Prediction of the Secondary Structure Content of Globular Proteins based on their Primary Sequences by the Artificial Neural Network

Takahiro SUZUKI*

Abstract

The secondary structure of a protein is of utmost importance to its biological activity. The artificial neural network (ANN) and multiple linear regression (MLR) calculations have been applied to predict the content of α-helix and β-strand of a globular protein based on its primary sequence. The amino acid composition and auto-correlation functions based on the hydrophobicity profile of the primary sequence have been taken into account in the modeling. For three-layer ANN models with back-propagation to predict the secondary structure content, the input pattern was a string of real numbers representing the protein's amino acid composition and auto-correlation functions. The output pattern was the content of α-helix and β-strand. The database of 261 proteins by Eisenhaber et al. [*Proteins*, **25**, 157-168; 169-179 (1996)] was used to make models. After dividing the total database of 261 proteins into training (131) and testing (130), good correlation with the secondary structure content from the MLR models and a significant improved fitting of the ANN over the MLR models was observed. The robustness of the neural model was successfully examined by the leave-half-out cross-validation procedure. The prediction of the secondary content may play an important role in the prediction of the protein's structure.

Key words: α-helix content, β-strand content, neural network, protein, primary sequence

1. Introduction

Proteins and polypeptide are made up of elementary building blocks or the amino acids and just 20 kinds of different amino acids occur in proteins except for some special cases. These amino acids are arranged sequentially in a protein and the sequence is called the *primary structure*. This linear sequence folds and turns into a unique three-dimensional structure containing global features that are referred to as the *secondary structure*. There are

* Natural Science Laboratory, Toyo University, 5-28-20 Hakusan, Bunkyo-ku, Tokyo 112-8606, Japan

three types of secondary structures, α-helix, β-strand, and random coil. In an α-helix structure, the protein chain turns continuously in the same direction to form a "spiral"; in a β-strand, two or more parts of the same chain are aligned parallel in space; the random coil collects all the other more or less irregular three-dimensional arrangements of amino acids. All of these can be found in one protein. Amino acids in a protein are generically called *residues*.

The biological activity of proteins depends primarily on their spatial conformation and knowledge of protein structures plays a key role in understanding their functions. Therefore, the prediction of the tertiary structure of a protein from knowledge of its amino acid sequence (Anfisen, 1973) is one of the most challenging problems in molecular biology. Although present knowledge is still insufficient to predict the tertiary architecture of any protein correctly, *a priori* knowledge of the secondary structure content (i.e., the percentage of residues in different secondary structural states) of a protein structure is useful. Firstly, the secondary structure content provides an intuitive description of protein structure; secondly, the secondary structure prediction from the primary sequence can be significantly improved by incorporating the effect of the secondary structure content; thirdly, information on the secondary structure content could reduce the scope of searching conformation space or set a good staring points for the energy minimization process in the prediction of the tertiary structure.

The secondary structural content can be determined by spectroscopic methods such as IR Raman spectroscopy (Bussian & Sander, 1989) and circular dichroism (CD) spectroscopy in the UV absorption range (Sreerama & Woody, 1994). The characterization of a protein structure by its secondary structural content is an intermediate level between the complete description of the secondary structural states of every residue and the simple assignment of a secondary structural class (folding types all α-helix, all β-strand, $\alpha+\beta$ or α/β, irregular). However, unfortunately, there are no general experimental methods suitable for all proteins and the accuracy of the experimental methods is sometimes not so good. Besides those methods are laborious, time-consuming and require the use of suitable single crystals which can not be easily obtained for several proteins. Therefore, theoretical prediction of the secondary structural contents of a protein is needed.

Several attempts have been made to predict protein secondary structure content from the conventional amino acid composition. The aim is to predict a reduced representation of a protein structure, the percentage of residues defined as helix or strand, from a reduced representation of its sequence, the proportion of each of the 20 different amino acids. The multiple linear regression (MLR) scheme to predict the secondary structure content of a protein on the basis of its amino acid composition was introduced (Krigbaum & Knutton, 1973). The most widely used traditional approach is the method of Chou and Fasman (1974a, b), which allows one to predict whether a certain amino acid is part of an α-helix, a β-strand, or random coil from the amino acid sequence. The basic assumption of in the work

of Chou and Fasman is that the identities of an amino acid and its neighbors determine the secondary structure of that neighborhood with about 50-53% correctness.

The MLR approach has been improved by Muskal and Kim (1992) based on a larger database of proteins and the result was compared with that of artificial neural network (ANN) approach. In the ANN approach, the protein's amino acid composition, the molecular weight and the presence or absence of the heme group in a protein were used as input data. The method utilized two neural networks placed in tandem to predict secondary structure content: they reported that these two networks together gave prediction errors as low as 5.0 and 5.6% for helix and strand content, respectively, on a set of protein crystal structures reported to show little sequence homology to those used in the training of the network. They concluded that the tandem ANN scheme gave better predictions than other methods, MLR, a non-hidden node network, and a secondary structure assignment analysis. A new analytic vector decomposition method to predict the secondary structural contents of a protein relying on its amino acid composition has been developed (Eisenhaber et al., 1996a). An improved MLR method to predict the contents of α-helix and β-strand of a globular protein based on its primary sequence and structural class has been proposed (Zhang et al., 2001).

However, the exercise showed that existing models performed poorly in predicting the secondary structure content. One of the reasons might come from the fact that most models used linear relationships between the secondary structure content and the amino acid composition and omit of nonlinear effects and the coupling effect of the frequencies of different amino acids. In this study, artificial neural network (ANN) modeling was applied to make a new prediction scheme for the secondary structure content of globular proteins from their amino acid sequences.

2. Materials and Methods

2.1. Database

For the development of a prediction approach based on the computational techniques, it is important to choose a representative database. A database of 261 proteins was chosen (Eisenhaber et al., 1996a, b) and 131 proteins ware used for training set and 130 proteins were for testing set. The PDB codes for the 261 proteins with resolution better than or equal to 2.0 Å are listed in Tables 1 and 2. The protein structural class and secondary structure contents for each protein were obtained by the standard DSSP method (Kabsch & Sander, 1983). No two proteins in the set share more than 35% sequence identity over a length of more than 80 residues. Although the original database of Eisenhaber et al. includes 262 proteins, one protein (PDB code: 3bcl-) was removed because of too many nonstandard amino acids in its primary sequence.

Table 1. The PDB codes of 131 proteins in the training set.

153L-_-1	1ALK-A-1	1BAM-_-1	1BSR-A-1	1CMB-A-1	1DBS-_-1	1EDB-_-1	1FRD-_-1	1GLG-_-1	1HBG-_-1	1HUW-_-1	1LIS-_-1	1MNG-A-1
1AAJ-_-1	1AMP-_-1	1BBH-A-1	1BTC-_-1	1COB-A-1	1DDT-_-1	1ENJ-_-1	1FRP-A-1	1GLP-A-1	1HBI-A-1	1I1B-_-1	1LKI-_-1	1MNI-A-1
1AAZ-A-1	1AOZ-A-1	1BBP-A-1	1BTL-_-1	1CON-A-1	1DHI-A-1	1EPT-B-2	1FUS-_-1	2GLT-_-1	1HBQ-_-1	1IAG-_-1	1LMB-4-4	1MNS-_-1
1ABK-_-1	1APB-_-1	1BCX-_-1	1CBS-_-1	1COT-_-1	1DMB-_-1	1ESL-_-1	1FX1-_-1	1GMP-A-1	1HDS-A-1	1ICM-_-1	1LOB-A-1	1MOL-A-1
1ACF-_-1	1APM-E-1	1BGC-_-1	1CDG-_-1	1CP4-_-1	1DOB-_-1	1EZM-_-1	1GCS-_-1	1GOB-_-1	1HDS-B-2	1IDS-A-1	1LTS-D-1	1MPP-_-1
1ACX-_-1	1APT-E-1	1BGH-_-1	1CDM-A-1	1CPC-A-1	1DRF-_-1	1FDD-_-1	1GCT-A-1	1GOF-_-1	1HFC-_-1	1IHS-H-2	1LTS-A-6	1MRH-_-1
1ADL-_-1	1ARB-_-1	1BMD-A-1	1CEL-A-1	1CPC-B-2	1DRK-_-1	1FKB-_-1	1GD1-O-1	1GOX-_-1	1HIP-_-1	1KNB-_-1	1MBA-_-1	1MUA-_-1
1AIZ-A-1	1ARP-_-1	1BPQ-_-1	1CEW-I-1	1CSC-_-1	1DSB-A-1	1FLP-_-1	1GDI-_-1	1GP1-A-1	1HLE-A-1	1L49-_-1	1MDC-_-1	1NAR-_-1
1ALC-_-1	1AST-_-1	1BRS-F-6	1CFB-_-1	1CSE-E-1	1EAS-_-1	1FNA-_-1	1GER-A-1	1GPB-_-1	1HNA-_-1	1LCF-_-1	1MFE-H-2	1NBA-B-2
1ALD-_-1	1AYA-A-1	1BSE-A-1	1CHM-A-1	1CYO-_-1	1ECA-_-1	1FNB-_-1	1GKY-_-1	1GPR-_-1	1HTR-B-2	1LEC-_-1	1MLD-A-1	1NBV-L-1
1NHK-L-2												

Table 2. The PDB codes of 130 proteins in the test set.

1NHQ-_-1	1OYA-_-1	1PPF-E-1	1SHA-A-1	1THV-_-1	1WHT-A-1	2BAT-_-1	2CLR-A-1	2FCR-_-1	2MHR-_-1	2PKC-_-1	2SPC-A-1	3RUB-L-1
1NIS-_-1	1PBP-_-1	1PRN-_-1	1SLT-A-1	1TIB-_-1	1WHT-B-2	2BBK-H-1	2CLR-B-2	2FGF-_-1	2MSB-A-1	2PLT-_-1	2TRX-A-1	3RUB-S-2
1NPC-_-1	1PDA-_-1	1REC-_-1	1SMR-A-1	1TML-_-1	1XIB-_-1	2BBK-L-2	2CST-A-1	2HPE-A-1	2NAC-A-1	2POR-_-1	351C-_-1	3TGL-_-1
1NSC-A-1	1PGS-_-1	1RIS-_-1	1SRE-A-1	1TPH-1-1	1YEA-_-1	2BLT-A-1	2CUT-_-1	2HPR-_-1	2OHX-A-1	2RAN-_-1	3COX-_-1	4BLM-A-1
1OFV-_-1	1PHP-_-1	1RSY-_-1	1SRP-_-1	1TRB-_-1	1YTB-A-1	2BMH-A-1	2CY3-_-1	3HSC-_-1	2P07-_-1	2RMC-A-1	3CPA-_-1	4CLA-_-1
1OLB-A-1	1PII-_-1	1RVA-A-1	1TAD-A-1	1TRK-A-1	256B-A-1	2BOP-A-1	2CYP-_-1	2HTS-_-1	2PAZ-_-1	2RSP-B-2	3DFR-_-1	4ENL-_-1
1OMD-_-1	1PIP-A-1	1SAC-A-1	1TCA-_-1	1TRO-A-1	2ACU-_-1	2C2C-_-1	2CYR-_-1	2IHL-_-1	2PGD-_-1	2SCP-A-1	3DNI-_-1	5RUB-B-2
1ONC-_-1	1POC-_-1	1SBP-_-1	1TEN-_-1	1TTB-A-1	2AK3-B-2	2CCY-A-1	2EBN-_-1	2LAO-_-1	2PIA-_-1	2SIC-I-2	3FXN-_-1	6Q21-A-1
1OPA-A-1	1POH-_-1	1SGC-_-1	1TFG-_-1	1TYS-_-1	2APR-_-1	2CHE-_-1	2EXO-_-1	2LHB-_-1	2PKA-A-1	2SIL-_-1	3PGA-4-4	9LDT-A-1
1OVA-C-3	1PPA-_-1	1SGT-_-1	1THG-_-1	1UKZ-_-1	2AYH-_-1	2CHS-A-1	2FBJ-H-2	2MCM-_-1	2PKA-B-2	2SNS-_-1	3RP2-A-1	2CWG-A-1

2.2. Autocorrelation function

The amino acid composition of a protein is represented by 20 numbers. The amino acid composition is related to the structural classes, the thermal stability, the cellular positions, and the secondary structural content. However, the amino acid composition is not enough for explaining the secondary structural content from the primary sequences. Therefore, the auto-correlation function based on the hydrophobicity of residues to reflect the profile of the hydrophobicity indices of residues along the sequence was used. To calculate the auto-correlation functions, replace each residue in the primary sequence by its hydrophobicity index, respectively. The hydrophobicity index of Fauchere & Pliska (1983) which is listed in Table 3 was employed. The hydrophobicity index is a set of 20 numerical values representing

Table 3. Hydrophobicity index used in this study.

Amino acid	Hydrophobicity indices*
Ala(A)	0.42
Cys(C)	1.34
Asp(D)	−1.05
Glu(E)	0.87
Phe(F)	2.44
Gly(G)	0.00
His(H)	0.18
Ile(I)	2.46
Lys(K)	−1.35
Leu(L)	2.32
Met(M)	1.68
Asn(N)	−0.82
Pro(P)	0.98
Gln(Q)	−0.30
Arg(R)	−1.37
Ser(S)	−0.05
Thr(T)	0.35
Val(V)	1.66
Trp(W)	3.07
Tyr(Y)	1.31

*: The unit is kcal mol^{-1}.

the hydrophobicity of the 20 amino acids. The replacement results in a numerical sequence

$$h_1, h_2, \ldots\ldots, h_N \quad (1)$$

where h_i is the hydrophobicity index for the ith residue and N is the number of residues of the query protein. For some nonstandard amino acids in the sequence, they are simply assigned to zero.

The autocorrelation function r_n of the sequence (1) is defined as (Cornette et al., 1987)

$$r_n = \frac{1}{N-n}\sum_{i=1}^{N=n} h_i h_{i+n}, \qquad n=1,2,3,\ldots\ldots, \quad (2)$$

where h_i is the hydrophobicity index for the i-th residue in the primary sequence. According to the study by Zhang et al.(1998), 10 terms (n =1 to 10) of the auto-correlation functions lead to an optimal result or minimum prediction errors.

2.3. The artificial neural network (ANN) modeling

The artificial neural networks (ANNs) have an inherent ability to provide non-linear and cross product terms for QSAR modeling. The ANNs may be especially useful when a rigid theoretical basis and/or mathematical relationship to describe a phenomenon to be modeled is not available in advance.

From many ANN approaches, both different in architecture and in learning algorithms, the three-layer ANNs with the back-propagation of errors were employed in this study. Since the theory and practical application of the ANN are popular, an explanation of the methodology can be delegated to the literature (Zupan & Gasteiger, 1999). The most commonly used log sigmoid transfer function and the delta rule for the error correction formula were used in the networks. The ANN calculations were carried out by using our in-house program.

2.4. The multiple linear regression (MLR) model

The MLR model was used here to find significant descriptors for ANN model and to compare its modeling performance with those by ANN models. It is assumed that the contents of α-helix and β-strand of a protein which are denoted by f_α and f_β, respectively, can be expressed as linear functions of its amino acid composition and the auto-correlation function:

$$f_\alpha = \sum_{i=1}^{20} a_i x_i + \sum_{i=1}^{10} a_j r_j + a_0 + e_k \quad (3)$$

$$f_\beta = \sum_{i=1}^{20} b_i x_i + \sum_{i=1}^{10} b_j r_j + b_0 + e_k \quad (4)$$

where x_i (i=1 to 20) are the frequencies of occurrence of the 20 amino acids for a protein, r_j (j=1 to 10) are the auto-correlation functions of the sequence, a_i, a_0, a_j, b_i, b_j and b_0 are the regression coefficients to be determined by the data, and e_k the deviations or residuals. The 20 amino acids are ordered alphabetically according to their single-letter codes. They

correspond to A(Ala), C(Cys), D(Asp), E(Glu), F(Phe), G(Gly), H(His), I(Ile), K(Lys), L(Leu), M(Met), N(Asp), P(Pro), Q(Glu), R(Arg), S(Ser), T(Thr), V(Val), W(Trp), and Y(Tyr), respectively. For example, x_1 represents the frequency of occurrence for alanine (A), x_2 for cystein (C), and so forth.

Therefore, there are totally 30 input parameters including 20 amino acid frequencies and ten auto-correlation functions in the regression formulae. The importance of the all parameters or descriptors was evaluated by statistical analysis and just significant parameters were included in the final models.

2.5 Statistical Parameters

The calibration and the prediction quality of the MLR and ANN modeling results for both training and test sets were evaluated using the following parameters: the squared correlation coefficient R^2,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{fit})^2}{\sum_{i=1}^{n}(y_i - y_{mean})^2} \tag{5}$$

and average absolute error AAE,

$$AAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^{fit}| \tag{6}$$

where n is the number of proteins. In above formulas, y_i represents the observed target value for the ith protein, y_{mean} denotes the associated mean, and y_i^{fit} represents the calculated target value using the ANN and MLR model.

3. Results and Discussion

3.1. MLR models

On the basis of the data in the training set, the following two MLR models for each class were obtained for the training set of 131 proteins, using a set of 30 descriptors. by variable selection based on the T-square criterion (Mager, 1992):

$$f_\alpha = 0.087x_1 - 0.029x_6 + 0.040x_9 + 0.066x_{10} - 0.043x_{13} - 0.035x_{17} - 0.047x_{18} - 0.514r_2 + 0.201r_3 + 0.334r_4 - 0.106r_6 + 0.129r_7 + 0.128 \tag{7}$$

$$f_\beta = -0.075x_1 - 0.056x_2 - 0.095x_7 - 0.009x_9 + 0.052x_{17} + 0.107x_{18} + 0.150x_{19} - 0.159r_1 + 0.295r_2 - 0.175r_3 - 0.231r_4 + 0.235 \tag{8}$$

The statistical analysis to find the significant descriptors was carried out using a heuristic method based on the linear regression technique. This procedure is based on the scale

forward selection technique (Mager, 1992). The analyses revealed that 12 and 11 descriptors were proved to be the significant model parameters for the contents of α-helix and β-strand, respectively. The alternative regression models with other combinations of descriptors gave slightly inferior fit on the given set of proteins. Therefore, the significant descriptors in above models were implemented as the inputs for a neural network. The results are compared with those by NN models as below.

3.2. ANN models

The architecture of the neural network model used in this study is shown in Fig. 1. The network consists of three layers: an input layer, a hidden layer and an output layer. Each layer is comprised of individual processing units called neurons and represented here as circles. Since the number of hidden neurons strongly influences the predictive quality of the derived model, the number of hidden neurons was set to be variable. Both input and hidden layers have an additional neuron, termed a bias neuron, respectively.

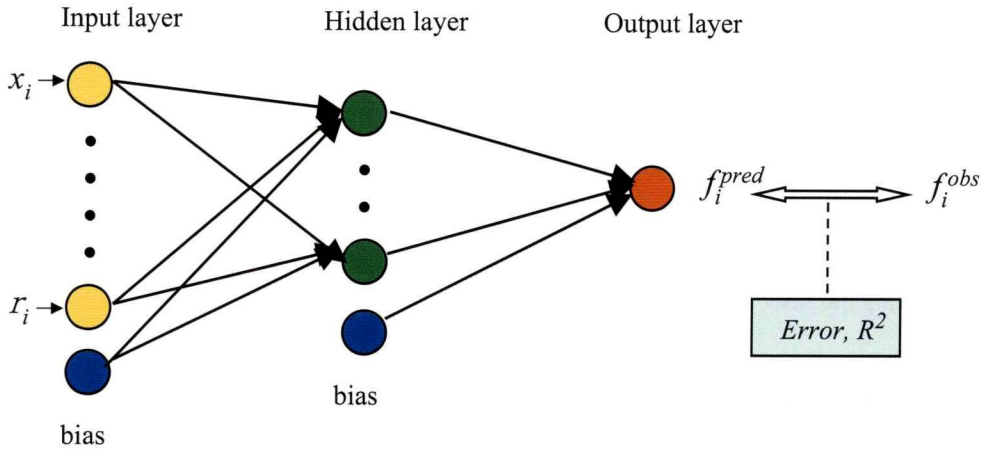

Fig. 1. Architecture of the three-layer neural network model used in this study.

As described above, in the MLR analysis 12 descriptors in the first model for f_α [see equation (7)], respectively, 11 descriptors in the second model for f_β [see equation (8)] were found to be effective for the given training set proteins. The same descriptors were implemented as inputs for the ANN to compare the results with those obtained by MLR models. Besides, other three sets of alternative descriptors from all original 30 descriptors were employed as inputs for f_α and f_β, respectively. (Tables 4 and 5).

These input nodes plus a bias, a varying number of hidden-layer neurons (between two to ten plus one bias), and a single output neuron corresponding to the contents of α-helix and β-strand of a protein are included in the ANN model. The statistical quality of the ANN and

Table 4. Input data for the neural network modeling of α-helix content.

Numbers	Descriptors
7	x_1, x_6, x_9, x_{10}, x_{13}, x_{17}, x_{18}
12	x_1, x_6, x_9, x_{10}, x_{13}, x_{17}, x_{18}, r_2, r_3, r_4, r_6, r_7
20	x_1 to x_{20}
30	x_1 to x_{20} + r_1 to r_{10}

Table 5. Input data for the neural network modeling of β-strand content.

Numbers	Descriptors
7	x_1, x_2, x_7, x_9, x_{17}, x_{18}, x_{19}
11	x_1, x_2, x_7, x_9, x_{17}, x_{18}, x_{19}, r_1, r_2, r_3, r_4
20	x_1 to x_{20}
24	x_1 to x_{20} + r_1 to r_4

MLR modeling results for training and test sets were evaluated based on both squared correlation coefficient R^2 or R and average absolute error AAE.

The best architecture was determined to be at four hidden-layer neurons plus a bias for both f_α and f_β modeling. Figures 2 and 3 shows the effects of the choice of input parameters on the R values as a function of training steps for the training set and test set, respectively, together with the results by MLR models using same sets of descriptors. In the training of the ANN models, as can be seen from Fig.3, R values reached their maximum at 10000 epochs for f_α and 7500 epochs for f_β, respectively. The quality of fitting, the values of R having more than 0.9 were attained for both f_α and f_β with the ANN model, was found to be better than that obtained with the corresponding MLR models. This improvement is probably due to the ANNs ability to include interactions among input descriptors as well as non-linearities. The prediction results obtained by the ANN and MLR models are compared in Fig. 4. The prediction performance of two types of models is just comparable with compared to those reported in previous works.

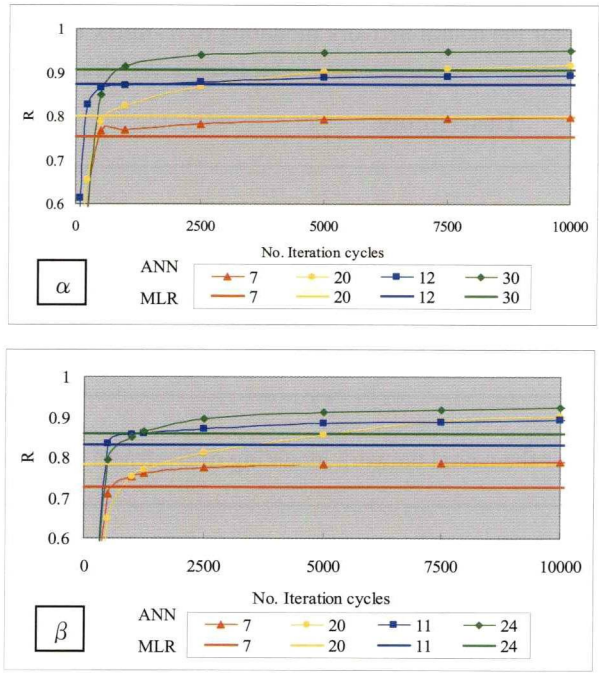

Fig.2. Effects of training iteration cycles on the performance of the neural network models with different sets of inputs for the training set of 131 proteins.

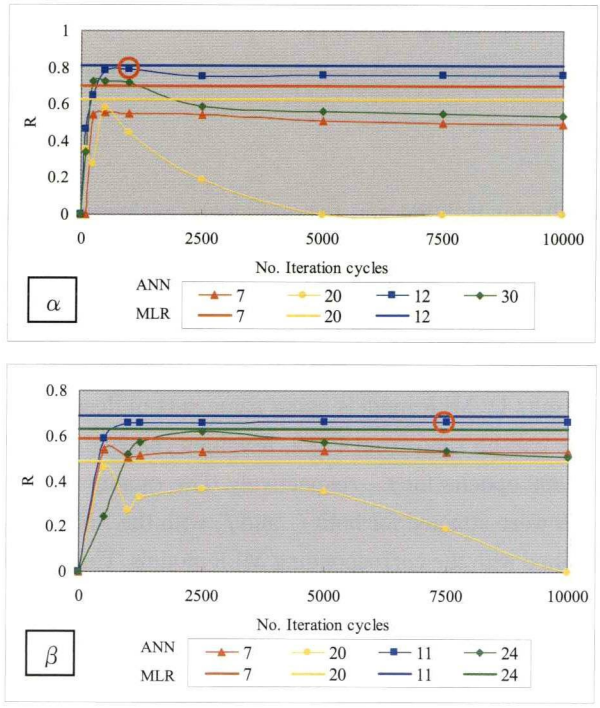

Fig.3. Effects of training iteration cycles on the performance of the neural network models with different sets of inputs for the test set of 130 proteins.

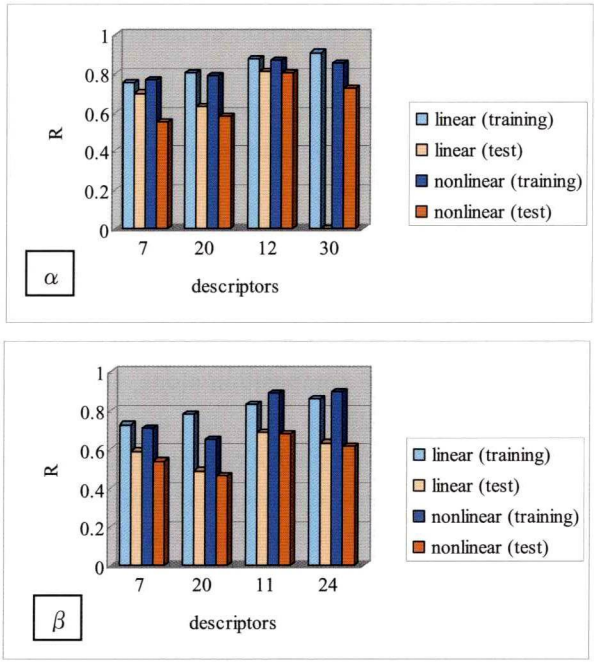

Fig.4. Results of the calibration and prediction for 231 proteins by linear (MLR) and nonlinear (ANN) models with different sets of descriptors.

Conclusions

The results obtained in this paper demonstrate that it is possible to generate robust neural network models capable of prediction of the secondary structural contents of proteins. The inputs required for the models were just the information on the amino acid composition and auto-correlation functions based on the hydrophobicity profile of the primary sequence. The MLR approach leads to a better interpretation of the contribution of individual terms, but ANNs can exact more information from the data than statistical methods, especially where non-linear relationships are involved. The present method would be of not only practical value, but also some theoretical interest.

References

Anfisen, C. B., 1973. Principles that govern the folding of protein chains, *Scinence* **181**, 223-230.

Bussian, B.M., Sander, C.,1989. How to determine protein secondary structure in solution by Raman spectroscopy: Practical guide and test case DNase I, *Biochemistry* **28**, 4271-4277.

Chou, P.Y., Fasman, G. D., 1974a. Conformational parameters for amino acids in helical, β-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**, 222-241.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., DeLisi, C., 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, *J. Mol. Biol.* **195**, 695-685.

Eisenhaber,F., Imperiale, F., Argos, P., Frommel, C., 1996a. Prediction of secondary structural contents of proteins from their amino acid composition alone, I. New analytic vector decomposition methods, *Proteins* **25**, 157-168.

Eisenhaber,F., Frommel, C., Argos, P., 1996b. Prediction of secondary structural contents of proteins from their amino acid composition alone, II. The paradox with secondary structural class, *Proteins* **25**, 169-179.

Fauchere, J.L., Pliska, V., 1983. Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369-375.

Kabsch W., Sander, C., 1983. Dictionary of protein secondary structures: Pattern of recognition of hydrogen-bounded and geometrical features. *Biopolymers* **22**, 2577-2637.

Krigbaum, W. R., Knutton, S.P., 1973. Prediction of the amounts of secondary structure in a globular protein from its amino acid composition, *Proc. Nat. Acad. Sci.USA* **70**, 2809-2813.

Mager, P. P., 1992. In *QSAR in Design of Bioactive Compounds.* Proc. 2nd Telesymp. Med. Chem., M. Kuchar (ed.), J.R. Prous, Barcelona, pp.446-469.

Muskal, S. M., Kim, S.-H., 1992. Predicting protein secondary structural content: A tandem neural network approach, *J. Mol.Biol.* **225**, 713-717.

Sreerama, N., Woody, R.W., 1994. Protein secondary structure from circular dichroism spectroscopy. *J. Mol. Biol.* **242**, 497-507.

Zhang, C.-T., Lin, Z.-S., Zhang, Z., Yan, M., 1998. Prediction of the helix/strand content of globular proteins based on their primary sequences. *Protein Engng.* **11**, 971-979.

Zhang, Z., Sun Z.-R., Zhang, C.-T., 2001. A new approach to predict the helix/strand content of globular proteins, *J. Theor. Biol.* **208**, 65-78.

Zupan, J., Gasteiger, J., 1999. *Neural Networks for Chemistry and Drug Design,* 2nd Edition, Wiley-VCH, Weinheim.