Application of Greedy Randomized Adaptive Search Procedure to the Biclustering of Gene Expression Data

Shyama Das
Department of Computer Science
Cochin University of Science and
Technology, Kochi, Kerala, India

Sumam Mary Idicula
Department of Computer Science,
Cochin University of Science and
Technology, Kochi, Kerala, India

ABSTRACT

Microarray technology demands the development of data mining algorithms for extracting useful and novel patterns. A bicluster of a gene expression dataset is a local pattern such that the genes in the bicluster exhibit similar expression patterns through a subset of conditions. In this study biclusters are detected in two steps. In the first step high quality bicluster seeds are generated using K-Means clustering algorithm. These seeds are then enlarged using a multistart metaheuristic method Greedy Randomized Adaptive Search Procedure (GRASP).In GRASP there are two phases, construction and local search. The Experimental results on the benchmark datasets demonstrate that GRASP is capable of identifying high qua;ity biclusters compared to many of the already existing biclustering algorithms. Moreover far better biclusters are obtained in this algorithm compared to the already existing algorithms based on the same GRASP metaheuristics. In this study GRASP is applied for the first time to identify biclusters from Human Lymphoma dataset.

Categories and Subject Descriptors

H.2. [**Database Management**] Database applications: Data Mining, J.3 [**Computer Applications]**: Life and Medical Sciences

General Terms

Algorithms, Measurement, Experimentation.

Keywords

Gene expression data, greedy randomized adaptive search procedure, K-Means clustering, biclustering, data mining.

1. INTRODUCTION

The relative abundance of mRNA of a gene is called the expression level of a gene. This is measured using DNA microarray technology which revolutionized gene expression study by simultaneously measuring the expression levels of thousands of genes in a single experiment. The data generated by these experiments high dimensional matrix contain thousands of rows (genes) and hundreds of conditions. The experimental conditions can be patients, tissue types, different time points etc. particular gene under a specific condition. Each entry in this matrix is a real number which denotes the expression level of a gene. Genes participating in the same biological process will have similar expression patterns. Clustering is the suitable mining method for identifying these patterns.

Mining various patterns from microarray data is a vital research problem in bioinformatics and clinical research. Being typically high-dimensional, mining functional and class information from such large volumes of data is a crucial event. Hence it calls for appropriate mining strategies. Data Mining techniques in gene expression data comprise of clustering genes by their expression under multiple conditions, classification of a new gene given the expression of other genes with known classification, clustering of conditions based on the expression of many genes, and the classification of a new sample given the expression of genes under that experimental condition. Clustering is the most widely used data mining technique for analyzing gene expression data to group similar genes or conditions. Clustering of co-expressed gene into biologically meaningful groups assists in inferring the biological role of an unknown gene that is co-expressed with a known gene.

However clustering has got its own limitations. Clustering is based on the assumption that all the related genes behave similarly across all the measured conditions. It may reveal the genes which are very closely co-regulated along the entire column. Based on a general understanding of the cellular process, the subsets of genes are co-regulated and co-expressed under certain experimental conditions. But they behave almost independently under other conditions. Moreover clustering partitions the genes into disjoint sets i.e. each gene is associated with a single biological function, which is in contradiction to the biological system [1].

In order to make the clustering model more flexible and to overcome the difficulties associated with clustering the concept of biclustering was introduced. Biclustering is clustering applied in two dimensions, i.e. along the row and column, simultaneously. This approach identifies the genes which show similar expression levels under a specific subset of experimental conditions. The objective is to discover maximal subgroups of genes and subgroups of conditions. Such genes express highly correlated activities over a range of conditions. Biclustering was first introduced by Hartigan who called it direct clustering [2].

Cheng and Church were the first to apply biclustering to gene expression data [3]. Biclustering is a powerful analytical tool when some genes have multiple functions and experimental conditions are diverse.

In this work an algorithm is developed for biclustering gene expression data using GRASP which is a semi-greedy, multi-start metaheuristics which alternates between construction and local search phase to find a globally optimal solution. Initially high quality bicluster seeds are generated using K-Means and they are enlarged using GRASP.

2. BICLUSTERS WITH COHERENT VALUES

A bicluster is a submatrix of the gene expression data matrix. A bicluster of a gene expression dataset is a subset of genes which exhibit similar expression patterns along a subset of conditions. Let X={I1,I2,....IN} be the set of genes and Y={J1,...JM} be the set of conditions in the gene expression dataset. The dataset can be viewed as an NxM matrix A of real numbers. A bicluster is a submatrix B of A and if the size of B is IxJ, then I is a subset of rows X of A, and J is a subset of the columns Y of A. The rows and columns of the bicluster B need not be contiguous as in the expression matrix A. There are four types of biclusters namely biclusters with constant values, biclusters with constant values on rows or columns, biclusters with coherent values, and biclusters with coherent evolutions.Biclusters with coherent values are identified in this work. They are biologically more relevant than biclusters with constant values .The degree of coherence is measured by MSR or Hscore. It is the sum of the squared residue score. The residue score of an element *bij* in a submatrix *B* is defined as *RS(bij)=bij-bIj-biJ+bIJ.*

Hence Hscore or MSR of bicluster *B* is

$$\mathrm{MSR}(B) = \frac{1}{|I||J|} \sum i \in I, j \in J \, (\mathrm{RS}(bij))^2$$

where *I* denotes the row set, *J* denotes the column set, *bij* denotes the element in a submatrix, *biJ* denotes the ith row mean, *bIj* denotes the *j*th column mean, and *bIJ* denotes the mean of the whole bicluster. If the MSR of a matrix is less than certain threshold δ then it is a bicluster and called δ bicluster where δ is the MSR threshold. The value of δ depends on the dataset. For Yeast dataset the value of δ is 300 and for Lymphoma dataset the value of δ is 1200. There is correlation in the matrix if the MSR value is low. The volume of a bicluster or the bicluster size is the product of number of rows and the number of columns in the bicluster. The larger the volume and the smaller the MSR of the bicluster, greater is the quality of the bicluster.

2.1 Encoding of Bicluster

Each bicluster is represented by a binary string of fixed length n+m, where *n* and *m* are the number of genes and conditions of the microarray dataset, respectively. The first n bits is associated to n genes, the following m bits to m conditions. If a bit is set to 1, it means that the corresponding gene or condition belongs to the bicluster; otherwise it does not. This encoding presents the advantage of having fixed size [4].

3. DESCRIPTION OF THE ALGORITHM

In this study biclustering problem is solved using the multistart metaheuristic method greedy randomized adaptive search procedure (GRASP). The algorithm has two major phases. In the first phase, an initial set of seed biclusters are generated using K-Means one way clustering algorithm. The second phase is used to enlarge the seeds by adding more rows and columns using GRASP. Greedy seed growing strategy makes a choice that optimizes a local gain in the hope that this choice will lead to a globally good solution. This will produce only local optimal solutions. Metaheuristic methods have the potential to escape from local minima. Moreover GRASP is semi-greedy. Hence it can combine the advantages of both greedy and random solution constructions.

3.1 Seed Finding

A good seed of a bicluster is a small bicluster with a possibility of accommodating more genes and conditions within the given Hscore threshold. In this algorithm a simple seed finding technique is used [5]. For finding seeds K-Means clustering algorithm is used. K-Means is a partitional clustering algorithm that generates a specific number of disjoint, flat or non-hierarchical clusters. In K-Means clustering algorithm distance measure is a parameter that specifies how the distance between data points in the clustering input is measured. The various distance measures used are Euclidean, Manhattan, Mahanbolis, Cosine angle distance, Hamming distance etc. Here cosine angle distance is selected as the distance measure. First of all gene and condition clusters are obtained from the K-Means one way clustering algorithm.

That is genes in the dataset are partitioned into n gene clusters. Some of the clusters will contain more than 10 genes. They are further divided into groups based on cosine angle distance from the cluster centre so that each group contains at most 10 genes. Similarly conditions in the dataset are partitioned into m clusters and each cluster containing more than 5 conditions is further divided based on cosine angle distance from the cluster center so that each group contains at most 5 conditions. Assume that there are p gene clusters and q condition clusters. All combinations of these p gene clusters and q condition clusters are found. Hscore value for all these combinations are calculated and those with hscore value below a certain threshold is selected as seeds. Thus the gene expression data matrix is partitioned into fixed sized tightly co-regulated submatrices. The Lymphoma dataset is partitioned into 200 gene clusters and 15 condition clusters. The Yeast dataset is partitioned into 140 gene clusters and 3 condition clusters [4].

3.2 Seed Growing using Greedy Randomized Adaptive Search Procedure (GRASP)

GRASP is a multi-start metaheuristics to solve combinatorial Optimization problems. GRASP is an iterative randomized sampling method consisting of two phases: construction and local search. The construction phase will generate a feasible solution, whose neighborhood will be investigated until a local minimum is identified during the process of local search phase. The best overall solution will be reserved as the result. In the construction phase a feasible solution is iteratively developed by adding one element at a time. During each iteration in the construction phase a group of candidate elements are generated by all the elements that can be incorporated into the partial solution under construction without eliminating feasibility. The choice of the next element for incorporation is solved by the evaluation of all candidate elements in accordance with a greedy evaluation function [6].

This greedy function stands for the incremental increase in the cost function due to the incorporation of this element into the solution which is under construction. The evaluation of the elements by this function will result in the creation of a restricted candidate elements (RCL) produced by the best elements. This means that, those elements whose incorporation to the current partial solution will result in the smallest incremental costs. This makes the greedy aspect of the algorithm. The element which is to be included in the partial solution is randomly chosen from those in the RCL. This makes the probabilistic aspect of the heuristic algorithm. Once the selected element is included in the partial solution, the candidate list is reconstituted and the incremental costs are recalculated. This makes the adaptive aspect of the heuristic algorithm.

The solutions generated by the greedy randomized construction are not always optimal even in the case of simple neighborhoods. The local search phase can make the constructed solution better. A local search algorithm functions in an iterative manner by consecutively replacing the current solution by an enhanced solution in the neighborhood of the existing solution. It completes its function process when no better solution is identified in the neighborhood.

When GRASP is applied to gene expression data, conditions and genes are added to the seed in the construction phase. To this effect genes or conditions which are not included in the bicluster are identified. From this list the candidate list is generated by the genes or conditions whose inclusion in the bicluster will not exceed the hscore value above the selected threshold. The candidate list is dynamic in the sense it varies in accordance with the variation of bicluster size. From the candidate lists the best elements are chosen and another list is generated which is known as the restricted controlled list or RCL. The RCL contains genes or conditions which when added result in hscore increment less than a threshold known as RCL threshold which in turn is calculated by the formula hscoremin+ α (hscoremax-hscoremin). Hscoremax is the maximum Hscore value when a candidate is added and hscoremin is the minimum Hscore value of the added candidate for a particular iteration. The value of α varies from 0 to1. The case α=0 corresponds to pure greedy algorithm, while α=1is equivalent to a random construction. Thus the parameter α can control the amounts of greediness and randomness in the algorithm.

From the RCL an element is chosen at random and added to the bicluster. This requires the updation of candidate list and the process is continued. The construction and local search is continued alternately till the Mean square residue score of the bicluster reaches the given threshold. Here the neighborhood search is implemented using best improving strategy. To get biclusters having more conditions gene list and condition list are maintained separately and construction phase is executed in the condition list initially and then followed by the gene list.

```
Procedure construct_candidatelist (bicluster, δ)

    Bicluster1←bicluster;

    notinlist← the list of Genes or Conditions
    not included in the bicluster

    notinlistcount← noofelements(notinlist)

    For i=1:notinlistcount

    Hscorelist[i]=hscore(Bicluster1 U
    notinlist[i])

    End(for)

    Candidatelist={}

    For i=1:notinlistcount

    If Hscorelist[i]< δ

    Candidatelist=candidatelist U Notinlist[i]

    End(for)

    end(construct_candidatelist)
```

```
Procedure BuildRCL(bicluster,C)

   // C is the candidate list
       Sminhscore = inf
     Smaxhscore = -inf
     nocan=noofelements(C)
 for I =1: nocan do
    calculate H[i]← MSR{ bicluster U C[i]}

       if H[i ]<Sminhscore
          Sminhscore=H[i]
```

```
            Endif
            if  H[i ]>Smaxhscore
              Smaxhscore=H[i]
            Endif
   Endfor

  RCLthresh=Sminhscore+α*(Sminhscore-
                                  Sminhscore
     RCL={}
     For i=1:nocan
         If  H[i]<RCLthresh
         RCL=RCL U{ C[i]}
        endif
    end(for)
   end BuildRCL
```

```
Procedure Greedy_Randomized_Construct (Seed)

bicluster←seed;

While solution construction notdone

cand←construct _candidatelist (bicluster, δ)

RCL←BuildRCL (bicluster,cand)

Select an element S from RCL at random

bicluster=bicluster U{S}

Update G or C

End(while)

End(Greedy_Randomized_Construct)
```

```
      Procedure Local_Search(bicluster)

 While there exists s∈ genelist or conditionlist
 such that  hscore(biclusterU s)<hscoer(bicluster) do

 bicluster={bicluster U s}

  end(while)
 end(Local_Search)
```

4. RESULTS

4.1 Datasets Used

We have implemented the proposed algorithm in Matlab and tested on bench mark data set namely Yeast Saccharomyces Cerevisiae cell cycle expression dataset. The yeast dataset is based on Tavozoie et al [7]. The expression values were transformed by scaling and logarithm x→100 log (10^5x) and the result was matrix of integers in the range 0 and 600. Missing values are represented by -1. Human B-cell Lymphoma expression data contain 4026 genes and 96 conditions. The dataset was downloaded from the website for supplementary information for the article by Alizadeh et al. (2000) [8]. The expression levels were reported as log ratios. After scaling by a factor of 100 the values in the dataset are integers in the range -750 to 650. There are 47,639 (12.3%) missing values in the Lymphoma dataset. Missing values were represented by 999. In the Lymphoma dataset missing values are replaced by random numbers between -800 and 800 as in ref [3].The datasets are obtained from http://arep.med.harvard.edu/biclustering

4.2 Bicluster Plots for Yeast Dataset

Eight biclusters obtained using GRASP is shown in Figure 1. Here biclusters with all 17 conditions are obtained. From the bicluster plots which show strikingly similar upregulation and down regulation it is concluded that GRASP is an ideal method to identify biclusters of large volume from gene expression data.

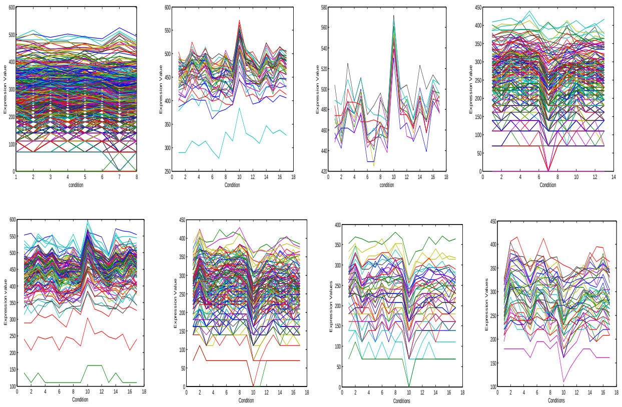

Figure 1. Six biclusters found for the Yeast dataset. Bicluster labels are (a), (b), (c), (d), (e) and (f) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 1 using bicluster label. All the means squared residues are lower than 215.

Table 1. Information about biclusters of Figure 1

Label	Rows	Columns	Volume	MSR
(a)	783	8	6264	215.0790
(b)	42	17	714	121.6900
(c)	12	17	204	69.9591
(d)	208	13	2704	193.6400

(e)	108	17	1836	200.7372
(f)	140	17	2380	200.0088
(g)	47	17	799	145.3612
(h)	44	17	748	163.9544

In the above table the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the last column contains the mean squared residue of the bicluster.

4.3 Bicluster Plots for Human Lymphoma Dataset

In Figure 2 eight biclusters obtained using GRASP are shown. A biclusters with maximum 89 conditions is obtained using this method. From the bicluster plots it is clear that biclusters show strikingly similar up-regulation and down regulation.

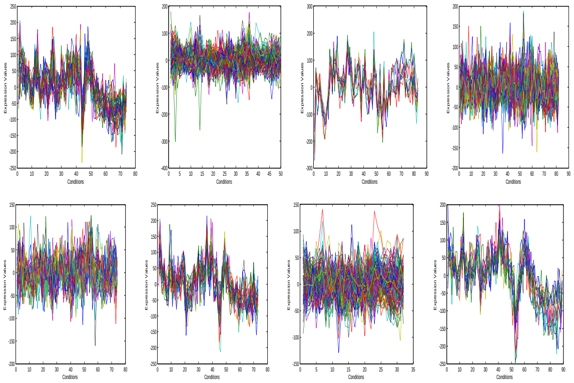

Figure 2. Eight biclusters found for the Lymphoma Dataset. Bicluster labels are (p), (q), (r), (s), (t), (u), (v) and (w) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about the biclusters can be obtained from Table 2 using bicluster label. All the means squared residues of the biclusters are lower than 1200.

Table 2. Information about biclusters of Figure 2

Label	Rows	Columns	Volume	MSR
(p1)	16	89	1424	1196.9
(q1)	38	74	2812	1189.8
(r1)	175	50	8750	1075.2
(s1)	10	83	830	1182.1
(t1)	62	82	5084	1197.3
(u1)	34	74	2516	1019.5
(v1)	24	73	1752	1197.9
(w1)	132	32	4224	751.9

In the table given above the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the last column contains the mean squared residue or Hscore of the bicluster.

5. BIOLOGICAL SIGNIFICANCE

Biclusters can be evaluated using prior biological knowledge [9]. Existence of biclusters comprising a significant proportion of those genes that considered similar biologically is a proof that a specific biclustering technique produces biologically relevant results. Biological relevance of biclusters obtained using GRASP algorithm is verified using a bicluster of size 30x17. GO annotation database can be used to determine the biological significance of biclusters. In this database genes are assigned to three structured controlled vocabularies. Gene products are described in terms of associated biological process, components and molecular functions in a species-independent manner. To evaluate the statistical significance for the genes in each bicluster p-values are used. P-values indicate the extent to which the genes in the bicluster match with the different GO categories. Smaller p-values indicates better match. . P-values can be calculated using a cumulative hypergeometric distribution. The probability p for finding at least k genes, from a particular GO category (function, process or component) within a cluster of size n, is calculated as

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$$

where f is the total number of genes within a category and g is the total number of genes within the genome. Yeast genome gene ontology term finder [10] is a database available in the Internet which can be used to evaluate the biological significance of biclusters. In the bicluster selected for testing the biological significance there are 30 genes namely YAL003W, YAL038W, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR106W, YBR118W, YBR189W,YCR013C

YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR382W, YDR385W.

Figure 3 shows the significant GO terms for the set of 30 genes along with their p values. It shows the branching of a generalized molecular function into sub-functions like structural molecule activity, binding and enzyme regulator activity. These activities are clustered using genes to produce the final result. Figure 3 is obtained when gene ontology database [10] is searched by entering the names of genes and selecting function ontology.

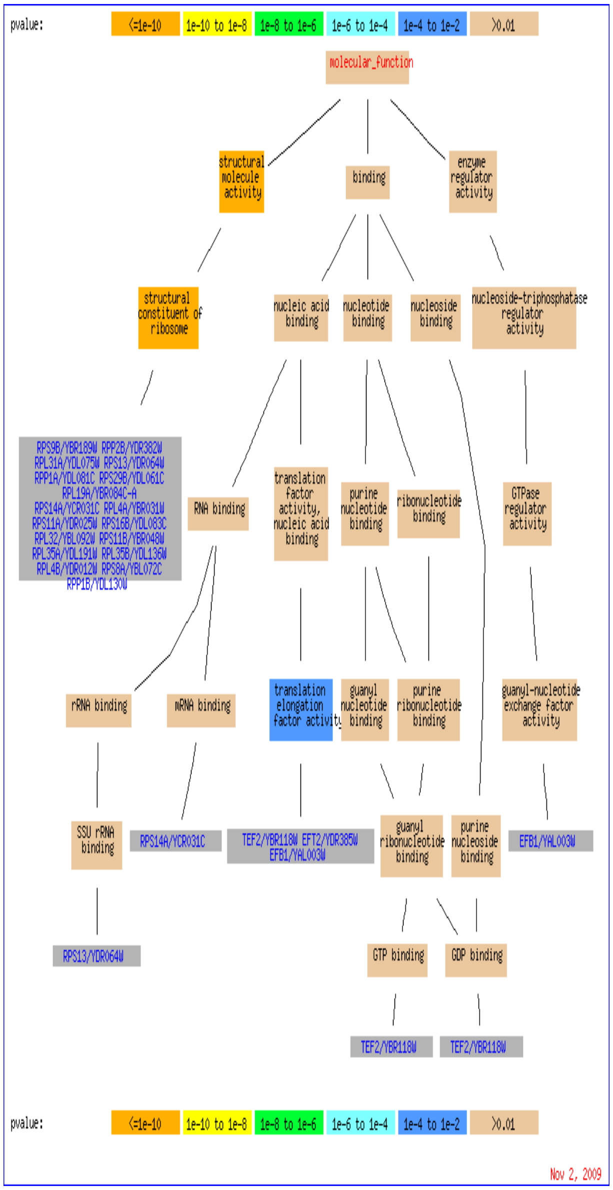
pvalue:
<=1e-10
1e-10 to 1e-8
1e-8 to 1e-6
1e-6 to 1e-4
1e-4 to 1e-2
>0.01
molecular_function
structural molecule activity
binding
enzyme regulator activity
structural constituent of ribosome
nucleic acid binding
nucleotide binding
nucleoside binding
nucleoside-triphosphatase regulator activity
RPS9B/YBR189W RPP2B/YDR382W
RPL31A/YDL075W RPS13/YDR064W
RPP1A/YDL081C RPS29B/YDL061C
RPL19A/YBR084C-A
RPS14A/YCR031C RPL4A/YBR031W
RPS11A/YDR025W RPS16B/YDL083C
RPL32/YBL092W RPS11B/YBR048W
RPL35A/YDL191W RPL35B/YDL136W
RPL4B/YDR012W RPS8A/YBL072C
RPP1B/YDL130W
RNA binding
translation factor activity, nucleic acid binding
purine nucleotide binding
ribonucleotide binding
GTPase regulator activity
rRNA binding
mRNA binding
translation elongation factor activity
guanyl nucleotide binding
purine ribonucleotide binding
guanyl-nucleotide exchange factor activity
SSU rRNA binding
RPS14A/YCR031C
TEF2/YBR118W EFT2/YDR385W
EFB1/YAL003W
guanyl ribonucleotide binding
purine nucleoside binding
EFB1/YAL003W
RPS13/YDR064W
GTP binding
GDP binding
TEF2/YBR118W
TEF2/YBR118W
pvalue:
<=1e-10
1e-10 to 1e-8
1e-8 to 1e-6
1e-6 to 1e-4
1e-4 to 1e-2
>0.01
Nov 2, 2009

Figure 3. Sample of 30 genes for Yeast data, with corresponding GO terms and their parents for function ontology

The Table 3 given below shows the significant GO terms used to describe the set of 30 genes of the bicluster for the process, function and component ontologies. The common terms are described with increasing order of p-values or decreasing order of significance. In Table 2 the first entry of the second column with the title process contains the tuple Translation (22,8.73e-15) which means that 22 out of the 30 genes of the bicluster are involved in the process of translation and their p-value is 8.73e -15. The 22 genes are YAL003W, YBL072C, YBL092W, YBR031W, YBR048W, YBR084C-A, YBR118W, YBR189W, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL229W, YDR012W, YDR025W, YDR064W, YDR382W and YDR385W. Second entry indicates that 23 out of 30 genes are involved in cellular protein metabolic process. In the table the first entry of the column with the title Function contains the entry structural constituent of ribosome (18,4.61e-19).That means 18 genes are annotated to this fuction. This proves that the bicluster contains biologically similar genes and the method used here is capable of identifying biologically significant biclusters.

Table 3. Significant shared GO terms (process, function, component) of the 30 genes in a bicluster obtained using GRASP algorithm

No.of genes	Process	Function	Component
30	Translation (22, 8.73e-15) Cellular protein metabolic process (23, 3.16e-09) protein metabolic process (23, 6.81e-09) Macromolecule metabolic Process (26, 7.53e-06)	structural constituent of ribosome (18, 4.61e-19) structural molecule activity (18, 1.02e-15) Translation elongation factor (3, 0.00024)	cytosolic ribosome (18,1.15e-20) ribosome (21,6.55e-20) ribonucleo protein complex (23,1.81e-16) Cytoplasmic Part (25,3.08e-05)

6. COMPARISON

Results obtained by related algorithms such as GRASP[11], RGRASP[12], CGRASP[13], SEBI [14], Cheng and Church's algorithm (CC) [3], and the algorithm FLOC by Yang et al. [15] and DBF [16] etc compared with GRASP in this study on Yeast dataset are given in Table 4. All the algorithms listed in Table 4 are having MSR value more or less equal to 200, even though the maximum limit of δ is 300. SEBI (Sequential Evolutionary Biclustering) is based on evolutionary algorithms. CC algorithm used greedy approach by removing rows and columns from the from the gene expression data matrix to find a bicluster. The model of bicluster proposed by Cheng and Church was generalized by Yang et al (2003). They developed FLOC which is a probabilistic algorithm that can discover a set of possibly overlapping biclusters simultaneously. Zhang et al proposed Deterministic Biclustering with frequent pattern mining (DBF). In DBF good quality biclusters seeds are generated using frequent pattern mining. These seeds are then enlarged by adding more genes or conditions.

In the case of GRASP algorithm presented here all fields are better than GRASP in [11], RGRSP [12], CGRASP [13], SEBI, CC, FLOC and DBF except that DBF is having lower value for average residue score. For Yeast dataset biclusters with all 17 conditions are obtained in this method. Maximum number of conditions obtained for Lymphoma dataset is 89. In metaheuristic methods like multi-objective evolutionary computation [17] the maximum number of conditions obtained is only 11 for Yeast dataset and 49 for the Lymphoma dataset. For the Yeast dataset the maximum number of genes obtained for this algorithm in all the 17 conditions is 140 with Hscore value 200.0088. The maximum number of genes for 17 conditions is obtained by multi-objective PSO [18] is the maximum available in the literature published so far. They obtained 141 genes for 17 conditions with Hscore value 203.25.

Table 4. Performance comparison between GRASP and other algorithms for Yeast Dataset

Algori thm	Avg. Residue	Avg. Num. Genes	Avg. Num. Cond.	Avg. Vol.	Largest Bicluster
GRASP in this study	166.85	215.50	14.83	2350.33	6264
GRASP [11]	188.57	30.00	14.00	430.33	1335
RGRASP [12]	182.34	21.25	13.13	283.38	854
CGRASP [13]	187.05	18.20	12.20	215.40	319
SEBI	205.18	13.61	15.25	209.92	1394
CC	204.29	166.71	12.09	1576.98	4485
FLOC	187.54	195.00	12.80	1825.78	2000
DBF	114.70	188.00	11.00	1627.20	4000

As is clear from the above table the average mean squared residue, the average number of genes and conditions, average volume and largest bicluster size are compared for various algorithms. For the average mean squared residue field lower values are better where as higher values are better for all other fields.

The Table 5 given below provides comparison of results obtained by various biclustering algorithms for the Human lymphoma dataset. In this study GRASP is applied for the first time to Human Lymphoma dataset. Cardinality based GRASP [19] and reactive GRASP [20] are applied to find biclusters from Lymphoma data.

Table 5. Performance comparison between GRASP and other algorithms for Human Lymphoma dataset

Algorithm	Avg. gene. Num.	Avg. cond.num .	Avg. Volume	Avg. MSR
GRASP	61.38	69.63	3424.00	1101.35
SEBI	14.07	43.57	615.84	1028.84
CC	269.22	24.50	4595.98	850.04

7. CONCLUSION

In this paper the GRASP metaheuristics is uaed for finding biclusters in gene expression data. In the first step K-Means algorithm is used to group rows and columns of the data matrix separately. Then they are combined to produce small tightly coregulated submatrices. Then these seeds are enlarged using GRASP. The algorithm is implemented on both benchmark datasets. The results obtained for Yeast dataset prove that the GRASP algorithm performs better than the other approaches. Here the biclusters discovered are larger having more genes and conditions with low Hscore value. In short GRASP method identifies high quality biclusters which manifest strikingly similar up-regulations and down-regulations under a set of experimental conditions that can be inspected visually by using the bicluster plots. The quality of the biclusters identified by the GRASP metaheuristics in this study is far better than the already existing biclustering algorithms. Moreover far better biclusters are obtained in this algorithm compared to the already existing algorithms based on the same GRASP metaheuristics. In the already existing works based on the same GRASP metaheuristics not even a single bicluster is identified with all 17 conditions. On the other hand in this work many biclusters with all 17 conditions are identified. Furthermore in this study GRASP is applied for the first time to the Human Lymphoma dataset.

8. REFERENCES

[1] Madeira S. C. and Oliveira A. L., “Biclustering algorithms for Biological Data analysis: a survey” IEEE Transactions on computational biology and bioinformatics, pp. 24-45, 2004.

[2] J. A. Hartigan, “Direct clustering of Data Matrix”, Journal of the American Statistical Association Vol.67, no.337, pp. 123-129, 1972.

[3] Yizong Cheng and George M. Church, “Biclustering of expression data”, Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology. pp. 93-103, 2000.

[4] Anupam Chakraborty and Hitashyam Maka “Biclustering of Gene Expression Data Using GeneticAlgorithm” Proceedings of Computation Intelligence in Bioinformatics and Computational Biology CIBCB, pp. 1-8, 2005.

[5] Chakraborty A. and Maka H., “Biclustering of gene expression data by simulated annealing”,HPCASIA ’05, pp. 627-632, 2005.

[6] Feo TA and Resende MGC "Greedy randomizedadaptive search procedures",Journal of Global Optimization vol 6 1995. pp. 109-133.

[7] Tavazoie S., Hughes J. D., Campbell M. J., Cho R. J. and Church G. M., "Systematic determination of genetic network architecture", Nat. Genet., vol.22, no.3 pp. 281-285, 1999.

[8] Alizadeh, A. A. et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature Vol.43,no. 6769, pp. 503-11, 2000.

[9] Amos Tanay, Roded Sharan and Ron Shamir, "Discovering Statistically significant Biclusters in Gene Expression Data," Bioinformatics, 2000. Vol. 18, Suppl 1, pp.S136-44.

[10] SGD GO Termfinder [http://db.yeastgenome.org/cgi bin/ GO/ goTermFinder]

[11] Dharan S and Nair AS, "Biclustering of gene expression data using greedy randomized adaptive search procedure", IEEE TENCON 2008. pp. 1-5.

[12] Smitha Dharan, Achuthsankar S. Nair, Biclustering of Gene expression Data using Reactive Greedy Randomized Adaptive Search Procedure", BMC Bioinformatics, 2009. Vol. 10, Suppl 1: s27

[13] Smitha Dharan and Achuthsankar S Nair, "Cardinality based Greedy Randomized Adaptive Search Algorithm for the detection of biclusters in Microarray gene expression data", *Proc. Int. Conf. Advanced Computing and Communication Technologies for High Performance Applications*, 2008, Vol. 1, pp. 244-248.

[14] Federico Divina and Jesus S. Aguilar-Ruize, "Biclustering of Expression Data with Evolutionary computation", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, pp. 590-602, 2006.

[15] J. Yang, H. Wang, W. Wang and P. Yu, "Enhanced Biclustering on Expression Data", Proc. Third IEEE Symp. BioInformatics and BioEng. (BIBE'03), pp. 321-327, 2003.

[16] Z. Zhang, A. Teo, B. C. Ooi, K. L. Tan, "Mining deterministic biclusters in gene expression data", In: Proceedings of the fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), 2004, pp. 283-292, 2004.

[17] Banka H. and Mitra S., "Multi-objective evolutionary biclustering of gene expression data", Journal of Pattern Recognition, Vol.39. pp. 2464-2477, 2006.

[18] Junwan Liu, Zhoujun Lia and Feifei Liu "Multi-objective Particle Swarm Optimization Biclustering of Microarray Data", IEEE International Conference on Bioinformatics and Biomedicine, pp. 363-366, 2008.

[19] Shyama Das and Sumam Mary Idicula "Application of Cardinality based Grasp to the Biclustering of Gene Expression Data" International Journal of Computer Applications. 2010.

[20] Shyama Das and Sumam Mary Idicula "Application of Reactive Grasp to the Biclustering of Gene Expression Data" Proceedings of the ACM International Syposium on Biocomputing, 2010.