International Journal of Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-2, Issue-4)

Research Article | **April 2013**

Protein Family Recognition based on Fuzzy Logic

Bassam M. El-Zaghmouri *
Computer Information Systems Dept
Jerash University, Jordan

Marwan AL-abed Abu-zanona
Department of Computer Science
Imam Muhammad Ibn Saud Islamic University, KSA

Abstract- In the rise rabid research related to biometrics, bio-informatics and genome; many researches, fields, and issues are still undergoing any uncertainties. One of the hottest areas in this field of research is the proteins informatics, that is relates the protein data with the modern information technology and it includes portions mapping and classification. This paper contributes an intelligent system which consists of adaptive neuro-fuzzy computations that is able to recognize and classify the proteins in families. An intelligent trainer will be structured based on Perceptron neural network in order to build an intelligent fuzzy inference system that is capable of predicting and classifying that data into categories according to the function of each protein. The structured system preprocesses that data set and extracts unique features from it. The system was built using a highly developed programming language. This paper will clearly show the results that such system achievement about 92% of accuracy when over 1000 inputs sequence of the validation sample was processed.

Keywords: Pattern, Fuzzy Logic, Estimation, Protein Recognition, Genome.

1. INTRODUCTION

The characterization of protein functions could be processed in many approaches. Generally, unknown proteins are being considered as functions of infer protein whose sequences are based on similarities that could make proteins annotating in the data set. That process is ambiguous and considered to be very complex. The vast annotation or genome was propagated across other data sets of molecular. Those data sets are being used to identify the protein family in contrast (Marcotte, *et a*l., 2001). Proteins clustering (family recognition) results in important clues architecture, role of metabolic and activities. The clustering of proteins due to their families returns many benefits for genomes of large scale annotation (i.e. modification of the proteins identification which is very hard to characterize, helps to keep the data set family based promoting propagation and annotating. Also, it ensures efficient meaning of biological relevant information retrieval out of large data sets, and gets a scope of underlying families of genome) (Wu, *et a*l., 2002) . The DNA, RNA and all proteins are organized in sequence forms. The sequence determines the similarities between those proteins. When transforming part of the sequence or even overall sequence, some distortion may occur, this makes the sequence recognition more difficult (Marcotte, et al., 2001).

The proteins sequencing consists of twenty amino-acids, where the determination of the sequence of those amino acids is commonly known as protein sequence. The protein structure and functions clustering in specified living organisms represent an important stage of many medical processes, such as the internal process working of the cell. For example, it is important in diagnosing patient case, bio analysis and medical studies (Wu, *et a*l., 2002). The modern computer systems that deal with specific information (informatics) represent a basic block of each engineering, scientific, accounting, financial, political, and other issues. The use of informatics made a hope in all fields of science and humanities. One important field of using informatics is the use of biology and medical information in computerized systems to increase the reliability of those systems. The fact that humans should adapt and keep up to date with all tools and techniques to make their process and job easier, accurate, precise, fast, and repetitive , leads to promoting the use of computer information systems in wide fields of applications where, no such criteria could be obtained without informatics (Ashburner, *et a*l., 2000) Two methods are being used traditionally by the researchers to sequence the proteins; Edman degradation and mass spectrometry reactions. The use of intelligent classifier benefits in gaining more precise data from the constrained data and information, which can be handled as preliminary data (Wolstencroft, *et a*l., 2007) Many classifiers were designed to automatically distinguish the protein sequence family. Some of them work in special cases or under constraints. Some of them have limitations in data size and accuracy (Wu, 2003). Some of them cause problems in memory and processing power (Wolstencroft, 2007). This leads to the need of designing a suitable computer system that is capable of handling large numbers of data in a relatively fast time, accurate result, and generalized for common protein sequences (Hoff, 2007). Some proteins can be classified within more than one family based on their behavior or sequence. And over all proteins sequence length is not always equal. The contribution of this thesis are as follows using one adaptive system that can handle the suggested protein families; handle the variable length of sequences, retreat the complex data structure and extract the unique features from it using wavelet transformation techniques. Since it is difficult to estimate the functions of proteins due to their large similarity and structure, the computer information adaptation is required in such cases. That motivation and the fact that no accurate research can recognize and classify protein sequence in right way led to the start of this research (Wolstencroft, 2007). Specially, when adapting an intelligent neuro-fuzzy inference system, the accuracy that could be achieved should be better than any other classical computing or even intelligent single performance intelligent predictor / classifier (Russell, 2005). Actually, combining neural network solid training algorithms with the high performance fuzzy logic classifier will build a complete

neuro-fuzzy inference system that takes historical data of proteins sequence and convert it to fuzzy membership functions. When running the system, it will convert any sequence protein data inputs, into an output in the form of numbers to be able to classify the protein families according to their functions. This paper studies the implementation of a fuzzy system that is adaptive and capable to classification the input protein sequences in clear known protein families, in order to achieving high accuracy of estimation. This will be done behind the dealing with large number of data under memory constraints and an efficient program. The twenty amino acids that construct the protein are represented in table1 compressed by the use of annotation of IUPAC (International Union of Pure and Applied Chemistry).

Table (1): Table of Amino Acids

Amino acid	Codons	Compressed	Amino acid	Codons	Compressed
Ala/A	GCU, GCC, GCA, GCG	GCN	Leu/L	UUA, UUG, CUU, CUC, CUA, CUG	YUR, CUN
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	CGN, MGR	Lys/K	AAA, AAG	AAR
Asn/N	AAU, AAC	AAY	Met/M	AUG	
Asp/D	GAU, GAC	GAY	Phe/F	UUU, UUC	UUY
Cys/C	UGU, UGC	UGY	Pro/P	CCU, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	Ser/S	UCU, UCC, UCA, UCG, AGU, AGC	UCN, AGY
Glu/E	GAA, GAG	GAR	Thr/T	ACU, ACC, ACA, ACG	ACN
Gly/G	GGU, GGC, GGA, GGG	GGN	Trp/W	UGG	
His/H	CAU, CAC	CAY	Tyr/Y	UAU, UAC	UAY
Ile/I	AUU, AUC, AUA	AUH	Val/V	GUU, GUC, GUA, GUG	GUN
START	AUG		STOP	UAA, UGA, UAG	UAR, URA

2. LITERAL REVIEW

The current researches face the problem of uncertainty in the large scale data of portions and protein sequences. This problem necessitates that the recognition and classification be artificially intelligent and not analytical. Most researchers concentrate on classical methods for classification (i.e. statistical analysis) not artificial intelligent techniques problem solving. A good work was presented in (Wu, *et a*l., 2002)where the genomic sequence accumulation acceleration of bio data had been represented by the pressing need to implement statistical methods and advanced infrastructure bioinformatics for a large data set and efficient annotating of the protein and discovery of biologic knowledge discovery. The problem of annotation of the protein has been driven by classification and was addressed to generate rules to handle the attribution of evidence. The rules are connected with an integrated expert system to cover different behaviors of the protein. The authors presented the knowledge in that research to be combined of two databases; one is analysis of sequence tools, while the other is the graphical interface. This research was supposed to verify the errors of identification sensitivity, consistency, systematic errors of annotating in detection, rich annotating errors, and the errors of experimental distinction. However, we cannot completely handle those errors because of the expert system complexity, large computational power and the high consumption cost. In 2001, (Marcotte, *et al* ,2001) presented an approach that received some attention that helped modify the previous works. It extracted the interactions of protein-protein and found a solution of relationships between them. Marcotte named these statistical approaches named "bag of words." But the problem of this paper is that, it assumed very limited data sets. In (Hoff, 2007) the authors showed that the modeling of the relations can be done by multiplications of the factors node-wise. The generalized multiplicative modeling is a latent-class models which could encode the relationships of transitivity.

A large molecule made from a long chain consists of smaller components known as nucleotides is the Deoxyribo Nucleic Acid (DNA). Each sequence of DNA is constructed from an amino acids sequence, whereas, that sequence has a function performance code known as Ribo Nucleic Acid (RNA) (Ashburner, *et al*,2000) The DNA is holding the information over all its structure in gene structure segments. The functions of protein sequence are determined by all genes. Those functions vary from building the basic structures to controlling reactions, (Wolstencroft, *et al*,2007)

3. METHODOLOGY

Today, Fuzzy Logic applications are employed in numerous commercial devices; including control systems, estimators, forecasters, etc. Therefore, it has many benefits in comparison with the conventional estimators, controllers, and classifiers. The Fuzzy Logic is Multi Input Multi Output (MIMO) Systems (Xu, *et al*., 2006).

This paper adapts a suggeno-type fuzzy logic system in order to recognize or classify the input matrix which is entered as features form. This fuzzy system consists of five input variables each one is considering a step membership function. Those input variables are obtained from preprocessing of the bulk historical data.

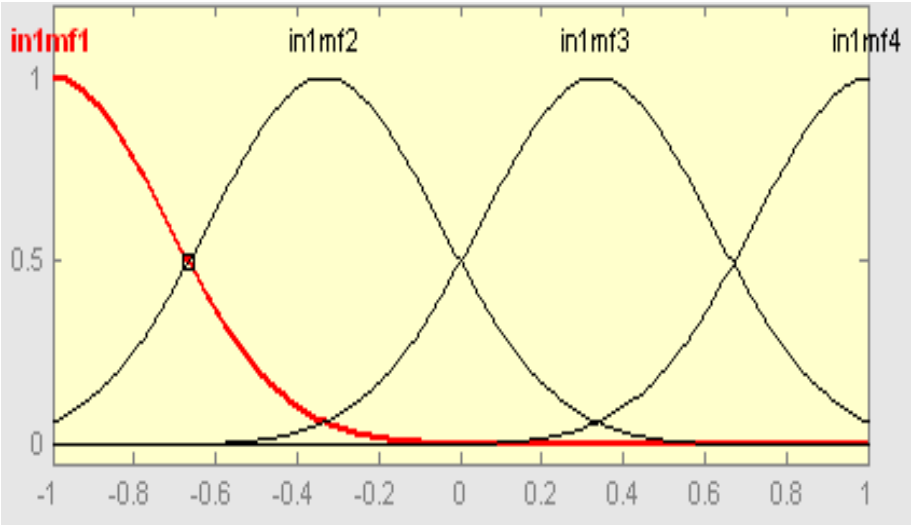

Figure (2) internal structure of membership functions of the input variables

The fuzzy inference system (FIS) that is generated after training is then ready to use in Running Mode. While running, the coefficients that are generated by data that is preprocessed in the designed module are entered to the FIS and thus, it will output the estimated ID that represents the Speech Segments ID's. The use of fuzzy logic gets the many advantages over the use of single neural network. It comprises minimal structure which proposes less memory, fast anticipation and estimation, ease to build and modify, also, it could model the uncertainties in the membership functions and rules. Although the neural networks have many advantages over the fuzzy logic; it contains more accuracy and precise outputs over different input / output ranges. Another benefit of the neural network over the fuzzy logic is the validity over data out of training range. The fuzzy logic anticipates the data that comes out of the range of input data for training, and estimates it to the max input value by normalization. Whereas, the neural network applies all bias and weight arrays and estimates the outputs directly, this makes the output more meaningful and increases the precision. Two stages of preprocessing of data are adapted in this paper. The first stage is handling the data arrangement and organization in an acceptable form to be used as input vectors for the neural network.

The second stage is normalization of the data. The normalization is a process to make the data shares specified references (minimum value and maximum value). The normalization is being done using the equation 1.

$$Xout = \frac{\text{Xin}}{\text{Xmax}} \qquad (1)$$

Where Xmax is being calculated from equation 2

$$Xmax = \sum_{k=0}^{n} \mathrm{Xk} \qquad (2)$$

Where n is the number of input set, Xk is the k^{th} element of the input set, Xin is the element to be normalized, Xout is normalized element.

Table 2 shows the encoding numbers for the testing set used in this thesis.

Table (2): Output protein sequence encoding

Reference Codon Sequence	Code in Information Input
ATC	1
CTA	2
GAT	3
GGA	4
ATT	5
ACG	6
GTG	7
GAG	8
CCG	9
GCG	10
CTC	11
CAA	12
GCC	13
CGC	14
TGG	15
TTG	16
TTG	17
CCT	18
CAG	19
CGT	20

4. RESULTS

The experimental data collection collects data from biological experiments and computational analysis. Many experiments handle protein sequence, genome, gene expressions, etc. The tested experimental data almost saved in data base library to be used by different researchers and analytical specialists. These data bases are distributed over the internet have very high importance in bio science, and are used by medical experts. In addition, they have a unique importance for computer information systems developers and researchers. The Protein Information Resources shares the bio information data base with different famous online resources. It localizes very accurate and comprehensive data sets.

For the used neural networks and fuzzy logic systems, the main parameters are the architecture (number of hidden layers and number of neurons/hidden layer with respect neural networks, also the membership functions and rules with respect to fuzzy logic), in addition to training factors (learning rate, stopping condition,…etc). The modulation selection is done through trial and error, at which different trials were tested to tune the classifier parameters. To find the best parameters, the following is used:

1. Divide the data set into training set, and test set.
2. Select architecture and training parameters.
3. Train the model using the training set.
4. Evaluate the model using the testing set.
5. Repeat steps 2 through 4 using different architectures and training parameters.
6. Select the best model and train it using the training set.
7. Assess the final model using the testing set.

Once the model selection and training is completed, its generalization performance needs to be evaluated on previously unknown data set. The most popular methods for evaluating the generalization performance is to split the entire training data into two partitions, where the first partition is used for actual training and the second partition is used for verifying the performance of the algorithm. The performance on this latter data set is then used as an estimate of the algorithm's performance. Different approaches could be used (Yoshua, 2003).

One approach is to use the entire data set for testing and part of it for training to select the classifier and estimate the error rate. This approach suffers from the over-fit of training data, leading to error rate estimate which will be overly optimistic (lower than the true error rate).

Holdout Method (split the training data into disjoint subsets) for a single train-and-test experiment, the holdout estimate of error rate will be misleading if unfortunate split happens. The limitations of the holdout can be overcome with a family of re-sampling methods at the expense of more computations cross-validation, random-sub sampling, K-Fold cross-validation, leave-one-out cross-validation.

Hence the goal of this paper is to design an intelligent system that is able to classify the protein sequence from its behavior. So, the neuro-fuzzy system that is designed is subjected to be trained and tested in cross validation process in order to measure the classification error of the contributed system, in addition to the validity of the system design assumptions in different training-running schemes. And the cross validation is the validity test technique. The successive process is done through the following steps:

1. From the overall data set , select 40% of that data to be training set, and the other 60% will considered to be testing set. Note that, the selection of training set and testing set is being done randomly.
2. Train the fuzzy system using neural network trainer. Consider the training set that was selected above in training.
3. Run the neuro-fuzzy recognition program on the training set that is selected previously and record the recognition result.
4. Run the neuro-fuzzy recognition program on the testing set that is selected previously, and record the recognition result.
5. Re-select 40% of overall data set as training and the remaining 60% are being considered to be testing set. Note that, the selection is being done randomly in all trials, so the reselected data will differ from the last selected sets.
6. Repeat step three and four.
7. Repeat steps five and six until getting five different training and testing records.
8. Calculate the error by taking the means of the five experiments.

Table 3 shows the result that is obtained by previously described experiments considering 1000 record of data set in care.

Table (3): Validation result for 500 set of data records

Recognition Result	Accuracy using Mean Square Error (MSE)	Accuracy using Mean Percentage Absolute Error (MAPE)
40% that is used for training	0.938	95%
60% that is selected for testing	0.907	91%
100% of the data set	0.913	92%

Table 3 shows that, the result obtained is very good and could be considered to be an amazing result while it is computationally generated. Also, this result is not final as it could be changed depending on the training data set and testing data set.

This contributed method makes a high modification in accuracy performance with respect to the most commonly used techniques in such field. The time performance that designed and modified in this paper is not much reasonable than other

researches, this comes from that; this technique depends on artificial technique to achieve the accuracy principle and precision, while the other researches depends on statistical calculations. Even though, the time that gotten in this method is reasonable and reliable.

5. CONCLUSION

This research is concerning with building a protein sequence classifier using artificially intelligent computations. The benefit of using artificial intelligence is connected with preprocessing of the protein input to automatically estimate the protein family sequence, thus anticipating many behaviors of such sequences and bio informatics. The modern genome researches focus on adapting the artificial intelligence techniques to handle the problems of accuracy, precision, reproducibility, and repeatability, where. Such criteria are so hard to be achieved by human experience in most many cases of the real word. The demands of some complex systems that require complex computations vary from system to another (i.e. the size of the data) that needs to be handled represents a challenge in addition to other challenges. The work in huge input set or huge data set is like a puzzle that is being needed to be solved in order to suggest suitable solutions. The bio informatics is adaptive systems and data styles considered to be hard to solve their estimate-ability without modern adaptive expert systems and artificial intelligence techniques and specifications which minimize the uncertainty and realize the processing input, output, and rule generations of such cases. The uncertainty in bio informatics represents the biggest problem faced. For protein sequence estimation or classifications, it is not easy to model the uncertainty and the generation of rules that describe such systems is not an easy task. In fact, the uncertainty of protein information system is not easy to be modeled, so fuzzy logic could be considered to be the best predictor that would be used in such problem solving strategies.

The accuracy obtained by this research is 92% of MAPE. Even though this is a very good accuracy for the contributed research in comparison with the related researches in such field, but also, it should be improved to be better.

REFERENCES

[1] U., Turi D. and Stevens R.. “A Little Semantic Web Goes a Long Way in Biology”, Computer Science Journal, 2007.

[2] Kay Russell. Biometric Authentication, Technical Report, CSO, the resource for security executives, 2005. Accessed 18-5-2010.

[3] Xu Z., Tresp V., Yu K. and Kriegel H.-P.. “Infinite hidden relational models”, International Conference on Uncertainty in Artificial Intelligence, 2006.

[4] Hoff P.. “Multiplicative latent factor models for description and prediction of social networks.” Computational and Mathematical Organization Theory, 2007.

[5] Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., J. M., Cherry A. P. avis, Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A. B., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M. and Sherlock G..“Gene Ontology: tool for the unification of biology”. Nature Genetics, 25:25–29, 2000.

[6] Wu Cathy H., Huang Hongzhan, Yeh Lai-Su L. and Barker Winona C., “Protein family classification and functional annotation”, Computational Biology and Chemistry 27 (2003) 37/47, ELSEVER 2002.

[7] Bengio Yoshua and Grandvalet Yves. “No Unbiased Estimator of the Variance of K-Fold Cross-Validation", IRO Technical Report TR-2003-1234, May 21st, 2003.

[8] Marcotte, et al., 2001Marcotte E. M., Xenarios I. and Eisenberg D.. “Mining literature for protein–protein interactions”. Bioinformatics, 17:359–363. 2001.