length, from which, 283 (29%) have been improved by our method. The improved database is publicly available at http://cbp.ut.ac.ir/iPROSITE.

Number of improvements for different values of k are presented in Fig. 1, where, as mentioned above, *k* is the minimum acceptable appearance frequency in PROSITE, for new amino acid sets. This chart shows that the improvement is not very sensitive to small changes of this parameter.

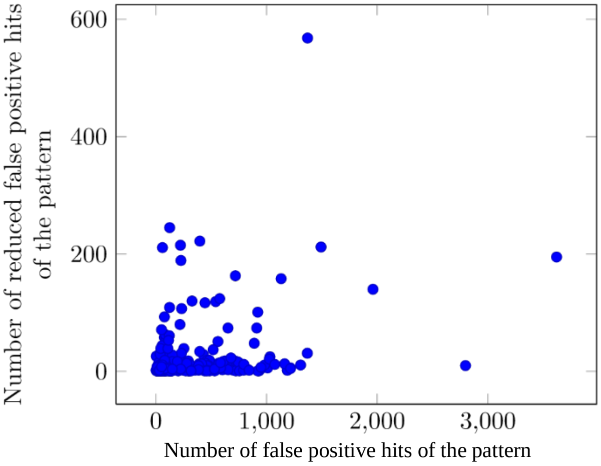

Figure 1. If we only let the pr ovided method to choose from acceptable sets which are appearing in the PROSITE patterns at least k times, for different values of k, the number of improvements are presented.

The method which is presented in this paper, removed 2438 false positive hits (out of 6102) and removed 2429 false positive sequences (out of 6050) from the PROSITE database. The number of reduced false positives versus the number of false positives of the original pattern is represented in Fig. 2. Clearly, the number of reduced false positives for a pattern is less than or equal to the number of false positives of the original pattern. Thus, there should be no point above the line with slope 1 originating from the origin. Indeed, there are many points lying near this line. These points represent the patterns that originally have some false positives, but after improvement they do not have any false positives. This shows the high efficiency of the proposed method.

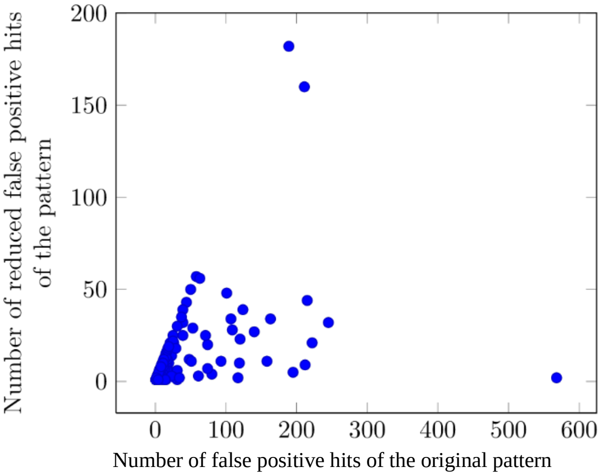

Figure 2. Number of reduced false positives versus number of false positives of the original pattern.

In comparison to related works, Skrabanek and Niv (6) studied the patterns for which the information of secondary structure was available. From these 763 pattern, they increased the selectivity of 26 patterns. Via and Helmer-Citterich (7), and Lin, et al. (8), studied 8 and 12 patterns, respectively. They improved these patterns in selectivity or sensitivity, by considering tertiary structures. Milledge, et al. (9) combines the information of PROSITE patterns with the information which is provided by structural databases to obtain better results. They claimed that they have improved 90% of protein patterns for which enough tertiary structure is available. All these works rely on the information of the secondary or tertiary structure of the query protein. In other words, these methods are not able to assign function to proteins for which we do not have secondary or tertiary structure. In contrast, we are able to search our obtained patterns in protein sequences without considering any further information.

The number of reduced false positive hits versus the original number of true positive hits is represented in Fig. 3. For a PROSITE pattern, consider all possible protein subsequences that match to this pattern. Number of such subsequences is equal to the multiplication of number of acceptable amino acids for pattern positions. For example, the pattern “[LIVM](2)-x-R-L-[DE]-x(4)-R-L-E” has $4^2 \times 20 \times 2 \times 20^4 = 102{,}400{,}000$ possible matching protein subsequences. We name this number as the *acceptable volume* of the pattern. However, not all these possible subsequences appear in real proteins. If real protein sequences have been distributed randomly, the number of real proteins that match to a pattern is proportionally related to the acceptable volume of the pattern. Obviously, an improved pattern, in comparison to the original pattern, has a smaller acceptable volume. Therefore, if real protein sequences have been distributed randomly, the ratio of reduced hits should be equal to the ratio of reduced acceptable volumes. On the other hand, this ratio is equal to the number of true positive hits divided by the number of reduced hits. Thus, we expect that, the points in Fig. 3 would lie on a straight line, which is not the case here. It shows that not only the distribution of true positive hits is not random, but also, true positive hits are concentrated around the center of the cubes corresponding to patterns.

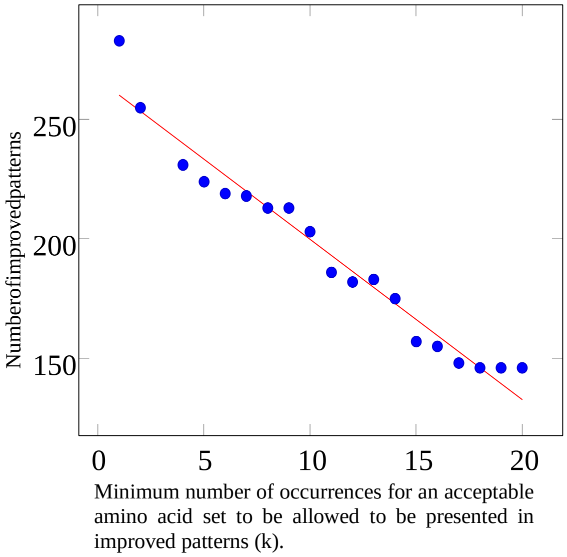

Figure 3. Number of reduced false positives versus number of true positive hits of the original pattern.

The technique which we have provided for improvement of PROSITE patterns is based on the format of PROSITE *pattern*s. This technique could not be extended to other databases or even position specific matrix patterns in PROSITE, in a straightforward manner. As a future work, we will provide similar techniques for different pattern databases, such as Pfam and BLOCKS. Also, we will provide other techniques for improving PROSITE database by reducing number of false negatives.

References

1. Attwood,T. (2000) The role of pattern databases in sequence analysis. *Brief. Bioinform.*, **1**, 45-59.
2. Sigrist,C. and Cerutti,L. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161-D166.
3. Attwood,T. and Bradley,P. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400-402.
4. Punta,M., Coggill,P., Eberhardt,R., Mistry,J., Tate,J. and Boursnell,C. (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290-D301.
5. Hunter,S., Jones,P. and Mitchell,A. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*
6. Skrabanek,L. and Niv,M. (2008) Scan2S: Increasing the precision of PROSITE pattern motifs using secondary structure constraints. *Proteins*, **72**, 1138-1147.
7. Via,A. and Helmer-Citterich,M. (2004) A structural study for the optimisation of functional motifs encoded in protein sequences. *BMC. bioinformatics.*, **5**, 50.
8. Lin,K., Wright,J. and Lim,C. (2000) Conformational analysis of long spacers in PROSITE patterns. *J. Mol. Biol.*, **299**, 537-548.
9. Milledge,T., Khuri,S., Wei,X., Yang,C., Zheng,G. and Narasimhan,G. (2005) Sequence Structure Patterns: Discovery and Applications, Proc CBGI'05.
10. Magrane,M. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*.