Indian Journal of Science and Technology, Vol 8(14), DOI: 10.17485/ijst/2015/v8i14/70701, July 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Mining the Amino Acid Dominance in Gene Sequences

V. Balamurugan[1*] and T. Marimuthu[2]

[1]Department of Information Technology, AMET University, Chennai - 603112, Tamil Nadu, India; bala_vm@yahoo.com
[2]Manonmaniam Sundaranar University, Tirunelveli - 627012, Tamil Nadu, India; mastersvksmca@gmail.com

Abstract

In the recent period, the classification techniques are widely applied in the field of Bioinformatics. The proposed Amino Acid Component based Classification algorithm adopts Iterative Dichotomiser3 classifier. The algorithm consists of two phases viz. attribute selection and component based classification. In the attribute selection phase the dominating amino acids and deficiencies in amino acids that cause the diseases are found. The second phase finds the components of amino acids which spread the diseases in the specified sequence. The experiments were carried out on the gene sequence of dengue virus which is available on the NCBI online biological database and the accuracy of the proposed algorithm is calculated as 90.744%. The proposed classification algorithm is compared with the traditional benchmark algorithms such as Naive Bayes, ID3, Random Forest, Multilayer Perceptron and J48. The result of this work can be used by the drug designers to predict new viral diseases.

Keywords: Amino Acid Components, Classification, Entropy, Information Gain

1. Introduction

Data mining is the process of extracting meaningful or interesting patterns from the large amount of data. There are several techniques in data mining such as classification, clustering, association rule mining, and regression that are used to extract the required information. Classification is a method of categorizing the given input items into a predefined group called class. As the manual data classification is a time consuming process, there is a need for automatic classification techniques. Machine learning techniques like neural networks, genetic algorithms, artificial intelligence are found useful in the classification of large data set. The primary step in the classification process is to build a classifier on the basis of known cases with multiple attributes. The classifier thus found is used to predict the new cases with similar attributes. In other words, the classifier allots the label for the new case. Data classification process has two phases viz. training phase and testing phase. There are several classification techniques such as Bayesian classification, decision tree induction, Self Organizing Map (SOM), etc in practice. Furthermore, the classification techniques that are based on rough set, fuzzy set, neural networks and Genetic algorithm are available in the literature[11]. In the recent period, the classification of genome sequences has been attempted by many researchers as this will be helpful in the treatment of many diseases[1]. Though this technique can be applied on any sequence, this work concentrates on the dengue virus genomic data.

Cells of the human body have a central core called Nucleus, which are packaged in units known as

*Author for correspondence

Chromosomes. Human beings have 23 pairs of chromosomes, which are together known as Genome. Genes are specific region of the genomes, which are the molecular unit of heredity of a living organism. Gene sequence contains a sequence of nucleic and amino acids. Nucleic acid consists of a chain of linked units called Nucleotide. Nucleic acid sequence has the combination of nucleotide bases within Deoxyribo Nucleic Acid (DNA) or Ribo Nucleic Acid (RNA). DNA is a chain of four types of nucleotide bases Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). A sample DNA sequence may look like TCCTGATAAGTCAGTGTCTCCT. RNA is represented as the combination of four nucleotide bases Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). A sample RNA sequence may look like UCCUGAUAAGUCAGUGUCUCCU.

DNA and RNA play a major role in the formation of proteins. The constituents of Proteins are amino acids which are represented using 20 English letters except B,J,O,U, X and Z. A sample protein sequence may look like CFPUEQGHILDCLKSTFEWEGHILDWES. Protein sequences are shorter than DNA Sequences[3].

Genome sequencing projects are currently producing a large amount of unclassified new sequences and cause the rapid development of protein sequence databases[9]. The classification of these data into functional groups or amino acid components has become one of the principal research objectives in proteomics[7].

The proposed Amino Acid Component based Classification (AACC) algorithm automatically classifies the protein sequences based on amino acids components with high accuracy rate. Initially, the AACC algorithm reads the given input gene sequence and makes the numeric conversion based on the length of the given sequence. The next step is to select the attribute for performing the classification based on entropy. Entropy is a measure based on information gain. Further it finds the dominating components and deficiencies in amino acids in the given sequence. The AACC algorithm will be helpful for drug designers in predicting new viral diseases.

In Section 2, the works related to the classification algorithms on amino acid sequences are outlined. Section 3 demonstrates the methodologies related to the proposed algorithm. Section 4 contains the experimental results that are obtained using dengue gene sequence. Finally, Section 5 offers the conclusion and the scope for further extension of this work.

2. Literature Review

In the last two decades, there are several research contributions in the field of bioinformatics which are available in the literature[18]. This section reviews the works related to the classification of amino acid sequences.

2.1 Classification Algorithms

Rajeswari et al.[17] have reviewed several classification techniques such as Self Organizing Map (SOM), Artificial Neural Network (ANN), and Extreme Learning Machine (ELM). Their study focused mainly on the data classification algorithm using machine-learning techniques.

Gupta et al.[8] have highlighted the importance of classification techniques in data mining. They applied the techniques such as data preprocessing, classification, clustering, regression, association rule mining, and visualization using the Waikato Environment for Knowledge Analysis (WEKA) tool. They applied the classification algorithms such as decision tree, Support Vector Machine (SVM), 1-Nearest Neighbor (1-NN), and Naive Bayes (NB). Their results indicated that the SVM had better accuracy. From the execution point of view, the NB classification algorithm was found to be faster.

Barnaghi et al.[4] have applied the classification techniques on a medical data set in order to diagnose the liver disorder. They applied four classification methods that include decision tree, Bayesian algorithms, neural network, and rough sets. For evaluation, they used WEKA open source tool and Rosette tool. They observed that the Rosette tool performs better for the rough sets compared to the WEKA tool. Rough set is a set that has no clear boundary. The data set was classified into two classes: class 0 and class 1. Their observation revealed that the neural networks classifier performed better in terms of accuracy and Multilayer Perceptron (MLP) was suitable for large data set.

Chopra[6] has classified disabled students of Indira Gandhi National Open University (IGNOU) into three different groups viz. area, type of disability and region. His objective was to provide group learning environment for the students with similar nature. They have justified the choice of attributes on the basis that students with similar type of disability would have similar problems and solutions. They used the enrollment data as the training set. Iterative Dichotomiser3 (ID3) classifier was used for classification.

Priyanga et al.[16] have proposed the cancer risks prediction system based on data mining techniques. The system estimated the risks of the breast, skin and lung cancers. They used WEKA tool for validating their system and then analyzed the performances of decision tree algorithms such as J48 and ID3. The J48 algorithm recursively classified data until it has been classified perfectly. The ID3 classifier leads to the construction of the decision tree that can be used to classify the data set. They also applied the Naive Bayes classifier for classifying the dataset. Naive Bayes makes the assumption of a class conditional independence which means the occurrence or non-occurrence of the previous assumptions. They observed that the accuracy of J48, ID3, and Naive Bayes were 98.16%, 100% and 86.23% respectively for the prediction of breast cancer and 98.31%, 100% and 89.03% respectively for the prediction of lung cancer. They concluded that ID3 algorithm provided an accurate result for all types of cancer data set.

2.2 Biological Classification

Osman et al.[14] have developed a hybrid learning algorithm called Neural Network Enzyme Classification (NNEC) to classify an enzyme found in Protein Data Bank (PDB) to a given family of enzymes. NNEC was developed based on MLP with hybrid learning algorithm, combining the genetic algorithm and back propagation. They used the dataset that consists of 6 enzyme super families that were extracted from PDB. They used 3200 enzymes in total as the training and testing samples. For all the samples, length of the sequence was ranging from 40 to 600 amino acids. They classified the sequences using NNEC algorithm with 6 families such as Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases. The accuracy rate was 72.94% for 6 classes.

Mansoori et al.[13] have defined the interpretable fuzzy rules for assigning amino acid sequence into an appropriate protein super family. They have used the distribution of amino acid sequence as features. These features were the occurrence probabilities of six exchange groups in the sequences. They used 6-letter exchange group method. Exchange groups are the commonly used piece of information. There are six groups of amino acids which represent high evolutionary similarity. The 6 letter exchange groups are {e1, e2, e3, e4, e5, e6} where e1 = {H,R,K}, e2 = {D, E, N, Q}, e3 = {C}, e4 = {S, T, P, A, G}, e5 = {M, I, L, V} and e6 = {F, Y, W}. Consider, the protein sequence PVKVPTKPKV which is represented by e4e5e1e5e4e4e1e4e1e5 and hence the exchange group encoding for this sequence is: three for e1, four for e4 and three for e5. Each feature value is scaled to a probability by dividing the number of the class member with length of the sequence PVKVPTKPKV. The probability obtained for the feature vectors e1, e2, e3, e4, e5 and e6 are 0.3, 0.0, 0.0, 0.4, 0.3 and 0.0 respectively. They designed a fuzzy rule based system for the classification of protein sequences and observed that the accuracy was low compared to the other classifiers.

Leon et al.[12] have examined the problem of classifying protein fold structure without sequence similarity by using classification techniques. They used Naive Bayes classifier, Instance based classifier and k-Nearest Neighbor algorithm. They analyzed the performances of the algorithms and observed that the instance based classifier algorithm produced 100% accuracy on the testing sets whereas; the k-Nearest Neighbor algorithm produced 90%. In addition, the Naive Bayes algorithm performed poorly on the training set.

Golmohammadi et al.[10] proposed a novel computational method to predict the type of unclassified membrane proteins based on their sequence. They compared their method with other methods such as k* classifier, Ensemble of neural network, Fuzzy k-Nearest Neighbor (kNN), SVM, Weighted SVM, augmented covariant discriminate, co-variant discriminate, port lock, least hamming distance, and least Euclidean distance. The main feature of their method was the inclusion of seven feature sets to encode a protein sequence. The proposed method correctly predicted the membrane protein type 86.9% of the time when tested on the independent data set.

Exarchos et al.[7] have used Sequential Pattern Mining (SPM) for sequence based fold recognition. One of the most efficient SPM algorithms, cSPADE has been employed for protein primary structure prediction. To validate the proposed classifier, an appropriate group of primary protein sequences was taken from the Protein Data Bank (PDB). The cSPADE algorithm found the frequent sequences with the constraints, such as minimum and maximum gap between sequence items. This approach classified unknown proteins in 17 candidate folds based on sequential pattern mining. The proposed method exhibited an overall accuracy of 35.9% while the overall accuracy of Split and Merge algorithm was 35%.

Othman et al.[15] have investigated the performances of different classification algorithms using WEKA tool kit

for breast cancer. They have used 6291 data with a dimension of 699 rows and 9 columns. They used 75% of overall data for training and the rest for testing. The best algorithm, based on the breast cancer data, was found to be Bayes network classifier with an accuracy of 89.71% and the total time taken to build the model was 0.19 seconds.

Aizenberg et al.[2] have used Multi Layer neural network, based on Multi Valued Neurons (MLMVN), for solving several classification problems in bioinformatics. They developed a model with multiple classes of microarray gene expression and breast cancer diagnostics. The observed classification rate was 95.94%.

3. Methodology

3.1 Problem Description

The viral infection affects the protein sequence and spread out specific amino acids in that protein sequence. The diagnosis of viral diseases is a complicated process in medical domain. The core objective of this work is to classify the affected amino acid components and to find the dominating components and deficiencies of amino acids in the given sequence.

3.2 Amino Acid Component Based Classification

The AACC algorithm is based on the ID3 classifier. The protein sequence is a sequence made of English alphabets[5]. The sequence is classified into six components such as Sulfur, Aromatic, Aliphatic, Acidic, Basic and Neutral. This process is represented in Figure 1.

The existing methods classify the protein sequence based on their families. The proposed method, classifies the protein sequence into six components.

Apart from the classification, the proposed method finds the dominance and deficiencies of amino acids in the affected sequence. Gene sequence is given as the input to the proposed AACC algorithm. The AACC comprises of two phases, namely attribute selection and component based classification. The pseudo code of the overall algorithm is illustrated in the Figure 2.

At the end of these two phases the input sequence is classified into the components of amino acids. In addition, the dominance and deficiencies of amino acids are also computed.

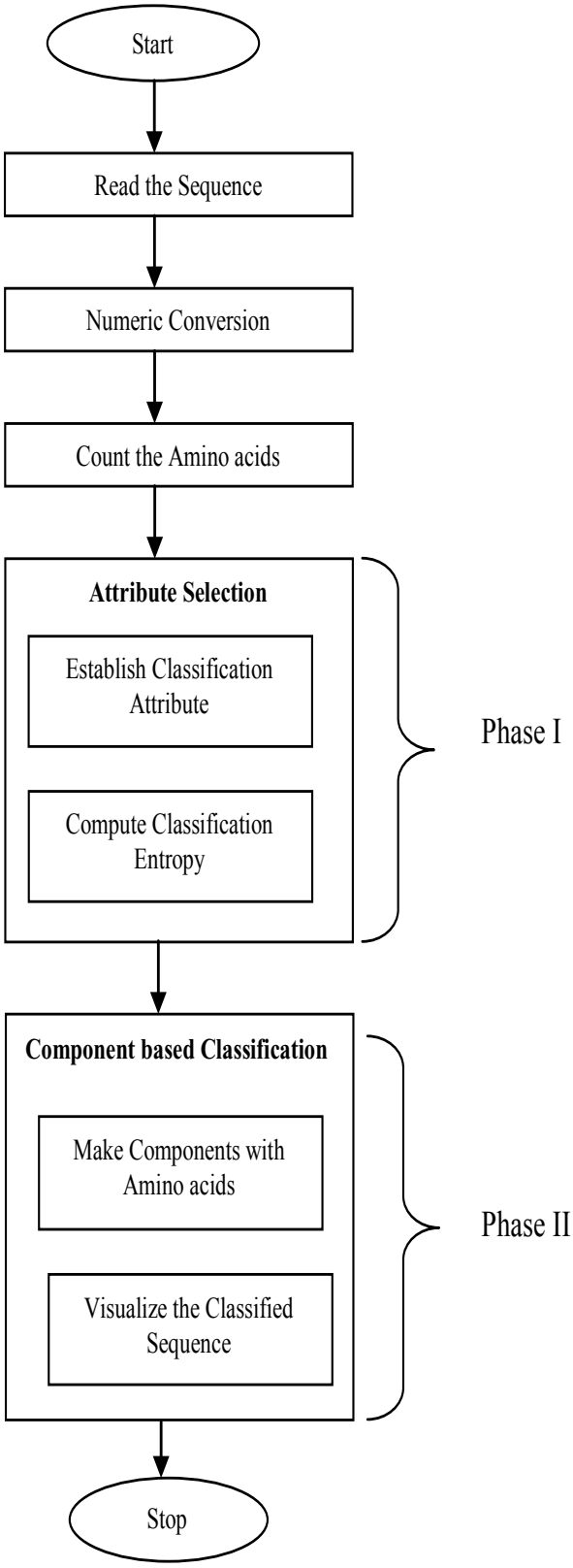

Figure 1. Block Diagram of AACC.

Input : *Gene Sequence with amino acids.*

Output : *Components of amino acids, dominance and deficiencies of amino acids, classification tree.*

Begin

Step 1 : *Read the sequence.*

Step 2 : *Convert gene sequence into numeric using random log value.*

Step 3 : *Count the amino acids (characters) in sequence.*

Step 4 : *Select the attribute as dominating and deficiencies of amino acids in the sequence by using their entropy values.*

Step 5 : *Make components with the combination of amino acids.*

Step 6 : *Visualize the classified sequence as components of amino acids and dominating and deficiencies of amino acids in the sequence.*

End

Figure 2. AACC Algorithm.

3.3 Attribute Selection

Attribute selection is the process of selecting a subset of relevant features for the construction of the classifier. In AACC algorithm, the attribute selection is the process of selecting the dominance and deficiencies of amino acids in the gene sequence using the ID3 classifier.

ID3 classifier is one of the decision tree classification algorithms for constructing a decision tree based on the information theory. Information theory is useful in finding the information content of an attribute in terms of entropy and information gain. The ID3 classifier uses these measures for selecting the attributes while constructing the decision tree.

3.3.1 Entropy

Entropy[11] is the measure of randomness in data or is the measurement of uncertainty for a piece of information. The value of the entropy is calculated using equation (1).

$$H(X) = -\sum_{k=0}^{n} P(X_k) \log P(X_k) \ldots (1)$$

Where, H(X) is the entropy of information X and P(X) denotes the probability of information X and n is the total number of attributes.

3.3.2 Information Gain

Information gain11 is the measure of expectation value of the conditional probability distribution. Information gain is sometimes called as mutual information. The information gain is calculated using equation (2).

$$Information\ Gain\ (T, a) = H(T) - H(\frac{a}{T}) \ldots (2)$$

where, T is the training sample and a is an attribute. H (T) is an entropy value of an attribute a and $H(\frac{a}{T})$ is the attribute available on the training sample. The mutual information or information gain is equal to the total entropy for an attribute.

The AACC calculates the entropy values and information gain for each amino acid in the given sequence. For the construction of the model, Dengue Serotype sequences are used as training data set. To construct the decision tree, the amino acids with the highest and lowest entropy values are chosen to be the root node and branch node respectively. The amino acid with the highest entropy value denotes the deficiencies of that amino acid. Similarly, the amino acid with the lowest entropy value denotes the dominating amino acid in that sequence.

3.4 Component Based Classification

Protein classification is the process of classifying the protein sequence based on their amino acid components Sulfur, Neutral, Alphatic, Aromatic, Basic and Acidic. To do this the researcher needs to know the components of the proteins. There exists several classification algorithms, such as SVM, decision tree and functionality based classifications for classifying the protein sequence into their amino acid components.

The AACC algorithm, classifies the protein sequence into components of amino acids. Neutral component produces the Asparagine, Serine, Threonine and Glutamine. Sulfur component produces the Cytosine and Methoine. Alphatic component produces the Leucine, Isoleucine, Glycine, Valine and Alanine. The Basic component produces the amino acids Arginine and Lysine. Acidic component produces Glutamic and Aspartic acids. Aromatic component produces the Phenylalanine, Tryptophan and Tyrosine.

In phase II the algorithm reads the sequence and count the number of amino acids in the sequence. Then it shows the corresponding components and visualizes the sequences as indicated in Table 1.

Table 1. Amino Acid Components and Amino Acids

Amino Acid Component	Amino Acids
Neutral	Asparagine, Serine, Threonine Glutamine
Sulfur	Cytosine, Methoine
Alphatic	Leucine, Isoleucine, Glycine, Valine, Alanine
Basic	Arginine, Lysine
Acidic	Glutamic, Aspartic
Aromatic	Phenylalanine, Tryptophan, Tyrosine

3.5 Classification Accuracy

Classification accuracy is the factor to assess the performances of any classification algorithm. For amino acid component based classification algorithm, the accuracy is calculated using equation (3).

$$ClassificationAccuracy = \frac{NCA}{TNA} \times 100 \ldots (3)$$

Where, NCA is the Number of Classified Amino acids and TNA is the Total Number of Amino acids in the sequence.

4. Experimental Results

In this section, the performance analyses of AACC algorithm is carried out in terms of classifier accuracy with several data sets and are compared with the existing bench mark algorithms.

4.1 Dataset for Experiments

The dengue virus sequences are used as training data set for AACC algorithm. For the training process, the protein sequence of National Centre for Bioinformatics (NCBI) is used. It contains 103 lines with 10,169 amino acids.

The dengue virus type I sequence contains 49 lines with 3391 characters. Each character depicts amino acids. Dengue virus type I sequence contains 49 lines with 3391 amino acids. Dengue virus type III sequence contains

RENSLSGVEGEGLHKLGYILRDISKIPGGNMYADDTAGWDTRITE
DDLQNEARITDIMEPEHALLAKSIFKLTYQNKVVRVQRPAKNGTV
MDVISRRDQRGSGQVGTYGLNTFTNMEAQLIRQMESEGIFSPSEL
ETPNLAERVLDWLEKYGVERLKRMAISGDDCVVKPIDDRFATAL
TALNDMGKVRKDIPQWEPSKGWNDWQQVPFCSHHFHQLIMKDG
REIVVPCRNQDELVGRARVSQGAGWSLRETACLGKSYAQMWQL
MYFHRRDLRLAANAICSAVPVDWIPTSRTTWSICLGKSYAQMWQ
LMYFHRSELET

Figure 3. Partial Amino Acid Sequence.

49 lines with 3390 amino acids. Dengue virus type IV sequence contains 49 lines with 3387 amino acids. The Figure 3 shows the part of the sequence of dengue serotype I.

4.2 Calculation of Entropy and Information Gain

The first step is to count the total number of amino acids in the Dengue serotype I sequence. There are 3489 amino acids in that sequence. The Table 2 depicts the counts of all the components in the given input dengue serotype I sequence.

Table 2. Counts of Components

Components	Counts
Neutral	694
Sulfur	183
Aliphatic	1255
Basic	393
Acidic	363
Aromatic	266

The entropy of overall sequence is computed using Equation (1) as below:

$$Entropy(OverallSequence) = -\frac{694}{3489}\log\left(\frac{694}{3489}\right) - \frac{183}{3489}\log\left(\frac{183}{3489}\right) - \frac{266}{3489}\log\left(\frac{266}{3489}\right) - \frac{363}{3489}\log\left(\frac{363}{3489}\right) - \frac{1255}{3489}\log\left(\frac{1255}{3489}\right) - \frac{393}{3489}\log\left(\frac{393}{3489}\right)$$

$$= 0.6641 \text{ Bits}$$

Table 3. Computation of Entropy Values

Amino Acid Components	Amino acids	Entropy ofAmino acids in Components	Entropy of Components in overall sequence
Neutral (N)	Asparagine (N)	0.0278	0.5764
	Serine (S)	0.0306	
	Threonine (T)	0.0318	
	Glutamine (Q)	0.0245	
Sulfur (S)	Cysteine (C)	0.0106	0.269382
	Methionine (M)	0.0075	
Alphatic (A)	Leucine (L)	0.0542	0.694
	Isoleucine (I)	0.0480	
	Glycine (G)	0.0519	
	Valine (V)	0.04712	
	Alanine (A)	0.04832	
Basic (B)	Arginine (R)	0.04014	0.60832
	Lysine (K)	0.028382	
Acidic (C)	Glutamic (E)	0.02101	0.63154
	Aspartic (D)	0.044692	
Aromatic (R)	Phenylalanine (F)	0.04414	0.30832
	Tryptophan (W)	0.03424	
	Tyrosine (Y)	0.02821	

The dominance and deficiencies of amino acids in dengue virus I can be found by computing the entropy of each component. The dominating amino acids will have the lowest entropy and the deficiencies of amino acids will have the highest entropy.

For example, the entropy value for Neutral components in dengue serotype I can be found using Equation (1). The Neutral component contains Asparagine, Serine, Threonine and Glutamine. Among the 694 Neutral components, the entropy of serine acid, whose count is 189, is computed as below:

$$Entropy(Serineacid) = \frac{694}{3489} \times \left(-\frac{189}{694}\log\left(\frac{189}{694}\right)\right)$$

$$= 0.0306 \text{ Bits.}$$

Similarly entropy of all amino acids in the dengue virus - I sequence can be found. The entropy values for all amino acids measured in terms of bits are listed in Table 3.

Then the information gain is calculated for all amino acids in dengue virus I. The information gain is the measurement of gain value of entropy for each amino acid and it is computed as the difference between the Entropy (overall sequence) and the Entropy (component). For example the Information gain of the Alanine is computed as:

$$InformationGain(Alanine) = 0.6641 - 0.04832$$

$$= \mathbf{0.6158\ Bits}$$

Table 4. Information Gain for all amino acids

Name of the Amino Acid Components	Name of the Amino Acids	Information Gain (Amino acids) in overall sequence	Information Gain (Amino Acid Components) in overall sequence
Neutral (N)	Asparagine (N)	0.6363	0.5486
	Serine (S)	0.6335	0.5458
	Threonine (T)	0.6323	0.5446
	Glutamine (Q)	0.6396	0.5519
Sulfur (S)	Cysteine (C)	0.6535	0.2588
	Methionine (M)	0.6566	0.261882
Alphatic (A)	Leucine (L)	0.609	0.6398
	Isoleucine (I)	0.6161	0.646
	Glycine (G)	0.6122	0.6421
	Valine (V)	0.6170	0.64688
	Alanine (A)	0.6158	0.64568
Basic (B)	Arginine (R)	0.62396	0.56818
	Lysine (K)	0.635718	0.579938
Acidic (C)	Glutamic (E)	0.64309	0.61053
	Aspartic (D)	0.619408	0.586848
Aromatic (R)	Phenylalanine (F)	0.61996	0.26418
	Tryptophan (W)	0.62986	0.27408
	Tyrosine (Y)	0.63589	0.28011

The information gain for all the amino acids in overall sequence and in components, are given in the Table 4 in the unit of bits.

In Table 4, the dominating amino acids are highlighted in bold letters which have the lowest information gain value. Deficiencies of amino acids are found by selecting the highest information gain value for every component.

At end of the algorithm, the dominating acids such as Threonine (T), Cysteine (C), Leucine (L), Arginine (R) Asparatic acid (D) and Phenylalanine (F) are identified. Also, the deficient amino acids are Glutamine (Q), Methionine (M), Valine (V), Lysine (K), Glutamic (E), and Tyrosine (Y).

4.3 Experimental Results

The AACC algorithm allocates the class labels such as Neutral, Sulfur, Alphatic, Basic, Acidic and Aromatic to the amino acid sequence. Initially, the dengue virus I sequence was fed as an input to the algorithm. In the next step, the total number of amino acids in that sequence is computed, followed by counts of the individual amino acid. The total count of that sequence is 3489. The counts of the individual amino acids were obtained as: Cysteine: 58, Aspartic acid: 135, Glutamic acid: 228, Phenylalanine: 99, Glycine: 283, Isoleucine: 196, Lysine: 220, Leucine: 320, Methionine: 122, Asparagine: 135, Glutamine: 112,

Arginine: 173, Serine: 178, Threonine: 278, Valine: 227, Tryptophan: 95, Tyrosine: 72 and Alanine: 236.

The counts of the components were computed by adding the counts of the appropriate amino acids. Figure 4 illustrates the resultant components and their counts.

The components are classified by plotting the components in the plot window. The components of the dengue serotype I sequence are plotted using appropriate colors. The Figure 5 represents the classified components in dengue serotype I sequence. The visualization of all the resultant amino acids can be done by computing the logarithmic values for all the counts of amino acids and then plotting all the amino acids in the dengue serotype I sequence. The Figure 6 depicts all the amino acids in the dengue serotype I sequence.

In Figure 5 every color plot depicts a component of amino acids. The numerical values of amino acids are obtained from the logarithmic values for the count of all the amino acids in the dengue virus sequence.

4.4 Classification Accuracy

Classification accuracy is the factor that indicates the efficiency of the classifier. Classification accuracy is computed as follows:

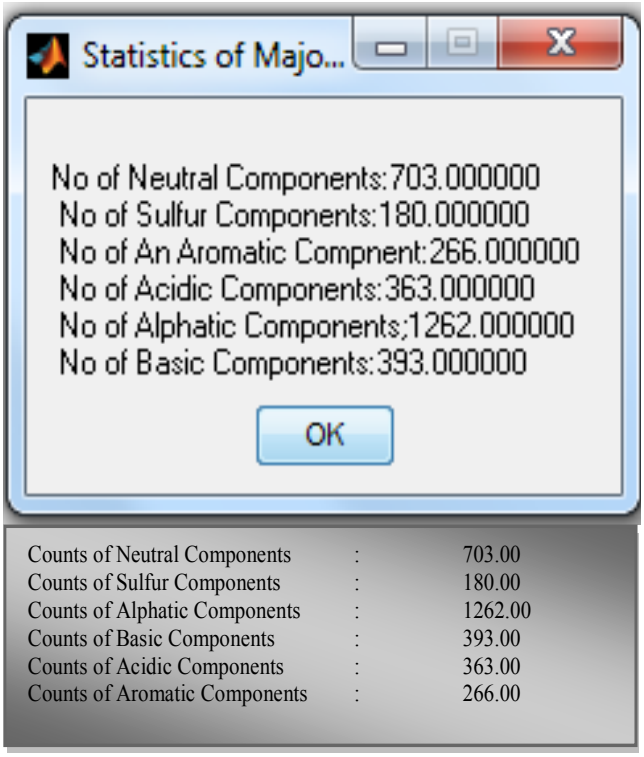

Figure 4. Six components of input sequence.

Total Number of Amino acids in the sequence=3489

Total Number of classified amino acids in the sequence=3167

$$Classsification\ accuracy = \frac{3167}{3489} \times 100$$

$$= 90.744\%$$

AACC algorithm provides 90.744% accuracy for the dengue serotype I sequence.

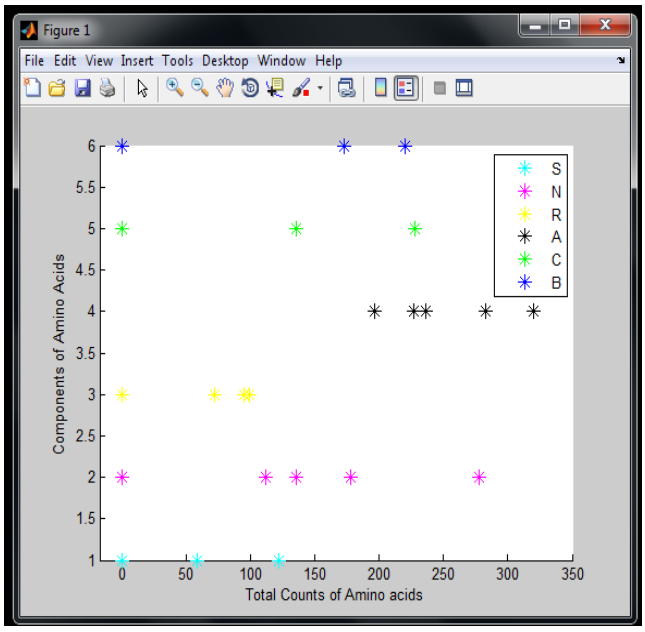

Figure 5. Classified Components in Dengue Virus I.

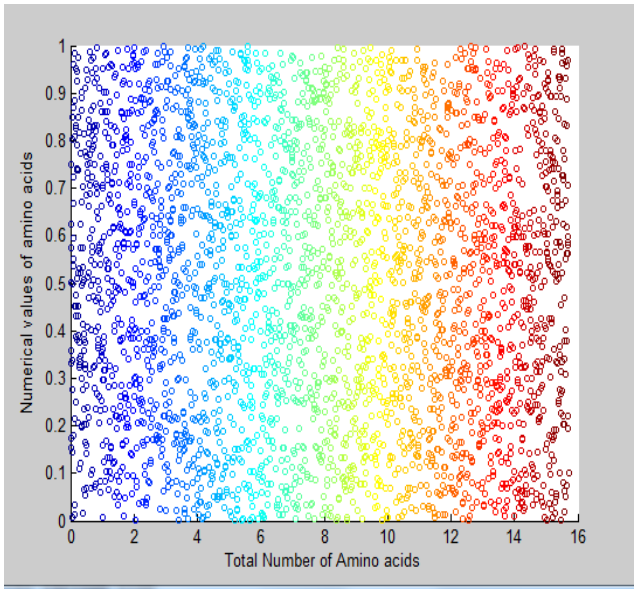

Figure 6. Classified Amino acids.

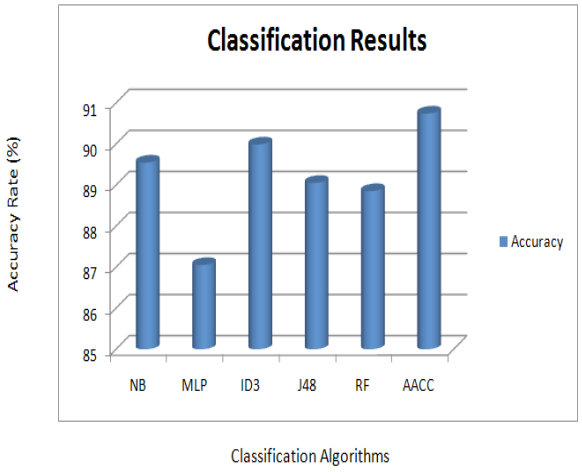

Figure 7. Comparative analysis of classification algorithms.

4.5 Comparative Analysis

The comparative analysis of AACC with other bench mark classification algorithms, such as Iterative Dichotomiser 3(ID3), Naive Bayes, Multi-Layer Perceptron, Random Forest (RF) and J48, is furnished in Figure 7.

5. Conclusion

The proposed AACC algorithm classified the given gene sequence based on the dominating amino acid components. The experiments were carried out by using the gene sequence of dengue virus. The accuracy of AACC was measured as 90.744%. It was compared with other traditional bench mark algorithms such as NB, ID3, J48, RF and MLP. The result of the comparative analysis revealed that AACC performed better than other methods while considering accuracy along with error rates. The results of this work can be used to design the drug for new viral diseases.

6. References

1. Archana S, Elangovan K. Survey of classification techniques in data mining. International Journal of Computer Science and Mobile Applications. 2013; 2(2):1024–33.
2. Aizenberg I, Zurada MJ. Solving selected classification problems in bioinformatics using multilayer neural network based on multi-valued neurons. Proceedings of International Conference on Artificial Neural Networks; New York, USA. 2007; 68(3):874–83.
3. Bashyam MD, Hasnain SE. The human genome sequence: Impact on health care. Indian Journal of Medical Research. 2003; 117(21):43–65.
4. Barnaghi PM, Sahzadi VA, Bakar AA. A comparative study for various methods of classification. International Conference on Information and Computer Networks. 2012; 27(2):875–81.
5. Bhardwaj R, Vatta S. Implementation of ID3 algorithm. International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(6):2277–81.
6. Chopra N. Enrolment data of disabled students of IGNOU: A case study using ID3. International Journal of Engineering Sciences. 2013; 8(3):222–8.
7. Exarchos TP, Papaloukas C, Lampros C, Fotiadis DI. Protein classification using sequential pattern mining. Proceedings of 28th IEEE EMBS Annual International Conference; New York, USA. 2006; 21(13):194–200.
8. Gupta M, Agarwal N. Classification Techniques and Analysis. Proceedings of National Conference on Computational Instrumentation; Chandigarh, India. 2010; 5(3):120–8.
9. Gengalakshmi S. Study on data mining for modeling biological system. Proceedings of National Conference on Innovations in communicative World; Coimbatore, India. 2012; 1(1):741–4.
10. Golmohammadi SK, Kurgan L, Crowley B, Reformat M. Amino acid sequence based method for prediction of cell membrane protein types. International Journal of Hybrid Information Technology. 2008; 1(2):108–15.
11. Han J, Kamber M. Data mining: Concepts and techniques. 2nd ed. Elsevier Publications; 2011. ISBN: 978-81-312-0535-81.
12. Leon F, Aignatoaies BI, Zaharia MH. Performance analysis of algorithms for protein structure classification. IEEE 20th International Workshop on Database and Expert Systems Applications of Computer Society; USA. 2009.
13. Mansoori EG, Zolghadri MJ, Katebi SD, Mohabatkar H, Boostani R, Sadreddini MH. Generating fuzzy rules for protein classification. Iranian Journal of Fuzzy Systems. 2008; 5(2):21–33.
14. Osman MH, Liong CY, Hashim I. Hybrid learning algorithm in neural network system for enzyme classification. International Journal of Advance Soft Computing Applications. 2010; 2(2):122–8.
15. Othman MF, Shan Yau TM. Comparison of different classification techniques using WEKA for breast cancer. Proceedings of International Federation of Medical and Biological Engineering. 2007; 1(1):56–61.

16. Priyanga A, Prakasam S. Effectiveness of data mining based cancer prediction system. International Journal of Computer Applications. 2013; 83(10):975–87.
17. Rajeswari J, Chandra E. A survey on data classification using machine learning techniques. International Journal of Engineering Science and Technology. 2011; 3(10):56–76.
18. Tao Li, Chengliang Z, Mitsunori O. A comparative study of feature selection and multiclass classification methods on gene expression. International Journal of Bioinformatics. 2004; 20(15):2429–37.