Available online at www.sciencedirect.com
ScienceDirect
Biomolecular Engineering 24 (2007) 237–243
Biomolecular Engineering
www.elsevier.com/locate/geneanabioeng

Application of expert networks for predicting proteins secondary structure

Sarit Sivan [a,*], Orna Filo [a], Hava Siegelmann [b]

[a] Department of Biomedical Engineering, Technion, Israel Institute of Technology, IIT, Haifa 32000, Israel
[b] Department of Computer Science, University of Massachusetts Amherst, Amherst MA 01003, United States

Received 5 November 2006; received in revised form 5 December 2006; accepted 6 December 2006

Abstract

The present study utilizes expert neural networks for the prediction of proteins secondary structure. We use three independent networks, one for each structure (alpha, beta and coil) as the first-level processing unit; decision upon the chosen structure for each residue is carried out by a second-level, post-processing unit, which utilizes the Chou and Fasman frequency values F_α and F_β in order to strengthen and/or deplete the probability of the specific structure under investigation. The highest prediction case was 76%.

Our method requires primitive computational means and a relatively small training set, while still been comparable to previous work. It is not meant to be an alternative to the determination of secondary structure by means of free energy minimization, integration of dynamic equations of motion or crystallography, which are expensive, time-consuming and complicated, but to provide additional constrains, which might be considered and incorporated into larger computing setups in order to reduce the initial search space for the above methods.

Keywords: Proteins; Secondary structure prediction; Expert neural networks; Chou and Fasman frequency parameters

1. Introduction

The knowledge of protein secondary structure is essential for the understanding of both the mechanisms of folding and the biological activity of proteins. X-ray diffraction has been successful in elucidating the three dimensional structure of many crystallized proteins. Although this method can be very accurate, it is expensive and time-consuming. Furthermore many membranes and ribosomal proteins have not yet yielded suitable crystals, so that other approaches must be explored to give the structural information required. Since experimental evidence shows that the native conformation of a protein is coded within its amino acid sequence (Anfinsen et al., 1961), many efforts have been made to predict the protein secondary and tertiary structure from the sequence data.

Following the pioneering work of Pauling and Corey (1951), which suggests that proteins form certain local conformations as helices and strands, 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). These methods exploit, in different ways, the correlation between amino acid and the local secondary structure, i.e. neighbors effect of no more than 10 amino acids away. The average success of these methods is 50–53% on three types of secondary structures (alpha-helix, beta-sheet, and coil) (Nishikawa, 1983; Kabsch and Sander, 1983a,b). Secondary structure predictions have been performed by various methods. These methods make use of the physicochemical characteristics of the amino acids (Lim, 1974; Ptitsyn and Finkelstein, 1983), sequence homology (Levin et al., 1986; Nishikawa and Ooi, 1986; Zvelebil et al., 1986), pattern matching (Cohen et al., 1983, 1986; Taylor and Thornton, 1983; Rooman et al., 1989; King and Sternberg, 1990; Presnell et al., 1992), 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; Biou et al., 1988; Di Francesco et al., 1997; Fasman, 1989; Garratt et al., 1991; Muggleton et al., 1992), and neural network (Bohr et al., 1988, 1993; Qian and Sejnowski, 1988; Holley and Karplus, 1989;

* Corresponding author at: Julius Silver Institute of Biomedical Sciences, Department of Biomedical Engineering, Technion, Israel Institute of Technology, IIT, Haifa 32000, Israel. Tel.: +972 4 8294150; fax: +972 4 8294599.
E-mail address: sarit@bm.technion.ac.il (S. Sivan).

1389-0344/$ – see front matter
doi:10.1016/j.bioeng.2006.12.001

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).

A promising approach in the area of secondary structure prediction is the use of neural network methods (Bohm, 1996). One of the first examples for this method used 48 proteins in the learning dataset, in order to teach the relationship between primary sequence and secondary structure to the neural network (Holley and Karplus, 1989). The overall accuracies achieved in this study and in a similar one (Qian and Sejnowski, 1988) were 63% and 64.3%, respectively, which had no major improvement compared with traditional methods of secondary structure prediction by statistical and knowledge-based methods.

Following the pioneering work of Qian and Sejnowski (1988), many new computational techniques involving neural networks for the prediction of proteins secondary structure were introduced (Holley and Karplus, 1989; Rost and Sander, 1993a,b, 1994; Hua and Sun, 2001; Armano et al., 2005; Lee et al., 2006; Huang et al., 2005; Ceroni et al., 2005; Ruan et al., 2005; Wood and Hirst, 2005; Meiler and Baker, 2003; Hering et al., 2003; Cai et al., 2002, 2003; Kaur and Raghava, 2003; Shepherd et al., 1999, 2003; Pal and Basu, 2001; Petersen et al., 2000; Cuff and Barton, 2000; Chandonia and Karplus, 1995, 1996, 1999; Kawabata and Doi, 1997; Barlow, 1995; Salamov and Solovyev, 1995); the average prediction accuracy achieved varies between 70% and 80%. In order to improve prediction accuracy, several studies applied sophisticated network structures such as hierarchical (Jordan and Jacobs, 1994; Huang et al., 2005; Barlow, 1995), cascade (Wood and Hirst, 2005) and multiple experts networks (Armano et al., 2005). Others combined additional structural information in the network input, for example, amino acid composition (Lee et al., 2006), interaction graphs (Ceroni et al., 2005), tertiary (Meiler and Baker, 2003; Chandonia and Karplus, 1995) and secondary (Rost and Sander, 1993a,b; Shepherd et al., 1999) structure information, information on the probabilities of residues buried in the protein core or on the protein surface (Vieth et al., 1992) and multiple sequence alignment profiles (Rost and Sander, 1993a,b, 1994; Cuff and Barton, 2000). Numerous methods involve preprocessing of protein sequence data using Fourier transform (Shepherd et al., 2003) and binary word encoding (Kawabata and Doi, 1997). Other approaches such as adaptive neuro-fuzzy inference system (Hering et al., 2003) and nearest neighbor algorithm (Salamov and Solovyev, 1995) combine additional classification algorithms with neural networks. Decoding the networks output in order to estimate the probability of finding a secondary structure at a specific position (Chandonia and Karplus, 1999) also provides more accurate prediction.

Our approach is to use three independent expert neural networks, one for each structure (alpha, beta and coil) as the first-level processing unit; decision upon the chosen structure for each residue is carried out by a second-level, post-processing unit, which utilizes the Chou and Fasman statistical frequency values F_α and F_β. This architecture takes into account the 'neighbors' effect and in turn, strengthens and/or depletes the probability of any structure under investigation to be part of a specific secondary structure.

Despite the simplicity of the networks presented in this work, they have the ability to deal with complex classification problems. This advantage was accomplished by separation of the comprehensive problem into three sub-classification items. Implementation of divide-and-conquer algorithms to deal with a complex problem by dividing it into simpler problems whose solutions can be combined to yield an answer to the complex problem was suggested by Jordan and Jacobs (1994).

2. Methods

2.1. Database

The secondary structure assignment used in this study was based on the work of Kabsch and Sander (1983a,b). Their DSSP program was used to classify known structures in the Brookhaven Protein Data Bank (BPDB) as helices and sheets. Residues that are neither helices nor sheets are classified as coil. Following Qian and Sejnowski (1988), we selected a representative sample of proteins that limited the number of almost identical sequence, such as the similar types of hemoglobin.

2.2. Network formulation and training

Three expert nets were applied in this work; each structure (alpha, beta and coil) is represented by a separate network (Fig. 1). All the networks used were feed-forward nets utilizing the back-propagation algorithm and the Sigmoid-Logistic as their activation function. Calculations were carried out using MATLAB. The input vector for each expert net encodes a moving window

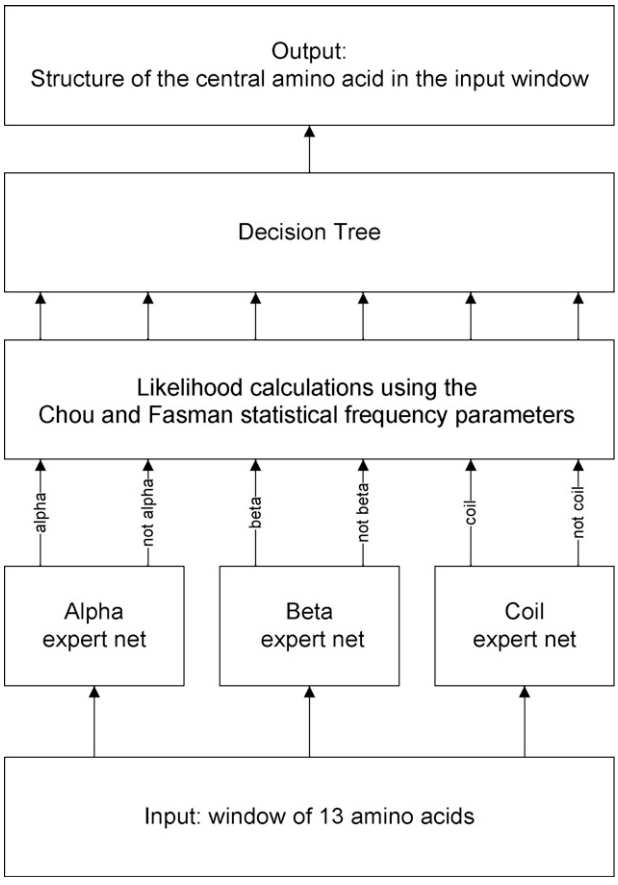

Fig. 1. A schematic description of the expert neural network used for the prediction of proteins secondary structure.

of 13 amino acids, in the protein's amino acid sequence. Prediction of the secondary structures was made for the central amino acid in this window. A binary encoding scheme is used for the network input. In this scheme, each amino acid at each window position is encoded by a group of 21 characters: one for each possible amino acid type at that position and one to provide a null input used when a moving window overlaps the edge of the protein. In each group of 21 characters, the input corresponding to the amino acid type is set to 1 and all other inputs are set to 0. Thus, the input consists of 13 groups of 21 characters each. The final output vector contains three output units, where each output neuron represents a different secondary structure (alpha, beta or coil).

2.3. The alpha, beta and coil expert networks

The expert networks have one or two hidden layers and two output neurons. The secondary structure is encoded in the output layers as follows:

- Alpha-expert: (1, 0) = (helix); (0, 1) = (not helix).
- Beta-expert: (1, 0) = (beta); (0, 1) = (not beta).
- Coil-expert: (1, 0) = (coil); (0, 1) = (not coil).

This encoding scheme is useful to prevent cases in which an amino acid might be classified into more than one structure (Havsteen, 1966).

The proteins listed in Tables 1 and 2 were used as training sets for the alpha and beta-expert nets, respectively; these proteins are non-homologous (comprises less than 17% identity in sequence). Training of the coil-expert net was carried out with the proteins listed in Table 2. These training sets are composed of %α = 49.8, %β = 4.4 and %coil = 45.8 for the α-expert net, and of %α = 4.1, %β = 40.1 and %coil = 55.8 for both the beta and coil expert nets. The size of these training sets (i.e. number of residues used for training) are nearly identical (997 residues for the alpha-expert net and 946 residues for the beta and coil expert nets). Each of the training sets was designed to give the maximum accuracy by strengthening the fraction (%α or %β or %coil) corresponds to the structure under investigation. In these training sets, for each structure, emphasis was given to the positive channel since the negative one is partially covered by the other expert networks. Moreover, the maximal percentage of a specific structure is limited by its natural average frequency in proteins.

All calculations at the training stage were carried out using adaptive learning rates in the following manner: if the new error exceeds the old one by more than 4%, the new weights, biases, outputs and errors are discarded and the learning rate is decreased; otherwise, the new weights are kept. On the other hand, if the new error is less than the old one, learning rate is increased by 5%. Initial learning rate was 0.01 and initialization of the weights was in the range of (−1, 1).

Table 1
List of proteins used for as the training set of the alpha-expert net

PDB code	Protein name	No. of residues	%α	%β	%Coil
1CPV	Calcium binding parvalbumin B	108	48.1	5.6	46.3
256B	Cytochrome B5 (oxidized)	85	24.7	24.7	50.6
251C	Cytochrome C551 (oxidized)	82	45.1	0	54.9
1FDX	Ferredoxin (*Peptococus aerogenes*)	54	9.3	7.4	83.3
1PPT	Avian pancreatic polypeptide	36	50	0	50
1GCN	Glucagon (pH 6–7)	29	48.3	0	51.7
1INS	Insulin (A and B chains)	51	43.1	5.9	53
7LZM	Lysozyme (Hen egg white, Triclinic)	129	29.5	7.8	62.7
1ECD	Hemoglobin (Erythrocruorin Deoxy)	136	71.3	0	28.7
2MHB	Hemoglobin (Horse, Aquo Met)	287	67.2	0	32.8
Total		997	49.8	4.4	45.8

Table 2
List of proteins used as the training sets for the beta and coil expert nets

PDB code	Protein name	No. of residues	%α	%β	%Coil
2PAB	Prealbumin (human plasma)	114	7	49.1	43.9
1ALP	Alpha lytic protease	198	3.5	42.9	53.6
1SGA	Proteinase A (*Streptomyces griseus*)	181	6.1	39.8	54.1
1EST	Tosyl-Elastase	240	5.4	34.2	60.4
2SOD	Cu–Zn super-oxide dismutase	151	0	38.4	61.6
1NXB	Neurotoxin B	62	0	41.9	58.1
Total		946	4.1	40.1	55.8

2.4. Integration of the three expert networks

The results for each of the three expert networks were further processed according to options 1–5 listed below. These options make use of the Chou and Fasman frequency parameters F_α and F_β (Wu and Kabat, 1971, 1973; Chou and Fasman, 1974a,b; Chou et al., 1972; Lewis and Scheraga, 1971), where $F_\alpha = f_\alpha / \langle f_\alpha \rangle$ and $F_\beta = f_\beta / \langle f_\beta \rangle$ are the helix and beta-sheet conformational parameters, respectively. f_α and f_β are the frequency of residues in the helix and beta regions. $\langle f_\alpha \rangle$ and $\langle f_\beta \rangle$ are the average frequency of residues in the helix and beta regions. Output from the expert nets was multiplied by F_α or F_β or by any other combination of which according to options 1–5 (below) in order to strengthen and/or deplete the probability of the specific structure under investigation. In all cases studied, the maximal value obtained was chosen as the most suitable structure for the central amino acid in the window.

The following options were applied:

1. Selecting the maximal value for one of three possible structures of the positive channels with no further process.
2. Output from the alpha and beta-expert nets was multiplied by F_α and F_β.
3. Output from the alpha and beta expert nets were multiplied by F_α and F_β, respectively, whereas output from the coil-expert net was divided by the average sum of F_α and F_β.
4. Output from the alpha-expert nets was multiplied by $F_\alpha \times 0.7$, output from the beta-expert net was multiplied by F_β and output from the coil expert net was divided by the absolute difference of $F_\alpha \times 0.7$ and F_β.
5. Output from the positive channels was processed as in option 4; in addition, the negative channels were divided by the same factors used for multiplication in option 4.

For all the above options, the suitable structure for the central amino acid in the window was chosen using a decision tree.

2.5. Network testing procedure

In the integrated system, the output consists of three units, each representing one of the possible secondary structures for the central amino acid. For a given input and set of weights and biases, the actual computed output will be a set of three numbers in the range 0–1. The secondary structure chosen was the output with the highest value. By doing so, undesired cases of predicting more than one possible structure for each amino acid were prevented. This problem evolves whenever actual outputs are converted to predictions with the use of threshold values (Ceroni et al., 2005). The non-homologous proteins (comprises less than 15% sequence identity) listed in Table 3 were used as the testing set, having a similar percentage of alpha and beta.

2.6. Performance measures

The most commonly used performance measure is a simple success rate, representing the positive successes. In this study, performance was measured for the three different expert networks as well as for the final integrated network.

Table 3
List of proteins used for testing the nets

PDB code	Protein name	No. of residues	%α	%β	%Coil
1FXC	Ferredoxin (*Spirulina platensis*)	98	0	0	100
1PCY	Plastocyanin	99	4	35.3	60.7
1LZM	Lysozyme (bacteriophage T4)	164	50.6	8.5	40.9
2ACT	Actinidin	218	25.7	18.3	56
1FAB	Lambda immunoglubulin FAB	426	0	42.5	57.5
1GPD	D-Glyceraldehyde-3-phosphate dehydrogenase	333	21.6	22.2	56.2
2GRS	Glutathione reductase	461	27.1	18.6	54.3
1HBL	Leghemoglobin (acetate, met)	153	69.3	0	30.7
1OVO	Ovomucoid third domain	56	17.8	21.4	60.8
Total		2008	22.7	22	55.3

The success rate of each expert network is reported as Q_1:

$$Q_1 = \frac{P_s}{N_s} \tag{1}$$

where N_s is the total number of the predicted residues of type s and P_s is the number of correctly predicted secondary structures of type s.

The success rate of the integrated network is measured by the index, Q_3, defined as the percentage of correctly predicted residues for all three types of secondary structures:

$$Q_3 = \frac{P_\alpha + P_\beta + P_{\text{coil}}}{N} \tag{2}$$

where N is the total number of the predicted residues and P_α, P_β and P_{coil} are the number of correctly predicted secondary structures of types alpha, beta and coil, respectively.

Q_3 refers to the ''positive'' successes in determining the secondary structures. For each data it can count up to 3 times if it is perfect. However, the ''negative'' successes, namely correct rejections are also considered as vital information. Therefore an additional index, $\tilde{Q}_3$, was defined:

$$\tilde{Q}_3 = \frac{\sum_1^N Y_{\alpha,i} + Y_{\beta,i} + Y_{\text{coil},i}}{3N} \tag{3}$$

where $Y_{\alpha,i}$ $Y_{\beta,i}$ and $Y_{\text{coil},i}$ equal to 1 for true positive or true negative predictions or 0 for false positive or false negative predictions for the ith residue and N is the total number of the predicted residues.

In order to take over prediction into account and to give a more meaningful measure in terms of each specific secondary structure, we calculated the Matthew's correlation coefficient (Szent-Gyorgyi and Cohen, 1957), for the three different structures:

$$C_s = \frac{p_s n_s - u_s o_s}{\sqrt{(n_s + u_s)(n_s + o_s)(p_s + u_s)(p_s + o_s)}} \tag{4}$$

where s is the alpha or beta or coil for helix, beta and coil, respectively; p_s the number of positive cases that were correctly predicted, for type s; n_s the number of negative cases that were correctly rejected, for type s; o_s the number of over-predicted cases (false positive), for type s; u_s the number of under-predicted cases (misses), for type s.

Table 5
Summary of results for the integrated system performance

Option	Q_3	$\tilde{Q}_3$	C_α	C_β	C_{coil}
Option 1	0.474	0.649	0.141	0.081	0.292
Option 2	0.536	0.691	0.117	0.220	0.268
Option 3	0.540	0.693	0.100	0.220	0.264
Option 4	0.551	0.701	0.114	0.238	0.258
Option 5 (decision tree)	0.560	0.701	0.121	0.247	0.258

Q_3, $\tilde{Q}_3$: prediction accuracy values; C_α, C_β and C_{coil} are the correlation coefficients for alpha, beta and coil, respectively.

3. Results and discussion

Kabsch and Sander (1983a,b) classified protein secondary structures into eight categories: three types of helices, two types of beta structures, two types of turns, and one for coil. We used only three structures: alpha-helix, beta-sheet, and coil. This choice was based on the results reported by Sasagawa and Tajima (1993), in which lower values of Q_3 (32–34%) were obtained with eight-structure classification compared to a three-structure classification. Alpha, beta and coil expert networks were trained and tested as independent ones (Fig. 1). Several tests were performed to determine the performance of the networks using the prediction accuracy measure, Q_1 (Eq. (1)). The results are presented in Table 4. One can see that for the α-expert net, two hidden layers have no advantage over one hidden layer. This is not the case for the coil expert; here, the best results were obtained with 2 hidden layers and 10 neurons in each layer. This is also the case for the beta-expert net.

The integrated system performance was assessed by the prediction accuracy values, Q_3 and $\tilde{Q}_3$ and the Matthew's correlation coefficients for each of the five different options. The Q_3 index refers only to the ''positive'' successes in determining the secondary structures. However, we believe that the ''negative'' successes, namely correct rejections, should also be considered as important information for first screening. Therefore, we defined an additional index, $\tilde{Q}_3$, in which true negative cases are also accounted for. Results are presented in Tables 5 and 6. As expected, better prediction rate (of up to 30%) was achieved with $\tilde{Q}_3$ compared to Q_3.

Table 4
Summary of the results for α-expert networks with one hidden layer

Expert network type	N_1	N_2	SSE	Q_1 (α or β or coil)	Q_1 (not α or not β or not coil)
Alpha	40	–	0.01	0.694	0.401
	10	10	0.01	0.642	0.460
Beta	10	–	8	0.588	0.574
	10	10	0.02	0.606	0.610
Coil	60	–	2	0.630	0.528
	10	10	0.003	0.674	0.510

N_1, N_2: number of neurons in the first and second hidden layers, respectively; SSE: sum of squared errors; Q_1: prediction accuracy value (Eq. (1)).

Table 6
Prediction statistics for the proteins testing set

PDB code	Protein name	Q_3	$\tilde{Q}_3$	C_α	C_β	C_{coil}
1FXC	Ferredoxin (*Spirulina platensis*)	0.756	0.843	–	–	–
1PCY	Plastocyanin	0.657	0.771	−0.061	0.336	0.389
1LZM	Lysozyme (bacteriophage T4)	0.439	0.626	0.037	0.267	0.237
2ACT	Actinidin	0.610	0.740	0.196	0.315	0.370
1FAB	Lambda immunoglubulin FAB	0.622	0.750	–	0.285	0.319
1GPD	D-Glyceraldehyde-3-phosphate dehydrogenase	0.528	0.686	0.146	0.226	0.165
2GRS	Glutathione reductase	0.534	0.690	0.066	0.318	0.215
1HBL	Leghemoglobin (acetate, met)	0.372	0.582	0.160	–	0.208
1OVO	Ovomucoid third domain	0.607	0.738	0.096	0.134	0.133

Q_3, $\tilde{Q}_3$: prediction accuracy values; SSE: sum of squared errors; C_α, C_β and C_{coil} are the correlation coefficients for alpha, beta and coil, respectively. Results were obtained after normalization according to option 5.

Integration of the three expert networks utilizes the Chou and Fasman frequency parameters, F_α and F_β that were used as statistical factors. The best prediction values Q_3 and $\tilde{Q}_3$ obtained using the integrated system were 56% compared to 70%, respectively (Table 5). When the prediction was carried out for each of the testing sets, the best values of Q_3 and $\tilde{Q}_3$ obtained were 76% and 84%, respectively (Table 6). Values for Q_3 varied between 37% and 76%, compared to higher prediction values obtained using $\tilde{Q}_3$ (between 58% and 84%). The fact that in both cases prediction values varies within a large range emphasizes the dependence of the results in the characteristics of the testing set. Our best prediction value ($Q_3 = 76\%$) is comparable to the average values obtained by other prediction algorithms, e.g. multiple experts (Armano et al., 2005), dihedral angles (Wood and Hirst, 2005) and multiple sequence alignment (Kaur and Raghava, 2003; Cuff and Barton, 2000; Rost and Sander, 1994) which all use a significantly bigger training sets. In addition, accuracy of 59–69% (Table 4) in the prediction of a specific structure can be achieved by using each of the expert networks separately. This is very useful whenever a determination of one structure is needed.

In the work presented by Kneller et al. (1990), proteins were subdivided into structural classes based on the knowledge of their sequence. Ruggiero et al. (1993) predicted secondary structure after classifying the tested protein into the appropriate group according to its α-helix content. Here, no earlier information for the prediction of secondary structure is required.

The suggested system obtained apparently a "lower" average value of Q_3 compared to the work of Qian and Sejnowski (1988) for example. In their study, saturation was achieved after training with about 8000 residues (9 times the size of the training set we used). However, our results coincide with those obtained by Qian and Sejnowski for a smaller training set of 1000 residues, the size used in this work. Therefore, since the improvement of prediction accuracy with large database is known to give better results (Chandonia and Karplus, 1995), we have the basis to believe that increase of the training set will end up with better results of Q_3 and of the correlation coefficients values.

4. Conclusion

Using a simple expert neural networks, trained with a small protein database, and followed by a post-processing unit which utilizes the Chou and Fasman frequency values F_α and F_β, it is possible to provide preliminary constrains, which are useful as a first screening step toward the determination of protein secondary structure of a given protein. Yet, in the future, this architecture should be using a larger training and testing protein sets.

Acknowledgements

One of the authors (Sarit Sivan) wishes to acknowledge the support provided by the Julius and Dorothea Harband Fellowship and to thank Dr. Daniel Ripoll for his help with the sequence homology.

References

Anfinsen, C.B., Haber, E., Sela, M., White, F.H., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 47, 1309–1314.

Armano, G., Mancosu, G., Milanesi, L., Orro, A., Saba, M., Vargiu, E., 2005. A hybrid genetic-neural system for predicting secondary structure. BMC Bioinformatics 6 (Suppl. 4), S3.

Barlow, T.W., 1995. Feed-forward neural networks for secondary structure prediction. J. Mol. Graph. 13, 175–183.

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.

Bohm, G., 1996. New approaches in molecular structure prediction. Biophys. Chem. 59, 1–32.

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 homology by neural networks. The alpha-helices in rhodopsin. FEBS Lett. 241, 223–228.

Cai, Y.D., Liu, X.J., Chou, K.C., 2003. Prediction of protein secondary structure content by artificial neural network. J. Comput. Chem. 24, 727–731.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Artificial neural network method for predicting protein secondary structure content. Comput. Chem. 26, 347–350.

Ceroni, A., Frasconi, P., Pollastri, G., 2005. Learning protein secondary structure from sequential and relational data. Neural Netw. 18, 1029–1039.

Chandonia, J.M., Karplus, M., 1995. Neural networks for secondary structure and structural class predictions. Protein Sci. 4, 275–285.

Chandonia, J.M., Karplus, M., 1996. The importance of larger data sets for protein secondary structure prediction with neural networks. Protein Sci. 5, 768–774.

Chandonia, J.M., Karplus, M., 1999. New methods for accurate prediction of protein secondary structure. Proteins 35, 293–306.

Chou, P.Y., Fasman, G.D., 1974a. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13, 211–222.

Chou, P.Y., Fasman, G.D., 1974b. Prediction of protein conformation. Biochemistry 13, 222–245.

Chou, P.Y., Wells, M., Fasman, G.D., 1972. Conformational studies on copolymers of hydroxypropyl-L-glutamine and L-leucine. Circular dichroism studies. Biochemistry 11, 3028–3043.

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.

Cuff, J.A., Barton, G.J., 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40, 502–511.

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.

Garnier, J., Osguthorpe, D.J., Robson, B., 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120, 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.

Havsteen, B.H., 1966. A study of the correlation between the amino acid composition and the helical content of proteins. J. Theor. Biol. 10, 1–10.

Hering, J.A., Innocent, P.R., Haris, P.I., 2003. Neuro-fuzzy structural classification of proteins for improved protein secondary structure prediction. Proteomics 3, 1646–1675.

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., Sun, Z., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.

Huang, X., Huang, D.S., Zhang, G.Z., Zhu, Y.P., Li, Y.X., 2005. Prediction of protein secondary structure using improved two-level neural network architecture. Protein Pept. Lett. 12, 805–811.

Jordan, M.I., Jacobs, R.A., 1994. Hierarchical mixtures of experts and the EM algorithm. Neur. Comp. 181–214.

Kabsch, W., Sander, C., 1983a. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

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

Kaur, H., Raghava, G.P., 2003. Prediction of beta-turns in proteins from multiple alignment using neural network. Protein Sci. 12, 627–634.

Kawabata, T., Doi, J., 1997. Improvement of protein secondary structure prediction using binary word encoding. Proteins 27, 36–46.

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.

Lee, S., Lee, B.C., Kim, D., 2006. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62, 1107–1114.

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.

Lewis, P.N., Scheraga, H.A., 1971. Predictions of structural homologies in cytochrome *c* proteins. Arch. Biochem. Biophys. 144, 576–583.

Lim, V.J., 1974. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. J. Mol. Biol. 88, 873–894.

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.

Meiler, J., Baker, D., 2003. Coupled prediction of protein secondary and tertiary structure. Proc. Natl. Acad. Sci. U.S.A. 100, 12105–12110.

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

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 homology applied to the prediction of protein secondary structures and joint prediction with existing methods. Biochim. Biophys. Acta 871, 45–54.

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

Pal, L., Basu, G., 2001. Neural network prediction of 3(10)-helices in proteins. Indian J. Biochem. Biophys. 38, 107–114.

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.

Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P., Lund, O., 2000. Prediction of protein secondary structure at 80% accuracy. Proteins 41, 17–20.

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.

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.

Rost, B., Sander, C., 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19, 55–72.

Ruan, J., Wang, K., Yang, J., Kurgan, L.A., Cios, K., 2005. Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. Artif. Intell. Med. 35, 19–35.

Ruggiero, C., Sacile, R., Rauch, G., 1993. Peptides secondary structure prediction with neural networks: a criterion for building appropriate learning sets. IEEE Trans. Biomed. Eng. 40, 1114–1121.

Salamov, A.A., Solovyev, V.V., 1995. Prediction of protein secondary structure by combining nearest neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247, 11–15.

Sasagawa, F., Tajima, K., 1993. Prediction of protein secondary structures by a neural network. Comput. Appl. Biosci. 9, 147–152.

Shepherd, A.J., Gorse, D., Thornthon, J.M., 1999. Prediction of the location and type of beta-turns in proteins using neural networks. Protein Sci. 8, 1045–1055.

Shepherd, A.J., Gorse, D., Thornton, J.M., 2003. A novel approach to the recognition of protein architecture from sequence using Fourier analysis and neural networks. Proteins 50, 290–302.

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.

Vieth, M., Kolinski, A., Skolnick, J., Sikorski, A., 1992. Prediction of protein secondary structure by neural networks: encoding short and long range patterns of amino acid packing. Acta. Biochim. Pol. 39, 369–392.

Wood, M.J., Hirst, J.D., 2005. Protein secondary structure prediction with dihedral angles. Proteins 59, 476–481.

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 homologous sequences. J. Mol. Biol. 195, 957–961.