VOL4 NO. 1 March 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

IUS Soft Computing Research Group

Protein Secondary Structure Prediction Based on Physicochemical Features and PSSM by KNN

Faruk Berat Akcesme
International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
fakcesme@ius.edu.ba

Article Info

Article history:
Article received on February 2015
Received in revised from March 2015

Keywords:
Protein secondary structure prediction; PSSM profiles; physicochemical features; PDB25 dataset

Abstract

In this paper, we propose a protein secondary structure prediction method based on the k-nearest neighborhood (KNN) technique with position-specific scoring matrix (PSSM) profiles, propensity matrix of amino acids in three conformations (HEC) and three physicochemical features; hydrophobicity, net charges, and side chain mass. First, the KNN with the optimal k-value is found. Then, the Euclidean distance of 26-dimensional data for each amino acid of a protein, to the data vectors of all other proteins are computed. The conformations of the nearest seven amino acids are pooled. Majority of the pooled votes is given to the amino acid of the quarry protein as the conformation H, E, or C. Finally, we use a filter to refine the predicted results from KNN. After filtering, the accuracy of the prediction goes up to the level of 90% for some proteins. This validates that considering PSSM, the propensity matrix, and physicochemical features may exhibit better performance.

1. INTRODUCTION

Although it is probably true that one can infer protein properties by given protein primary structure, current state of the art approaches are not able to implement this in practice. There is many different approaches and algorithms which are designed to predict the secondary structure of protein from it's know primary sequence but no algorithm can predict with desirable accuracy. In this paper protein secondary structure are investigated based on protein primary structure and its physicochemical properties.

A protein primary sequence is composed of 20 different kinds of amino acids. Each of them is denoted by a different letter in the Latin alphabet as shown below.

In this paper protein secondary structure are investigated based on protein primary structure and its physicochemical properties.Due to the differences of their side chain sizes, shapes, reactivity, and the ability to form hydrogen bonds, the secondary structure of a protein sequence comes from different folding of amino acids into helices, sheets and coils (Chou, and Fasman, 1978; Garnier et. al., 1978). Furthermore, owing to the differences of the side chain sizes, the number of electric charges, coupled with the affinity for water, the tertiary structures of protein sequences are not all the same, even their primary sequences are similar. Thus, the exploration of molecular

structures on protein sequences is divided into primary, secondary, tertiary, and even quaternary structures (Huang, and Chen, 2013).

Table 1 Names and symbols of 20 amino acids

#	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

Through x-ray analysis, given a protein primary sequence, its corresponding secondary structure may be obtained as follows.

Primary sequence:

MKRESHKHAEQARRNRLAVALHELASLIPAEWKQQ
NVSAAPSKATTVEAACRYIRHLQQNGST

Secondary structure:

CTHHHHHHHHHHHHIIIHHHHBBBBHTTEEESSGGGT
SSSCCSSSHHHHHHH---HHHHTEECC

Eight secondary structure types appear in a secondary sequence; H(α-helix), G(310-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend), and -(rest). The eight structure classes are usually reduced to three classes of helix (H), sheet (E), and coil (C). In this paper the reduction is performed as follows.
1) H, G and I to H;
2) E to E;
3) The rest to C

2. FEATURE EXTRACTION

Five relevant kinds of features are extracted from protein sequences to predict protein secondary structure; i.e., 1) conformation parameters,
2) Position specific scoring matrix (PSSM) profiles,
3) Net charge,
4) Hydrophobic, and
5) Side chain mass.

2.1 Extracting Primary and Secondary Sequences:

Amino acid primary and secondary structure was extracted from the PDB website

(http://www.rcsb.org/pdb/home/home.do) using the PDB codes of 25PDB. Then, we can further extract five different features from amino acid sequences as follows.

2.2 Propensity matrix:

Intrinsic properties of amino acids enable us to figure out their tendency for being in certain conformation. The main idea of using propensity table is to get benefits from amino acid properties and find out statistically significant contribution to prediction capacity. In general, protein secondary structure is divided into three types: α-helix (H), β-sheet (E), and coil (C), so that there are three values for each amino acid. In the feature extraction, all the conformation parameters are calculated from a data set. The conformation parameters for each amino acid S_{ij} are defined as follows:

$$S_{ij} = \frac{a_{ij}}{a_i}, i = 1,2,..,30; j = 1,2,3. \quad (1)$$

In this formula, i indicates the 20 amino acids, and j indicates the 3 types of secondary structure: H, E, and C. Here, a_i is the amount of the ith amino acid in a data set whereas a_{ij} is the amount of the ith amino acids with the jth secondary structure.

Table 2. Conformation parameters for each amino acid in a data set of 20.289 proteins

"o"	"H"	"E"	"C"
"A"	0.51	0.16	0.33
"R"	0.44	0.19	0.37
"N"	0.29	0.12	0.58
"D"	0.33	0.11	0.56
"C"	0.32	0.28	0.40
"Q"	0.47	0.15	0.38
"H"	0.30	0.19	0.50
"G"	0.17	0.13	0.70
"E"	0.50	0.14	0.37
"I"	0.39	0.36	0.25
"L"	0.49	0.23	0.28
"K"	0.42	0.16	0.42
"M"	0.43	0.20	0.37
"F"	0.38	0.30	0.32
"P"	0.19	0.09	0.72
"S"	0.29	0.16	0.54
"T"	0.28	0.25	0.47
"W"	0.41	0.27	0.32
"Y"	0.37	0.29	0.33
"V"	0.33	0.40	0.28

The conformation parameters for each amino acid in a data set of 20347 proteins are shown in Table 2. The reason of using conformation parameters as features is that the folding of each residue has something to do with forming a specific structure.

2.3 PSSM Profiles

PSSM profiles are generated by PSIBLAST (Position Specific Iterative-Basic Local Alignment Search Tool) program (Alteschul et al., 1997). Since PSSM profiles are involved with biological evolution, we consider them as features in our work. A PSSM profile has L×20 elements, where L is the length of a query sequence. These profiles are then used as the input features to feed an SVM, employing a sliding window method.

The position weight matrix was introduced by American geneticist Gary Stormo and colleagues in 1982 (Gary.S et all, 1982). PSSM has found good alternative to consensus sequence. Consensus sequences had previously been used to represent patterns in biological sequences, but had difficulties in the prediction of new occurrences of these patterns. First, a database containing all known sequences (or non-redundant database) is selected. Then, low complexity regions are removed from the nr database. Finally, PSI-BLAST program is used to query each sequence in 25PDB, and generates PSSM profiles after three iterations. Here, multiple sequence alignment (MSA) and BLOSUM62 matrix (Henikoff, and Henikoff, 1992) are used in this process.

2.4 Net Charges

One of the physical properties of amino acids is their charges. Five of the amino acids are charged amino acids: R, D, E, H, and K. Residues which have similar electric charge repel each other and it interrupts the hydrogen bonds in the main chain of amino acids. It prevents the formation of α-helix. In addition, continues β-sheet formation are not possible when the residues have similar charges. This physical property of amino acids helps to predict secondary structure of proteins. Net charge of each amino acid can be obtained from from Amino Acid index database (Kawashima, et. al, 1999; Kawashima, and Kanehisa, 2000; Kawashima, et. al, 2008; Nakai, et. al., 1998; Tomii, and Kanehisa, 1996), as shown in Table 3.

Table 3. Net charge of amino acids

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

2.5 Hydrophobicity

Some of the amino acids do not like to reside in an aqueous environment and they called hydrophobic amino acids. They are generally seen buried within the hydrophobic core of protein since for protein folding, polar residues prefer to stay outside of protein in order to prevent non polar residues from exposing to polar solvent. Hydrophobic protein can be used as one of the parameter to predict the secondary structure of proteins. In α-helix, generally hydrophobic segments are followed by hydrophilic segment. Unlike α-helix, β-sheet structure is affected by the environment due to its structural characteristics so it is not a case in β-sheets. The hydrophobic values of amino acids can also be obtained from Amino Acid index database (or AAindex) as shown in Table 4. Positive values indicated more hyrophobicity.

Table 4. Hydrophobic values of amino acids

Aacids	Hydphb	Aacids	Hydphb
A	1.8	L	3.8
R	-4.5	K	-3.9
N	-3.5	M	1.9
D	-3.5	F	2.8
C	2.5	P	-1.6
E	-3.5	S	-0.8
Q	-3.5	T	-0.7
G	-0.4	W	-0.9
H	-3.2	Y	-1.3
I	4.5	V	4.2

2.6 Side Chain Mass

Although the basic structure as shown in Fig. 3 is the same for 20 amino acids, the size of the side chain R group still influences structure folding. Side chains of amino acids are the structural elements which make amino acids different. These unique R groups influencing the conformation of protein secondary structure and they can give a clue to predict the secondary structural element depends on their existence in certain position. The site chain R group form in the outside of the main chain of α-helix structure but when large R groups distributed continuously, they can make α-helix structure unstable. For instance, proline is composed of 5 atoms in a ring, which is difficult to form hydrogen bonds. In addition, generally it is observed that R group of B-sheet structure is smaller than those of other structure. Side chain mass is considered one of the important features that can contribute to predict secondary structure of proteins.

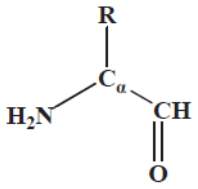

Figure 1. Basic structure of amino acids.

Table 5. Side chain mass of amino acids

Aacids	mass	Aacids	mass
A	15.0347	L	57.1151
R	100.1431	K	72.1297
N	58.0597	M	75.1483
D	59.0445	F	91.1323
C	47.0947	P	41.0725
E	73.0713	S	31.0341
Q	72.0865	T	45.0609
G	1.0079	W	130.1689
H	81.0969	Y	107.1317
I	57.1151	V	43.0883

3. K-NEAREST NEIGHBOR TECHNIQUE (KNN)

The KNN used in the experiments is a classifier for predicting the secondary structure H, E, and C. Three-fold cross-validation is employed on the 25PDB data set to find the optimal neighbor number k.

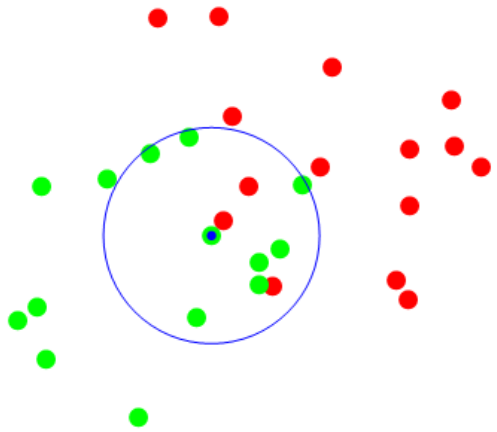

Figure 2. Majority votes of seven nearest neighbors of the point in the center of the circle is green.

Here, the distance of the data vectors are first measured by Euclidean distance. Then other distance measures are also used for comparison.

Filter

It is not possible for amino acid to form α-helix or β-sheet alone. Incorrect predicted results should be eliminated by replacement with reasonable conformation if single conformation exists in the predicted sequences. For the current scanning window (i-1, i, i+1) in the predicted secondary structure, two possible structures could happen at position i:

Case H: if str(i-1) and str(i+1) are H, then str(i) is not changed; otherwise, extend the examined segment to (i-3, i-2, i- 1, i, i+1, i+2, i+3) and replace str(i) with the majority structure in the examined segment.

Case E: if str(i-1) or str(i+1) is E, then str(i) is not changed; otherwise, extend the examined segment to (i-3, i-2, i- 1, i, i+1, i+2, i+3) and replace str(i) with the majority structure in the examined segment.

4. EXPERIMENTS

A. *Data Set*

Many different dataset are used for predicting secondary structure of proteins, such as RS126 (Rost, and Sander, 1993), CB513 (Cuff, and Barton, 1999), CASP (Moult, et. al., 1995), EVA (Eyrich, et. al., 2001). The 25PDB dataset selected for our studies the similarity between sequences of 25PDB is less than 25%. 25 PDB designed for predicting protein classes but it is found useful for predicting the secondary structure of protein since similarity is very small, this let us to predict secondary structure of protein more accurately.

25PDB contain 1674 amino acid sequences and it can be downloaded from

http://biomine.ece.ualberta.ca/SCPRED/SCPRED.htm

B. *Performance Measures*

Two kinds of performance measures are frequently used in protein secondary structure prediction; i.e., Q3 or three-state overall per-residue accuracy. Q3 is a residue based measure of three-structure overall percent-age of correctly predicted residues, which can be represented as Formula (2).

$$Q_3 = \frac{N_H+N_E+N_C}{N} \quad (2)$$

where N is the total number of predicted residues, N_H is the correctly classified secondary structure for helix, N_E for sheet, and N_C for coil.

C. *Experimental Results*

In this section, first we expose the accuracy in secondary structure prediction by charts which shows the frequencies of proteins at each accuracy level. It is seen that in all-α, and all-β protein classes up to 90% accuracy is achieved (Figure 3., and Figure 4.). For α+β, and α/β protein classes this drops up to 80% accuracy (Figure 5., and Figure 6.).

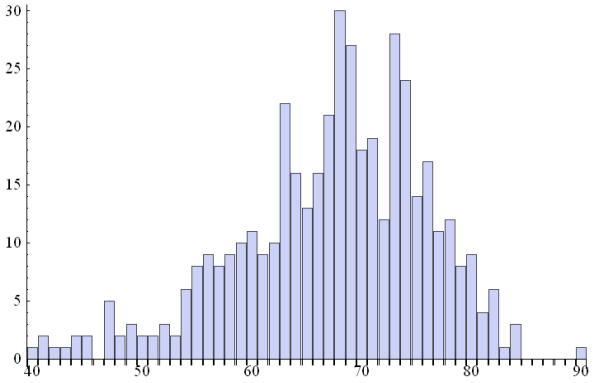

Figure 3. In all-α protein classes up to 89% secondary structure prediction accuracy is achieved

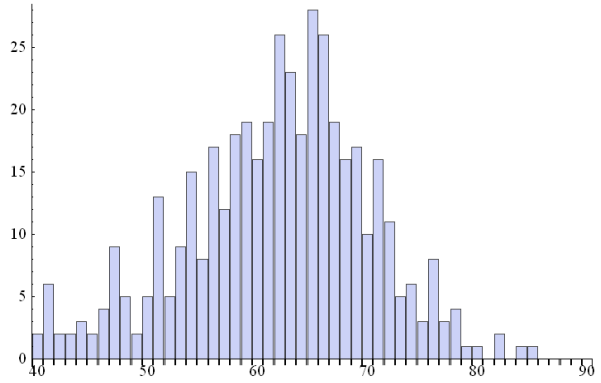

Figure 4. In all-β protein classes up to 86% secondary structure prediction accuracy is achieved

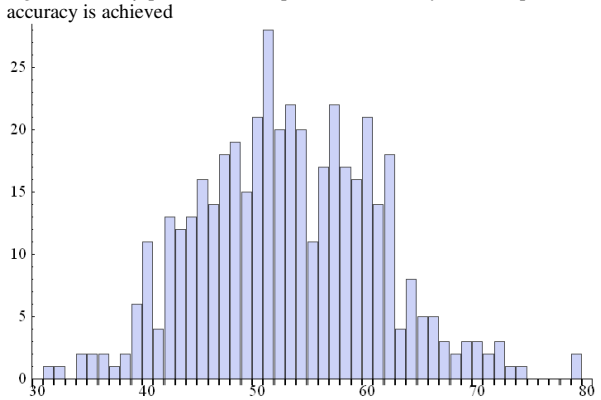

Figure 5. In α+β protein classes up to 80% secondary structure prediction accuracy is achieved

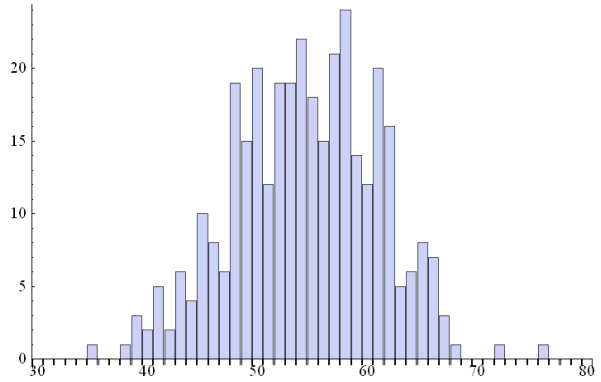

Figure 6. In α/β protein classes up to 77% secondary structure prediction accuracy is achieved

D. Filtering Effect

For all-α protein class, filtering of outputs improved the mean accuracy from 65.73% to 67.09%. The frequencies of the percentage increases in accuracy are shown in Figure 7 below.

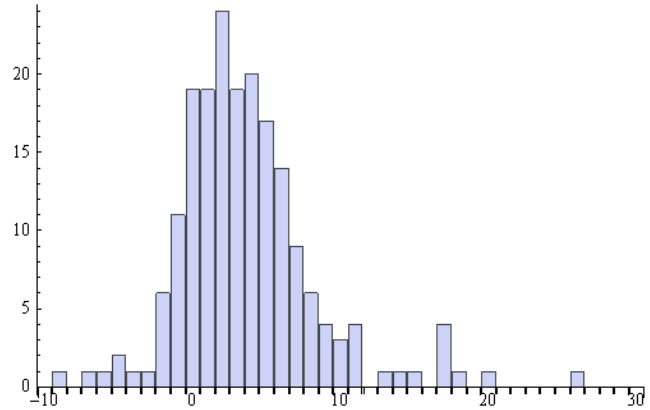

Figure 7. The effect of filtering for all-α protein class. Frequencies of the percentage increases in accuracy is on the vertical axis.

For all-β protein class, filtering of outputs improved the mean accuracy from 59.60% to 61.60%. The frequencies of the percentage increases in accuracy are shown in Figure 8 below.

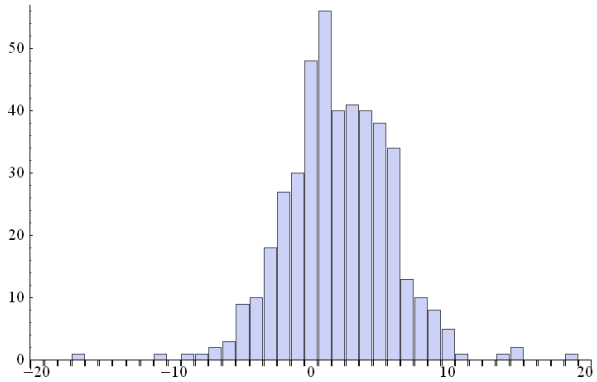

Figure 8. The effect of filtering for all-β protein class. Frequencies of the percentage increases in accuracy is on the vertical axis.

For α+β protein class, filtering of outputs improved the mean accuracy from 49.97% to 52.80%. The frequencies of the percentage increases in accuracy are shown in Figure 9 below.

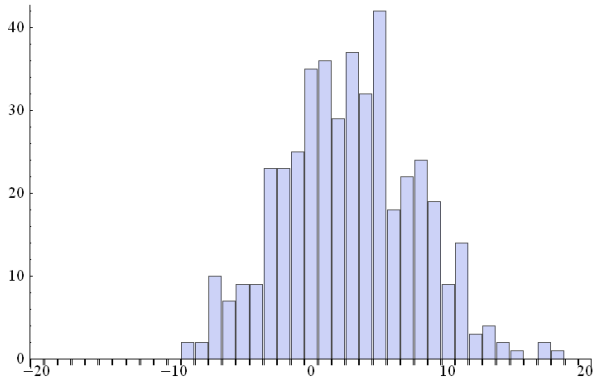

Figure 9. The effect of filtering for α+β protein class. Frequencies of the percentage increases in accuracy is on the vertical axis.

For α/β protein class, filtering of outputs improved the mean accuracy from 52.07% to 54.36%. The frequencies of the percentage increases in accuracy are shown in Figure 10 below.

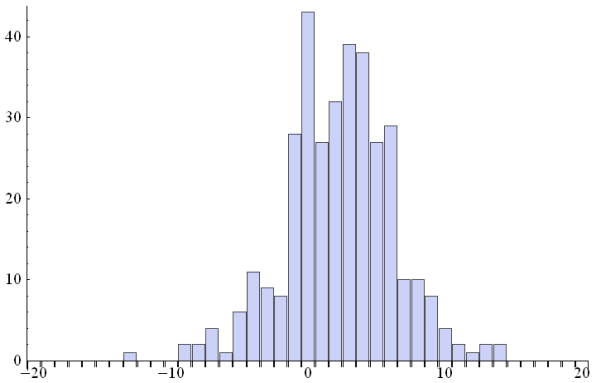

Figure 10. The effect of filtering for α+β protein class. Frequencies of the percentage increases in accuracy is on the vertical axis.

Average accuracies without the filter and with the filter are given in Table 6.

Table 6. Average accuracies without the filter and with the filter, and improvements due to filtering.

	All-α	All-β	α+β	α/β
Filtered	67.09	61.60	52.80	54.36
Not Filtered	65.73	59.60	49.97	52.07
Improvement	+1.36	+2.00	+2.83	+2.29

5. CONCLUSIONS

In this paper, we propose a protein secondary structure prediction method using PSSM profiles and four physicochemical features, including conformation parameters, net charges, hydrophobic, and side chain mass. In the experiments, the KNN with the optimal neighbor size k found first. Then, the majority of the conformations of the k neighbors of a given amino acid in a certain class is given to this amino acid as secondary structure.

Finally, we use the filter to refine the predicted results from the KNN. Although the tool KNN is the simplest one of all methods, we succeeded accuracy in secondary structure prediction of proteins up to 90% for the 25PDB data set. In summary, considering these physicochemical features and PSSM matrix, results in better performances.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, Vol. 25, No. 17, pp. 3389-3402.

Chou, P. Y., and Fasman, G. D. (1978) "Empirical predictions of protein conformation," Annual Review Biochemistry, Vol. 47, pp. 251-276.

Cuff, J. A., and Barton, G. J. (1999) "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction," Proteins: Structure, Function, and Bioinformatics, Vol. 34, No. 4, pp. 508-519.

Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2001) "EVA: continuous automatic evaluation of protein structure prediction servers," Bioinformatics, Vol. 17, No. 12, pp. 1242-1243.

Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) "Analysis and implications of simple methods for predicting the secondary structure of globular proteins," Journal of Molecular Biology, Vol. 120, No. 1, pp. 97-120.

Henikoff, S., and Henikoff, J. G. (1992) "Amino acid substitution matrices from protein blocks," Proceedings of the National Academy of Sciences U.S.A., Vol. 89, No. 22, pp. 10915-10919.

Huang, Y. F., and Chen, S. Y. (2013) Protein Secondary Structure Prediction Based on Physicochemical Features and PSSM by SVM, 978-1-4673-5875-0/13 IEEE

Kawashima, S., Ogata, H., and Kanehisa, M. (1999) "AAindex: amino acid index database," Nucleic Acids Research, Vol. 27, No. 1, pp. 368-369.

Kawashima, S., and Kanehisa, M. (2000) "AAindex: amino acid index database," Nucleic Acids Research, Vol. 28, No. 1, pp. 374.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008) "AAindex: amino acid index database, progress report 2008," Nucleic Acids Research, Vol. 36, No. 1, pp. D202-D205.

Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995) "A large-scale experiment to assess protein structure prediction methods," Proteins: Structure, Function, and Bioinformatics, Vol. 23, No. 3, pp. ii- iv.

Nakai, K., Kidera, A., and Kanehisa, M. (1988) "Cluster analysis of amino acid indices for prediction of protein structure and function," Protein Engineering Design and Selection, Vol. 2, No. 2, pp. 93-100.

Rost, B., and Sander, C. (1993) "Prediction of secondary structure at better than 70% accuracy," Journal of Molecular Biology, Vol. 232, No. 2, pp. 584-599.

Stormo, Gary D.; Schneider, Thomas D.; Gold, Larry; Ehrenfeucht, Andrzej (1982). "Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli". Nucleic Acids Research 10 (9): 2997–3011.

Tomii, K., and Kanehisa, M. (1996) "Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins," Protein Engineering Design and Selection, Vol. 9, No. 1, pp. 27-36.