

Systematic and Automated Discovery of Patterns in PROSITE Families

Reece Hart, Ajay K. Royyuru, Gustavo Stolovitzky, and Andrea Califano

IBM Computational Biology Center, TJ Watson Research Center

PO Box 704, Yorktown Heights, NY 10598

email: {rkh, ajayr, gustavo, acal}@us.ibm.com

Abstract

PROSITE is a method for protein classification which relies on a database of biologically significant sites and patterns in protein sequences. Most patterns in PROSITE have been gathered by a labor intensive combination of experimental characterization of functional residues and sequence alignment. In this paper we present a new and efficient supervised learning procedure, based on the Splash deterministic pattern discovery algorithm and on a framework to assess the statistical significance of patterns. We demonstrate its application to the fully automatic discovery of patterns in 974 PROSITE families. For these families, Splash generates patterns with better specificity and/or sensitivity in 28%, identical statistics in 48%, and worse statistics in 15% of the cases; for the remaining families, patterns exhibited mixed behavior. Second, we have characterized the amount of overlap, on the sequences, between newly discovered patterns and those in PROSITE. In about 75% of the cases, Splash patterns identify sequence sites that overlap more than 50% with those reported in PROSITE. Of the 272 patterns which perform strictly better than the corresponding PROSITE pattern, 178 show more than 70% overlap with the PROSITE pattern. Third, our results suggest that the statistical significance of discovered patterns correlates well with their biological significance. Finally, we use the trypsin subfamily of serine proteases to illustrate the use of this method to exhaustively discover all motifs in a family that are statistically and biologically significant. The complete analysis is sufficiently rapid, taking less than a day for all PROSITE families, to enable the use this methodology for routine curation of existing motif and profile databases.

1. Introduction

Rapid advancement in sequencing technology and exponential growth in genomic databases are spurring the development of techniques for the identification of sequence motifs and sequence classification. This is commonly accomplished by defining sequence signatures that distinguish all members of the respective family from the complete sequence database and allow the classification of new proteins into these families [1]. Definition of the sequence signatures can range from simple consensus patterns in sequences, often called sequence motifs, to more elaborate and rigorous descriptors, termed position specific scoring matrices or profiles, to Hidden Markov Models [2]. Currently, there are several well curated and established compilations of sequence motifs, such as PROSITE [3], PRINTS [4], PFAM [5], and BLOCKS [6]. The latter is a comprehensive non-redundant database of blocks derived from several databases of sequence motifs and profiles.

The differences between each approach depend on the method used to generate sequence motifs and profiles, on whether these are used individually or jointly, on whether they are gapped or ungapped, and on whether they have to match exactly or through a scoring function. A statistical framework for combining motif scores has been proposed [7] and can be used for sequence classification using multiple PSSM models.

Among these databases, PROSITE is especially relevant because of the high biological significance of the reported patterns. The PROSITE database version 15.0 contains extensively annotated collections of 1352 motifs grouped into 1014 protein families. Each PROSITE entry stems from a set of protein sequences grouped by an expert, using biological information which is provided as documentation. For almost all entries, PROSITE provides a sequence motif that characterizes the functionally relevant residues of a protein family. These are obtained by selecting regions of sequences that have a documented functional significance and by performing multiple sequence alignment over these selected regions to identify consensus patterns. Beyond the inherent utility of the simple consensus patterns, PROSITE also serves as a very useful database of seed locations to guide the automated development of more complex descriptors, as in BLOCKS. Similar seed entries are created semi-manually for other curated databases as well, such as PRINTS and PFAM. Efforts are underway to integrate these into a single resource called InterPro (www.ebi.ac.uk/interpro). Curation of these databases is a

labor-intensive task, which is increasingly challenged by the rapid explosion of data in genomic repositories. Efforts to automate the seed generation process have resulted in several new methodologies [8] for the identification of conserved sequence motifs such as MEME [9], the Gibbs Sampler [10], Pratt [11], EMOTIF [12], Splash [13], and Teiresias [14].

Due to the high quality of the annotations in PROSITE, it is useful to systematically and objectively compare the results of any new technique with those reported by PROSITE, both in terms of sensitivity and specificity (false positives and false negatives) and of the ability to identify regions that are biologically significant [15].

Here we report on the systematic application and evaluation of Splash [13], a deterministic pattern discovery algorithm, in combination with a framework for the analysis of the statistical significance of patterns [16] and demonstrate its application to the identification of highly conserved motifs in protein families. In Section 3, we report the results of the analysis of each one of the 974 protein families associated with one or more PROSITE motifs. In Section 3.2, we study the sensitivity and specificity of the automatically discovered patterns. In Section 3.3, by correlating the statistical significance of patterns with their performance, we show that this statistical criterion is meaningful. Indeed, as shown in Section 3.4, it often leads to the identification of regions in protein sequence that are biologically significant. The latter is determined by measuring the overlap between automatically discovered patterns and PROSITE patterns. In Section 3.5 we illustrate instances where our procedure suggests refinements of the biologically relevant PROSITE patterns that improve their sensitivity and/or specificity. Some examples in which PROSITE patterns perform better than Splash patterns will be discussed as well to determine how to further refine this procedure. Finally, we report on the exhaustive discovery and analysis of the motifs that occur in at least 40% of the sequences in the trypsin family. These are compared with motifs reported in PROSITE and BLOCKS.

This paper demonstrates the utility of automated pattern discovery methods for the maintenance of current databases. We emphasize that the entire process is automated and performed identically for all PROSITE families; we have made no effort to tune parameters for specific families. Although we have used PROSITE families for analysis and comparison, we have not made use of the pattern information contained in a PROSITE record in the automated pattern discovery process. The tabulated analysis of all PROSITE families will be made available on our web site at the time of publication.

2. Methods

PROSITE regular expressions are described in [3]. These patterns, although defined rigorously for matching, are not easily generalized to a consistent set of rules for automatic pattern discovery purposes. For instance there is no apparent *a priori* rule for the selection of residues in square brackets. Rather, as observed in [12] patterns are refined *a posteriori* to improve performance.

In this paper we will focus exclusively on rigid patterns, that is patterns that, in the PROSITE representation, do not contain variable length gaps such as $-x(a, b)-$, where “x” is the wildcard character and a and b are two integers. The sets of residues that are considered similar and can be included within square brackets is determined via a substitution matrix $M(i, j)$, such as PAM [17] or BLOSUM [18]. For each i -th residue a similarity set $K(i)$ is defined by selecting each other residue j such that $M(i, j) > m_0$. Throughout this paper, we will use BLOSUM 50 as a substitution matrix and $m_0 = 0$. This results in the following set:

TABLE 1. Similarity sets for BLOSUM 50 with a threshold $m_0 = 0$

$K(A) = \{AS\}$	$K(D) = \{DENBZ\}$	$K(I) = \{ILMV\}$	$K(M) = \{MILV\}$
$K(S) = \{STNA\}$	$K(Y) = \{YHFW\}$	$K(R) = \{RKQ\}$	$K(C) = \{C\}$
$K(G) = \{G\}$	$K(L) = \{LIMFV\}$	$K(F) = \{FYLW\}$	$K(T) = \{TS\}$
$K(V) = \{VILM\}$	$K(N) = \{NDHSB\}$	$K(Q) = \{QEKRHZ\}$	$K(H) = \{HYQN\}$
$K(K) = \{KRQEZ\}$	$K(P) = \{P\}$	$K(W) = \{WYF\}$	$K(E) = \{EDQKBZ\}$

This set is augmented by each individual amino acid, if not already present in a group by itself, such as it is the case for cysteine, glycine, and proline. Characters from this augmented set, Ψ , will be called *tokens*. These are divided into *identity tokens*, such as L , and *similarity tokens*, such as $[FYLW]$. Patterns composed of tokens

from Ψ and wildcards will be called similarity patterns. For convenience we shall represent the wildcard as a “.” (dot) rather than a “x.” $\mathbf{x}(a)$, therefore, becomes a string with a dots.

2.1 Pattern Discovery

Splash [13] can deterministically discover all maximal similarity patterns of the form $\Psi(\Psi \cup \{.\})^* \Psi$, which occur in at least $j \geq j_0 \geq 2$ independent protein sequences in a protein sequence database. Here, j is called the pattern support. Patterns with k tokens and l total characters are called kl -patterns. kl -patterns with support j are called jkl -patterns.

Density Constraint: To restrict the number of possible patterns and avoid creating patterns that are too sparse, which tend to have low statistical significance, we impose that patterns contain at least k_0 tokens over any stretch of l_0 consecutive characters starting with a token. Because k_0/l_0 can be interpreted as a token density per unit length, we call this the *density constraint*.

Biologically significant patterns with very low densities are not frequent. For instance, PROSITE reports only one rigid pattern with more than 20 consecutive wildcards (PS001254). This is because long stretches of wildcards occur over loops, where insertions and deletions are likely. Very sparse and flexible patterns are not frequent either. The PROSITE database contains only 8 flexible patterns with more than 20 consecutive wildcards.

Maximal Patterns: Splash reports only patterns that are maximal. A pattern is said to be maximal if no token can be added to it without reducing the number of matches in the database. For example, given the two sequences WQKCCDDV and RCKCDCIRKKA, C.C.[ILMV] is a maximal pattern, which satisfies the density constraint $k_0 = 2, l_0 = 4$.

2.2 Pattern Statistics

Given a kl -pattern π , the probability that it occurs at least once in a sequence of length L is $p = 1 - (1 - \rho_\pi)^{L-l}$, where

$$\rho_\pi = \prod_{i=1 \dots k} \wp(v_i) \quad (2.1)$$

is the probability of pattern π with the k tokens $v_1 \dots v_k$ to occur in a sequence and

$$\wp(v) = \sum_{a \in \Psi} f(a) \quad (2.2)$$

is the probability of a token $v \in \Psi$ to occur in sequence at a given position. In the previous Equation, $f(a)$ is the relative frequency with which amino acid a occurs in the matching sequence.

As shown previously [4], the average number of maximal jkl -patterns that satisfies the $\langle l_0, k_0 \rangle$ density constraint, appearing in a random database composed of n sequences of length L , is given by

$$\langle n_{jkl} \rangle = N_0(k, l) \binom{n}{j} \left\langle \frac{p^j (1-p)^{n-j}}{\rho_\pi} \langle p_{in} \rangle \langle p_{out} \rangle^2 \right\rangle \quad (2.3)$$

The outer angular brackets refer to an average with respect to the matching probability of a generic pattern π , and $N_0(k, l)$ is the number of kl -patterns that satisfy the density constraint. Also, p_{in} and p_{out} are respectively the probability that a given jkl -pattern is maximal in composition and length for pattern π . From this analysis it is possible to estimate the probability that any discovered pattern would have occurred in a random database of similar size and composition. This probability is close to a binomial distribution and as such its mean and variance are related. Therefore, it is possible to compute a z-score using only the above result as:

$$z = \frac{n_{jkl} - \langle n_{jkl} \rangle}{\sigma_{n_{jkl}}} \quad (2.4)$$

where n_{jkl} is the number of discovered jkl -patterns. Full details of this analysis, which is in excellent agreement with experimental data, are available in [16].

2.3 Motif Discovery

We describe an algorithm to identify patterns in a set of protein sequences which have high support and which are statistically significant, that is, they occur with unexpectedly high frequency across the set. Specific values of the parameters used by the algorithm will be discussed in the Results Section.

Let us define as input an arbitrary set of protein sequences, called the database. Pattern discovery is performed using the Splash algorithm by progressively decreasing both the density constraint and the minimum support j_0 , in that order, until at least n_0 patterns are reported. Patterns are reported only if their z-score is greater or equal to a predefined threshold z_0 and the expected number of random matches in SWISS-PROT Rel. 36 is less than a predefined fraction r_0 of the number of sequences in the database. The latter is estimated by the formula $N_{SP} \prod p_i$, where p_i is the probability of the i -th token in the pattern, extrapolated from its frequency in SWISS-PROT Rel. 36, and N_{SP} is the number of residues in that database.

An initial density constraint (k_0, l_{min}) and minimum support j_0 (equal to 100% of the sequences in the database) are chosen. If fewer than n_0 of patterns are reported, the density constraint is reduced by progressively increasing the value of l_0 . If the value l_{max} is exceeded without discovering at least n_0 patterns, the minimum support j_0 is decreased and the procedure is repeated. If a predefined support threshold j_{min} is reached, without any pattern being discovered, the procedure is halted and no pattern is reported.

Due to the deterministic nature of Splash, if a pattern is discovered for a given support and density, it would also be reported for any minimum density and minimum support that are equal or lower. Therefore, one may decide to discover patterns using directly a very low density and support. This, however, would be very computationally inefficient. Due to the combinatorial nature of the pattern formation process, anytime a pattern has a high support, it generates an exponential number of slight variation with smaller support. For instance, if the pattern C.C occurs in 100 sequences, any of the patterns C.CA, CAC, AC.C, etc. will also likely occur by random chance with a ratio equal to the probability of alanine. Given the number of potential positions compatible with the density constraint and the number of potential tokens Ψ , it is easy to see how this could generate a tremendous number of patterns that are all slight random variations of a non-random pattern C.C. The number of patterns explodes even further if the region where the pattern occurs is highly correlated.

As a result, once the support drops below that of a statistically relevant pattern, the number of patterns can start to grow dramatically. For instance, 140 patterns are found in about 20 seconds with a minimum support equal to 95% and a density ($k_0 = 4, l_0 = 16$) in the 123 protein sequences defined by the PROSITE pattern PS00010 (including one identical to PS00010). If the minimum support is dropped to 85%, 342 patterns are found in 75 seconds. At 70%, 5132 patterns are found in about 3 minutes. This iterative procedure finds the pattern with the highest support and statistical significance above a prefixed threshold, if it exists, in a computationally efficient manner.

2.4 Pattern Score

All discovered patterns are sorted by their support and secondarily by their z-score. Then, for each one of the top n_1 scoring patterns, any other pattern with an equal or greater z-score is also selected. These patterns form the *discovery set*. For each one of the patterns in the discovery set, we calculate the number of false positives n_{fp} and false negatives n_{fn} , using criteria identical to that of PROSITE Rel 15. The value n_{fn} is defined as the number of sequences in the input database that do not contain the target pattern. The value n_{fp} is defined as the number of sequences in SWISS-PROT Rel. 36 that contain the target pattern but which are not in the input database. We define a penalty score S_p (lower scores are better) that is used to compare the performance of different patterns:

$$S_p = \alpha_{fn}n_{fn} + \alpha_{fp}n_{fp}; \alpha_{fn} + \alpha_{fp} = 1. \quad (2.5)$$

Here, α_{fn} and α_{fp} are two positive constants used to weight the desired relative importance of false negatives and false positives. For instance, $\alpha_{fp} = 0, \alpha_{fn} = 1$ yields the most sensitive pattern, regardless of its specificity.

Another relative measure of performance is the overlap between a target pattern and one that is known to be biologically significant. Let us assume that π is a biologically significant pattern and that $L(\pi) = \{l_1(\pi), l_2(\pi), \dots\}$, which we call the locus, is the list of positions where it occurs in the database. Given a target pattern π' , with locus $L(\pi') = \{l_1(\pi'), l_2(\pi'), \dots\}$, we compute the overlap with π as follows. First the most likely relative offset of π' and π is computed by histogramming the signed relative offset $\delta_{i,j} = l_i(\pi') - l_j(\pi)$. If a relative offset is supported by at least 50% of the positions where the pattern with the lower support occurs, then the *ratio of overlap* is defined as:

$$o_{ps} = \frac{|[0, l_\pi] \cap [\delta_{i,j}, \delta_{i,j} + l_{\pi'}]|}{l_\pi} \quad (2.6)$$

where l_π and $l_{\pi'}$ are, the total number of characters, including tokens and wildcards in π and π' respectively. If $o_{ps} = 1$, then π occurs within the boundaries of π' . If $o_{ps} = 0$, then the two patterns are incident on different regions of the sequences.

2.5 Discovery of Multiple Motifs

More than one conserved pattern can be discovered as follows. The previous procedure is applied until at least n_0 patterns are discovered. The pattern with the lowest penalty score S_p is selected and reported. All occurrences of that pattern in the database are masked so its tokens can no longer form other patterns. The procedure is resumed with density and support equal to those of the last run until another set of at least n_0 patterns is discovered. This is further repeated until the minimum support drops below a predefined threshold j_{min} . See Section 3.6 for an example of this approach. If the sequence set contains multiple disjoint motifs that are statistically significant and have support greater than j_{min} , this procedure is guaranteed to find all of them.

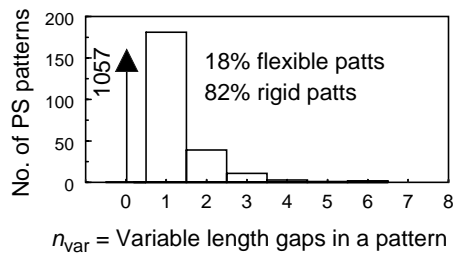


Fig. 1: Number of PROSITE patterns with n_{var} variable length gaps.

3. Results

In this section we first report on the statistical analysis of some interesting properties of the PROSITE patterns, such as their density and flexibility. Then, we study the sensitivity and specificity of the automatically generated patterns compared to those in PROSITE. We then use the overlap ratio, defined in Section 2.4, to measure the success of the identification of protein regions deemed biologically relevant, using only statistical criteria. Finally we exhaustively discover and study all statistically significant patterns for the trypsin family.

3.1 PROSITE Statistics

In Fig. 1, we report on the overall flexibility of PROSITE patterns. As shown, the vast majority of patterns (1057 in bin 0) is rigid. Only 18% of PROSITE patterns contain one or more variable length gap.

Many PROSITE patterns contain tokens that are supersets of those in Table 1. We can however *project* a PROSITE onto a similarity pattern by replacing any such element by an “x”. The idea is to assess how many patterns in PROSITE would still be dense enough that they could be discovered by Splash, even after having been projected onto a similarity pattern. In that case, the deterministic nature of the algorithm guarantees their discovery. To do that, we study projected PROSITE patterns with respect to the density constraint. This will be used to validate our choice of density constraints for the motif discovery procedure. In Figs. 2(a-d) we histogram the number of projected PROSITE patterns that satisfy the (k_0, l_0) density constraint, as a function of l_0 , for $k_0 = 2, 3, 4, 5$. The cumulative (curve) is also plotted. The first bin, not included in the cumulative,

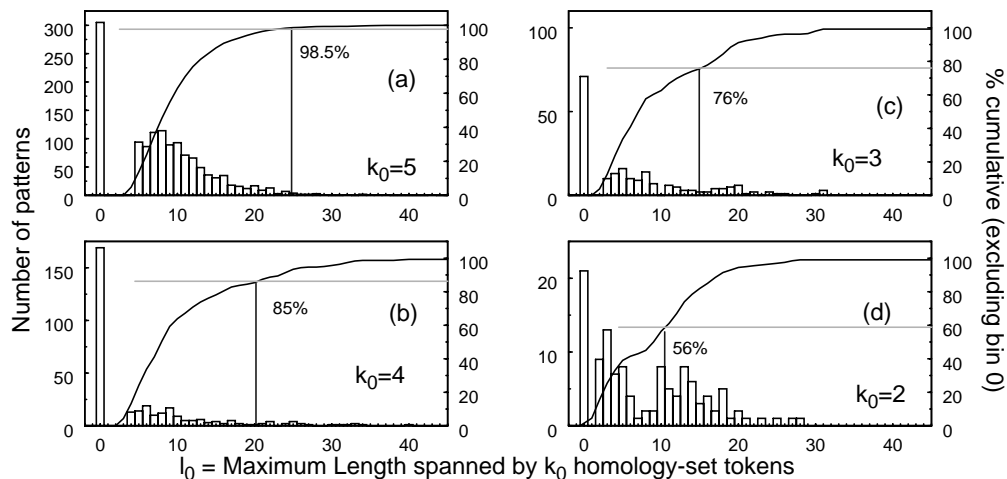


Fig. 2: Number (bins) and cumulative (curves) of patterns that satisfy a (k_0, l_0) density constraint plotted as a function of l_0 . Lines show the cumulative at reported threshold.

(a) $k_0 = 5$
 (b) $k_0 = 4$
 (c) $k_0 = 3$
 (d) $k_0 = 2$.

includes projected patterns that contain fewer than k_0 tokens. If a pattern is flexible, only the most dense rigid subcomponent is analyzed.

In Fig. 2(a), we start by analyzing the 989 projected patterns that have at least $k_0 = 5$ tokens. 974 of these (98.5%) satisfy the density constraint ($l_0 = 25, k_0 = 5$). 320 patterns remain, which either do not contain enough tokens or are too sparse. In Fig. 2(b), we analyze 151 of these that have at least 4 tokens. 128 of them (85%) satisfy the density constraint ($l_0 = 20, k_0 = 4$). 192 patterns are still left. In Fig. 2(c) we analyze 121 of these that have at least 3 tokens. 92 of them (76%) satisfy the density constraint ($l_0 = 15, k_0 = 3$). There are only 100 projected PROSITE patterns that would still elude us. In Fig. 2(d), we analyze this last set. 44 of these (56%) satisfy the density constraint ($l_0 = 10, k_0 = 2$). Of the remaining 64 patterns, the 21 in the first bin are virtually undetectable by any automatic pattern discovery algorithms because they contain either one or no token in each one of their rigid component. In all cases, we have used a window size $l_0 = 5k_0$ as the cutoff.

This suggests that by performing pattern discovery with an appropriate choice of density constraint, one could find a large number (95%) of projected PROSITE patterns. We do not know how well these projected patterns will perform in terms of false positives and false negatives. This will be discussed in the following sections.

3.2 Sensitivity and Specificity

We report the results of the analysis of all 974 families in PROSITE 15.0, which are associated with one or more PROSITE patterns that have a 'Data bank Reference' (DR) field. There are 9 families defined by patterns which do not have DR information. These cannot be used to assess sensitivity and specificity. Of these 974 families, 18.4% are associated with a variable length pattern.

For each family, the procedure described in Section 2.3 has been performed with the following choice of parameters: $k_0 = 4, l_0 = 8, 16, 32$, j_0 is decreased by 5% of the total number of sequences in the set until $j_{min} = 2$ is reached. The minimum z-score is $z_0 = 10^3$, The ratio of the maximum expected number of false positives is $r_0 = 0.1$ (10% of the number of sequences in the database). The minimum number of reported patterns at which the procedure is halted is $n_0 = 100$. The size of the initial report set is $n_1 = 20$. This choice of density is consistent with the results of Section 3.1 and could in principle allow the discovery of all but 169 patterns that have fewer than $k_0 = 4$ tokens.

For each pattern in PROSITE we select all reported true positives and false negatives as the training set. Partials are not considered because they are not reported as true positives in PROSITE when matched. For PS00334, for instance, this results in a group of 31 sequences which includes 30 true positives and 1 false negative. If multiple PROSITE patterns are reported for the same set of proteins, i.e., PS00639 and PS00640, only the patterns with the lowest number of false negatives is considered for this comparison. In this case, PS00639. As seen in Section 2.5. the pattern discovery procedure can be used to extract more than one conserved pattern. In this first part of the analysis, for sake of clarity, we will limit ourselves only to the comparison of the single, most conserved pattern across the entire set, both for PROSITE and for Splash.

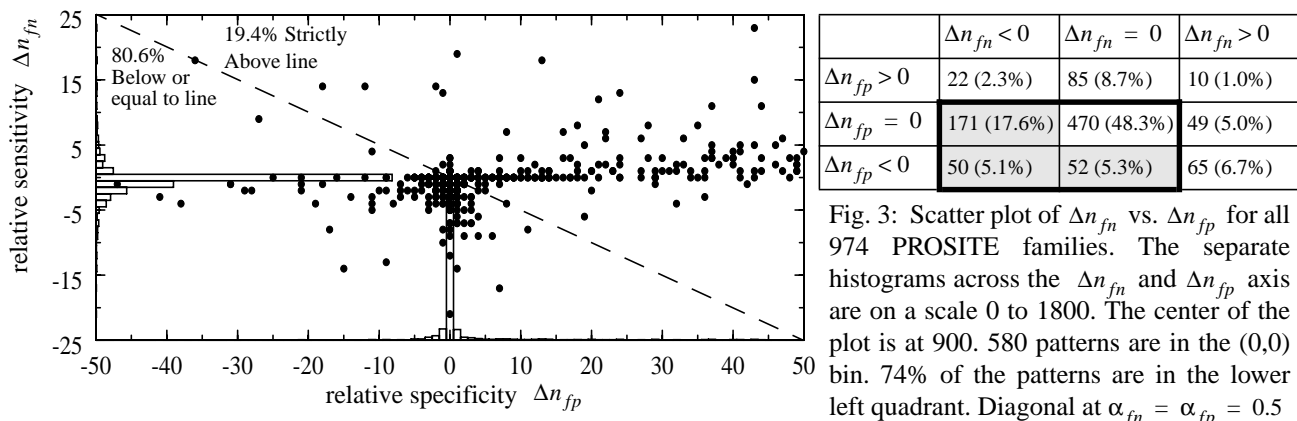


Fig. 3: Scatter plot of Δn_{fn} vs. Δn_{fp} for all 974 PROSITE families. The separate histograms across the Δn_{fn} and Δn_{fp} axis are on a scale 0 to 1800. The center of the plot is at 900. 580 patterns are in the (0,0) bin. 74% of the patterns are in the lower left quadrant. Diagonal at $\alpha_{fn} = \alpha_{fp} = 0.5$

The first set of results compares the specificity and sensitivity of Splash generated patterns compared to PROSITE pattern. These will be called P-patterns and S-patterns, respectively. For each S-pattern in the *discovery set*, we compute the difference Δn_{fp} (false positives) and Δn_{fn} (false negatives) with respect to the corresponding P-pattern. Negative values mean that the S-pattern outperforms the P-patterns (lower false positives/negatives) and vice versa. Then, we determine if there is any S-pattern in the *discovery set* for which, simultaneously, $\Delta n_{fn} \leq 0$ and $\Delta n_{fp} \leq 0$. If more than one is found, the one with the lowest Δn_{fn} is selected. If none is found, the best one according to formula Equation (2.5), with $\alpha_{fn} = \alpha_{fp} = 0.5$ is used. Other values for the two constants can be used as well. This process uniquely selects a single pattern as the top scoring one. Results for all 974 PROSITE families are reported in the table of Fig. 3, where the number and percent of patterns is tabulated based on the respective negative, null, or positive values of Δn_{fn} and Δn_{fp} .

Fig. 3 also plots a scatter graph of the values of Δn_{fn} and Δn_{fp} across all 974 families. Points below the dotted line ($\alpha_{fn} = \alpha_{fp} = 0.5$) correspond to S-patterns that perform at least as well as their PROSITE counterpart according to the P_s penalty score. The histogram of the scatter plot for both Δn_{fn} and Δn_{fp} is also reported. This shows that most patterns are accumulated in a few bins around the center of the plot. The associated table shows that for 76.3% of the families (thick rectangle or lower left quadrant) S-patterns perform at least as well as the corresponding P-patterns. For 28% of the families, the S-patterns strictly outperform the P-patterns. If the ranking is done using Equation (2.5), The number of S-patterns that perform as well or better than P-patterns increases, as some of the 9% mixed cases (top left and lower right) can now be compared in an objective fashion. In particular, for $\alpha_{fn} = \alpha_{fp} = 0.5$, 80.6% of the S-Pattern perform at least as well as their corresponding P-Pattern.

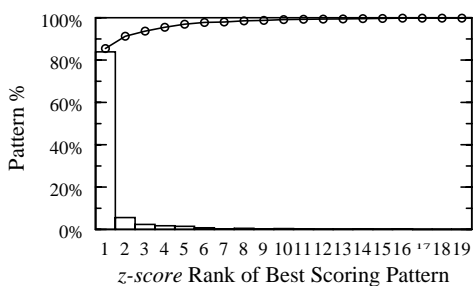


Fig. 4: Patterns% per sorted z-score rank

3.3 Statistical Significance

The plot in Fig. 4 shows the percent of patterns that have the lowest value of S_p , for $\alpha_{fn} = \alpha_{fp} = 0.5$, versus the rank of their corresponding z-score, sorted by the rank. It shows that more than 90% of the top ranking patterns have either the best or the second best z-score. This validates the use of the z-score as a criteria for automatic pattern selection.

3.4 Biological Significance

The following set of results sets a quantitative basis for the likelihood that statistically significant patterns may also be biologically significant. In this case, because of the high quality of the annotation of PROSITE, we will assume that the patterns contained in this databases are associated to sequence regions that are important from a biological perspective. The analysis, then, is aimed at characterizing the degree of overlap between the top ranking S-patterns, which as shown in the previous section have a high statistical significance, and the corresponding P-pattern, which we assume to have high biological significance.

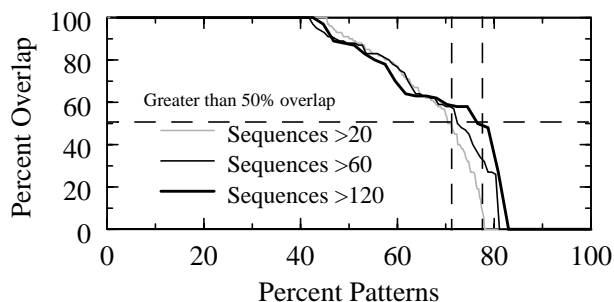


Fig. 5: Cumulative of patterns with overlap better than y

elements. The relatively small improvement hints that about 20 sequences may be sufficient to identify biologically relevant regions from purely statistical criteria. Overall, Figs. 4 and 5 show a remarkable relationship between patterns that are identified based on purely statistical criteria (z-score) and those in PROSITE, which are assumed to have a significant biological role. This suggests that patterns generated by our methodology would be useful as seeds for further refinement with PSSM or profile HMM.

3.5 Analysis of the Results

Of the 273 families where Splash patterns have better or equivalent sensitivity and specificity compared to PROSITE patterns (shaded region in the table of Fig. 3), 178 have an overlap of at least 70% between the two patterns. These are clear instances where our procedure suggests refinement of the biologically relevant PROSITE patterns that improve their sensitivity and/or specificity. Table 2 lists a few examples.

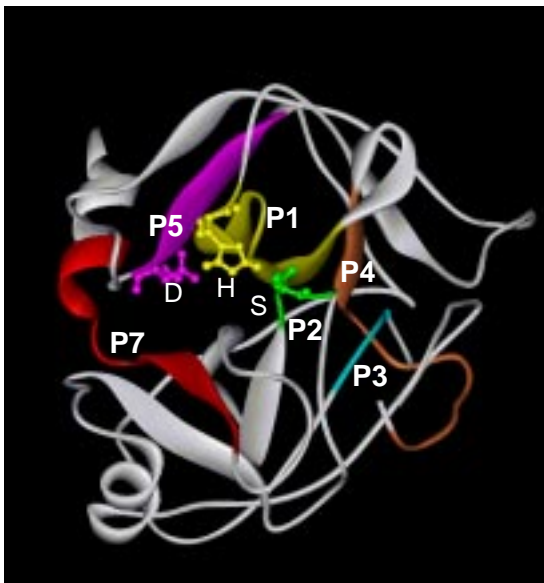
TABLE 2. A few examples of Splash patterns that outperform PROSITE patterns and which have high ratio of overlap

PROSITE	Size	o_{ps}	n_{fn}^P	Δn_{fn}	n_{fp}^P	Δn_{fp}	Splash Pattern	
PS00027	613	0.6	37	-20	6	-1	[ILMV]...[ILMFV]..W[FWY].N.R	Homeobox domain
PS00267	50	0.8	33	-13	9	-9	P.[RQEK]F.G[ILM]M	Tachykinin
PS00080	39	1.0	12	-12	0	0	H.H[ILMV]..[QH]...G[ILMV]	Multicopper oxidases
PS00678	259	0.9	49	-12	92	-62	[ILMFV].[ST]...D...[ILMV][RQK].W	WD-40 repeats
PS00675	64	1.0	13	-11	80	-79	[ILMV].[ILMFV].G..G[RQK]...[AS]...H	Sigma-54 interaction
PS00153	44	0.8	8	-8	1	-1	[ILMV][LFY]...[DQE].....[AS]...[AS]M.[ANST][AS]..N...[ILMV]...[LY]...N...Q...[IV]T.E[IL].[DE][IV]..G	ATP synthase gamma
PS00490	28	1.0	10	-7	0	0	[ILMFV][ILMV]..[DE]...[ANST].[ANST]...[AS]D..L.....[DE]	Prokaryotic molb-dopterin oxidoreductases
PS00432	205	0.8	12	-6	0	0	S..[ANST].L.[ST][LF]..[MV].[IV].[RK].[DE][FY]	Actins
PS00306	28	1.0	5	-5	3	-3	M[RK][LFV][ILFV][IV][LF].C[LF].[AT]..[ILFV]A	Casins alpha/beta

Analysis of protein families where Splash patterns do not match the PROSITE patterns in sensitivity or specificity reveals the following causes. 176 (18%) PROSITE families are described by a flexible motif. In 70 of these, Splash patterns do not score as well as PROSITE patterns. This suggests that the present approach could be further improved by using Splash to discover flexible patterns when the false negative ratio is high. We note that in 78 families, Splash patterns, in spite of being rigid, have fewer or equivalent number of false negatives and false positives, although the overlap of these with the PROSITE patterns is varied.

Another significant cause is the larger grouping of ambiguous and acceptable residues within square brackets PROSITE patterns, compared to our definition of similarity tokens (Table 1). For example, PS00061 is defined by [LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-{PC}-[SAGFR]-[LIVMSTAGD]-x(2)-[LIVMFYW]-x(3)-[LIVMFYWGAPTHQ]-[GSACQRHM] in PROSITE, while the Splash pattern for this family is G.[ILMV]...[AS].....Y..[ANST]K. The similarity tokens used by Splash fail to capture the extent of allowed ambiguity in some residues. Construction of a PSSM or profile HMM seeded by the Splash pattern would provide the necessary sensitivity in such instances.

Fig. 5 plots how many (in percent) of the top scoring patterns (on the x axis) have a ratio of overlap larger than a given percent (on the y axis). The three lines correspond to (a) All families with at least 20 elements, (b) all families with at least 60 elements, and (c) all families with at least 120 elements. About 72% of the S-patterns, for families with 20 or more elements, overlap at least 50% with their corresponding PROSITE patterns. That is, they tend to identify the same region of the protein sequence. This ratio increases to about 77% for families with at least 120



3.6 Multiple Pattern Analysis

In the previous sections we have purposefully limited our analysis to the single best pattern according to the S_p penalty score. However, as described in Section 2.5, the efficiency of the algorithm allows discovery of several patterns that are independently conserved within a family, down to very low values of the support.

We illustrate this by studying the trypsin family of serine proteases. The catalytic activity of these proteases is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine, together known as the catalytic triad. The residues forming the catalytic triad occur in well separated regions of the sequence but are in close spatial proximity in the structure. PROSITE reports only two patterns for this family: PS00134 at the histidine and PS00135 at the serine. Residues of the catalytic triad for the sequence TRYP_PIG (PDB code 1AKS) are his57, asp102, and ser195.

By following the procedure described in Section 2.5, the following 11 patterns are discovered in less than 2 minutes, using a Pentium II 266MHz PC, for the trypsin family, containing 269 proteins. These have a support of at least 40% of the sequences in the family.

Id	Pattern	Support	z-score	BLOCKS
P1	C [ILMV] [ILMV] [ST] A . HC	268	10.0 e+300	BL00134A
P2	G [DE] SGG	274	5.93 e+100	BL00134B
P3	[ANST] G [HFWY] G	261	6.55 e+20	
P4	G P [HFWY] . . . [ILMFV]	253	5.23 e+017	
P5	D [ILMFV] . L [ILMV] [RQEHK] [ILMV]	236	10.0 e+300	
P6	G [ILMFV] . [ANST] . G	253	3.16 e+017	
P7	P . . [FY] . . [ILMV] W [ILMV]	216	8.05 e+057	BL00134C
P8	L [RQEHK] [ILMFV] C	212	3.21 e+006	
P9	[ANST] . . [ILMFV] LP	201	4.63 e+004	
P10	[ILMFV] C [ANST] G	187	3.07 e+003	
P11	[ILMFV] . LG . [NQHY] [NDHS]	139	1.27 e+021	

As expected, the three patterns with the highest z-score (1,2, and 5) correspond to the three catalytic residues. These are shown in bold. BLOCKS currently reports only three of these 11 patterns and misses the other patterns. In particular, pattern 5, which contains the third conserved catalytic residue in the active site and also is one of the 3 highest Z-scoring patterns, is not reported by either BLOCKS or PROSITE. Our method of exhaustive discovery and analysis of conserved regions in protein sequences, efficiently and with a rigorous statistical basis, provides a more comprehensive set of sequence motifs. Discovered patterns, rather than being used directly as regular expressions, could be used to generate position specific scoring matrices, such as BLOCKS. This would allow the number of statistically and biologically significant BLOCKS to be increased without a significant computational load. In particular, the deterministic nature of the pattern discovery component could further improve the sensitivity of BLOCKS. The collection of statistically relevant patterns could be used in conjunction to perform sequence classification.

4. Conclusions

A significant advantage of this approach is its efficiency and scalability. The full set of 974 protein families in PROSITE can be processed in about 12 hours of CPU time on a conventional workstation. Used systematically, therefore, this approach could significantly reduce the labor-intensive component of generating and maintaining PROSITE-like databases. This methodology can be used to significantly and systematically extend the

number of known motifs that uniquely characterize a protein family. Once one or more statistically significant motifs have been identified for a given protein family, these can be used to seed the generation of corresponding PSSM. Alternatively, the collection of motifs incident on a family can be assembled as a multiple sequence alignment using MUSCA [19]. The latter can then be used to produce a profile HMM.

We present a protocol of deterministic pattern discovery and demonstrate its application to automatically and successfully provide patterns that perform as well or better than those in PROSITE, for 72% of protein families in PROSITE. By applying this method to exhaustively discover all statistically relevant patterns in the trypsin family, we obtain a comprehensive set of sequence motifs, including one that contains a functionally critical conserved residue that does not have a corresponding pattern in PROSITE or BLOCKS. This can be a valuable tool to help maintain current databases and significantly increase the number of known, biologically significant motifs, PSSM, or HMM models. The efficiency of the protocol allows it to be incorporated into a daily or weekly update procedure for many existing motif databases.

5. References

- [1] P. Bork, E. V. Koonin. Protein sequence motifs. *Curr. Opin. Struct. Biol.* 6: 366-376 (1996).
- [2] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of protein and nucleic acids.* Cambridge University Press, 1998.
- [3] K. Hoffman, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Research* 27: 215-219 (1999).
- [4] T. K. Attwood, D. R. Flower, A. P. Lewis, J. E. Mabey, S. R. Morgan, P. Scordis, J. N. Selley, and W. Wright. PRINTS prepares for the new millenium. *Nucleic Acids Research* 27: 220-225 (1999).
- [5] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, R. D. Finn, and E. L. L. Sonnhammer. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Research* 27: 260-262 (1999).
- [6] S. Henikoff, J. G. Henikoff, and S. Pietrokovski. Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15: 471-479 (1999).
- [7] T.L.Bailey and M.Gribskov, "Methods and statistics for combining motif match scores," *Journal of Computational Biology*, Vol. 5, pp. 211-221, 1998.
- [8] A.Brazma et al.: "Approaches to the Automatic Discovery of Patterns in Biosequences," *J.Comp.Biol.* 5(2):279-305, 1998
- [9] Bailey, T.L. and Elkan, C., "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." in *Proc. of 2nd ISMB Conf.*, 28-36, AAAI Press, Menlo Park (1994)
- [10] Neuwald, Liu, & Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats," *Protein Science* 4, 1618-1632. 1995.
- [11] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* 13:509-522 (1997).
- [12] Nevill-Manning, C.G. Wu, T.D. & Brutlag, D.L. "Highly Specific Protein Sequence Motifs for Genome Analysis," *Proc. Natl. Acad. Sci. USA*, 95(11):5865-5871, (1998)
- [13] A. Califano, "SPLASH: Structural Pattern Localization Algorithm by Sequential Histograms," communicated to *Bioinformatics*, also communicated to RECOMB 2000.
- [14] I.Rigoutsos and A.Floratos: "Motif Discovery without Alignment or Enumeration," in S.Istrail, P.Pevzner, and M.S.Waterman editors, *Proc. 2nd annual ACM Intl. Conf. on Comp. Mol. Biol, RECOMB98*, 221-227, 1998
- [15] S. Henikoff, and J.G. Henikoff. "Automated assembly of protein blocks for database searching." *Nucleic Acids Research*, 19: 6565-6572 (1991).
- [16] G.Stolovitzky and A.Califano: "Pattern Statistics in Biological Datasets," IBM RC , also subm. to "Discrete Applied Mathematics Series," ed. P.Pevzner
- [17] R.M.Schwartz and M.O.Dayhoff: "Matrices for Detecting Distant Relationships," *Atlas of Protein Sequence and Structure*, 1978, ed. Dayhoff, M.O., pp. 353-358.
- [18] S.Henikoff and J.G.Henikoff: "Amino acid substitution matrices from protein blocks." *Proc. Natl. Acad. Sci. USA*. 89: 10915-10919, Nov., 1992.
- [19] L.Parida, A.Floratos, and I.Rigoutsos, "An Approximate Algorithm for Alignment of Multiple Sequences using Motif Discovery," *J. Comb. Opt.* 3:(2/3):247-275, 1999.