JJBS

Jordan Journal of Biological Sciences

*Volume 10, Number 2,June 2017*
*ISSN 1995-6673*
*Pages 73 - 78*

# Neural Network Based Prediction of 3D Protein Structure as a Function of Enzyme Family Type and Amino Acid Sequences

Eyad M. Hamad[1*], Nathir A. Rawashdeh[2], Mohammad F. Khanfar[3], Eslam N. Al-Qasem[1], Samer I. Al-Gharabli[3**]

[1]*Biomedical Engineering Department, School of Applied Medical Sciences, German Jordanian University. Amman, Jordan*
[2]*Mechatronics Engineering Department, School of Applied Technical Sciences, German Jordanian University. Amman, Jordan*
[3]*Pharmaceutical and Chemical Engineering Department, School of Applied Medical Sciences, German Jordanian University. Amman, Jordan*

*Received: March 26, 2017 Revised: April 25, 2017 Accepted:May 3, 2017*

## Abstract

Prediction of dihedral angles from amino acid sequences based on the neural network to predict protein structure is promising in the field of bioinformatics. The present proposed study presents a prediction tool for 3-Dimensional (3D) protein structure as a function of enzyme family types and amino acid sequences. 11 different families of enzymes were investigated amounting to 97 enzymes in total. Correlation of sequence with geometry coordinates as a function of amino acid descriptors and family class were generated through a neural network to predict coordinates. The structural-features of dissected triplets show significant influence on family type. R-values for the inter-family dataset as well as feature selection were not satisfying. In contrast, R-values around 0.8 were achieved in the case of intra-family prediction. Furthermore, about 55 % of features were eliminated with a limited negative influence of 13% on the R-value. We believe that the present study provides a promising prediction method that advance computational methods in bioinformatics, especially to predict 3D protein structure as a function of enzyme family type and amino acid sequences. However, intra-family prediction probability is higher while using only one type of analysis based on the dihedral angles of turn structures of enzyme families.

**Keywords**: Neural Network, 3D Protein, prediction, Enzyme Family, Amino acid sequences.

## 1. Introduction

The complete genetic blueprint of a human being is now available for implementing new effective therapeutic strategies (Piccoli *et al*., 2013; Singh *et al*., 2013; Zini, 2005). Human DNA information is a powerful tool used to explore the role of genetic codes in pathogen formation and in the development of several diseases that form the majority of health problems worldwide, like cancer, diabetes, cardiovascular and others(Csermely *et al*., 2013; Mathkour and Ahmad, 2010; Oakley *et al*., 2008). This valuable trove of data has limitations in understanding higher order protein structures and particularly in translating protein molecular functionality from linear codes (Friedberg, 2006).

The knowledge of predicting three-dimensional structure of a protein can be used, on one hand, in drug design and in understanding biological mechanisms of protein function. X-ray, NMR, and, to some extent, electron microscopy are methods used to measure protein folding and surface topography. These methods however are limited in decoding the structure of many vital proteins classes (Pavlopoulou and Michalopoulos, 2011).

On the other hand, structure prediction from amino acid sequence requires the development of complex algorithms and is dependent on the millions of data points extracted experimentally to solve protein structures (Mills *et al*., 2015). In addition, algorithms should also be able to predict newly discovered or yet unrevealed structures (Kryshtafovych and Fidelis, 2009).

Deciphering algorithms of how protein structure is predicted as a function of primary sequence is no longer a purely academic problem, but can be used as a powerful method leading to effective drug design (Ahsanullah *et al*., 2012; El-Dahshan *et al*., 2014; Pavlopoulou and Michalopoulos, 2011).

Although the level of complexity between the primary sequence and final structure is relatively high, integrity and synchronization of protein building blocks, i.e., utilization of amino acid sequence to determine the folding process and the final 3D structure (Babu *et al*., 2011; Dokholyan, 2006; Liu *et al*., 2011). However, protein structure has been reported to be

[*] Corresponding author e-mail: eyad.hamad@gju.edu.jo.
[**] Corresponding author e-mail: samer.gharabli@gju.edu.jo

classified on three levels: primary, secondary, and tertiary structure (Zhang, 2009).

The first level consists of the sequence of amino acids making a linear structure, whereas the secondary structures depict the kinks and folding process where alpha helices and beta sheets are formed (Sikder and Zomaya, 2005). Although, the tertiary structure can be understood by the means of algorithms to be generated that includes the "turns" weave secondary structure fragments and lay the orientation of a whole ribbon in 3D space (Kryshtafovych and Fidelis, 2009).

One of the challenges in science is to predict the coordinate of these structure fragments (Grana *et al*., 2005; Liwo *et al*., 2011). In the present paper, "turns" that form the tertiary structure extracted from several enzyme families were investigated and correlated to structure of the description of the 3D structures of the studied proteins.

Correlation of sequence with geometry coordinates as a function of amino acid descriptors and family class were generated through a neural network tool to predict spatial configurations of the acids.

## 2. Methods

### 2.1. Database Mining

Eleven different families of enzymes were selected and investigated including EC 1.1.1.X (were X = 1, 2, 3, 8, 9, 10, 14, 17, 18, 21, or 22) with a grand total of 97 enzymes.

Structures were extracted from the Expert Protein Analysis System (ExPASy) a bioinformatics resource as well as from the protein data bank (Artimo *et al*., 2012). Secondary structures consisting of α-helices and β-strands were removed from the Protein Data Bank (PDB) file leaving only the turns.

These turns were recorded as a PDB file format and processed with the Ramachandran algorithm to calculate and assign phi, psi, and omega angles. The resulting dataset under investigation consisted of 17225 amino acid turn-examples in total, from 11 families, with their corresponding phi, psi, and omega angles. Each of the 20 amino acid types were assigned an identification number from 1 to 20, and saved in the neural network input feature vector as a descriptor, i.e., feature.

**Figure 1.** A schematic illustration of Homoserine Dehydrogenase enzymes where turns (loops) attach beta-sheets and alpha helices.

The present work's algorithm focuses on mapping the phi, psi, and omega angles of central amino acids in chains of 3 (triplets) in each of the 57 enzymes, i.e., amino acid chains, from the eleven families mentioned above.

The neural network input feature vector was constructed programmatically by scanning the enzyme chains for amino acid triplets, called (aa$i_{-1}$, $a_i$, and $aa_{i+1}$) as shown in Fig.2.

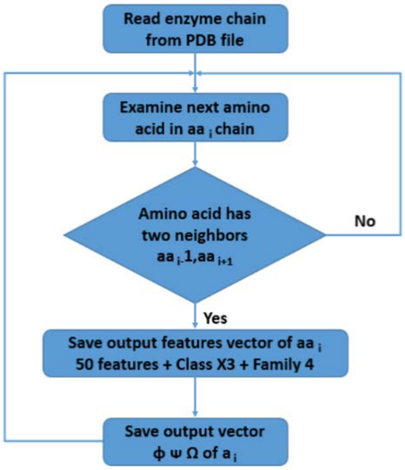
Read enzyme chain from PDB file
Examine next amino acid in aa $_i$ chain
Amino acid has two neighbors aa $_{i-}$1,aa $_{i+1}$
No
Yes
Save output features vector of aa $_i$ 50 features + Class X3 + Family 4
Save output vector ϕ ψ Ω of a $_i$

**Figure 2.** A Flow diagram data processing.

For each central amino acid $a_i$ found, a feature vector was constructed containing: Fifty descriptors of $a_i$; class number of $aa_{i-1}$; class number of $aa_i$; class number of $aa_{i+1}$; family number of enzyme chain containing the triplet.

The output vector representing the central amino acid $a_i$ is composed of the phi, psi, and omega dihedral angles of the amino acid $a_i$ inside the enzyme turn as illustrated in Fig.3. The fifty descriptors of the triplet center amino acid $a_i$ consisted of three groups: 15 electronic properties, 17 steric properties, and 18 hydrophobic properties which can be discussed later.

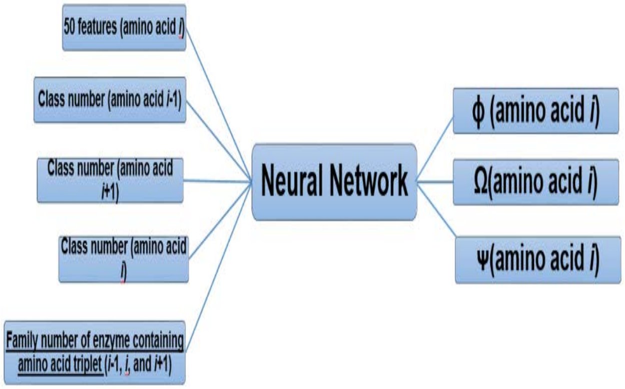
50 features (amino acid i)
Class number (amino acid i-1)
Class number (amino acid i+1)
Class number (amino acid i)
Family number of enzyme containing amino acid triplet (i-1, i, and i+1)
Neural Network
ϕ (amino acid i)
Ω(amino acid i)
Ψ(amino acid i)

**Figure 3.** A schematic illustration ofneural network configuration.

### *2.2. Artificial Neural Network*

The type of neural network (Bishop, 1995) implemented in the present work is a feed-forward Levenberg-Marquardt back propagation (Nawi *et al*., 2013), that illustrates using the gradient descent method with momentum weight and a bias learning function (Kumar and Minz, 2014). It consists of three layers of perceptrons, i.e., nodes: input; hidden; output. Each feature value in the input feature vector is connected to each of the input layer perceptron's by a multiplicative weight (Bishop, 1995).

In Fig. 4, the weights are implied parts of the arches shown. This means that the number of input layer nodes equals the length of the input feature vector, which was 53 or 54 depending on the experiment conducted.

Each hidden layer node processes the weighted sum of inputs to produce an output, which in turn feeds, via weight into each output layer node. In the present work, the number of hidden layer nodes was chosen to be equal to the number of input nodes. The output layer consists of three nodes producing three outputs, which are the angles phi, psi, and omega.

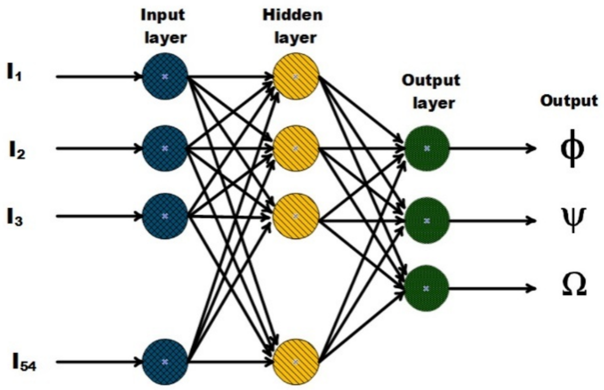

**Figure 4.** A schematic illustration of Feed-forward back-propagation neural network structure.

The data set was randomly divided into three parts: 60% for training the neural network; 20% for validating it and stop training before over fitting; and 20% for independent testing. The network performance was measured using the Mean Squared Error (MSE) (Kumar and Minz, 2014).

## 3. Results

### *3.1. Feature Descriptions*

Physical properties of amino acids are the foremost players in building the final or even the dynamic 3D structure of a protein. Classically, interactions were classified in three groups: electronic, steric, and hydrophobic.

Fifty features have been employed in the present work classified as follows: 15, 17, and 18 features represent electronic, the steric and the hydrophobic properties, respectively. Detailed description of these features can be found in the work conducted by Mei *et al.* (Mei *et al*., 2005).

The total number of amino acid features used to predict the three angles, i.e., 3D structure is 53. These include the 50 features described above in addition to the labels of the amino acid under investigation and its two sequence (before and after) neighbors.

### *3.2. Enzyme Family Mapping*

In order to validate the utility of the 53 selected features, a test was performed to sort out family types based on the mentioned features. Eleven family labels were used as output of the pattern recognition network. The dataset was composed of around 10 enzyme examples of each of the 11 family types. Each of the enzyme examples was processed to produce amino acid triplets with 53 features for each triplet's central amino acid, i.e., the neural network input vector. Thus, the neural network consisted of 53 inputs and 11 outputs.

The numbers of hidden layer nudes were 10 and the dataset was divided into three parts: 70% for training; 15 % for validation and prevention of over training; and 15 % for independent performance testing.

Performance was evaluated using two measures: Mean Square Error (MSE) between outputs and targets and the confusion matrix percentage of correct classification, these were 0.078 and 76.4 %, respectively.

### *3.3. Intra-Family Structure Prediction*

The original dataset was divided by family type and inside each family, neural networks were trained to predict the dihedral angles phi, psi, and omega of the central amino acid in the triplets based on 53 descriptors. Structural elements were predicted for each family in a separate training set.

Table 1 shows the regression coefficient (R) as a measure of alignment validation. In general, EC 1.1.1.X families where X = 1, 2, 8, 9, 10, 14, 17, and 18 shows an R-value above 0.5. EC 1.1.1.X with X=10 shows the highest value of 0.8 where the poorest value was recorded for X=21. The training regression is shown in Fig. 5 for the best Enzyme family.

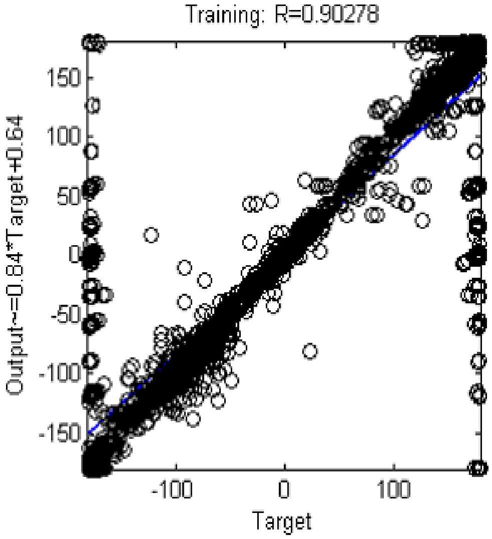

**Figure 5.** Training regression result for EC 1.1.1.10 with 53 features as input and 53 hidden layer nodes

**Table. 1**. Final results of neural network training before feature extraction (sorted by performance)

| Enzyme's family | R-value | Number of iterations | Training time (min) |
|---|---|---|---|
| EC 1.1.1.10 | 0.84147 | 131 | 7:22 |
| EC 1.1.1.9 | 0.79310 | 124 | 6:20 |
| EC 1.1.1 14 | 0.75090 | 132 | 6:20 |
| EC 1.1.1.1 | 0.74688 | 196 | 14:48 |
| EC 1.1.1.2 | 0.69972 | 95 | 4:52 |
| EC 1.1.1.8 | 0.55979 | 74 | 3:33 |
| EC 1.1.1.18 | 0.52505 | 89 | 6:34 |
| All Families | 0.50942 | 131 | 25:11 |
| EC 1.1.1.17 | 0.50326 | 74 | 3:44 |
| EC 1.1.1.3 | 0.48436 | 73 | 3:32 |
| EC 1.1.1.22 | 0.35053 | 52 | 2:30 |
| EC 1.1.1.21 | 0.31822 | 69 | 2:53 |

### *3.3.1. Inter-Family Structure Prediction*

Based on the intra-family result, additional training was performed where the family labels were added to the input features (making 54 features in total), and the amino acid triplets' structures were predicted across all families.

Results are summarized in row number eight in Table 1 that shows low performance, which may indicate the demand to extend the size of the input feature vector. In addition, the structure of the central element of the amino acid triplets varies with family type.

### *3.3.2. Feature Selection*

Feature selection can be defined as a process of feature-selection, or an applicant subset of features. In order to set the evaluation criteria, few feature subsets are used. The present study enables a promising method that predicts the dihedral angles of turn structures by the means of feed-forward Levenberg-Marquardt neural network. Feed-Forward selection enable finding weaker subset of features, due to the face that weaker features are not assessed while subset selection (Kumar and Minz, 2014).

Several neural networks have been trained to predict the structure of the central element of amino acid triplets across variable number of features and hidden layer nodes. All of the networks employ an input vector length of 54 features as in the intra family structure prediction described earlier.

Results are summarized in Table 2 and show no significant configuration that outperforms the base case of 54 with 50 hidden layer nodes. As and additional effort, the training parameters were modified for two cases to test whether the results could be improved.

The number of training epochs was raised from 1000 to 2000, and the failure checks were changed from 66 to 2000. The results were R = 0.33062 for 32 features and 50 hidden layer nodes, R = 0.47658 for 43 features and 50 hidden layer nodes.

In the first case, the performance (relative to 54 features and 50 nodes) decreased probably due to over fitting. In the second case, the performance is slightly worse than the base case, probably due to the elimination of 10 features.

**Table. 2.** Network performance with varying feature node and number (Shown is the total R-value)

| | No. of Features | | | | | |
|---|---|---|---|---|---|---|
| No. of Nodes | 5 | 16 | 22 | 32 | 43 | 54 |
| 25 | 0.46448 | 0.46619 | 0.46643 | 0.46893 | 0.46488 | 0.49363 |
| 50 | 0.47322 | 0.46858 | 0.47634 | 0.46514 | 0.47146 | 0.49657 |
| 100 | 0.47517 | 0.47678 | 0.47509 | 0.46804 | 0.28371 | 0.50777 |
| 200 | 0.42271 | 0.25977 | 0.46628 | 0.31304 | 0.44277 | 0.51903 |

The results of inter-family predictions in Table 2 indicate dependence of the amino acid triplet structures on the family type of which they belong. Thus, the potential of feature reduction was investigated from the EC 1.1.1.10 family only, using 53 features (with family label removed) and 50 hidden layer nodes. The R-value for this case was 0.80192. The objective was to reduce the number of features, without significantly lowering the R-value.

In order to reduce the number of features, the sum of absolute perceptron weights for each of the 53 inputs was computed. A threshold of 15 was chosen (as shown in Fig. 6) to discard all features with an absolute sum of feature weights below this threshold.

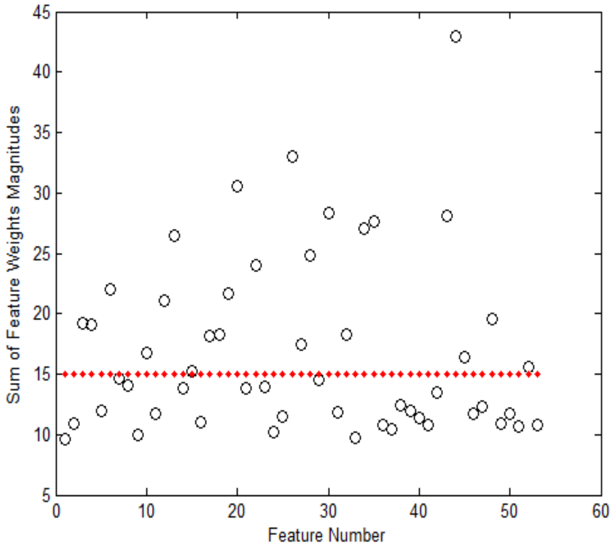

**Figure 6**. Feature extraction based on the sum of absolute perceptron weights for each of the 53 inputs where 15 was used as a threshold

The number of discarded, i.e., minor features, was 29, which is about 54.7% of the original 53 features. The R-value was 0.69366, which can be considered good because it represents a performance degradation of only about 13.5% when compared to the R value of 0.80192 for the 53 feature case.

The major features that remained are listed in the input vector:

[3, 4, 6, 10, 12, 13, 15, 17, 18, 19, 20, 22, 26, 27, 28, 30, 32, 34, 35, 43, 44, 45, 48, 52]

These major features are summarized descriptively as follows: 9 features were found in the hydrophobic properties including:

- solvation free energy
- Melting point
- Number of full nonbonding orbitals
- Retention coefficient in HPLC, *pH* 2.1
- Retention coefficient at *pH* 2,
- $R_f$ for 1-N-(4-nitrobenzofurazono)-amino acids in ethyl acetate/pyridine/water
- Hydration potential or free energy of transfer from vapor phase to water
- Log D
- Partition coefficient at pH 7.1 for acetylamide derivatives of amino acids in octanol/water
- $dG = ¼RT \ln f$ , where f = ¼fraction buried/accessible amino acids.

Other 10 features appeared in steric properties:

- Average volume of buried residue
- Residue accessible surface area in tri-peptide
- Normalized van der Waals volume
- Average accessible surface area
- Distance between $C_\alpha$ and centroid of side chain
- Side-chain angle
- Radius of gyration of side chain
- van der Waals parameter epsilon
- value of $\theta (i)$
- Substituent van der Waals volume.

And four features originated from electronic properties:

- Negative charge
- Polarity
- Net charge
- Electron-ion interaction potential values

Furthermore, one of the major features was the type of central amino acid in the sub group ($a_i$).

## 4. Conclusions

The present study gives a method to predict the dihedral angles of turn structures by feed-forward Levenberg-Marquardt neural network. The datasets for training and testing the network are PDBs of eleven different families of enzymes from Expert Protein Analysis System (ExPASy) and protein data bank. Secondary structures consisting of α-helices and β-strands were removed from the PDB file leaving only the turns.

A feature vector containing around 53 parameters was constructed for each central amino acid of amino acid triplets. This vector is used as input for neural network.

Ninety-seven enzyme families were selected and preconditioned to be an input vector for a feed forward back propagation neural network. The dihedral angles of only the turns in the 3D structures were predicted after training.

The structural features of dissected triplets show significant influence on family type. R-values for the inter-family data set as well as feature selection were not satisfying. In contrast, the R-values of about 0.8 were achieved in the case of intra-family prediction.

In addition, it is believed that the structural features of dissected triplets show significant influence on family type. About 55 % of features can be eliminated with relatively less negative influence of 13% on the R value. The present paper can provide promising useful prediction method that can advance the computational methods in bioinformatics, especially about the prediction of 3D protein structure as a function of enzyme family type and amino acid sequences.

To the best of our knowledge, many researchers have established various methods to predict protein structure. However, the intra-family prediction probability is higher when only one type of analysis based on the dihedral angles of turn structures of enzyme families is used. Therefore, biochemical experiments can be used for validation of the proposed prediction method that will enable reliable and shorter time experiments in the field of bioinformatics.

## Conflict of Interest Statement (COI)

The authors declare that they have no conflict of interest.

## References

Ahsanullah, Al-Gharabli, S. I. & Rademann, J. 2012. Soluble Peptidyl Phosphoranes For Metal-Free, Stereoselective Ligations In Organic And Aqueous Solution. *Organic Letters,* 14**,** 14-17.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., Duvaud, S., Flegel, V., Fortier, A. & Gasteiger, E. 2012. Expasy: Sib Bioinformatics Resource Portal. *Nucleic Acids Research***,** Gks400.

Babu, V., Uthayakumar, M., Vaishnavi, M. K., Senthilkumar, R., Shankar, M., Archana, C., Priya, S. S. & Sekar, K. 2011. Rps: Repeats In Protein Sequences. *J. Appl. Crystallogr.,* 44**,** 647-650.

Bishop, C. M. 1995. *Neural Networks For Pattern Recognition*, Oxford University Press.

Csermely, P., Korcsmaros, T., Kiss, H. J., London, G. & Nussinov, R. 2013. Structure And Dynamics Of Molecular Networks: A Novel Paradigm Of Drug Discovery: A Comprehensive Review. *Pharmacol Ther,* 138**,** 333-408.

Dokholyan, N. V. 2006. Studies Of Folding And Misfolding Using Simplified Models. *Current Opinion In Structural Biology,* 16**,** 79-85.

El-Dahshan, A., Al-Gharabli, S. I., Radetzki, S., Al-Tel, T. H., Kumar, P. & Rademann, J. 2014. Flexible, Polymer-Supported Synthesis Of Sphingosine Derivatives Provides Ceramides With Enhanced Biological Activity. *Bioorganic & Medicinal Chemistry,* 22**,** 5506-5512.

Friedberg, I. 2006. Automated Protein Function Prediction—The Genomic Challenge. *Briefings In Bioinformatics,* 7**,** 225-242.

Grana, O., Baker, D., Maccallum, R. M., Meiler, J., Punta, M., Rost, B., Tress, M. L. & Valencia, A. 2005. Casp6 Assessment Of Contact Prediction. *Proteins: Struct., Funct., Bioinf.,* 61**,** 214-224.

Kryshtafovych, A. & Fidelis, K. 2009. Protein Structure Prediction And Model Quality Assessment. *Drug Discov Today,* 14**,** 386-93.

Kumar, V. & Minz, S. 2014. Feature Selection. *Smartcr,* 4**,** 211-229.

Liu, T., Tang, G. W. & Capriotti, E. 2011. Comparative Modeling: The State Of The Art And Protein Drug Target Structure Prediction. *Comb Chem High Throughput Screen,* 14**,** 532-47.

Liwo, A., He, Y. & Scheraga, H. A. 2011. Coarse-Grained Force Field: General Folding Theory. *Phys. Chem. Chem. Phys.,* 13**,** 16890-16901.

Mathkour, H. & Ahmad, M. A Comprehensive Survey On Genome Sequence Analysis. In Bioinformatics And Biomedical Technology 2010. Institute Of Electrical And Electronics Engineers, 14-18.

Mei, H., Liao, Z. H., Zhou, Y. & Li, S. Z. 2005. A New Set Of Amino Acid Descriptors And Its Application In Peptide Qsars. *Peptide Science,* 80**,** 775-786.

Mills, C. L., Beuning, P. J. & Ondrechen, M. J. 2015. Biochemical Functional Predictions For Protein Structures Of Unknown Or Uncertain Function. *Computational And Structural Biotechnology Journal,* 13**,** 182-191.

Nawi, N. M., Khan, A. & Rehman, M. 2013. A New Levenberg Marquardt Based Back Propagation Algorithm Trained With Cuckoo Search. *Procedia Technology,* 11**,** 18-23.

Oakley, M. T., Barthel, D., Bykov, Y., Garibaldi, J. M., Burke, E. K., Krasnogor, N. & Hirst, J. D. 2008. Search Strategies In Structural Bioinformatics. *Curr. Protein Pept. Sci.,* 9**,** 260-274.

Pavlopoulou, A. & Michalopoulos, I. 2011. State-Of-The-Art Bioinformatics Protein Structure Prediction Tools (Review). *Int. J. Mol. Med.,* 28**,** 295-310.

Piccoli, S., Suku, E., Garonzi, M. & Giorgetti, A. 2013. Genome-Wide Membrane Protein Structure Prediction. *Curr. Genomics,* 14**,** 324-329.

Sikder, A. R. & Zomaya, A. Y. 2005. An Overview Of Protein-Folding Techniques: Issues And Perspectives. *Int. J. Bioinf. Res. Appl.,* 1**,** 121-143.

Singh, A. K., Bhargava, A., Kaur, G., Sharma, A. & Misra, K. 2013. *Bioinformatics Tools And Resources For Cancer Diagnosis And Drug Development,* Nova Science Publishers, Inc.

Zhang, Y. 2009. Protein Structure Prediction: When Is It Useful? *Curr. Opin. Struct. Biol.,* 19**,** 145-155.

Zini, G. 2005. Artificial Intelligence In Hematology. *Hematology,* 10**,** 393-400.