Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 3 Number 2 (2010) pp. 175–184

http://www.ripublication.com/jcib.htm

Classification of Genomes based on Subclass Characterization

B.V. Dhandra[1] and S.S. Patil[2*]

[1]Department of Computer Science, Gulbarga University, Gulbarga, 585 106, Karnataka State, India, E-mail: dhandra_b_v@yahoo.co.in
[2]Department of Computer Science, University of Agricultural Sciences, Bangalore, 560 065, Karnataka State, India,
**Corresponding Author E-mail: spatilsuasb@gamail.com*

Abstract

Classification of large set of sequences of the grass genomes is a major challenging task in functional genomes. The presence of motifs in grass genome chains can make the possible prediction of the functional behavior. The correlation between grass genome properties and their motifs is not always obvious, since more than one motif may exist within a genome chain. Due to the complexity of this association most data mining algorithms are either non efficient or time consuming. Hence, we attempt to reduce the time complexity of classification of large dataset of grass genomes sequences, using a divide and conquer technique. First, data are split into 'p' equal multiple subsets by preserving the original data distribution in each set. Multiple models are created by using the data sets as independent training sets. Second, classification technique is applied to each module for constructing the classes. Finally, the outputs of these modules are combined to produce the final classification rule. The methodology is tested on three different datasets containing various grass genomes. Experimental results indicate that the proposed method reduces the time complexity keeping the classification accuracy level comparable to Nearest Neighbor Classifier (NNC) algorithm.

Keywords: Divide and Conquer (DAC), Classification Accuracy (CA), Grass genome.

Introduction

The development of advanced and specialized bioinformatics tools has led to

revolutionary changes in the analysis of biological sequences. Visualizing and predicting molecular structures and functions, separating DNA sequences according to grass genome coding regions, classifying grass genomes, detecting weak similarities have to rely on computational methods. The continuous increase in size of biological databases address the need of new, computation-sensitive data mining techniques, but also present unique opportunity for new fields of inquiry; among these one of the ambitious goal of bioinformatics is the prediction of the functional behavior of grass genomes. Genomes are large molecules composed by the base sequence of the nucleotides. Grass genome sequence classification is a major direction to prediction of functional behavior. Sequence classification enables to form genome families/groups possessing common behavioral characteristics and structural similarities. Classification of a grass genome into multiple families with different similarity levels - makes the procedure even harder, due to its increasing complexity. Grass genome functionality prediction could hardly be achieved were it not for *motifs*, short amino acid sequences of specific order, which appear in grass genome chains and play a decisive role in grass genome behavior. Although a straightforward mapping between motifs and grass genome properties is hard to achieve due to the presence of multiple motifs in each grass genome chain, they can facilitate prediction of grass genome functionality, if the latter is considered to be derived by the combining effect of many, either conflicting or consistent motifs. To overcome these problems, the occurrence of a particular motif in a grass genome chain is obtained. The overall procedure of motifs identification and detection in a genome sequence can be carried out using *unsupervised* learning technique. Then Divide-and-Conquer (DAC) technique employed to classify a given set of grass genome sequences into families and groups, which reflects common functional characteristics and evolutionary relatedness.

In this direction many researchers contributed to protein classification. A comprehensive and well-documented database created by experts in the field of bioinformatics, such as the PROSITE [11] the PFAM [6] or the PRINTS [4] databases can facilitate motif search in a way that satisfies these requirements. DNA and RNA molecules or detecting weak similarities have come to rely vitally on computational methods [5]. A generalization of the clustering problem, referred to as the projected clustering problem, in which the subsets of dimensions selected are specific to the clusters themselves[1]. A projected clustering concept of using extended cluster feature vectors in order to make the algorithm scalable for very large databases [2]. The masking of null values and discovered biclusters with random numbers may result in the phenomenon of *random interference* which in turn impacts the discovery of high quality biclusters. The divide and conquer approach is used to align of Sum-of-Pairs of multiple genome sequences and improve the alignment approach [21][22] [24]. have made an attempt to find the optimal alignments of multiple protein and DNA sequences[14]. have precisely sequenced the maize repeat annotation genome shotgun sequences using divide and conquer technique[20]. have considered the associated (fixed size) attribute vector for genomic string data, dividing and conquering the machine learning problem of ortholog detection[15]. This is seen as an analogy problem. The algorithm called "minimum conflict phylogeny estimation",

proposed by [13] estimates the conflict from the root to the leaves by heuristically searching for a minimum-conflict split and tackling the resulting two subsets in the same way. Related literature evidences many data mining algorithms that utilize the motifs present in genome sequences to perform genome classification originating from the field of pattern recognition and artificial intelligence, as seen in [9] and [10]. They include many different techniques such as decision trees, statistical models, neural networks Classification [19], [7], protein grid classification [17] and subclass unknown interactions of some gene pairs. The quality of the prototypes is evaluated using the Classification Accuracy obtained from the testing data set [3].

A novel technique has been proposed based on the concept of Grid and parallel computing. The combination of the two leading techniques may help to overcome the computational difficulties often encountered in genome classification problems. DAC divides fragments of grass genome sequences of large data sets into k/n subsets. The proposed "divide and conquer" approach comprising three steps: (1) Grass Genome data from FAST format based database are divided into multiple disjoint data sets, each one preserving the original data distribution. (2) Individual models have been generated for each subset, which includes motif generation based on leader algorithm and clustering using nearest neighborhood technique. (3)The classes of the modules obtained are merged appropriately for final classification. The proposed method is experimented with various data sets of grass genomes for their classification. Experimental results show that the proposed method outperforms than the NNC method and would be efficient technique for classification of grass genomes.

Methodology

The main goal of the DAC classification methodology is to utilize the existing data mining algorithms in a parallel-enabled environment, such as the grid, to create a grass genomes classification model. The proposed method encompasses three major steps. The architecture of the proposed method illustrated in Fig-1. The first step of the grid methodology is to divide the original single dataset into multiple datasets. The single input dataset is then split into multiple datasets of the same format. The process makes sure that each new dataset contains a unique subset of the original data. This allows for the independent use of them and therefore enables parallel processing to extract a useful knowledge module from each one of the datasets. The Round-Robin technique divides the grass genome sequence dataset into 'p' number of sub datasets which randomly distribute and allocates the data into p partitions as subsets Si, for i =1 to p of equal size. This is required for the algorithm as input. Experimentally, it is found that the method is robust in the class allocation, both for different number of splits and for varying number of classes involved. The M numbers of common motifs are generated from the set of sequences using dynamic algorithm with threshold technique and motif size, the threshold value is used either creating a new motif or assigning to the existing one. Based on generated motifs a frequency table of size NxM is formed, where N is number of sequences and M is number of motifs. Using the frequency table, clustering is done with hamming distance, Cosine and Jaccard similarity metric.

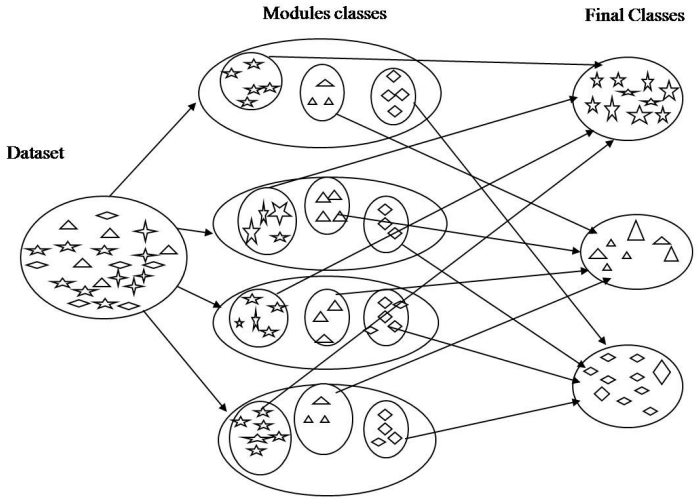

Figure 1: Over all architecture of divide and conquer method.

Next classification, consider a subset as a module. In each module, generating clusters Clj for j=1 to Ki, Ki is the size of the i^{th} class contains very close sequences using the leader algorithm with conservative threshold value. In this, leader algorithm has been used for classification, Divide and conquers with grid classification [15], [16]. After splitting the original dataset into subsets, the next step in the DAC methodology is the creation of each subset into individual knowledge modules. Each module involves training and testing of the classifier. Since each dataset contains a disjoint subset of the original data, they can be processed in parallel for time efficiency. Leader algorithm has been used in this phase for classification with hamming distance metric. In order to facilitate an efficient way to train the multiple modules simultaneously, the training phase makes use of the DAC resources.

Grid computing is cutting of a large number of genome sequences into shorter numbers of sequences segments. The secondary structures of the segments can be predicted individually by different module of computations and the individual predictions for the small pieces can be assembled to give a predicted structure for the original class. The advantage of the grid computing approach is that it can accommodate a variety of existing and new prediction algorithms in a heterogeneous module. However, the challenge lies in the necessity of ensuring that the predicted results of the smaller pieces are sufficiently consistent with one another so that they can be assembled to generate a reasonable structure for the original family. The Grid concept of genome sequence dataset provides a distributed computing infrastructure for advanced science and engineering [10]. It aims to facilitate flexible, secured, coordinated and controlled resource sharing among computers. The distributed computing power data storage and assistance of virtual organization and dynamic collection of individuals / groups. The grid offers the resources needed to run multiple training processes, thus reducing the total time and cost of the classification procedure. The Grid infrastructure is then responsible for assigning suitable resources according to the description in the DAC and queue the appropriate grid node after the

successful execution of the training process, the resulting outputs are returned. This process is repeated for multiple training processes, each one of them assigned to grid node for execution. In contrary to other parallel classification techniques, the DAC methodology is independent of the actual data mining algorithm, thus providing a degree of freedom to the methodology. Also, due to the fact that the training dataset was equivalent to the actual data representation, the final knowledge modules are also of equivalent accuracy.

Algorithm: DAC Classification distinguish the steps
Input:
S: A set of N number of sequences s1, s2, ……..sN
Threshold: Threshold for generating motifs
DistKey: distance key is threshold for creating a new cluster
Num-subs: Number of sub-sets in given data set

Output:
C: Cluster c1, c2 ……. ck

Algorithm:
Generating the number of motifs m1, m2, …….. mM from a given set of sequences using dynamic program according to threshold and fixed length (L).
m1=substring of size L from fist sequence beginning substring of the
M=1 // number of motifs
do until end of last sequence
sub_str=next substring of size L
for j=1 to no. of motifs(M)
find nearest motif (J) exist for sub_str
end loop j
if nearest motif exist then
sub_str belongs to Jth motif
else
M=M+1
Create a new Mth motif
end if
end do
Generating motif frequency table using above motifs from given set of sequences.
Freq-table (i,j) = number of times the jth motif appears in the ith sequences
Dividing the given data into num_subs number of subsets (disjoins) Round-Robin
For i=1 to num_subs
Classifying the subsets into Ci classes using leader classifier with hamming distance measure and Dist_key
End loop i
Merging the all generated classes in step 3
Merging the similar classes to obtain final classes c1, c2 ……. ck using leader classifies with hamming distance measure and Dist_key.

Finally, the outputs of multiple modules are combined into a single unified module to produce final module and the final module is tested both on the original and test datasets. In this step, the combined multiple knowledge modules were extracted in the previous level, which is treated as a new module and final classes are obtained. The overall efficiency of the module is calculated by testing it on the original dataset.

Experimental Results and Discussions

Tabel 1: Comparision with NNC algorithm of DAC experimental results with Cross Validation.

Classification methods	Grass genome Sequences Dataset (GSD)			Time in sec		No of Class	CA in %
	Total	Training	Testing	Training	Testing		
NNC algorithm	20000	12000	8000	634	548	40	99.75
DAC-Module	20000	12000	8000	46.8	55.81	44	99.82
DAC-Final	44	29	15	0.18	0.011	4	96.59

In order to evaluate the proposed method with the NNC method, we choose grass genome dataset of 136 sequences of 6 families with 8 species from NCBI databank as benchmark for cross validation. Our method shows 87.5% accuracy against the actual classes and results are presented in Table-1. To evaluate time complexity, we have chosen a large dataset of 20000 genome sequences. The proposed method is faster than NNC algorithm as the time complexity in training and testing is 46.98 and 55.81 respectively where as NNC algorithm require training and testing 634 and 548 respectively. The DAC algorithm divides the dataset into sub modules based on round robin method. Each sub modules takes negligible amount of time to classify between the classes and the obtained classes of each module are appropriate merged with classes. This technique radically improves the time complexity for minimizing the accessibility of every sequence.

The performance of the algorithm is assessed and evaluated extensively in terms of speed and accuracy on the two datasets with algorithms proposed by [20] and [13]. Our algorithm outperforms their algorithms in terms of time complexity and classification accuracy and is presented in Table-2. The proposed and NNC algorithms are tested and the results are presented in Table-2. The first, contain the 6 most populated grass genome classes of 20000 grass genome sequences belonging to 22 different grass genome classes. The merged module contains the 4 most populated classes. In this test, it is found that the proposed algorithm outperforms the leader algorithm. Further, we have analyzed our method with very large dataset that can be processed as a single dataset by DAC. In addition to these datasets, we have experiments on nine more large size datasets to analyze the consistency of the proposed technique and the results are presented in Table-3.

Table 2: Comparison of NNC algorithm with DAC experimental results for large size data with time complexity and classification accuracy.

Classification methods	Grass genome Sequences Dataset (GSD)			Time in sec		No of Class	CA in %
	Total	Training	Testing	Training	Testing		
DAC-Module Classification	20000	12000	8000	33	39	22	99.97
	20000	12000	8000	68	106	34	99.89
	20000	12000	8000	76	207	271	98.95
	20000	12000	8000	82	32	50	99.53
	20000	12000	8000	43	73	238	98.04
DAC-Final Classification (DAC Module No of class are total GSD)	22	15	7	0.1	0.01	6	100
	34	23	11	0.1	0.02	3	98.88
	271	181	90	1	0.01	5	98.95
	50	33	17	0.1	0.01	4	98.85
	238	159	79	1	1	32	98.22
NNC algorithm Classification	20000	12000	8000	669	383	2	99.98
	20000	12000	8000	2356	2323	29	99.35
	20000	12000	8000	2366	2287	29	99.98
	20000	12000	8000	1696	1733	35	99.94
	20000	12000	8000	634	548	7	99.98

(Note: Training data of DAC Final class is 2/3 of Total Number of class of Module Classification)

Table 3: Time complexity and classification accuracy of DAC for increasing size data.

	DAC-Module Classification						DAC-Final Classification				
GSD	Total data	Training data	Training Time(s)	Testing Time(s)	Total no of Class	CA in %	Training Data	Training Time(s)	Testing Time(s)	c-sub class	CA in %
GSD 1	5000	3000	18	21	71	98.10	49	0.1	0.01	4	100
GSD 2	10000	6000	123	90	91	99.99	60	0.1	0.01	14	96.66
GSD 3	15000	9000	206	161	91	99.88	63	0.1	0.01	14	96.66
GSD 4	20000	12000	33	39	22	99. 97	15	0.1	0.1	6	100
GSD 5	25000	15000	36	26	9	99.94	6	0.1	0.01	4	100
GSD 6	30000	18000	55	30	9	99.53	6	0.1	0.01	4	100
GSD 7	35000	21000	125	88	9	99.23	6	0.1	0.01	4	100
GSD 8	40000	24000	146	102	9	99.33	6	0.1	0.01	4	100
GSD 9	45000	27000	69	47	9	99.40	6	0.1	0.01	4	100

(Note: Training data of DAC merge class is 2/3 of Total Number of class)

DAC algorithm produces consistent results on large datasets also. Large dataset of grass genome sequences have high time complexity in leader algorithm where in DAC algorithm is radically reduced to twenty times less in training and ten times less in testing though there is no significant difference in CA. The numbers of splits for each dataset are different to keep similar sub dataset size in order to have comparable results. The experimental results show a substantial improvement and also indicate

that the divide and conquer method reduces the processing time as seen from the tables, while the accuracy is fairly constant.

Conclusions

We have presented a novel approach for grass genome classification based on grid and parallel computing concept using Leader classifier and hamming distance. A grass genome dataset is divided into multiple disjoint sets, where each one preserves the original class distribution. The new sets are then mined in parallel for knowledge, using leader classification algorithm, and the extracted knowledge modules are combined procedure to final classes. Results indicate that the proposed method is time efficient and shows that overall accuracy is comparable with other methods. It must be noted that the parallelization of the procedure allows for the processing of much larger datasets as compared to other techniques. The representatives of the Subclass help in improving the CA and hence the Class–Subclass algorithm performs better than the leader algorithm. The hamming distance, approach to obtain similarities as the inner product of the vectors representing the motifs enables one to use linear algebra techniques, to reduce the cost of computation of similarities, and at the same time, keep the error as low as possible. By also taking into account the frequency of motifs in the sequence, the errors can be further reduced.

References

[1] Aggarwal, C.C., Procopiuc, C., Wolf, J., Yu, P.S., and Park, J.S. 1999, Fast algorithm for projected clustering, *ACM SIGMOD* Vol. 28(2), 61-72

[2] Aggarwal, C.C., Yu, P.S., 2000, Finding generalized projected cluster in high dimensional spaces, ACM, SIGMOD, 70-81.

[3] Ananthanarayana, V.S., Murty, M.N., and Subramanian, D.K., 2001, " Efficient clustering of large data sets", Pattern Recognition Letters., 34, pp. 2561–2563.

[4] Attwood, T.K., Croning, M.D.R., Flower, D.R, Lewis, A.P., Mebey, J.E, Scodis, P., Selley, J., and Wright, W., 2000, "PRINT-S:the database formerly known as PRINTS", Nucleic Acids Res., 28, pp.225-227.

[5] Baldi, P.F., and Burnak, S., 2001, Bioinformatics: A Machine Learning Approach", The MIT Press Cambridge, MA.

[6] Beteman, A., Berniy, E., Durbin, R., Eddy S.R., Howem, K.L., and Sonnhammer, E.L.L., 2000, "The Pfam protein families database", Nucleic Acids Res.28, pp.263-266.

[7] Bishop, C., 1995, Neural Networks for Pattern Recognition", Oxford University Press, New York.

[8] Cheng-Long Chuang, Chih-Hung Jen, Chung_Ming Chen, & Grace S., Shieh., 2008. A Pattern recognition approach to Inter time- Lagged genetic interactions", *Bioninfornatics, 24(9),* 1183-1190

[9] Diplaris, S., Tsoumakas, G., Mitkas, and P.A., Vishavas, I., 2005, "Protein Classification with multiple algorithms", In Proc. of 10th Panhellenic

Conference in Informatics, Volos Greece,21-23November, Spring-Verlag, LNCS 3746, 446- 456.

[10] Duda, R., Hart, P., and Stork, D., 2002, Pattern Classification", second ed. John Wiley.

[11] Falquet, L., Pagani, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofman, K., and Bairoach, A., 2002, "The PROSITE database, its status in 2002", Nucleic Acids Res.30, 235-238.

[12] Foster, I., and Kesselman, S., 2003, The Grid 2: Blueprint for a New Computing Infrastructure", 2nd Edition, Morgan Kaufman

[13] Fullen Gorge, Johann-Wolfgang Wagele and Robert Giegerich, 2001, "Minimum conflict: a divide and conquer approach to phylogeny estimation", Bioinformatics, 17(12):1168-1178

[14] Gupta, S.K., Kececioglu, J.D., and SchÃffer,J , A.A. ,1995," Improving the Practical Space and Time Efficiency of the Shortest-Paths Approach to Sum-of-Pairs Multiple Sequence Alignment", Comp. Biol., 2(3), 459-472.

[15] Ming Ouyang, John Case, and Joan Burnside, Divide and Conquer Machine Learning for a Genomics Analogy Problem", Proceedings of the 4th International Conference on Discovery Science, pp. 290 - 303

[16] Patil, S.S., Dhandra, B.V., and Angadi, U.B., 2009, "Efficient Scheme for Classifying Grass Genomes", Proceedings of the Computational Biology: Word Congress on Engineering and Computer Science, WCECS2009, UC, Berkely, San Francisco, USA, Vol. I, pp. 28-31.

[17] Polychroniadou, H. E., Psomopoulos, F. E., and Mitkas, P. A., 2006, "G-Class: A Divide and Conquer Application for Grid Protein Classification", Proce. 2nd ADMKD 2006: Workshop on Data Mining and Knowledge Discovery, ADBIS 2006: The 10th East-European Conference on Advances in Databases and Information Systems, Thessaloniki, Greece, pp. 1-12.

[18] Pujari, A.K., 2002. *Data Mining Techniques*". University Press (India), Pvt. Ltd.

[19] Spath, H., 1980, Cluster Analysis Algorithms for Data Reduction and Classification". Ellis Horwood, Chichester, UK.

[20] Stefan Kurtz, Apurva Narechania, Joshua C Stein and Doreen Ware. (2008) 'A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes"', BMC Genomics, Vol. 9, No.517,pp.1-18

[21] Stoye, J., Moulton, V., and Dress, A.W.M., 1997a, "DCA: An Efficient Implementation of the Divide-and-Conquer Multiple Sequence Alignment Algorithm", CABIOS 13(6), pp.625-626.

[22] Stoye, J., Perrey, S.W., Dress, A.W.M., 1997b, "Improving the Divide-and-Conquer Approach to Sum-of-Pairs Multiple Sequence Alignment", Appl. Math. Lett. 10(2), pp. 67-73.

[23] Vijaya P. A., M Narasimha Murty , and D.K.Subramanian, 2003, "An efficient incremental protein sequence clustering algorithm", Proceedings of IEEE TENCON, Bangalore, India, pp. 409-413.

[24] Yang, J., H. Wang, and W. Wang, Yu.P.S., An Improved Biclustering Method of Analyzing Gene Expression Profiles, Intl. Journal on Artificial Intelligence Tools, Vol. 14, No. 5, Oct. 2005, pp.771-790.