

Homology Modeling of Myoglobin using Adaptive Neurofuzzy Systems

Achuthsankar S. Nair,
Hon. Director, Centre for Bioinformatics,
University of Kerala,
Karyavattom,
Trivandrum -695581,
Kerala
India
Email: sankar.achuth@gmail.com
Tel: 91-471-2542220

Koshy P. Vaidyan,
Tata Consultancy Services,
Technopark,
Karyavattom,
Trivandrum - 695581,
Kerala
India
Email: koshypvaidyan@yahoo.com
Tel: 91-471-2312551

Keywords: Protein Tertiary Structure Prediction, Neurofuzzy Systems, ANFIS, Nonlinear System Identification

Abstract

The problem of nonlinear system identification as applied to protein folding problem is discussed in this paper. The mapping of amino acid sequence and the atomic coordinates of the alpha carbon atoms of the amino acids in a protein is typically a nonlinear problem. System identification is performed using adaptive neurofuzzy techniques. Various ANFIS models are created and the model with the least error is selected. The protein 'Human Myoglobin Mutant (PDB Id: 2MM1)' and its homologue 'Pig Metmyoglobin (PDB Id: 1MYH)' have been used for the creation and training of the ANFIS model. The tertiary structure of 3 proteins 'Myoglobin (Horse Heart) wild type complexed with nitrosoethane (PDB Id: 1NPG)', 'Loggerhead sea turtle Myoglobin (PDB Id: 1LHS)' and 'MetMyoglobin from Yellowfin Tuna (PDB Id: 1MYT)' have been predicted using the same ANFIS model. It is found that the root mean square errors in the prediction of the tertiary structures of the 3 proteins considered in this study are 4.32, 3.04 and 3.00 respectively.

1. Introduction

The problem of determining the mathematical model of an unknown system by observing its input-output data is generally referred to as System Identification [1]. This method is very widely used in a number of practical applications in the areas of Communications, Control Systems, Signal Processing, Chemical Process Control, and Biological Processes. Systems can be broadly classified as linear and nonlinear. Strictly speaking, linear systems do not exist in practice, since all physical systems are nonlinear to some extent [2]. Linear Systems are idealized models developed for the simplicity of design, and there exist a large number of analytical and graphical techniques for their design. Nonlinear systems, on the other hand, are difficult to treat mathematically, and there are no general methods available for solving a wide class of nonlinear systems [2]. However, most systems can be better described by nonlinear models which are able to depict the behaviour of the system over the entire operating range, whereas linear models approximate the behaviour of the system around a given operating point. In the last few decades, a considerable amount of research has been conducted on the modeling and identification of nonlinear systems. Most of the methods of nonlinear system identification are based on parametrized nonlinear models such as Wiener-Hammerstein models, Volterra series, Wavelet Networks, Neural Networks, etc. In these, the parameter estimation can be performed using nonadaptive techniques such as least squares methods and higher order statistics-based methods, and also adaptive techniques such as the backpropagation algorithm and adaptive gradient learning [3].

The Protein Folding Problem can be visualized in the framework of Nonlinear System Identification. The hypothesis is that the function of the protein is determined by its structure, and this, in turn, is determined by the amino acid sequence of the protein. In this paper, the authors attempt to develop a system, which, by

taking the amino acid sequence of a protein as the input, provides, as indicated in Fig. 1, the atomic coordinates of the alpha carbon atoms of the amino acids in the folded structure, as the output. The system is to be constructed from the input-output data, which is obtained from the Protein Data Bank (PDB). This system, which predicts the protein structure, should have discrete signals at the input and output to facilitate the application of traditional signal processing and nonlinear system identification techniques. This approach of converting symbol sequences to digital sequences has been referred to as genomic signal processing in genomic studies. Filtering of such genomic digital signals have been used to characterize genomes and locate coding regions [4], [5], [6]. This work also draws inspiration from this approach.

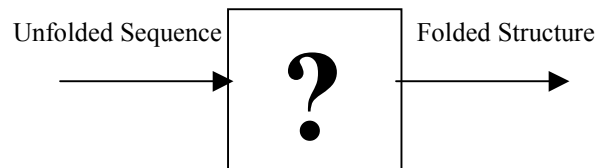


Fig. 1. System to Predict Protein Structure

The input to the system, which is the amino acid sequence of the protein, is to be converted to a discrete signal. The symbol to discrete signal mapping can be done using the various physico-chemical properties of amino acids, some of which are given in Table 1. A similar method for transforming an amino acid sequence into a numerical sequence is elaborated in [10], where each amino acid is represented by the value of Electron Ion Interaction Potential (EIIP).

Table 1: Physico-Chemical Properties of Amino Acids

Amino Acid	Water-Oil Ratio	Molecular Weight	EIIP	Alpha Propensity	Beta Propensity
A	-1.60	89	0.0373	1.41	0.72
C	-2.00	121	0.0829	0.66	1.40
D	9.20	132	0.1263	0.99	0.39
E	8.20	146	0.0058	1.59	0.52
F	-3.70	165	0.0946	1.16	1.33
G	-1.00	75	0.0050	0.43	0.58
H	3.00	156	0.0242	1.05	0.80
I	-3.10	131	0.0000	1.09	1.67
K	8.80	147	0.0371	1.23	0.69
L	-2.80	131	0.0000	1.34	1.22
M	-3.40	149	0.0823	1.30	1.14
N	4.80	132	0.0036	0.76	0.48
P	0.20	116	0.0198	0.34	0.31
Q	4.10	146	0.0761	1.27	0.98
R	12.30	175	0.0959	1.21	0.84
S	-0.60	105	0.0829	0.57	0.96
T	-1.20	119	0.0941	0.76	1.17
V	-2.60	117	0.0057	0.90	1.87
W	-1.90	203	0.0548	1.02	1.35
Y	0.70	181	0.0516	0.74	1.45

This paper attempts to predict the tertiary structure of the protein. The methods of prediction can be generally classified into two categories: Comparative Modeling and Ab-initio Modeling. Of these, the first is based on known homologous structures and currently, the most accurate and reliable 3-D Structure

Prediction is based on this type of modeling. The method specified in this paper comes under the category of Comparative Modeling, and it visualizes protein tertiary structure prediction as a nonlinear system identification problem. This method, which is expected to provide good insight into the system, may be a good candidate for investigation in protein folding, so that better prediction accuracy may be obtained.

2. Methods for Nonlinear System Identification

Some of the currently used techniques for nonlinear system identification include Fuzzy Logic, Neural Networks, Genetic Algorithms etc. Fuzzy Systems, which are universal approximators, can be used to model nonlinear systems. Fuzzy Estimation and Identification has been discussed in detail in [7]. Training Fuzzy Systems to perform system identification is described in [9], and the use of Takagi-Sugeno type fuzzy systems to represent nonlinear systems is elaborated in [7]. Likewise Neural Networks can also be used for modeling systems that are nonlinear [3], and have shown excellent capability as compared to classical methods. Neurofuzzy systems combine the advantages of fuzzy systems (e.g. the ease of incorporating expert knowledge), and those of neural networks (e.g. the learning capability)[1],[8]. Nonlinear System Identification, using neurofuzzy systems such as Adaptive Network based Fuzzy Inference System (ANFIS), is elaborated in [1].

ANFIS has five layers [1], [8]. In layer 1, every node is a square node $O_i^1 = \mu_{A_i}(x)$ where x is the input to node i and A_i is the linguistic label (small, large etc) associated with this node function.

$$\mu_{A_i}(x) = \frac{1}{1 + \left[\left(\frac{x - c_i}{a_i} \right)^2 \right]^{b_i}} \quad (1)$$

The second layer which consists of circle nodes multiplies the incoming signals and sends the products out. For instance,

$$w_i = \mu_{A_i}(x) * \mu_{B_i}(y), \quad i = 1, 2, \dots \quad (2)$$

Each node output represents the firing strength of the rule. The outputs of all these nodes are summed and the reciprocal of this value is taken.

Every node in the third layer is a circle node. The i^{th} node calculates the ratio of the i^{th} rule's firing strength to the sum of all the rule's firing strength.

$$\bar{w}_i = \frac{w_i}{\sum w_i}, \quad i = 1, 2, \dots \quad (3)$$

The outputs of this layer are called normalized firing strengths.

Every node i in the fourth layer is a square node with a node function.

$$O_i^4 = \bar{w}_i * f_i = \bar{w}_i * (p_i x + q_i y + r_i) \quad (4)$$

where, w_i is the output of the layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set. Parameters in this layer are referred to as 'consequent parameters'.

The single node in the fifth layer computes the output as the summation of all incoming signals.

$$O_i^5 = \text{overall input} = \sum_i \bar{w}_i * f_i = \frac{\sum_i w_i * f_i}{\sum_i w_i} \quad (5)$$

ANFIS has two sets of parameters: the premise parameters and consequent parameters, as already discussed. [1],[8]. A hybrid learning algorithm is used, where the consequent parameters are evaluated using Least Squares Estimate (LSE) in the forward pass and the premise parameters are evaluated using the Gradient Descent method in the backward pass.

Different ANFIS models representing the system are created and the root mean square error of the prediction results is evaluated and the model having the least root mean square error is selected.

3. Implementation

The proteins considered in this study are given in the table below.

Table 2: Proteins considered in the study

S.No	Protein	PDB Id	Remarks
1	Human Myoglobin Mutant	2MM1	Used for Training
2	Pig Metmyoglobin	1MYH	Used for Training
3	Myoglobin (Horse Heart)	1NPG	Used for Prediction
4	Loggerhead sea turtle Myoglobin	1LHS	Used for Prediction
5	MetMyoglobin from Yellowfin Tuna	1MYT	Used for Prediction

The protein ‘Myoglobin (Horse Heart) (PDB Id: 1NPG)’ has more than 80% similarity and ‘Loggerhead sea turtle Myoglobin (PDB Id: 1LHS)’ has similarity between 50 to 80% and ‘MetMyoglobin from Yellowfin Tuna (PDB Id: 1MYT)’ has less than 50% similarity with ‘Human Myoglobin Mutant (PDB Id: 2MM1)’

The signal representation of the amino acids using Water Oil Ratio mapping becomes the input sequence For example, the amino acid sequence of one of the proteins used for training is given below.

PDB ID: 2MM1 (HUMAN MYOGLOBIN MUTANT)

GLSDGEWQLVLNVWVGKVEADIPGHGQEVLRIRLFKGGHPETLEKFDRLFHKLKSEDEMKASEDLKHH
GATVLTALGGILKKGHHEAEIKPLAQSHATKHKIPVKYLEFISEAIIQVLQSKHPGDFGADAQGA
MNKALELFRKDMASNYKELGFQG

This is converted to a signal by the values of ‘Water Oil Ratio’ given in Table1 and is shown in fig.2.

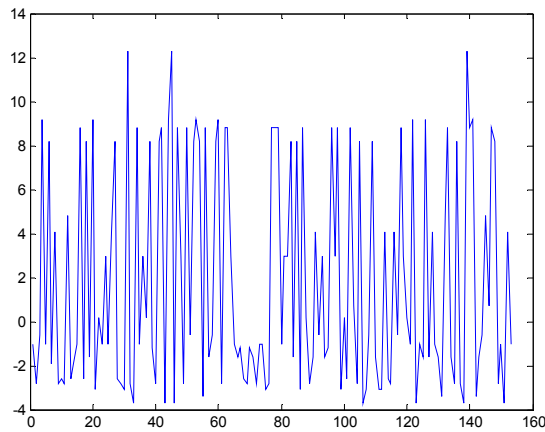


Fig.2 Representation of the Input Signal

The atomic coordinates of the alpha carbon atoms of the amino acids, as already stated, becomes the output sequence. Initially the x coordinates only are considered. The work can be extended to y coordinates and z coordinates as well.

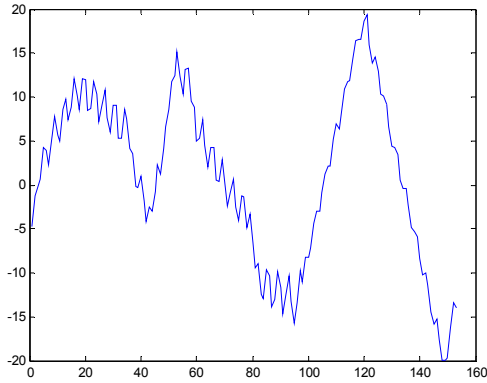


Fig.3. Representation of the output signal

The x coordinates of the alpha carbon atoms of the amino acids are completely independent of the type of amino acid. There is no correlation between the input sequence and the output sequence. Since the principle of superposition does not hold good, the system to be identified can be considered as a nonlinear system and nonlinear system identification techniques can be applied.

ANFIS is used for system identification and there is a limitation in the number of inputs to ANFIS since the computational complexity is to be reduced. The input sequence is considered as the vector $u[k]$ and the output sequence is indicated by the sequence $y[k]$, where $k = 1, 2, 3, \dots$. In this study, ANFIS is assumed to have 3 inputs.

We assume that there are 10 input candidates. Hence, 3 inputs are to be selected from 10 input candidates. For dynamical system identification, the inputs should not come from either of the following two sets of input candidates exclusively:

$$Y = \{y(k-1), y(k-2), y(k-3), y(k-4)\}$$

$$U = \{u(k-1), u(k-2), u(k-3), u(k-4), u(k-5), u(k-6)\}$$

A reasonable guess would be to take two inputs from Y and one from U to form the inputs to ANFIS. There are 36 ways of making this selection. Each of the models are evaluated based on the prediction results and the Root Mean Square Error (RMSE) is calculated for the prediction results for each model. The ANFIS model with the least error is selected.

The simulation is done using Simulink and the block diagram is given below.

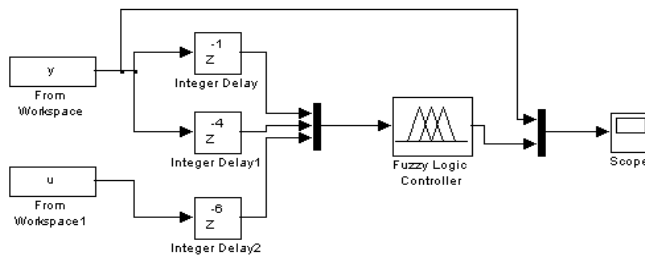


Fig.4 Simulink Block Diagram for checking Prediction Results

4. Prediction Results

For the proteins ‘Myoglobin (Horse Heart) (PDB Id: 1NPG)’, ‘Loggerhead sea turtle Myoglobin (PDB Id: 1LHS)’, ‘MetMyoglobin from Yellowfin Tuna (PDB Id: 1MYT)’, the training and prediction results are shown in figs. 5(a), (b) and (c).

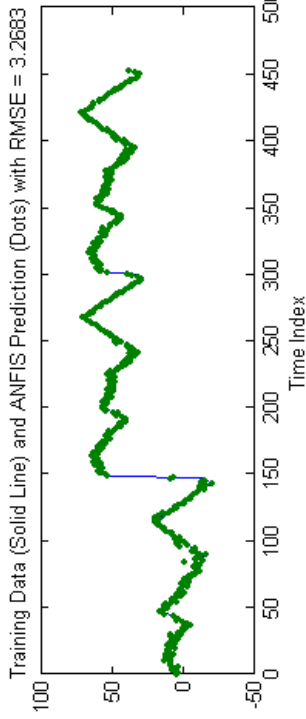


Fig.5(a) Training and Prediction Results for the protein with PDB Id '1NPG'

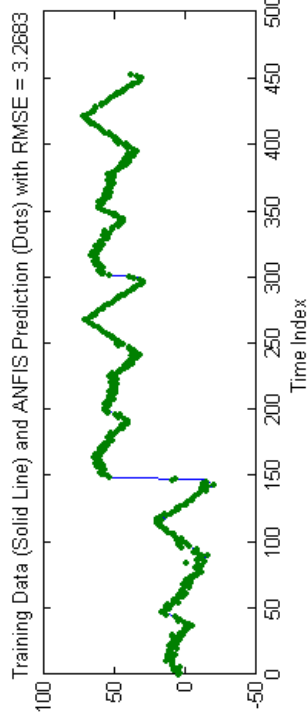


Fig.5(b) Training and Prediction Results for the protein with PDB Id '1LHS'

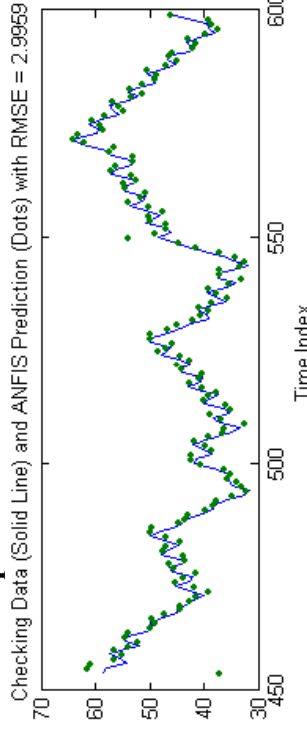
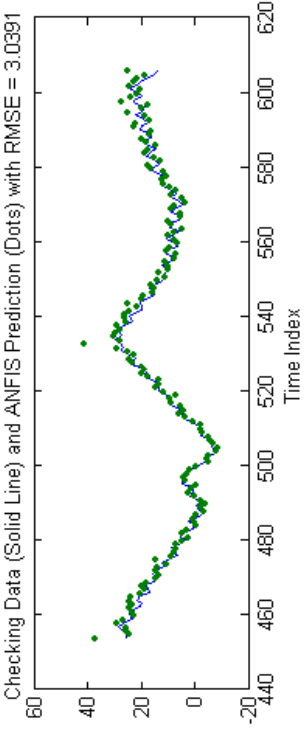
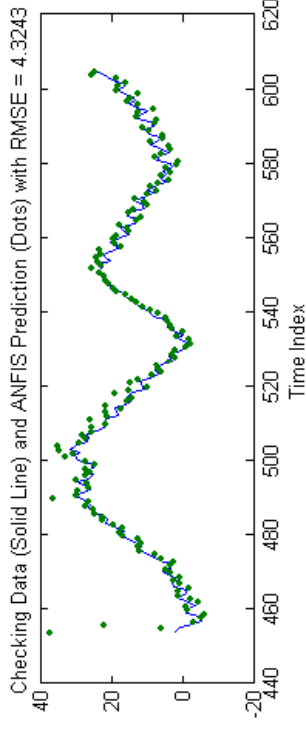
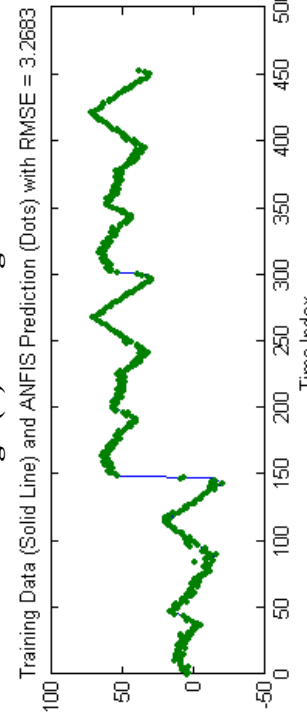


Fig.5(c) Training and Prediction Results for the protein with PDB Id '1MYT'

5. Discussions and Conclusion

The prediction results are summarized in the table given below. The similarity of proteins is also indicated.

Table 3: Summary of Prediction Results

S.No.	Protein	PDB Id	Similarity with Human Myoglobin	Root Mean Square Error (Prediction)	Remarks
1	Myoglobin (Horse Heart)	1NPG	Above 80%	4.3243	Indicated in Fig.5(a)
2	Loggerhead sea turtle Myoglobin	1LHS	Between 50 and 80%	3.0391	Indicated in Fig. 5(b)
3	MetMyoglobin from Yellowfin Tuna	1MYT	Below 50%	2.9959	Indicated in Fig. 5(c)

A nonlinear system, which performs the mapping between the amino acid sequence and the atomic coordinates, has been identified using ANFIS. This system has been used for the learning and prediction of protein folding. The prediction of tertiary structures of 3 proteins has been made with an accuracy of RMSE less than 4.4 as indicated in the results. The neurofuzzy system is also able to predict, even when there is a sharp turn in the coordinates, which indicates very good prediction capability. The input signal is derived from the amino acid by considering one of the properties of amino acids. One of the authors' strong hypothesis is that the efficiency of the model will increase with the biological relevance of the parameter chosen. The authors are hence continuing with such an exploration to see if they can further enhance the prediction results summarized above. In the present study, all the proteins selected belong to the Myoglobin family. Extensive studies are required to identify the prediction capability when the proteins belong to entirely different families. The authors also strongly feel that if more number of homologous proteins are used for learning, the prediction results will improve.

References

- [1] J.S.R Jang, C.-T.Sun, E. Mizutani, "Neurofuzzy and Soft Computing", Prentice Hall of India Pvt. Ltd., 1997.
- [2] Benjamin C. Kuo, "Automatic Control Systems", Sixth Edition, Prentice Hall of India Pvt. Ltd., 1993.
- [3] Mohamed Ibnkahla, "Statistical Analysis of Neural Network Modeling and Identification of Nonlinear Systems with Memory", IEEE Transactions on Signal Processing, Vol. 50, No. 6, June 2002.
- [4] Achuthsankar S. Nair, Mahalakshmi T, "Visualization of Genomic Data using Inter-Nucleotide Distance Signals", GSP 2005 – International Conference on Genomic Signal Processing, Romania, July 2005.
- [5] Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R., "Prediction of probable genes by Fourier analysis of genome sequences", CABIOS, Vol.113, pp.263-270,1997.
- [6] Vaidyanathan P.P., Yoon B.J., "The role of signal processing concepts genomics and proteomics", Journal of Franklin Institute, Special Issue on enomics, 2004.
- [7] Kevin M. Passino, Stephen Yurkovich, "Fuzzy Control", Addison Wesley, 1998.
- [8] Jyh-Shing Roger Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System", *IEEE Transactions on Systems, Man and Cybernetics*, Vol.23, No.3., May/June 1993.
- [9] Laukonen E.G., Passino K.M., "Training Fuzzy Systems to Perform Estimation and Identification," *Engineering Applications of Artificial Intelligence*, Vol. 8, No. 5, pp.499–514, 1995.
- [10] Irena Cosic, "Macromolecular Bioactivity: Is It Resonant Interaction Between acromolecules? – Theory and Applications", IEEE Transactions on Biomedical Engineering, Vol.41, No.12, December 1994.