VOL 6 NO.1 March 2017 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
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

Accuracy of Identical Subsequences Based Protein Secondary Structure Prediction

Faruk B. Akcesme, Muhamed Adilovic, 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; madilovic@ius.edu.ba; mcan@ius.edu.ba

Article Info

Article history:
Article received on 11 Jan. 2017
Received in revised form 17 Feb.2017

Keywords:
Protein Secondary Structure Prediction;
PDB, Sequence similarity

Abstract

Chou, and Fasman developed the first empirical prediction method to predict secondary structure of proteins from their amino acid sequences. Subsequently, a more sophisticated GOR method has been developed. Although it became very popular among biologists, their accuracy was only slightly better than random. A significant improvement in prediction accuracy >70% has been achieved by 'second generation' methods such as PHD, SAM-T98, and PSIPRED, which utilized information concerning sequence conservation. Only recently F. B. Akcesme developed a local similarity based method to obtain an accuracy >90%in secondary structure prediction of any new protein. In this article we examined the possibility of sequence similarity based secondary structure prediction of proteins. To deal with this issue, all proteins of PDB dataset are searched for identical subsequences in the other larger proteins of PDB dataset. It is seen that around 17% of proteins in the PDB dataset have identical subsequences in other larger proteins of PDB dataset. When the secondary structures of proteins are assigned as the corresponding secondary structures of identical parts in other larger proteins, the average prediction accuracy is found to be 90.39 %. Therefore, we concluded that an unknown protein has a chance of 17 % to have an identical subsequence in a larger protein in Protein Data Bank (PDB), and there is a possibility that its secondary structure be predicted with around 90% accuracy with this method.

1. INTRODUCTION

For the understanding of both the mechanisms of folding and the biological function of proteins the knowledge of protein structures is essential. To predict the secondary and tertiary structures of proteins, X-ray diffraction has been successfully used for many crystallized proteins. This method is highly accurate, while it is expensive and time-consuming. But many membrane and ribosomal proteins have not yet been crystallized.

Although it is widely believed that the native conformation of a protein is determined by its amino acid sequence(Anfinsen et al., 1961),many unsuccessful efforts have been made to predict the protein secondary and tertiary structures from the protein sequence data.

In (1951),Pauling and Corey suggested that proteins form certain local conformations as helices and strands. Then many workers used different methods to predict protein secondary structure (Szent-Gyorgyi and Cohen,1957; Periti et al., 1967; Ptitsyn, 1969; Pain and Robson, 1970;Robson and Pain, 1971). In most of these researches, the correlation between amino acid sequences and the local secondary structure is used. The effect of neighbors 7-19 amino acids away are taken into account. The average success of these methods could not go much over 50% on three types of secondary structures (alpha-helix, beta sheet, and coil) (Nishikawa, 1983; Kabsch and Sander,1983a,b).

Some attempts were also made to improve the accuracy of secondary structure prediction using the physicochemical properties of the amino acids (Lim, 1974; Ptitsyn and Finkelstein, 1983), statistical analyses of proteins with known structure (Wu and Kabat, 1971, 1973; Chou and Fasman, 1974a,b; Nagano, 1977; Garnier et al., 1978; Maxfield and Scheraga, 1979; Gibrat et al., 1987; Holley L. H., and Karplus M., 1989;Biou et al., 1988; Di Francesco et al., 1997; Fasman, 1989; Garratt et al., 1991; Muggleton et al., 1992), neural networks (Bohr et al., 1988, 1993; Qian and Sejnowski, 1988; Holley and Karplus, 1989; Kneller et al., 1990; Hirst and Sternberg, 1992; Maclin and Shavlik, 1993; Stolorz et al., 1992; Zhang et al., 1992; Rost and Sander, 1993a,b, Chandonia, andKarplus M.,1999, Hua, and Sun2001, Sivanet. al. 2007,Li, and Yu 2016, Rashid et. al. 2016), and pattern matching (Cohen et al., 1983, 1986; Taylor and Thornton, 1983; Rooman et al., 1989; King and Sternberg, 1990; Presnell et al., 1992).

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 118.000solved 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 is zero 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), and 3D prediction tools first try to find identical proteins in PDB, before they start *de novo* predictions.

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), (Levin et al., 1986; Nishikawa and Ooi, 1986; Zvelebil et al., 1986), until recently, no attempts have been made to use the identical chain based prediction technique in protein secondary structure prediction. To date, there has been no systematic analysis about its possibility. The exploration of this issue provides the motivation for this work.

In this paper, using a search tool, we first analyzed pair wise secondary structure similarities of all 80,552 non redundant proteins in PDB. For each protein in PDB we find proteins that contain the query protein as a subsequence. Then secondary structure prediction of a query protein is made adopting the corresponding secondary structure sequence of that subsequence as the secondary structure sequence of the query protein.

2. METHODS

The protein secondary structure prediction procedure presented in this work consists of two steps: Identification of a protein that contains the query protein as a subsequence, and the prediction of the secondary structure of the query protein by the use of the secondary structure of that subsequence.

Identical subsequence Identification.

Proteins that contain the query protein as a subsequence are identified from the solved protein structures in the PDB.

Secondary Structure Prediction of the Query Protein

Address of the first amino acid where the primary sequence of the query protein starts in the bigger protein is noted. From the secondary structure sequence of thebigger 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 protein. If more than two host subsequences do exist for the query, their consensus is then predicted as the secondary structure of the query.

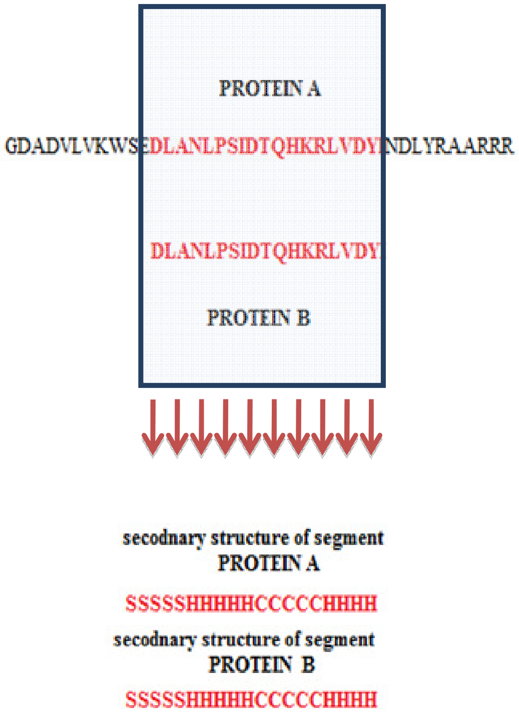
PROTEIN A
GDADVLVKWSEDLANLPSIDTQHKRLVDYINDLYRAARR
DLANLPSIDTQHKRLVDY
PROTEIN B
secodnary structure of segment
PROTEIN A
SSSSSHHHHHCCCCCHHHH
secodnary structure of segment
PROTEIN B
SSSSSHHHHHCCCCCHHHH

Figure 1. Protein A is a protein that contains protein B as a subsequence. 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 (Akcesme, and Can,2016a).

3. RESULTS AND DISCUSSION

Using a search tool, for each protein in PDB we find proteins that contain these proteins as a subsequence. It is found that 13,913 proteins out of 80,552 proteins have at least one identical subsequence in other proteins. The number of identical domains for these 13,913 proteins are distributed as in Figure 2.

Then as secondary structure prediction of query proteins, the corresponding secondary structure sequence of the identical are taken. Since query proteins are all known proteins, their secondary structures are taken from the PDB, and compared by the predicted secondary structures. The distribution of the accuracies of predictions are given in Figure 3.

Table 1. NH; number of large proteins with identical subsequence with the query, NP number of proteins that have this number of identical subsequence.

NH	NP	NH	NP	NH	NP	NH	NP
1	8960	11	41	21	10	31	4
2	2361	12	33	22	7	32	3
3	1011	13	20	23	3	33	2
4	500	14	15	24	5	34	2
5	303	15	17	25	3	35	2
6	172	16	18	26	1	36	3
7	146	17	15	27	3	37	0
8	99	18	3	28	4	38	0
9	65	19	7	29	1	39	2
10	53	20	7	30	3	40	2

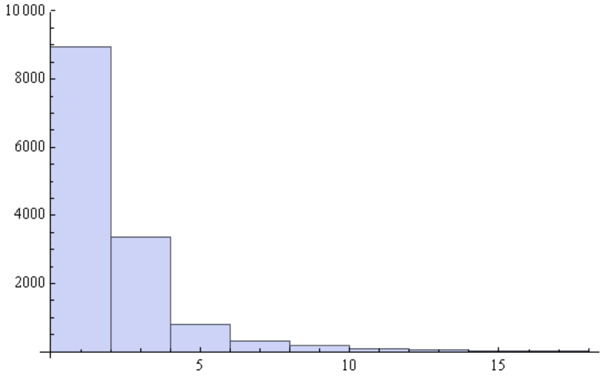

Figure 2. Histogram for the number of the proteins (vertical) with given number of identical subsequences (1-20 horizontal).

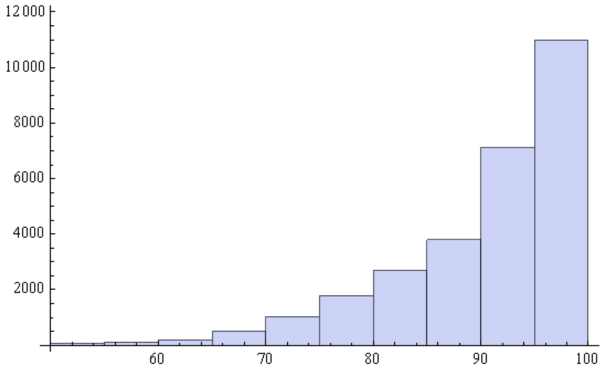

Figure 3. Histogram for the number of the proteins (vertical) with given accuracy of secondary structure predictions (50%-100% horizontal). Around 1% of predictions have less than 50% accuracy.

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 around 17% of proteins in the PDB dataset have identical subsequences in other larger proteins of PDB dataset. When the secondary structures of proteins are assigned as the corresponding secondary structures of identical parts in other larger proteins, the average prediction accuracy is found to be 90.39 %. The percentage of predictions with accuracy less than 50% is only around 1%.

83% of proteins in PDB do not have identical subsequence in larger proteins in PDB itself. Although average prediction accuracy is high enough, for the secondary structure of a query protein, there is at most 17% chance to be predicted in this way. In his PhD thesis F. B. Akcesme (Akcesme, 2016), the possibility of secondary structure prediction with much higher accuracy (mean is more than 80% for all PDB proteins) by the use of smaller conserved segments is discussed.

This work also sheds some light on the accuracy of identical based tertiary structure predictions. Inaccuracy of the identical subsequence based secondary structure predictions undoubtedly set an upper boundary for the identical based tertiary structure predictions.

5. FURTHER WORK

For 1% of proteins that have an identical domain in PDB proteins, secondary structures are predicted with less than 50% accuracy. The reason of this low accuracy is due to the loose relation between sequence and structure for these proteins. This observation must be analyzed in a separate article. On the other hand, the implications of inaccuracy of the sequence similarity based secondary structure predictions onthe sequence similarity based tertiary structure predictions must also be investigated.

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 Thesis at IUS).

Akcesme, F.B, and Can, M. (2016a) Protein Secondary Structure Prediction Using Super-chains in PDB, SEJSC, Vol. 5, No1, 1-4.

Akcesme, F.B, and Can, M. (2016b) 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.

Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J., 1988. Secondary structure prediction: combination of three different methods. Protein Eng. 2, 185–191.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M., Fredholm, H., Lautrup, B., Petersen, S.B., 1993. Protein structures from distance inequalities. J. Mol. Biol. 231, 861–869.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M., Lautrup, B., Norskov, L., Olsen, O.H., Petersen, S.B., 1988. Protein secondary structure and sequence similarity by neural networks. The alpha-helices in rhodopsin. FEBS Lett. 241, 223–228.

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.

Chandonia J.-M., Karplus M., (1999) New Methods for Accurate Prediction of Protein Secondary Structure, PROTEINS: Structure, Function, and Genetics 35:293–306.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J., 1986. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25, 266–275.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J., 1983. Secondary structure assignment for alpha/beta proteins by a combinatorial approach. Biochemistry 22, 4894–4904.

Di Francesco, D., Garnier, J., Munson, P.J., 1997. Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. J. Mol. Biol. 267, 446–463.

Fasman, G.F., 1989. The development of the prediction of protein structure. In: Fasman, G.F. (Ed.), Prediction of Protein Structure and the Principle of Protein Conformation. Plenum Press, New York, pp. 193–613.

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.

Garratt, R.C., Thornton, J.M., Taylor, W.R., 1991. An extension of secondary structure prediction towards the prediction of tertiary structure. FEBS Lett. 280, 141–146.

Gibrat, J.F., Garnier, J., Robson, B., 1987. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol. 198, 425–443.

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.

Hirst, J.D., Sternberg, M.J., 1992. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry 31, 7211–7218.

Holley, L.H., Karplus, M., 1989. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. U.S.A. 86, 152–156.

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.

Holley L. H., and Karplus M., (1989) Protein secondary structure prediction with a neural network, Proc. Nati. Acad. Sci. USA Vol. 86, pp. 152-156.

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.

Kabsch, W., Sander, C., 1983b. How good are predictions of protein secondary structure? FEBS Lett. 155, 179–182.

King, R.D., Sternberg, M.J., 1990. Machine learning approach for the prediction of protein secondary structure. J. Mol. Biol. 216, 441–457.

Kneller, D.G., Cohen, F.E., Langridge, R., 1990. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214, 171–182.

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.

Levin, J.M., Robson, B., Garnier, J., 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett. 205, 303–308.

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

Li, Z., Yu, Y. (2016) Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16) pp. 2560-2567

Maclin, R., Shavlik, J.W., 1993. Using knowledge-based neural networks to improve algorithms. Refining the Chou–Fasman algorithm for protein folding. Machine Learning 11, 195–215.

Maxfield, F., Scheraga, H., 1979. Improvements in the prediction of protein backbone topography by reduction of statistical errors. Biochemistry 18,
697–704.

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.

Muggleton, S., King, R.D., Sternberg, M.J., 1992. Protein secondary structure prediction using logic-based machine learning. Protein Eng. 5, 647–657.

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.

Nagano, K., 1977. Triplet information in helix prediction applied to the analysis of super-secondary structures. J. Mol. Biol. 109, 251–274.

Nishikawa, K., 1983. Assessment of secondary-structure prediction of proteins. Comparison of computerized Chou–Fasman method with others. Biochim. Biophys. Acta 748, 285–299.

Nishikawa, K., Ooi, T., 1986. Amino acid sequence sequence similarity applied to the prediction of protein secondary structures and joint prediction with existing methods. Biochim. Biophys. Acta 871, 45–54.

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.

Pauling, L., Corey, R.B., 1951. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl. Acad. Sci. U.S.A. 37, 729–740.

Periti, P.F., Quagliarotti, G., Liquori, A.M., 1967. Recognition of alpha-helical segments in proteins of known primary structure. J. Mol. Biol. 24, 313–322.

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

Presnell, S.R., Cohen, B.I., Cohen, F.E., 1992. A segment-based approach to protein secondary structure prediction. Biochemistry 31, 983–993.

Ptitsyn, O.B., 1969. Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. J. Mol. Biol. 42, 501–510.

Ptitsyn, O.B., Finkelstein, A.V., 1983. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 22, 15–25.

Qian, N., Sejnowski, J., 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202, 865–884.

Rashid S., Saraswathi S., Kloczkowski A., Sundaram S., and Kolinski A., (2016) Protein secondary structure prediction using a small training set (compact model) combined with a Complex-valued neural network approach, Rashid et al. BMC Bioinformatics 17:362-380.

Robson, B., Pain, R.H., 1971. Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. J. Mol. Biol. 58, 237–259.

Rooman, M.J., Wodak, S.J., Thornton, J.M., 1989. Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. Protein Eng. 3, 23–27.

Rost, B., Sander, C., 1993a. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232, 584–599.

Rost, B., Sander, C., 1993b. Secondary structure prediction of all-helical proteins in two states. Protein Eng. 6, 831–836.

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.

Sivan S., Filo O., Siegelmann H., (2007) Application of expert networks for predicting proteins secondary structure, Biomolecular Engineering 24, 237–243.

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

Stolorz, P., Lapedes, A., Xia, Y., 1992. Predicting protein secondary structure using neural net and statistical methods. J. Mol. Biol. 225, 363–377.

Szent-Gyorgyi, A.G., Cohen, C., 1957. Role of proline in polypeptide chain configuration of proteins. Science 126, 697–698.

Taylor, W.R., Thornton, J.M., 1983. Prediction of super-secondary structure in proteins. Nature 301, 540–542.

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.

Wu, T.T., Kabat, E.A., 1971. An attempt to locate the non-helical and permissively helical sequences of proteins: application to the variable regions of immunoglobulin light and heavy chains. Proc. Natl. Acad. Sci. U.S.A. 68, 1501–1506.

Wu, T.T., Kabat, E.A., 1973. An attempt to evaluate the influence of neighboring amino acids (n " 1) and (n + 1) on the backbone conformation of amino acid (n) in proteins. Use in predicting the three-dimensional structure of the polypeptide backbone of other proteins. J. Mol. Biol. 75,13–31.

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

Zvelebil, M.J., Barton, G.J., Taylor, W.R., Sternberg, M.J., 1986. Prediction of protein secondary structure and active sites using the alignment of identicalous sequences. J. Mol. Biol. 195, 957–961

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.