ARTICLE
BIOINFORMATICS
DOI: 10.5504/BBEQ.2012.0017

S-MOTIFS AS A NEW APPROACH TO SECONDARY STRUCTURE PREDICTION: COMPARISON WITH STATE OF THE ART METHODS

Ivan Popov
AgroBioInstitute, Sofia, Bulgaria
Correspondence to: Ivan Popov
E-mail: popov.bioinfo@gmail.com

ABSTRACT

The development of protein structure prediction still has not reached the point when there can be only small improvements in the quality of the results. However, the methods for secondary structure prediction are much closer to this leveling point than the much more complex field of tertiary (3D) structure prediction. This paper presents a novel approach to the simpler of the two problems: the assignment of secondary structure elements to a sequence of amino acid residues. The proposed method offers high quality prediction in 70% of the tested cases. This, and the novel S-motif approach it uses, make the method a viable part of any consensus prediction method that may be developed in the future.

Biotechnol. & Biotechnol. Eq. 2012, **26**(3), 3016-3020

Keywords: proteins structure prediction, secondary structure, s-motifs

Introduction

Structure prediction tries to bridge the gap between the huge amount of sequenced protein chains and the relatively small number of resolved protein structures. While the prediction of tertiary structure is still highly restricted (6) and multi-chain protein complexes can be predicted via docking, the prediction of the secondary structure (SS) of amino acid sequences has been in development for several decades. A large number of methods have been proposed, each of them with a different approach to the problem. Recently results have approached 80% accuracy, with the possibility of further improvement by adding long range interactions in the prediction process (14). This means that four out of every five residues have their structure properly assigned from the three possible states – alpha helix, beta strand, or coil. Some state-of-the-art methods use multiple sequence alignment directly or build amino acid profiles of the homologues of the query sequence (9, 12, 15), while others include complex mathematical models like the Artificial Neural Networks that are commonly used (2, 9, 12, 14, 15, 17).

Here we present a method for secondary structure prediction that can also infer additional structural information about the protein sequence. As it is a novel approach that differs from all known methods, we discuss its possible use in consensus predictions and prediction cross validation. The method breaks known protein molecules into pairs of secondary structure elements, called S-motifs, which are used to infer the probable structure of the query sequence. It gives results with quality comparable to the best achieved so far, which shows that it can be used together with other prediction methods without fear of lowering the quality of prediction.

Materials and Methods

Motif definition

S-motifs are a way to represent the secondary structure of a protein. Each motif consists of two secondary structure elements connected by a loop region. They were first used to explore the diversity of the loop regions in available protein structures (5), and later, to characterize the novel protein folds that were added to the databases concerned with the classification of protein structures (SCOP, CATH) (4, 11, 13).

The basic features of an S-motif are shown in **Fig. 1**. These parameters describe the relative position of the two secondary structure elements to each other. There are four main geometrical parameters that define every motif uniquely: 1) **D**, the length of the vector connecting the end of the first SS element and the start of the second one, 2) **θ**, the angle between the main axes of the two elements, 3) **δ**, the angle between the axis of the first element and the vector, and 4) **ρ**, the angle between the axis of the second element and the norm to the plane formed by the axis of the first element and the vector between the two. The parameters are binned as described by Fernandez-Fuentes et al. (5), creating groups of similar S-motifs that share their geometry and the type of the structural elements that they consist of.

An important property of S-motifs is the fact that every consecutive pair of them shares one of the secondary structure elements. That is, every second element of every motif, except for the last in the sequence, is also the first element of the next motif. The classical direction for amino acid chains (N-terminus to C-terminus) is used. Also, a motif is not tied to the particular sequence that one can encounter in a specific protein. There are many occurrences for every S-motif in many different protein

molecules, and most of them differ by their sequence and the length of the two secondary structure elements.

Protein test set

In this work the new method was tested on protein sequences with prevalent beta-strand secondary structure. The dataset of 56 protein chains was selected from the latest release of the Protein Data Bank (www.pdb.org) (1) by the time of writing of this paper (September, 2011) with the following restrictions:

- chain length of 150 to 450 aminoacids;
- at least 90% of secondary structure elements are beta-strands;
- maximum 70% sequence identity in the data set (redundancy threshold).

The length restriction is made in order to lower the chance of possible protein chains with only one secondary structure element, as they do not contain a full S-motif. Seventy percent of the structures in the PDB satisfy this condition. As we wanted to test only beta proteins, we set a minimum amount of beta-strands for the contents of the protein. Finally the redundancy of the set was reduced using the maximum identity feature of PDB for a total of 56 protein structures.

S-motif database

A partly redundant dataset from the PDB database was used to build the starting set of S-motifs used for prediction. PDB sequences were filtered with a 90% identity cut-off.

The S-motif database that was built for the method contains links to the amino acid sequences of the known motifs, and is used to find the secondary structure elements that best match the query sequence. It is divided in four major parts, defined by the type of secondary structure that makes up the motifs in them. These parts are HH, HE, EH, and EE, where H stands for alpha-helix and E stands for beta-strand, according to the general notation of secondary structure.

Motifs are further divided by their geometrical bins, that is, the range of values each of the four parameters take. Each bin is treated as a single type of motif and is used as a reference to all the motif instances with these particular geometrical parameters. The links to the actual amino acid sequences are indexed by length for faster access when searching the database.

Motif matching

Motif sequences are matched to the query by direct comparison of identity. No gaps are allowed. A threshold of 65% was selected as a proper level of identity for the selection of matching motifs. Higher threshold levels detected only the query sequence motifs for some of the test proteins. Lower thresholds increased the number of detected motifs that did not match the type of secondary structure of the query sequence and increased the error in the results (not shown). This effect was observed very strongly when the size of the window (for the initial step, see below) or the length of the element (for every following step) was small.

Search methodology

The prediction is done in steps that overlap known motif sequences with the query sequence. Every search starts with an initiation step that aims to match a starting S-motif to the beginning of the sequence. This step uses a moving window of variable length. As we have no information about the first motif, the first step does an exhaustive search of the motif sequences, retrieving those that match the window in the current iteration with a quality above the threshold. An initial pool of starting motifs is built, and all its members are considered for the next step.

The overlap of S-motifs is utilized to "grow" the chain of motifs to the end of the query sequence. The second element of every matching instance of a motif is taken in turn as a template, in order to determine what part of the query sequence should be used in the second step. Then a search is performed in the respective half of the database that matches the type of the second element; for example if it is a helix, then only the HH and HE parts are searched, as only those contain motifs that start with a helix. Matching motifs that are returned by the second step are pooled and the process is repeated until the end of the query sequence is reached or no matching motifs can be found to continue the chain.

As a final processing step the secondary structure of every position in the query is determined by consensus, using the matching motifs that were detected during the search.

Results and Discussion

The new approach is compared here with two of the best prediction softwares available, representative of the older statistical approaches – GOR IV (8) and the new and widely used Neural Networks (Jnet) (3). The predictions with the new method were done locally with an automated script, while the available web interface was used for the other two methods. Results are summarized in **Table 1**. The software using Neural Networks is much more effective at predicting secondary structure than the direct knowledge based approach. The new S-motif method stands in the middle, having both the low quality results of GOR IV and the very high quality predictions of Jnet, which leads to the higher standard deviation of results. The average prediction quality of the method is comparable to that of Jnet.

TABLE 1

Summary of the results for the prediction of the test protein set by the three methods

Method	Average quality	Quality range	Standard deviation
S-motifs	**76.99**	**26.26 – 100.0**	**19.45**
Jnet	76.95	55.87 – 90.28	6.29
GOR IV	42.31	22.58 – 73.44	10.21

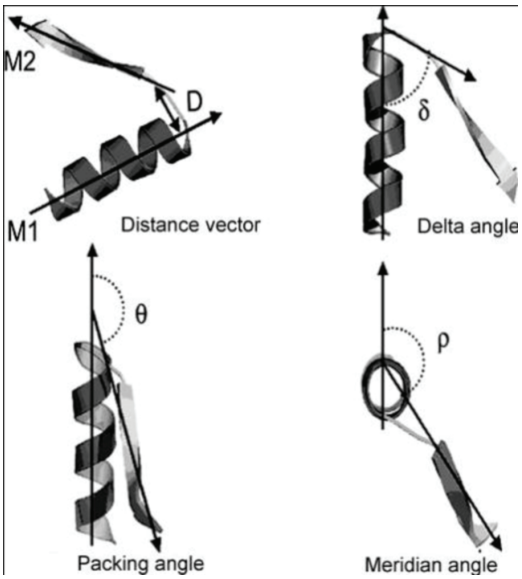

Fig. 1. The geometric parameters of an S-motif. M1 and M2 are the main axes of the two SS elements, and are defined as the vector of the shortest of the principal moments of inertia of the element. Figure adopted from Fernandez-Fuentes et al. (5).

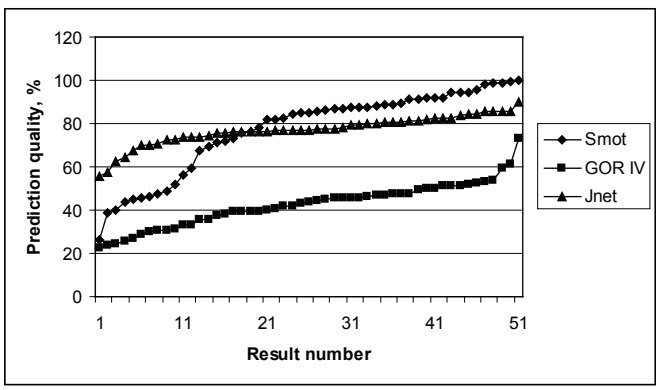

Fig. 2. Plot of the prediction quality of the three methods. Results are ordered by quality for each of the methods, result numbers do **not** correspond to a particular protein sequence.

A more visual plot of the results can be seen in **Fig. 2**, where the predictions of the whole set for each method are ordered by quality. The average prediction quality corresponds to the area under each of the curves. It can be easily seen that the 20 to 25% lower end predictions are the reason why the new method does not outperform Jnet. These predictions are a result of the current restrictions of the method, and the new content included in the PDB database between the building of the S-motif dataset and the current moment when the test was carried out. Currently the length of the starting element of the first S-motif is limited to 20 amino acids to decrease search times in the initial step. This makes the algorithm miss motifs with longer elements, and if no other motif that contains a smaller element is found, the prediction contains no motifs. On the other hand, several thousand new structures were introduced into the PDB, widening its sequence base, and not all proteins in the test set were represented in the S-motif database. This factor has an even higher impact because of the type of test set that was selected for the comparison, as all-beta sequences contain a large number of amino acid residues with specific physical-chemical properties.

Specific characteristics of the algorithm

No gaps can be introduced in the sequence comparison as they cannot be transferred meaningfully into a structural comparison. This sets the approach apart from regular alignment-based methods that simply search for the best homologous sequence on which to model the structure of the query. S-motifs carry information about both the secondary and tertiary structure of the protein sequence. As a result when the method compares the query to a possible S-motif the "alignment" carries structural information as well. The ability to introduce gaps is removed on purpose because there is no way for the method to predict the possible change in overall tertiary structure if one of the elements of the motif has different length, or if a certain amino acid is present at a certain position. Such a change could mean a motif with a different set of geometric parameters, which leads to a different set of motif sequences from the known protein structures.

Apart from the decrease of quality when a lower matching threshold is used, there is also a performance issue connected with this particular setting. When a very low threshold for the matching is set, there are a large number of returned motifs for every initial motif in the pool in the second and every consecutive step. This leads to a geometrical increase in running time when the threshold is lowered. On the other hand, a high threshold may stop the building of the S-motif chain too early, or may even filter any initial matches, so that no prediction is made.

In the current sequence base of the PDB there are around 8000 sequence fragments that constitute both an alpha-helix in some structures, and a beta-strand in others. About 95% of these fragments have a length of 3, 4 or 5 amino acids. At the time these numbers were determined there were 1 260 000 secondary structure elements in all the chains in the PDB, and 204 000 had sequences from that group. This leads to ambiguity when detecting the secondary structure just by the sequence and increases the chance of error in 1 out of every 6 predicted elements. This is of course the upper bound of the error, as most elements will have a predominant secondary structure which by consensus can mask the other, less frequent state of the fragment. The predominant structure will be properly predicted most of the time while errors will arise when the low frequency structure is encountered. The method presented here has an advantage when this type of error is considered. Motifs are found by their first element, which in most cases is already predetermined in its secondary structure by the previous motif in the chain. Apart from the case of starting motifs, this removes

the probability of selecting the wrong structure. Even when a previous motif is not present and hits with the wrong type of SS can appear in the initial pool, the matching is done over the whole motif sequence, which includes the loop region and the second SS element. The frequency at which two consecutive ambiguous fragments are encountered is much lower, and the matching of the loop region lowers the probability of getting a wrong match even further.

Even if a motif with the wrong SS type is selected, there are two factors that prevent it from lowering the quality of the prediction. The first one is the fact that in the end a consensus is still used to select the proper structure type, which will decrease the error rate based on the frequency with which the non-predominant structure type is encountered, as stated above. The second factor is the matching of the next motif in the chain. In the case of an ambiguous fragment in the second element of the last matched motif, we will have both types of structure available for that fragment in the pool of selected S-motifs. All of these motifs will be considered for the next step of elongation of the motif chain, and so both the motifs with the correct and the wrong structure will be matched to the next part of the query sequence. Fragments with the wrong structure type will have a different structural neighborhood in their respective protein folds, which leads to large differences in the loop regions and the next SS element, both in structure and in sequence. Motifs containing those fragments will not be continued in the next step, which limits the error in the consensus sequence of the particular element.

Further development

We are considering replacing the initial step of the algorithm with a BLAST search with the sequence window against a database of all first elements of motifs that start a protein structure. This may include the first elements of all motifs, irrelevant of whether they start a structure, in order to take into account the possibility that an unknown protein may start with a rare S-motif. As BLAST has been widely used in sequence alignment for the last two decades it may be possible that it may perform better in terms of speed than our current exhaustive search method. Again, adjustments should be made to the settings of the algorithm not to allow creating gaps in the alignment, as these are not acceptable when comparing structures.

Currently our method gives a slightly broader spread of prediction qualities compared to Jnet. We are, however, working on improving the lower quality predictions and bringing the quality range closer to that of NN methods, thus giving higher reliability.

Another way of preventing wrong predictions may be the determination of acceptable geometric parameters for every combination of two or more S-motifs. As there is spatial information included in every motif, certain combinations may be structurally impossible even though a sequence match is found. This is a development that is considered in the light of the possible use of the method as a step in tertiary structure prediction.

As the determination of the geometrical parameters requires computational time, the S-motif dataset was built once and was not updated for the last six months, which led to a partly lower quality of prediction. An update of the sequence base of the current S-motifs will increase the overall quality of prediction of the method. Ways to update the database are considered an important feature for the future development of the software.

Conclusions

The average quality of the results from state-of-the-art methods is slowly increasing past 70% correctly predicted residues. In theory the maximal average quality for this particular problem is thought to be around 80%. This is due to the errors in defining the exact starting and ending position of secondary structure elements in the sequence. Even when the ends of the elements are determined from the experimental protein structures, the different algorithms – DSSP (10), DEFINE (16), STRIDE (7), give slightly different results. The final quality of secondary structure prediction will always depend on the quality of the data used for the training of the methods. The new method presented here gives an improvement of quality with an average of correct predictions around 75% and an increase in the upper bound of prediction quality compared to one of the best methods available. The lower quality results that are observed in some cases are partly due to the introduction of novel protein sequences in the PDB, that are not represented by the training set used to build the S-motif database. However, when there are representative sequences in the database, the actual average quality of prediction is closer to 85%. It is the author's belief that using the S-motif model in the prediction of secondary structure can push the upper boundary of prediction quality and introduce a novel approach to use in consensus predictions and the validation of other methods.

Furthermore, the new method also infers additional structural information about the protein in the form of the respective S-motif geometry. This is why it is expected to be suited as a step to be followed up by other prediction methods that aim at the protein's tertiary structure.

REFERENCES

1. **Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.** (2000) Nucleic Acids Res., **28**, 235-242.
2. **Cole C., Barber J.D., Barton G.J.** (2009) Nucleic Acids Res., **36**, W197-W201.
3. **Cuff J.A., Barton G.J.** (2000) PROTEINS: Structure, Function and Genetics, **40**, 502-511.
4. **Fernandez-Fuentes N., Dybas J.M., Fiser A.** (2010) PLoS Comput. Biol., **6**(4), e1000750.
5. **Fernandez-Fuentes N., Oliva B., Fiser A.** (2006) Nucleic Acids Res., **34**(7), 2085-2097.

6. **Floudas C.A.** (2007) Biotechnol. Bioeng., **97**(2), 207-213.
7. **Frishman D., Argos P.** (1995) Proteins, **23**(4), 566-579.
8. **Garnier J., Gibrat J.-F., Robson B.** (1996) In: Methods in Enzymology (R.F. Doolittle, Ed.), **266**, 540-553.
9. **Jones D.T.** (1999) J. Mol. Biol., **292**, 195-202.
10. **Kabsch W., Sander C.** (1983) Biopolymers, **22**, 2577-2637.
11. **Murzin A.G., Brenner S.E., Hubbard T., Chothia C.** (1995) J. Mol. Biol., **247**, 536-540.
12. **Ouali M., King R.D.** (2000) Protein Sci., **9**, 1162-1176.
13. **Pearl F., Bennett C., Orengo C.A.** (2004) In: Dictionary of Bioinformatics and Computational Biology (J.M. Hancock, M.J. Zvelebil, Eds.), Wiley.
14. **Pirovano W., Heringa J.** (2010) Methods Mol. Biol., **609**, 327-348.
15. **Przybylski D., Rost B.** (2002) Proteins, **46**, 197-205.
16. **Richards F.M., Kundrot C.E.** (1988) Proteins, **3**(2), 71-84.
17. **Rost B., Sander C.** (1993) J. Mol. Biol., **232**, 584-599.