VOL5 NO.1March 2016 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

A Promising Similarity Based Secondary Structure Prediction Method

Faruk B.Akcesme, Mehmet Can
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
HrasnickaCesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
fakcesme@ius.edu.ba; mcan@ius.edu.ba

Article Info

Article history:
Article received on 1 Jan. 2016
Received in revised form 7 Feb.2016

Keywords:
Protein Secondary Structure Prediction;
PDB, similarity

Abstract

Assigning secondary structure to amino acid sequences is challenging task due to complexity of protein folding. Protein structures differ enormously. Here, we analyze the mapping between amino acid sequence and secondary structure in a set of 80,592 non-redundant protein chains from the PDB (Protein Data Bank). To identify local conserved regions, we restricted our attention only to the components of these structures of window sizes from 7, to 45. In this article, we examined the issue of how far secondary structure of proteins can be predicted based on the similar segments of solved structures currently deposited in PDB. It is seen that for almost all proteins, secondary structures can be predicted with a mean accuracy of 92%. This accuracy is the highest in the literature.

1. INTRODUCTION

Protein structure prediction at several levels is the main problem of today's molecular biology. The protein secondary structure prediction is the most attractive subset of this problem. It does not attempt to predict the entire three-dimensional structure, but instead concentrates on the local conformational regularities. Secondary structure prediction problem has a long history. First algorithms are formulated around 45 years ago (Pain and Robson, 1970; Finkelstein and Ptitsyn, 1971; Chou and Fasman, 1974; Lim, 1974; Garnier et al., 1978) and new ones emerging every year till today. The most important achievement in the development of the secondary structure prediction algorithms is to provide some information about the fold of the target protein. There are some attempts to predict the fold directly from the predicted secondary structure patterns (Rost, 1995; Di Francesco et al., 1997). Recently secondary structure prediction became an important tool in augmentation of the fold recognition methods (Bowie et al., 1991; Ouzounis et al., 1993; Wilmanns and Eisenberg, 1993; Chou and Zhang, 1994; Yi and Lander,1994; Alexandrov et al., 1996; Fischer and Eisenberg, 1996; Jaroszewski, et al., 1998).

Although protein folding problem is on a long way to solution, there are protein structure prediction methods, such as comparative modeling and threading that work well, but unfortunately only for some special cases. These methods are based on the paradigm that proteins with similar sequences fold to similar structures and usually have similar functions (Sali et al., 1990). Traditional methods of sequence analysis define the similarity on the level of amino acids treated as letters from an abstract alphabet (Waterman, 1995, Lin et al., 2010). Threading methods extend the notion of protein sequence similarity by including the compatibility between interaction preferences of amino acids in different protein structures expressed in propensity matrices. Underlying rules of protein folding is widely accepted as a secret of the nature, and researchers benefit from the reality that, as long as they have similar sequences, they should fold to similar structures.

The usefulness of these analogy based methods, since in nineties; even the most optimistic estimate was claiming that only 10% of all existing protein topologies are currently known (Chothia, 1992). Today this percentage is not improved significantly. But the low percentage of known structures is not so severe for local structure prediction any more. Almost all regular structures are known and thus there is no limit on the accuracy of analogy based secondary structure prediction algorithms.

Analogy based approaches are referred to as the nearest-neighbor algorithm (Levin et al., 1986; Nishikawa and Ooi, 1986; Salzberg and Cost, 1992; Zhang et al., 1992; Geourjon and Deleage, 1994; Yi and Lander, 1994; Salamov and Solovyev, 1995). The core of all these procedures is the definition of similarity, just as in the fold recognition methods.

In this study, to reduce the search space MAX-MATCH (Akcesme 2016) is used as the similarity measure between two proteins. For segment similarity of two segments of equal lengths, it reduces to the similarity derived from Hamming Distances.

2. METHODS

In a previous article (Akcesme, and Can 2016) it has been shown that around 20% of proteins in a non redundant subset of 80,529 protein chains from PDB dataset have larger protein chains that contain them as subsequences, and using the secondary structure sequences of these larger protein chains, it is possible to predict the structures of these smaller proteins with an accuracy more than 85% in average.

In this study, it is postulated that for the remaining 80% of proteins that do not have larger proteins to include them as subsequences, shorter amino acid segments with high similarity, may have higher similarity in corresponding secondary structure segments.

For several datasets, as segment lengths, window sizes from 7 to 45 are tried. No significant differences in accuracies are seen. For this study, window size 17 is adopted. To a segment of amino acid sequence of a query protein from those datasets, all sequence segments of the proteins in the PDB database are visited. The addresses of the ones which have mismatches from 0 to 10 to the query segment are collected. Then from these addresses, the secondary structure sequence segments of the same size are cut (Figure 2).

A plot research on several thousand proteins revealed the surface in Figure 1. Coordinate axis in the front represents the sequence mismatches between two segments of equal lengths of 17 residues. Coordinate axis on the left represents mismatches of the corresponding to the two secondary sequence segments. The height is the number of segment pairs with mentioned properties.

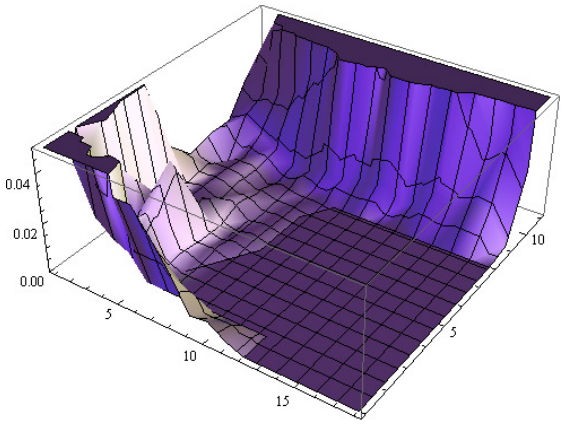

Figure 1. Plateau that shows the relationship between structures and sequences.

Front wall tells that segments that have fully matched amino acid sequences mostly have highly matched secondary structures. Mismatches of secondary structures are not more than 12 for a segment size 17 along this wall. The left wall tells that segments that have amino acid sequences with mismatches from 1 to 10 may still have fully matched secondary structure segments. The importance of the left wall relies on the fact that segments that have amino acid sequences with mismatches from 2 to 8 hardly have matched secondary structures with more than five mismatches. The front corner represents the number of segment pairs whose primary as well as secondary sequences are in full match. The back wall shows that if amino acid sequences of two segment of length 17 have 10 mismatches, then secondary structures of these segments may have from all mismatch levels from 0 to 16.

The above facts about the plateau, suggest that there are a large number of segment pairs whose primary as well as secondary sequences are in high matches. In this study, this idea is used for the secondary structure prediction of the unknown proteins. For each segment of length of a window size of a query protein, segments of the same size whose amino acid chains have mismatches of levels 0-10 with the query segment are searched in the database. The secondary structure sequence of the query segment is assigned as the consensus of those secondary structure sequences of matching segments.

The success of the technique presented here relies on the left wall fact of the plateau. It suggests that there are large numbers of segment pairs whose primary sequences are in several mismatch levels of 1-10 while the corresponding secondary segments are in a higher match position. To benefit from this property, for each segment of length of a window size of a query protein, segments of the same size whose amino acid chains are in mismatch levels of 1-10 with the query segment are searched in the database. The secondary structure sequence of the query segment is assigned as the consensus of those secondary structure

sequences of matching segments at all mismatch level separately, and for each level of mismatch, a prediction for the secondary structure query is established.

The consensus of these secondary structure segments are assigned as a vote for the predicted secondary structure segment corresponding to the query segment. When this is done for all segments of size 17 of the query protein, overlapping votes are aggregated to obtain a prediction for the secondary structure sequence of the whole protein.

Secondary Structure Prediction of the Query Protein

A segment length and a mismatch level are fixed. Then from the amino acid sequence of the query protein a segment is taken of the prescribed length. Amino acid sequences of all proteins in the PDB database except the query protein, and proteins that contain the query protein as a sub chain are visited.

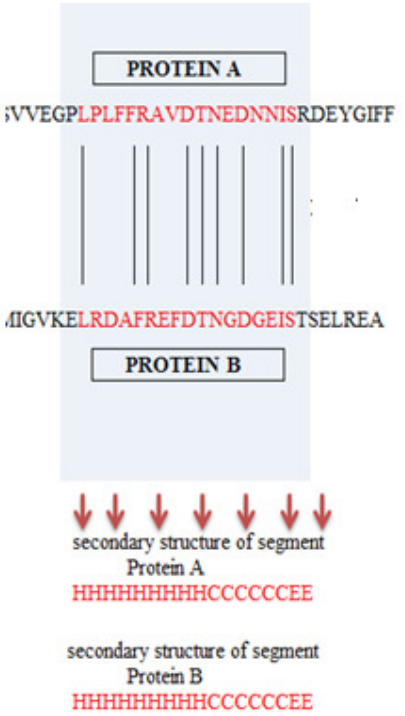

Figure 2. A segment length 17 and a mismatch level 9 is fixed. Then from the amino acid sequence of the query protein (PROTEIN A) a segment is taken of the prescribed length. Amino acid sequences of all proteins in the PDB database except the query protein, and proteins that contain the query as a subchain (PROTEIN B) are visited. The addresses of segments of the same sizes and have the set number of mismatches with the segment from query are noted.

In Figure 2. the fixed length is 17, and mismatch level is 9. Then addresses at the secondary structure sequences of the host proteins (protein B) are visited. At the address, the segment of the same length as the query protein segment is extracted. This segment is taken as the predicted secondary structure of the query segment. The consensus of these secondary structure segments are assigned as a vote for the predicted secondary structure segment corresponding to the query segment. When this is done for all segments of size 17 of the query protein, overlapping votes are aggregated to obtain a prediction for the secondary structure sequence of the whole protein.

Table 1, and Table 2. in the below summarizes this technique for the proteins 1Q2U (Huai et al. 2003), 1BROA (Hecht, H. J. 1994) , and 1 PKYA : 1 – 69 (Mattevi, A. et al. 1995), of the 1189 dataset. First column is the number of mismatches; second column gives the number of the proteins in the PDB dataset that have segments which make mismatches given in the first column. Third column is the total number of segments supplied by the proteins in the second column.

Protein 1Q2U has 287 residues in its amino acid chain. Fourth column gives the number of residues left undecided by the segments with given mismatches in the first column. Fifth column gives the average mismatches of collected secondary structure segments. The last column is the achieved accuracies of secondary structure prediction using the secondary structure sequence segments of target proteins.

Table 1. Summary of the results obtained for protein 645 of the 699 dataset, 1Q2U.

mma	hosts	segments	undec	avmm	accur
0	2	13	194	1.20	77.42
1	2	103	101	0.40	81.18
2	2	93	38	0.81	84.34
3	2	66	63	0.90	80.36
4	2	51	79	0.25	79.81
5	2	70	68	0.08	77.63
6	1	46	111	0.00	75.00
7	2	36	111	1.26	75.00
8	71	227	40	0.80	62.35
9	438	1297	0	0.64	50.17
10	2408	13723	0	0.98	59.58

Comparison of first and last columns gives an idea about the relation between primary and secondary structures of a protein, and sheds some light on the question: how far amino acid sequence determines the secondary structure? As revealed by the plateau in Figure 1, corresponding to the high mismatch levels in the primary sequence of column 1, are not reflected to the fifth column which is the average secondary structure mismatches of collected segments. Because of the high match levels of secondary structure segments, the high accuracies in column six are achieved.

A proper aggregation of the decisions made at every mismatch level, gives an overall accuracy of 81.18% for this protein (Akcesme, F.B., PhD Thesis 2016).

Table 2. Summary of the results obtained for protein 1BROA of the 1189 dataset.

mma	hosts	segments	undec	avmm	acc
0	5	1135	0	2.68	86.28
1	8	187	103	2.00	90.23
2	3	46	140	1.38	88.32
3	4	65	100	1.65	84.18
4	8	114	80	2.00	84.26
5	10	124	40	1.83	84.39
6	11	191	30	1.85	83.40
7	11	203	16	1.59	84.67
8	13	253	26	1.53	88.05
9	39	331	13	2.93	86.36
10	174	1067	2	7.96	68.70

Protein 1BROA has 277 residues in its amino acid chain. As in Table 1., the fourth column gives the number of residues left undecided by the segments with given mismatches in the first column. Fifth column gives the average mismatches of collected secondary structure segments. The last column is the achieved accuracies of secondary structure prediction using the secondary structure sequence segments of target proteins. A proper aggregation of the decisions made at every mismatch level, gives an overall accuracy of 86.28% for this protein.

Table 3. Summary of the results obtained for protein 1PKYA: 1 – 69 of the 1189 dataset.

pmm	hosts	segments	undec	avsmm	acc
0	0	0	□	□	□
1	0	0	□	□	□
2	2	3	34	0.47	85.71
3	3	10	25	0.17	88.64
4	3	16	24	0.47	84.44
5	32	94	24	0.26	88.89
6	42	219	0	0.36	94.20
7	46	210	0	0.31	92.75
8	49	469	1	0.32	94.12
9	100	559	0	0.73	86.96
10	347	1412	0	0.76	63.77

Protein 1PKYA:1– 69 has 69 residues. The fourth column gives the number of residues left undecided by the segments with given mismatches in the first column. Fifth column is the average mismatches of collected secondary structure segments. The last column is the achieved accuracies of secondary structure prediction using the secondary structure sequence segments of target proteins. A proper aggregation of the decisions made at every mismatch level, gives an overall accuracy of 86.28% for this protein.

Table 3. also shows the contribution of the amino acid chain segments in high mismatches with the query segments. In PDB database there are no primary sequence segments that have 0, and 1 mismatch with any 17 residue segment of the query protein. Primary sequence segments with 6, 7, and 8 mismatches bring information which enables secondary structure of the query with accuracies over 90%.

This ability of segments with high mismatches is reflected from the own plateau of this protein as shown in Figure 3.

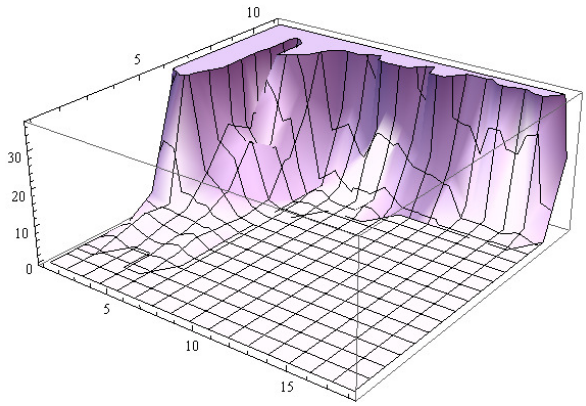

Figure 3. Back corner of the plateau for the protein 1PKYA: 1 – 69 shows the contribution of the amino acid chain segments with high mismatches with the query segments.

3. RESULTS AND DISCUSSION

The method explained in the above applied to dataset which is composed of 500 randomly selected protein chains from PDB dataset. Method is repeated for segments with 1-11 mismatches to the query as well. Predictions made at each mismatch level then aggregated by giving priorities to smaller mismatches.
The technique is repeated for well known datasets such that CB513, FC699, 640, 25PDB, SCOP, and 1189 s well. The number of proteins in databases, and their mean accuracies are given in Table 4.

Table 4. Summary of the accuracies of secondary structure prediction using the secondary structure sequence segments of target proteins.

Data Set	Proteins	% Accuracy
CB513	513	90.20
FC699	858	72.74
640	640	82.63
25PDB	1670	79.19
1189	1092	80.69
Random	500	88.48
Average		82.33

4. CONCLUDING REMARKS

In this article, we examined the issue of how far secondary structure of proteins can be predicted based on the similar sequence segments of solved structures currently deposited in PDB and query chains. It is seen that secondary structures of proteins from the selected as well as randomly chosen datasets can be predicted with a mean accuracy of 82%. This accuracy is the highest in the literature.

REFERENCES

Akcesme, F.B. (2016) A Sequence Segments Similarity based Protein Secondary Structure Prediction Method by the Use of the Relationship between Primary and Secondary Structure of Proteins, International University of Sarajevo (PhD dissertation, in preparation).

Akcesme, F.B., and Can, M. (2016) Is Protein diversity in PDB Library Complete Enough for Similarity Based Secondary Structure Prediction? , SEJSC, Vol. 5, No1, 1-4.

Anfinsen C. (1973) Principles that govern the folding of protein chains. Science;181:223–230.

Matthews B. (1995) Studies on protein stability with T4 lysozyme. Adv Protein Chem;46:249–278.

Alexandrov, N. N., Nussinov, R., & Zimmer, R. M. (1996). Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In Pac. Symp. Biocomput (Vol. 96, pp. 53-72).

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164 –170.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. Nature, 357(6379), 543.

Chou, K. C., & Zhang, C. T. (1994). Predicting protein folding types by distance functions that make allowances for amino acid interactions. Journal of Biological Chemistry, 269(35), 22014-22020.

Di Francesco, V., Garnier, J., & Munson, P. J. (1997). Protein topology recognition from secondary structure sequences: Application of the hidden Markov models to the alpha class proteins. Journal of Molecular Biology,267(2), 446-463.

Finkelstein, A. V., & Ptitsyn, O. B. (1971). Statistical analysis of the correlation among amino acid residues in helical, β-stractural and non-regular regions of globular proteins. Journal of molecular biology, 62(3), 613-624.

Fischer, D., & Eisenberg, D. (1996). Protein fold recognition using sequence derived predictions. Protein Science, 5(5), 947-955.

Garnier, J., Osguthorpe, D.J., and Robson, B. (1978) Analysis and implications of simple methods for predicting the secondary structure of globular proteins, Journal of Molecular Biology, Vol. 120, No. 1, pp. 97-120.

Geourjon, C., & Deleage, G. (1994). SOPM: a self-optimized method for protein secondary structure prediction. Protein Engineering, 7(2), 157-164.

Hecht, H. J., Sobek, H., Haag, T., Pfeifer, O., & Van Pée, K. H. (1994). The metal-ion-free oxidoreductase from Streptomyces aureofaciens has an α/β hydrolase fold. Nature Structural & Molecular Biology, 1(8), 532-537.

Huai, Q., Sun, Y., Wang, H., Chin, L. S., Li, L., Robinson, H., & Ke, H. (2003). Crystal structure of DJ□1/RS and implication on familial Parkinson's disease 1. FEBS letters, 549(1-3), 171-175.

Levin, J. M., Robson, B., & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS letters, 205(2), 303-308.

Lim, V. I. (1974). Algorithms for prediction of α-helical and β-structural regions in globular proteins. Journal of Molecular Biology, 88(4), 873-894.

Lin, H.N., Sung, TY., Ho, SY., Hsu, WL., (2010) Improving protein secondary structure prediction based on short subsequences with local structure similarity, BMC Genomics, 11(Suppl 4):S4

Mattevi, A., Valentini, G., Rizzi, M., Speranza, M. L., Bolognesi, M., & Coda, A. (1995). Crystal structure of Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition. Structure, 3(7), 729-741.

Nishikawa, K., & Ooi, T. (1986). Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 871(1), 45-54.

Ouzounis, C., Sander, C., Scharf, M., & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures.Journal of molecular biology, 232(3), 805-825.

Pain, R. H., & Robson, B. (1970). Analysis of the code relating sequence to secondary structure in proteins. Nature, 227, 62-63.

Rost, B., Sander, C., Casadio, R., & Fariselli, P. (1995). Transmembrane helices predicted at 95% accuracy. Protein Science, 4(3), 521-533.

Salamov, A. A., & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. Journal of molecular biology, 247(1), 11-15.

Šali, A., Overington, J. P., Johnson, M. S., & Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modelling and design. Trends in biochemical sciences, 15(6), 235-240.

Salzberg, S., & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. Journal of molecular biology, 227(2), 371-374.

Waterman,M.S. (1995) Introduction to Computational Biology: Maps, Sequences and Genomes (Interdisciplinary Statistics). Chapman and Hall, London.

Wilmanns, M., & Eisenberg, D. (1993). Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold. Proceedings of the National Academy of Sciences, 90(4), 1379-1383.

Yi, T. M., & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). Protein Science, 3(8), 1315-1328.

Zhang,X., Mesirov,J.P. and Waltz,D.L. (1992) Hybrid system for protein secondary structure prediction., J. Mol. Biol., 225, 1049–1063.