Prediction of Protein Secondary-Structure by Monte Carlo Simulation

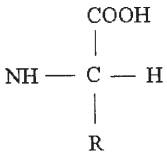
ALEXANDRU WOINAROSCHY*
Politehnica University of Bucharest, Dept. of Chemical Engineering, str. Polizu 1-5, 011061, Bucharest, Romania

Proteins have four structural categories. The primary structure is the amino-acid sequence of the polypeptide chain. The secondary structure is the conformation, representing of the backbone (α-helix or β-sheet). The knowledge of protein structure has a paramount theoretical and practical importance (e.g. cancer disease) and a huge effort of research was devoted to this subject. Despite the fact that several methods were developed for protein secondary-structure prediction, there are no consensuses of their results. In this paper was proposed an new, original, method to investigate the influence of the number of amino acids and the percentage contents in the twenty amino acids for the prediction of protein secondary-structure, respectively Monte Carlo simulation using a multilayer neural networks. The method is very promising in connection with the use of large data bases.

Keywords: protein secondary-structure, Monte Carlo simulation, neural networks

The knowledge of protein structure has a paramount theoretical and practical importance (e.g. cancer disease) and a huge effort of research was devoted to this subject.

Proteins are complex polymers composed of a series of amino acids attached by peptide bonds. There are 20 different amino acids present in proteins, each with a different residual group R:



The properties of the residual groups, in conjunction with their structural positions, define the solution properties of the protein. The structures of the twenty amino acids [1] are presented in figure 1.

Amino acids fall into five categories: aliphatic, nonpolar, aromatic, polar and charged. Table 1 lists the twenty amino acids in order of ascending hydrophobicity, measured on the basis of solubility in various solvents [2].

Proteins have four structural categories [2]. The primary structure is the amino-acid sequence of the polypeptide chain. The secondary structure is the conformation, representing of the backbone (α-helix or β-sheet). The tertiary structure is the three-dimensional conformation, representing how the secondary structure folds to obtain the most favorable thermodynamic state, with hydrophobic residues on the interior and hydrophilic residues on the exterior. The quaternary structure is the arrangement of the aggregation of several polypeptide chains.

The α-helix is a tight coil with 3.6 amino acids per turn, stabilized by hydrogen bonds between the NH and CO

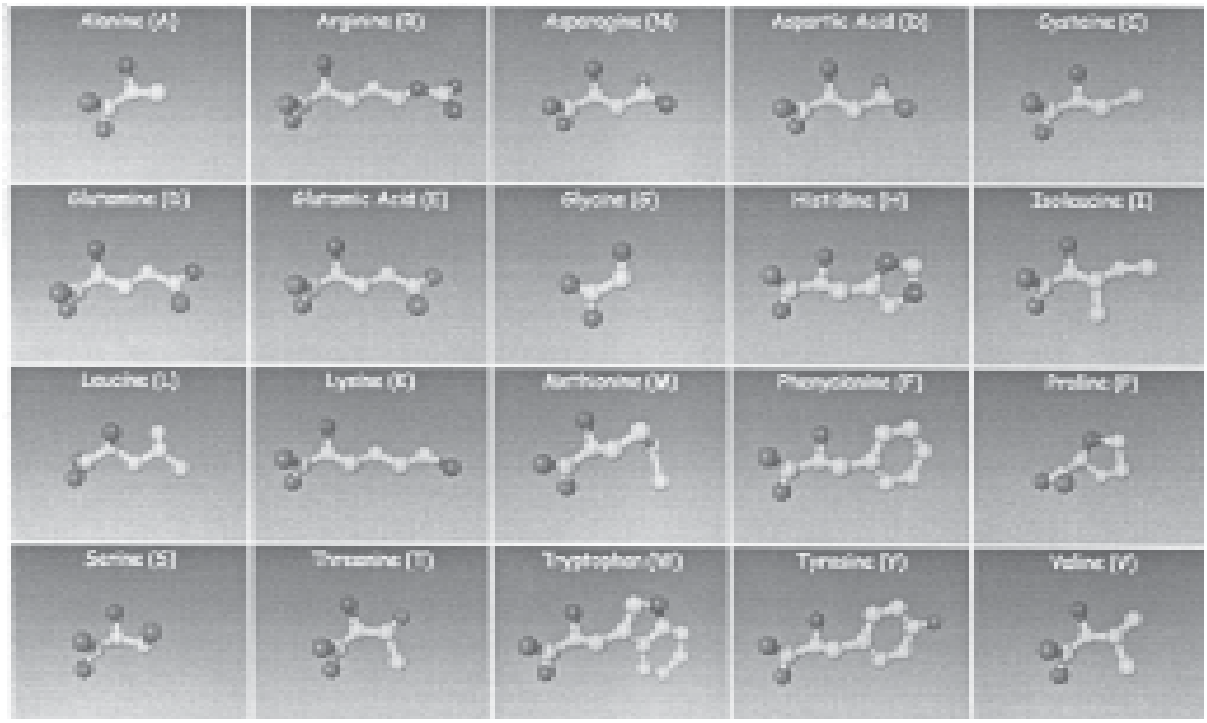


Fig. 1. The structure of the twenty amino acids from proteins [1]

* email: a_woinaroschy@chim.upb.ro

Table 1
THE AMINO ACIDS CHARACTERISTICS

Amino acid	Free energy (kcal/mole)	Residue category	Amino acid	Free energy (kcal/mole)	Residue category
Phenylalanine	3.7	aromatic	Serine	0.6	polar
Methonine	3.4	nonpolar	Proline	-0.2	nonpolar
Isoleucine	3.1	aliphatic	Tyrosine	-0.7	aromatic
Leucine	2.8	aliphatic	Histidine	-3.0	aromatic
Valine	2.6	aliphatic	Glutamine	-4.1	polar
Cysteine	2.0	nonpolar	Asparagine	-4.8	polar
Tryptophan	1.9	aromatic	Glutamic acid	-8.2	charged (-)
Alanine	1.6	aliphatic	Lysine	-8.8	Charged (+)
Threonine	1.2	polar	Aspartic acid	-9.2	Charged (-)
Glycine	1.0	nonpolar	Arginine	-12.3	Charged (+)

groups. All hydrogen bounds in the α-helix point the same direction, with the NH groups towards the N-terminal side and the CO groups towards C-terminal side. Polarity differences between the CO and NH groups cause an overall dipole moment along the helical axis. The helix structure occurs most often at the protein's surface, when on side of the α-helix is hydrophobic and the other hydrophilic [3]. In figure 2 is presented [4] a side view of an α-helix of alanine residues in atomic detail. The protein chain runs upwards, i.e., its N-terminus is at the bottom and its C-terminus at the top of the figure.

The β-sheet conformation contains amino acid strands, approximately 5 to 10 units in length, and aligned side by side [3]. The structure exists in either parallel or antiparallel form. In parallel form, all strands run in the same direction, while in the antiparallel form, the strands alternate. Unlike the α-helix, the hydrogen bounds between the NH and CO groups alternate directions in both β-sheet forms. Therefore, no hydrophobic or hydrophilic side exists at the protein surface. In figures 3.a and 3.b are presented [4] the β-sheet secondary structures of proteins. The hydrogen bonding patterns are represented by dotted lines, in an antiparallel β-sheet (fig. 3.a) and in parallel β-sheet (fig. 2.b).

Usual methods for protein secondary-structure prediction [4]

Early methods of secondary-structure prediction were based on the helix- or sheet-forming propensities of individual amino acids, sometimes coupled with rules for estimating the free energy of forming secondary structure elements. Such methods are typically ~60% accurate in

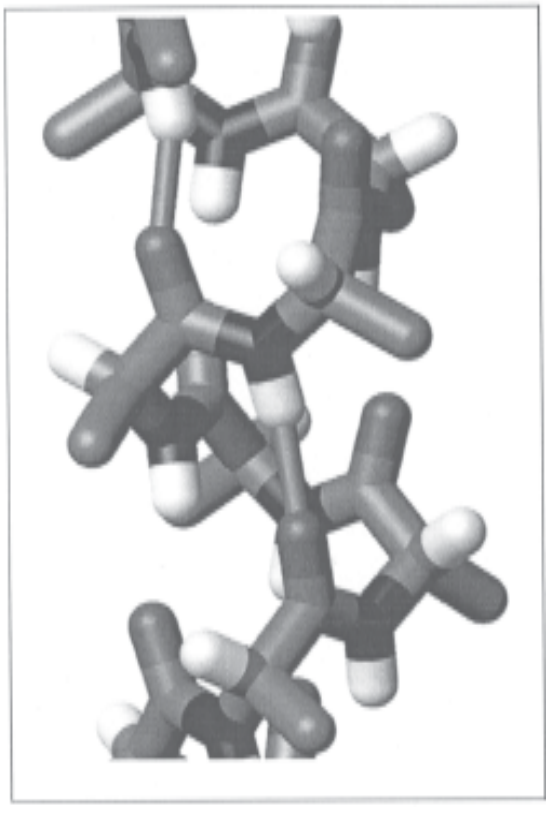


Fig. 2. A α-helix of alanine residues

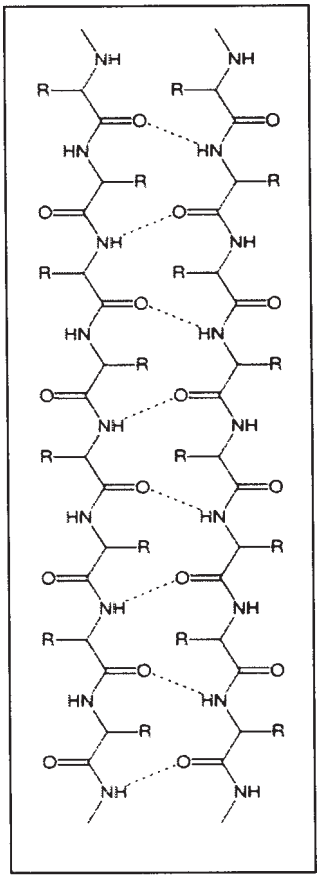


Fig. 3a. Antiparallel β-sheet

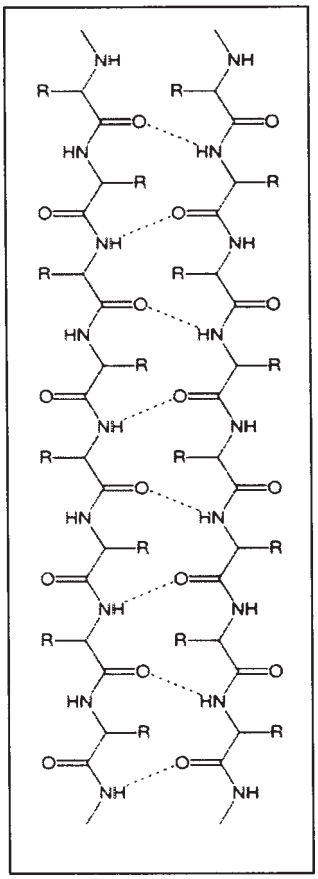


Fig. 3b.Parallel β-sheet

predicting which of the three states (helix/sheet/coil) a residue adopts. A significant increase in accuracy (to nearly ~80%) was made by exploiting multiple sequence alignment; knowing the full distribution of amino acids that occurs at a position (and in its vicinity, typically ~7 residues on either side) throughout evolution provides a much better picture of the structural tendencies near that position. For illustration, a given protein might have a glycine at a given position, which by itself might suggest a random coil there. However, multiple sequence alignment might reveal that helix-favoring amino acids occur at that position (and nearby positions) in 95% of homologous proteins spanning nearly a billion years of evolution. Moreover, by examining the average hydrophobicity at that and nearby positions, the same alignment might also suggest a pattern of residue solvent accessibility consistent with an α-helix. Taken together, these factors would suggest that the glycine of the original protein adopts α-helical structure, rather than random coil.

Secondary-structure prediction methods are continuously benchmarked, e.g., in the EVA (http://cubic.bioc.columbia.edu/eva/sec/res_sec.html) experiment. Based on approx. 270 weeks of testing, the most accurate methods at present are:

-PsiPRED (http://bioninf.cs.ucl.ac.uk./psipred/psiforum.html);

-PROF (www.predict_protein.org);

-SABLE = protein structure prediction server used to predict Solvent AccessiBiLitiEs secondary structure and transmembrane domains for proteins of unknown structure (http://sable.cchmc.org).

Interestingly, it does not seem to be possible to improve upon these methods by taking a consensus of them. The chief area for improvement appears to be the prediction of β-strands; residues confidently predicted as β-strand are likely to be so, but the methods are apt to overlook some β-strand segments.

Chou-Fasman method for protein secondary-structure prediction [5]

Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to those numbers presented in table 2.

The algorithm contains the following steps:

Assign all of the residues in the peptide the appropriate set of parameters.

Scan through the peptide and identify regions where 4 out of 6 contiguous residues have P(a-helix) > 100. That region is declared an alpha-helix. Extend the helix in both directions until a set of four contiguous residues that have an average P(a-helix) < 100 is reached. That is declared the end of the helix. If the segment defined by this procedure is longer than 5 residues and the average P(a-helix) > P(b-sheet) for that segment, the segment can be assigned as a helix.

Repeat this procedure to locate all of the helical regions in the sequence.

Scan through the peptide and identify a region where 3 out of 5 of the residues have a value of P(b-sheet) > 100. That region is declared as a beta-sheet. Extend the sheet in both directions until a set of four contiguous residues that have an average P(b-sheet) < 100 is reached. That is declared the end of the beta-sheet. Any segment of the region located by this procedure is assigned as a beta-sheet if the average P(b-sheet) > 105 and the average P(b-sheet) > P(a-helix) for that region.

Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if the average P(a-helix) > P(b-sheet) for that region. It is a beta sheet if the average P(b-sheet) > P(a-helix) for that region.

To identify a bend at residue number j, calculate the following value

$$p(t) = f(j)f(j+1)f(j+2)f(j+3)$$

where the f(j+1) value for the j+1 residue is used, the f(j+2) value for the j+2 residue is used and the f(j+3) value for the j+3 residue is used. If: (1) p(t) > 0.000075; (2) the average value for P(turn) > 1.00 in the tetrapeptide; and (3) the averages for the tetrapeptide obey the inequality P(a-helix) < P(turn) > P(b-sheet), then a beta-turn is predicted at that location.

Protein secondary-structure prediction by neural networks

Several types of methods are used to combine all the available data to form a 3-state prediction, including neural networks [3, 6], hidden Markov models and support vector machines.

Amino acid	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.11	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.19	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Table 2
CHOU-FASMAN PARAMETERS

Table 3a
CHOU-FASMAN PROTEIN DATABASE

α proteins	% α-helix	% β-sheets	β proteins	% α-helix	% β-sheets
Ca binding parvalbumin	56.5	4.6	a-Chymotripsin	8.3	40.2
Cytochrome b-52	78.6	0.0	Concanavalin	2.1	57.9
Cytochrome c	42.7	5.8	Elastase	7.5	49.2
Cytochrome c-2	43.8	3.6	Erabutoxin B	0.0	50.0
Cytochrome c-550	39.6	4.5	Immunoglobulin Fab (V_H and C_H) (human)	2.3	59.1
Cytochrome c-555	36.1	0.0	Immunoglobulin Fab (V_H and C_1) (human)	2.4	58.7
Hemerythrin (Met-)	64.6	0.0	Immunoglobulin MCG (human)	10.7	65.7
Hemerythrin (Myo-)	68.6	0.0	Immunoglobulin REI (human)	4.6	56.5
Hemerythrin (G.gouldi)	71.7	0.0	Penicillopepsin	8.7	41.8
Hemoglobin, α (human)	77.3	0.0	Prealbumin	6.3	45.7
Hemoglobin, β (human)	76.7	0.0	Protease A	8.3	51.9
Hemoglobin, α (horse)	77.3	0.0	Protease B	4.9	56.2
Hemoglobin, β (horse)	78.8	0.0	Rubredoxide Dismutase	0.0	25.9
Hemoglobin (glycera)	76.2	0.0	Superoxide Dismutase	4.6	50.3
Hemoglobin (lamprey)	79.1	0.0	Trypsin	0.0	41.3
Hemoglobin (migde larva)	83.1	0.0			
Hemoglobin,γ (human)	77.4	0.0			
Myglobin (seal)	79.1	0.0			
Myglobin (sperm whale)	79.1	0.0			

In this work was used and developed the protein secondary-structure prediction with neural networks proposed by Baughman and Liu [3]. It was used the four structural classes of proteins, as defined by Chou and Fasman [5], as output categories:

-α proteins: predominantly α-helix regions with little or no β-sheets;

-β proteins: predominantly β-sheets with minimal or no β-helix regions;

-α + β proteins: α-helices and α-sheets clustered in separate domains;

-α/β proteins: alternating α-helices and β-sheets.

It was used the protein database from Chou and Fasman [5], listed in table 3.a and 3.b which contains 19α proteins, 15β proteins, 14α +β proteins, and and 16 α/β proteins.

The training data for the classification neural networks consisted in the matrix file *protcls.nna* given by Baughman and Liu [3]. The rows of this matrix correspond to the proteins from table 3.a and 3.b, whilst the columns contain the normalized number of amino acids in the protein, the percentage contents in the twenty amino acids, and the codes of the four protein classes as follows:

{1 0 0 0} : α proteins
{0 1 0 0} : β proteins
{0 0 1 0} : α + β proteins
{0 0 0 1} : α /β proteins

According to this structure of the training data, the input layer of the neural network has 22 neurons (corresponding to the normalized number of amino acids, the percentage contents in the twenty amino acids and the bias node). The output layer of the neural network contains four neurons in correspondence with the codes of the four protein classes. Similar with Baughman and Liu [3], the best results in the training of the neural network was reached when one hidden layer with 30 nodes were used. For sigmoid transfer function and δ-learning rule with cumulative update of interneuron connections weights, after 100,000 learning iterations, the network response was a perfect correct classification.

Monte Carlo simulation

Using the classification neural network, the influence of the number of amino acids and the percentage contents in the twenty amino acids on the affiliation to a protein class was investigated by Monte Carlo simulation using CRYSTAL BALL®. During a simulation, Crystal Ball ranks the assumptions according to their importance to each forecast cell and indicate which assumptions are the most important or least important in the model [7]. The assumptions for the signal variables (respectively the inputs of neural network) have a stochastic uniform distribution between minimum and maximum values. The forecasts correspond to the outputs of neural network, respectively

Table 3.b.
CHOU-FASMAN PROTEIN DATABASE

α - β proteins	% α-helix	% β-sheets	α /β proteins	% α-helix	% β-sheets
Actinidin	28.0	14.2	Adenylate kinase	54.1	12.4
Cytochrome b-5	46.2	28.0	Alcohol dehydrogenase	28.3	30.8
Ferredoxin	24.1	31.5	Carbonic anhydrase B	18.9	27.7
High-protein iron protein	11.8	15.3	Carbonic anhydrase C	20.5	26.6
Insulin	49.0	23.5	Carboxypeptidase A	35.2	14.7
Lysozyme (bacteriophage T4)	65.2	12.2	Carboxypeptidase B	31.4	17.7
Lysozyme (chicken)	41.9	17.1	Dihydrofolate reductase	17.6	30.8
Papain	26.4	14.2	Flavodoxim	36.2	26.8
Phospholipase A-2	49.6	9.8	Glyceraldehyde 3-phosphate dehydrogenase (lobster)	32.7	34.5
Ribonuclease S	25.0	44.4	Glyceraldehyde 3-phosphate dehydrogenase (B stearotherm)	31.1	26.7
Staphylococcal nuclease	25.5	28.9	Lactate dehydrogenase	40.4	24.0
Subtilisin inhibitor	17.7	33.6	Phosphoglycerate kinase	40.9	23.8
Thermolysin	35.4	20.6	Rhodanese	41.0	14.3
Trypsin inhibitor	19.0	27.6	Subtilisin BPN	29.1	20.0
			Thiordoxin	48.2	27.8
			Thiose phosphate isomerase	54.0	20.2

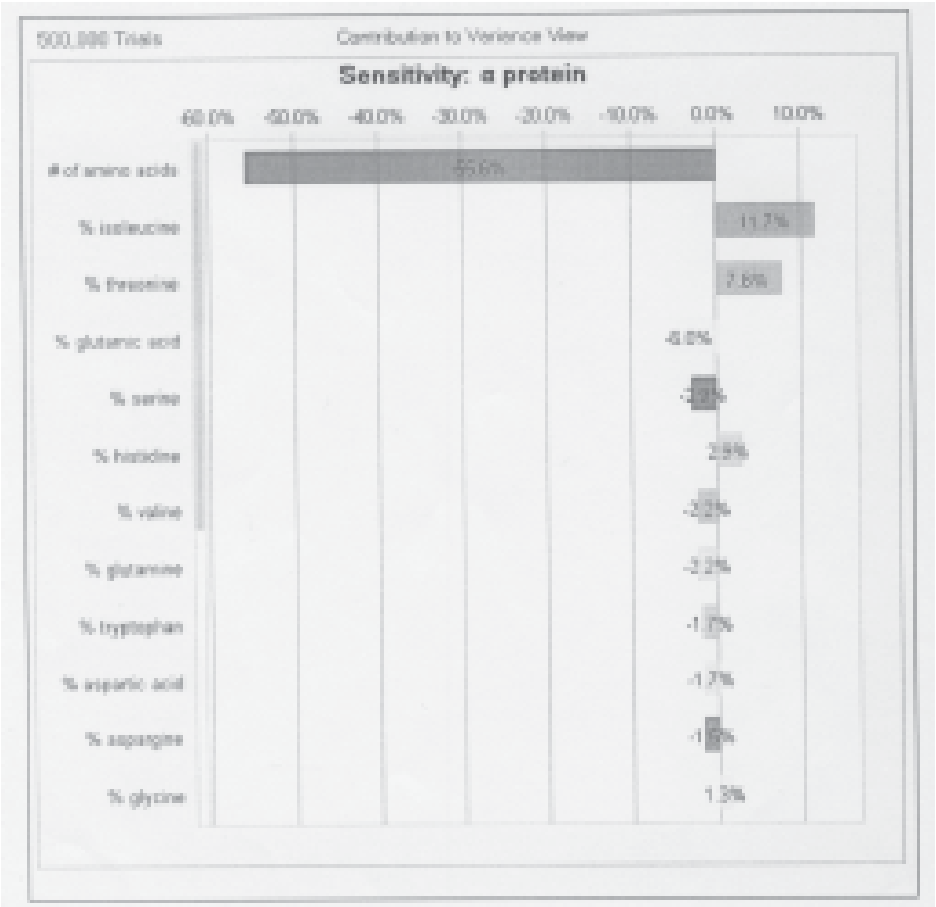


Fig. 4. Sensitivity chart for α protein structure

Table 4
PERCENTAGE CONTRIBUTION TO VARIANCE

	With # amino acids				Without # amino acids			
Variable	α	β	α + β	α /β	α	β	α + β	α /β
# amino acids	-55.6	-7.9	-18.8	74.8	-	-	-	-
% alanine	-	-36.1	21.5	-	-	-39.2	26.5	-
% arginine	-	2.2	-1.5	-	-	2.4	-1.9	-
% asparagine	-1.5	-	-	4.2	-3.5	-	-	16.6
% aspartic acid	-1.7	-	-	4.7	-3.7	-	-	18.6
% cysteine	-	1.6	-	-	-	1.8	-1.6	-
% glutamine	-2.2	2.3	-2.7	3.0	-4.9	2.5	-3.3	12
% glutamic acid	-6	-	4.4	-	-13.5	-	5.4	-
% glycine	1.3	-0.9	-1.5	0.3	2.8	-1.0	-1.8	1.2
% histidine	2.9	-	-	-0.6	6.5	-0.3	-	-3.2
% isoleucine	11.7	-	-9.4	-1.2	26.3	-	-11.7	-4.6
% leucine	-	-0.9	15.2	-1.0	-2.5	-1.0	18.7	-4.1
% lysine	-	-	-	-	-	-	-	-
% methionine	-	-0.4	-	-	-	-0.4	-	-
% phenylalanine	-	-	-	-	-	-	-	-
% proline	-	12.8	-4.7	-0.9	-	13.8	-5.8	-3.4
% serine	-2.9	29.6	-	-4.8	-6.5	32.1	-	-19.2
% threonine	7.8	-2.9	-2.6	-	17.5	-3.1	-3.2	1.2
% tryptophan	-1.7	-	7.9	-0.3	-3.9	-	9.8	-1.2
% tyrosine	-	-	-	-	-	-	-	-
% valine	-2.2	-1.7	5	3.4	-5.1	-1.8	6.2	13.5

the codes of the four protein classes. It was used 500,000 simulation trials. The corresponding results are given by Crystal Ball sensitivity charts. In figure 4 is presented one of these sensitivity charts, respectively for α protein structure. Negative values of a percentage contribution to variance indicate an inverse proportionality. The complete results of the Monte Carlo sensitivity analysis are given in table 4. Here are also presented the simulation results without consideration of the influence of the normalized number of amino acids.

Results and Discussion

According to the results presented in table 4, a high number of amino acids in protein molecule will be very favorable for an α /β protein structure and strong no favorable for an á protein structure. The number of amino acids has a moderate negative influence to α + β protein structure and a small negative influence to β protein structure.

For α protein structure are favorable the presence of isoleucine and threonine, for β protein structure serine and proline, for α + β protein structure alanine, leucine and tryptophan, and for α/β protein structure aspartic acid and aspargine. There is some disagreement with reference [4] and future investigations must confirm or infirm these results.

Conclusions

Despite the fact that several methods were developed for protein secondary-structure prediction, there are no consensuses of their results. Here was proposed a new, original method to investigate the influence of the number of amino acids and the percentage contents in the twenty amino acids to predict protein secondary-structure, respectively Monte Carlo simulation using a multilayer neural networks. The corresponding results are not strong consistent, especially because a small data base was used, but the method is very promising in connection with the use of large data bases. Accurate secondary-structure prediction is a key element in the prediction of tertiary structure, in all but the simplest (homology modeling) cases.

References

1.BETANCOURT M., Protein Structure Prediction, Databases & Neural Networks, www.ccr.buffalo.edu/Workshop03/mb-ccr-proteins.pdf
2.BRANDEN C., TOOZE T., Introduction to Protein Structure, Garland Publishing Inc., New York, 1991
3.BAUGHMAN D.R., LIU Y.A., Neural Networks in Bioprocesing and Chemical Engineering, Academic Press, New York, 1995
4.*** WIKIPEDIA, http://en.wikipedia.org/wiki/Protein_secondary_structure
5.CHOU P.Y., FASMAN G.D., Prediction of Protein Structure and the Principles of Protein Conformation, Plenium Press, New York, 1989
6.SHEPHERD, A., Protein Secondary Structure Prediction with Neural Networks Tutorial, www.biochem.ucl.ac.uk/~shepherd/sspred_tutorial/ss-index.html
7.*** DECISIONEERING® INC., Crystal Ball® 7.2 User Manual, 2005

Manuscript received: 15.08.20097