

Scan2S: Increasing the precision of PROSITE pattern motifs using secondary structure constraints

Lucy Skrabanek^{1,2*} and Masha Y. Niv^{3*}

¹Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York 10021

²HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, New York 10021

³Faculty of Agricultural, Food, and Environmental Quality Sciences, Institute of Biochemistry, Food Science, and Nutrition, The Hebrew University of Jerusalem, Rehovot 76100, Israel

ABSTRACT

Sequence signature databases such as PROSITE, which include protein pattern motifs indicative of a protein's function, are widely used for function prediction studies, cellular localization annotation, and sequence classification. Correct annotation relies on high precision of the motifs. We present a new and general approach for increasing the precision of established protein pattern motifs by including secondary structure constraints (SSCs). We use Scan2S, the first sequence motif-scanning program to optionally include SSCs, to augment PROSITE pattern motifs. The constraints were derived from either the DSSP secondary structure assignment or the PSIPRED predictions for PROSITE-documented true positive hits. The secondary structure-augmented motifs were scanned against all SwissProt sequences, for which secondary structure predictions were precalculated. Against this dataset, motifs with PSIPRED-derived SSCs exhibited improved performance over motifs with DSSP-derived constraints. The precision of 763 of the 782 PSIPRED-augmented motifs remained unchanged or increased compared to the original motifs; 26 motifs showed an absolute precision increase of 10–30%. We provide the complete set of augmented motifs and the Scan2S program at <http://physiology.med.cornell.edu/go/scan2s>. Our results suggest a general protocol for increasing the precision of protein pattern detection via the inclusion of SSCs.

Proteins 2008; 72:1138–1147.
© 2008 Wiley-Liss, Inc.

Key words: protein motif; regular expression; pattern; secondary structure constraint.

INTRODUCTION

The functional annotation of proteins is an important challenge in molecular biology. The classical way of accomplishing this involves expensive and time-consuming mutagenesis studies, in order to determine the residues comprising the functional site(s). In the last decade, following the explosion of genomic and structural information, automated computational methods have become indispensable for the prediction of protein function. Computational tools based on sequence, structure, gene-neighbor analysis, or functional links have been developed and used separately or in combination.¹

Among the widely applied sequence-based methods are those that rely on detecting a signature pattern associated with a particular family or a specific function (such as cell-compartment targeting, protein–protein interaction, or regulation by post-translational modifications) in the queried sequences. Patterns are typically described as profiles (weighted matrices) or motifs (regular expressions specifying allowed residues at particular positions).

The use of profiles is becoming increasingly sophisticated, with the introduction of profile–profile matching,² evolutionary information,³ and inclusion of secondary structure information.⁴ Advantages of profiles over patterns include greater sensitivity and quantitative scoring. At the same time, the straightforward simplicity of protein motifs, as well as their ability to describe confined patterns pertinent to enzymatic activities, explain their abundance^{5,6} and sustained usefulness, documented in multiple citations of motif databases and other works, for example, Loose *et al.*⁷ Several approaches have been recently introduced to detect and refine pattern motifs using, for example, evolutionary information⁸ or 3D structural data.^{9,10} We used secondary structure information to refine pattern motifs. The approach was inspired by works

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Lucy Skrabanek, Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, NY 10021. E-mail: las2017@med.cornell.edu or Masha Niv, Faculty of Agricultural, Food, and Environmental Quality Sciences, Institute of Biochemistry, Food Science, and Nutrition, The Hebrew University of Jerusalem, Rehovot 76100, Israel. E-mail: niv@agri.huji.ac.il

Received 26 September 2007; Revised 4 January 2008; Accepted 23 January 2008
Published online 4 March 2008 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/prot.22008

illustrating that secondary structure information is useful in threading tasks,^{11,12} domain assignment,^{13,14} protein sequence alignment,^{15–18} and homolog detection^{4,19,20} and by our recent results showing that the inclusion of secondary structure constraints (SSCs) increases the precision of detection of Type II REases.²¹ These enzymes belong to the “midnight zone” (below 15%) of sequence similarity,²² and are variable even at the level of organization of their secondary structure elements (SSEs).²³ Nevertheless, a conserved core with characteristic physical properties and SSEs was identified and enabled motif derivation.²¹

Here, we test the idea that inclusion of SSCs may serve as a general tool for increasing motif precision, by systematically refining PROSITE pattern motifs using secondary structure information. PROSITE is the first pattern database, which enjoys frequent use and continues to evolve. Patterns in PROSITE are linked to documentation, briefly describing the protein family or domain they are designed to detect, a list of true and false positive matches in the SwissProt database, and experimental structures of proteins that match the motif (linked to the Protein Data Bank²⁴). We use the predominant secondary structures of the original motif-matching positions in the PROSITE-documented true positives to obtain SSCs. Secondary structures can be assigned to a protein in one of two ways: from experimental data (such as X-ray or NMR) or by prediction. We used DSSP²⁵ to generate secondary structure assignments for all true positive hits of PROSITE pattern motifs that have associated experimental structures, and PSIPRED²⁶ to generate secondary structure predictions for the same set of proteins. Using each of these secondary structure assignments, we derived sets of SSCs for each PROSITE motif that has true positives with associated experimental structures (termed DSSP-augmented and PSIPRED-augmented motifs). All the augmented motifs were then scanned against all SwissProt sequences, for which we precalculated secondary structures using PSIPRED. We show that precision for most of the motifs increased or remained unchanged for both methods of SSC derivation; that the improvements are most dramatic for motifs of low original precision, and that PSIPRED-augmented motifs perform better than DSSP-augmented ones. The improved performance of PSIPRED-augmented pattern motifs suggests a general applicability of the protocol for increasing protein motif precision by the inclusion of SSCs, using only sets of known true positives and without the need for experimental structural data.

MATERIALS AND METHODS

Scan2S

We developed Scan2S, a sequence pattern motif-scanning program that optionally includes SSCs. All the con-

ventions recognized by the Java 5.0 regex package can be used in the construction of the pattern. Each position in the pattern is followed by its SSC, for example, [FY]H means that a phenylalanine or tyrosine must be found on a helix. Conversely, one can use the “not” operator to indicate that a residue is not allowed to occur on a certain SSE, for example, P[^H] means that the proline in this motif must not lie on a helix. Where there are no SSCs, this is indicated by a period (the wildcard character), for example, [ILV]. means that the residue at that position can be an isoleucine, leucine, or valine, and that there is no SSC imposed. This last notation is used when representing unannotated PROSITE motifs in Scan2S syntax. There is no restriction on the length of the motif pattern or the number of groups. When components of the pattern can be repeated, the number of times they can appear is listed in brackets, for example, (?.[^H]){2,3} indicates a two- or three-residue set, in which any residue is allowed (indicated by the period) on a stretch of structure that must not be a helix.

Calculation of secondary structure for SwissProt (queried) proteins

We used SwissProt release 50.7 as the protein dataset to be queried. Since both the sequence and the secondary structure information of the protein dataset to be queried are required as input to Scan2S, and since experimental structures are available for only some of the proteins in SwissProt, we generated secondary structure predictions for all 232,345 sequences using PSIPRED version 2.5. PSIPRED²⁶ is among the best protein secondary structure prediction programs available, with a per residue accuracy of 78%.²⁷ We were unable to generate secondary structure predictions for three proteins (TITIN_HUMAN—34,350 aa, TITIN_DROME—18,074 aa, DIG1_CAEEL—13,100 aa) due to their length, and these were removed from further analyses. Scan2S then combines the protein sequence with the secondary structure prediction, producing an output in which the allowed amino acids for each position are immediately followed by its secondary structure assignment. It should be noted that the calculation of the secondary structure for SwissProt is done only once, and is then subsequently used for all the searches. The precalculated set is available at <http://physiology.med.cornell.edu/go/scan2s>.

Selection of PROSITE pattern motifs

Release 19.35 of the PROSITE database holds 1,331 patterns. For this analysis, we used several subsets of the motifs in PROSITE (detailed in Table I). To compare the performance of constraints derived from known structures to constraints derived from secondary structure predictions, we looked at motif patterns that are associated with at least four experimental structures. There are

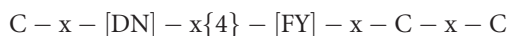
Table I*Description of Datasets Used in Our Analysis*

Dataset name	Dataset description	No. of motifs
prosite_all_pdb	All PROSITE motifs that have at least four associated experimentally determined structures	782
prosite_95_pdb	All PROSITE motifs that have at least four associated experimentally determined structures, and that show a precision of 95% or less (subset of prosite_all_pdb)	152
Prosite_75_all	All PROSITE motifs that show a precision of 75% or less, regardless of whether they have any experimental structures associated with them	47

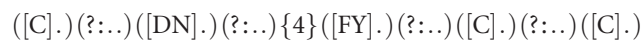
782 such motif patterns in the prosite_all_pdb dataset. We created a subset of 152 motifs that have an original precision of less than or equal to 95%. The prosite_75_all dataset holds all (47) PROSITE motifs with a precision of $\leq 75\%$ (with or without associated PDB files), the subset for which an increase in precision is most needed.

Translation of PROSITE motifs to Scan2S format

In translating PROSITE motifs into Scan2S syntax, every position was supplemented with a nonspecific SSC. Thus, for example, motif PS00010 (the pattern for the aspartic acid and asparagine hydroxylation site) in PROSITE syntax is



and the translated Scan2S syntax is



where the notation $([C.])$ indicates a cysteine match on any SSE, while $\{?\dots\}\{4\}$ indicates a stretch of any four residues on any SSE. No SSCs are included in this motif, as indicated by the period or wild-card character at each SSE position in the motif. Such motifs are hereafter referred to as “unannotated Scan2S motifs.”

One PROSITE motif could not be translated into Scan2S syntax: PS00267, the tachykinin family signature, is defined in PROSITE as being (a) either six residues long, ending with a glycine, or (b) five residues long, but appearing only at the end of the sequence. In this case, we constructed two different motifs, and combined the results.

Annotation of secondary structure for motif augmentation

We calculated the secondary structure in two ways: using DSSP²⁵ and using PSIPRED.²⁶ DSSP: DSSP computes the secondary structure of a protein, given its 3D

coordinates in the PDB file. We ran DSSP on each experimental structure associated with at least one motif in our dataset. We extracted the FASTA sequence from the SEQRES field in the PDB file, and computed the corresponding secondary structure, taking proper account of component chains. Gaps in the sequence were filled with Xs, and corresponding gaps in the structure were filled with dashes (indicating no secondary structure annotation).

PSIPRED: For every sequence having a structure file used for DSSP assignment as described above, its secondary structure was also predicted from the sequence using PSIPRED 2.5, to enable comparison between the DSSP- and PSIPRED-derived motifs. Additionally, for every motif in the prosite_75_all dataset, PSIPRED was used to predict the secondary structure for each PROSITE-documented true positive.

Generation of secondary structure-augmented motifs

We used the predominant secondary structures of original pattern-matching positions in the documented true positives to obtain SSCs. Each motif in our datasets was translated into Scan2S syntax (“unannotated Scan2S motif”), and the list of PDB files associated with each pattern was retrieved. A secondary structure was assigned to each PDB sequence, once using DSSP and once using PSIPRED, as described above. Alignments of all PDB-associated true positive matches for a given unannotated motif were generated from the Scan2S output, by extracting all the sequence matches to the motif, and their corresponding secondary structures.

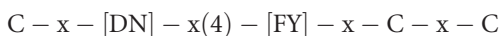
For each position in the alignment, the occurrence of each type of SSE was calculated. The total number of occurrences of each SSE is reported in the Results section and in Supplementary data. SSCs for each matched position in a motif were generated based on several sets of rules shown in Table II. Rule_100 adds as a constraint the SSE that occurs 100% of the time at a given position. In Rule_0, if a SSE never occurs in a position, a negative constraint is added (i.e., if a helix is never seen at a particular position, the SSC will be $\wedge H$). Rule_0_100 combines both Rule_0 and Rule 100. Rule_90, Rule_80, and Rule_80_ext are less stringent, and are based on the idea that if a position typically has a helical nature, it is unlikely to be found on a strand, and vice-versa. Rule_80 (Rule_90) adds an $\wedge H$ constraint if an extended strand (E) is found in at least 80% (90%) of the true matches at a given position, or $\wedge E$ if an H is found in at least 80% (90%) of the true matches at a given position. Rule_80_ext adds further constraints to Rule_80, namely, at positions that are annotated as S (a bend) in at least 80% of the true positive matches, $\wedge H$ is added as a SSC; at positions that have a helical annotation of between 50 and 80%, $\wedge E$ is added as a constraint (Table II). Posi-

Table II
Criteria for Generating Secondary Structure Constraint Annotations

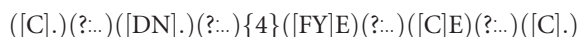
Rule	Criteria used to generate SSCs		
Rule_0	If 0% H then ^H	If 0% E then ^E	If 0% H and 0% E then ^HE
Rule_100	If 100% H then H	If 100% E then E	
Rule_0_100		Combination of 0 rule and 100 rule	
Rule_90	If ≥90% H then ^E	If ≥90% E then ^H	
Rule_80	If ≥80% H then ^E	If ≥80% E then ^H	
Rule_80_ext	The same as 80 rule and if ≥80% S then ^H, if <80% H && >50% H + G + I then ^E		

The secondary structure assignments H, G, and I represent different forms of helices (α -helix, 3/10 helix and π -helix), S indicates a bend, and E represents an extended strand, such as in a β -sheet.

tions that fulfilled none of the criteria in Table II were not considered to lie on conserved SSEs, and no SSC was introduced. As an example, the PROSITE PS00010 motif reads



The PSIPRED-augmented motif for PS00010 using Rule_100 reads



Both the fifth and seventh elements of the motif fall on a strand in 100% of the 52 associated PDB files. The secondary constraints were added accordingly.

For each motif, 13 separate Scan2S motifs were constructed: the translated unannotated Scan2S motif, and the 12 motifs augmented with SSCs according to the above-described rule sets (Table II), generated using either DSSP or PSIPRED secondary structure annotations for the PROSITE motif hits.

We also generated PSIPRED-augmented motifs for dataset_75_all, that is, all PROSITE motifs, with or without associated experimental structures, with a precision of $\leq 75\%$. There are 14 motifs in this dataset that were not present in the original 782 motifs, because they were not associated with a sufficient number of experimental structures. For all motifs in dataset_75_all, the constraint derivation was based on the PSIPRED predictions of all true positive SwissProt matches annotated in PROSITE.

Calculation of precision, recall, and F-measure

We calculated the precision (or specificity, i.e., the likelihood that a match is a true match), recall (or sensitivity, i.e., the percentage of all true matches returned in a dataset), and F-measure (weighted harmonic mean of precision and recall) for each motif using the following equations^{28,29}:

Precision (specificity): $TP/(TP + FP)$

Recall (sensitivity): $TP/(TP + FN)$

F-measure (or F1): $(2 * precision * recall)/(precision + recall)$

where TP is the number of true positives matched in a dataset and FP is the number of false positives or incorrect matches. FN is the number of known false negatives (true matches that were not found). The total number of real occurrences of the motif in the dataset is given by $TP + FN$.

The numbers and identities of the TP, FP, and FN were extracted from each PROSITE entry. Matches found by Scan2S for each unannotated and annotated motif were compared against the lists of TP and FP in the corresponding PROSITE entry. Since the Scan2S motifs used in this analysis add constraints to the PROSITE motifs, no new matches can possibly be found. Therefore, the number of false negatives for Scan2S is calculated as the sum of FN reported for the PROSITE motif, and the difference between the number of TP reported for the PROSITE and the number of TP found by Scan2S. The results from the unannotated motifs were practically identical to those listed in PROSITE.

RESULTS

Secondary structures associated with PROSITE motifs

We started out by checking whether the PROSITE motifs correspond to conserved SSEs. To do this, we analyzed the frequency of the SSE at each position of all PROSITE motifs in the true positive hits, as documented by PROSITE. For motifs for which experimental structures were solved for at least four true positives (prosite_all_pdb set motifs), the frequencies of the SSEs occurring in all motif positions were assigned by both DSSP²⁵ analysis of these structures and PSIPRED²⁶ prediction of the SSEs of the corresponding sequences. We chose PSIPRED as a representative method for secondary structure prediction because, while many new methods

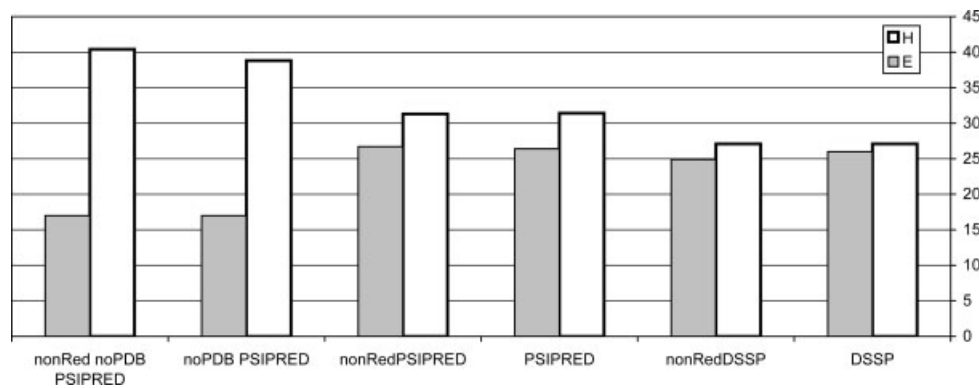


Figure 1

Frequency of occurrence of SSE “H” and “E” in true positive matches of the PROSITE motifs. The percent of “H” (helix, shown in white) and “E” (extended strand, shown in gray) in positions that match the motifs in documented true positives are shown for different datasets: all matches to motifs without associated PDB matches, matched against the redundant and the nonredundant SwissProt 50.7 release (nonRed noPDB PSIPRED and noPDB PSIPRED, respectively); matches to motifs with associated PDB files, first for SSEs obtained by PSIPRED (the nonredundant and the full set of matches) and then for SSEs obtained by DSSP (the nonredundant and the full set of matches).

have been developed to perform this task, in particular, methods that utilize 3D information^{30,31} and consensus servers,³² PSIPRED remains one of the most popular ones with a high Q3 measure (the percentage of correctly classified residues).^{27,30} PSIPRED also has the advantage of being distributed as free stand-alone code which is important for large dataset analysis. The documented true positives of the prosite_all_pdb motifs, corresponding to 39,202 sequences (640,623 individual positions), were analyzed. To account for a possible bias toward particular secondary structures due to over-representation of similar or identical sequences, redundancy was reduced to 90% using the Cd-hit program with default parameters.³¹ A total of 4,725 sequences (81,851 individual positions) in the nonredundant set were also analyzed. In addition, PSIPRED predictions of 27,822 sequences (509,191 individual positions) associated with the 537 remaining motifs (279 of these having no associated PDB files at all), were analyzed.

Figure 1 shows the percentages of the helical and extended strand SSEs occurring in the motif positions in (from left to right): all matches in the SwissProt 50.7 release without associated PDB matches and with redundancy removed (nonRed noPDB PSIPRED); the redundant set (noPDB PSIPRED); the PDB-associated files, first for SSEs obtained by PSIPRED (the nonredundant and the full set of matches) and then for SSEs obtained by DSSP (the nonredundant and the full set of matches). Additional information can be found in the Supplementary data.

Several trends are apparent in Figure 1: (a) A significant percentage of the motif positions occur on helices or on extended strands. (b) There is almost no difference in the relative numbers of SSEs for the full and the non-

redundant results. (c) The ratio of helical to strand positions is larger in the non-PDB set compared to the PDB-associated sequences. This is in agreement with the finding that secondary structure predictions for structurally uncharacterized regions contain, on average, more helical structures than the PDB,³³ and that the average frequency in genomes is 17% for strands and 40% for helices.³⁴ (d) The PSIPRED predictions are in general agreement with DSSP assignment, but the helical content is overestimated by PSIPRED.

Further analysis showed that 89%/80%/80% (respectively) of the H/E/“-” DSSP-assigned positions are correctly predicted to be H/E/C (helix/extended strand/coil) positions by PSIPRED. Therefore, 11% of the DSSP-assigned helices and 20% of the DSSP-assigned strands were incorrectly predicted by PSIPRED. Additional SSEs (T [hydrogen-bonded turn], G [3/10 helix], I [pi helix], B [residue in isolated β -bridge], and S [bend], which are defined in DSSP but not in PSIPRED) were distributed between the H, E, and C definitions of PSIPRED, causing a slight overall overestimation of helix occurrences. A summary of the differences between the PSIPRED predictions and the DSSP assignments for the different SSEs can be found in Figure S2 in the Supplementary data.

Because of the discrepancy between the DSSP and PSIPRED SSE assignments, and because of lack of experimental structures for most genome sequences, we derived rules using both DSSP- and PSIPRED-based SSEs.

SSE-augmented motifs

Positions that matched the motifs in the true positives were analyzed in terms of secondary structure conservation. We found that motif positions in the PDB dataset

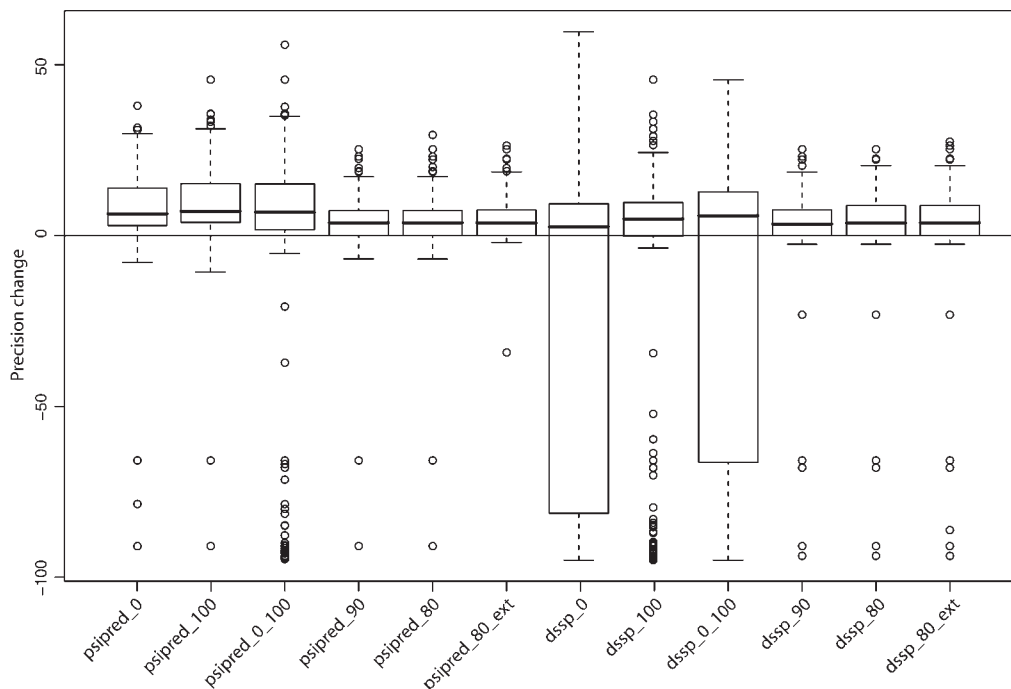
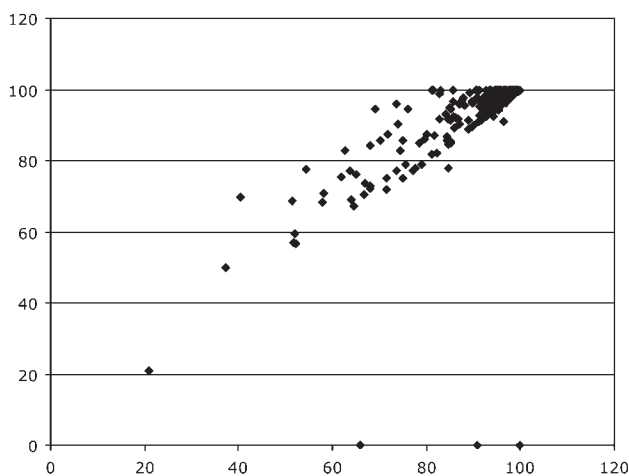


Figure 2

Box plots of the absolute percent change in precision for each method of generating augmented motifs, compared to the unannotated motifs, for the *prosite_95_pdb* dataset. The median is shown by a solid black line. The box boundaries are determined by the first and third quartile values. The whiskers on each plot extend out to 1.5 times the IQR.

Scan2sPrecision
for PSIPRED (Rule_80)



Scan2sPrecision
for DSSP (Rule_80)

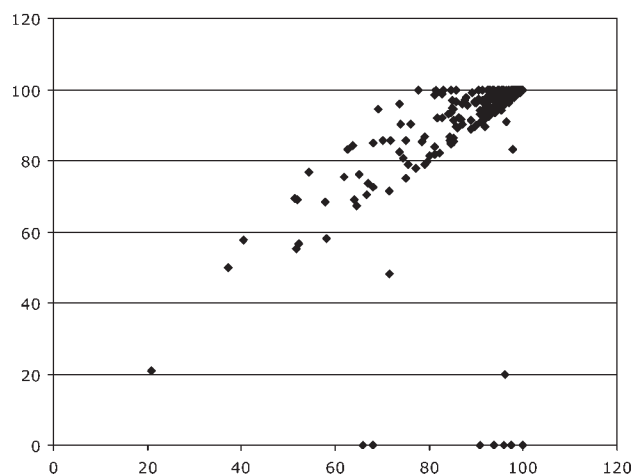


Figure 3

Precision of augmented motifs as a function of precision of the original PROSITE motifs, for *prosite_95_pdb*: Left panel: Rule_80 motifs based on PSIPRED-predicted SSEs; Right panel: Rule_80 motifs based on DSSP-assigned SSEs.

tend to be “all or nothing”: about 90% of all motif positions that match PDB-associated sequences are either always on a helix or never on a helix, and about 90% of all motif positions are either always on a strand or never on a strand. In contrast, only 70–80% of the matches in non-PDB sequences were “all or nothing,” in accordance with the somewhat different nature of PDB and non-PDB sequences.³³ Based on this analysis, we derived several sets of SSCs, as detailed in the Materials and Methods section. PROSITE pattern motifs that had at least four associated experimental structures were augmented with both constraints based on experimental secondary structure (assigned by DSSP), and constraints based on PSIPRED predictions. Ideally, the added constraints should decrease the number of false positives without decreasing the number of true positives. To determine the effect of the addition of SSCs to the motifs, we compared the precision and F-measure of the augmented motifs. The results are presented in Figure 2 and in Supplementary data.

We found that all tested methods of SSC derivation on average increase the precision of the motifs. For example, out of the 782 PSIPRED- (DSSP-) augmented motifs derived with Rule_80_ext, 763 (726) had the same or increased precision compared to the original motifs. All medians of precision showed an increase of between 2.5% and 7.2%, and the first and third quartile values [depicted by box boundaries and describing 50% of the motifs, Fig. 2(A)] were all positive. Remarkably, the DSSP-based motifs showed a much larger variability in increasing precision. The difference between the PSIPRED- and DSSP-based motifs was even more striking when comparing the F-measure (the weighted harmonic mean between precision and recall), in particular for the more stringent sets of rules (Supplementary Figs. S3 and S4). Overall, PSIPRED-based rules performed well for both the full set of prosite_all_pdb and the set of suboptimal precision motifs, prosite_95_pdb, in terms of both precision and F1-measure (see Fig. 3 and Supplementary data). In the following, we detail the results for a representative set of rules, Rule_80. Motifs augmented according to all the tested rule sets are available at <http://physiology.med.cornell.edu/go/scan2s>.

To understand the factors that determine the level of success of the secondary structure annotation in improving precision, we plotted the precision increase against various parameters, such as number of SSCs, number of available experimental structures, number of true positives, motif length, and so forth (not shown). We found that the most important factor in determining the success of secondary structure annotation is the precision of the original motif. Figure 3 plots the precision of the annotated motif versus the precision of the original PROSITE motif for Rule_80 motifs obtained by two methods of generating secondary structure annotations. Several interesting results emerged (see Fig. 3): (1) The precision for most of the motifs with an

original precision of $\leq 95\%$ was increased or remained unchanged for both methods of SSC derivation. (2) The PSIPRED-derived motifs performed better than the DSSP-derived ones. The precision of several DSSP-augmented motifs with originally high precision dropped to zero. In these cases, the DSSP-augmented motifs did not retrieve any true positives, because the predicted secondary structure of the queried sequences differs from the experimental secondary structure used to derive the motifs. There were no such cases for the PSIPRED-augmented motifs, since the SSEs used for constraints' derivation were assigned by the same method used for the queried sequences. PSIPRED-based motifs were very favorable: only three motifs showed significant precision decrease (see Fig. 3). The practical significance of this result is that when documented true positives are available, it is possible to increase the precision of pattern motifs using secondary structure predictions (i.e., no experimental structures are needed). This conclusion is

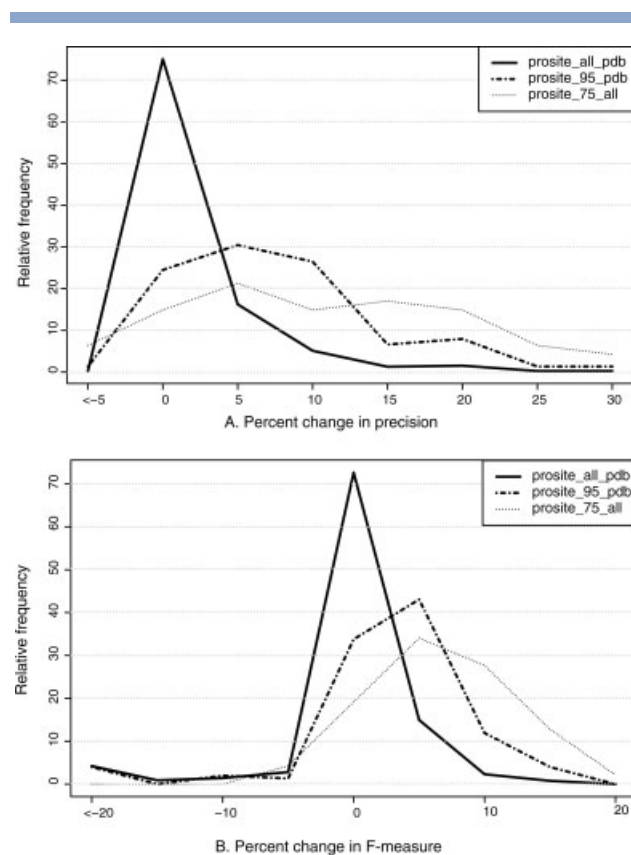
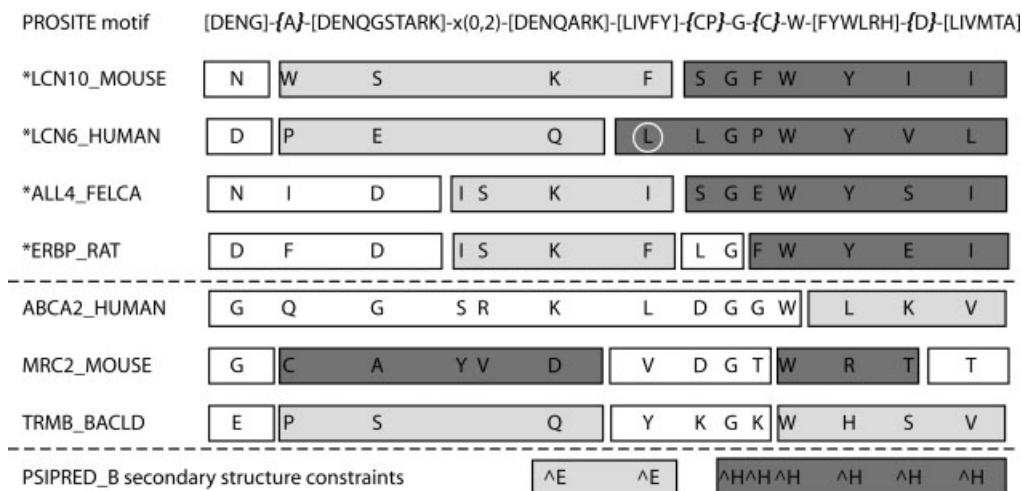


Figure 4

Histograms of changes in absolute precision (panel A) and F-measure (panel B) of PSIPRED-augmented Rule_80 motifs. These histograms are shown in line form instead of bar form for clarity and ease of comparison. In panel 3A, the leftmost bin contains all motifs whose precision decreased by at least 5%; similarly, in panel 3B, the leftmost bin contains all motifs whose F-measure decreased by at least 20%. The precision for about 75% of the PSIPRED-based Rule_80 motifs for the prosite_all_pdb dataset are unchanged. The prosite_75_all dataset shows the highest percentage of motifs whose absolute precision increased by 15% or more.

**Figure 5**

The lipocalin motif. This figure illustrates the secondary structure annotations for a small selection of proteins that match the PROSITE motif for the lipocalin family. The first row is the original motif PS00213 in PROSITE syntax. The secondary structure constraints were derived using Rule_80_ext with PSIPRED predictions. The nonshaded box represents a coil, the light gray box a helix, and the dark gray box a strand. The first four proteins are members of the lipocalin protein family (asterisks). The three nonasterisked proteins contain the PROSITE motif, but obviously do not share the same structure as the lipocalin proteins in the region of the motif. This figure also highlights an example of mistakenly eliminating a true positive: LCN6_HUMAN, which belongs to the lipocalin family, is eliminated since the first leucine (circled) is predicted to fall on a strand, whereas the motif is constrained to nonstrands. All non-lipocalin family members in this figure are correctly eliminated.

further supported by results for motifs derived for all PROSITE patterns of low precision.

Since PSIPRED-augmented motifs performed very well, we could also augment motifs that do not have associated experimental structures. We used Rule_80 to annotate the PROSITE pattern motifs, whose original precision was of $\leq 75\%$ (prosite_75_all). To further highlight the effect of original precision and F-measure on the augmented motifs' performance, we show the performance of the motifs for prosite_95_pdb (the subset of prosite_all_pdb with original precision of less than or equal to 95%) as well as prosite_75_all (the set of motifs with precision of less than or equal to 75%, with or without associated structures):

The histograms in Figure 4 show the distribution of absolute changes in precision and F-measure for the two datasets, binned into ranges of 5%. In the prosite_95_pdb set, 26 of the total 152 motifs showed an increase in precision that was greater than 10%. In the prosite_75_all dataset, 20 of the total 47 motifs showed an increase in precision of between 10 and 30%. Overall, the Rule_80 motifs significantly increased both the precision and the F-measure of a large proportion of the PROSITE motifs, with an original precision of $\leq 75\%$.

CONCLUSIONS

Our results indicated that many pattern motif positions have well-conserved secondary structures. We there-

fore derived motif constraints based on either predicted or observed secondary structures of known true positives. The current number of sequences is still larger than the number of associated structures or even of high-quality homology models.³⁵ Therefore, the SSEs of the queried databases are likely to be based on sequences and obtained by prediction methods (e.g., PSIPRED or others). Because of systematic prediction errors in secondary structure assignment (see Supplementary data), it was not surprising that PSIPRED-based motifs are better suited to querying sequence datasets with PSIPRED-calculated SSEs.

We showed that the secondary structure annotation of PROSITE pattern motifs having an original precision of 95% or less results in increased precision of those motifs. We also showed that the improvement obtained by secondary structure annotation increases for motifs of originally low precision. Thus, secondary structure annotation can be used to improve motifs of suboptimal precision (when documented true positives are available) without the need for experimental structures, as we illustrated for a set of low-precision PROSITE motifs (prosite_75_all). However, for patterns of unknown original precision, no factor could be found that would predict the effect of augmentation with secondary structure. Therefore, further work is needed before secondary structure augmentation of novel motifs can be recommended as a routine approach.

It is worth examining in detail some of the annotated PROSITE patterns that performed exceptionally well. For example, the Rule_80_ext PS00213 motif for the lipocalin

signature showed a 26% precision increase. The PS00213 motif is based on the conserved region SCR1 (strand A and the 3/10-like helix preceding it).³⁶ The Rule_80_ext annotation reflects the structure of SCR1, by introducing ^E constraints at the 310 helix positions, and ^H constraints at the positions that lie on strand A (see Fig. 5). While only 28 of 90 true positives were lost, these annotations were enough to eliminate 102 of the 133 (77%) false positives that come from families as dissimilar as ATP-binding cassettes to mannose receptors to tRNA methyltransferases, none of which share the same secondary structure at the site of the sequence motif detailed by PS00213.

Other noteworthy examples include PS00089 (ribonucleotide reductase large subunit signature), for which Rule_80_ext eliminated all false positives (corresponding to Na(+)-translocating NADH-quinone reductases), while retaining all 56 true positives, resulting in a 19% increase in precision; PS00589 (PTS HPR component serine phosphorylation site signature), which showed a 25% increase in precision by eliminating 26 of 30 false positives, without any loss of true positives; and PS00639 (eukaryotic thiol (cysteine) proteases histidine active site), where constraints increased the precision of the motif by 20%, losing one true positive but eliminating 75 of 117 false positives.

In summary, we introduce a new generation of pattern motifs by augmenting the sequence motif with SSCs. Once the secondary structure predictions for the queried sequences (in this study, this included all SwissProt sequences) have been precomputed and read in, the performance of Scan2S is very fast (about a second per motif) and is therefore useful for genome-wide annotations. The particular rules for constraint derivation can be further refined and readily explored using the automated Scan2S suite of programs. We obtained tens of motifs whose precision was at least 10% higher than that of the original PROSITE motifs. The refined motifs, the precalculated PSIPRED predictions for SwissProt release 50.7 and the program are available at <http://physiology.med.cornell.edu/go/scan2s>, and we hope that these will be useful in function-prediction efforts.

ACKNOWLEDGMENTS

The authors thank Jason R. Banfelder for many helpful discussions and suggestions, and Steve W. Lockless for comments on the manuscript.

REFERENCES

1. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* 2005;15:275–284.
2. Sadreyev RI, Tang M, Kim BH, Grishin NV. COMPASS server for remote homology inference. *Nucleic Acids Res* 2007;35:W653–W658 (Web Server issue).
3. Gutman R, Berezin C, Wollman R, Rosenberg Y, Ben-Tal N. Quasi-MotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res* 2005;33:W255–W261 (Web Server issue).
4. Söding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;21:951–960.
5. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. *Nucleic Acids Res* 2006;34:D227–D230 (Database issue).
6. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingdsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 2003;31:3625–3630.
7. Loose C, Jensen K, Rigoutsos I, Stephanopoulos G. A linguistic model for the rational design of antimicrobial peptides. *Nature* 2006;443:867–869.
8. Davey NE, Edwards RJ, Shields DC. The SLIMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res* 2007;35:W455–W459 (Web Server issue).
9. Via A, Helmer-Citterich M. A structural study for the optimisation of functional motifs encoded in protein sequences. *BMC Bioinformatics* 2004;5:50.
10. Lin KY, Wright J, Lim C. Conformational analysis of long spacers in PROSITE patterns. *J Mol Biol* 2000;299:537–548.
11. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000;299:499–520.
12. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 2003;19:874–881.
13. Gewehr JE, Zimmer R. SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics* 2006;22:181–187.
14. Marsden RL, McGuffin LJ, Jones DT. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* 2002;11:2814–2824.
15. Fontana P, Bindewald E, Toppo S, Velasco R, Valle G, Tosatto SC. The SSEA server for protein secondary structure alignment. *Bioinformatics* 2005;21:393–395.
16. Kim NK, Xie J. Protein multiple alignment incorporating primary and secondary structure information. *J Comp Biol* 2006;13:1615–1629.
17. Qiu J, Elber R. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. *Proteins* 2006;62:881–891.
18. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res* 2005;33:W289–W294 (Web Server issue).
19. Kosinski J, Feder M, Bujnicki JM. The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics* 2005;6:172.
20. Wallqvist A, Fukunishi Y, Murphy LR, Fadel A, Levy RM. Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics* 2000;16:988–1002.
21. Niv MY, Skrabanek L, Roberts RJ, Scheraga HA, Weinstein H. Identification of GATC- and CCGG-recognizing type II REases and their putative specificity-determining positions using Scan2S—a novel motif scan algorithm with optional secondary structure constraints. *Proteins*, in press.
22. Bujnicki JM. Crystallographic and bioinformatic studies on restriction endonucleases: inference of evolutionary relationships in the “midnight zone” of homology. *Curr Protein Pept Sci* 2003;4:327–337.

23. Niv MY, Ripoll D, Vila JA, Liwo A, Vanamee ES, Aggarwal AK, Weinstein H, Scheraga HA. Topology of type II REases revisited; structural classes and the common conserved core. *Nucleic Acids Res* 2007;35:2227–2237.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242.
25. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
26. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
27. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. *Proteins* 2001; (Suppl 5):192–199.
28. Hand DJ, Mannila H, Smyth P. Principles of data mining. Cambridge, MA: MIT Press; 2001.
29. van Rijsbergen CJ. Information retrieval. London: Butterworths; 1979.
30. Lin K, Simossis VA, Taylor WR, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 2005;21:152–159.
31. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22:1658–1659.
32. Cheng HT, Sen TZ, Jernigan RL, Kloczkowski A. Consensus data mining (CDM) protein secondary structure prediction server: combining GOR v and fragment database mining (FDM). *Bioinformatics* 2007;23:2628–2630.
33. Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des* 1998;3:497–512.
34. Gerstein M, Hegyi H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev* 1998; 22:277–304.
35. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 2006;34:D291–D295 (Database issue).
36. Flower DR, North AC, Attwood TK. Structure and sequence relationships in the lipocalins and related proteins. *Protein Sci* 1993;2: 753–761.