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

Protein Secondary Structure Prediction using Super-chains in PDB

Faruk B. Akcesme, Mehmet Can
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
Hrasnicka Cesta 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, Super chains

Abstract

The completeness of the protein structures in the current Protein Data Bank (PDB) library for use in secondary structure prediction of unknown structure of protein is examined. To deal with this issue, randomly several 1000 protein chains batches are chosen from PDB. For each protein chain in the batch of PDB dataset that who contain the query protein chain as a subsequence are identified and named as a super-chain and prediction of the secondary structure of the query protein is performed by the use of the corresponding sub sequences of the secondary structure sequence of these chains. The technique is repeated for well known datasets such that CB513, FC699, 640, 25PDB, SCOP, and 1189 as well. It is seen that sequences of around 18% of proteins in the batch are present in other chains of PDB dataset. The average prediction accuracy of this method is found to be 80%. Therefore an unknown protein has a chance of 20% to have a super-chain in Protein Data Bank (PDB), and if a protein has a super-chain in the PDB database, there is a possibility that its secondary structure be predicted with around 80% accuracy.

1. INTRODUCTION

For several decades the prediction of protein secondary structure has been studied. At the beginning, the statistical analysis of secondary structure was done for a single amino acid. The most representative technique is the Chou-Fasman method (Chou-Fasman 1978), and the accuracy was hardly 50%. Next, the statistical analysis for amino acid segments came usually with 9~21 amino acid residues. Predicting the structure of central residues based on an amino acid segment improved accuracy. The most representative technique is the GOR method (Garnier, et.al., 1978), and the accuracy increases more than 10%, about 63%. After a decade, the prediction methods on protein secondary structure evolved into using machine learning algorithms (Rost, and Sander 1993, Hua, and Sun 2001, Kim, and Park 2003, Guo et.al. 2004).

In the late 1990`s one of the most famous algorithm PSIPRED was introduced by David Jones. He used the PSI-BLAST which is running for finding similarities to the query and generates intermediate PSI-BLAST profile; position- specific scoring matrices (PSSM). Rather than extracting the sequences, Jones used this intermediate profile as a direct input to two-stage neural network. The accuracy of using PSSM to predict secondary structure has reached between 70~80% accuracy (Jones, 1999).

To the date of December 30, 2003, more than 23,000 solved protein structures have been deposited in the Brookhaven Protein Data Bank (PDB) (Berman, et. al. 2000). This number kept increasing, with 300 new entries added each month at that time. Today there are more than 115.000 (118280) solved protein structures in PDB.
To benefit from the huge size of PDB, methods include comparative modeling (Sali et. al. 1993, Fiser et. al. 2000)

and threading (Bowi et. al. 1991, Jones 1999, Fiser et. al. 2000, Skolnick, et. al. 2004), which are designed to infer an unknown tertiary structure based on solved, similarly folded protein structures in the PDB are developed. Because an accurate theory for the prediction of protein structure on the basis of physical principles does not yet exist, comparative modeling/threading approaches were the only reliable strategy for high-resolution tertiary structure prediction (Moult et. al. 1999, 2001, 2003). On the other hand, the percentage of new folds in these new entries, the topology of which has never been seen in the current PDB library, keeps decreasing. The percentage of new folds was 27% in 1995 but 5% in 2001; number of new unique fold iszero since 2008 (PDB statistics) The apparent saturation of new folds immediately raises an important question: (Zhang, and Skolnick, 2005) , Is the current structure library already complete enough to in principle solve the protein tertiary structure prediction problem at low-to-moderate resolutions?

By means of a variety of structure comparison tools (Taylor et. al. 1994, Holm, and Sander, 1995, Gibrat, et. al. 1996, Shindyalov et. al. 1998), this issue has been partially addressed by many authors (Murzin, et. al. 1995, Orengo, et. al. 1997, Yang, and Honig, 2000, Harrison, , et. al. 2002, Kihara, and Skolnick, 2003).

Although protein secondary structure prediction problem is addressed decades before tertiary structure prediction, it is interesting that, except some pioneering works (Rychlewski, and Godzik, 1997, Lin, et. al., 2010), until recently, no attempt have been done to use the same technique in protein secondary structure prediction as in tertiary structure. To date, there has been no systematic demonstration that this is possible. The exploration of this issue provides the motivation for this work.

In this paper, using a search tool, we first randomly choose samples of 1000 proteins from 80,552 non redundant proteins of PDB. Also most of the famous datasets are visited and their proteins are chosen as queries for the same purpose. For each protein in the sample we find host sequences that contain the amino acid sequence of the query protein as a subsequence. Then secondary structure prediction of a query protein is done using the corresponding secondary structure sequence of the host protein. The technique is repeated for well known datasets.

2. METHODS

The protein secondary structure prediction procedure presented in this work consists of two steps: Identification of a protein chain (super-chain) that contains the query protein as a sub-chain, and prediction of the secondary structure of the query protein is predicted by the use of the corresponding segment of the secondary structure sequence of the super-chain.

Super-Chain Identification.

Super-chain that contains the query protein as a sub-chain are identified from the solved protein structures in the PDB by searching super-chains that contain the query protein as a sub-chain. Because of the huge size of the PDB dataset, each time 1000 proteins are chosen at random. For each protein in the batch, after dropping this protein from the PDB dataset, super-chains are searched.

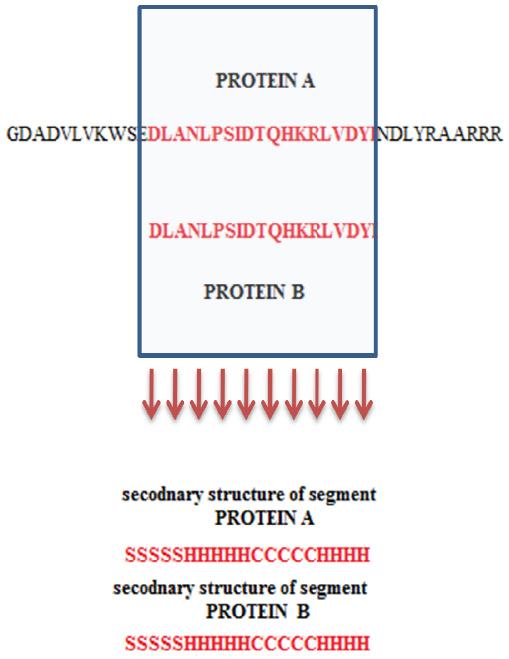

Figure 1. Protein A is a superchain that contains protein B. From the secondary structure sequence of the host protein (protein B), starting from the same address, the subsequence of the same length as the query protein is extracted. This subsequence is taken as the predicted secondary structure of the query.

Secondary Structure Prediction of the Query Protein

Address of the first amino acid where the primary sequence of the query protein starts in the super-chain is noted. From the secondary structure sequence of the host protein, starting from the noted address, the subsequence of the same length as the query is extracted as seen in Figure 1. This secondary structure segment is taken as the predicted secondary structure of the query. If more than two host super-chains do exist for the query, their consensus is the predicted secondary structure of the query.

3. RESULTS AND DISCUSSION

Ten samples of size 1000 proteins from 80.552 non redundant proteins of PDB are randomly chosen. For each

protein in the sample super-chains that include the amino acid sequence of the query protein as a subsequence are found. Then secondary structure prediction of a query protein is done using the corresponding secondary structure sequence of the host protein. The percentage of sample proteins that have a host in database, and mean accuracies of predictions are given in Table 1.

Table 1. The percentage of random sample proteins of batch size 1000, to have a host in database, and their mean accuracies of predicted secondary structure.

Sample	% Hosts	% Accuracy
1	18.7	67.32
2	17.9	89.80
3	15.7	86.89
4	16.3	70.12
5	17.4	90.59
6	23.1	90.91
7	17.9	89.80
8	15.7	86.89
9	19.7	75.46
10	18.2	71.33
Average	18.06	81.91

The technique is also repeated for well known datasets such that CB513, FC699, 640, 25PDB, SCOP, and 1189. The percentage of sample proteins that have a host in database, and their mean accuracies are given in Table 2.

Table 2. Summary of the accuracies of secondary structure prediction using the secondary structure sequences of host proteins

Data Set	Proteins	% Super	% Accuracy
CB513	513	55.17	94.50
FC699	858	71.91	79.19
640	640	58.13	82.72
25PDB	1670	29.76	79.19
SCOP	10294	11.42	90.91
1189	1092	51.10	81.70
Average		23.25	84.67

4. CONCLUDING REMARKS

In this article, we examined the issue of how far secondary structure of proteins can be predicted based on the set of solved structures currently deposited in PDB. It is seen that for around 20% of proteins, secondary structures can be predicted with a mean of 82%. This accuracy around 3% higher for specially designed sets of proteins.

5. FURTHER WORK

80% of proteins in PDB do not have host super-chain in PDB. Although prediction accuracies high enough, for the secondary structure of a query protein there is only 20% chance to be predicted in this way. In another article (Akcesme, 2016), we'll discuss the possibility of secondary structure prediction by the use of smaller conserved segments.

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) A Promising Similarity Based Secondary Structure Prediction Method, SEJSC, Vol. 5, No1, 15-18.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., & Bourne, P. E. (2000). The protein data bank. Nucleic acids research,28(1), 235-242.

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.

Fiser, A., Do, R. K. & Sali, A. (2000) Modeling of loops in protein structures. Protein Sci. 9, 1753–1773.

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.

Guo, J., Chen, H.,Sun, Z., and Lin, Y. (2004) A novel method for protein secondary structure prediction using dual-layer SVM and profiles," Proteins: Structure, Function, and Bioinformatics, Vol. 54, No. 4, pp. 738-743.

Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. (2002) A fast method for reliably recognising the fold of a protein structure, J. Mol. Biol. 323, 909 –926.

Hua, S., and Sun, Z. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," Journal of Molecular Biology, Vol. 308, No. 2, pp. 397-407.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology, 292(2), 195-202.

Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287, 797– 815.

Kihara, D. & Skolnick, J. (2003) The PDB is a covering set of small protein structures. J. Mol. Biol. 334, 793–802.

Kim, H., and Park, H., (2003) Protein secondary structure prediction based on an improved support vector machines approach, Protein Engineering Design and Selection, Vol. 16, No. 8, pp. 553-560.

Lin, HN., 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

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999) Critical assessment of methods of protein structure prediction (CASP) — round x, Proteins 37, Suppl. 3, 2–6.

Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV, Proteins 45, Suppl. 5, 2–7.

Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 53, Suppl. 6, 334 – 339.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536 –540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997) CATH--a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

Peter Y. Chou and Gerald D. Fasman, "Empirical predictions of protein conformation," Annual Review Biochemistry, Vol. 47, pp. 251-276, 1978.

Rost, B., and Sander, C. (1993) Prediction of secondary structure at better than 70% accuracy, Journal of Molecular Biology, Vol. 232, No. 2, pp. 584-599.

Rychlewski, L., and Godzik, A. (1997) Secondary structure prediction using segment similarity, Protein Engineering vol.10 no.10 pp.1143–1153, 1997

Sali, A. & Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779 – 815.

Shindyalov, I. N., and Bourne, P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739 –747.

Skolnick, J., Kihara, D. & Zhang, Y. (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm, Proteins 56, 502–518.

Yang, A. S. & Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance J. Mol. Biol. 301, 665– 678.

Zhang, Y., and Skolnick, J. (2005) The protein structure prediction problem could be solved using the current PDB library, PNAS, vol. 102, no. 4,1029 –1034.