VOL.2 NO.2September 2013 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.com.ba

IUS Soft Computing Research Group

Regression Analysis to Predict the Secondary Structure of Proteins

Betul Akcesme and Faruk B. Akcesme
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
betul.cicek@yahoo.com; fakcesme@ius.edu.ba

Article Info

Article history:
Article received Sep.2014
Received in revised form Oct 2014

Keywords:
Secondary structure; Conformation of proteins; Statistical methods

Abstract

A method is presented for protein secondary structure prediction based on the use of multidimensional regression. 200 proteins are chosen from RCSB Protein Database. Their secondary structures obtained through x-ray crystallography analyses are downloaded from the same source. Primary and secondary structure of proteins are concatenated separately to create a sequence of 169 026 residues. First 150 000 of the amino acid residues and corresponding secondary structures are chosen to create a regression model. The remaining 19 026 residues are used for testing. Since we expect three outputs α-helices "S", β-sheets "H", and coiled coils "C", our regression modes consists of $3 \times 20 \times 23$ parameters. These parameters are tuned and a correct classification rate of 62.50% is achieved on the test data. Furthermore, the performance of the regression model compared with online secondary structure estimation algorithms on 14 unused proteins, and the performance of the regression model is found comparable with the online estimation tools.

1. INTRODUCTION

Large-scale sequencing projects produced a large number of protein sequences. In 1993 the number was 26,000 (Bairoch & Boeckmann, 1963; Ewbank & Creighton, 1992) sequences, but before the end of the century the number easily past the 500,000 limit. Today, at the end of the year 2014 the number reached to 546,790.

To compare the number of known proteins sequences, the number of proteins which is known by structure is still very limited, in 1993 it was at about 1000 (Bernstein et al., 1977). Today it reached at 105,025 increased efforts focused on narrowing the widening gap. The most reliable prediction of the structure of new proteins is done by detection of significant similarities to proteins of known structure (Taylor & Orengo, 1989; Sander & Schneider, 1991; Vriend & Sander, 1991). But only about one-seventh of new sequences have similarities to known structures (Bork et al., 1992) in the years 1993.

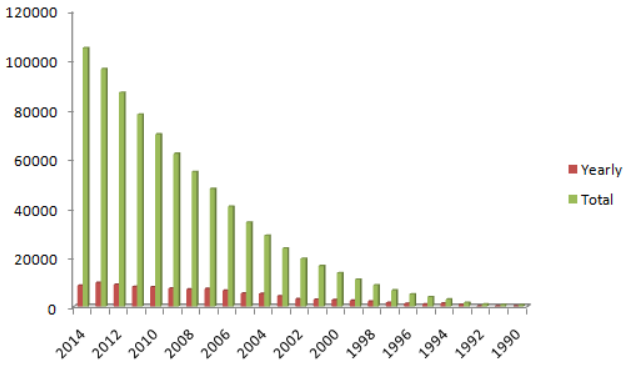

Figure 1. Number of proteins whose structures are known

Attempts to predict structure from sequence by physical simulation techniques, such as molecular dynamics (Momany et al., 1975; Karplus & Petsko, 1990), have fallen far short of solving the task of finding the "hidden" relation between the primary and tertiary structure. Although the folding process may require catalysts such as chaperonins (Hubbard & Sander, 1991), the basic hypothesis that the three dimensional (tertiary) structure of a protein is uniquely determined by its sequence of amino acids (primary structure) appears to remain valid (Anfinsen et al., 1963; Ewbank & Creighton, 1992). A simple reduction of the prediction problem is the projection of the three-dimensional structure onto one dimension, i.e. onto a string of secondary structure assignments for each residue.

Secondary structure predictions have been performed by various methods (Szent-Györgyi & Cohen, 1957; Periti et al., 1967; Ptitsyn, 1969; Pain & Robson, 1970; Robson & Pain, 1971), ever since Pauling suggested that proteins form certain local conformational patterns like helices and strands (Pauling & Corey, 1951; Pauling et al., 1951). The different algorithms can be approximately grouped into those using (1) statistical information (Nagano, 1973; Chou & Fasman, 1974; Nagano & Hasegawa, 1975; Garnier et al., 1978; Schulz & Schirmer, 1979; Levin et al., 1986; Gibrat et al., 1987; Biou et al., 1988; Kanehisa, 1988; Levin & Garnier, 1988; Fasman, 1989; Garrett et al., 1991; Muggleton et al., 1992); (2) physico-chemical properties (Lim, 1974; Ptitsyn & Finkelstein, 1983); (3) Sequence patterns (Cohen et al., 1983, 1986; Taylor & Thornton, 1983; Rooman et al ., 1989, 1991; Sternberg & King, 1990; Rooman & Wodak, 1991; Presnell et al., 1992); (4) multi-layered (or neural) networks (Bohr et al., 1988, 1990; Qian & Sejnowski, 1988; Holley & Karplus, 1989; Bossa & Pascarella, 1990; Kneller et al., 1990; Hirst & Sternberg, 1992; Maclin & Shavlik, 1993; Stolorz et al., 1992; Zhang et al., 1992); and (5) evolutionary conservation (Maxfield & Scheraga, 1979; Zvelebil et al., 1987; Frampton et al., 1989; Benner & Gerloff, 1990; Barton et al., 1991; Niermann & Kirschner, 1991; Ouzounis & Melvin, 1991; Musacchio et al., 1992; Russell et al ., 1992; Gibson et al., 1993).

One of the problems of these prediction methods is that the formation of secondary structure elements is only to a certain degree due to sequentially local interaction of amino acids (Nagano & Hasegawa, 1975; Taylor, 1988; Zhong et al., 1992). However, most methods known to date do rely on local information. For the 1980's these methods have hovered around 60 to 64% in overall three-state accuracy. Some methods predicted, e.g. β-strands, only 12 percentage points better than the chance value of 33 % (Biou et al., 1988). In 1990's, the reported overall accuracy of 66,5% (Zhang et al., 1992) and single examples of predictions of proteins of unknown structure have generated enthusiasm in the field (Barton et al., 1991; Benner et al., 1992; Rost & Sander, 1992; Russell et al., 1992). At those times it was claimed that predictions cannot be better than 65(± 2) % (Garnier, 1992).

In 1993, B. Rost, and C. Sander (Rost & Sander, 1993) presented the results of an in-depth analysis of the performance of multi-layered (neural) networks. By appropriately processing the information about structure contained in a multiple sequence alignment, it proves possible to increase the accuracy of secondary structure prediction above 70%.

Following decades brought new ideas. In his comprehensive review B. Rost (Rost, 2001) summarized the state of art at the beginning of 2000's. In his report there was at least five methods that pass the 75% correct classification limit. He concludes saying: 88% is a limit, but shall we ever reach close to there?

In this paper we check the validity of the basic hypothesis that the secondary, and three dimensional tertiary structure of a protein is uniquely determined by its sequence of amino acids, that is its primary structure.

The amount of variability in the secondary structure conformation of proteins at each residue suggests its relative importance and possible functions. Variability of outcomes at identical environments has also been a central concern in statistics. It would seem natural, then, to apply statistical methods to study structural variability in protein structures. In this paper, we undertake such an approach. We use the most classic field of statistical analysis that is regression to analyze secondary structures of a family of multiple protein structures (Zar, 2010; Ho, 2013). We assume that variations in protein structure can be represented by a statistical formulation.Our formulation can be solved using techniques from regression analysis to obtain a model with high generalization power.

2. FORMULATION OF THE PROBLEM

To estimate the conformation of the protein at a given residue, we consider 6 right and 6 left neighbors of this residue. Our hypothesis is that the conformation at the central residue is determined by these neighbors and by itself.

Primary structure: DETTALVCDNGSG
Secondary structure: CCCCCCSSSSSS

Figure 2 Primary and secondary structures of a protein of length 13 residues.

(a) Database

Primary structures of 200 proteins are obtained from the PDB website. Secondary structures of these proteins are obtained in the form of the x-ray crystallography analyses in three conformations helix "h", sheet "s" , and others ".". Others are interpreted as coils "c".

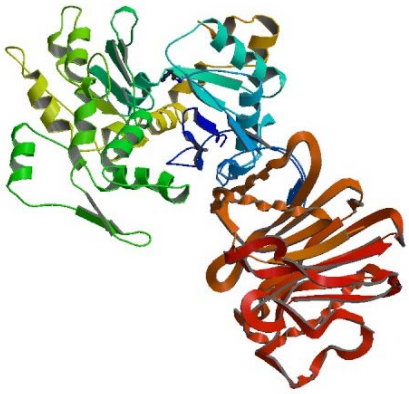

Figure 3 α-helices, β-sheets, and coils on the same picture

(b) Symbols for Amino Acids

Proteins are chains in the three dimensional space built from smaller chemical molecules called amino acids. There are 20 different amino acids. Each of them is denoted by a different letter in the Latin alphabet as shown below.

#	Amino acid	Chemical	alphabet
1	Alanine	Ala	A
2	Arginine	Arg	R
3	Asparagine	Asn	N
4	Aspartic acid	Asp	D
5	Cysteine	Cys	C
6	Glutamine	Gln	Q
7	Glutamic acid	Glu	E
8	Glycine	Gly	G
9	Histidine	His	H
10	Isoleucine	Ile	I
11	Leucine	Leu	L
12	Lysine	Lys	K
13	Methionine	Met	M
14	Phenylalanine	Phe	F
15	Proline	Pro	P
16	Serine	Ser	S
17	Threonine	Thr	T
18	Tryptophan	Trp	W
19	Tyrosine	Tyr	Y
20	Valine	Val	V

Table 1 Names and symbols of 20 amino acids

Based on the protein chain it is easy to create its relevant sequence of amino acids replacing an amino acid in chain by its code in Latin alphabet. As a result a word on the amino acids' alphabet is received. This word can be called a protein primary structure on the condition that letters in this word are in the same order as amino acids in the protein chain are.

A secondary structure of a protein is a subsequence of amino acids coming from the relevant protein. These sub-chains form in the three dimensional space regular structures which are the same in shape for different proteins. In the analysis, a similar representation for the secondary structures as for the primary ones has been used. A secondary structure is represented by a word on the relevant alphabet of secondary structures – each kind of a secondary structure has its own unique letter α-helix, H; β-sheet S, and coil C. An alphabet of secondary structures consisting of three different secondary structures has been considered in the analysis.

(c) Coding the Data

In this paper, data corresponding to an amino acid consists of 6 right, and 6 left neighboring amino acids of this amino acid in the primary chain of the protein as in Table 2. In the second row, secondary structure conformations of these neighboring amino acids are given.

A	E	E	K	E	A	**V**	L	G	L	W	G	K
H	H	H	H	H	E	**E**	E	E	C	C	C	E

Table 2 Six right, and six left neighboring amino acids of the amino acid V, and their conformations.

Secondary structure letters H, E, and C are coded as in the table below;

H	**E**	**C**
1	0	0
0	1	0
0	0	1

Table 3 Codes for secondary structure letters H, E, and C.

The data corresponding to an amino acid is coded by a 20×13 matrix:

	A	E	E	K	E	A	**V**	L	G	L	W	G	K
A	1	0	0	0	0	1	0	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	1
D	0	0	0	0	0	0	0	0	0	0	1	1	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	1	1	0	1	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1	0	0	1	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	1	0	1	0	0	0
K	0	0	0	1	0	0	0	0	0	0	0	0	1
M	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	1	0	0	0	0	0	0

Table 4 Code of the data corresponding to the central amino acid V.

3. THE MULTIPLE-REGRESSION EQUATION

A simple linear regression for a population of paired variables is the relationship

$$Y_i = \alpha + \beta X_i \quad (1)$$

In this relationship, Y_i and X_i represent the dependent and independent variables, respectively; β is the regression coefficient in the sampled population; and α, the Y intercept, is the predicted value of Y in the population when X is zero. And the subscript i in this equation indicates the i[th] pair of X and Y data in the sample.

In some situations, however, Y may be considered dependent upon more than one variable. Thus,

$$Y_i = \alpha + \beta_{11}X_{11i} + \beta_{12}X_{12i} + \cdots + \beta_{nm}X_{nmi} \quad (2)$$

or, more succinctly,

$$Y_i = \alpha + \sum_{k=1}^{n} \beta_{jk}X_{jki} \quad (3)$$

in the existence of *n* independent variables.

In the particular multiple regression model of this article, we have three sets of one dependent variable and 20×13 independent variables.

The population parameters $\beta_{11}, \beta_{12}, \dots, \beta_{nm}$ are termed partial regression coefficients because each expresses only part of the dependence relationship; β_{kj} expresses how much Y would change for a unit change in X_{kj}, if all other independent variables were held constant. It is sometimes said that β_{kj} is a measure of the relationship of Y to X_{kj} after controlling other independent variables; that is, it is a measure of the extent to which Y is related to X_{kj} after removing the effects of other independent variables. The Y intercept, α, is the value of Y when all $X_{11}, X_{12}, \dots, X_{nm}$ are zero.

A regression with $n \times m$ independent variables defines an $n \times m$ dimensional surface, sometimes referred to as a "response surface" or "hyperplane."

The population data whose relationship is described by Equation (2) will probably not all lie exactly on a plane, so this equation may be expressed as

$$Y_i = \alpha + \beta_{11}X_{11i} + \beta_{12}X_{12i} + \cdots + \beta_{nm}X_{nmi} + \epsilon_i \quad (4)$$

ϵ_i, the " residual," or " error," is the amount by which Y_i differs from what is predicted by $\alpha + \beta_{11}X_{11i} + \beta_{12}X_{12i} + \cdots + \beta_{nm}X_{nmi}$, where the sum of all ϵ_i's is zero, the ϵ_i's are assumed to be normally distributed.

If we sample the population containing the $n \times m + 1$variables $Y, X_{11}, X_{12}, \dots, X_{nm}$ in Equation (3), we can compute sample statistics to estimate the population parameters in the model.

The multiple-regression function derived from a sample of data would be

$$\hat{Y}_i = a + b_{11}X_{11i} + b_{12}X_{12i} + \cdots + b_{nm}X_{nmi} \quad (5)$$

The sample statistics $a, b_{11}, \dots, b_{nn}$ are estimates of the population parameters $\alpha, \beta_{11}, \beta_{12}, \dots, \beta_{nm}$, respectively, where each partial regression coefficient b_{ij} is the expected change in Y in the population for a change of one unit in X_{ij} if all of the other $n \times m - 1$ independent variables are held constant, and a is the expected population value of Y when each X_{ij} is zero.

Theoretically, in multiple-regression analyses there is no limit to $n \times m$, the number of independent variables (X_{ij}) that can be proposed as influencing the dependent variable (Y), as long as the size of the data $N \geq n \times m + 2$. At least $n + 2$ data points are required to perform a multiple regression analysis, where *n* is the number of independent variables determining each data point.

The criterion for defining the *"best fit"* multiple regression equation is most commonly that of least squares, which represents the regression equation with the minimum residual sum of N squares :

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \quad (6)$$

From Equation (4) the objective function to be minimized can be written as

$$F(a, b_{11}, \dots, b_{nm}) = \sum_{i=1}^{N} \big(Y_i - (a + b_{11}X_{11i} + \cdots + b_{nm}X_{nmi})\big)^2 \quad (7)$$

Minimum of this differentiable function is at the points where the gradient vanishes:

$$\frac{\partial F(a, b_{11}, \dots, b_{nm})}{\partial a} = 0, \quad (8)$$

$$\frac{\partial F(a, b_{11}, \dots, b_{nm})}{\partial b_{ij}} = 0, i = 1,2, \dots, n, j = 1,2, \dots, m.$$

which leads

$$Na + b_{11}\sum_{i=1}^{N} X_{11i} + \dots b_{nm}\sum_{i=1}^{N} X_{nmi} = \sum_{i=1}^{N} Y_i \quad (9)$$

$$a\sum_{i=1}^{N} X_{kji} + b_{11}\sum_{i=1}^{N} X_{kji}X_{1i} + \dots b_{nm}\sum_{i=1}^{N} X_{kji}X_{nmi} = \sum_{i=1}^{N} X_{kji}Y_i$$

$k = 1,2,\dots,n,\ j = 1,2,\dots,m.$

For $n \times m + 1$ unknowns $a, b_{11}, \dots, b_{nm}$, we have $n \times m + 1$ linear equations in (9). After flatting the matrix of unknowns to a vector with $n \times m + 1$ components, $\alpha, \beta_{11}, \beta_{12}, \dots, \beta_{nm}$, the coefficient matrix becomes

$$A = \begin{bmatrix} N & \sum_{i=1}^{N} X_{11i} & \cdots & \sum_{i=1}^{N} X_{nmi} \\ \sum_{i=1}^{N} X_{11i} & \sum_{i=1}^{N} X_{11i}X_{11i} & \cdots & \sum_{i=1}^{N} X_{11i}X_{nmi} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{N} X_{nmi} & \sum_{i=1}^{N} X_{nmi}X_{1i} & \cdots & \sum_{i=1}^{N} X_{nmi}X_{nmi} \end{bmatrix} \quad (10)$$

and the right hand side vector of the linear system of equations which has $n \times m + 1$ components is

$$c = \begin{bmatrix} \sum_{i=1}^{N} Y_i \\ \sum_{i=1}^{N} X_{11i}Y_i \\ \dots \\ \sum_{i=1}^{N} X_{nmi}Y_i \end{bmatrix} \quad (11)$$

3. IMPLEMENTATION OF THE MULTIPLE-REGRESSION MODEL

In our data we have three types of conformations, H, S, and C. Therefore we have three different dependent variables. Accordingly, we look for three different multivariate regression models for each of them. First dependent variable has the value 1 for H, and zero for S, and C, second dependent variable has the value 1 for S, and zero for H, and C, and the third has the value 1 for C, and zero for H, and S.

Training Data

First 150000 of the amino acid residues and corresponding secondary structures of around 170 proteins are concatenated to form a long string of amino acids. Then from this string 13-tuples are formed, and amino acids occurring in right and left neighborhoods of the central amino acid, together with the central amino acid are coded as shown in Table 4. These 20×13 matrices are the values of independent variables. The value of the dependent variable depends on the conformation of the central amino acid. For this data, the matrix A in (10), which is the same for all three models is computed. The right hand side vector c in (11) depends on the values of the dependent variable, hence three right hand side vectors are computed for three models. For each model, solving the system $[A|c]$ the three sets of regression coefficients $a, b_{11}, \dots, b_{nm}$ are found.

Testing Data

Using the remaining 19026 residues of the concatenated proteins, the testing data is coded and prepared as in for the training data. Each testing data is sent to the three models and the value of the dependent variable is computed. The model that produces the largest output, determines the conformation of the central amino acid of the data considered. Then the prediction of the regression model and the true conformations are compared to find the confusion matrix, and success in the estimation of the conformations of the testing data as helix, sheet, and coil. Correct classification rates of the training and testing data are given in Table 5.

	Training %	Testing %
Regression Analysis	58.84	62.50

Table 5 Correct classification rates on the training and testing data

4. RESULTS AND DISCUSSION

To compare the robustness of the system with the ones that exist as free excess tools in the web, we have chosen 14 additional proteins from NCBI Protein database with their secondary structure estimates through x-ray analysis. The secondary structures of these proteins are obtained through the tools given in Chou-Fasman website[1]. Experiment is made using Chou-Fasman (C-F), and Neural Network (ANN) estimates. Comparison of the regression results of this paper and results from these experiments are seen in Table 6.

	PDB Codes	CF	NN	RA
1	2W6K	0.48	0.59	0.57
2	2V4Y	0.51	0.67	0.67
3	3BHJ	0.52	0.59	0.62
4	3BBU	0.68	0.60	0.62
5	3BH8	0.47	0.50	0.52
6	3BL9	0.40	0.48	0.48
7	2ZPE	0.52	0.71	0.66
8	3BGM	0.43	0.71	0.72
9	3CR3	0.49	0.72	0.68
10	2Q8S	0.53	0.65	0.70
	Average	0.50	0.61	0.62

Table 6 Correctness of the estimates for the secondary structure of three experiments using Chou-Fasman, Neural Network, and regression model.

[1] http://cib.cf.ocha.ac.jp/bitool/MIX/

These results show that regression analysis which relies on a database of 200 proteins has a estimation power that is comparable with the famous online estimation tools.

REFERENCES

Anfinsen, C. B., Epstein, C. J. & Goldberger, R. F. (1963). The genetic control of tertiary protein structure: studies with model systems. Cold Spring Harbor Symp. Quant. Biol. 28, 439-449.

Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. Nucl. Acids Res. 20, 2019-2022.

Barton, G. J., Newman, R. H., Freemont, P. S. & Crumpton, M. J. (1991). Amino acid sequence analysis of the annexin super-gene family of proteins. Eur. J. Biochem. 198, 749-760.

Benner, S. A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Advan. En,. Reg. 31, 121-181.

Benner, S. A .. Cohen, M. A. & Gerloff, D. (1992). Correct structure prediction? Nature (London), 359, 781.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macro-molecular structures. J. Mol. Biol. 112, 535- 542.

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. J.,Lautrup, B., Nerskov, L., Olsen, 0. H. & Petersen, 8. B. (1988). Protein secondary structure and homology by neural networks. FE BS Letters, 241, 223-228.

Bohr, H., Bohr, J., Brunak, S., Fredholm, H., Lautrup, B. & Petersen, S. B. (1990). A novel approach to prediction of the 3dimensional structures of protein backbones by neural networks. FEES Letters, 261, 43-46.

Bork, P., Ouzonis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). What's in a genome? Nature (London) , 358, 287.

Bossa, F. & Pascarella, S. (1990). PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. CABI OS, 5 , 319-320.

Chou, P. Y. & Fasman, U. D. (1974). Prediction of protein conformation. Biochemistry, 13, 211-215.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R., J. (1983). Secondary structure assignment for α/β proteins by a combinatorial approach. Biochemistry, 22, 4894-4904.

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.

Ewbank, J. ,T. & Creighton, T. E. (1992). Protein folding by stages. Curr. Opin. Struct. Biol. 2, 347-349.

Fasman, G. F. (1989). Protein conformation prediction. In Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G. D., ed.), pp. 193-316, Plenum, New York and London.

Frampton, J., Leutz, A., Gibson, T. J. & Graf, T. (1989). DNA-binding domain ancestry. Nature (London), 342. 134.

Garnier, J. (1993). Prediction of protein structure. In Biological Sequences: Finding Structure and Function by Neural Networks (Brunak, S., ed.), Institute for Scientific Interchange Foundation, Torino, Italy, J. Mol. Biol. 232,584-599.

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.

Garrett, R. C., Thornton, J. M. & Taylor, W. R. (1991). An extension of secondary structure prediction towards the prediction of the tertiary structure. PE BS Letters, 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. M al. Biol. 198, 425-443.

Gibson, T. J., Thompson, J. D. & Abagyan, R. A. (1993). Proposed structure for the DNA-binding domain of the helix-loop-helix family of eukaryotic gene regulatory proteins. Protein Eng. 6, 41-50.

Hirst, J. D. & Sternberg, M. J. E. (1992). Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. Biochemistry, 31, 615-623.

Ho, R. (2013) Handbook of Univariate and Multivariate Data Analysis with IBM SPSS, Chapman and Hall/CRC.

Holley, H. L. & Karplus, M. (1989). Protein secondary structure prediction •with a neural network. Proc. Nat. Acad . Sci., U.S.A. 86, 152-156.

Hubbard, T. J. P. & Sander, C. (1991). The role of heat-shock and chaperone proteins in protein folding: possible molecular mechanisms. Protein Eng. 4, 711-717.

Kanehisa, M. (1988). A multivariate analysis method for discriminating protein secondary structural segments. Protein Eng. 2, 87-92.

Karplus, M. & Petsko, G. A. (1990). Molecular dynamics simulations in biology. Nature (London), 347, 631-639.

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.

Levin, J. M. & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. Biochim. Biophys. Acta , 955, 283-295.

Levin, J. M., Robson, B. & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Letters, 205, 303-308.

Lim, V. T. (1974). Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J . Mol. Biol. 88, 857-872.

Maclin, R. & Shavlik, J. W. (1993). Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. Machine Learning, Multistrategy Learning, 87-107.

Maxfield, F. R. & Scheraga, H. A. (1979). Improvements in the prediction of protein topography by reduction of statistical errors. Biochemistry, 18, 697-704.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in poly-peptides. ,J . Phys. Chem. 79, 2361-2381.

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

Musacchio, A., Gibson, T., Lehto, V.-P. & Saraste, M. (1992). SH3 - an abundant protein domain in search of a function. FE BS Letters, 307, 55-61.

Nagano, K. (1973). Logical analysis of the mechanism of protein folding. J. Mol. Biol. 15, 401-420.

Nagano, K. & Hasegawa, K. (1975). Logical analysis of the mechanism of protein folding. J . Mol. Biol. 94, 257-281.

Niermann, '1\ & Kirschner, K. (1991). Improv ing the prediction of secondary structure of "TIM-barrel" enzymes (Corrigendum). Protein Eng. 4, 359-370.

Ouzounis, C. A. & Melvin, W. T. (1991). Primary and secondary structural patterns in eukaryotic cyto- chrome P-450 families correspond to structures of the helix-rich domain of Pseudomonas putida cytochrome P-450cam. Eur. J. Biochem. 198, 307-315.

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

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

Pauling, L., Corey, R. R. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Nat. Acad. Sci., U.S.A . 37, 205.

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

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

Ptitsyn, 0. 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, 0. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. Biopolymers, 22, 15-25.

Qian. N. & Sejnowski, T., 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. (1991). Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. Proteins: Struct. Funct. Genet. 9, 69-78.

Rooman, M. J., Wodak, S. & Thornton, ,J. M. (1989). Amino acid sequence tem plates derived from recur- rent turn motifs in proteins: critical evaluation of their predictive power. Protein Eng. 3, 23-27

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. Nature (London), 360, 540.

Rost, B. & Sander, C. (1993). Prediction of Protein Secondary Structure at Better than 70% Accuracy. J . Mol. Biol. 232, 584-599

Rost, B., (2001). Review: Protein Secondary Structure Prediction Continues to Rise, Journal of Structural Biology 134, 204–218.

Russell, R. B., Breed, J. & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. FEBS Letters, 304, 15-20.

Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. Proteins: Struct. Funct. Genet. 9, 56-68.

Schulz, G. E. & 8chirmer, R. H. (1979). Principles of Protein Structure, Springer, New 'York.

Sternberg, M. J. E. & K ing, R. D. (1990). Machine learning approach for the prediction of protein secondary structure. J. Mol. Biol. 216. 441-457.

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

Szent-Györgyi, A. G. & Cohen, C. (1957). Role of proline in polypeptide chain configuration of proteins. Science, 126. 697.

Taylor, W. R.. & Orengo, C. A. (1989). A holistic approach to protein structure alignment. Protein Eng. 2, 505-519.

Taylor, W. R. & Thornton, J. M. (1988). Prediction of super-secondary structure in proteins. Nature (London), 301. 540-542.

Vriend, G. & Sander, C. {1991). Detection of common three-dimensional substructures in proteins. Proteins: Struct. Funct. Genet. 11, 52-58.

Zar, J. H. (2010) Biostatistical Analysis, Prentice Hall, N. Jersey USA.

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

Zhong, L., Johnson, W. & Curtis, J. (1992). Environment affects amino acid preference for secondary structure. Proc. Nat. Acad. Sci., U .S .A. 89, 4462--4465.

Zvelebil. M. J., Barton, G. J., Taylor, W. R. & Sternberg, 1. J. E. (1987). Prediction of protein secondary structure and active sites using alignment of homologous sequences. J. Mol. Biol. 195, 957-961.