Prediction of Protein Secondary Structure using Artificial Neural Network

MN Vamsi Thalatam[1], P Venkata Rao[2], KVSRP Varma[3], NVR Murty[4] ,Allam Apparao[5]

1. Associate Professor, Dept.of MCA, GVPCollege for Degree & PGCourses, Visakhapatnam, India.
2. Associate Professor, Dept.of MCA, GVPCollege for Degree & PGCourses, Visakhapatnam, India.
3. Assistant Professor, Dept.of CSE, GITAM University, Visakhapatnam, India
4. Associate Professor, Dept.ofMCA, GVPCollege for Degree & PGCourses, Visakhapatnam, India.
5. Vice-Chancellor, JNTUK, Kakinada. India

Abstract:

Structural information can provide insight into protein function, and therefore, high- accuracy prediction of protein structure from its sequence is highly desirable. We predicted the secondary structure of query protein sequence using Artificial Neural Network available in Neurosolutions. The structure is described in terms of Alpha Helix (H), Extended Strand (E) and Random Coil (C). The results are displayed in tabular forms. By proper training of the network on the data related to structure we tested and predicted the secondary structure of the query protein.

Introduction:

Proteins are essential to biological processes. Protein function can be understood in terms of its structure. With the completion of many large Genomes enormous amount of amino acid data is available from which one can predict the secondary structure of the query sequence. Although the prediction of tertiary structure is one of the ultimate goals of protein science, the prediction of secondary structure from sequence is still a more feasible intermediate step in this direction. Furthermore, some knowledge of the secondary structure can serve as an input for prediction [1]. Instead of predicting the full three-dimensional structure, it is much easier to predict simplified aspects of structure, namely the key structural elements of the protein and the location of these elements not in the three-dimensional space but along the protein amino acid sequence. This reduces the complex three-dimensional problem to a much simpler one-dimensional problem. The fundamental elements of the secondary structure of proteins are α-helices, β-sheets, coils, and turns. All these elements can be easily observed in the crystal three dimensional structure of proteins in the PDB [1]. According to the DSSP classification, there are eight elements of secondary structure assignment denoted by letters: H(a-helix),E(extended b-strand),G(310 helix), I (p-helix), B(bridge, a single residue b-strand),T(b-turn), S (bend), and C (coil).But for existing method of the secondary structure prediction, instead usually only three states are predicted: α-helix (H), extended (β-sheet) (E), and coil (C)[1].

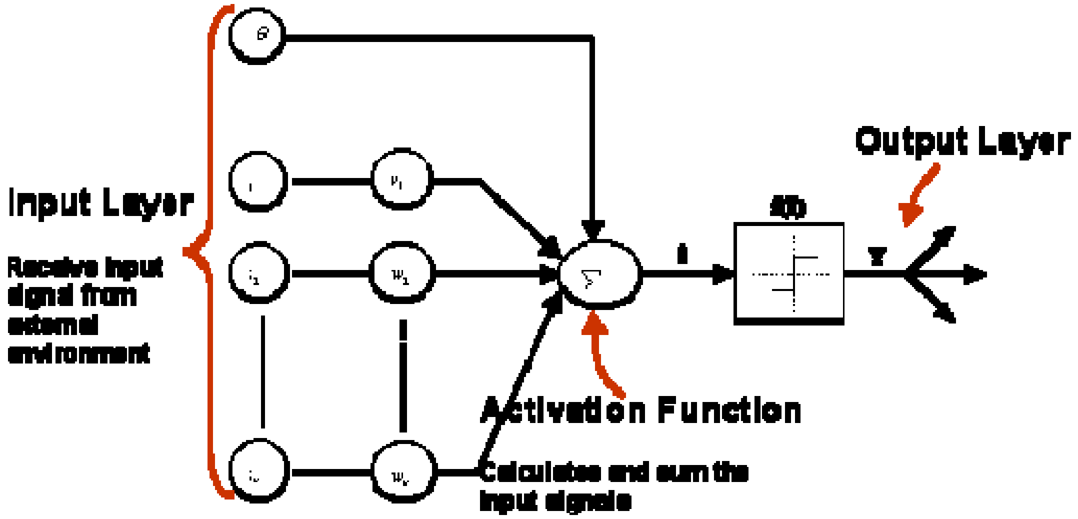

Input Layer
Receive input signal from external environment
θ
i
i_1
i_k
w_1
w_k
Σ
t
f(t)
Y
Output Layer
Activation Function
Calculates and sum the input signals

Figure 1 : McCulloch-Pitts Neuron Model

Neural Network role in the Prediction of the structure:

Artificial Neural Network could be define as an interconnected of simple processing element whose functionality is based on the biological neuron.

Simple neuron (Figure 1) introduced by McCulloch and Pitts in 1940s, consists of input layer, activation function, and output layer [6].Input layer receive input signal from external environment (or other neuron). Activation function is the neuron internal states that calculates and sum the input signals. The signals are then transmitted to output layer. The input layer, activation function and output layer in artificial neuron are similar to the function of dendrites, soma and axon in biological neuron.

Training the Network

Training the network is time consuming. It usually learns after several epochs, depending on how large the network is. Thus, large network required more training time compared to the smaller one. Basically, the network is trained for several epochs and stopped after reaching the maximum epoch. For the same reason minimum error tolerance is used provided that the differences between network output and known outcome are less than the specified value [2]. We could also stop the training after the network meets certain stopping criteria. During training the network might learn too much. This problem is referred to as overfitting. Overfitting is a critical problem in most all standard NNs architecture. Furthermore, NNs and other AI machine learning models are prone to overfitting [3]. One of the solutions is early stopping [4], but this approach need more critical intention as this problem is harder than expected [3]. The stopping criterion is also another issue to consider in preventing overfitting [5]. Hence, for this problem during training, validation set is used instead of training data set. After a few epochs the network is tested with the validation data. The training is stopped as soon as the error on validation set increases rapidly higher than the last time it was checked [5]. Figure 2 shows that the training should stop at time t when validation error starts to increase.

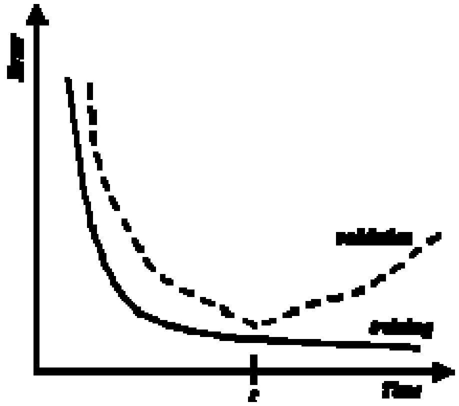


Figure 2: Training and validation curve

Constructing a program for Neural Network is not a difficult task. Basically, it was only several steps of algorithms that are easily followed even by novice practitioners. However, preparing the network for training is a difficult task since the network dealing with a large amount of data. Another problem is when to stop the training? Over training could cause memorization where the network might simply memorize the data patterns and might fail to recognize other set of patterns. Thus, early stopping is recommended to ensure that the network learn accordingly.

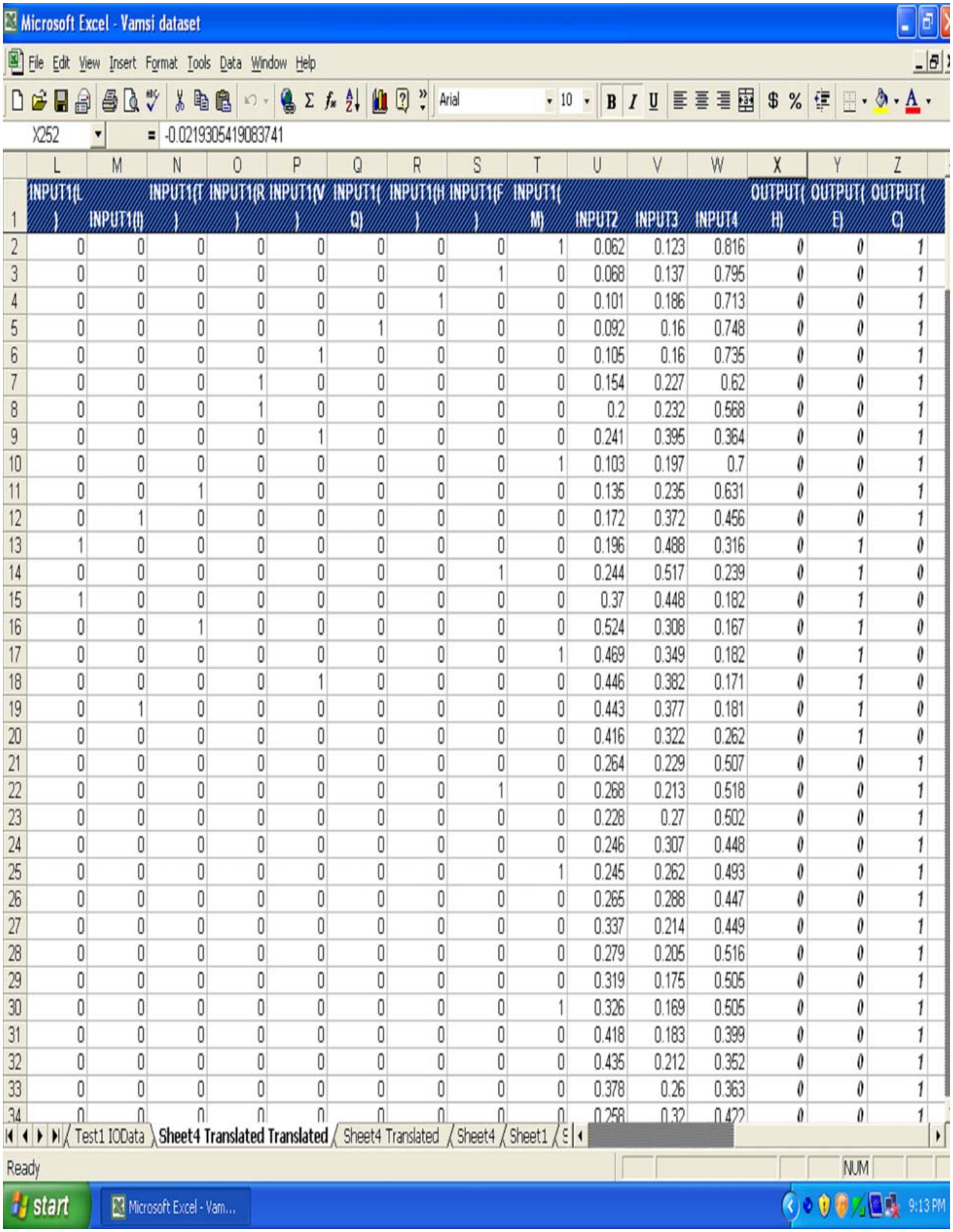
Methods and Materials:

The query protein sequence is retrieved from NCBI repository in FASTA format and then submits this sequence to GOR V tool and uses this data for the secondary structure prediction using neural network. Trained this data in the neural network to classify the three parameters in the data ,i.e α-helix(H), β-sheet(E) and Coil (C) in terms of either 0 or 1. There are 256 Rows Containing the data of GOR V output is obtained for the given sequence in which the columns labeled as Input 1, Input 2, Input 3, Input 4 will serve as inputs to the neural network and the columns labeled Output will serve as the desired outputs. Here Input 1, which is given as symbolic data at the initial stage, is converted to numerical data (0s & 1s). Before train a neural network, it needs to know which columns to use as inputs and which columns to use as the desired outputs. To accomplish this, we selected the columns that we want to use as inputs and then selected "Tag Data! Column as Input". Similarly, to tag the desired output columns, we selected the corresponding columns then select "Tag Data|Column as Desired. The rows of data must also be tagged before a training process can be run. Rows can be tagged as "Training", "Cross Validation", "Testing", or "Production". However, cross validation is a very useful tool for preventing over-training, so in most cases we will also want to tag a portion of our data as "Cross Validation". The rows of data to use for testing the trained network ("Testing") or producing the network output for new data ("Production") can be tagged before or after the training process is run.

Results and Discussions:

Secondary structure prediction of query sequence is analyzed using the artificial neural network by considering the output given by GOR V tool as the input for the neural network. The three parameters α-helix (H), β-sheet (E) and Coil (C) in terms of either 0 or 1are tabulated for all 256 rows. To tag rows of data we select the appropriate rows then we select the corresponding tagging operation. The first 150 rows we have used for training 1000 times and the next 50 rows used for cross validation. At last we found that the network did a good job of classifying the samples that were used to train the network. In the next step we tested the networks performance on the 50 rows of data, which were tagged as "Testing", after testing the data set through the network, we produced a report summarizing the network classification performance. And finally we have given 5 rows for producing new results for the given data and the results are very fine and the network is able to produce new results for any given data, so production phase also tested successfully. All the results are tabulated.

Table:1 to 3-The predicted output for given input

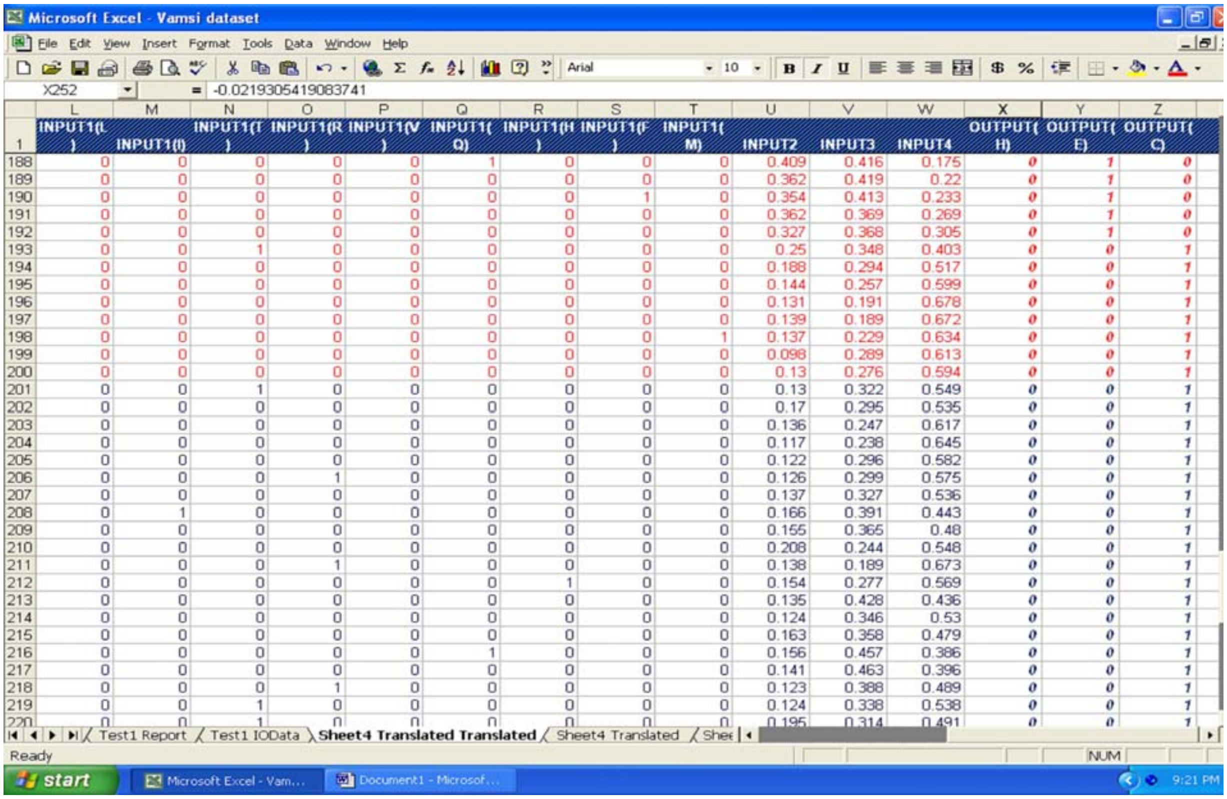

Microsoft Excel - Vamsi dataset
File Edit View Insert Format Tools Data Window Help
Arial
10
Test1 IOData
Sheet4 Translated Translated
Sheet4 Translated
Sheet4
Sheet1
Ready
NUM
start
Microsoft Excel - Vam...
9:13 PM

X252 = -0.0219305419083741

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	INPUT1(L)	INPUT1(I)	INPUT1(T)	INPUT1(R)	INPUT1(V)	INPUT1(Q)	INPUT1(H)	INPUT1(F)	INPUT1(M)	INPUT2	INPUT3	INPUT4	OUTPUT(H)	OUTPUT(E)	OUTPUT(C)
2	0	0	0	0	0	0	0	0	1	0.062	0.123	0.816	0	0	1
3	0	0	0	0	0	0	0	1	0	0.068	0.137	0.795	0	0	1
4	0	0	0	0	0	0	1	0	0	0.101	0.186	0.713	0	0	1
5	0	0	0	0	0	1	0	0	0	0.092	0.16	0.748	0	0	1
6	0	0	0	0	1	0	0	0	0	0.105	0.16	0.735	0	0	1
7	0	0	0	1	0	0	0	0	0	0.154	0.227	0.62	0	0	1
8	0	0	0	1	0	0	0	0	0	0.2	0.232	0.568	0	0	1
9	0	0	0	0	1	0	0	0	0	0.241	0.395	0.364	0	0	1
10	0	0	0	0	0	0	0	0	1	0.103	0.197	0.7	0	0	1
11	0	0	1	0	0	0	0	0	0	0.135	0.235	0.631	0	0	1
12	0	1	0	0	0	0	0	0	0	0.172	0.372	0.456	0	0	1
13	1	0	0	0	0	0	0	0	0	0.196	0.488	0.316	0	1	0
14	0	0	0	0	0	0	0	1	0	0.244	0.517	0.239	0	1	0
15	1	0	0	0	0	0	0	0	0	0.37	0.448	0.182	0	1	0
16	0	0	1	0	0	0	0	0	0	0.524	0.308	0.167	0	1	0
17	0	0	0	0	0	0	0	0	1	0.469	0.349	0.182	0	1	0
18	0	0	0	0	1	0	0	0	0	0.446	0.382	0.171	0	1	0
19	0	1	0	0	0	0	0	0	0	0.443	0.377	0.181	0	1	0
20	0	0	0	0	0	0	0	0	0	0.416	0.322	0.262	0	1	0
21	0	0	0	0	0	0	0	0	0	0.264	0.229	0.507	0	0	1
22	0	0	0	0	0	0	0	1	0	0.268	0.213	0.518	0	0	1
23	0	0	0	0	0	0	0	0	0	0.228	0.27	0.502	0	0	1
24	0	0	0	0	0	0	0	0	0	0.246	0.307	0.448	0	0	1
25	0	0	0	0	0	0	0	0	1	0.245	0.262	0.493	0	0	1
26	0	0	0	0	0	0	0	0	0	0.265	0.288	0.447	0	0	1
27	0	0	0	0	0	0	0	0	0	0.337	0.214	0.449	0	0	1
28	0	0	0	0	0	0	0	0	0	0.279	0.205	0.516	0	0	1
29	0	0	0	0	0	0	0	0	0	0.319	0.175	0.505	0	0	1
30	0	0	0	0	0	0	0	0	1	0.326	0.169	0.505	0	0	1
31	0	0	0	0	0	0	0	0	0	0.418	0.183	0.399	0	0	1
32	0	0	0	0	0	0	0	0	0	0.435	0.212	0.352	0	0	1
33	0	0	0	0	0	0	0	0	0	0.378	0.26	0.363	0	0	1
34	0	0	0	0	0	0	0	0	0	0.258	0.32	0.422	0	0	1

Table:1

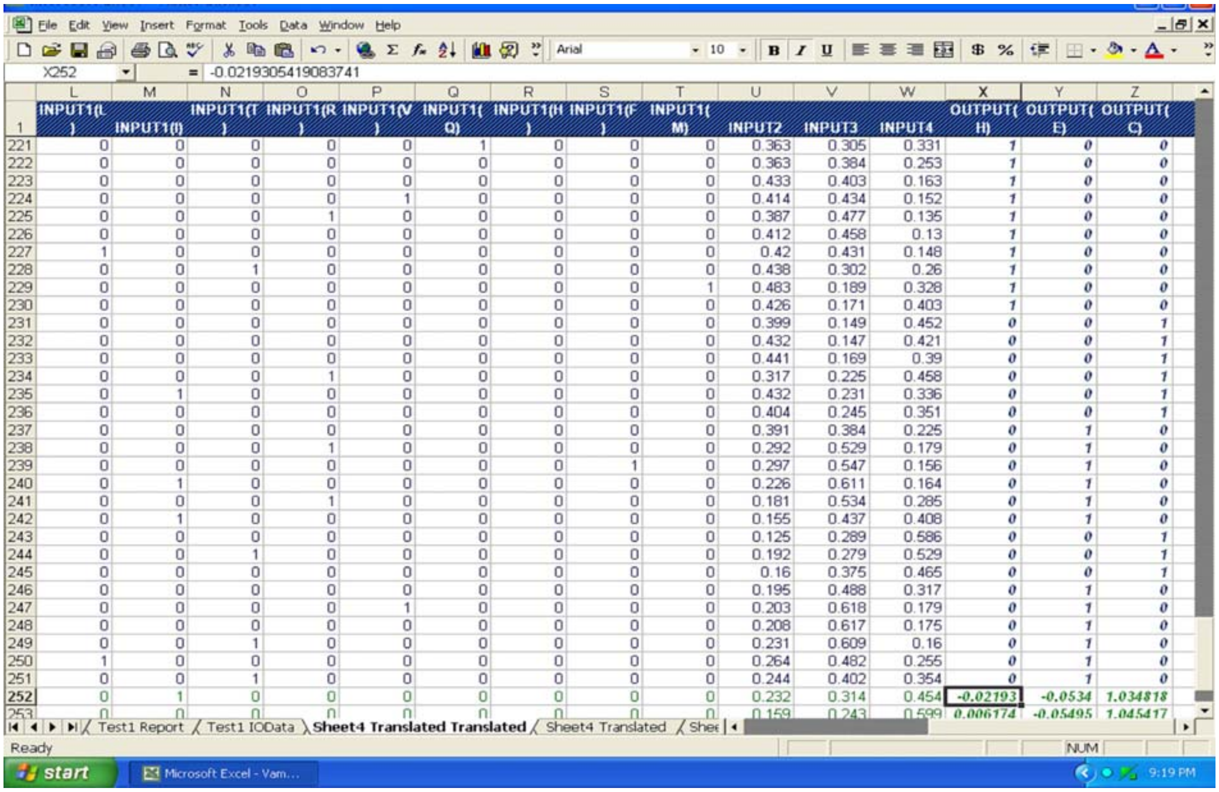
Table:2



X252 = -0.0219305419083741

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	INPUT1(L)	INPUT1(I)	INPUT1(T)	INPUT1(R)	INPUT1(V)	INPUT1(Q)	INPUT1(H)	INPUT1(F)	INPUT1(M)	INPUT2	INPUT3	INPUT4	OUTPUT(H)	OUTPUT(E)	OUTPUT(C)
188	0	0	0	0	0	1	0	0	0	0.409	0.416	0.175	0	1	0
189	0	0	0	0	0	0	0	0	0	0.362	0.419	0.22	0	1	0
190	0	0	0	0	0	0	0	1	0	0.354	0.413	0.233	0	1	0
191	0	0	0	0	0	0	0	0	0	0.362	0.369	0.269	0	1	0
192	0	0	0	0	0	0	0	0	0	0.327	0.368	0.305	0	1	0
193	0	0	1	0	0	0	0	0	0	0.25	0.348	0.403	0	0	1
194	0	0	0	0	0	0	0	0	0	0.188	0.294	0.517	0	0	1
195	0	0	0	0	0	0	0	0	0	0.144	0.257	0.599	0	0	1
196	0	0	0	0	0	0	0	0	0	0.131	0.191	0.678	0	0	1
197	0	0	0	0	0	0	0	0	0	0.139	0.189	0.672	0	0	1
198	0	0	0	0	0	0	0	0	1	0.137	0.229	0.634	0	0	1
199	0	0	0	0	0	0	0	0	0	0.098	0.289	0.613	0	0	1
200	0	0	0	0	0	0	0	0	0	0.13	0.276	0.594	0	0	1
201	0	0	1	0	0	0	0	0	0	0.13	0.322	0.549	0	0	1
202	0	0	0	0	0	0	0	0	0	0.17	0.295	0.535	0	0	1
203	0	0	0	0	0	0	0	0	0	0.136	0.247	0.617	0	0	1
204	0	0	0	0	0	0	0	0	0	0.117	0.238	0.645	0	0	1
205	0	0	0	0	0	0	0	0	0	0.122	0.296	0.582	0	0	1
206	0	0	0	1	0	0	0	0	0	0.126	0.299	0.575	0	0	1
207	0	0	0	0	0	0	0	0	0	0.137	0.327	0.536	0	0	1
208	0	1	0	0	0	0	0	0	0	0.166	0.391	0.443	0	0	1
209	0	0	0	0	0	0	0	0	0	0.155	0.365	0.48	0	0	1
210	0	0	0	0	0	0	0	0	0	0.208	0.244	0.548	0	0	1
211	0	0	0	1	0	0	0	0	0	0.138	0.189	0.673	0	0	1
212	0	0	0	0	0	0	1	0	0	0.154	0.277	0.569	0	0	1
213	0	0	0	0	0	0	0	0	0	0.135	0.428	0.436	0	0	1
214	0	0	0	0	0	0	0	0	0	0.124	0.346	0.53	0	0	1
215	0	0	0	0	0	0	0	0	0	0.163	0.358	0.479	0	0	1
216	0	0	0	0	0	1	0	0	0	0.156	0.457	0.386	0	0	1
217	0	0	0	0	0	0	0	0	0	0.141	0.463	0.396	0	0	1
218	0	0	0	1	0	0	0	0	0	0.123	0.388	0.489	0	0	1
219	0	0	1	0	0	0	0	0	0	0.124	0.338	0.538	0	0	1
220	0	0	1	0	0	0	0	0	0	0.195	0.314	0.491	0	0	1

Table:3



X252 = -0.0219305419083741

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	INPUT1(L)	INPUT1(I)	INPUT1(T)	INPUT1(R)	INPUT1(V)	INPUT1(Q)	INPUT1(H)	INPUT1(F)	INPUT1(M)	INPUT2	INPUT3	INPUT4	OUTPUT(H)	OUTPUT(E)	OUTPUT(C)
221	0	0	0	0	0	1	0	0	0	0.363	0.305	0.331	1	0	0
222	0	0	0	0	0	0	0	0	0	0.363	0.384	0.253	1	0	0
223	0	0	0	0	0	0	0	0	0	0.433	0.403	0.163	1	0	0
224	0	0	0	0	1	0	0	0	0	0.414	0.434	0.152	1	0	0
225	0	0	0	1	0	0	0	0	0	0.387	0.477	0.135	1	0	0
226	0	0	0	0	0	0	0	0	0	0.412	0.458	0.13	1	0	0
227	1	0	0	0	0	0	0	0	0	0.42	0.431	0.148	1	0	0
228	0	0	1	0	0	0	0	0	0	0.438	0.302	0.26	1	0	0
229	0	0	0	0	0	0	0	0	1	0.483	0.189	0.328	1	0	0
230	0	0	0	0	0	0	0	0	0	0.426	0.171	0.403	1	0	0
231	0	0	0	0	0	0	0	0	0	0.399	0.149	0.452	0	0	1
232	0	0	0	0	0	0	0	0	0	0.432	0.147	0.421	0	0	1
233	0	0	0	0	0	0	0	0	0	0.441	0.169	0.39	0	0	1
234	0	0	0	1	0	0	0	0	0	0.317	0.225	0.458	0	0	1
235	0	1	0	0	0	0	0	0	0	0.432	0.231	0.336	0	0	1
236	0	0	0	0	0	0	0	0	0	0.404	0.245	0.351	0	0	1
237	0	0	0	0	0	0	0	0	0	0.391	0.384	0.225	0	1	0
238	0	0	0	1	0	0	0	0	0	0.292	0.529	0.179	0	1	0
239	0	0	0	0	0	0	0	1	0	0.297	0.547	0.156	0	1	0
240	0	1	0	0	0	0	0	0	0	0.226	0.611	0.164	0	1	0
241	0	0	0	1	0	0	0	0	0	0.181	0.534	0.285	0	1	0
242	0	1	0	0	0	0	0	0	0	0.155	0.437	0.408	0	1	0
243	0	0	0	0	0	0	0	0	0	0.125	0.289	0.586	0	0	1
244	0	0	1	0	0	0	0	0	0	0.192	0.279	0.529	0	0	1
245	0	0	0	0	0	0	0	0	0	0.16	0.375	0.465	0	0	1
246	0	0	0	0	0	0	0	0	0	0.195	0.488	0.317	0	1	0
247	0	0	0	0	1	0	0	0	0	0.203	0.618	0.179	0	1	0
248	0	0	0	0	0	0	0	0	0	0.208	0.617	0.175	0	1	0
249	0	0	1	0	0	0	0	0	0	0.231	0.609	0.16	0	1	0
250	1	0	0	0	0	0	0	0	0	0.264	0.482	0.255	0	1	0
251	0	0	1	0	0	0	0	0	0	0.244	0.402	0.354	0	1	0
252	0	1	0	0	0	0	0	0	0	0.232	0.314	0.454	-0.02193	-0.0534	1.034818
253	0	0	0	0	0	0	0	0	0	0.159	0.243	0.599	0.006174	-0.05495	1.045417

Fig 3:Neural Network

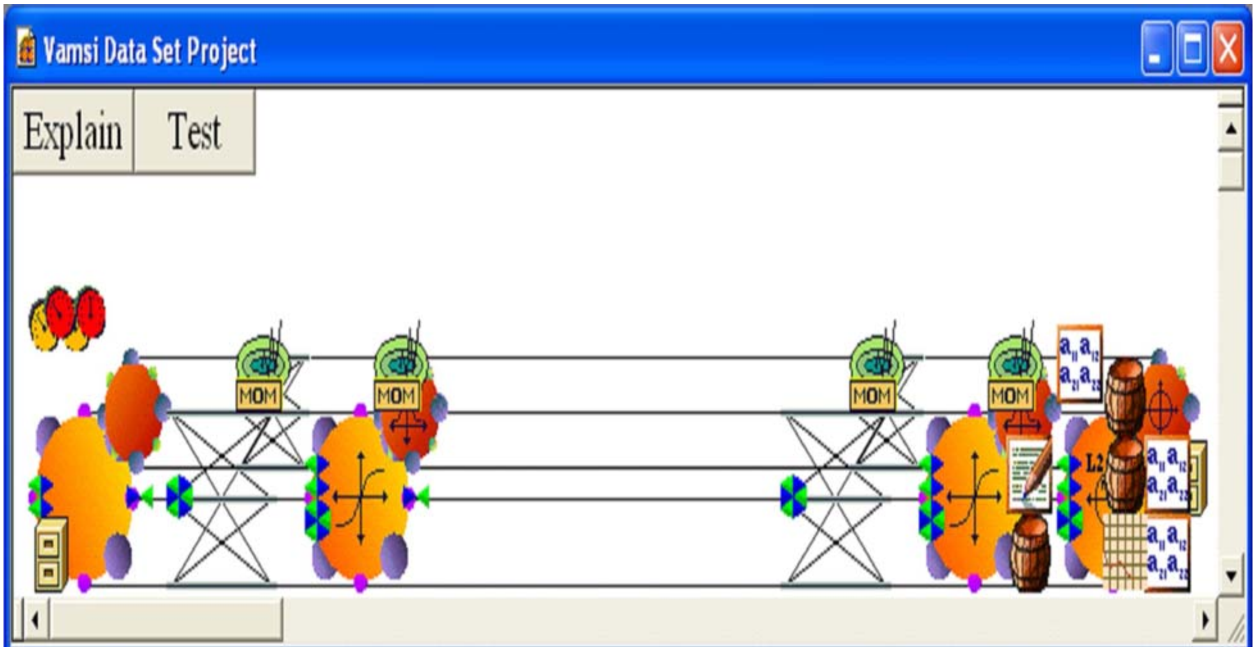
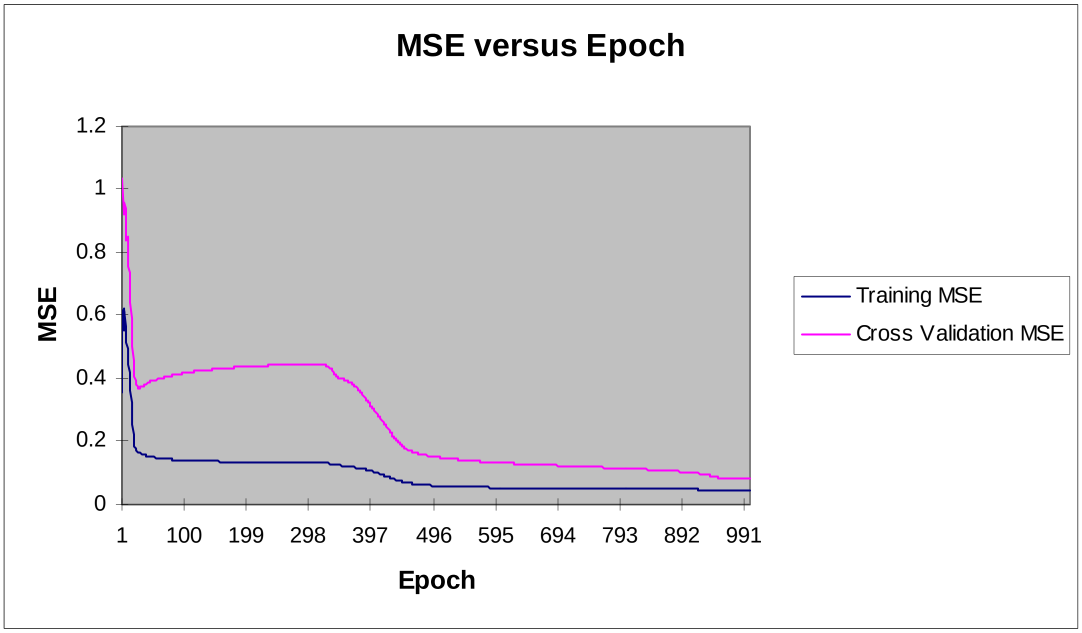


Fig 4: MSE Vs Epoch



Conclusion:

Structural prediction of the query sequence reveals the functionality of the protein. The prediction of secondary structure with reference to neural network works well and the analysis can concludes that it is possible to predict the secondary structure of any protein sequence by using this neural network. The further analysis on the neural network output describes the how the MSE and epoch can vary for the given set of data. The system is expected to enhance further in which it can accept the protein sequence directly and then predict the final secondary structure.

Reference:

[1] A. Kloczkowski,K.-L. Ting,R.L. Jernigan, J. Garnier,” Combining the GORVAlgorithmWith Evolutionary Information for Protein Secondary Structure PredictionFromAmino Acid Sequence. PROTEINS: Structure, Function, and Genetics 49:154–166 (2002).

[2] Pofahl, W. E., Walczak, S. M., Rhone, E., and Izenberg, S. D. (1998). Use of an Artificial Neural Network to Predict Length of Stay in Acute Pancreatitis. American Surgeon, Sep98, Vol. 64 Issue 9, (pp: 868 – 872).

[3] Lawrence, S., Giles, C. L., and Tsoi, A. C. (1997). Lessons in Neural Network Training: Training May be Harder than Expected. Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97, (pp. 540-545), Menlo Park, California: AAAI Press.

[4] Sarle, W. (1995). Stopped Training and Other Remedies for Overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, (pp. 352-360). Retrieved March 18, 2002. From World Wide Web: ftp://ftp.sas.com/pub/neural/

[5] Prechelt, L. (1998). Early Stopping-but when? Neural Networks: Tricks of the trade, (pp. 55-69).Retrieved March 28, 2002. from World Wide Web: http://wwwipd.ira.uka.de/~prechelt/Biblio/

[6] Wan Hussain Wan Ishak , “Notes on Neural Networks Learning and Training”from World Wide Web: http://www.generation5.org/content/2004/NNTrLr.asp.