EXCLI Journal 2009;8:66-73 ISSN 1611-2156
Received: March 4, 2009, accepted: March 21, 2009, published: March 25, 2009

Original article:

ADDITION OF CONTACT NUMBER INFORMATION CAN IMPROVE PROTEIN SECONDARY STRUCTURE PREDICTION BY NEURAL NETWORKS

Amir Lakizadeh[1*], Sayed-Amir Marashi[2]

[1] Department of Computer Science, Faculty of Science, Qom University, Qom, Iran
[2] International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC), Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, Berlin, Germany
[*] Corresponding author (e-mail address: lakizadeh@khayam.ut.ac.ir)

ABSTRACT

Prediction of protein secondary structures is one of the oldest problems in Bioinformatics. Although several different methods have been proposed to tackle this problem, none of these methods are perfect. Recently, it is proposed that addition of other structural information like accessible surface area of residues or prior information about protein structural class can significantly improve the prediction of secondary structures. In this work, we propose that contact number information can be considered as another useful source of information for improvement of secondary structure prediction. Since contact number, i. e. the number of other amino acid residues in the structural neighbourhood of a certain residue, depends on the secondary structure of the residue, we conjectured that contact number data can improve secondary structure prediction. We used two closely related neural networks to predict secondary structures. The only difference in the neural networks was that one of them was also provided with residue contact numbers as an additional input. Results suggested that addition of contact number information can result in a small, but significant improvement in prediction of secondary structures in proteins. Our results suggest that residue contact numbers can be used as a rich source of information for improvement of protein secondary structure prediction.

Keywords: Protein secondary structure; neural networks; contact number

INTRODUCTION

One of the oldest problems in bioinformatics is to accurately predict the structure of proteins from their amino acid sequences. At the end of the 50's and the beginning of the 60's, it was suggested that protein sequence can uniquely determine the three dimensional structure of a protein (Anfinsen and Haber, 1960; Anfinsen et al., 1961; Sela et al., 1957; White, 1961). If this is true, all the information needed for protein structure prediction is already present in protein sequence. At least, protein secondary structure (2S), which is almost stable regardless of protein flexibility and motions, must be predictable. Since the existing protein 3D structures contain some inaccuracies, there is a theoretical limit for accuracy of protein structure prediction methods (Huang and Wang, 2002), which is estimated to be 88 % (Rost, 2003), or maybe up to 90-95 % (Pollastri et al., 2007).

Although the problem of protein 2S prediction is extensively studied (Floudas et al., 2006), the accuracy of state-of-the-art methods is still far below the theoretical limit (Dor and Zhou, 2007; Liu et al., 2008; Montgomerie et al., 2006; Reyaz-Ahmed and Zhang, 2008). This might be due to the fact that sequence alone does not provide (directly) all the necessary information for

EXCLI Journal 2009;8:66-73 ISSN 1611-2156
Received: March 4, 2009, accepted: March 21, 2009, published: March 25, 2009

2S formation. For example, amino acid propensities for different secondary structures depend on different factors, like: the organisms from which the proteins are selected (Marashi et al., 2007), protein structural class (Costantini et al., 2006), and relative solvent accessibility of the residues (Cohen et al., 1993; Han and Baker, 1996; Kabsch and Sander, 1984; Minor and Kim, 1996; Sudarsanam, 1998; Zhong and Johnson, 1992). Structural class (Costantini et al., 2007; Yüksektepea et al., 2008) and relative solvent accessibility (Adamczak et al., 2005; Macdonald and Johnson, 2001; Momen-Roknabadi et al., 2008; Zhu and Blundell, 1996) have already been used for improving 2S prediction in proteins.

In this work, we present the novel idea of incorporating contact number (CN) information for 2S prediction. The idea comes from the fact that residues within regular secondary structures in proteins have distinct contact numbers compared to other residues in proteins. Therefore, it may be possible to exploit this additional information to help secondary structure prediction. Since residue contact number is also a predictable quantity, it might be possible in future to use predicted CN data to empower 2S prediction algorithms.

Table 1: List PDB chains used in this study

1A12A	1G8KA	1J1NA	1N97A	1QOPB	1UG6A	1Y4WA	2B0JA	2DQ6A	2I49A
1B43A	1G9GA	1J4AA	1NC5A	1QW9A	1UH4A	1Y7BA	2B0TA	2DSJA	2I4LA
1BS0A	1GDEA	1JFBA	1NE9A	1QZ9A	1UM0A	1Y7TA	2B3FA	2DVTA	2I5NC
1BUPA	1GK9B	1JIXA	1NOFA	1R17A	1URSA	1YDYA	2B5WA	2E7ZA	2INCA
1C1DA	1GNLA	1JNRA	1NR0A	1R6DA	1UVJA	1YFQA	2BF6A	2EX0A	2IVFA
1C3PA	1GOTB	1JQ5A	1NSZA	1R6XA	1UWKA	1YHLA	2BIBA	2EZ2A	2IW1A
1C96A	1GP6A	1JU3A	1NTHA	1R89A	1V0EA	1YHTA	2BJFA	2EZ9A	2IXSA
1CB8A	1GPUA	1JX6A	1NUYA	1RA0A	1V33A	1YIIA	2BJKA	2F2HA	2J1NA
1CCWB	1GQ8A	1K7WA	1O0SA	1RGZA	1V54A	1YJSA	2BJQA	2FBAA	2J6LA
1CHMA	1GQIA	1K92A	1O7JA	1RI6A	1V5VA	1YKDA	2BMOA	2FE8A	2JDID
1CIPA	1GU7A	1KA1A	1ODMA	1RJDA	1V6SA	1YRCA	2BO4A	2FF4A	2JE8A
1CVRA	1GUQA	1KMJA	1OFLA	1RK6A	1VEFA	1YT3A	2BWRA	2FGQX	2JEPA
1CZAN	1GWEA	1KMOA	1OK7A	1RWHA	1VFLA	1YU0A	2C0HA	2FMPA	2NVOA
1D0CA	1GXMA	1KOLA	1ON3A	1RYIA	1VYRA	1ZAIA	2C1VA	2FNUA	2NX9A
1D5TA	1H16A	1KWFA	1ONRA	1S1DA	1W23A	1ZB1A	2C31A	2FQXA	2O0JA
1DLJA	1H2WA	1KWGA	1OWLA	1S3EA	1W4XA	1ZCJA	2C5AA	2G50A	2O36A
1DPGA	1H6LA	1L7AA	1OX0A	1S95A	1W78A	1ZHXA	2C78A	2G5FA	2O4VA
1DQAA	1HBNB	1LC5A	1OXXK	1SU8A	1W99A	1ZJAA	2C81A	2G8JA	2O5VA
1DS1A	1HDHA	1LFWA	1OYGA	1SVMA	1WDPA	1ZJCA	2C82A	2GDQA	2O9CA
1DUVG	1HM9A	1LWDA	1OZ2A	1SYYA	1WMWA	1ZPDA	2CF5A	2GF3A	2OB3A
1E6UA	1HNJA	1LZLA	1P1JA	1T1UA	1WTJA	1ZSQA	2CN3A	2GJLA	2OITA
1EEXA	1HS6A	1M15A	1P1MA	1T2DA	1WVFA	1ZXXA	2CXNA	2GL5A	2OKTA
1ELUA	1HT6A	1M1NB	1PFVA	1T4BA	1WY2A	1ZZ1A	2CZCA	2GZ1A	2OQYA
1EU8A	1HX6A	1MTPA	1PO5A	1TBFA	1X1NA	1ZZGA	2D0OA	2H6FB	2OSXA
1EZWA	1HYOA	1MTYD	1Q16A	1TKIA	1X54A	2ACVA	2D29A	2H88A	2OX0A
1F20A	1HZ4A	1MUWA	1Q6ZA	1TXGA	1XFKA	2AEUA	2D3NA	2H9AA	2P02A
1FN9A	1I1QA	1MXRA	1Q7ZA	1U09A	1XO0A	2AHFA	2D54A	2HC9A	2P0WA
1FP2A	1I24A	1N40A	1QF5A	1U5UA	1XPMA	2AKZA	2D73A	2HDWA	2P1MB
1FS7A	1IB2A	1N4WA	1QHDA	1U6ZA	1XSZA	2AQJA	2DE3A	2HEKA	2P3ZA
1G5AA	1IO1A	1N62B	1QLMA	1U8VA	1XUUA	2AXQA	2DG1A	2HHVA	2UXYA
1G6SA	1IOMA	1N8KA	1QMGA	1UA4A	1Y1PA	2AZ4A	2DGKA	2HZLA	3THIA

EXCLI Journal 2009;8:66-73 ISSN 1611-2156
Received: March 4, 2009, accepted: March 21, 2009, published: March 25, 2009

MATERIALS AND METHODS

Dataset

A list of protein chains with mutual sequence identity less than 25 % and structural resolution smaller than 3 Å was taken from the Protein Data Bank, followed by a sequence culling procedure by PISCES (Wang and Dunbrack Jr., 2003, 2005). Chains with unknown structure regions were removed. The final dataset contained 310 protein chains with a total of 132 676 residues. The list of these proteins can be found in Table 1.

Secondary structure assignment

For each protein structure, secondary structures of its residues were assigned by DSSP (Kabsch and Sander, 1983). Since eight possible secondary structures are assigned by DSSP, we grouped them into the three states by converting [G, H, I] to H (i. e. helices), [B, E] to E (i. e. extended structures), and other structures (including T and S) to C (i. e. coils).

Calculating contact numbers

If Cα's of two residues are closer than 6 Å in space, we assumed the two residues to be in contact. Contact number of a residue is equal to the number of all other residues in the same chain that are in contact with this residue.

Neural networks, their inputs and their architectures

Following typical machine learning protocols, the classifiers discussed in this work assume that each residue is represented by a vector in a certain feature space defined by a set of attributes. These attributes include 20 values from the Position Specific Scoring Matrix (PSSM). PSSM is obtained from PSI-BLAST (Altschul et al., 1997) with three iterations of searching against non-redundant sequence database (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz). In addition, one parameter is used for describing the nonexistence of amino acids for some positions of the window centred at the edges of the sequence. In this method, we assume that the local structural environment and evolutionary context of each residue is characterized by a sliding window of 13-residue, with the residue of interest at position 7. The window of length 13 proved to be sufficient in our tests to achieve accuracies essentially identical to those with longer windows. Moreover, a longer window would imply a larger number of parameters to be optimized, increasing the risk of overfitting. Thus, we have a vector of 21 parameters for each residue and the total number of attributes for each input pattern in the method is 13*21=273. In other words, length of every input vector for training the neural networks is 273.

In this work, in order to show the impact of addition of CN on 2S prediction accuracy, two separate predictors consisting of two-level neural networks (i. e. a total of four neural networks) are used. In the first predictor, prediction was done merely based on the secondary structure patterns and multiple alignments of sequences, while in the second one, CN information was also used. In both predictors, following Rost and Sander classical protocol (Rost and Sander, 1993; Adamczak et al., 2005; Jones, 1999), our method for secondary structure prediction consists of two levels. The initial "sequence to structure" prediction uses information derived from the amino acid sequence and input vectors in the feature space (described in the previous section), while the final "structure to structure" prediction is based on the outcome of the first level prediction and allows us to correlate better predictions for neighbouring residues.

The first level is the same in both predictors, and consists of 25 small NNs. In the first step, we divided our protein dataset into 25 subsets. Therefore, each subset contained 12 distinct protein chains (except one subset, which contained 22 chains) and more than 5000 input patterns. In the next step, for each subset, a neural network was designed and then trained by the corresponding input patterns, which means that each of these NNs is trained by about 1/25 of the whole training set. All these networks

EXCLI Journal 2009;8:66-73 ISSN 1611-2156
Received: March 4, 2009, accepted: March 21, 2009, published: March 25, 2009

are four-layer feed-forward NNs, with 273 input neurons, 100 neurons in the first hidden layer, 50 neurons in the second hidden layer, 10 neurons in the third hidden layer, and 3 neurons in the output layer. The latter three neurons correspond to the three structural states, i. e. H, E and C. Quick back propagation algorithm was used in training these networks.

In the next step, the training of second level in both models was done by 25-fold cross validation. Briefly, in each iteration the following tasks were performed. First, one subset (and its corresponding NN) was chosen from the dataset. Afterwards, all of the remaining 24 datasets were used as the input for all of the remaining NNs in first level. Thus, each of the NNs resulted in three values for H, E and C. Subsequently, the average values of H, E and C were computed. For training the second level, a feed forward network was used, with Quick back propagation algorithm. In case of not using CN information, the network had three input neurons (i. e. for average H, E and C), 10 neurons in the hidden layer, and 3 neurons in the output layer. In case of using CN data, an additional neuron in the input layer was added, while the hidden and the output layers are similar. The values of the final three outputs of the second level present the criteria for deciding about the secondary structure of the residues. At the end of each iteration, the chosen subset is used as a "test set", to determine the accuracy of the predictor.

Altogether, each iteration results in the prediction of 2S of residues in one test set. Since a 25-fold cross validation is performed, finally we obtain 25 values as the performance values of the simple NN, and 25 corresponding values as the performance of the NN trained by addition of CNs.

Measuring the performance of prediction

The prediction quality was evaluated by three state percent accuracy (*Q3*), which can be computed as:

$$Q3 = \frac{N_H + N_E + N_C}{N_{tot}} \times 100$$

where N_H, N_E and N_C is the number of correctly predicted residues in helices, extended structures and coils, respectively. N_{tot} is the total number of residues in the corresponding subset of the dataset.

Statistical analysis

In our study, we performed 25-fold cross-validation. This means that for each of the 25 subsets, one performance value is obtained when CN information is used for training, and another performance value (for the same dataset) is obtained when CN information is not used. Therefore, for comparing the performance of the methods when CN information is used vs. not used, paired *t*-test can be applied. We used R-package (http://www.r-project.org/) to perform the statistical analysis.

RESULTS AND DISCUSSION

The purpose of this work is to see if addition of contact number information can improve protein secondary structure prediction. Therefore, two closely predictors consisting of two-level neural networks were designed: the first predictor is a typical one, designed for prediction of 2S merely based on the protein sequence, while the second predictor has one additional input node for Contact Number of the residues. During the training procedure, the first predictor is trained with sequence and its corresponding 2S, while the second predictor is trained by sequence, 2S and additionally CN.

In order to compare the predictive power of the two predictors, we performed a 25-fold cross validation test (see Methods). Briefly, the whole protein dataset is split into 25 subsets. In each iteration, the two predictors were independently trained by 24 datasets and then the 2S of the proteins in the remaining subset was predicted by the trained predictors.

Table 2 summarizes the results of this study. Apparently, in all cases a small improvement can be observed. Additionally,

EXCLI Journal 2009;8:66-73 ISSN 1611-2156
Received: March 4, 2009, accepted: March 21, 2009, published: March 25, 2009

this improvement in prediction is statistically significant ($P<10^{-8}$ in paired t-test). This means that CN information can significantly improve the prediction of protein 2S.

Figure 1 illustrates the distribution of Q3 scores when CN information is used or not used for training the predictors. The evident shift in the accuracy of prediction performance proves that addition of CN information can help the neural network based predictor to better learn the 2S patterns of proteins.

Table 2: A comparison between performances of the neural network for prediction of protein secondary structure when contact number (CN) information is used or not used for training the networks.

Subset	With CN	Without CN
1	68.588	68.089
2	70.491	69.259
3	70.905	70.714
4	72.285	71.554
5	72.332	71.692
6	72.374	71.968
7	72.451	72.231
8	73.266	72.491
9	73.586	72.958
10	73.601	73.033
11	73.740	73.151
12	73.876	73.191
13	73.935	73.250
14	74.063	73.484
15	74.191	73.560
16	74.261	73.703
17	74.362	73.768
18	74.558	73.820
19	75.030	73.834
20	75.263	74.537
21	75.453	74.667
22	75.498	75.177
23	75.935	75.273
24	75.966	75.368
25	76.731	76.200

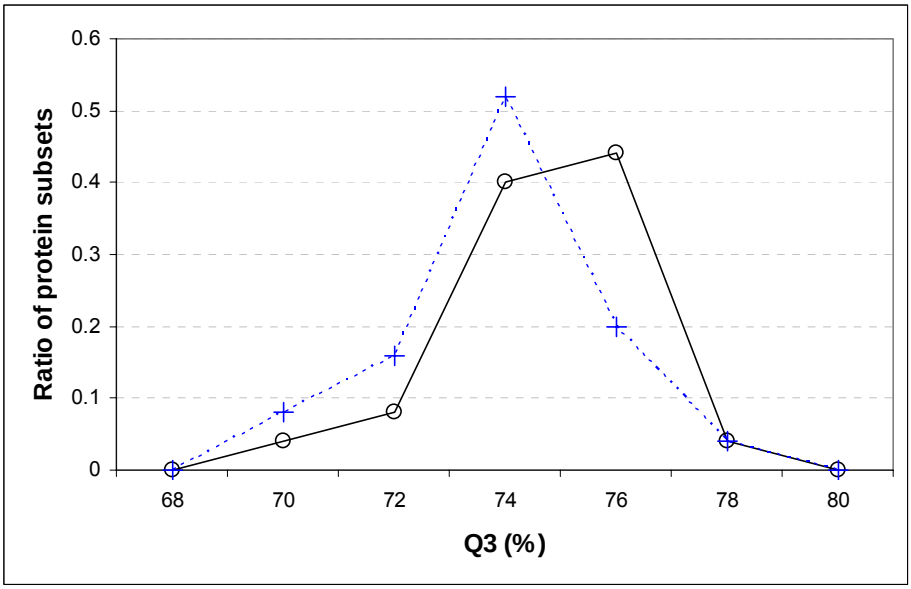

Figure 1: Comparison of the performance of the designed NNs for prediction of protein secondary structure, when contact number information is used (○) or not used (+). In each case, the proportion of the 25 datasets that have a certain performance, Q3, is shown.

In order to show the reproducibility of the results, the experiment was repeated five times after randomly shuffling the proteins in the subsets. In all cases, the prediction was improved significantly by the addition of CN information (*P*-values ranged from 10^{-3} to 10^{-8} in paired t-test). This shows that improvement of 2S prediction after addition of CN is robust and it will not be influenced by changing the protein training set.

CONCLUSION

Previous studies have been shown that additional structural information can help in the improvement of current protein structure prediction methods. Here, using neural networks based predictors it is shown that contact number can also be used as a rich source of information for improvement of secondary structure prediction. It might be possible to use a combination of contact numbers, accessible surface areas, protein structural classes, and other probable structural data to improve the prediction of secondary structures in proteins. Finally, this work suggests a demand for high-quality contact number prediction algorithms, which can provide the CN information in the real-world version of the problem with no information about the actual values of CN.

ABBREVIATIONS

2S: Secondary Structure; CN: Contact Number; NN: Neural Network

ACKNOWLEDGEMENTS

We are grateful to Dr. M. Sadeghi (NIGEB, Tehran) and Dr. A. Nowzari-Dalini (University of Tehran) for their fruitful comments.

REFERENCES

Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:467-75.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.

Anfinsen CB, Haber E. Studies on the reduction and re-formation of protein disulfide bonds. J Biol Chem 1960;236:1361-3.

Anfinsen CB, Haber E, Sela M, White FH. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 1961;47:1309-14.

Cohen BI, Presnell SR, Cohen FE. Origins of structural diversity within sequentially identical hexapeptides. Protein Sci 1993;2: 2134-45.

Costantini S, Colonna G, Facchiano AM. Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys Res Commun 2006;342:441-51.

Costantini S, Colonna G, Facchiano AM. PreSSAPro: A software for the prediction of secondary structure by amino acid properties. Comput Biol Chem 2007;31:389-92.

Dor O, Zhou Y. Achieving 80 % ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 2007;66:838-45.

Floudas CA, Fung HK, McAllister SR, Monnigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem Eng Sci 2006;61:966-88.

Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA 1996;93:5814-8.

Huang J-T, Wang M-T. Secondary structural wobble: The limits of protein prediction accuracy. Biochem Biophys Res Commun 2002;294:621-5.

Jones DT. Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol 1999;292:195-202.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-637.

Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA 1984;81:1075-8.

Liu K-H, Xia J-F, Li X. Efficient ensemble schemes for protein secondary structure prediction. Protein Peptide Lett 2008;15: 488-93.

Macdonald JR, Johnson WC. Environmental features are important in determining protein secondary structure. Protein Sci 2001;10:1172-7.

Marashi S-A, Behrouzi R, Pezeshk H. Adaptation of proteins to different environments: A comparison of proteome structural properties in Bacillus subtilis and Escherichia coli. J Theor Biol 2007;244:127-32.

Minor DL, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature 1996;380:730-4.

Momen-Roknabadi A, Sadeghi M, Pezeshk H, Marashi S-A. Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinformatics 2008;9:357.

Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 2006;7:301.

Pollastri G, Martin AJM, Mooney C, Vullo A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. BMC Bioinformatics 2007;8: 201.

Reyaz-Ahmed A, Zhang Y. A new SVM-based decision fusion method using multiple granular windows for protein secondary structure prediction. Lect Notes Comput Sci 2008;5009:324-31.

Rost B. Rising accuracy of protein secondary structure prediction. In: Chasman D, editor. Protein structure determination, analysis, and modeling for drug discovery. New York: Dekker, 2003;pp 207-49.

Rost B, Sander C. Prediction of protein secondary structure at better than 70 % accuracy. J Mol Biol 1993;232:584-99.

Sela M, White FH, Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. Science 1957;125:691-2.

Sudarsanam S. Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations. Proteins 1998;30:228-31.

Wang G, Dunbrack Jr. RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589-91.

Wang G, Dunbrack Jr. RL. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33:W94-W98.

White FH. Regeneration of native secondary and tertiary structures by air oxidation of reduced ribonuclease. J Biol Chem 1961; 236:1353-60.

Yüksektepea FÜ, Yılmaza Ö, Türkay M. Prediction of secondary structures of proteins using a two-stage method. Comput Chem Eng 2008;32:78-88.

Zhong L, Johnson WC. Environment affects amino acid preference for secondary structure. Proc Natl Acad Sci USA 1992; 89:4462-5.

Zhu Z-Y, Blundell TL. The use of amino acid patterns of classified helices and strands in secondary structure prediction. J Mol Biol 1996;260:261-76.