Development of Predictor for Sequence Derived Features from Amino Acid Sequence using Associate Rule Mining

Manpreet Singh mpreet78@yahoo.com
Department of Information Technology
Guru Nanak Dev Engineering College
Ludhiana, Punjab, India - 141006

Gurvinder Singh gsbawa71@yahoo.com
Department of Computer Science and Engineering
Guru Nanak Dev University
Amritsar, Punjab, India

Abstract

Drug Discovery process include target identification i.e. to identify a target protein whose inhibition can destroy the pathogen. In testing phase, clinical and pre-clinical trials are done on the animals and then on humans. After the discovery process, the drug or medicine is made available for public use. But if the testing of the drug is ineffective or unable to yield the appropriate results, then the whole process need to be repeated. This makes the first stage of drug discovery the most important than the other stages. The present work will assist in the process of drug discovery.

The present work involves the development of a model that extracts the sequence derived features from the given amino acid sequence using associative rule mining. Associative rule mining is a data mining technique useful to identify related items and to develop rules. In the present work, various parameters of the amino acid sequence are studied that affect the sequence-derived features and some of the equations and algorithms are implemented. Input is given through text file and collective results are obtained. MATLAB environment is used for the implementation. The results are compared with the previous bioinformatics tools. The model developed assists in protein class prediction process which assists drug discoverers in the drug discovery process.

Keywords: Drug Discovery, Associative Rule Mining, Amino Acid, Sequence Derived Features.

1. ASSOCIATIVE RULE MINING

Association rule mining is a common data mining technique, which can be used to produce interesting patterns or rules [2]. Association rule mining involves counting frequent patterns (or associations) in large databases, reporting all that exist above a minimum frequency threshold known as the 'support' e.g. analyzing supermarket basket data, where a supermarket would want to see which products are frequently bought together. Such an association might be "if a customer buys biscuits and patty then they are 80% likely to buy coffee". Rule support and confidence are two measures of rule interestingness. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold [5].

1.1 Association Rules

Association rules are required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. To achieve this, association rule generation is a two-step process. First, minimum support is applied to find all frequent itemsets in a database. In a second step, these frequent itemsets and the minimum confidence constraint are used to form rules. While the second step is straight forward, the first step needs more attention. The problem is defined as: Let I be a set of n binary attributes called items. Let D be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form where X, Y C I and X ∩Y= θ. The sets of items (for short item sets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

Finding all frequent itemsets in a database is difficult since it involves searching all possible itemsets (item combinations). The set of possible itemsets is the power set over I and has size 2n – 1 (excluding the empty set which is not a valid itemset). Although the size of the powerset grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent itemset also all its subsets are frequent and thus for an infrequent itemset, all its supersets must be infrequent [13][16].

2. PROTEIN AND PROTEIN FUNCTIONS

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function and regulation of the body's tissues and organs.

Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function [6-8].

As amino acids band together in chains to form the stuff from which our life is born. It's a two-step process: Amino acids get together and form peptides or polypeptides. It is from these groupings that proteins are made. Commonly recognized amino acids include glutamine, glycine, phenylalanine, tryptophan, and valine. Three of those — phenylalanine, tryptophan, and valine — are essential amino acids for humans; the others are isoleucine, leucine, lysine, methionine, and threonine. The essential amino acids cannot be synthesized by the body; instead, they must be ingested through food.

They serve as enzymatic catalysts, are used as transport molecules (hemoglobin transports oxygen) and storage molecules (iron is stored in the liver as a complex with the protein ferritin), they are used in movement (proteins are the major component of muscles), they are needed for mechanical support (skin and bone contain collagen-a fibrous protein), they mediate cell responses (rhodopsin is a protein in the eye which is used for vision), antibody proteins are needed for immune protection; control of growth and cell differentiation uses proteins (hormones) [4].

3. SEQUENCE DERIVED FEATURES

Sequence derived features are the various features of protein which are used to predict protein class. Sequence derived features are very important in protein class prediction as these are the input to the HPF predictor. SDF's can be derived from a given set of amino-acid (protein) sequences using various web-based bioinformatics tools [12]. The various sequence derived features are as given below:

3.1 Extinction Coefficient (Eprotein)

Extinction Coefficient is a protein parameter that is commonly used in the laboratory for determining the protein concentration in a solution by spectrophotometry. It describes to what extent light is absorbed by the protein and depends upon the protein size and composition as well as the wavelength of the light. For proteins measured in water at wavelength of 280nm, the value of the Extinction coefficient can be determined from the composition of Tyrosine, Tryptophan and Cystine.

Mathematically:

$$Eprotein = Ntyr * Etyr + Ntrp * Etrp + Ncys * Ecys \qquad (1)$$

Where Etyr=1490, Etrp =5500, Ecys =125 are the Extinction coefficients of the individual amino acid residues.

3.2 Absorbance (Optical Density)

For proteins measured in water at wavelength of 280nm the absorbance can be determined by the ratio of Extinction coefficient and the molecular weight of the protein. It is a representation of a material's light blocking ability.

Mathematically:

$$\text{Absorbance} = \text{Eprotein} / \text{Molecular Weight} \qquad (2)$$

3.3 Number of Negatively Charged Residues (Nneg)

This can be calculated from the composition of Aspartic acid and Glutamic acid.

3.4 Number of Positively Charged Residues (Npos)

This can be calculated from the composition of ARginine and Lysine.

3.5 Aliphatic Index (AI)

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular.

Mathematically:

$$AI = Xala + a * Xval + b * (Xile + Xleu) \qquad (3)$$

Where Xala , Xval , Xile and Xleu are the mole percentages of alanine, valine, isoleucine and leucine respectively. Coefficients a and b are the relative volume of valine side chain and of leu/ile side chains to the side chain of alanine i.e. a = 2.9 and b = 3.9.

3.6 Compute IP/mol weight

It calculates the isoelectric point by molecular weight of the input amino acid sequence. IP stands for isoelectric point of the input amino acid sequence. Mol weight stands for molecular weight of the input amio acid sequence.

4. LITERATURE SURVEY

Jensen et al. (2002) proposed the human protein function from post-translational modifications and localization features. The prediction method involved the use of sequence derived features for human protein function prediction. The posttranslational modifications (PTMs) are the changes that occur to the protein after its production by the process of translation. They extracted the sequence derived features from the different servers like Expasy, PSORT as discussed in section 5. Fourteen features were extracted from the amino acid sequences [6].

Al-Shahib et al. (2007)

Calculated the frequency, total number of each amino acid and the set of amino acids for the input protein seqnence. To encode distributional features, they also determined the number and size of continuous stretches of each amino acid or amino acid set. They subdivided every protein into four equally sized fragments and calculated the same feature values for each fragment and combination of fragments. In addition, the other features like the secondary structure was predicted using Prof [10], the position of putative transmembrane helices using TMHMM [21] and of disordered regions using DisEMBL [15]. The features were used for protein function prediction [1].

Kanakubo et al. (2007)

Stated that association rule mining was one of the most important issues in data mining. With Apriori methods, the problem becomes incomputable when the total number of items are large. On the other hand, bottom-up approaches such as artificial life approaches were opposite of the top-down approaches of searches covering all transactions and may provide new methods of breaking away from the completeness of searches in conventional algorithms. Here, an artificial life data mining technique was proposed in which one transaction was considered as one individual and association rules were accumulated by the interaction of randomly selected individuals. The proposed algorithm was compared to other methods in application to a large scale actual dataset and it was verified that its performance was greatly superior to that of the method using transaction data virtually divided and that of apriori method by sampling approach, thus demonstrating its usefulness [9].

Gupta et al. (2008)
Proposed a novel feature vector based on physicochemical property of amino acids for prediction protein structural classes. They presented a wavelet-based time-series technique for extracting features from mapped amino acid sequence and a fixed length feature vector for classification is constructed. Wavelet transform is a technique that decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. The proposed feature vector contains information about the variability of ten physiochemical properties of protein sequences over different scales. The variability of physiochemical properties was represented in terms of wavelet variance [14].

Jaiswal et al. (2011)
Studied that the identification of specific target proteins for any diseased condition involves extensive characterization of the potentially involved proteins. Members of a protein family demonstrating comparable features may show certain unusual features when implicated in a pathological condition. They studied the Human matrix metalloproteinase (MMP) family of endopeptidases and discovered their role in various pathological conditions such as arthritis, atherosclerosis, cancer, liver fibrosis, cardio-vascular and neurodegenerative disorders, little is known about the specific involvement of members of the large MMP family in diseases. They hypothesized that cysteine rich and highly thermostable MMPs might be key players in diseased conditions and hence signify the importance of sequence derived features [3].

5. FEATURE EXTRACTION TOOLS

There are various Bioinformatics Tools for obtaining Sequence derived features (SDFs) [11]. These are as follows:

5.1 NetNGlyc 1.0 Server

It predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-ser/Thr sequins [18].

5.2 PSORT Server

It is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source origin e.g. Gram-negative bacteria, as inputs. Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information [23].

5.3 TMHMM Server

It is a program for predicting transmembrane helices based on a hidden Markov model. It reads a FASTA formatted protein sequence and predicts locations of transmembrane, intracellular and extracellular regions. (http://www.cbs.dtu.dk/services/TMHMM/) [21].

5.4 NetOGlyc Server

It produces neural network predictions of mucin type GalNAc O-Glycosylation sites in mammalian proteins [19].

5.5 Signal-P server

It predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models [20].

5.6 Expasy Server

It computes various physico-chemical properties of protein like iso-electric point, extinction coefficient, optical density etc [22].

5.7 PROFEAT

It is a web server for computing commonly-used structural and physicochemical features of proteins and peptides from amino acid sequence. It includes amino acid composition and dipeptide composition, normalized Moreau–Broto autocorrelation, Moran autocorrelation and Geary autocorrelation, composition, transition and distribution, sequence-order-coupling number, quasi-sequence-order [17].

6. ALGORITHM FOR PREDICTING SEQUENCE DERIVED FEATURES

Among various Sequence derived features are Extinction Coefficient, Aliphatic Index, Absorbance, No. of negatively charged residues, No. of positively charged residues, Compute Iso-electric point/molecular weight. These SDF's are integrated and computed in one platform using Associative rule mining.

In the existing techniques the various categories of features were not computed for the same input by any single tool rather different tools are available for different categories as discussed in section 5. The methods for extracting multiple features of relevant to the density of the amino acids were also developed [1][14]. The present work focuses on developing the single tool for extracting the features of multiple categories from the single input file. The algorithm predicts the Extinction Coefficient, Aliphatic Index, Absorbance, No. of negatively charged residues, No. of positively charged residues, Compute Iso-electric point/molecular weight from the input amino acid sequence. All these features are computed by giving single input sequence file which is type of string made from possible combinations of 20 characters representing the 20 amino acids. It is possible by integrating all these computations by using associative rule mining. The amino acid sequences may contain thousands of amino acid i.e characters in a string. Thus extracting any sequence derived feature from the sequence is a compute intensive problem. Computing multiple features from single sequence is even more herculean task. Each computation has some common part that is called the intersection. The intersection and union operations are utilized to give the integrated results from the single input file.

The different algorithms for individual feature prediction are shown here for clear interpretation. Fig. 1 shows the steps involved in Extinction Coefficient Prediction. Fig. 2 and Fig. 3 shows the algorithm for prediction of negatively charged and positively charged residues respectively. Fig. 4 shows the predictor for Iso-electric point and molecular weight for the amino acid chain. Fig. 5 shows the Aliphatic Index prediction and Fig. 6 shows the Absorbance prediction for the input file.

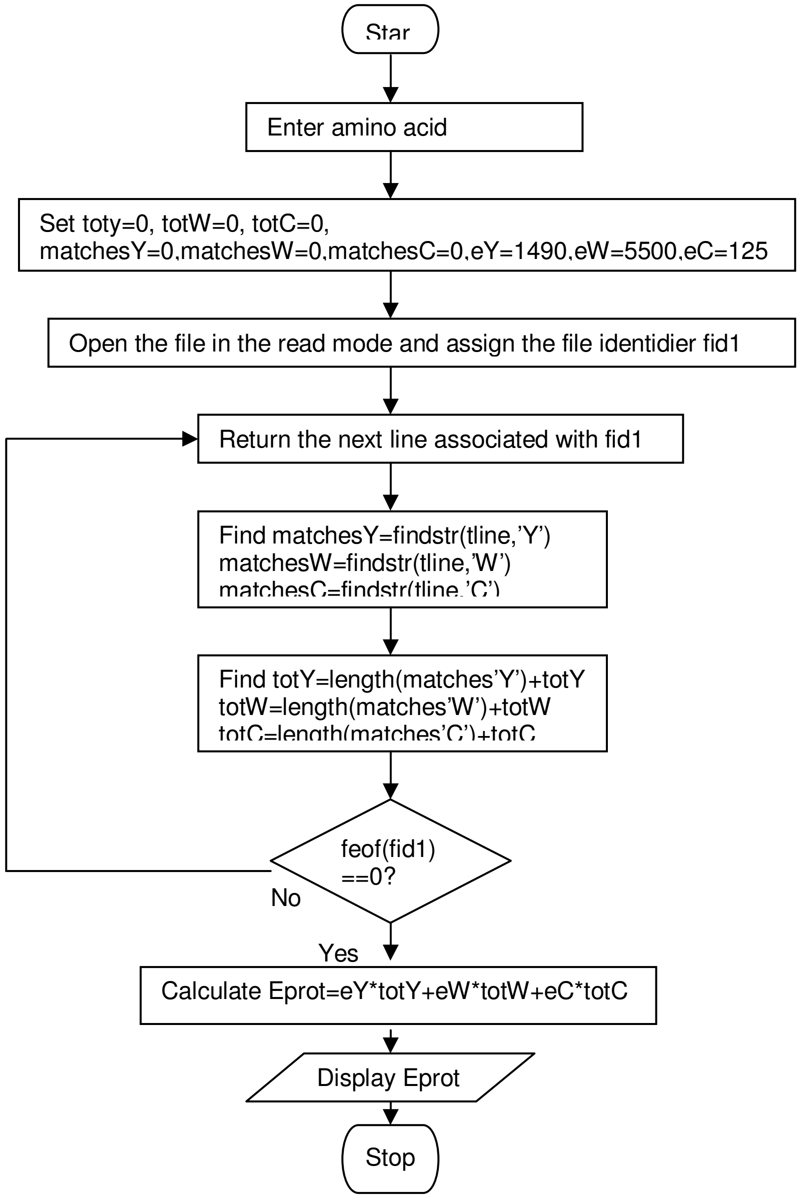
Star
Enter amino acid
Set toty=0, totW=0, totC=0,
matchesY=0,matchesW=0,matchesC=0,eY=1490,eW=5500,eC=125
Open the file in the read mode and assign the file identidier fid1
Return the next line associated with fid1
Find matchesY=findstr(tline,'Y')
matchesW=findstr(tline,'W')
matchesC=findstr(tline,'C')
Find totY=length(matches'Y')+totY
totW=length(matches'W')+totW
totC=length(matches'C')+totC
feof(fid1)
==0?
No
Yes
Calculate Eprot=eY*totY+eW*totW+eC*totC
Display Eprot
Stop

FIGURE 1: Prediction of Extinction Coefficient.

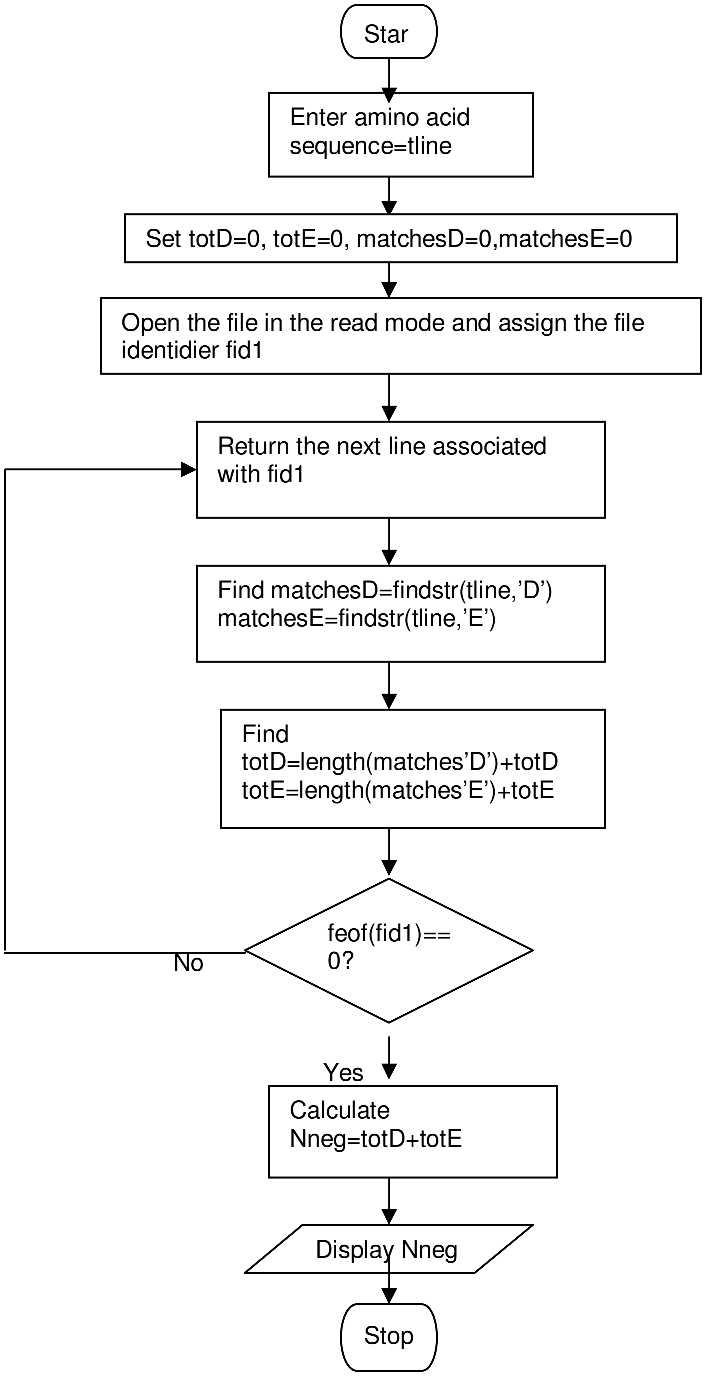
Star
Enter amino acid sequence=tline
Set totD=0, totE=0, matchesD=0,matchesE=0
Open the file in the read mode and assign the file identidier fid1
Return the next line associated with fid1
Find matchesD=findstr(tline,'D')
matchesE=findstr(tline,'E')
Find
totD=length(matches'D')+totD
totE=length(matches'E')+totE
feof(fid1)== 0?
No
Yes
Calculate
Nneg=totD+totE
Display Nneg
Stop

FIGURE 2: Prediction of Negatively charged residues

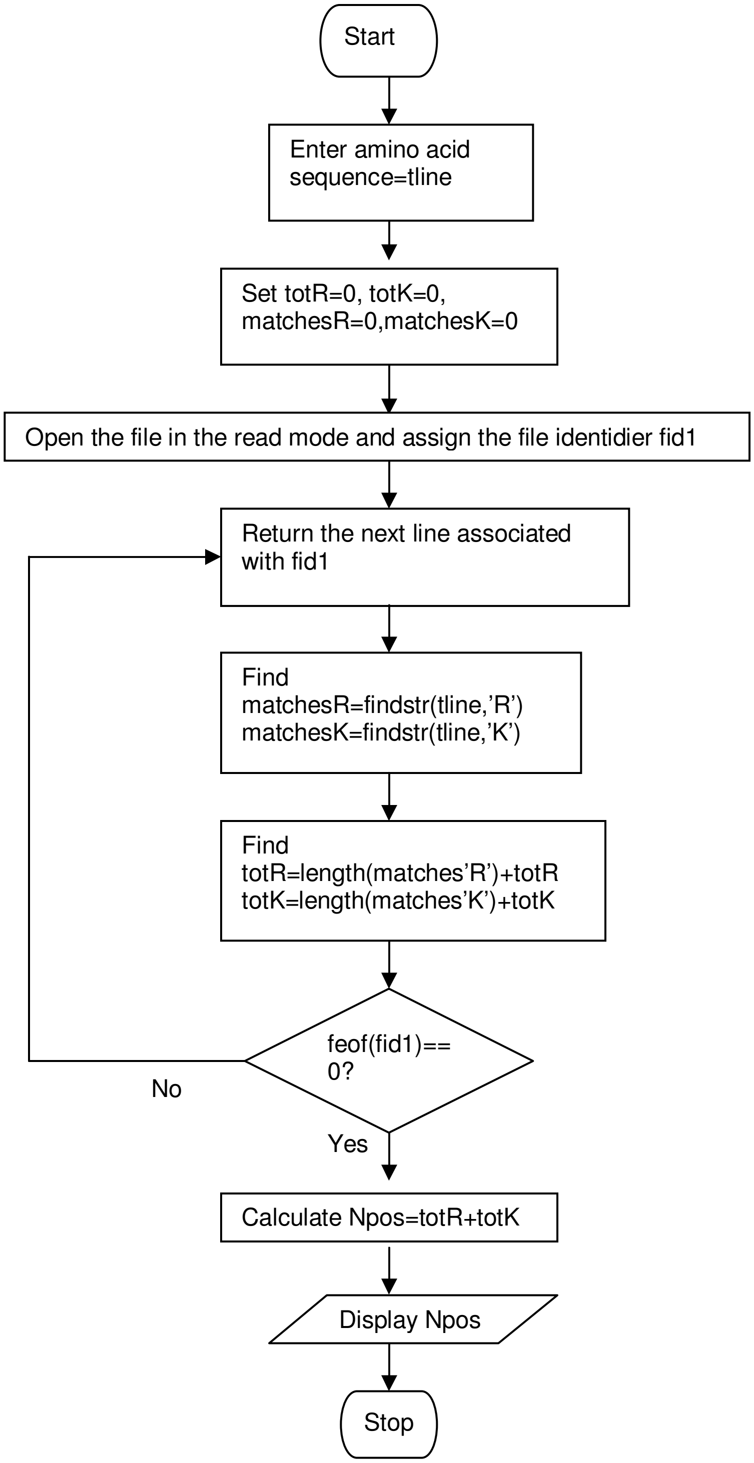
Start
Enter amino acid sequence=tline
Set totR=0, totK=0, matchesR=0,matchesK=0
Open the file in the read mode and assign the file identidier fid1
Return the next line associated with fid1
Find matchesR=findstr(tline,'R') matchesK=findstr(tline,'K')
Find totR=length(matches'R')+totR totK=length(matches'K')+totK
feof(fid1)== 0?
No
Yes
Calculate Npos=totR+totK
Display Npos
Stop

FIGURE 3: Prediction of Positively charged residues

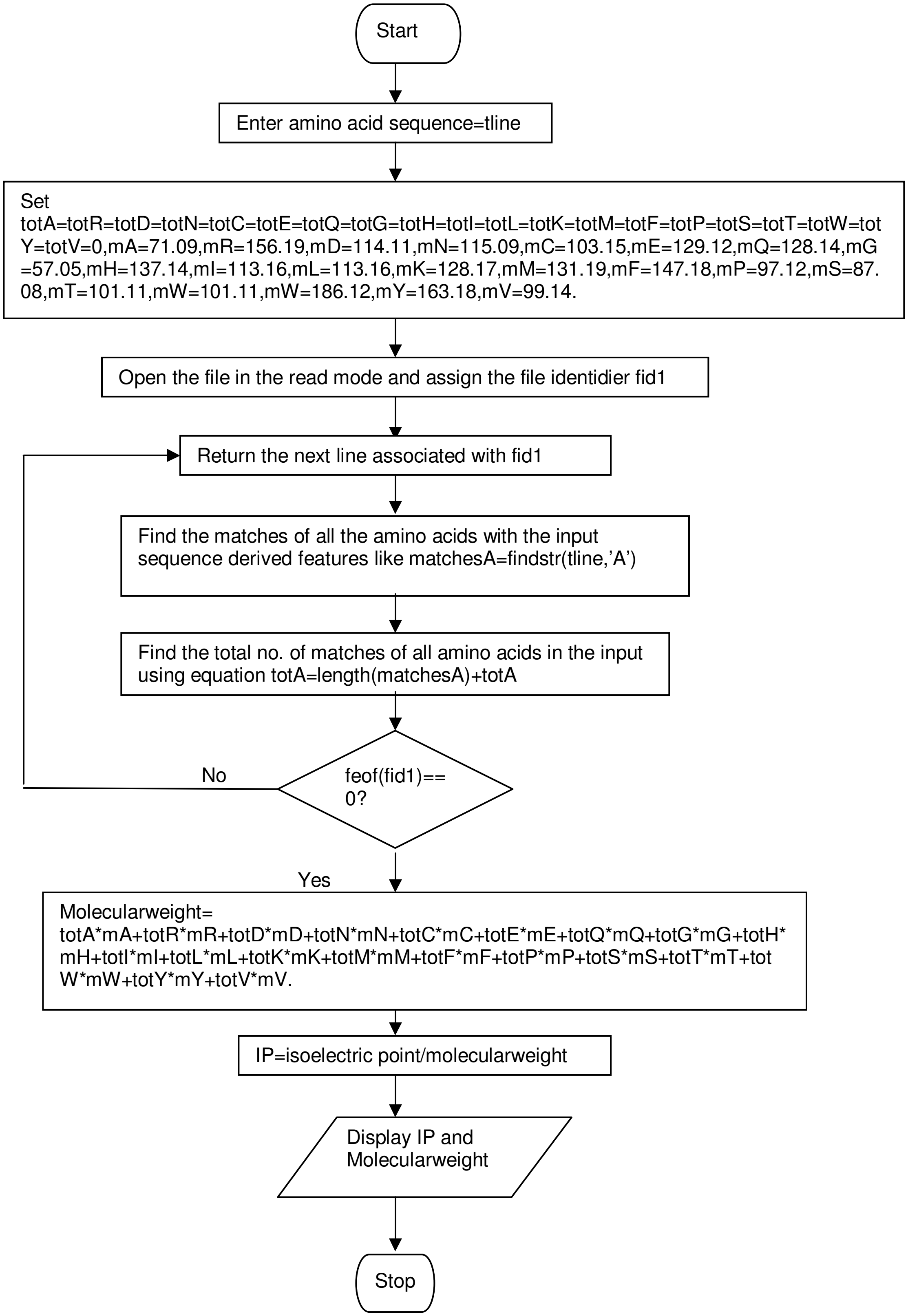
Start
Enter amino acid sequence=tline
Set
totA=totR=totD=totN=totC=totE=totQ=totG=totH=totI=totL=totK=totM=totF=totP=totS=totT=totW=totY=totV=0,mA=71.09,mR=156.19,mD=114.11,mN=115.09,mC=103.15,mE=129.12,mQ=128.14,mG=57.05,mH=137.14,mI=113.16,mL=113.16,mK=128.17,mM=131.19,mF=147.18,mP=97.12,mS=87.08,mT=101.11,mW=101.11,mW=186.12,mY=163.18,mV=99.14.
Open the file in the read mode and assign the file identidier fid1
Return the next line associated with fid1
Find the matches of all the amino acids with the input sequence derived features like matchesA=findstr(tline,'A')
Find the total no. of matches of all amino acids in the input using equation totA=length(matchesA)+totA
No
feof(fid1)== 0?
Yes
Molecularweight=
totA*mA+totR*mR+totD*mD+totN*mN+totC*mC+totE*mE+totQ*mQ+totG*mG+totH*mH+totI*mI+totL*mL+totK*mK+totM*mM+totF*mF+totP*mP+totS*mS+totT*mT+totW*mW+totY*mY+totV*mV.
IP=isoelectric point/molecularweight
Display IP and Molecularweight
Stop

FIGURE 4: Prediction of Iso-electric point/Mol weight

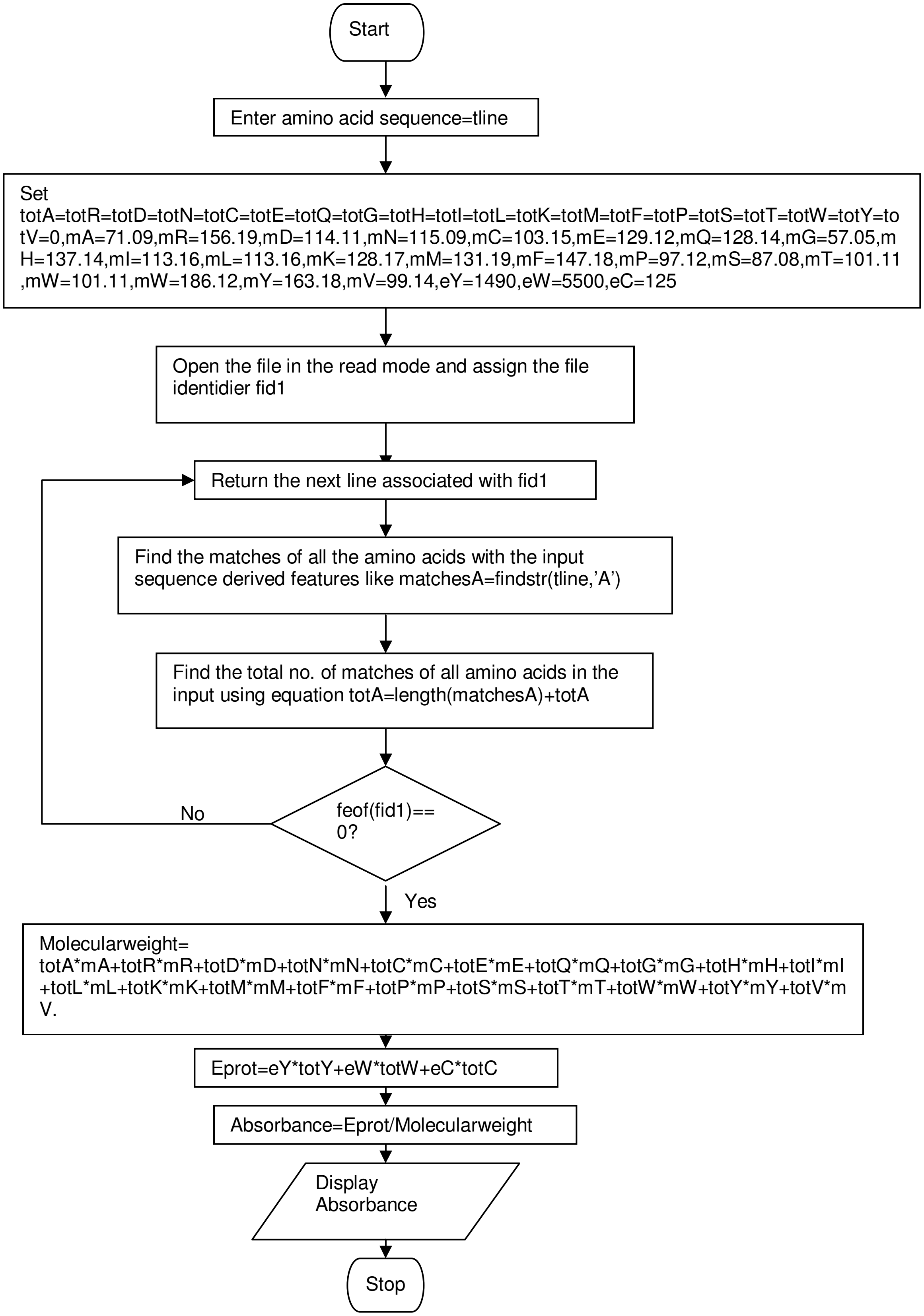
Start
Enter amino acid sequence=tline
Set
totA=totR=totD=totN=totC=totE=totQ=totG=totH=totI=totL=totK=totM=totF=totP=totS=totT=totW=totY=totV=0,mA=71.09,mR=156.19,mD=114.11,mN=115.09,mC=103.15,mE=129.12,mQ=128.14,mG=57.05,mH=137.14,mI=113.16,mL=113.16,mK=128.17,mM=131.19,mF=147.18,mP=97.12,mS=87.08,mT=101.11,mW=101.11,mW=186.12,mY=163.18,mV=99.14,eY=1490,eW=5500,eC=125
Open the file in the read mode and assign the file identidier fid1
Return the next line associated with fid1
Find the matches of all the amino acids with the input sequence derived features like matchesA=findstr(tline,'A')
Find the total no. of matches of all amino acids in the input using equation totA=length(matchesA)+totA
No
feof(fid1)== 0?
Yes
Molecularweight= totA*mA+totR*mR+totD*mD+totN*mN+totC*mC+totE*mE+totQ*mQ+totG*mG+totH*mH+totI*mI+totL*mL+totK*mK+totM*mM+totF*mF+totP*mP+totS*mS+totT*mT+totW*mW+totY*mY+totV*mV.
Eprot=eY*totY+eW*totW+eC*totC
Absorbance=Eprot/Molecularweight
Display Absorbance
Stop

FIGURE 5: Prediction of Absorbance

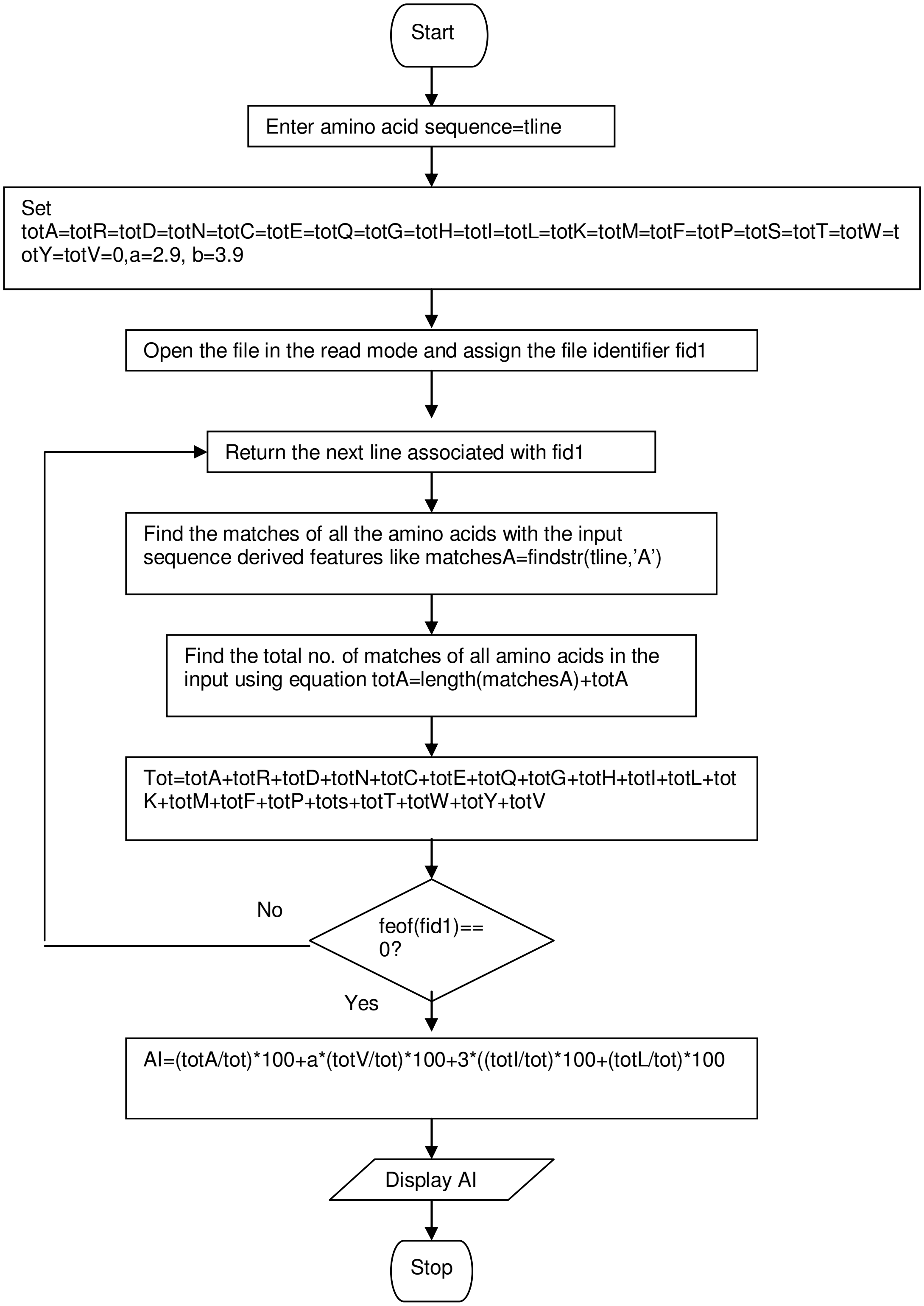
Start
Enter amino acid sequence=tline
Set totA=totR=totD=totN=totC=totE=totQ=totG=totH=totI=totL=totK=totM=totF=totP=totS=totT=totW=totY=totV=0,a=2.9, b=3.9
Open the file in the read mode and assign the file identifier fid1
Return the next line associated with fid1
Find the matches of all the amino acids with the input sequence derived features like matchesA=findstr(tline,'A')
Find the total no. of matches of all amino acids in the input using equation totA=length(matchesA)+totA
Tot=totA+totR+totD+totN+totC+totE+totQ+totG+totH+totI+totL+totK+totM+totF+totP+tots+totT+totW+totY+totV
No
feof(fid1)==0?
Yes
AI=(totA/tot)*100+a*(totV/tot)*100+3*((totI/tot)*100+(totL/tot)*100
Display AI
Stop

FIGURE 6: Prediction of Aliphatic Index

7. RESULTS AND DISCUSSION

In the present work, the amino acid sequence is used as input to predict the sequence derived features. Some of these sequence derived features are produced from amino acid sequence by exploring different parameters. In this present work, output will be displayed if the blank spaces are included in the input sequence and output will not be displayed if lowercase alphabets of amino acid sequence are given as input. The respective results are shown in table 1.
The input of amino acid sequence in ex1.txt file is given below:

```
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGT
NLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYY
TIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVL
ALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYF
SIILAFLPIAGXIENY
```

Sr. No.	Sequence Derived Features obtained	Values
1.	No. of Amino Acids	283
2.	Molecular weight	3.1955e+504
3.	Number of Negatively charged residues	15
4.	Number of positively charged residues	11
5.	Extinction Coefficient	66475
6.	Absorbance	2.0803
7.	Aliphatic Index	119.6113
8.	Compute IP/Mw	1.2048e-004

TABLE 1: Sequence Derived Features Produced

Calculations on ex1.txt file:
Number of Negatively charged residues:
Composition of Aspartic acid (D) and Glutamic acid (E) for ex1.txt:
No. of Aspartic acid in ex1=9
No. of Glutamic acid in ex1=6
Number of Negatively charged residues for ex1 =9+6=15
Number of positively charged residues:
Composition of Arginine (R) and Lysine (K) for ex1.txt:
No. of Arginine in ex1=5
No. of Lysine in ex1=6
Number of Positively charged residues for ex1 =5+6=11
Extinction Coefficient:
The equation to calculate Extinction Coefficient is as given below:
Eprotein = Ntyr * Etyr + Ntrp * Etrp+ Ncys * Ecys
Where Etyr =1490, Etrp =5500, Ecys = 125
No. of Tyrosine (Y) in ex1=Ntyr= 15

No. of Tryptophan (W) in ex1=Ntrp = 8
No. of Cystine (C) in ex1=Ncys= 1
Putting these values in the above equation:
Eprotein=15*1490+8*5500+1*125 = 66475
Aliphatic Index:
The equation to calculate the Aliphatic Index as given below:
AI = Xala + a * Xval + b * (Xile + Xleu)
Where a = 2.9, b = 3.9
Mole Percentage of Alanine(A) in ex1= Xala =(17/283)*100=6.007
Mole Percentage of Valine (V) in ex1= Xval = (10/283)*100=3.533
Mole Percentage of Isoleucine(I) in ex1= Xile =(25/283)*100=8.834
Mole Percentage of Leucine (L) in ex1= Xleu = (50/283)*100=17.666
Putting these values in the above equation:
AI= 6.007+2.9*3.533+3.9(8.834+17.666) =0.9542
Absorbance/Optical Density:
The equation to calculate the Absorbance is as given below:
Absorbance = Eprotein / Molecular Weight
Eprotein for ex1 (as calculated above) =66475
Molecular weight for ex1=3.1955e+504
Putting these values in the above equation:
Absorbance =66475/3.1955e+504=2.0803
Compute IP/Mw:
Calculate Isoelectric of the amino acid sequence/Molecular weight
Isoelectric point of ex1=24.742e
Molecular weight of ex1=3.1955e+504
Compute IP/Mw=1.2048e-004

8. CONCLUSION AND FUTURE SCOPE

This present work is designed and implemented for prediction of sequence derived features which are used for protein class prediction that is further useful in drug discovery process. In this various sequence derived features are studied and integrated using Associative rule mining. The model is very simple to use and no manual work is involved. The results have been verified by comparing their output with the previously available tools. The future scope for further work is listed below:
a) Other 2-D and 3-D protein structure prediction algorithms can be included to predict the protein function.
b) Protein Class Prediction can be done from the input sequence itself.

9. REFERENCES

1. A. Al-Shahib, R. Breitling, and D. R. Gilbert "*Predicting protein function by machine learning on amino acid sequences – a critical evaluation*" BMC Genomics, 8:1-10, 2007

2. A. Clare. "*Machine learning and data mining for yeast functional genomics*", Ph.D. thesis, University of Wales, February 2003

3. A. Jaiswal, A. Chhabra, U. Malhotra, S. Kohli, V. Rani "*Comparative analysis of human matrix metalloproteinases: Emerging therapeutic targets in diseases*" Bioinformation 6(1): 23-30, 2011

4. D. Krane and M. Raymer. "*Fundamental Concepts of Bioinformatics*", Pearson Education, New Delhi, pp.1-314 (2006)

5. J. Han and M. Kamber. "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, pp. 226-229 (2004)

6. L. Jensen. "*Prediction of Protein Function from Sequence Derived Protein Features*", Ph.D. thesis, Technical University of Denmark, 2002

7. L. Jensen, M. Skovgaard and S. Brunak. "*Prediction of Novel Archaeal Enzymes from Sequence Derived Features*", Protein Science, 11: 2894-2898, 2002

8. L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H.H. Starfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Knudsen, A. Krogh, A. Valencia and S. Brunak "*Prediction of Human Protein Function from Post-Translational Modifications and Localization Features*" Journal of Molecular Biology, 319(5): 1257-1265, 2002

9. M. Kanakubo and M. Hagiwara. "*Speed up technique for Associative rule mining based on an Artificial Algorithm*", GRC book on granular computing, 38(12):318-323, 2007

10. M. Ouali, R.D. King "*Cascaded multiple classifiers for secondary structure prediction*" Prot Sci., 9:1162–1176, 2000

11. M. Singh, P. Singh and P.K, Wadhwa "*Human Protein Function Prediction using Decision Tree Induction*" International Journal of Computer Science and Network Security, USA, 7(4):92-98, 2007

12. M. Singh, Wadhwa P.K., Surinder Kaur "*Predicting Protein Function using Decision Tree*" World Academy of Science, Engineering and Technology, 39:350-353, 2008

13. R. Agrawal, T. Imielinski and A. Swami. "*Mining Association Rules Between Sets of Items in Large Databases*", SIGMOD ACM Conference, 22(2):207-216, 1993

14. R. Gupta, A. Mittal, and K. Singh. "*Time series based feature extraction approach for prediction of protein structural class*", EURASIP Journal, 8(1): 1-7, 2008

15. R. Linding, L. J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell "*Protein disorder prediction: implications for structural proteomics*" Structure, 11:1453-1459, 2003

16. Veenu Mangat "*Swarm Intelligence Based Technique for Rule Mining in the Medical Domain*" International Journal of Computer Applications, 4(1):19-24, July 2010

17. Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, Y.Z. Chen. "*PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence*" Nucleic Acids Res, 34:W32-W37, 2008

18. http://www.cbs.dtu.dk/services/NetNGlyc/

19. http://www.cbs.dtu.dk/services/NetOGlyc/

20. http://www.cbs.dtu.dk/services/SignalP/

21. http://www.cbs.dtu.dk/services/TMHMM/

22. http://expasy.org/

23. http://psort.hgc.jp/