Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction

Manpreet Singh
Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, Punjab, INDIA

Dr. Gurvinder Singh
Department of Computer Science and Engineering, Guru Nanak Dev University, Amritsar, Punjab, INDIA

ABSTRACT
In the present paper, the cluster analysis as a form of unsupervised learning is implemented for human protein class prediction. The data related to human protein is accessed from Human Protein Reference Database (HPRD). From HPRD, the sequences related to ten molecular classes are obtained. For each of the molecular class five amino acid sequences are obtained. Then with the help of various web based tools, SDFs (Sequence derived Features) are extracted for each sequence. By analyzing the variation in the values of the obtained SDFs, priorities are assigned to them. Because each sequence has some value for each of the SDF, so obtained data is a complete weighted bipartite graph consisting of two independent set of nodes i.e. one set of all the sequences and second of all SDFs. Then bipartite graph is represented into the memory with adjacency weight matrix. On the basis of values of input SDFs and by considering priority of each of the SDF, clusters of the data available in the adjacency matrix are generated. Then those clusters are backtracked to predict the class of the entered sequence.

General Terms
Bioinformatics, Machine Learning, Human Protein Class Prediction.

Keywords
Protein class prediction, cluster analysis, bipartite graph

1. INTRODUCTION
Protein class prediction is helpful in the process of drug discovery. In drug discovery, it is very difficult to find out the complementary protein for each protein individually. But if the class of the protein will be known for which the drug is to be discovered then it will become very easy to find the complementary protein sequence which can be attached to the active site of the protein to stop it to expand. So class prediction of a protein helps to enhance the process of drug discovery. [1-2]

1.1 Previous Techniques
The computational techniques developed for predicting the structure and functions of unknown proteins are as follows:

- The QM/MM scheme i.e. the Quantum Mechanical/Molecular Mechanical scheme is used by software named GAMESS (General Atomic and Molecular Electronic Structure System) to predict an unknown protein. It requires a large computer memory to perform mathematical calculations and it runs on Linux operating system.
- Software named as SWISS-Model is available for automated building of the theoretical structural models of a given protein (amino-acid sequence) based on the known proteins' structures.
- Classifiers, for example, neural networks, decision trees etc. learn classification rules from the given training data which are used to predict functions of unknown proteins. [3-4]

1.2 Cluster Analysis
Cluster analyzes the data objects without consulting a class label. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to in other clusters. Each cluster that is formed can be viewed as a class of objects from which rules can be derived. [5-6]

2. DATA SOURCE
The data related to human protein is accessed from Human Protein Reference Database (HPRD) [7]. The HPRD represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data. HPRD has been created using an object oriented database in Zope, an open source web application server that provides versatility in query functions and allows data to be displayed dynamically. It includes approximately 163 classes of protein functions. The database provides information about protein function under the heading 'molecular class' covering all the major protein function categories [3].

From HPRD, the sequences related to ten molecular classes are obtained. These classes are: Defensin (Def), Heat Shock Protein (HSP), Voltage Gated Channel (VGC), Cell Surface Receptor (CSR), DNA Repair Protein (DRP), Aminopeptidase (Ami), Decarboxylase (Dec), G-Protein (GP), RNA Binding Protein (RBP) and Transport/Cargo Protein (T/CP). For each of the molecular class five amino acid sequences are obtained.

3. SEQUENCE DERIVED FEATURES (SDFs)

Sequence derived features are the various features of protein which are used to predict human protein function. Sequence derived features are very important in protein prediction as these are the input to the HPF predictor as labeled vector. Some of the sequence derived features are:

a) Number of negatively charged residues.

b) Number of positively charged residues.

c) Extinction coefficients.

d) Instability index.

e) Aliphatic index.

f) Grand average of hydropathicity (GRAVY).

g) Tyrosine.

h) Serine.

i) Threonine.

j) Max cleavage site probability.

k) Predicted Transmembrane Helices (PredHel).

SDF's can be derived from a given set of amino-acid (protein) sequences. Following are the various bioinformatics Tools for obtaining sequence derived features (SDFs):

- NetNGlyc server predicts N-Glycosylation sites in human proteins using artificial neural networks.
- PSORT server is a computer program for the prediction of protein localization sites in cells. It analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possiblity for the input protein to be localized at each candidate site with additional information.
- TMHMM server is a program for predicting transmembrane helices based on a hidden Markov model.
- NetOGlyc server predicts the O-GalNAc (mucin type) glycosylation sites in mammalian proteins.
- Signal-P server 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.
- ExPASy ProtParam server computes various physico-chemical properties of protein like isoelectric point, extinction coefficient. [3] [8-10]

Values of the SDFs for each of the ten protein classes for five sequences are stored. The data for defensin class is shown in table 1.

Table 1: Values of SDFs for Defensin

SDFs	Amino Acid Sequences				
	94AA	100AA	97AA	326AA	68AA
Nneg	9	12	9	37	2
Npos	9	9	11	35	7
Exc1	15845	6335	7365	56880	6335
Exc2	15470	5960	6990	56380	5960
I.Index	49.71	55.34	46.1	22.99	32.91
A.Index	102.02	73.4	97.53	66.78	77.5
GRAVY	0.285	-0.169	0.178	-0.425	0.157
S	3	0	0	0	0
T	0	0	0	0	0
Ser	0	4	1	7	0
Thr	7	0	0	2	0
Tyr	0	1	0	3	1
MeanS	0.579	0.889	0.963	0.886	0.943
D	0.353	0.881	0.928	0.813	0.852
Prob.	0.181	0.870	0.962	0.836	0.818
ExPAA	0.64	0.02	1.64	0.30	0.00
PredHel	0	0	0	0	0

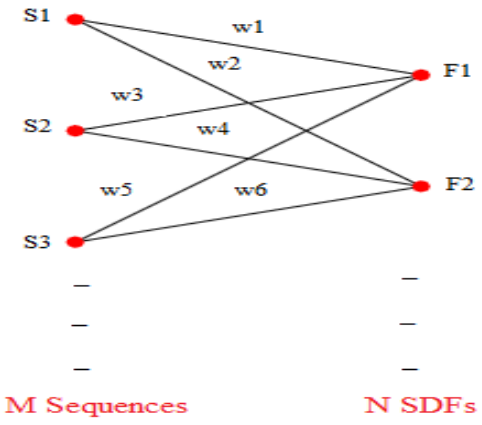

Fig. 1: Protein Sequences - SDFs Graph

3.1 Bipartite Graph Data

In the obtained data, each sequence has some value for each of the sequence-derived feature. So if amino acid sequences and SDFs are considered as two independent set of nodes, then each node from the first set is connected to each of the node in the second set. So obtained data is a complete weighted bipartite graph G = (V, E, W) consisting of V vertices (V = M U N i.e. M = no. of protein sequences and N = no. of sequence derived features), E edges (where each edge is connecting one node from each set M and N) and W weights (each edge is assigned

weight w equal to the value of the feature for the respective sequence) as shown in Fig. 1. Total no of edges are equal to total no of weights i.e. M * N. [11]

3.2 Adjacency Matrix Representation

As complete bipartite graph can be represented into memory with help of d adjacency matrix, so Protein Sequence – SDFs graph G = (V, E, W) where V = M U N, is stored into the memory with weighted adjacency matrix A (M * N matrix):

Where Aij = Wij for Eij; i: 1 to M, j: 1 to N

A consists of M rows (equal to the number of protein sequences) and N columns (equal to the number of SDFs).

3.3 Priority and Score of SDFs

For allocating score to each of the SDF, the variation in the values of each of the SDFs is studied for each functional class. If a particular SDF varies very highly from class to class, it is considered as high priority feature, because it can distinguish one class from another class easily. And the features with less variation in their values are considered as low priority features.

On the basis of priority score is allocated to each SDF. Priority and score allocated to each of the SDF is shown in Table 2. This score is utilized to find the highest score sequence to predict the class of the entered sequence. This concept of priority is not considered in any of the available prediction technique.

3.4 Range of SDFs

Range of the SDFs is used to find similarity between entered values of the SDFs and the sequences stored into the database. All the sequences which are in the specified range for the considered feature are considered similar to the entered sequence for that particular feature. For allocating range to each of the SDF:

- The difference in the values of each of the SDFs is studied for each sequence and functional class.
- Values of four lowest priority features, PredHel, S, T and Probability are repeated heavily from sequence to sequence and from class to class, so these features are supposed to match exactly to find the similar sequences so no range is allocated to them.

For the remaining features, differences in the values are studied thoroughly for each and range is allocated to each feature so that it should give rise to least overlap in the values from sequence to sequence or from class to class. Range allocated to each SDF is shown in Table 3.

Table 2. Priority and Score of SDFs

SDF	Priority	Score
PredHel	17	1
S	16	2
Probability	15	3
T	14	4
ExpAA	13	5
GRAVY	12	6
Inst. Index	11	7
A.Index	10	8
D	9	9
Mean S	8	10
Tyr	7	11
Thr	6	12
Ser	5	13
Exc2	4	14
Exc1	3	15
Npos	2	16
Nneg	1	17

Table 3. Range of SDFs

SDF	Range
PredHel	0
S	0
Probability	0
T	0
ExpAA	0.04
GRAVY	0.004
Inst. Index	0.04
A.Index	0.04
D	0.004
Mean S	0.004
Tyr	2
Thr	2
Ser	2
Exc2	10
Exc1	10
Npos	2
Nneg	2

4. PROTEIN CLASS PREDICTOR

The adjacency matrix A and value of each SDFs of unknown sequence will be given as input to the protein class predictor. The protein class predictor will first generate the clusters and then by applying the backtracking i.e. starting with the highest priority cluster and traveling backward it will find the sequence with the highest score. The class of the sequence with the highest score will be the class of the unknown sequence. How clusters are generated and backtracking proceeds is explained here.

4.1 Cluster Generation

Matrix A originally consists of un-clustered sequences, so agglomerative clustering (i.e. Hierarchical Clustering) is applied

to generate clusters. For cluster generation initially credit of each sequence in the matrix is set to zero. Then first SDF of unknown sequence is compared with the respective feature of each sequence in matrix A. If corresponding value of sequence is in the specified range of that particular SDF then the credit of the sequence is incremented by the score of that SDF and the sequence is included in the cluster of the SDF being compared. The same process is repeated for each of the SDF of unknown sequence. If respective feature value of a particular sequence is in the specified range of the current SDF being compared but it is already included in a cluster then priorities of the features are compared. If the priority of the current feature is greater than previous one then the sequence is shifted from the previous cluster to current cluster and credit of the sequence is again incremented by the score of the current SDF. Because each SDF's value is checked individually against the corresponding entry in the matrix A so number of generated clusters is always equal to N (i.e. no of SDFs) number of sequences in each of the cluster vary as it depend upon the number of entries in A which are in the specified range of the entered SDF of unknown sequence. As said, sequence can travel from one cluster to next high priority cluster and every time sequence enters a cluster the credit of the sequence is incremented by the score of that feature so if a cluster has travelled through all the low priority clusters its credit will have maximum possible value in that cluster.

4.2 Backtracking

In back tracking first highest priority feature's cluster is searched to find the sequence with the highest credit. If the credit of the obtained sequence is equal to the maximum possible credit in that cluster then this is the sequence with the maximum credit in all the cluster so need no to search further. Otherwise, next highest priority cluster is searched. Credit value of the sequence obtained in this cluster is compared with the previous cluster, if the current value is greater than the maximum credit sequence of the previous cluster then now this is the maximum credit sequence. Again its credit is checked to see whether it is equal to the maximum possible credit in this cluster, if yes then search is stopped here because maximum credit sequence is obtained, otherwise in the same way search is performed on the remaining clusters. But if the credit of the previous cluster is greater then that value is considered as highest and backtracking proceeds in the same manner.

Flow – Chart of protein class predictor is shown in Fig 2 and symbols used are shown in Table 4.

Table 4. Symbols used in Flow Chart of Fig. 2

Symbol	Meaning
M	Number of sequences
N	Number of SDFs
C	Credit of Sequence
S	Score of SDF
P	Priority of SDF
Pprev	Priority of previous cluster
Maxc	Current maximum credit
Maxp	Previous maximum credit
Mp	Maximum possible credit
R	Range of SDF
I, j and K	Counters

5. RESULTS AND DISCUSSION

To determine protein class of the following amino acid sequence, various SDFs are extracted with the help of the web based tools.

MIGQKTLYSF FSPSPARKRH APSPEPAVQG
TGVAGVPEES GDAAAIPAKK APAGQEEPGT
PPSSPLSAEQ LDRIQRNKAA ALLRLAARNV
PVGFGESWKK HLSGEFGKPY FIKLMGFVAE
ERKHYTVYPP PHQVFTWTQM CDIKDVKVVI
LGQDPYHGPN QAHGLCFSVQ RPVPPPPSLE
NIYKELSTDI EDFVHPGHGD LSGWAKQGVL
LLNAVLTVRA HQANSHKERG WEQFTDAVVS
WLNQNSNGLV FLLWGSYAQK KGSAIDRKRH
HVLQTAHPSP LSVYRGFFGC RHFSKTNELL
QKSGKKPIDW KEL

These sequence derived features will be given as input to the protein class predictor. For the given values of all features, this technique will match the value each of the entered feature with the respective value of all the sequences in the database and if the value of a sequence will be in the specified range of the entered value then that sequence will be included in the cluster of that feature. And every time a sequence will enter a cluster its credit will be incremented by the score of that feature. So in this manner clusters for all the SDFs will be generated. After all the clusters will be generated they will be backtracked to find the highest credit sequence. Starting with the highest priority cluster, sequence with the maximum credit will be determined from the current cluster. If the maximum credit obtained from the current cluster will be greater than the previous cluster then new sequence will become the maximum credit sequence otherwise previous sequence will remain the highest credit sequence.

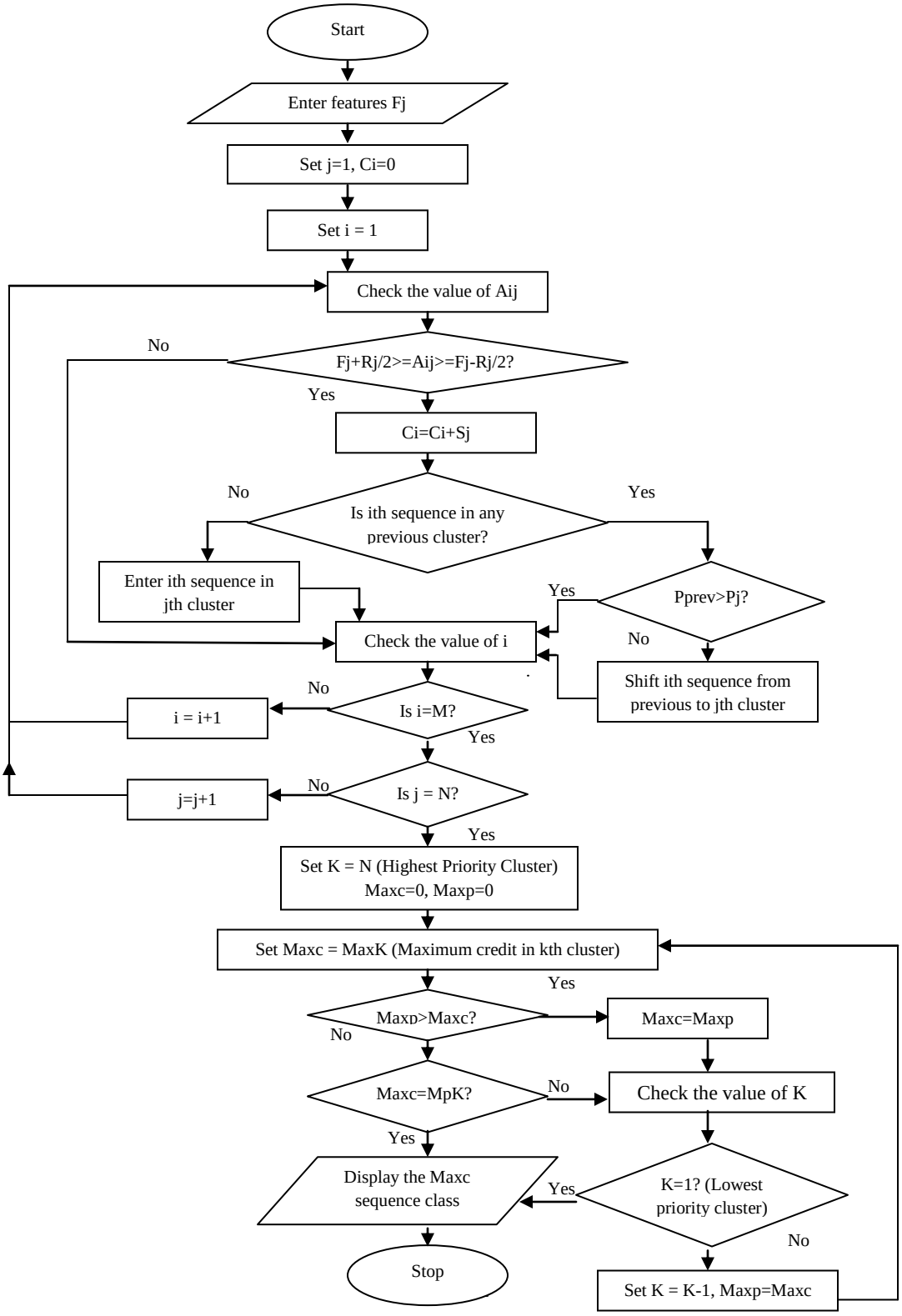

Fig 2: Flow-Chart of Protein Class Predictor

While traveling from higher priority clusters to lower priority, in each of the cluster if the maximum score of the current credit will be greater than all the previous clusters and also equal to the maximum possible credit of that cluster then the sequence with this credit will be the maximum credit sequence in the whole database and need not to travel further. Then the class of the sequence with maximum credit will be included in the prediction result. The SDF values obtained are shown in table 5. For the data considered here, the predicted class is Heat Shock Protein.

Table 5. Input SDF values to class predictor

Serial No.	Feature Name	Value
1	Nneg	28
2	Npos	37
3	Exc1	50545
4	Exc2	50420
5	Instability Index	47.19
6	Aliphatic Index	73.55
7	GRAVY	-0.488
8	S	1
9	T	2
10	Ser	15
11	Thr	4
12	Tyr	2
13	Mean S	0.078
14	D	0.67
15	Probability	0.000
16	ExpAA	0.03
17	PredHel	0

6. CONCLUSION

The model is very simple to use. The drug discoverer has to simply give the values of sequence derived features of unknown sequence and will get the resulted protein class. The classification process is using five protein classes, ten sequences for each protein classes i.e. 50 sequences and 17 sequence derived features. In the present technique, priorities are allocated to the various features. High priority features contribute more than the low priority features in the final classification. This concept of priority is not considered in any of the available prediction technique. In the present technique, backtracking was used which insures which insures accurate results. In future, the system can be made more user friendly by computing the results directly from the input amino acid sequence. This can be achieved by making the SDF server for computing all the required SDFs from amino acids.

7. REFERENCES

[1] Friedberg I. 2006. “Automated Protein Function Prediction-the genomic challenge”, Briefings in Bioinformatics, Vol. 7, No. 3.January 2006, pp. 225-242.

[2] Krane, D. and Raymer, M. 2006. Fundamental Concepts of Bioinformatics, Pearson Education, New Delhi.

[3] Singh Manpreet, Singh Parvinder and Wadhwa Parminder Kaur, 2007. “Human Protein Function Prediction using Decision Tree Induction”, International Journal of Computer Science and Network Security, Vol. 7, No. 4, pp. 92-98.

[4] Singh Manpreet, Wadhwa P.K., Kaur Surinder, 2008. “Predicting Protein Function using Decision Tree” World Academy of Science, Engineering and Technology, issue 39, pp. 350-353.

[5] Kaur Reet Kamal, Kaur Manjot, Kaur Amanjot. 2010. "Using Cluster Analysis for Protein Secondary Structure Prediction" International Journal of Computer Applications, Vol. 4, No. 12, August 2010, pp. 20-22.

[6] Singh Manpreet, Singh Gurvinder and Kahlon Karanjeet Singh, 2009. “Analyzing the Cluster for Protein Sequence Alignment”, PCTE Journal of Computer Sciences, Vol. 6, issue 1, 2009, pp. 74-83.

[7] Human Protein Reference Database (HPRD) http://www.hprd.org/moleculeClass

[8] Jensen L. 2002. Prediction of Protein Function from Sequence Derived Protein Features, Ph.D. thesis, Technical University of Denmark.

[9] Jensen L., Skovgaard M. and Brunak S. 2002. “Prediction of Novel Archaeal Enzymes from Sequence Derived Features”, Protein Science, Vol. 11, pp. 2894-2898.

[10] Jensen L.J., Gupta R., Blom N., Devos D., Tamames J., Kesmir C., Nielsen H., Stærfeldt H.H., Rapacki K., Workman C., Andersen C.A.F., Knudsen S., Krogh A., Valencia A. and Brunak S. 2002. “Prediction of Human Protein Function from Post-Translational Modifications and Localization Features”, Journal of Molecular Biology, Vol. 319(5), pp. 1257-1265.

[11] Charu C. A., Haixun W. 2010. Managing and Mining Graph Data, Springer.