

Research Article

Neural Network based Protein Sequence Classification

Farah Hanna Zawaideh

Irbid National University
Computer Information System Department

Dr.farahzawaideh@inu.edu.jo

ABSTRACT

An area of interest in biometrics is to make the medical and bio decisions that are related to human entity easier, accurate, repetitive, and more precise. The bio information has complex structure to be dealt with, analytically through mathematics. This paper deals with issues in this field which are still under uncertainties to minimize that uncertainty in one of the hottest areas in this field of research which is the proteins informatics, that is relates the protein data with the modern information technology and it includes portions mapping and classification. This paper focuses on the estimation and classification of the protein families. The classification of the protein families depends on their specific measurements. The sequence of proteins is an important issue for medical experts and represents a challenge in many cases of rising edge of the bio analysis field. The contribution of this paper is to adapt an artificial neural network based system in order to achieve high accuracy in protein classification that could be greater than 88 percent. **Copyright © AJCTA, all rights reserved.**

Keywords: Classification, Neural Networks, Protein Sequence, Bioinformatics.

1. INTRODUCTION

The DNA is storage of instructions of genetics and informatics that could control the functions of all organisms and their living developments. It contains the basic building ups of instructional and other components of the bio cells. Where RNA is considered to be one of those components [5]. The functions of RNA are to encode genetics and forward this informatics to protein synthesizer. The RNA sequence is transformed into proteins, which also controls the genome expressions, catalyzing the bio reactions and such [8]. The high productivity projects of genome are being parallel to the fast sequence of genome accumulation for huge organism's number. The specialists or biologists are subjected to utilize the data, analyze it, and identify the family of the proteins, where the proteins are usually implicitly included in sequences of genomes that are represented by a data of protein behaviors. The extensive experiments of biologists and medicine experts that direct the classification of proteins into families is not possible without modeling or efficient computing; especially for large scale genome and genome functionalities [1]. The composition of DNA takes the form of opposite strands in a helical shape. Its structure and arrangement represent the chromosome. The chromosome and genetic information could be extracted from the RNA. A sample series of CODON in a sample RNA messenger is shown in Figure-1. While C is bonded to G and also, G is being bonded to C, C is bonded to A, and A is bonded to U. Some information is not possible to be determined directly from the figure-1. It should be implicitly analyzed and recognized [1]. Bioinformatics is adaptation of computer information systems to handle biological and biomedical data sets includes biological analysis, classification, recognition, and processing. Merging data and technological tools enables to handle very large blocks of information, with complex structure and defined uncertainties. The main duties that are required from bioinformatics systems are to handle data sets, define processors or inference systems to handle the internal

structure of that data, and analyze that data to either get its internal characteristics in understandable form or to anticipate future behavior of specific attributes [2].

A public integrated sources of informatics related to proteins that can support the proteomic researches genome in scientific researched could be found in protein information resources (PIR). It makes the protein sequence data set available for protein annotation functionality of its sequence. This enlarges out of the sequence atlas of the proteins and their structure. The problem of annotating the proteins could be solved by classification and methods that are considered as rule based strategies. Also, it is cooperated with attributes evidence cooperated with knowledge integration bases structures implementation. This is commonly known as knowledge based structure identification. It is structured from two parts of data sets to make availability of complete sequences of proteins collection and rigid meaningful characteristics of information of the protein, in addition to analyzing the sequence by computerized tools introducing the bio informatics [1].

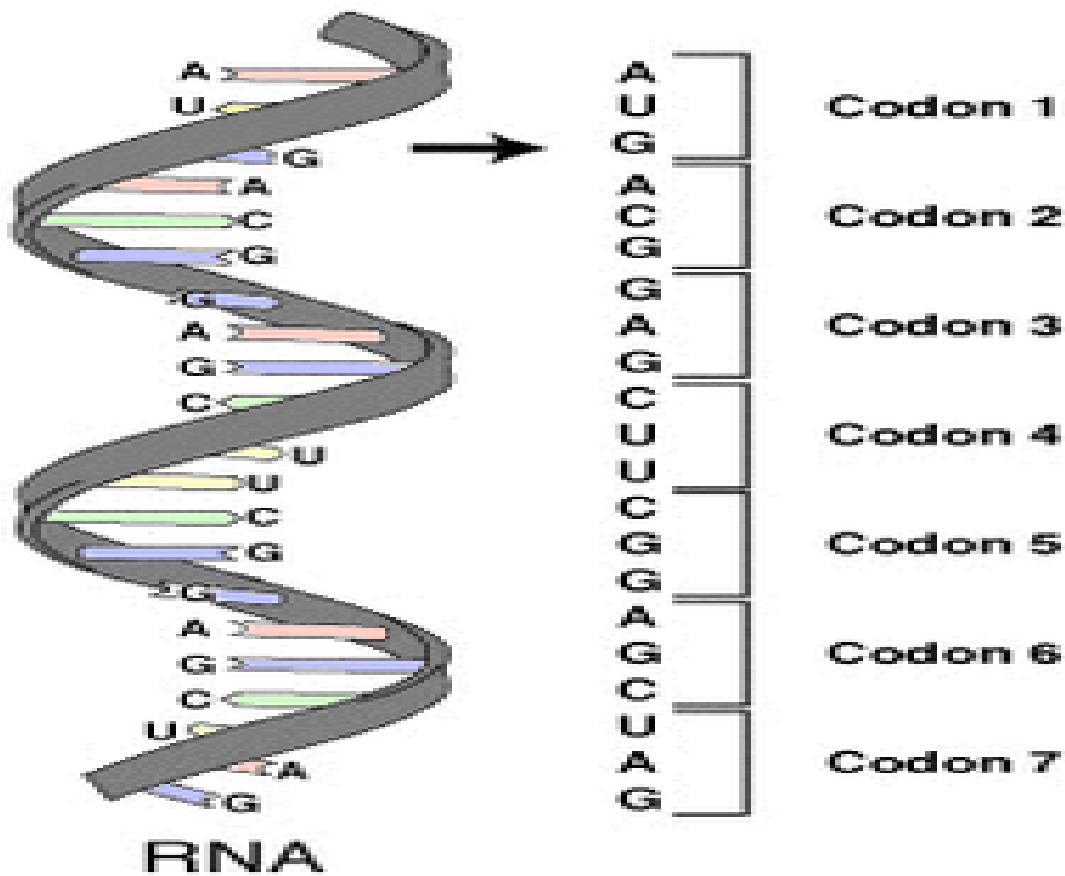


Figure 1: Coding of RNA for Ribonucleic Acid

This paper is dealing with protein informatics to analyze the sequence of proteins (i.e. such as DNA and amino-proteins) in order to determine the sequence type or protein family. This process includes very large data set of protein sequence and complex structure of the numeric data. This makes the process to be so hard to be achieved by traditional analytical mathematics or classical computing strategies [8]. The preprocessing of biological data and analyzing it is a major problem that facing the biologists. This paper improves the result extraction to be user friendly. Implementation of neural network gets the impact of artificial intelligent training that handles the large historical data; in addition, it builds the internal structure automatically without any analytical formulation. Neural network based intelligent returns the ability of intelligent prediction and forecasting, plus its high performance in clustering. The artificial neural network works as multi- input and multi -output recognizer, with a high performance in dealing with non-linearity intern the system and uncertainty of the input data and / or input data structure [1] , [7]. Multi-Layer Perceptron (MLP) neural network ensures high speed accuracy and precision due to its unique training performance and rigid structure [2]. The time that many researchers in such field are adapting different types

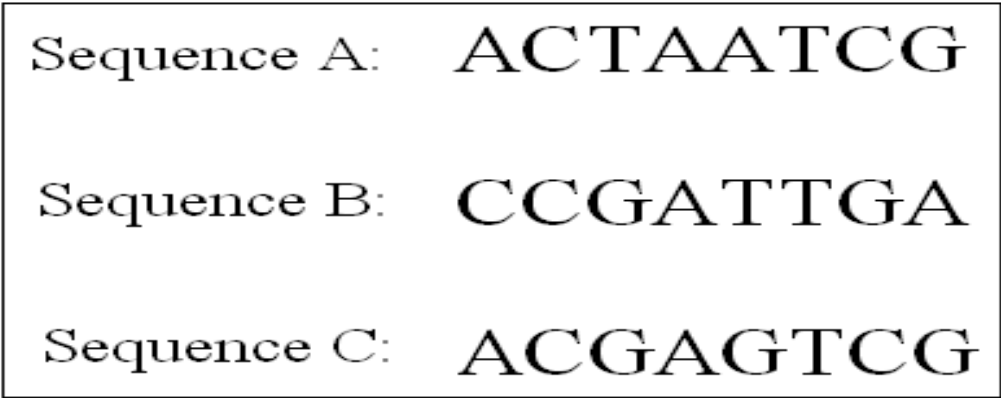
of neural network in different various training algorithm in order to minimize the complexities that concerns design multi-layer perceptron and its slow training rate, the MLP still the most rigid neural network that can overcome the flexibility problems in addition to handling very larger scale data.

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 and not artificial intelligent techniques problems [6]. The authors in [1] demonstrated the Semantics of state-of-the-art technology and how that technology could be adapted for e-Science applications. That research focused on the classification automates, especially for biological protein applications. The [1] research displayed recognition results were displayed in terms of quality comparability in order to achieve an expert system instead of a personal experience. The classification investigations of the data resulted was used in the informatics significant discovery. But that paper uses traditional statistics to overlap and predict behavior of the protein. This contributed technique overcomes many obstacles that are considered to be hard to solve, including the difficulty to allocate statistical functions in the data set response. The [5] work extracted the biological features related to the proteins, such as translation and transcription. So, the authors in [1] linked between pair wise relations of predetermined type and biological entities. But it didn't achieve accuracy results in addition to its high uncertainty in classifications. The authors in [3] demonstrated a relational model like the research demonstrated in [5], but it was based on the Latent-class model of relations. It demonstrated a common technique to classify the structure of block links of genome. Another similar technique was presented in [4]. It designed a model for link prediction in cascade design methodology. The cascade was structured from output-input prediction genome of protein sequences. Most researches concentrate on analytical and statistical approaches as described in this section. However, this paper concentrates on a modern approach to handle any number of input data set with large scale of data and protein families sets regardless of the basic modeling response.

3. METHODOLOGY

20 different amino acids could be extracted from specified genetic code. The sequence of amino acids is a polymer. One special type of protein is an enzyme that functions as a chemical reaction agent that becomes active during the cellular reactions [1]. A related number of proteins that are gathered together are called protein family, which may be considered to be a genome family. Where the family collects the proteins that have the same functions and similar sequences [2]. Since the protein sequence may be subjected to duplication, distortion, or deletion; nucleotides that form the sequence contain faults. Such faults take place during the cell formation or even the replication of sequence of DNA, and RNA. This will cause a gene to change the information being transformed into RNA and amino acids. The changes in the sequence make its analysis and study more difficult, leading to more difficulties in comparison and recognition. Figure-2 shows examples of sequence alignment [1].



Sequence A:	ACTAATCG
Sequence B:	CCGATTGA
Sequence C:	ACGAGTCG

Figure 2: Sample of Encoded Protein Sequences

The proposed model consists of two phases; the first phase is the preprocessing of input data, whether it is the data set that is used in building the system, or the input data to be recognized. The second phase is the intelligent recognizer which consists of

neural network. It is divided into two stages; the first is training while the system is being built. And the other is the running stage while the system is being subjected to recognizing the input pattern. The sample data that handled in this paper is gotten from the standard protein resources information database (PIR). The input data set consists of six values for each sequence that will be the output result. Figure-3 shows a sample of the data set used for training. This data set as shown in figure-3 is a raw material that needs to be processed before being used in feature extraction

Rs_id	ContigPositionStart(0 based)	ContigPositionEnd(0 based)	mRNA_start_position(0 based)	mRNA_end_position(0 based)	ReadingFrame	ReferenceCodon
669	1992391	1992391	3110	3110	1	ATC
79463492	2019323	2019323	993	993	2	CTA
78003756	6115827	6115827	275	275	1	GAT
1799931	6116515	6116515	963	963	2	GGA
1801280	6115999	6115999	447	447	2	ATT
79050330	6116236	6116236	684	684	2	ACG
56393504	6116496	6116496	944	944	1	GTG
2230180	6731080	6731080	1750	1750	1	GAG
28934585	6727211	6727211	344	344	2	CCG
3741056	11554761	11554761	88	88	1	GCG
76038779	14450818	14450818	766	766	1	CTC
1042714	9369399	9369399	317	317	1	CAA
28364013	25682021	25682021	308	308	1	GCC
28364012	25681388	25681388	941	941	1	GAG
36031925	25681076	25681076	1253	1253	1	CGC
4994	25681943	25681943	386	386	1	TGG
4994	25681943	25681943	386	386	1	TGG
79075278	20148200	20148200	1061	1061	1	GCC
78915065	20147046	20147046	832	832	3	TTG

Figure 3: Sample data set before processing

The sequence code is re-encoded by a number in order to make it suitable to be processed and handled in programming functions.

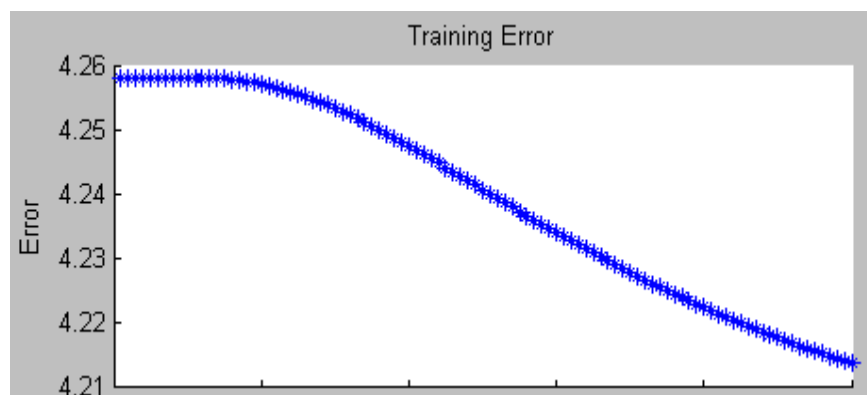


Figure 4: Final training process

While the training process is being completed, the neural network recognition system file is ready, and needs to be saved on the disk. Now, it is complete and ready for testing and validation procedures. The input data that is held for testing will pass the same steps during the training procedure, but the only difference is that the data will be smaller and there is no target attached

with it. Figure-4 shows the actual curve of training process. It represents the performance over epochs. The figure depicts a one stage of four training stages. Each stage represents an accumulating training process in order to achieve the reliability in training and adaptation of the network.

4. RESULTS

This paper handles data base supplied from the Protein Information Resources data base website. It covers different protein data sets that could be used for different applications and researches. The Protein Information Resources shares the bio information data base with different famous online resources. It localizes very accurate and comprehensive data sets. Cross validation is implied. The main advantages of cross validation algorithm are that all data instances are used for training as well as for testing (in different times) allowing full utilization of the data. Since the procedure will be repeated N times, the probability of an unusually unlucky or lucky partitioning will be reduced through averaging. The data set consists of 200 inputs of different and various data record for 20 protein sequence. The implemented data set is subjected to measure the validity of the contributed algorithm.

Hence, the cross validation is the validity test method; the successive process is done through the following steps – the implemented algorithm's steps –:

1. From the overall data set , select 80% of that data to be a training set, and the other 20% will considered to be testing set. Note that, the selection of training set and testing set is being done randomly.
2. Preprocessing of the data that will be used in training is expressed by normalization of the data records.
3. Train the neural network that its structure has been designed as described before consider the training set that selected in step one.
4. Once the neural network is gotten trained, run the simulation of network recognition program on the training set that is selected previously and record the recognition result.
5. Run the neural network recognition program on the testing set that is selected previously, and record the recognition result.
6. Re-select 80% of overall data set as training and the remaining 20% are being considered to be testing set. Note that, the selection is being done randomly, so the reselected data will differ from the last selected sets.
7. Repeat step three and four.
8. Repeat steps five and six until getting five different training and testing records.
9. Calculate the error by taking the means of the five experiments.

Table-1 shows the result that is obtained by previously described experiments and taking 200 record of data set in care. Note that, each experiment takes a relatively very short time. From table-1 it is clear 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. Actually, the training sets of about 200 different records are just enough for testing the validity of the proposed algorithm.

Table 1: Validation result for 200 set of data records

Validation result	Accuracy using	Accuracy using
	Mean Square Error (MSE)	Mean Percentage Absolute Error (MAPE)
80% that used for training	0.872	90%
20% that is selected for testing	0.816	83%
100% of the data set	0.861	88%

Another issue is the training data set. It is conventional that increasing the training set will increase the training performance, but that is not entirely true, because of the limitation of features that are implicitly included in the training data. Also, the training time is exponentially related to the size of training set, so the increase of training set should be taken as a spot point in this design. Depending on training results and validation tests for 20 sequences of the proteins, a one thousand of training sets will be the best size of data for more reliable future improvement. Computational power that should be supported by the CPU is expensive only in training process. But the fact that, this paper divided training into four accumulated stages those enables to clear memory and rest the resources to achieve high learning speed.

5. CONCLUSION

A sample data base that sourced from Protein Information Resources website to build and test the proposed system, whereas neural network based system is being built and trained adaptively to solve the target problem and achieve the motivation and goals of this paper. So, the results and work methodology procedure yield many conclusions. The rapid rise of researches in information systems make the desired and demands of informatics to be more accurate, precise, reproducible, and realistic. The objective of informatics systems is not easily performed with traditional computing; hence an increasing need of artificial intelligence the adaptability of the artificial neural network recognizer could be achieved by continuous training of that system. The continuous training is being built using neural network and its solid learning strategy. The neural network adaptability can meet the modern complex systems forecasting and recognition demands. The use of k-folds cross validation ensures that the contributed approach is a rigid system. But it is important to seek to make it adaptive and stable to achieve both accuracy and precision in terms of the number of reproductions. This paper takes a sample of protein resources data set and builds the system upon it. Although the testing and validation was done in a very rigid and well known procedure and results, the data size should be larger.

6. REFERENCES

- [1] Wolstencroft K., Brass A., Horrocks I., Lord P., Sattler U., Turi D. and Stevens R.. “A Little Semantic Web Goes a Long Way in Biology”, Computer Science Journal, 2007.
- [2] 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, ELSEVIER 2002.
- [3] Xu Z., Tresp V., Yu K. and Kriegel H.-P.. “Infinite hidden relational models”, International Conference on Uncertainty in Artificial Intelligence, 2006.
- [4] Andrews S. and Jebara T., “Structured Network Learning”, NIPS Workshop on Learning to Compare Examples, 2006.

- [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] Hoff P.. “Multiplicative latent factor models for description and prediction of social networks.” *Computational and Mathematical Organization Theory*, 2007.
- [7] Kotsiantis S. B., Zaharakis I. D. and Pintelas P. E..“Machine Learning: A Review of Classification and Combining Techniques”. *Artificial Intelligence Review*, 26 (3):159{190, November 2006.
- [8] Marcotte E. M., Xenarios I. and Eisenberg D.. “Mining literature for protein–protein interactions”. *Bioinformatics*, 17:359–363. 2001
- [9] 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.
- [10] Desobry F., Davy M., and Doncarli C.. An Online Kernel Change Detection Algorithm. *IEEE Transactions on Signal Processing*, 53(8), 2005a.