Institute for Mathematics and Scientific Computing
University of Graz

Sarajevo Summer School :
Mathematical Techniques in Modeling Physiological Systems
September 10-22, 2006

FEED-FORWARD NEURAL NETWORK FOR PROTEIN STRUCTURE PREDICTION

Faculty of Electrical Engineering, Sarajevo
Zikrija Avdagic@Elvir Purisevic

Project timetable:

- 2004-2006/Jan-Dec ,Sarajevo

Theoretical research :

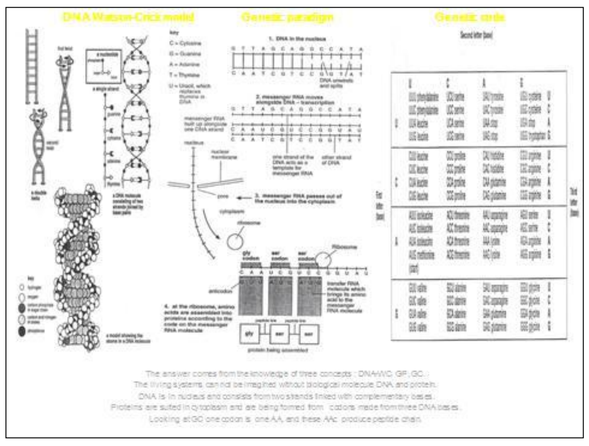
- Field: Artificial Intelligence
- Field: Bioinformatic (Genomic&Proteomic)

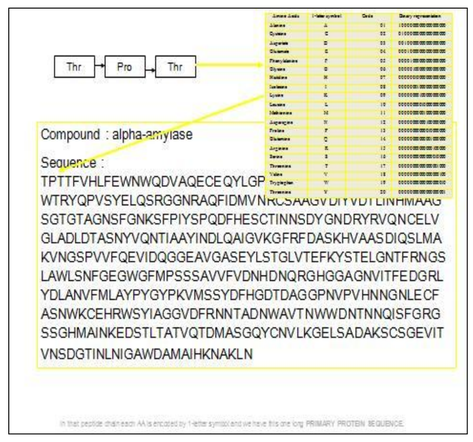
Practical implementation :

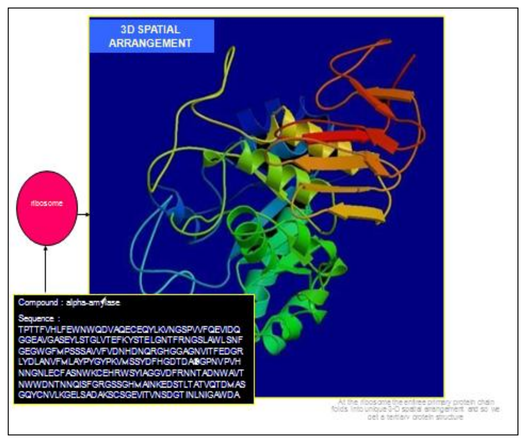
- Borland Delphi
- MATLAB-Neural Network Toolbox
- Databases (SWISS-PROT, PROSITE, PDBFINDER).

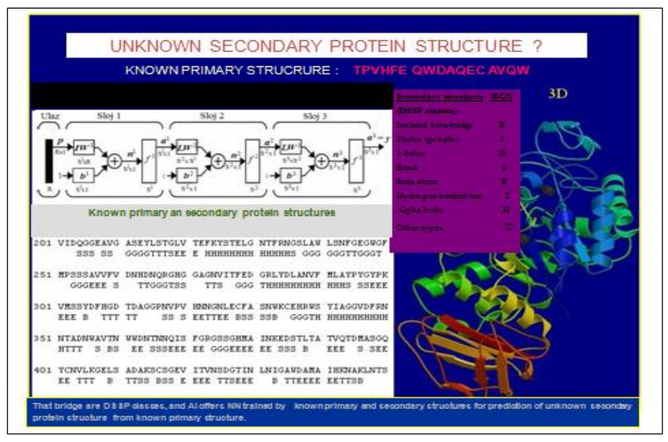
Project description

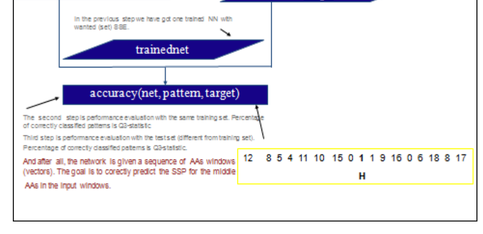
The problem of predicting protein structure based on amino acids sequence is one of the most interesting problem in molecular biology. Determination of protein structure based on experimental methods is very expensive, long term duration and request experts from different fields. For some types of protein, theoretic based methods for structure prediction are only one alternative. 3D protein structure determines the protein function and our first step to it is secondary protein structure prediction. In practise, the most successful structure prediction extracts patterns from data bases of known protein structures. Neural networks comprise a particular tool for patterns from data bases of known protein structure. In this project we implemented algorithm for prediction of secondary protein structure in Borland Delphi and MatLab – Neural Network Toolbox.

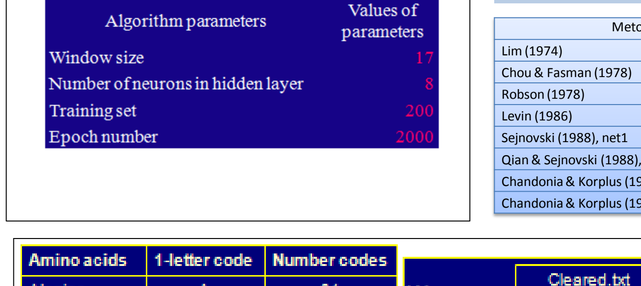






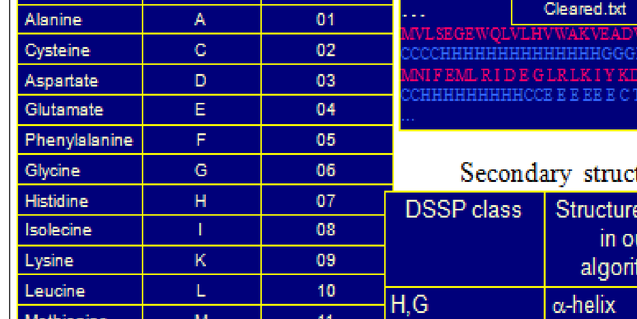


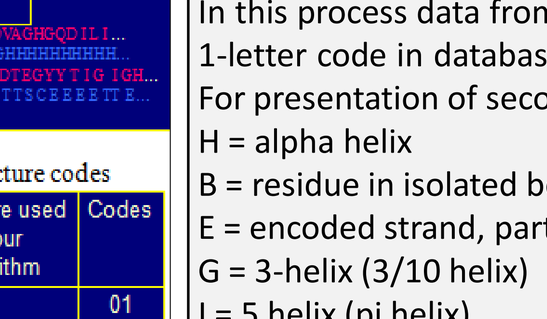


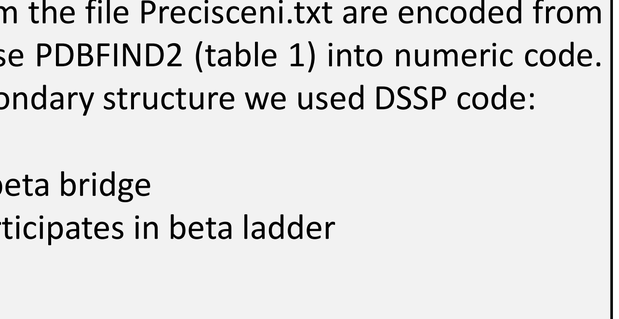


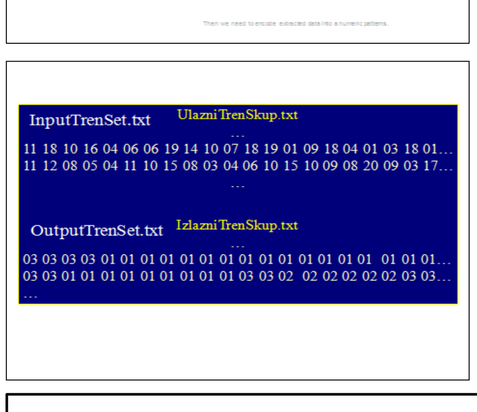
Application of an empirical approach to protein structure prediction is entirely dependent on the experimental databases which are available for analysis, generalization and extrapolation. Since all of the studies discussed below ara dependent on these databases, a brief discussion of their contents is appropriate. As a result of a large sequencing project (such as the Human Genome Project) data banks of protein sequences and structures are qrowing rapidly. All these database can be categorized according to type of protein sructure. So we have available primary sequence databases, composite databases and secondary structure databases. The best known secondary database is PROSITE established on SWISS-PROT database. At databases is stored a large amount of differenet data, but our algorithm need only fragments of protein primary sequence and corresponding secondary structures. PDBFIND2 data bases enables us to be effective extracting these data to our data files. Authors of this data base are Krieger E., Rob W.W. Hooft, Nabuurs S., and Gert V., and it is accessible by FTP server .

The implementation of this algorithm is provided with two software packages. The first one is API-EPE (Application interface used here for extracting, preparing amd encoding of data examples), and the second one is MATLAB & NN-TOOLBOX (software environment used here for the determination of input/output data training sets, the design of the neural network architecture, training/learning of NN-predictor, and the system performance evaluation.









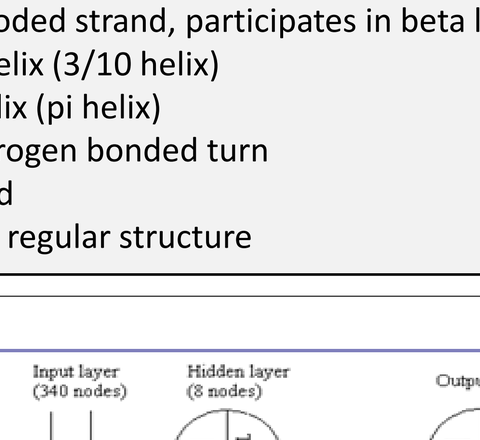
Parameter values for which we have the best prediction accuracy

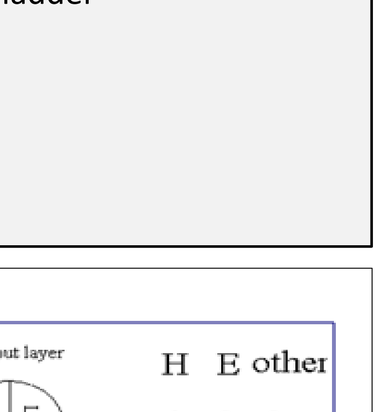
Q3
63.6261 %

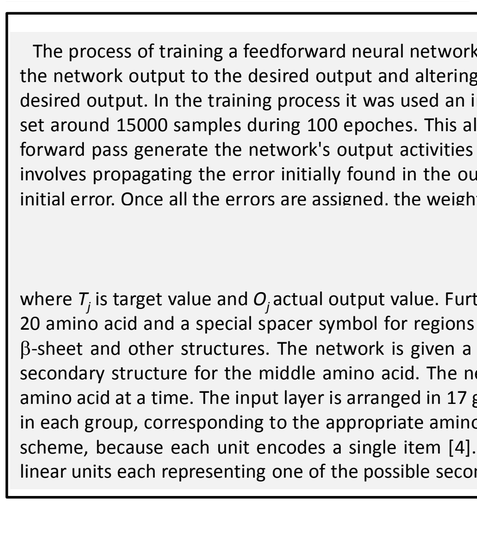
Algorithm parameters	Values of parameters
Window size	17
Number of neurons in hidden layer	8
Training set	200
Epoch number	2000

Comaprision with other methods
Avdagic&Purisevic
Q3
63.6261%

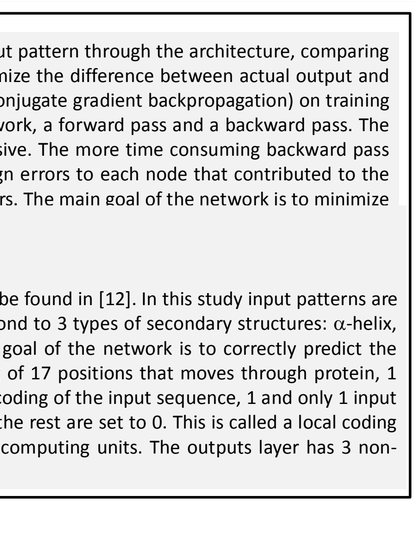
Metoda	Q3%
Lim (1974)	50
Chou & Fasman (1978)	50
Robson (1978)	53
Levin (1986)	59.7
Sejnovski (1988), net1	62.7
Qian & Sejnovski (1988), net2	64.3
Chandonia & Korplus (1995)	73.9
Chandonia & Korplus (1996)	80.2







Amino acids	1-letter code	Number codes
Alanine	A	01
Cysteine	C	02
Aspartate	D	03
Glutamate	E	04
Phenylalanine	F	05
Glycine	G	06
Histidine	H	07
Isolecine	I	08
Lysine	K	09
Leucine	L	10
Methionine	M	11
Asparagine	N	12
Proline	P	13
Glutamine	Q	14
Arginine	R	15
Serine	S	16
Threonine	T	17
Valine	V	18
Tryptophan	W	19
Threonine	Y	20



DSSP class	Structure used in our algorithm	Codes
H,G	α-helix	01
E	β-strand	02
B, I, S, T, C, L	other structure	03

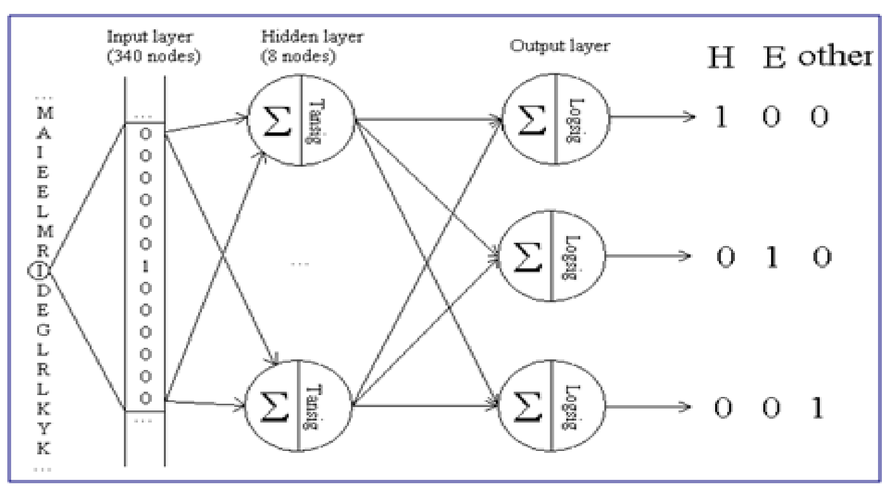
In this process data from the file Precisceni.txt are encoded from 1-letter code in database PDBFIND2 (table 1) into numeric code. For presentation of secondary structure we used DSSP code:

H = alpha helix
B = residue in isolated beta bridge
E = encoded strand, participates in beta ladder
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
T = hydrogen bonded turn
S = bend
C = non regular structure

The process of training a feedforward neural network involves presenting the network with on input pattern through the architecture, comparing the network output to the desired output and altering the weights in the direction in order to minimize the difference between actual output and desired output. In the training process it was used an improved backpropagation algorithm (scaled conjugate gradient backpropagation) on training set around 15000 samples during 100 epoches. This algorithm involves two passes through the network, a forward pass and a backward pass. The forward pass generate the network's output activities and it is generally the least computation intesive. The more time consuming backward pass involves propagating the error initially found in the output nodes back through the network to assign errors to each node that contributed to the initial error. Once all the errors are assigned, the weights are changed in order to minimize these errors. The main goal of the network is to minimize

$$E = \sum_{p} \sum_{j} (T_j - O_j)^2$$

where T_j is target value and O_j actual output value. Further details on backpropagation algorithm can be found in [12]. In this study input patterns are 20 amino acid and a special spacer symbol for regions between proteins; the target patterns corespond to 3 types of secondary structures: α-helix, β-sheet and other structures. The network is given a contiguous sequence of 17 amino acids. The goal of the network is to correctly predict the secondary structure for the middle amino acid. The network is presented with a window consisting of 17 positions that moves through protein, 1 amino acid at a time. The input layer is arranged in 17 groups. Each group has 20 units. For a local encoding of the input sequence, 1 and only 1 input in each group, corresponding to the appropriate amino acid at each position, is given a value 1, and the rest are set to 0. This is called a local coding scheme, because each unit encodes a single item [4]. The hidden layer is arranged in 8 non-linear computing units. The outputs layer has 3 non-linear units each representing one of the possible secondary structures for the centre amino acid.



MATRIX-VECTOR NOTATION OF NEURAL NETWORK