Bioinformation
www.bioinformation.net

open access
Hypothesis

Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties

Amit Kumar Banerjee, Sunita M., Naveen Mungara, Upadhyayula Suryanarayana Murty*

Bioinformatics Group, Biology Division, Indian Institute of Chemical Technology, Hyderabad-500607, A.P, India. U.S.N Murty-Email: murty_usn@yahoo.com, *corresponding author

Received January 12, 2010; revised March 02, 2010; accepted April 09, 2010; published April 30, 2010

Abstract:
Biological systems are highly organized and enormously coordinated maintaining greater complexity. The increment of secondary data generation and progress of modern mining techniques provided us an opportunity to discover hidden intra and inter relations among these non linear dataset. This will help in understanding the complex biological phenomenon with greater efficiency. In this paper we report comparative classification of Pyruvate Dehydrogenase protein sequences from bacterial sources based on 28 different physicochemical parameters (such as bulkiness, hydrophobicity, total positively and negatively charged residues, α helices, β strand etc.) and 20 type amino acid compositions. Logistic, MLP (Multi Layer Perceptron), SMO (Sequential Minimal Optimization), RBFN (Radial Basis Function Network) and SL (simple logistic) methods were compared in this study. MLP was found to be the best method with maximum average accuracy of 88.20%. Same dataset was subjected for clustering using 2*2 grid of a two dimensional SOM (Self Organizing Maps). Clustering analysis revealed the proximity of the unannotated sequences with the *Mycobacterium* and *Synechococcus* genus.

Keywords: Pyruvate Dehydrogenase, Data Mining, Clustering, KNIME, Self Organizing Maps (SOM).

Background:
Classification and clustering analyses are having their own significance in biological data mining. These mathematical techniques impart a long term relation with biological data analysis. With time, sophistication has increased in data generation and analytical methodologies. Basic biological taxonomical methods initiated with morphological identification and classification, now extended to the molecular level in the recent past **[01]**. Methodologically, the available techniques improved a lot with the help of modern statistical pure and hybrid strategies. Data mining techniques extracts complex pattern [**02, 03]** and relationship from a given set of data. As stated by Leslie Cauley, "This is referred to as data mining. They slice and dice these numbers a thousand different ways. They analyze patterns." Application of this classification and clustering techniques are well documented in the available literature **[04, 05, 06, and 07]**, where it has been extensively used with higher level of accuracy and efficiency for different kind of complex data sets. In the preceding years, this kind of methodology gained momentum due to the emergence of more interdisciplinary research. Artificial Neural Network (ANN) **[08, 09]** based methodologies and higher level multi variate regression techniques **[10, 11]** got special appreciation for handling more complex and non linear dataset. Biological datasets are complex due to its heterogeneous nature and are hard to classify. Recent applications deal with this complexity easily and do efficient classification for any kind of data **[12, 13]**. The advancement of the computational biology research have generated enormous amount of data in the recent decades. Mining of this chunk of data may unveil some unique evidence or clues on the pattern or the interrelationship. In the present study we have adopted an intelligent method dependent approach for classification and clustering the bacterial Pyruvate Dehydrogenase protein sequences according to their genera based on their physicochemical properties **[14, 15]**. Similar kind of studies we have reported in the recent past **[16, 17, and 18]**.

Methodology:
The aim of this study is to classify and cluster the protein sequences based on their 48 (28 different parameters and 20 amino acid compositions) physicochemical properties of Pyruvate dehydrogenase, extracted from National Centre for Biotechnological Information (NCBI) **[19]** public domain protein database, employing modern data mining approach and compare the efficiency of different methods while classifying the complex dataset. Pyruvate Dehydrogenase (PDH) superfamily **[20, 21, 22]** was selected out of several super families due to its enormous importance in the regular metabolism. A large number of sequences were collected initially and after complete preprocessing a final dataset of 95 sequences was prepared. The complete workflow is represented below in the **Figure 1.**

Data preprocessing:
All the collected sequences were verified manually and redundant sequences were removed from the collected dataset. Ample amount of sequences in the NCBI public domain database are present repetitively, all those repetitive sequences were sorted out and only the longest complete sequence was kept as a representative. Redundant sequences with unspecific amino acid one letter codons were not considered for the study.

Extraction of features:
Features were calculated by employing Protparam (http://www.expasy.ch/tools/protparam.html) and Protscale (http://www.expasy.ch/ tools/ protscale.html) servers available from EXPASY. A total of 48 parameters were computed by the above mentioned servers, they are as follows: bulkiness, recognition factors, hydrophobicity-Kyte & Dolittile, percent buried residues, ratio hetero, average flexibility, beta sheet-Chou Fasman, alpha helix-Chou Fasman, anti parallel beta strand, relative mutability, number of codons, polarity–Zimmerman, refractivity, transmembrane tendency, percent accessible residues, average area buried, beta turn-Chou Fasman, coil-Deleage & Roux , parallel beta strand, molecular weight, theoretical pI, different amino acid composition, extinction coefficient(All cys), extinction coefficient(No cys), total number of positive charges (Arg+Lys), total number of negative charges (Asp+Glu), aliphatic index, instability index and GRAVY. The extracted features obtained according to the used servers are listed in **Table 1 (see supplementary material).**

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

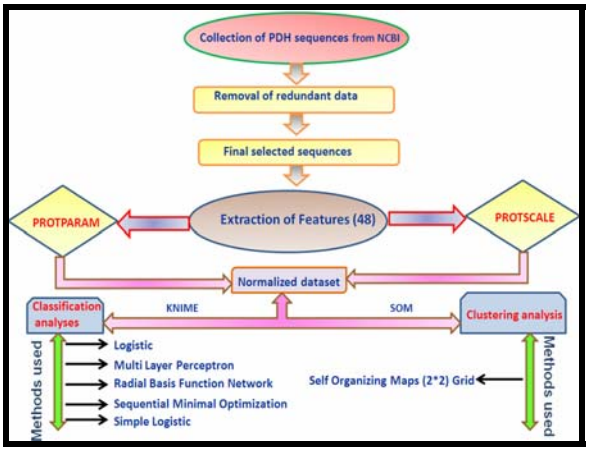

Figure 1: Workflow representation of the whole study.

Input data set preparation:
Preprocessed dataset prepared for this study contained 256 sequences of PDH family belonging to different prokaryotic and eukaryotic species. As this study was solely confined to the bacterial genera, sequences belonging to bacteria only were isolated. Moreover, only those genera were considered who have at least 5 sequences. The final input dataset contained a total of 95 sequences belonging to 5 different bacterial genera and an unannotated group. A total of 5 sub-datasets were prepared from the main dataset to remove any kind of bias from the analysis and each set was further divided into 2 subsets for training and test set respectively, maintaining a ratio of 70:30.The same dataset was used for clustering the sequences using Self Organizing Maps **[23, 24]** without any training and test set division as the algorithm itself takes care of data division during the calculation.

Classification with KNIME:
The input dataset was used for the classification using KNIME **[25]** software. It is a well known effective data mining tool which enables us to construct a proper mathematical model by using different classification and regression methods available in the tool. It is a modular data exploration platform that allows the user to visually create data flows. The functions, such as logistic method **[10, 26]**, Multi Layer Perceptron (MLP) **[27, 28]**, Radial Basis Functional Network (RBFN) **[29]**, Sequential Minimal Optimization (SMO) **[30]** and simple logistic (SL) were applied to classify the prepared data.

Clustering with Self-Organizing Maps (SOM):
A self-organizing map (SOM) is a modern clustering method which mimics the human brain in architecture. This algorithm possesses several input neurons containing the input data space. Topological properties of all the input neurons are preserved in the input space by using neighborhood function. The final winning neuron represents the low dimensional visualization of high dimensional data points. The output representation is done with different color codes which lie in various ranges of the data points. A two dimensional SOM was adopted to cluster the generated input data space of pyruvate dehydrogenase protein sequences. The following algorithm was used in this SOM:

Steps involved in the algorithm:
(1) Initialization: Randomly initialize a weight vector (Wi) for each neuron i *Wi = [wi; wi2; ... ; win]*; n denotes the dimension of input data. **(2)Sampling:** Select an input vector X=[*x1, x2, ... , xn*] **(3)Similarity matching:** Find the winning neuron whose weight vector best matches with the input vector j(t)= arg min {||X-Wi||} **(4)Updating:** Update weight vector of winning neuron, such that it becomes still closer to the input vector. Also, update weight vectors of neighbouring neurons-the further the neighbour, the lesser the degree of change. Wi(t+1)=Wi(t) +α(t) X hij(t)X [X(t)-Wi(t)] α (t): learning rate that decreases with time t, 0<α(t) ≤ 1 hij(t)= exp(-|| rj- ri || 2/2 X σ(t)2) ||rj-r i||2=distance between winning neuron and other neurons σ (t)=neighbourhood radius that decreases with time t. **(5)Continuation:** Repeat steps 2–4 until there is no change in weight vectors or up to certain number of iterations. For each input vector, find the best matching weight vector and allot the input vector to the corresponding neuron/cluster.

Data Normalization: To obtain unbiased results while ensuring equal importance to all parameters while clustering, data was normalized linearly such that value in each category ranged between 0 and 1. Normal Formula = (Original data value-Minimum data value)/ (Maximum data value-Minimum data value)

Results and discussion:
Proper classification has been the basic criteria in taxonomy from the dawn of scientific studies in biological sciences. It started with morphological and phenotypic differentiation and slowly with the advancement of molecular techniques inclined towards the specific molecular markers. Novel *in silico* and statistical methods aided accuracy and efficiency to these existing techniques and added a new dimension in biological data analysis **[09, 31]**. In this study we classified five genera and an unknown group of data with greater amount of accuracy level and later on clustered those considered sequences for better magnification and understanding of their exact comparative position.

The over all final curated data contains 95 sequences belonging to five different bacterial genus, they are, *Bacillus, Burkholderia, Geobacillus, Mycobacterium* and *Synechococcus* along with an unannotated sequence group.

Classification analysis with KNIME:
In KNIME, several methods with different functions like logistic, Multi Layer Perceptron (MLP), Radial Basis Function Network (RBFN), Sequential Minimal Optimization (SMO) and simple logistic are available. The program requires an input of training and test dataset. Rigorous training and testing exercise was performed with all the 5 data sets and classification efficiency was calculated for them applying the above mentioned methods.

The obtained results indicated that Multiple Layer Perceptron (MLP) method had yielded maximum average accuracy level of 88.2024 **(Figure 2)**. Though the other methods were also good and the optimum result remained over 80% of accuracy. Apart from the MLP method rest of the methods have shown more or less similar results which is lying within the range of 83.0196% and 83.7094% accuracy in average **(Figure 2).**

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

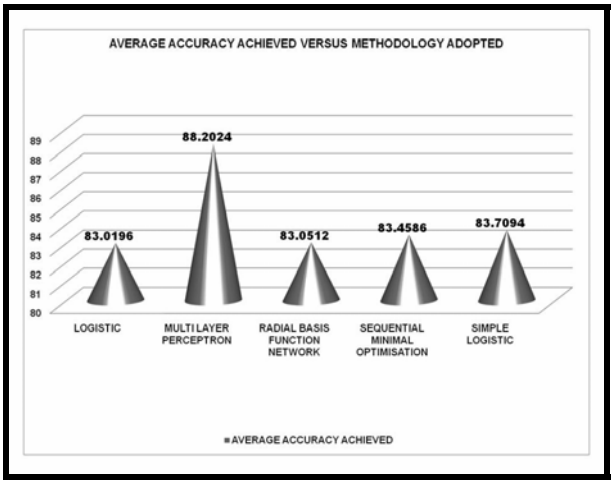

Figure 2: Average accuracy achieved versus methodology adopted.

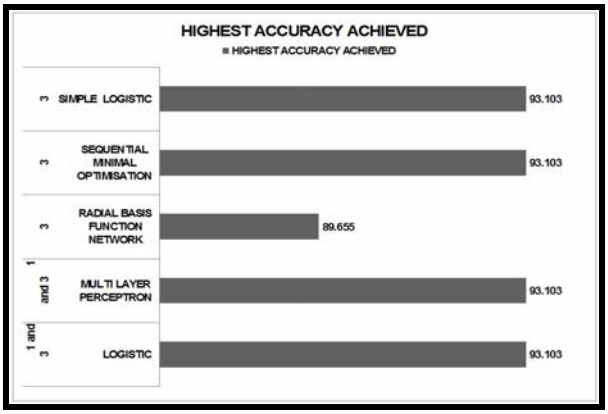

Figure 3: Representation of highest accuracy obtained with respective methodology adopted and dataset used.

Obtained highest accuracy for individual dataset achieved 93.103% of accuracy **(Figure.3)**. Except Radial Basis Function Network all method showed the exact same level of accuracy with different dataset. But in all the cases dataset 3 remain common while providing the maximum level of accuracy. Overall performance analysis exhibited consistency.

The detail of obtained proper classification, misclassification and error percentage are depicted in Figure 4. The classification analysis efficiently categorized the sequence data based on the parameter considered for this study. The methodology adopted was able to group the data successfully. To obtain further insight and recognize the orphan sequences (not annotated), help of clustering technique was taken. Initially, statistically mean distribution was calculated for each genus along with the unannotated group. Later on sophisticated Kohonen Map was used to get more insight.

The mean value for all the parameters were calculated for each genus considered for this study along with the unknown group. Higher level similarities were found in the calculated mean values **(data available from the authors upon request)** for the parameters of *Mycobacterium* and *Synechococcus* with the unknown sequences. *Mycobacterium* data and unknown data group showed similarity for the following parameters; Molecular weight, Hydrophobicity, % buried residues, Beta sheet, Polarity, Transmembrane tendency, Coil, Theoretical pI, Glutamine, Glycine, Leucine, Phenylalanine, Proline, Valine ,Total positive charge (Arg + Lys), Instability index and GRAVY.

Similar trend was observed in between *Synechococcus* and unknown data group. The following parameters showed similarity in the context of mean value; Bulkiness, Relative mutability, Number of codons , Transmembrane tendency, Average area buried, Glutamine, Glutamate, Leucine, Methionine, Serine, Threonine, Total negative charge (Asp+Glu). No resemblance was observed with the unknown data group and *Burkholderia* or *Geobacillus* genus. This obtained calculated statistical result suggests that some candidate sequences of the unknown group were in vicinity either with *Mycobacterium* or *Synechococcus* while maintaining a major distance with the other two genuses. To confirm this basic statistical insight, support of latest clustering approach was taken.

Cluster Analysis with SOM:

A two dimensional SOM has been employed to cluster the sequence data based on their respective genus. A 2*2 grid SOM was employed to cluster all the sequences based on their calculated physico-chemical parameters. The range of learning rate was tuned between .01 and .10. Iterative convergence of the output was restricted to 1, 00,000 iterations during the calculations. Four clusters were formed successfully. The distribution of all the 95 sequences are shown in the below pie chart (Figure 5) where the cluster (1, 1) is possessing 30 sequences, cluster (1, 2) 24, cluster (2, 1) 9 and cluster (2, 2) 32 sequences respectively.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

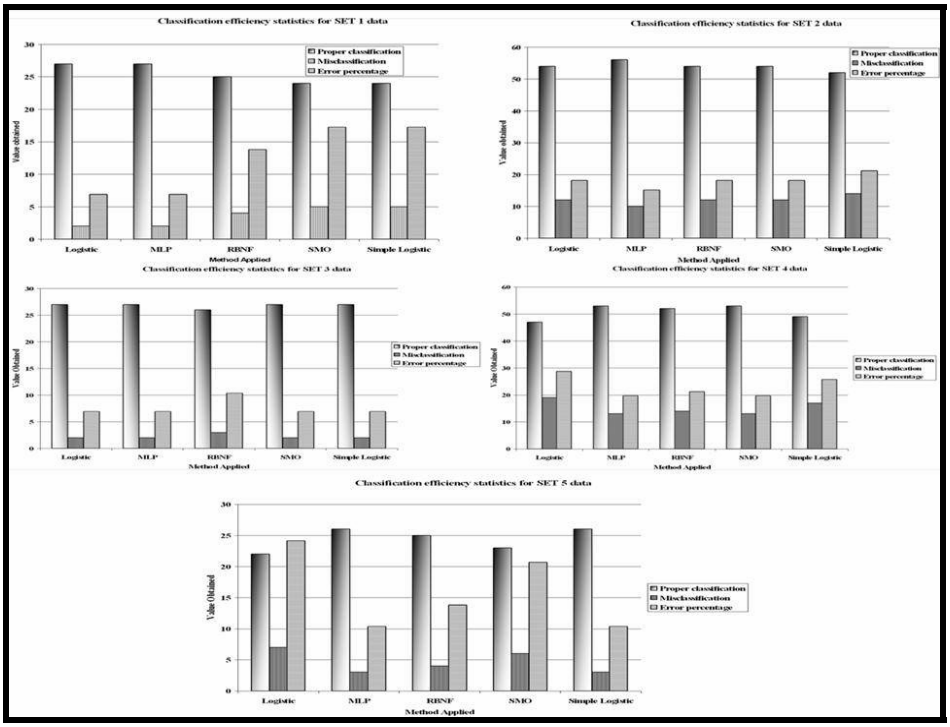

Figure 4: Classification efficiency of different methods with respect to dataset.

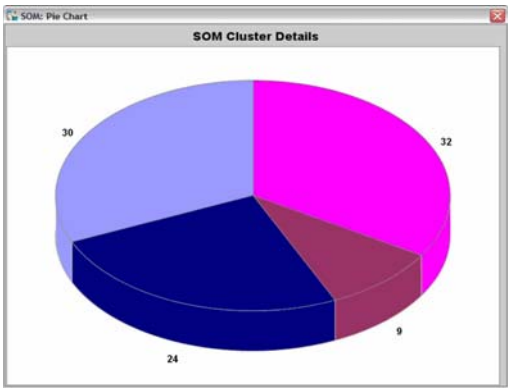

Figure 5: Distribution of sequences considered in this study according to their cluster formation.

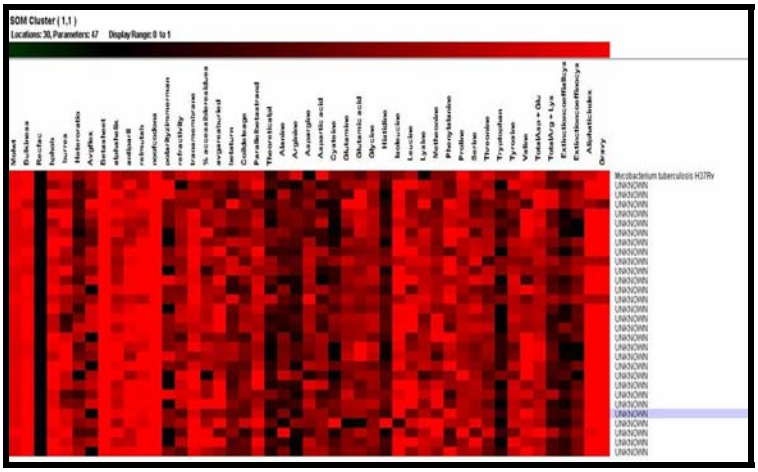

Figure 6: Visual representation of cluster 1, 1

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

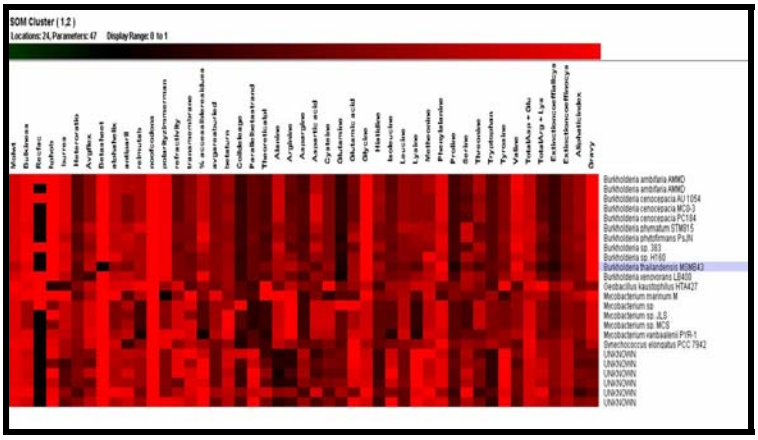

Figure 7: Visual representation of cluster 1, 2

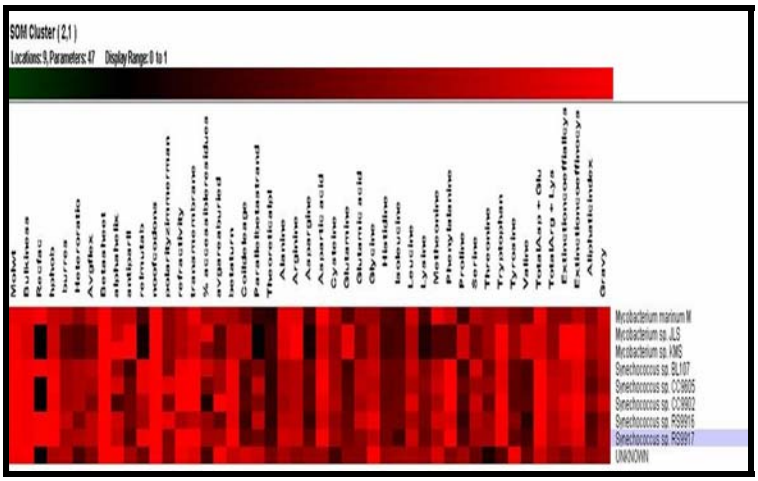

Figure 8: Visual representation of cluster 2, 1

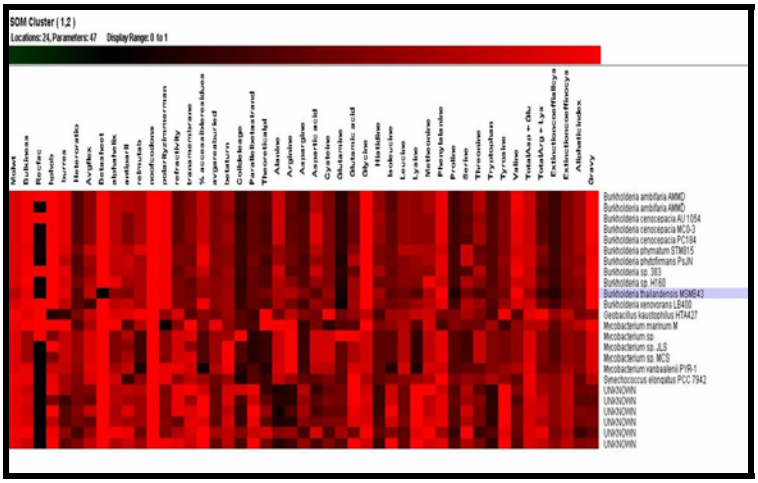

Figure 9: Visual representation of cluster 2, 2

Cluster (2, 1): The third cluster is the smallest one containing only 9 sequences **(Figure 8)**. The distribution of the sequences observed is sharing 3, 5 and 1 sequences from *Mycobacterium*, *Synechococcus* and unknown respectively. The uniqueness in this cluster is its maximum parametric similarity with each other which is easily visible from the above cluster compare to others. Parameters like molecular weight, bulkiness, beta sheet, number of codons are with complete agreement to each other. Higher level of similarity has been observed in hydrophobicity, polarity, total positive charges (Asp + Glu) and extinction coefficient.

Cluster (1, 1): The first cluster, i.e., cluster **(1, 1)** represented in Figure 6, was found to be unique according to the parameters it has clubbed together. This cluster was having only one *Mycobacterium*

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

sequence, rest of all is from the unannotated or unknown group. Few parameters in these sequences are having exact similarity, such as beta sheets and number of codons. All the values for these two parameters are near about 1. Some other parameters have also known high degree of equality like GRAVY, aliphatic index, molecular weight and bulkiness. Values of these attributes are again very similar and ranged near 1 with few minor deviations. On the contrary, parametric value of refractivity is quite low for all the sequences and it's in the vicinity to 0.

Cluster (1, 2): The second cluster contains 24 sequences overall. Among those 11 are from *Burkholderia,* 5 *Mycobacterium,* 6 unknown types and each *Geobacillus* and *Synechococcus* having a single representative sequence in this very cluster **(Figure 7).** All the unknown sequences in this cluster have been placed near *Synechococcus* sequence, suggesting their similarity with this very group. In this case also maximum similarity has been observed in case of number of codons followed by beta sheet with exception for one sequence, molecular weight, glycine content and bulkiness.

Cluster (2, 2): The final cluster is the largest one with 32 sequences **(Figure 9).** Though this is the largest cluster formed still *Synechococcus* and *Burkholderia* sequences are absent in this cluster. A total of 7 unknown sequences are clustered in vicinity to *Mycobacterium*, suggesting their similarity with *Mycobacterium.* In a similar fashion like the prior clusters molecular weight, beta sheet and number of codons have shown the identity in majority. Higher degree of similarity is observed in total (Asp + Glu) content.

In this study we were able to classify and cluster successfully a complex dataset containing 95 records and 29 attributes belonging to 5 different genuses and a complete set of orphan sequences. The overall accuracy achieved was more than 93% employing different classification methods and we were able to predict the position of the unknown sequences based on their distribution in different clusters. Most of them were in vicinity with *Mycobacterium* and *Synechococcus*.

Conclusion:

There are several bacterial species present in this universe but so many are to be discovered. Classification of those genus or species is of immense importance from taxonomical and molecular biology point of view. We have several molecular markers now to detect a particular species still it is cost effective and there are some deficiencies in our usual methodology which restricts us to confidently confirm in so many cases. Intelligent techniques may prove as an effective tool in this regard and considering of all statistical facts and complexity it can help us to reach a meaningful conclusion. In this study we have classified and clustered different microbial groups based upon their highly complex physicochemical properties with more than 80% accuracy successfully. Unannotated sequences were also classified and clustered which will help to determine their taxonomic position properly. As the heap of generated dataset is increasing day by day, in future this kind of unannotated data will increase. The approach adopted here to classify and cluster may become an effective tool in future for proper classification and clustering and determining the exact taxonomic position for such orphan data.

Acknowledgement: The authors are grateful to Dr. J.S.Yadav, Director, Indian Institute of Chemical Technology for his constant support and encouragement during the study. AKB thanks Council of Scientific and Industrial Research (CSIR), Govt. of India, for the Senior Research Fellowship (SRF).

Reference:

[1] H. Kito *et al. Biosci Biotechnol Biochem*., [Epub ahead of print] (2009) [PMID: 19352027]
[2] B.L. Adam *et al. Cancer Research*, 62: 3609 (2002)
[3] E.A. Madigan & O.L. Curet, *BMC Health Serv Res.,* 24:6:18 (2006) PMID: [16504115]
[4] G. Reibnegger *et al., Proc Natl Acad Sci U S A.,* 88:11426-30 (1991) [PMID: 1763057]
[5] V.A. Valera *et al. Ann Surg Oncol,.* 14:34-40 (2007) [PMID: 17024555]
[6] Valkonen V.P. *et al. Int. J. Epidemiol.,* 31:864-71 (2002) [PMID: 12177035]
[7] K. Huang & R.F. Murphy, *BMC Bioinformatics*, 18:5:78 (2004) [PMID: 15207009]
[8] G. Ball *et al. Bioinformatics,* 18:395-404 (2002) [PMID: 11934738]
[9] A.K. Banerjee *et al. Computational Biology and Chemistry*, 32:442-447 (2008) [PMID: 18838305]
[10] P.C. Austin, *Stat Med.,* 10:26:2937-57 (2007) [PMID: 17186501]
[11] Y.A. Chen *et al. BMC Bioinformatics,* 1:7:101 (2006) [PMID: 16509965]
[12] K.L. Lin *et al. IEEE Trans Nanobioscience,* 6:186-96 (2007) [PMID: 17695755]
[13] S.L. Dollhopf *et al. Microb Ecol.*, 42:495-505 (2001) [PMID: 12024232]
[14] J. Kyte & R.F. Doolittle, *J Mol Biol.,* 157:105-132 (1982) [PMID: 7108955]
[15] E. Gasteiger *et al.* In: Walker JM (ed) The proteomics protocols handbook. Humana New York 571-607 (2005)
[16] U.S.N. Murty *et al. J. Proteomics Bioinform.,* 2: 097-107 (2009)
[17] A.K. Banerjee *et al. Electronic Journal of Biology*, 4:27-33 (2008)
[18] U.S.N. Murty *et al. Interdisciplinary Sciences: Computational Life Sciences*, 1(3):173-178 (2009)
[19] E.W. Sayers *et al. Nucleic Acids Res*., 37:D5-15. (2009) [PMID: 18940862]
[20] M.S. Patel & T.E. Roche, *FASEB J*., 4:3224-33, (1990) [PMID: 2227213]
[21] M. Bowker-Kinley & K.M.Popov, *Biochem J.,* 15:344:47-53 (1999) [PMID: 10548532]
[22] R.A. Harris *et al. Adv Enzyme Regul.,* 41:269-88 (2001) [PMID: 11384751]
[23] H.C. Wang *et al. Mol. Biol. Evol.,* 18: 792-800 (2001) [PMID: 11319263]
[24] S. Mahony *et al. BMC Bioinformatics*, 5;5:23. (2004) [PMID: 15070404]
[25] http://www.knime.org/
[26] F.C. Mello *et al. BMC Public Health,* 23: 6:43 (2006) PMID: [16504086]
[27] V. Sindhwani, 15:937-48 (2004) [PMID: 15461085]
[28] E. Romero & J.M. Sopena, *IEEE Trans Neural Netw.* 19:431-41 (2008) [PMID: 18334363]
[29] X.M. Zhao *et al. Protein Pept Lett.,* 12:383-6 (2005) [PMID: 15907186]
[30] S.K. Shevade *et al.* IEEE Trans Neural Netw. 11:1188-93 (2000) [PMID: 18249845]
[31] M Juhola *et al. Acta Otolaryngol Suppl.*, 545:50-2 (2001) [PMID: 11677741]

Edited by P.Kangueane
Citation: Banerjee *et al.,* Bioinformation 4(10): 456-462 (2010)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)

Supplementary material:

Table.1: Calculated parameters by Protparam and Protscale.

Sl. No.	Tool Applied	Calculated Property
1	**Protscale**	Bulkiness
2		Recognition factors
3		Hydrophobicity-Kyte & Dolittile
4		Percent buried residues
5		Ratio hetero
6		Average flexibility
7		Beta Sheet-Chou Fasman
8		Alpha helix-Chou Fasman
9		Anti parallel beta strand
10		Relative mutability
11		Number of codons
12		Polarity–Zimmerman
13		Refractivity
14		Transmembrane tendency
15		Percent accessible residues
16		Average area buried
17		Beta turn-Chou Fasman
18		Coil-Deleage & Roux
19		Parallel beta strand
20	**Protparam**	Molecular weight
21		Theoretical pI
22		Different Amino Acid composition
23		Extinction coefficient(All cys)
24		Extinction coefficient(No cys)
25		Total number of positive charges (Arg+Lys)
26		Total number of negative charges (Asp+Glu)
27		Aliphatic index
28		Instability index
29		Gravy

List of Abbreviations:

SOM: Self Organizing Maps; PDH: Pyruvate Dehydrogenase; KNIME: Konstanz Information Miner; GRAVY: Grand Average of Hydropathicity; MLP: Multi Layer Perceptron; PI: Isoelectric Point; TPP: Thiamine pyrophosphate; FAD: Flavin Adenine Dinucleotide; NAD: Nicotinamide Adenine Dinucleotide; CoA: Coenzyme A; ANN: Artificial Neural Network; SMO: Sequential Minimal Optimization; RBFN: Radial Basis Function Network; SL: Simple Logistics; PDC: Pyruvate dehydrogenase complex.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(10): 456-462 (2010)