University of Ibadan
Journal of Science and Logics in ICT Research

A Computational Model for the Clustering of Protein Sequences Using a Self-Organizing Map and an Alignment-free Algorithm

[1]✉MAKOLO, A. U. and [2]MAMINOR, G. U.

University of Ibadan, Ibadan, Nigeria
aumakolo@gmail.com
gbubemimaminor@gmail.com

Abstract

Advancement in biotechnology has resulted in an increase in the rate at which biological data such as RNA, DNA and proteins are being sequenced. Inherent in the primary structures of proteins are features capable of providing information that can be used for classification using machine learning tools. In this study, a clustering model is designed for protein sequences using an alignment-free encoding technique and the Self Organizing Map (SOM). The model is an integration of an alignment-free encoding technique (Amino-acid Content Ratio (ACR) + Amino-acid Position Ratio (APR)) with the SOM algorithm. The encoding technique generates a 40 dimensional feature vector for each protein sequence which the SOM algorithm used to perform a clustering task. The SOM nodes are initialized randomly from the sample space which makes the ordering of the nodes faster. The model was implemented using the Java programming language and was evaluated using a data set of 500 sequences made up of five classes of Proteins (100 sequences each) which were collected from the UniProt Knowledgebase. Clustering of the data set was performed using learning rates of 0.1-0.9. A comparative analysis of the model against the use of only ACR encoding technique was also performed. The results showed that the model is valid and consistent in discovering quality protein clusters with a low standard error value of 0.2percent for Sensitivity test and a low standard error value of between 0.05-0.1percent with respect to specificity test. It also showed that the (ACR+APR) encoding technique is more sensitive and specific when compared to the ACR technique.

Keywords: *Self-Organizing map, Protein sequences, Alignment-free algorithm, Amino acid content ratio, Amino acid position ratio*

1. INTRODUCTION

The use of more efficient sequencing techniques has resulted in an increase in biological data such as DNA, RNA and Proteins and this has necessitated a need for more efficient soft computing tools for the analysis of these data [1].

Proteins play important roles in life such as providing structural support, storage, defence and transport to living organisms as well as performing enzymatic reactions in biological systems. Despite their diversity in function, they are all made up of the same basic components called amino acids .The amino acid sequence of a protein is represented using the symbols of the amino acids it is composed of, and these ordered symbols are referred to as the primary structure of a protein or simply a protein sequence. Each amino acid has its own characteristics and a combination of these amino acids in a particular order can influence the formation of the overall three-dimensional structure of the protein molecule [2].

Many machine learning tools exist for finding patterns in biological data and these can be used in predicting the classes which newly sequenced biological data belong. Machine Learning Algorithms can be classified into supervised and unsupervised learning algorithms. While supervised learning algorithms perform classification using labelled data, unsupervised learning algorithms have the

Makolo A. U. and Maminor G. U. (2020). A Computational Model for the Clustering of Protein Sequences using a Self-Organizing Map and an Alignment-free Algorithm. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 5 No. 1, pp. 1 – 10

ability to learn patterns from unlabelled data and form groups or classes based on the inherent structure or patterns in the data.

Clustering techniques are unsupervised learning algorithms which are exploratory in nature and are capable of grouping unlabelled data into clusters. These techniques analyse data by comparing the similarities of the data and grouping them into clusters based on the closeness of the data properties by using a distance measure such as the Euclidean distance. Comparison of similarities of biological sequences can be performed either by Sequence alignment methods or Alignment-free methods. For Protein sequences, sequence alignment methods require placing the protein sequences side by side and using a score function to determine similar sequences based on the number of deletion, substitution and insertion of amino acids in the compared protein sequences in order to detect the alignment with the best score. Sequence alignment methods such as dynamic programming can be misleading due to unequal length of sequences. They can also be slow and computationally expensive but usually produce optimal result. Alignment-free techniques for the analysis of protein sequences are less computationally expensive and are applicable to variable length sequences. They find application where computational speed and memory space are of essence [3] [4].

Clustering techniques can be useful as a pre-processing step for data classification [5]. The Self-Organizing Map (SOM) is an unsupervised learning algorithm. It is a model-based, competitive learning clustering tool [6] that is robust to noise, efficient and possesses strong visualization of the discovered clusters [4].The standard SOM algorithm cannot be used directly for the analysis of non-numeric data type such as protein sequences .It was designed to analyse continuous numeric attributes using the Euclidean distance for the comparison of object similarities [1].

The SOM has achieved a broad application in many fields such as pattern recognition, engineering systems, medical diagnosis, image segmentation [7], web document clustering [6], and bioinformatics [8]. In order to use the SOM algorithm for the analysis of protein sequences (non-numeric data type), a feature extraction technique (encoding) capable of preserving sequence information is therefore required prior to the performance of the clustering task. This will convert the features of the protein sequences into numeric data type.

This study develops clustering software (with graphical user interface) for protein sequences through the integration of an efficient alignment-free feature extraction technique with the SOM algorithm. The developed software model was evaluated using Sensitivity and Specificity test on benchmarked data set retrieved from UniProt Knowledgebase.

2. RELATED WORKS

Ahmad *et al*. [2] extracted protein sequence features using the one-gram model (amino acid content ratio). The Euclidean distance was used for the competitive learning phase of the SOM algorithm. The data set used for analyzing the model was made up of three families of proteins; 100 sequences of cytochrome c, 100 sequences of insulin, and 200 sequences of globin (made up of 100 sequences of hemoglobin alpha chain and 100 sequences of hemoglobin beta chain subfamilies). The study investigated the effect of a spread factor parameter to the node growth. Though the model was able to classify the sequences based on taxonomy, it was reported that some nodes had a mixture of two protein families.

Mohamed *et al*. [4] developed a SOM model for classification of DNA sequences using Needleman & Wunsch algorithm as the similarity function in competitive learning phase. Evolutionary techniques using crossover and mutation were integrated during each iteration to produce new features within the neighbor sequences of the winning unit. The dataset was also applied to other classifiers using WEKA and the proposed SOM algorithm showed a better performance in accuracy and precision than Random tree and Naïve Bayes algorithms.

Delgado *et al*. [1] proposed six different codification techniques based on Euclidean space for the analyses of genomic sequences. These were tested on the classical Kohonen SOM model and on the Growing Cell structures model using two different sets of sequences; 32 sequences of small subunit ribosomal RNA from organisms belonging to three domains of life and 44 sequences of the reverse transcriptase region of the *pol* gene of HIV *type1* belonging to different groups and subtypes. The sequences were initially aligned and then automatically coded into numeric vectors. The results showed that the most important factor affecting the accuracy of sequence clustering is the assignment of an extra weight to the presence of alignment-derived gaps.

In Hamel *et al*. [9], a model for the prediction of protein function from protein structure was developed. The protein features were extracted from the functional site in the proteins after alignment of the proteins. The three-dimensional coordinates of the alpha carbons of the functional site were unfolded into linear vectors and used as the feature vectors of the proteins. The analysis using SOM was performed using data from the Kinase family and the Ras superfamily.

Abdel-Azim [10] proposed the use of a probability density function (amino acid content ratio) for protein feature extraction and the similarity of the protein sequences were computed using Hellinger distance. The protein sequences were then clustered using hierarchical clustering algorithm. Two data sets made up of a mix of Influenza and Ebola virus and the second data set made up of only Influenza virus were clustered hierarchically.

Li *et. al.* [3] proposed a novel alignment-free algorithm for encoding protein sequences into a 440 dimensional feature vector that can be used in comparing the similarities of protein sequences using Euclidean distance. The 440 dimensional vectors consisted of a 400 dimensional Pseudo-Markov transition probability vector, a 20 dimensional content ratio vector, and a 20 dimensional position ratio vector of the amino acids in the sequence. The proposed algorithm was analysed statistically on two sets of data; the ND5 dataset and the F10 and G11 dataset, and the results obtained were seen to be consistent with protein sequence aligners like ClustalW.

Knowledge discovery from the clustering of protein sequences is an important preliminary task in bioinformatics. A review of literature in sequence clustering has shown that due to the limitations of sequence alignment techniques, alignment-free techniques can be adopted for feature extraction from biological sequences. Also, the data summarization and visualization ability of SOMs, propose to be an effective tool for knowledge discovery in biological sequences if an efficient feature extraction technique can be integrated into it for the analysis of non-numerical data such as protein amino acid sequence.

3. SOM ALGORITHM

The SOM algorithm is a competitive learning algorithm wherein all the nodes compete with each other for the current input data. The node with the closest similarity with the current data as measured by an appropriate distance measure is declared the winner. The weights of the winner node and its associated neighbouring nodes are adjusted in order to move them closer to the current input data [6]. The process is repeated for all the training data while decreasing the winner node neighbourhood radius until convergence or a stopping criterion is reached [5] [4]. After the SOM training, each of the nodes will represent a model of prototype vectors which have been organized into an optimally descriptive grid of the training data [11]. Analysis of a data set using the SOM algorithm begins with an initialization of the nodes and this is followed by the repetitive phases of competition, cooperation and adaptation.

3.1 Initialization

This is the starting phase of the SOM algorithm where random weights are assigned to all the nodes before the training begins. Initialization can be performed either randomly or from the sample space of the input vectors or linearly [5]. This research adopts the use of the Sample Initialization technique where SOM nodes are initialized by randomly selecting samples from the input data. This Data analysis approach was adopted because according to [11], random initialization of the nodes of a SOM is ineffective.

3.2 Competition

When an input sample data is passed into the SOM, the similarities of the nodes (using the node weights) and the current input sample (vectors) are computed based on a distance measure or discriminant function such as Euclidean, Manhattan, Hamming etc. The node with the smallest value is declared the winner of the competition and is activated.

The distance measure is an important component of a clustering algorithm and it is necessary that domain knowledge be applied in the choice of a distance measure for clustering technique [5].

The Euclidean distance was adopted for this work because all the vectors are in the same physical units [5] and [3] in their analysis on protein sequences using ND5 dataset and F10 and G11 dataset demonstrated that the Euclidean distance is more effective than Hamming distance in the separation of the protein families. The Euclidean distance is shown below where I is the current input vectors and W is the node weights for n number of features [3].

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(I_i - W_i)^2} \qquad (1)$$

3.3 Cooperation

The neighbours of the winner node (activated node) are defined and also activated by the winner node based on the current neighbourhood radius of the regression step and a grid of neighbours is set. The current neighbourhood radius ($\delta_{(t)}$) is calculated using the Gaussian function instead of the bubble function (winner-takes-all). This is because the Gaussian function retrieves more clusters than the bubble [5]. $\delta_{(t)}$ [12] is given as

$$\delta_{(t)} = \delta_{(0)} e^{-t/\gamma} \qquad (2)$$

where *t* is the current iteration (i.e. current step) and γ is the time constant [12] given as

$$\gamma = \frac{Total\ number\ of\ iterations}{Map\ radius(\delta_{(0)})} \qquad (3)$$

3.4 Adaptation

The weights of the winner node and its associated neighbours are adjusted to look more like the current input vectors of the data. The closer the neighbours are to the winner node the greater the degree of node weights(W) adjustment with respect to the input vectors(I).A learning rate(L) between 0 and 1 [5] is set at the beginning of the training(maximum value) but this decreases as the iteration number(t) increases . The decrease of the learning rate is achieved using the Gaussian function [12]:

$$L_{(t)} = L_{(0)}\, e^{-(\frac{t}{\gamma})} \qquad (4)$$

The time constant (γ) is also applied in the calculation of the current learning rate. The Learning rate is used to compute the neighbourhood function (NF) .The NF is like a smoothing or blurring kernel over the grid [11] and is dependent on the radius of the neighbour nodes (r_n), radius of the winner node (r_c) and the current neighbourhood radius $\delta_{(t)}$ .The NF is an important parameter used in the SOM node weights update rule [11]. It is given as:

$$NF = L_{(t)} e^{-((r_n - r_c)^2/2\delta_{(t)}^2)} \qquad (5)$$

The SOM node weight update rule is the equation that is used for the adjustment of the SOM node weights during the training phase [11]. This is given as:

$$W_{i(t+1)} = W_{i(t)} + NF_{(t)}(I_{i(t)} - W_{i(t)}) \qquad (6)$$

Where *W* is the node weights, *I* is the current input vectors, *i* is a feature vector, *t* is the current iteration and NF is the neighbourhood function.

4 FEATURE EXTRACTION USING AN ALIGNMENT-FREE ALGORITHM

A protein sequence is made up of the amino acid sequences that make up the polypeptide chains of the protein. The twenty basic amino acids and their symbols are shown below

1	A…………………	Alanine
2	C…………………	Cysteine
3	D…………………	Aspartic acid
4	E…………………	Glutamic acid
5	F…………………	Phenylalanine
6	G…………………	Glycine
7	H…………………	Histidine
8	I…………………..	Isoleucine
9	K…………………	Lysine
10	L………………….	Leucine
11	M…………………	Methionine
12	N…………………	Asparagine
13	P………………….	Proline
14	Q………………….	Glutamine
15	R…………………	Arginine
16	S…………………	Serine
17	T …………………	Threonine
18	V………………….	Valine

19	W…………………	Tryptophan
20	Y…………………	Tyrosine

The composition and order of the amino acids of a protein provide useful information for the classification of proteins. The total number of the different amino acids in a protein is always proportionally the same with proteins of the same type (e.g. Insulin) [13]. Based on the inherent Information of a protein's amino acid sequence composition, the function [3] is given as

$$ACR_I = \sum X_I / S_{\mathrm{L}} \qquad (7)$$

where S_{L} = Length of Amino acid sequence
X_{I} = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}
where 1 <=I<=20

was used to compute the Amino-acid Content Ratio(ACR)[3] for each of the twenty amino acids giving twenty features for a protein sequence. The function [11] is given as:

$$APR_I = \sum X_{I,P} / \sum_{n=1}^{L} S_n \qquad (8)$$

where S = Amino acid Sequence
L = Length of S
X = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}
where 1 <=I<=20 , $X_{\mathrm{I,P}}$ = Position of X_{I} in S

was used to compute the Amino-acid Position Ratio(APR) [3] of the twenty amino acids based on the order of occurrence in the sequence giving an additional twenty features for a protein sequence.

The ACR and APR therefore generates a total of forty features for each protein sequence. These features have numeric values which can be processed by the SOM algorithm. Figure 1 is a graphical representation of the proposed SOM model which uses an alignment-free algorithm for the encoding (feature extraction) of the protein sequences. This model was implemented into software called ProtSOM using Java programming language.

5. EVALUATION OF MODEL

5.1 Benchmark Data set

The benchmark data set used for the evaluation of the model was made up of five hundred sequences comprising of one hundred sequences (100) each of five classes of proteins which are Cytochrome b, Cytochrome c, Haemoglobin alpha chain (HBA), Haemoglobin beta chain (HBB) and Insulin.

These data were retrieved from the UniProt Knowledgebase. The five classes of proteins were combined together in a FASTA file by alternating them in the order of one sequence of cytochrome b followed by one sequence of HBA, followed by one sequence of Cytochrome c which is also followed by one sequence of HBB and finally followed by one sequence of Insulin. This order was repeated for the five hundred sequences giving each class of protein a form of identification using the last digit of the sequence position in the file. Table 1 shows the technique for identification of the classes of proteins in the SOM map.

Table 1: Class Identification of Protein Sequences

Class	Last digit of Sequence Position in file
Cytochrome b	1 or 6
Haemoglobin alpha chain(HBA)	2 or 7
Cytochrome c	3 or 8
Haemoglobin Beta chain(HBB)	4 or 9
Insulin	5 or 0

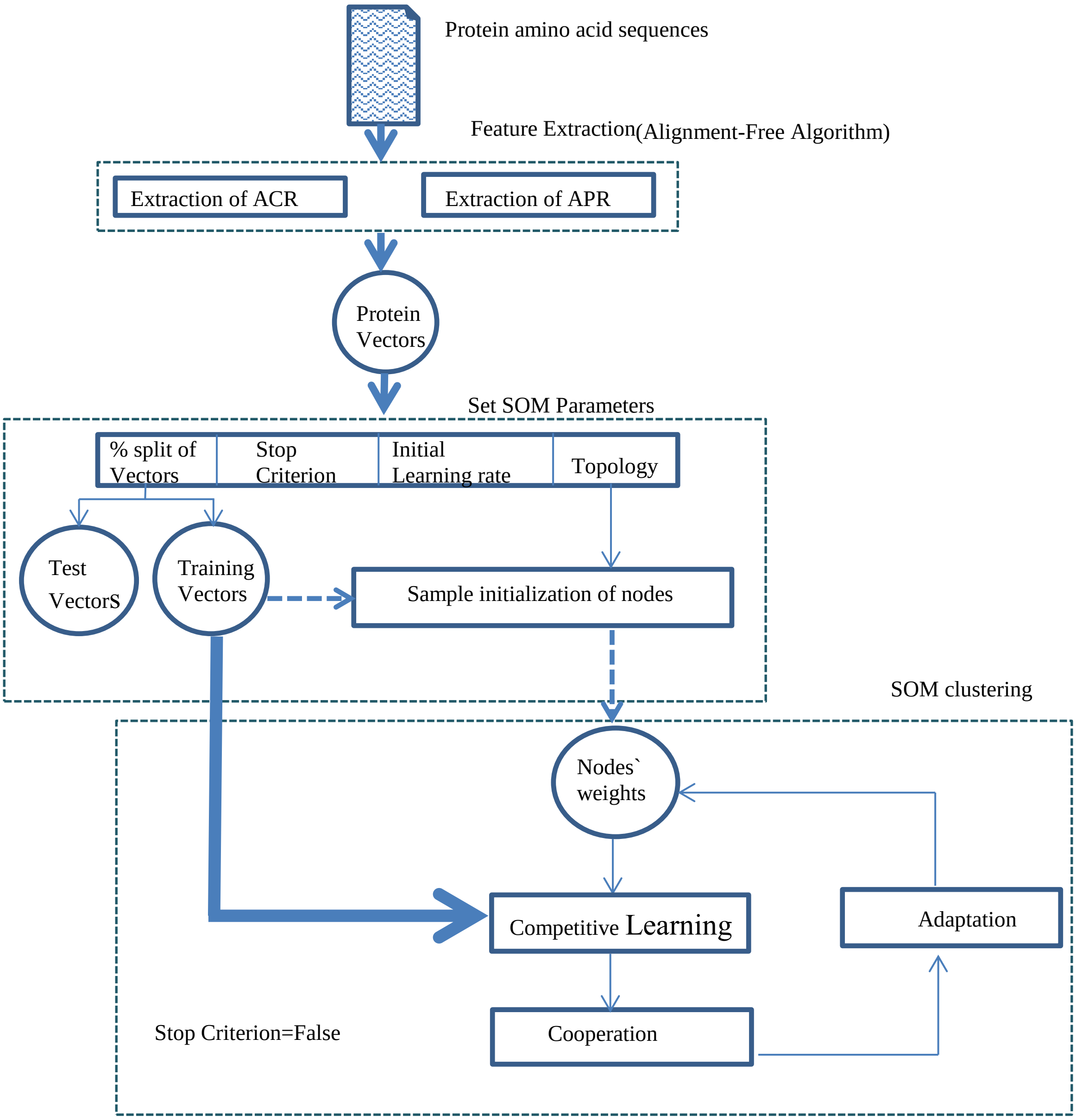

Figure 1: SOM model using an alignment-free algorithm for feature extraction

5.2 Clustering of Benchmark Data set

Clustering of the sequences was performed based on two models of encoding techniques which are:

1) Twenty feature vectors (ACR) –This is to be used for comparative analysis.
2) Forty feature vectors (ACR+APR)-This is our proposed model.

The data set was uploaded onto the ProtSOM software and the SOM Parameters were set as stated below for learning rates ranging from 0.1 to 0.9 for both models.

i. Topology of 1 X 5
ii. Neighbourhood radius of two.
iii. The same samples were used for the initialization of the nodes for all clustering activities. This was achieved by using the software option for constant initialization mode.

5.3 Evaluation Metrics

Figure 2 is a sample of the ProtSOM visualization using a one by five (1X5) SOM topology. A one-dimensional topology was adopted for ease of performance evaluation. Each protein sequence is represented using its serial number in the protein FASTA file

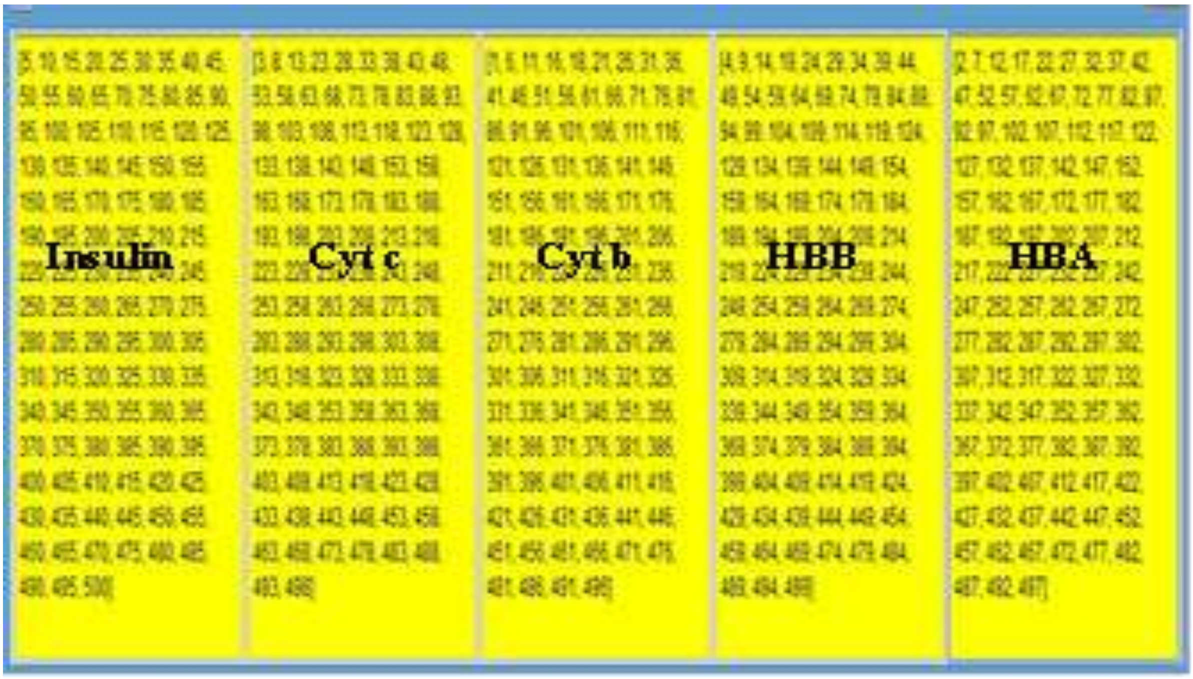

Figure 2: A One by Five (1X5) SOM topology of ProtSOM visualization

The ProtSOM model was evaluated using two statistical metrics of performance which are Sensitivity and Specificity. These metrics are used for binary classification problems but can also be adapted for multiclass problems [14].

5.3.1 *Sensitivity*

Sensitivity measures the ability of a classifier to correctly assign an object to its actual class. It is called true positive (TP) rate. Sensitivity of each class [15] can be calculated as:

$$Sensitivity = (TP * 100)/(TP + FN) \qquad (9)$$

For a given class A,

TP=True Positive (defined as objects of class A that are correctly classified as class A)

FN = False Negative (defined as all objects of class A which are not classified as class A).

5.3.2 *Specificity*

Specificity measures the ability of a classifier to correctly reject a given object as belonging to a class that it does not belong to. It is also called true negative (TN) rate. Specificity of each class can be calculated [15] as:

$$Specificity = (TN * 100)/(TN + FP) \qquad (10)$$

For a given class A,
TN= True Negative (defined as all non-class A objects that are not classified as class A).
FP=False Positive (defined as all non-class A objects wrongly classified as class A).

5.4 *Sensitivity Test Comparison of ACR and APR encoding technique*

Figure 3 is a plot of the Standard Error (SE) of Sensitivity Test of ACR and APR against learning rate.

5.5 *Specificity Test Comparison of ACR and APR encoding technique*

Figure 4 is a plot of the Standard Error (SE) of Specificity Test of ACR and APR against learning rate.

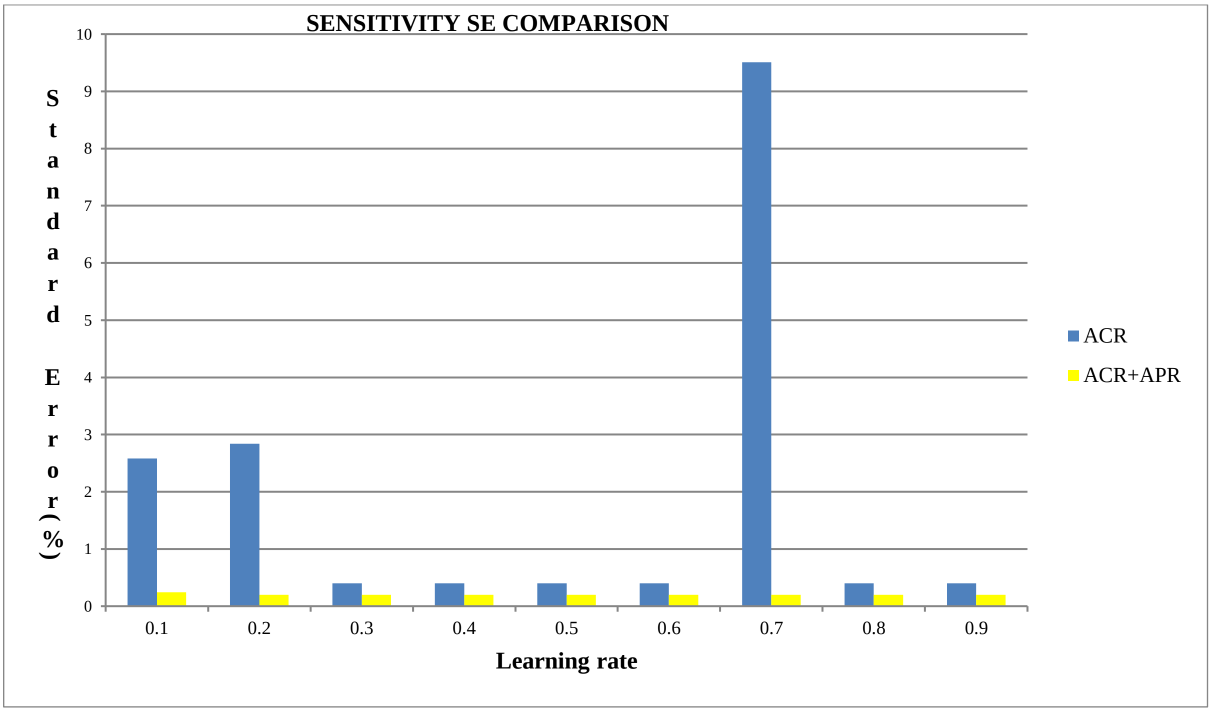

Figure 3: Sensitivity Test Comparison

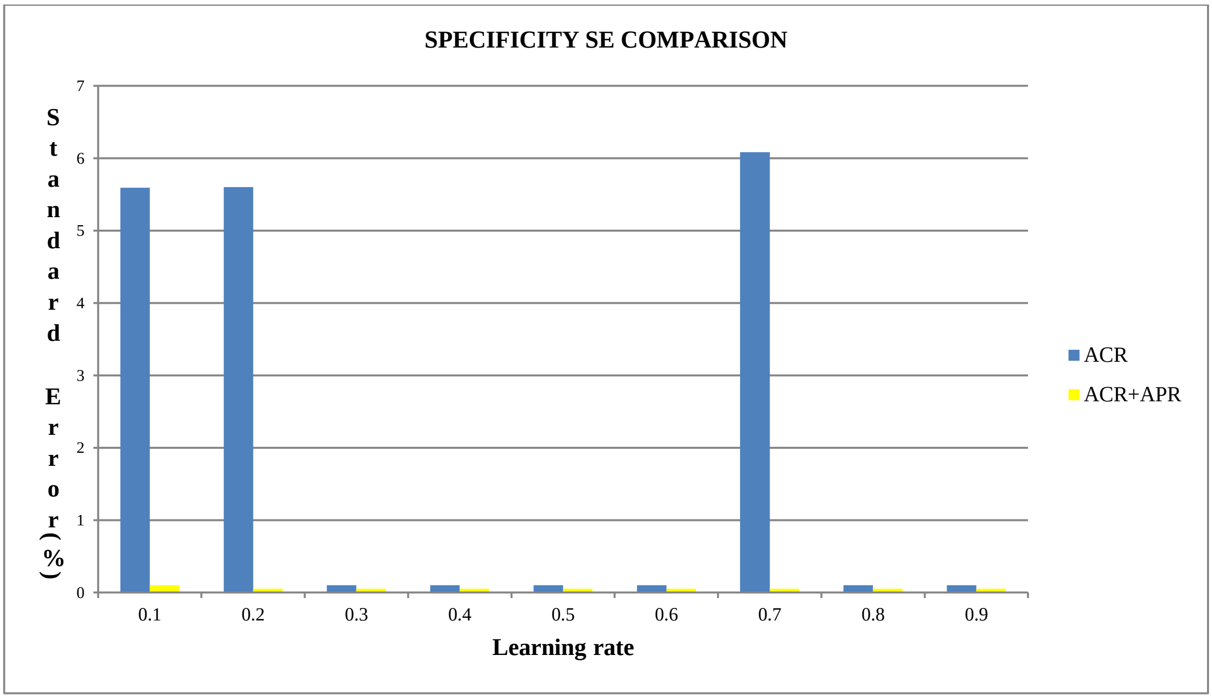

Figure 4: Specificity Test Comparison

5.6 Result Discussion

Table 2 Result Summary

	ACR Values (%)	ACR Range (%)	ACR + APR Values (%)	ACR + APR Range (%)
Sensitivity SE	0.4 – 9.5	9.1	0.2	0.0
Specificity SE	0.1 – 6.1	6.0	0.05 – 0.1	0.05

From Table 2, it is deduced that:

(i) Sensitivity Test for ACR+APR encoding technique had Standard Error values of 0.2% with a range of 0.0%. This implies that the feature extraction technique integrated with the Self Organizing Map has a **high validity rate.**

(ii) Specificity Test for ACR+APR encoding technique had Standard Error values of between 0.05% and 0.1%. This implies that the feature extraction technique integrated with the Self Organizing Map generates **high quality of clusters**.

The clustering activities performed using different values of learning rate showed that ProtSOM (using ACR+APR as the feature encoding technique for clustering Protein sequences) is **valid** and **consistent** in providing accurate clusters of protein sequences.

6. CONCLUSION

ACR has been used for encoding protein sequences in past works. This technique does not consider the amino acids positions in the sequence. The use of APR in addition to ACR is capable of giving better clusters. A comparison of both encoding techniques showed that the use of a combination of ACR and APR encoding technique has a better sensitivity and specificity than only the ACR technique.

The developed model (ProtSOM) which has a topology preserving ability can be adopted for the preliminary investigation of the family that proteins belong to, as well as show the closeness or relatedness of the clusters of proteins formed. The ProtSOM GUI eases the clustering task as well as provides an easily understandable visualization of the members of a given cluster of proteins.

References

[1] Delgado, S., Mora´n, F., Mora, A., Merelo, J.J. & Briones, C. (2014). A novel representation of genomic sequences for taxonomic clustering and visualization by means of self-organizing maps. *Bioinformatics 31*(5), 736-744.

[2] Ahmad, N., Alahakoon, D. & Chau, R. (2008). Classification of Protein Sequences using the Growing Self-Organizing Map. *4th International Conference on Information and Automation for Sustainability,* 167-172.doi:10.1109/ICIAFS.2008.4783969

[3] Li, Y., Song, T., Yang, J., Zhang ,Y. & Yang , J. (2016). An Alignment-Free Algorithm in Comparing the Similarity of Protein Sequences Based on Pseudo-Markov Transition Probabilities among Amino Acids. *PLoS ONE* 11(12):e0167430.doi:10.1371/journal.pone.0167430.

[4] Mohamed, M., Al-Mehdhar, A.A., Bamatraf, M. & Girgis, M.R. (2013). Enhanced Self-Organizing Map Neural Network for DNA Sequence Classification. *Intelligent Information Management 5*.25-33.

[5] Adeyemo, A.B. (2016). A Study of SOM Clustering Software Implementations. *Cori 2016 Proceedings of the 2nd Ibadan ACM International Conference on Computing Research and Innovations.*160-164.

[6] Han,J., and Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd Ed).(pp 383-434).San Francisco, USA: Elsevier.

[7] Chen, N. & Marques, N.C. (2005). An extension of Self-Organizing Maps to Categorical Data, *Proceedings of the 12th Portuguese conference on progress in Artificial Intelligence.*304-313.doi:10.1007/11595014_31.

[8] Gagandip, S. (2014). An Introduction to Self-Organizing Maps. http://home.cc.umanitoba.ca/~umsidh52/PLNT 7690/ presentation/SOM.html.

[9] Hamel, L., Lim, S. & Jaegle, S. (2016). Protein Structure-Function Analysis with Self-Organizing Maps. *Proceedings of the 17th International Conference on Bioinformatics & Computational Biology (BIOCOMP'16)*, 10-16, July 25-28, 2016, Las Vegas, Nevada, USA, ISBN: 1-60132-428-6, CSREA Press.

[10] Abdel-Azim, G. (2016). New Hierarchical Clustering Algorithm for Protein Sequences Based on Hellinger Distance. *Applied Mathematics & Information Sciences 10*(4). 1541-1549.

[11] Kohonen, T. 1999.The Self Organizing Map (SOM). http://www.cis.hut.fi/research/reports/quinquennial/ch1.pdf.

[12] Guthikonda, S.M. (2005). Kohonen Self-Organizing Maps. https://www.academia.edu/7880511/Kohonen-self-organizing-maps-shyam-guthikonda.

[13] Claverie, J. & Notredame, C. (2007). Bioinformatics for Dummies.2nd ed. Indiana : Wiley Publishing, Inc. Chapter 1:15.

[14] Pooja, M., & Pandey, P.N. (2011). A graph-based clustering method applied to protein sequences. *Bioinformation 6*(10). 372-374.

[15] Samanthi. (2013). Difference Between Sensitivity and Specificity. https://www.differencebetween.com/difference-between-sensitivity-and-vs-specificity