MEAN SQUARED RESIDUE BASED BICLUSTERING ALGORITHMS FOR THE ANALYSIS OF GENE EXPRESSION DATA

Thesis submitted by

SHYAMA DAS

In partial fulfilment of the requirements

For the award of the degree of

DOCTOR OF PHILOSOPHY

UNDER THE FACULTY OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

KOCHI – 682 022

INDIA

July 2011

Declaration

I hereby declare that the work presented in this thesis entitled **"Mean Squared Residue Based Biclustering Algorithms for the Analysis of Gene Expression Data"** is based on the original research work carried out by me in the Department of Computer Science, Cochin University of Science and Technology, Kochi – 682022, under the supervision and guidance of Dr. Sumam Mary Idicula, Professor, Department of Computer Science, Cochin University of Science and Technology, Kochi - 682022. The results presented in this thesis or parts of it have not been presented for the award of any other degree.

Kochi – 682022
July, 2011

SHYAMA DAS
Research Scholar

Certificate

This is to certify that the thesis entitled **"Mean Squared Residue Based Biclustering Algorithms for the Analysis of Gene Expression Data"** is a bonafide record of the research work carried out by Ms. Shyama Das in the Department of Computer Science, Cochin University of Science and Technology, Kochi – 682022, under my supervision and Guidance.

Kochi – 682022
July, 2011

Dr. Sumam Mary Idicula
Supervising Guide, Professor,
Department of Computer Science
Cochin University of Science and
Technology, Kochi-682022, Kerala.

Dedicated To
My Lord and Saviour
Jesus Christ

Acknowledgement

The author is deeply grateful to many who provided the support in carrying out the research work and the preparation of the thesis. The author offers foremost thanks and supreme glory to the God Almighty for providing the wisdom and health towards the completion of this research work.

The author expresses her sincere gratitude and appreciation to the supervising guide Dr.Sumam Mary Idicula, Professor, Department of Computer Science, Cochin University of Science and Technology, for her constant encouragement, support and guidance. In many difficult occasions she rendered the much needed mental support. Her strenuous effort in reviewing the research papers and the thesis, and her creative suggestions are highly appreciated.

The author expresses her profound feelings of gratitude and respect to Dr. K. Poulose Jacob, Professor and Head, Department of Computer Science, Cochin University of Science and Technology, for his constant encouragement and valuable suggestions. His sincerity, calmness and supportive attitude enabled the successful completion of this work.

The author is grateful to Dr.S.N.Omkar, Professor, Department of Aeorospace Engineering, IISc, Banglore, for introducing the area of Biologically Inspired Computing, through a short term course. The author also owes gratitude to Senthil Kumar, Research Scholar under the supervision of Dr.S.N.Omkar for his technical guidance and help. The author is also thankful to Dr. Mathew

Jacob. T, Professor, Department of Computer Science, IISc Banglore, for his suggestions. The author is also indebted to the Librarian and the students Pratheeksha, Lokesh, Saraswathy, Gilesh M.P of IISc Banglore for their help in various ways to conduct the literature survey.

The author's deepest gratitude and respect also goes to her husband Dr. K. George Joseph for his suggestions, advice and all the mental support. Without his help the completion of this work would not have been possible. The author is indebted to her children Jerusha and Jovana, who beared with her in spite of the lack of proper attention and care to them during the research work. At this time the author remembers her dear son Joshua. Inspite of the pain in heart due to his absence, Lord Jesus Christ strenghned her to complete the research work.

The author is very grateful to Dr.V.P.Devassia, Dr. Rekha K James, and Dr. Sheena Mathew for their cooperation, support and suggestions. The author is also grateful to Sudheer A.P., formerly collegue in the College of Engineering, Chengannur for his valuable suggestions regarding the research. The author is deeply indebted to her friends Dr.Sobha Cyrus, Nisha Kuruvilla, Philip Cherian, Prime Kumar and Renu George for their support and help. The author acknowledges the contribution of the technical and non-technical staff in the Department of Computer Science, Cochin University of Science and Technology. The author owes heartfelt thanks to her parents for their motivation, encouragement and support. The author is grateful to all who have been helpful in the completion of this work.

Shyama Das

Abstract

Computational Biology is the research area that contributes to the analysis of biological data through the development of algorithms which will address significant research problems. The data from molecular biology includes DNA, RNA, Protein and Gene expression data. **Gene Expression Data** provides the expression level of genes under different conditions. Gene expression is the process of transcribing the DNA sequences of a gene into mRNA sequences which in turn are later translated into proteins. The number of copies of mRNA produced is called the expression level of a gene. Gene expression data is organized in the form of a matrix. Rows in the matrix represent genes and columns in the matrix represent experimental conditions. Experimental conditions can be different tissue types or time points. Entries in the gene expression matrix are real values. Through the analysis of gene expression data it is possible to determine the behavioral patterns of genes such as similarity of their behavior, nature of their interaction, their respective contribution to the same pathways and so on. Similar expression patterns are exhibited by the genes participating in the same biological process. These patterns have immense relevance and application in bioinformatics and clinical research. These patterns are used in the medical domain for aid in more accurate diagnosis, prognosis, treatment planning, drug discovery and protein network analysis.

To identify various patterns from gene expression data, **data mining techniques** are essential. **Clustering** is an important data mining

technique for the analysis of gene expression data. To overcome the problems associated with clustering, biclustering is introduced. Biclustering refers to simultaneous clustering of both rows and columns of a data matrix. Clustering is a global model whereas biclustering is a local model. Discovering local expression patterns is essential for identifying many genetic pathways that are not apparent otherwise. It is therefore necessary to move beyond the clustering paradigm towards developing approaches which are capable of discovering local patterns in gene expression data.

A bicluster is a submatrix of the gene expression data matrix. The rows and columns in the submatrix need not be contiguous as in the gene expression data matrix. Biclusters are not disjoint. Computation of biclusters is costly because one will have to consider all the combinations of columns and rows in order to find out all the biclusters. The search space for the biclustering problem is 2^{m+n} where m and n are the number of genes and conditions respectively. Usually m+n is more than 3000. The biclustering problem is NP-hard. Biclustering is a powerful analytical tool for the biologist.

The research reported in this thesis addresses the problem of biclustering. Ten algorithms are developed for the identification of coherent biclusters from gene expression data. All these algorithms are making use of a measure called mean squared residue to search for biclusters. The objective here is to identify the biclusters of maximum size with the mean squared residue lower than a given threshold. All these

algorithms begin the search from tightly coregulated submatrices called the seeds. These seeds are generated by K-Means clustering algorithm.

The algorithms developed can be classified as constraint based, greedy and metaheuristic. Constraint based algorithms uses one or more of the various constraints namely the MSR threshold and the MSR difference threshold.The greedy approach makes a locally optimal choice at each stage with the objective of finding the global optimum. In metaheuristic approaches Particle Swarm Optimization (PSO) and variants of Greedy Randomized Adaptive Search Procedure (GRASP) are used for the identification of biclusters.

These algorithms are implemented on the Yeast and Lymphoma datasets. Biologically relevant and statistically significant biclusters are identified by all these algorithms which are validated by Gene Ontology database. All these algorithms are compared with some other biclustering algorithms. Algorithms developed in this work overcome some of the problems associated with the already existing algorithms. With the help of some of the algorithms which are developed in this work biclusters with very high row variance, which is higher than the row variance of any other algorithm using mean squared residue, are identified from both Yeast and Lymphoma data sets. Such biclusters which make significant change in the expression level are highly relevant biologically.

CONTENTS

Chapter 6

Chapter 7

…..ൠ…..

List of Tables

List of Figures

List of Abbreviations

ALL	-	Acute Lymphoblastic Leukemia
AML	-	Acute Myeloid Leukemia
AMR	-	Average Mean Squared Residue
ANG	-	Average Number of Genes
ANC	-	Average Number of Conditions
ANN	-	Artificial Neural Network
AV	-	Average Volume
CC	-	Cheng and Church algorithm
CGRASP	-	Cardinality based Greedy Randomized Adaptive Search Procedure
DBF	-	Deterministic Biclustering with Frequent pattern mining
DNA	-	Deoxyribonucleic Acid
FLOC	-	Flexible Overlapped Biclustering
GO	-	Gene Ontology
GRASP	-	Greedy Randomized Adaptive Search Procedure
ISA	-	Iterative Signature Algorithm
ISIMSRDT	-	Iterative Search with Incremental MSR difference Threshold
KNN	-	K-Nearest Neighbour
LB	-	Largest Bicluster
MDS	-	Multi Dimensional Scaling
MOEA	-	Multi Objective Evolutionary Algorithm
MOGAB	-	Multi-objective Genetic algorithm
mRNA	-	messenger RNA
MSR	-	Mean Squared Residue
MSRT	-	Mean Squared Residue Threshold
MSRDT	-	Mean Squared Residue Difference Threshold
OPSM	-	Order Preserving Submatrix Problem
PCA	-	Principle Component Analysis
PSO	-	Particle Swarm Optimization
RCL	-	Restricted Candidate List

RGRASP	-	Reactive Greedy Randomized Adaptive Search Procedure
RNA	-	Ribonucleic Acid
RWB	-	Random-Walk-based Biclustering
SAMBA	-	statistical-Algorithmic Method for Bicluster Analysis
SEBI	-	Sequential Evolutionary Biclustering
SGAB	-	Single Objective Genetic Algorithm for Biclustering
SGSC	-	Seed Growing using Separate Constraints for Genes and Conditions
SMOB	-	Sequential Multi-objective Biclustering
SVM	-	Support Vector Machine

Chapter 1

Introduction

Computational molecular biology deals with different kinds of biological data. Gene expression data is one among them. Hence some basics of molecular biology are given in this chapter. Gene expression data is the basic data used in this thesis. This chapter gives a brief description of microarray technology by which the gene expression data is measured. The chapter also describes the motivation for selecting the research problem, along with the goal, objectives, scope and contribution of the research work. The chapter also gives an overview of the research work detailed in this thesis.

1.1 Computational Molecular Biology

Molecular Biology is the most active field in biology today. An important part of molecular biology concerns the study of genetic material such as DNA, RNA, proteins, chromosomes and genes. In this chapter some basics of molecular biology are introduced for facilitating the understanding of the gene expression data, the data which underlies this thesis. Computational molecular biology [84] is an interdisciplinary subject involving fields as diverse as biology, computer science, information technology, mathematics, physics, statistics and chemistry.

1.2 Preliminaries from Molecular Biology

Cells are the basic building blocks of every organism. There is a central core in the cell called nucleus. Inside the nucleus there is an important molecule known as deoxyribonucleic acid (DNA). All living organisms contain DNA. All the information required for the development and functioning of an organism is encoded in the DNA molecule [3]. DNA molecules store the genetic information of an organism. These molecules are made of two polynucleotide chains (or strands) forming the double helix structure (Figure 1.1). The four nucleotides adenine (A), Cytosine (C), Guanine (G) and Thymine (T) are the building blocks of a DNA molecule. In the double stranded DNA one particularity is the complementary base pairing, i.e., a particular base on one strand binds only to a complementary base on the opposite strand. In other words, “A” binds only to “T”, and “C” to “G” (Figure 1.1). Inside the nucleus DNA is packaged in the form of chromosomes [57] or several

linear DNA molecules called chromosomes, are present in the cell nucleus. There are 24 distinct chromosomes for human beings [95]. They are together known as genome. RNA is molecule which is informationally similar to DNA. RNA is also made up of four nucleotides like DNA. But in RNA the Thymine (T) is replaced by another molecule called Uracil (U). Moreover RNA is single stranded where as DNA is double stranded. The major function of RNA is to selectively copy information from DNA and also to bring this information out of the nucleus for using it where it is intended to be [1]. A *gene is a segment of DNA, which contains the formula for the chemical composition of one particular protein* [4]. Proteins are the most important working molecules of life. Most of the biological processes which take place in a cell are carried out by proteins [40]. Proteins which are the final products of genes are vital to the functioning of cells. The structural components of the cells are constituted of proteins and they catalyze biochemical reactions.

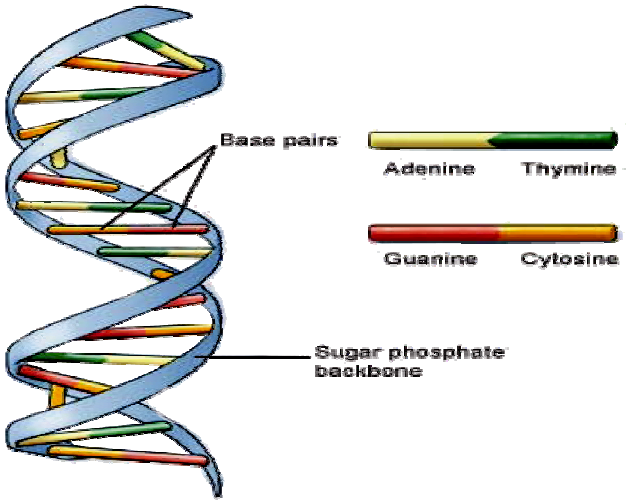


Figure 1.1 Example of a double stranded DNA molecule.

1.2.1 From DNA to Proteins

The process of producing a protein from the information in its corresponding gene in DNA in two phases such as transcription and translation is called *protein synthesis*. Gene expression is the process of transcribing a gene's DNA sequence into mRNA sequences, which in turn are later translated into proteins [105]. Messenger RNA (mRNA) is generated in a process called transcription. In short gene expression is the process by which the genetic information contained in the genes is translated into mRNA molecules and later into proteins. The number of copies of mRNA produced in the process of translation is called the expression level of the gene. The regulation of gene expression level is important for proper functioning of a cell. If the amount of protein required by the cell is more, then more copies of the corresponding mRNA molecule is produced. In short, the amount of specific mRNA copies produced by a gene refers to the activity of the gene. The more copies of mRNA produced, the higher the gene is expressed, and the more proteins will be generated. Genes with high abundance of mRNA copies are called up-regulated genes. On the other hand, if there are no or only a few specific mRNA copies are present, then the associated genes are called down-regulated genes. All the cells in a given multi-cellular organism carry the same genetic code. But the higher order species consist of highly specialized cell types, appearing in different locations of the body with different tasks. But the question arises as to why do the skin cells, nerve cells and blood cells, which all have the same genetic code, behave so differently? The answer is that different genes are active, or

expressed in the different cell types, making them produce their own specific set of proteins. The *expression profile* of a cell is the collected expression levels of all genes in the cell [58].

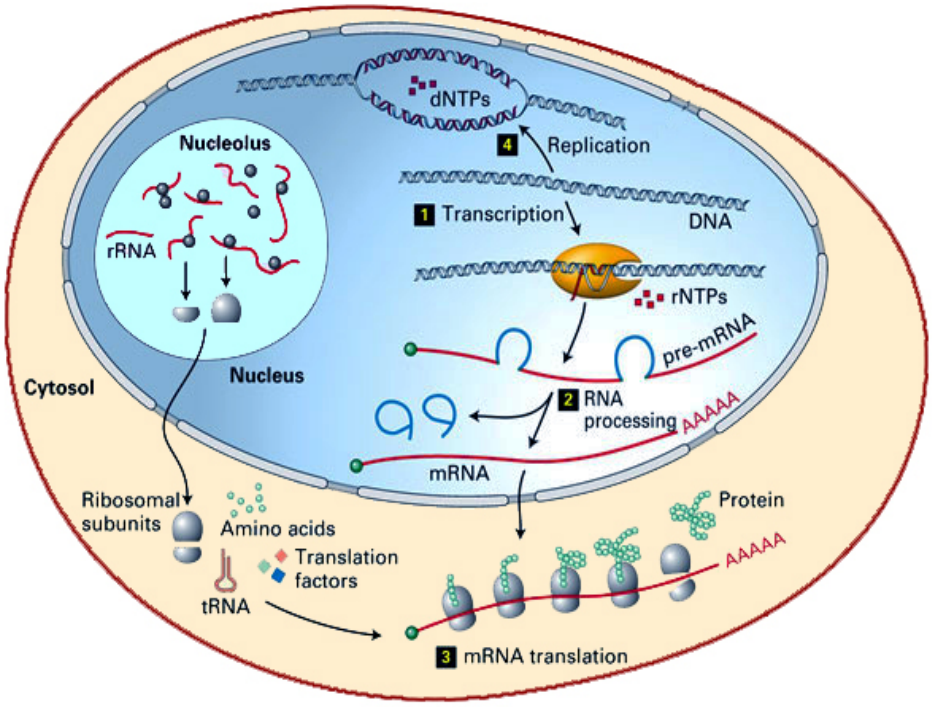


Figure 1.2 The main stages of gene expression. Step 1 corresponds to the transcription of DNA to RNA molecules. Step 2 corresponds to the translation of messenger RNA (mRNA) to protein molecules.

1.2.2 Measuring Gene Expression with Microarrays

Several microarray technologies have been developed to study gene expression regulation. A very popular microarray technology based on oligonucleotide chips is produced by the company Affymetrix. The other widely used microarray technology is cDNA-arrays. In both these

techniques the quantity of mRNA is measured based on hybridization [105]. DNA microarray is constituted of thin glass or nylon substrates. They contain specific DNA gene samples spotted in an array by a robotic printing device. Fluorescently labelled m-RNA from an experimental condition is spread onto the DNA gene samples in the array. This m-RNA hybridizes with some DNA gene samples depending on the double helical characteristics. Later a laser scans the array and the sensors for detecting the fluorescence levels using red and green dyes. The red and green dyes indicate the strength with which the sample expresses each gene. The logarithmic ratio between the two intensities of each dye is calculated and used as the gene expression data. The relative abundance of the spotted DNA sequences in a pair of DNA or RNA samples is measured by evaluating the differential hybridization of the two samples to the sequences in the array [21, 44, 95].

1.3 Motivation

Through the analysis of gene expression data it is possible to determine the behavioural patterns of genes such as similarity of their behaviour, nature of their interaction, their respective contribution to the same pathways and so on. Similar expression patterns are exhibited by the genes participating in the same biological process. These patterns have immense relevance and application in bioinformatics and clinical research. These patterns are used in the medical domain for aid in more accurate diagnosis, prognosis, treatment planning, drug discovery and protein network analysis. In this context some **research questions** arise.

How to identify the co-expressed genes? What are the computational methods that can be used to identify the co-expressed genes? What are the constraints to be considered while selecting the computational methods? How can we validate the results obtained from the computational methods in association with the biological annotations already available?

In order to identify various patterns from gene expression data, **data mining techniques** are essential. Major data mining techniques which can be applied for the analysis of gene expression data include, clustering, classification, association rule mining etc. **Clustering** is an important data mining technique for the analysis of gene expression data. However clustering has some disadvantages. To overcome the problems associated with clustering, **biclustering** is introduced. Clustering is a global model where as biclustering is a local model. Discovering such local expression patterns is essential for identifying many genetic pathways that are not apparent otherwise. It is therefore necessary to move beyond the clustering paradigm towards developing approaches which are capable of discovering local patterns in gene expression data.

1.4 Scope

The vast amount of data emerging from molecular biology, especially in the form of DNA, RNA, protein sequences and gene expression data demands the development of algorithms by computational scientists. In the context of gene expression data, **design and development of algorithms** can contribute towards the identification of biclusters with coherent values. Hence this study deals with the

development of algorithms for the identification of coherent biclusters from gene expression data. The degree of coherence is measured by mean squared residue. There are many algorithms for the identification of coherent biclusters from gene expression data. The algorithms developed in this thesis overcome some of the disadvantages associated with the existing algorithms.

1.5 Research Goal and Objectives

The research goal is to design and develop algorithms for finding coherent biclusters from gene expression data using different algorithm design techniques such as constraint based algorithms, greedy algorithm and metaheuristic algorithms. Hence the study is aimed at designing and developing biclustering algorithms. The objectives are:

- Compare the performance of these algorithms with the existing biclustering algorithms
- Validate the results with the biological annotations already available

1.6 Contribution

In this thesis ten algorithms are developed for the identification of coherent biclusters from gene expression data. In all the algorithms, biclusters are identified in two phases. They are seed finding phase and seed growing phase. In the seed finding phase seeds are generated. Seed is a tightly coregulated submatrix of the gene expression data matrix generated by K-Means clustering algorithm. All the algorithms mentioned

in the seed growing phase begin their search from these high quality seeds. More genes and conditions are added to these seeds in the seed growing phase. Each seed is grown separately by adding more genes and conditions. The next element to be selected and added depends on the algorithm used. The following algorithms were developed as part of the research work and they were used in the seed growing phase.

1. Mean Squared Residue Threshold (MSRT) algorithm
2. Mean Squared Residue Difference Threshold (MSRDT) algorithm
3. Iterative Search with incremental MSR Difference Threshold (ISIMSRDT) algorithm
4. Seed Growing using separate constraints (SGSC) algorithm
5. Algorithm based on greedy approach
6. Algorithm based on Greedy Randomized Adaptive Search Procedure (GRASP)
7. Algorithm based on Cardinality based Greedy Randomized Adaptive Search Procedure (CGRASP)
8. Algorithm based on Reactive Greedy Randomized Adaptive Search Procedure (RGRASP)
9. Algorithm based on Binary Particle Swarm Optimization (PSO)
10. Algorithm based on greedy - Binary Particle Swarm Optimization hybrid

These algorithms can be classified into three groups:

- Constraint based
- Greedy
- Metaheuristic algorithms

These algorithms are applied on both Yeast and Human Lymphoma datasets. The results obtained by all these algorithms are represented graphically by using the bicluster plots. The biologically significant biclusters are identified by all these algorithms. The results are compared with some of the already developed biclustering algorithms on the basis of bicluster size and mean squared residue and also the statistical significance. The statistical significance and biological relevance of the biclusters are also validated using gene ontology database. In these methods it is possible to obtain all kinds of biclusters. Some biclusters were obtained, whose row variance is greater than that of any algorithm using MSR, from both Yeast and Lymphoma datasets with the help of algorithms like MSRT and SGSC.

1.7 Layout of the Thesis

The layout of the thesis is as follows:

Chapter 1 is the introduction of the thesis.

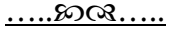
Chapter 2 provides a literature review of the various data mining techniques available for the analysis of gene expression data. A general description of the algorithms developed for the identification of coherent

biclusters, validation of the biclustering results using the biological annotations already available etc are also given in this chapter.

Chapter 3, 4 and 5 explain the algorithms developed as part of the research work. Chapter 3 describes all the constraint based algorithms namely MSRT, MSRDT, ISIMSRDT and SGSC. Chapter 4 describes the Greedy algorithm. Chapter 5 describes the metaheuristic algorithms namely GRASP, CGRASP, RGRASP, Binary PSO and also the Greedy-PSO hybrid. The description of algorithms, time complexity, different biclusters obtained from the datasets, significant biclusters obtained (biological validation), comparison of the algorithms with other biclustering algorithms are also given in the respective chapters.

Chapter 6 gives a performance evaluation of the MSR based algorithms and a consolidation of the research findings.

Chapter 7 contains conclusions and future work.



Chapter 2

Analysis of Gene Expression Data

This chapter provides a literature review of the existing data mining techniques for the analysis of gene expression data such as classification, dimensionality reduction, gene regulatory network analysis, association rule mining, clustering and biclustering. This chapter also gives a general description of the algorithms developed for the identification of coherent biclusters, and describes how their results can be validated using the already available biological annotations.

2.1 Gene Expression Data Analysis

Data mining is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [30, 41, 95]. Data mining techniques can be used for the analysis of Gene Expression data. Gene expression data has been analyzed in gene dimension as well as the condition dimension. There are a number of high-level analysis methods which have the common aim of extracting the biologically relevant patterns and information from the data. Clustering, classification, dimensionality reduction and other types of methods are all frequently applied in gene expression data analysis [3, 40, 46, 58, 89, 105]. This chapter reviews different types of data mining methods that are adopted to extract different types of information from gene expression data including biclustering which is the data mining technique used in this thesis. Moreover, the extracted structure needs validation, for example, while associating the results to prior knowledge which is often stored in large databases.

2.2 Classification

Classification is an important supervised data mining method for the analysis of gene expression data. The application of classification for microarray data include diagnosing cancer type from the expression pattern of a tumor sample, or predicting the biological function of genes based on their expression patterns. The samples are classified based on gene expression patterns into known categories based on morphology, known biological features, clinical outcomes, and so on. For

classification, the classifier is first trained on training samples, and then tested on test samples. Classification algorithms, explicitly or implicitly, identify variables, or functions of variables, that are good predictors of a class. After having been confirmed to have enough correctness, the classifier can classify samples of unknown class label. Classification approaches applied on gene expression data include decision tree [80], KNN [77], SVM [13], and artificial neural network [17]. Artificial neural networks are used for classification problems with more than two classes [68], while support vector machines are binary classifiers. For example they can classify healthy and cancerous tissue [43] or classify genes as belonging to a known functional group [23] or not. Binary classifiers can be extended to handle K classes. ANNs and SVMs are capable of learning non-linear decision functions. In SVMs this is made possible by a kernel transformation of the data. Classification methods like SVM and Neural network are effective in classifying test samples. For gene list based classifiers, the decision function is fixed and predefined. The set of variables on which it operates is learned from the data. For gene list based classifiers, since the genes by far outnumber the samples, introduces some difficulties. For example, the gene list based classifiers [49] classify genes based on the top discriminatory genes. All these approaches have some limitations when applied to gene expression data. A better alternative for gene expression data is the associative classification [18, 103] which makes the decision based on the most significant class association rules. Class association rules, are both informative and easy to understand.

2.3 Dimensionality Reduction

Dimensionality reduction methods select a subset of objects in such a way that important properties of the data are optimally conserved and thus provide a means of representing data in low dimensions. Dimensionality reduction methods are suitable for explorative data analysis. One of the main application of dimensionality reduction is for visualization of patterns in data. In gene expression data analysis, two dimensional or three-dimensional visualizations may be inspected for discovering outliers. Dimensionality reduction can also be used as a means of data quality control. Dimensionality reduction is used as a compressive preprocessing step prior to clustering or classification. This helps to filter out the noise and reduce the computational burden of subsequent methods. Some of the standard methods of dimensionality reduction used for gene expression data are *principal component analysis* and *multidimensional scaling* [6, 8, 54]. These methods are suitable for data patterns which are linear, and are not designed for data when the dependencies between variables are non-linear.

2.3.1 Principal Component Analysis (PCA)

PCA is a mathematical technique to pick out relevant patterns in the data, while reducing the effective dimensionality of gene-expression space without considerable loss of information. PCA is one of the techniques that include factor analysis, which provides a "projection" of complex datasets onto a reduced, easily visualized space. PCA finds those views which separate the data into groups. PCA creates a small number of

summary variables called principal components from a much larger set. These summary variables are used for visualization or for more complex statistical modelling. Creation of components and selection of the most representative (or principal) components are the two aspects of PCA. Components in the PCA are weighted averages of the original variables, which are uncorrelated with each other. Components are created by rotation of the original coordinates. The selection of the most representative (principal) components is based on the fraction of variability. An advantage of PCA is that redundant information (e.g., genes showing similar expression patterns across samples) can be represented by a single variable. A disadvantage is that sometimes the summary variables do not necessarily have a clear biological interpretation. This technique can be applied for both genes and conditions as a means of classification. PCAs are sometimes used to visually identify clusters. This may be successful, but there is, in general no guarantee that the data will cluster along the dimensions identified by the principal component. PCA is a powerful technique for the analysis of gene expression data when combined with other classification technique, such as k-means clustering [79] or self-organizing maps (SOM), which require the user to specify the number of clusters. PCA is widely used for the analysis of gene expression data [7, 50, 82, 108].

2.3.2 Multi-Dimensional Scaling (MDS)

Another technique for dimensionality reduction is Multidimensional scaling [71]. Multidimensional scaling identifies variables that are as consistent as possible with the observed distance matrix. This results in a

graphical representation of the objects as a 2D or 3D figure. User-specified options have an effect on the resulting representation generated by MDS. One of the most common methods for MDS is metric MDS or principal coordinates analysis. MDS finds application in cancer classification using microarrays [19, 69].

2.4 Gene Regulatory Network Analysis

All cells in an organism have the same genomic data. But the proteins synthesized in each cell vary according to cell type, time, and environmental factors. The activity of a cell depends on which genes are expressed, i.e., which genes are turned on, resulting in the active production of their respective proteins. By monitoring the expression levels of all genes within a cell simultaneously, it is possible to find out which genes are up-regulated, down-regulated, or not expressed under a specific condition and can also detect any correlations between the levels of expression of different genes. Using this information, it is possible to interpret the logic of gene regulation in a cell [33]. Genes interact with each other in regulatory networks. Therefore the most biologically authentic representation of the genes is, as a network which describes the functional relations between genes rather than as a number of clusters, or as a cloud of points in Euclidean space. The interactions between genes in the regulatory networks can be modeled in many ways [De Jong, 2002] [32]. They include simple *Boolean Networks* [63, 64, 65] to more complex regulatory networks such as random directed graphs and to detailed models such as *Stochastic Master Equation models* [11].

The study of gene networks is one of the subjects attracting more attention. The simplest approach for the identification of network is clustering the data and searching for regulatory control elements in all co-expressing genes [22, 99]. But the information provided by these approaches is limited to genes that are co-regulated. This will not identify a gene which is regulating another gene. In network inference, a model of the interactions between the genes, is constructed. Different models of gene regulation have been proposed. The simplest genetic regulatory network is the Boolean network. Boolean network was introduced by Kauffman in the late 1960s (Kauffman, 1969) [63]. The network is represented as a directed graph. If $G = (V, F)$ is the graph then V represents elements of the network, and F defines a topology of edges between the nodes and a set of Boolean functions. In the Boolean network each gene is modeled as either ON or OFF. The state of each gene at the next time step is determined by Boolean function of its input at the current time step. Even though the Boolean networks are simple, they are able to provide valuable insights in the behavior of gene interactions [64, 104]. They are used in the analysis of real gene expression data for the identification of drug targets for cancer therapy [55, 96].

Boolean network is useful in gene regulation studies. But the disadvantage is that the gene expression data is not binary but continuous. Moreover, the gene expression data is generally noisy and contain a high level of uncertainty. All these facts led to the proposal of various modifications on the basic Boolean network, such as the Noisy Boolean network [2], the Probabilistic Boolean network [87, 88], and the Hybrid

Boolean network. In the Hybrid Boolean network each gene has a continuously valued internal state, a Boolean external state [47, 48] or asynchronously updated logic with intermediate threshold values [101, 102]

2.5 Time Series Analysis

The goal of time series analysis is to find out genes that show similar trends over time within the same organism or sample type and to discover samples that are differentiated by such patterns. Time series analyses are often performed using regression. In this case time is the primary predictor variable and gene expression is the outcome [35, 92, 110].

2.6 Association Rule Mining

Association rule mining has attracted great interest since a rule provides a concise and insightful description of knowledge. It has already been applied for the analysis of biological data [26, 38, 61]. Powerful computational analysis tools are required to extract the most significant and reliable correlation between genes from high-dimensional gene expression data. Class association rule which is one of the most famous traditional data mining methods is the solution for the above requirements. Each row in the expression data matrix for finding class association rule mining corresponds to a sample or a condition, and each column corresponds to a gene. Conventional association mining methods [59, 81] use the item-wise searching strategy. Some of the current class association rule mining methods also use the same strategy [83]. A substantial amount of research in the field of association rule mining has demonstrated that accurate and inexpensive diagnosis is possible with

class association rules because of their informative nature. A class association rule can be defined as a set of items, or specifically a set of conjunctive gene expression level intervals (*antecedent*) with a single class label(*consequent*). The *general* form of a class association *rule* is: *gene*1[*a*1*; b*1], ...,*genen*[*an; bn*] → *class*, where *genei* is the name of the gene and [*ai; bi*] is its expression interval. For example, *X*95735 *at*[-*α,* 994] →*ALL* is one rule discovered from the gene expression profiles of ALL/AML tissues [105].

The unlabelled association rules can help discover the relationship between different genes and build the gene network [26]. Class association rules can relate gene expressions to their cellular environments or categories indicated by the class. Thus they can build accurate classifiers on gene expression datasets. Some of the association rule mining algorithms find the complete set of association rules satisfying user-specified constraints by discovering frequent (closed) patterns [59, 81].

2.7 Clustering

Clustering is an unsupervised learning technique. Cluster analysis is a fundamental technique in exploratory data analysis and pattern discovery. Cluster analysis is an important technique to partition objects that have many attributes (multi-dimensional data) into meaningful disjoint sub-groups. Clustering process groups together similar objects into clusters. The objects in each cluster are more similar to each other in the values of their attributes, than they are to objects in other groups.

Unlike classification, in cluster analysis the number of clusters is unknown. Clustering needs a similarity function to measure how similar two data points are. Mainly there are two types of clustering, partitional and hierarchical. Hierarchical techniques provide a series of successively nested clusters. Non-hierarchical techniques generally find a single partition, with no nesting. Both are used extensively in microarray analysis. In gene expression data analysis, clustering discovers groups of co-regulated genes or groups of samples.

2.7.1 Hierarchical Clustering

Hierarchical clustering is one among the most widely used technique in the analysis of gene expression data because of its simplicity and ease of visualization [39]. Hierarchical clustering can be classified as agglomerative or divisive. In the agglomerative approach initially all genes are considered as clusters. Then the distance matrix is calculated for all of the genes to be clustered. Two genes with the lowest distance from the distance matrix is selected and combined to form a single cluster. This process in which two selected clusters are merged to produce new clusters is continued until a single hierarchical tree is formed. There are several variations on hierarchical clustering which differ in the rules governing how distances or similarity is measured between clusters as they are constructed. Similarities between two clusters can be defined in a number of ways, such as single linkage, complete linkage and average linkage. In single linkage the largest similarity between any pair of objects in separate clusters is calculated. In complete linkage the smallest similarity

between any pair of objects in separate clusters is calculated. In average linkage, the average similarity between all pairs of objects in separate clusters is calculated. In hierarchical clustering the clustering is visualized as a cluster tree called a *dendrogram*. One problem with hierarchical clustering is that it is difficult to decide which clustering level in the dendrogram to choose. Another disadvantage is that different similarity measures yield very different cluster trees.

2.7.2 K-Means Clustering

K-means clustering [53] is a standard single level clustering algorithm. In K-means clustering, the goal is to break objects into groups that have low variance within clusters and large variance across clusters [46]. K-means clustering is a good alternative to hierarchical methods if there is advanced knowledge about the number of clusters. The K-means method does not have many parameters to assign. Tavazoie et al. uses K-means clustering in gene expression data analysis [99]. K-Means is the simplest clustering algorithm. It is the best known partitional clustering algorithm. The method is called K-means since each of the K clusters is represented by the mean of the objects. It is also called centroid method. Different distance measures like Euclidean, cosine angle distance etc. can be used in K-Means clustering. The K-Means method [51] may be described as follows:

1) Select the number of clusters. Let this number be K.
2) Pick K seeds as centroids of the Kclusters. The seeds may be picked randomly unless the user has some insight into the data.

3) Compute the distance of each object from each of the centroids.

4) Allocate each object to the cluster which is nearest to it based on the distance computed in the previous step.

5) Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.

6) Check if the stopping criterion has been met (e.g. the cluster membership is unchanged). If yes go to step 7. If not, go to step3.

7) [optional] One may decide to stop at this stage or to split a cluster or combine two clusters heuristically until a stopping criterion is met.

2.8 Biclustering

Biclustering is the data mining technique used in this thesis for the analysis of gene expression data. Biclustering is simultaneous clustering of both the rows and columns of a data matrix. Biclustering consists in simultaneous partitioning of the set of samples and the set of their attributes (features) into subsets (classes) [93]. A bicluster of a dataset D is a collection of pairs of gene and condition subsets B = *((*G1,C1*)*, *(*G2,C2*), . . . , (*G*r* ,C*r))* such that the collection *(*G1,G2*, . . . ,*G*r)* forms a partition of the set of genes, and the collection *(*C1,C2*, . . . ,*C*r)* forms a partition of the set of conditions [93]. In short a bicluster is a submatrix B of the gene expression data matrix D and if the size of B is IxJ, then I is a

subset of rows X of D, and J is a subset of the columns Y of D. The rows and columns of the bicluster B need not be contiguous as in the expression matrix D. It is not necessary that the identified submatrices to be disjoint or to cover the entire matrix. Biclustering is also known as co-clustering, bi-dimensional clustering and subspace clustering. Biclustering is a relatively young area, in contrast to its parent discipline, clustering, that has a very long history [98].

2.8.1 The Advantages of Biclustering over Clustering

Clustering is one of the important data mining techniques. However, applying clustering to gene expression data has some disadvantages. Many activation patterns are common to a group of genes only under specific experimental conditions. As per the general understanding of cellular process subsets of genes are co-regulated and co-expressed only under certain experimental conditions, but behave almost independently under other conditions. Discovering such local expression patterns may help to uncover many genetic pathways that are not apparent otherwise. It is therefore highly desirable to develop algorithmic approaches capable of discovering local patterns in gene expression data. Clustering is applied to either the rows or the columns of the data matrix, separately. Biclustering methods, on the other hand, perform clustering in two dimensions simultaneously. That means clustering methods derive a *global model* while biclustering algorithms produce a *local model*. When clustering is applied to gene expression data genes as well as conditions can be clustered. However, each gene in a bicluster is selected using only a subset of the conditions and each condition in a bicluster is selected using

only a subset of the genes. Biclustering thus performs simultaneous clustering of both rows and columns of the gene expression matrix, instead of clustering these two dimensions separately. In short unlike clustering algorithms, biclustering algorithms can identify groups of genes that show similar activity patterns under a specific subset of the experimental conditions. Biclustering is used when one or more of the following situations apply [74]:

1. A single gene may participate in multiple pathways that may or not be co-active under all conditions
2. Only a small set of the genes participates in a cellular process of interest.
3. An interesting cellular process is active only in a subset of the conditions.

2.8.2 Bicluster Types

An interesting criterion for evaluating a biclustering algorithm is the identification of the type of biclusters the algorithm is able to find. There are four major classes of biclusters:

1) Biclusters with constant values.
2) Biclusters with constant values on rows or columns.
3) Biclusters with coherent values.
4) Biclusters with coherent evolutions.

The simplest biclustering algorithms can identify biclusters with constant values. Figure 2.1 (a) gives an example of a constant bicluster.

Figure 2.1(b) is an example of a bicluster with constant rows. The bicluster in Figure 2.1(c) is an example of a bicluster with constant columns. More sophisticated biclustering approaches look for biclusters with coherent values on both rows and columns. Figure 2.1 (d) and (e) are examples of this type of bicluster. The last type of biclustering addresses the problem of finding biclusters with coherent evolutions. In coherent evolutions the elements of the matrix are considered as symbolic values and try to discover subsets of rows and subsets of columns with coherent behaviors without regarding the exact numeric values in the data matrix. Examples of these types of biclusters are given in Figures 2.1 (f) to (i) [74].

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

a) Constant Bicluster

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

b) Constant rows

1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0
1.0	2.0	3.0	4.0

c) Constant Columns

1.0	2.0	5.0	1.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

d) Coherent values – additive model

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

e) Coherent values – multiplicative model

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

f)Overall coherent Evolution

S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1
S1	S1	S1	S1

g) coherent Evolutions on rows

S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4
S1	S2	S3	S4

h) coherent Evolutions on columns

70	13	19	10
29	40	49	35
40	20	27	15
90	15	20	12

i) Coherent Evolutions on columns

Figure 2.1 Different types of biclusters

2.8.3 Biclusters with Coherent Values

Biclusters with coherent values are biologically more relevant than biclusters with constant values. Hence in this work biclusters with coherent values are identified. In this case the problem of biclustering can be formulated as follows: given a data matrix D, find a set of submatrices B1, B2,... Bn which satisfy some homogeneous characteristics or coherence. In order to identify the degree of coherence a measure called mean squared residue score or Hscore was introduced by Cheng and Church [29]. It is the sum of the squared residue score. The residue score of an element *bij* in a submatrix *B* is defined as

$$RS(bij)=bij-bIj-biJ+bIJ$$

Here *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. The residue score of an element *bij* provides the difference between the actual value and its expected value predicted from its row mean, column mean and bicluster mean. The residue of an element reveals its degree of coherence with the other elements of the bicluster it belongs to. Hence

from the value of residue score, the quality of the bicluster can be evaluated by computing the mean squared residue. That is Hscore or mean squared residue score of bicluster *B* is

$$MSR(B) = (\sum i \in I, j \in J (RS(bij))^2) / (|I| * |J|)$$

A submatrix *B* is called a δ bicluster if MSR(*B*)< δ for some δ >0. 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 .The value of δ is taken from Cheng and Church [29] and is calculated from the clustering experiments done by Tavazoie *et al.* [99]. Low MSR value denotes strong coherence in the bicluster. The volume of a bicluster or bicluster size is the product of the number of rows and the number of columns in the bicluster. The biclusters characterized by high values of row variance contains genes that display significant changes in their expression values under different conditions. Cheng and Church used row variance as an accompanying score to find out trivial biclusters. There is no threshold value for row variance in order to consider a bicluster as trivial. Row Variance of the bicluster B can be calculated using the formula

$$RowVar(B) = (\sum i \in I, j \in J (bij - biJ)^2) / (|I| * |J|)$$

The quality of the bicluster is always superior when the volume and row variance of the bicluster are larger, and when its mean squared residue is smaller.

2.8.3.1 Different Types of Biclusters Depending on Coherence and Row Variance

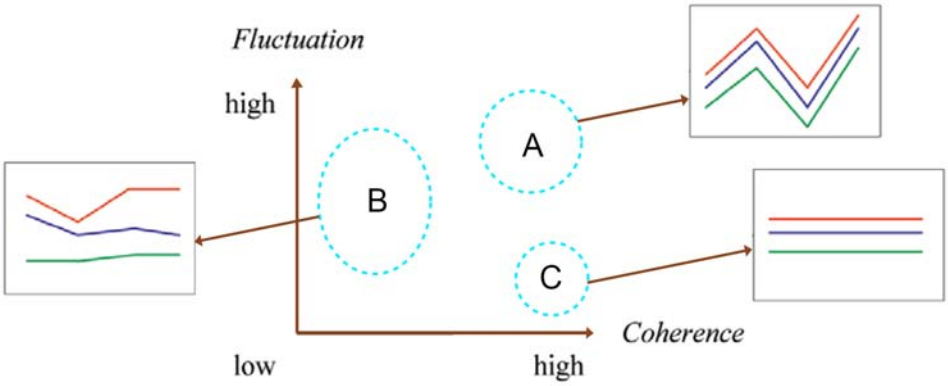


Figure 2.2 Characterization of biclusters based on coherence and row variance

Consider three biclusters A, B and C shown in Figure 2.2. The coherence of biclusters A and C are very high. But the coherence of bicluster C is low. The row variance of bicluster A is high since there is variation in the expression level of the genes. But in C there is no variation in the expression level. In applications like gene coregulation analysis, the biclusters in area A is the most interesting because similar behavior between highly expressed genes is much more important than that between two poorly expressed genes [45]. On the other hand, the flat biclusters in are C are important for applications such as the identification of marker genes. The biclusters in area B are less interesting because they have a lower level of coherence than those in area A or C [94]. Figure 2.2 is reproduced from [94].

2.8.4 Related Work

Various algorithm design techniques are used to address the biclustering problem including Iterative row and column clustering combination, Divide and Conquer, Greedy iterative search, Evolutionary or Metaheuristic algorithms. Iterative row and column clustering is a simpler way to perform biclustering. Here standard clustering methods are applied on the row and column dimensions separately and the result is combined to obtain biclusters. In divide and conquer strategy the problem is divided into small sub-problems, solve the sub-problems separately and then combine the solutions to get the final result. Divide and conquer algorithms are very fast. But the drawback of this approach when solving the biclustering problem is that in divide and conquer strategy as the data is divided, there is a possibility of splitting good biclusters before they can be identified. Greedy iterative search methods are based on the idea of creating biclusters by adding or removing rows or columns from them, using a criterion that maximizes a local gain [74]. They have the potential of being very fast. Metaheuristic algorithms are able to find global optimal solutions.

Computation of biclusters is costly because one will have to consider all the combinations of columns and rows in order to find out all the biclusters. The search space for the biclustering problem is 2^{m+n} where m and n are the number of genes and conditions respectively. Usually m+n is more than 3000. The biclustering problem is NP-hard. In a gene expression data matrix there are a number of biclusters with different

shapes. The biclustering algorithm should be capable of identifying these biclusters. Biclustering was first introduced by Hartigan who called it direct clustering [52]. Hartigan identifies two biclusters at a time. Cheng and Church were the first to apply biclustering to gene expression data [29]. In the approach taken by Cheng and Church, the rows or columns were deleted from the gene expression data matrix in order to find a bicluster. Their algorithm is based on the greedy strategy. Their algorithms are deterministic in the sense that repeated runs of them will not discover different biclusters, unless the discovered ones are masked. So the discovered bicluster is replaced by random values. These random values will interfere with the future discovery of biclusters, especially those that have overlap with the discovered ones. This problem is known as random interference. Yang et al. [106] generalized the model of bicluster proposed by Cheng and Church for incorporating null values and for removing random interference. They developed a probabilistic algorithm FLOC that can discover a set of possibly overlapping biclusters simultaneously. Zhang et al. presented Deterministic Biclustering with Frequent pattern mining (DBF) [109]. In DBF a set of good quality bicluster seeds are generated in the first phase based on frequent pattern mining. Then these biclusters are enlarged by adding more genes or conditions. Sequential Evolutionary Biclustering (SEBI) [36] is based on evolutionary algorithms. The objective of SEBI is to identify biclusters of maximum size, with MSR lower than a given δ, with relatively high row variance and with a low level of overlapping among the biclusters. Biclustering problem is also solved using global optimization techniques

like simulated annealing [25] in which the objective is to identify the bicluster with the maximum volume and low MSR. Tanay et al. [97] developed Statistical-Algorithmic Method for Bicluster analysis (SAMBA), in which statistically significant biclusters were identified using graph theoretic and statistical considerations. They defined a bicluster as a subset of genes that jointly respond across a subset of conditions, where a gene is termed as responding in some condition if its expression level changes significantly at that condition with respect to the normal level. Spectral biclustering approaches use techniques from linear algebra to identify bicluster structures in the gene expression data [70].

Recently biclustering problems are solved using multi-objective optimization methods. When searching for biclusters in microarray data, several objectives like the volume, mean squared residue and row variance are to be optimized simultaneously. Often these objectives are in conflict with each other. In multi-objective optimization problem there are a number of feasible solutions. In the work of Banka and Mitra the Multi Objective Evolutionary Algorithm (MOEA) is used for solving biclustering problem [15]. Here only the bicluster volume and MSR are optimized. Sequential Multi-objective Biclustering (SMOB) [37] also uses Multi-Objective EA for finding biclusters in gene expression data. In the work of Junwan Liu, Zhoujun Lia and Feifei Liu [62] multi-objective PSO is used for the identification of biclusters. Some more well known biclustering techniques are Random-Walk-based Biclustering (RWB) [9], SGAB [20], Order Preserving Submatrix algorithm OPSM) [16], iterative signature algorithm (ISA) [56], BiVisu [100] and Bimax [78]. MOGAB

was developed by malik et.al. Maulik et.al [75] solved biclustering problem using Multi-objective Genetic algorithm. Their objective was to identify coherent and nontrivial biclusters which should have low mean squared residue and high row variance. The Plaid model developed by Lazzeroni and Owen for the analysis of gene expression data uses a statistically inspired modelling approach [72]. Biclustering problem is also solved using GRASP variants [34, 90, 91] to identify biclusters from Yeast dataset. The RGRASP [91] uses this technique for the identification of significant biclusters.

2.8.5 Datasets Used

The algorithms are implemented in Matlab and the datasets used are Yeast and Lymphoma. The pre-processed datasets are downloaded from [107]. Experiments are also conducted on datasets by filtering out genes with small variance across conditions using 'genevarfilter' in Matlab.

2.8.5.1 Yeast Dataset

The Yeast dataset is based on Tavazoie et al. [99]. Yeast dataset consists of 2884 genes and 17 conditions. The values in the expression dataset are integers in the range 0 to 600. Missing values are represented by -1. A sample Yeast dataset is given in Appendix 7.

2.8.5.2 Lymphoma Dataset

Human B-cell Lymphoma expression dataset contains 4026 genes and 96 conditions. The dataset was downloaded from the website for supplementary information for the article by Alizadeh et al. [5]. 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. The datasets are obtained from [107]. In the Lymphoma dataset missing values are replaced by random numbers between -800 and 800 as in [29].

2.8.6 Biological Validation of Biclusters

Once high-level analysis methods have suggested some underlying structure in the data, these results need to be interpreted and validated in terms of biological significance. Prior biological knowledge can be used to evaluate the biological significance of biclusters obtained [97]. If the identified biclusters contain significant proportion of biologically similar genes, then it proves that the biclustering technique produces biologically relevant results. The biological significance can be verified using gene ontology database. In this database 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. If the p-value is smaller, then the match will be better. Yeast genome gene ontology term finder [85] is a database available in the Internet which can be used to evaluate the biological significance of biclusters. 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.

2.8.7 Biological Applications of Biclustering

Biclustering is applied when the data to be analyzed is a real valued matrix. The analysis of gene expression data plays a major role in our understanding of biological processes and systems including gene regulation, development, evolution and disease mechanisms [14]. Biclusters can be used to associate genes with specific clinical classes or for the classification of genes and samples, among other potentially interesting applications. The three main applications of biclustering approaches are identification of coregulated genes, gene functional annotation and sample classification [74]. Biclustering is also used in a number of other biomedical applications. In [73] biclustering was applied to drug activity data to associate common properties of chemical compounds with common groups of their descriptors. Moreover [72] presents the application of biclustering to the nutritional data. In this case each sample is associated with a certain food, while each feature is an attribute of the food. The goal was to form clusters of foods similar with respect to a subset of attributes.

2.9 General Description of all Algorithms Developed in this Thesis

2.9.1 Encoding of Bicluster

Each bicluster is encoded as a binary string of fixed length [28]. The length of the string is the sum of the number of rows and the number of columns of the gene expression data matrix. The first N bits represent genes and the next M bits represent conditions. A bit is set to one when the corresponding gene or condition is included in the bicluster. Otherwise the bit is set to zero. This representation is advantageous for node addition and node deletion.

2.9.2 Seed Generation Using K-Means Clustering Algorithm

A small tightly co-regulated submatrix of the gene expression data matrix with a low mean squared residue score is called the seed of the bicluster. Since the MSR value of the seed is lower than the threshold, there is a possibility of accommodating more genes and conditions within the given MSR threshold. The K-Means clustering algorithm is used for seed finding. The gene expression dataset is partitioned into n gene clusters and m sample clusters. Gene clusters having more than 10 genes are further divided according to the cosine angle distance from the cluster centre. Similarly each sample cluster having more than 5 samples is further divided into sets of 5 samples according to cosine angle distance from the cluster centre. The number of gene clusters having maximum 10 close genes is p and the number of sample clusters having maximum 5 conditions is q. The p gene clusters and q sample clusters are combined to

form p*q submatrices. The MSR value of these submatrices is calculated and those with MSR value below a certain limit are selected as seeds [27].

2.9.2.1 The Advantages of Using Seeds from K-Means

Using seeds from K-Means has some specific advantages.

1) Since the biclustering is a combinatorial optimization problem seed gives a good start and reduces the number of combinations.

2) There are different types of biclusters in a gene expression data. Some of them will be biclusters with very low variance and large volume. Genes with low variation in expression level are useful for marker gene identification. Some biclusters are with small volume and large row variance. In short there are different types of biclusters based on MSR, row variance and volume. When seeds from K-Means are used it is possible to get all types of biclusters.

3) MSR is biased towards biclusters with low row variance. But when seeds from K-Means are used, biclusters of large row variance can be obtained.

4) The problem of random interference can be avoided.

5) Some of the seeds will help the identification of the biclusters which cannot be identified by any other algorithm using MSR. After doing experiments with all these algorithms especially the MSRT and SGSC it is found that some conditions are

getting eliminated not because of the lack of coherence in the expression level but because of the significant increase in the expression level. Such conditions will increase the MSR above the threshold. Hence such conditions will get eliminated in all MSR based algorithms. With the seeds from K-Means algorithms such as SGSC or MSRT can identify such biclusters.

6) Seeds from K-Means help the identification of large number of biclusters. This eliminates the limitation of number of biclusters that can be identified by some of the algorithms.

2.9.3 Different Algorithms used in the Seed Growing Phase

More genes and conditions are added to the seed using different seed growing algorithms which are developed as part of this thesis work. Ten different algorithms are used for seed growing. Out of this 4 algorithms use different constraints. One uses the greedy approach. Other methods use metahueristic approaches GRASP and Particle Swarm Optimization (PSO). The Mean Squared Residue Threshold (MSRT) algorithm uses the only constraint mean squared residue threshold. Since biclustering is an optimization problem which is trying to optimize the MSR, Mean Squared Residue Difference Threshold (MSRDT) algorithm uses one more constraint namely the MSR difference threshold. In Iterative Search with Incremental MSR difference threshold (ISIMSRDT) algorithm, the MSR difference threshold value is incremented iteratively. While conducting these experiments it is found that the incremental

increase in genes is low, whereas the incremental increase in conditions are high. Hence an algorithm called SGSC which uses Separate Constraints for Genes and Conditions for finding biclusters from gene expression data is developed. In greedy approach the node with minimum incremental increase in MSR is selected for enlarging the seeds. Since greedy approaches have local minima problem metaheuristic methods like Greedy Randomized Adaptive Search Procedure (GRASP) is also used in the seed growing phase. The different variants of GRASP like basic GRASP, cardinality based GRASP and Reactive GRASP are used for finding biclusters. These three methods differ in the way the restricted candidate list is implemented. Particle swarm Optimization (PSO) which is an evolutionary computation based technique is also used for enlarging the seeds. One more approach which is a hybrid of greedy and PSO is also used for the identification of biclusters.

2.10 Summary

This chapter provides a literature survey of the various existing data mining techniques used for the analysis of gene expression data which includes classification, dimensionality reduction, clustering and biclustering etc. This thesis is concerned with the development of algorithms for the identification of coherent biclusters from gene expression data. A general description of the biclustering algorithms developed in this thesis is also given in this chapter.

Chapter 3

Constraint Based Algorithms

This chapter describes all the constraint based algorithms developed in this work for finding biclusters from gene expression data. A constraint is a condition which must be satisfied by the solution to an optimization problem. The constraint based algorithms are MSRT, MSRDT, ISIMSRDT and SGSC. These algorithms are used for enlarging the seeds obtained by K-Means clustering algorithm. In all these algorithms node addition follows node deletion if necessary. The condition in which the added node is deleted depends on the constraints used by the algorithm. The nodes are added sequentially. The description of the algorithms, Time complexity, different biclusters obtained from Yeast and Lymphoma datasets, significant biclusters obtained (biological validation), and the comparison of the algorithms with other biclustering algorithms are also given in this chapter. A comparison of all the constraint based algorithms is also given.

3.1 MSRT Algorithm

Mean Squared Residue (MSR) is used as the similarity measure to evaluate the coherence of the biclusters. There is a threshold value for the mean squared residue denoted by δ. This value depends on the dataset. The value of δ for the Yeast dataset is 300 and for the Lymphoma dataset, it is 1200. This algorithm is making use of MSR threshold value as the sole constraint for the identification of biclusters. Hence this algorithm is named MSRT algorithm.

In the MSRT algorithm genes or conditions can be added to the given seed one at a time. In this algorithm in order to enlarge the seeds, the conditions are searched first followed by the genes. Many factors are observed when a gene or condition is added to a seed for generating the final bicluster. After adding a gene or condition, the MSR value of the resulting bicluster reduces or increases. The variation in MSR caused by some of the genes or conditions will be very high. This algorithm is developed with the assumption that those genes or conditions having no coherence with the elements of the bicluster will create a large variation in MSR value when added to the bicluster which will be greater than the MSR threshold. Thus after adding one gene or condition the MSR value of the resulting submatrix is calculated in order to verify whether it exceeds the given MSR threshold. If it exceeds the given MSR threshold, it is removed from the bicluster. This process is continued till the last gene or condition is verified for the inclusion in the bicluster. This algorithm is deterministic in the sense that for a given threshold value of

MSR and for a given seed the execution of the algorithm will produce the same result. A pseudo code description of the algorithm is given below.

```
Algorithm MSRthreshold(seed, δ)
// δ denotes the MSR threshold
bicluster := seed;    j := 1;
While (j <= total _no_conditions)
if condition[ j] is not included in the bicluster
Add all elements of condition[j] corresponding to genes already included to the bicluster
calculate MSR
if (MSR> δ) remove elements of condition[ j] from the bicluster  and restore previous MSR value
endif
endif
j:= j+1 end(while)
i=1;
While (i <= total _no_ genes)
If gene[i] is not included in the bicluster
Add all elements of gene[i] corresponding to conditions already included to the bicluster;     calculate MSR
if MSR> δ
remove elements of  gene[i] from the bicluster restore the previous MSR value
endif
endif
i:= i+1
end(while)
return bicluster
end(MSRthreshold)
```

3.1.1 Time Complexity of the MSRT Algorithm

The basic operation for the identification of biclusters is the calculation of mean squared residue of a submatrix. Time complexity for calculating MSR is O(mn) where m and n are the number of genes and conditions respectively. In order to include a single gene or condition, MSR value is calculated once. There are m+n genes and conditions. Hence this calculation is performed at most m+n times. That means the worst case time complexity of the algorithm is O((m+n)mn).

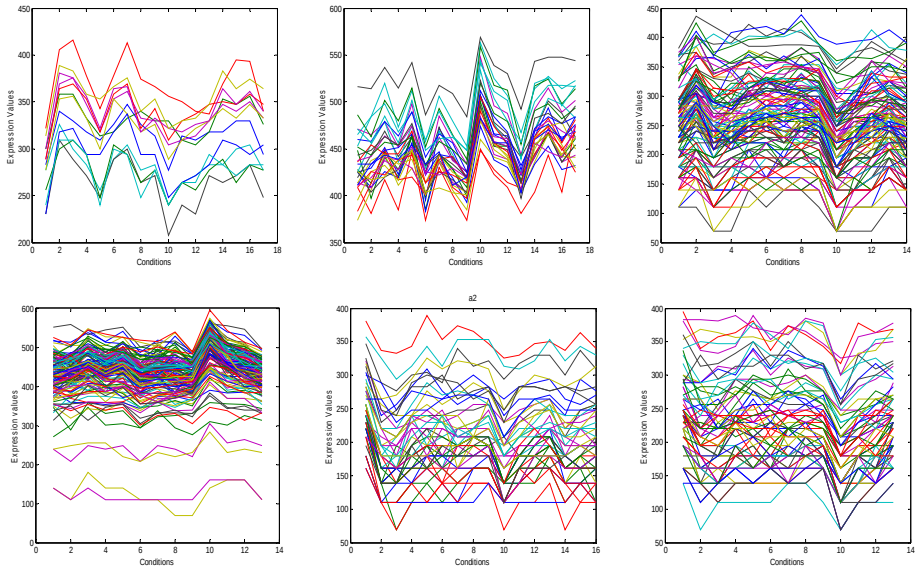
MSRT algorithm is very fast compared to evolutionary or metahueristic algorithms. The main operation for finding bicluster is the calculation of the MSR value of a submatrix. In this algorithm the number of submatrices whose MSR is to be calculated is at most m+n where m and n are the number of genes and conditions respectively. Usually m+n will be less than 4200 (total number of genes and conditions for the Lymphoma dataset which is the largest in this case). In the case of evolutionary algorithms the number of submatrices whose MSR is to be calculated is p*i where p is the number of populations and i is the number of iterations. For SEBI [36] and SMOB [37] the value of p*i is 20000.

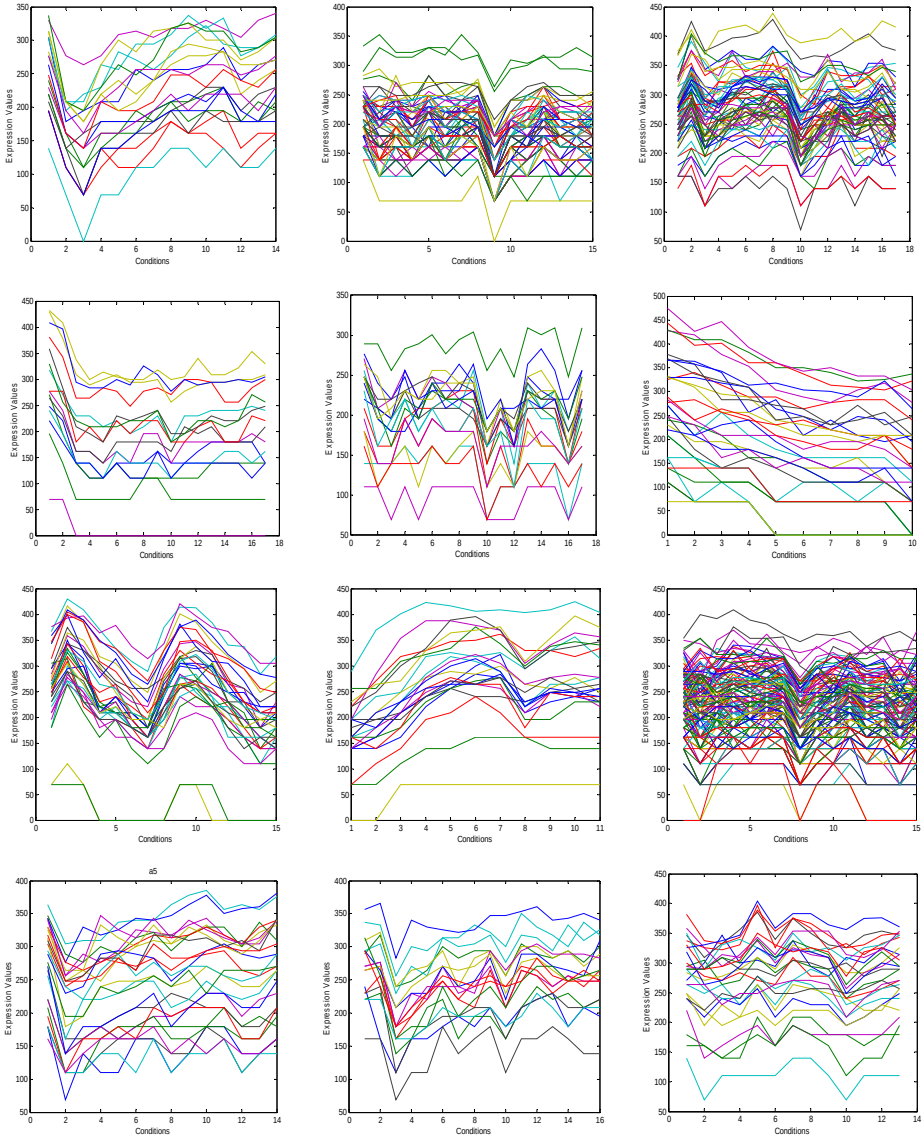
3.1.2 Experimental Results

3.1.2.1 Bicluster Plots for Yeast Dataset

In Figure 3.1 only twenty seven biclusters with different shapes found by the algorithm are shown. From the bicluster plots it is clear that

highly coherent biclusters are obtained using this method. When this algorithm is used, some of the seeds produce biclusters with row variance above 2000. In SEBI the attempt was to identify biclusters with high row variance by adjusting the fitness function. The minimum value of row variance they obtained for the biclusters in Yeast dataset was 317.23. In this study, all biclusters obtained are with row variance above 300. Biclusters with all 17 conditions are obtained using this method even though only seven such biclusters are given in the Figure 3.1. Experiments are conducted by setting the value of MSR threshold as 100, 200, and 300. Even though the MSR threshold value for Yeast dataset is 300, biclusters with low value of MSR are assumed to be more coherent. Hence lower values like 100 and 200 are also used.




Expression Values
Conditions
a5

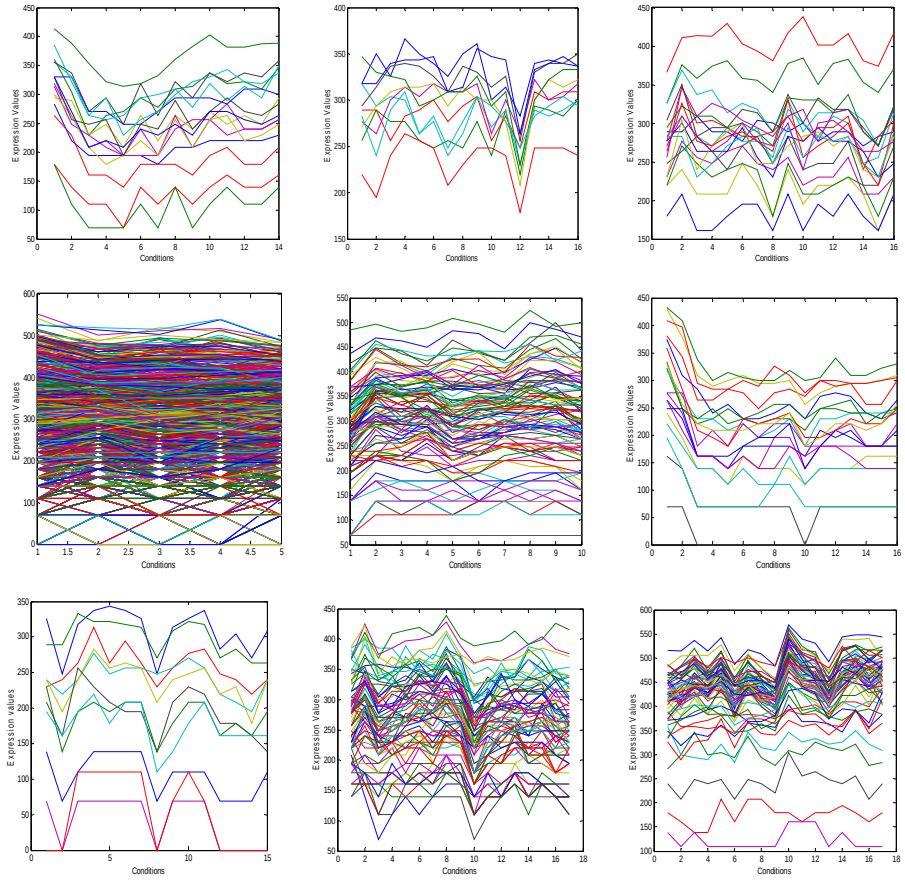


Figure 3.1 Twenty seven biclusters found for the Yeast dataset. Bicluster labels are (ya2), (yb2), (yc2), (yd2), (ye2), (yf2), (yg2), (yh2), (yi2), (yj2), (yk2), (yl2), (ym2), (yn2), (yo2), (yp2), (yq2), (yr2), (ys2), (yt2), (yu2), (yv2), (yw2), (yx2), (yy2), (yz2) and (ya12) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about the biclusters can be obtained from Table 3.1 using the bicluster label.

Table 3.1

Information about Biclusters of Figure 3.1

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
Ya2	13	17	221	99.41	505.91
Yb2	29	17	493	99.89	625.51
Yc2	114	14	1596	199.52	508.76
Yd2	124	13	1612	198.94	601.00
Ye2	49	16	784	199.46	513.27
Yf2	67	13	871	199.05	490.14
Yg2	20	14	280	198.00	1174.10
Yh2	69	15	1035	199.35	578.91
Yi2	67	17	1139	199.95	98.42
Yj2	16	17	272	197.21	115.10
Yk2	20	17	340	199.51	691.37
Yl2	31	10	310	292.16	2052.10
Ym2	33	15	495	299.26	2134.30
Yn2	22	11	242	297.63	1816.20
Yo2	137	15	2055	299.89	529.95
Yp2	26	14	364	199.03	611.65
Yq2	18	16	288	197.98	740.61
Yr2	26	13	338	199.13	378.97
Ys2	16	14	224	196.98	958.01
Yt2	11	16	176	194.38	501.86
Yu2	19	16	304	198.24	430.72
Yv2	1615	05	8075	299.71	308.95
Yw2	96	10	960	198.85	367.38
Yx2	20	16	320	197.57	1058.30
Yy2	11	15	165	273.63	958.06
Yz2	75	17	1275	199.95	459.01
ya12	57	17	969	199.09	618.64

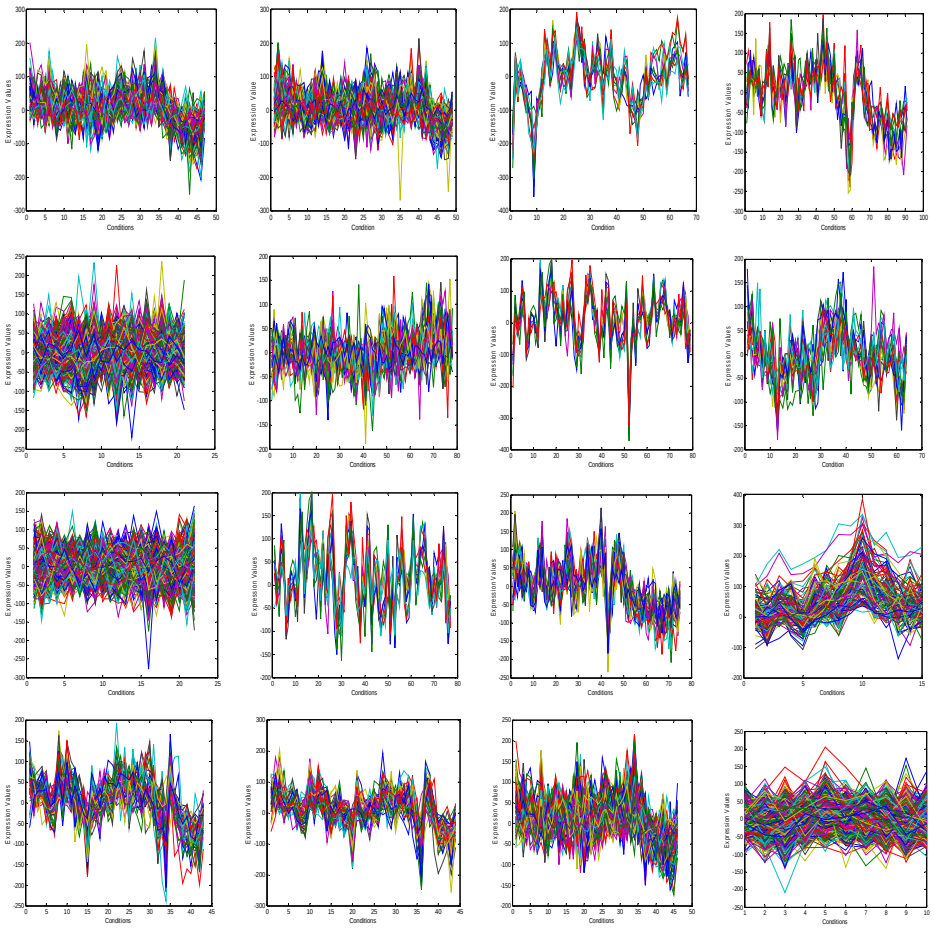
In the above table the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and the number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared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

Biclusters (ya2) and (yb2) are having very low value for MSR. Biclusters (yl2) and (ym2) are with row variance above 2000. In bicluster (yb2) the expression value of all the genes increases in unison under the tenth condition. A bicluster similar to (yb2) is obtained in SMOB but the number of genes and MSR value is 19 and 202.18 respectively. But for bicluster (yb2) the number of genes and MSR value is 29 and 99.8897 respectively. That means in bicluster (yb2) there are more number of genes and it is more tightly coregulated compared to the similar bicluster of SMOB. Shifting and scaling patterns [10] are clearly visible in biclusters (yd2), (yg2), and (yn2). In bicluster (yh2) the up-regulation and down- regulation in the genes are very small but frequent. In the biclusters (yd2) and (ym2) there are 3 and 2 sets of genes respectively. Biclusters with large number of genes having very few conditions (Yv2) are also obtained using this method.

3.1.2.2 Bicluster Plots for Human Lymphoma Dataset

In Figure 3.2 twenty eight biclusters found by the algorithm for the Lymphoma dataset are shown. The genes in the bicluster present a similar behaviour under a set of conditions. Biclusters like (la2), (lb2),

and (lb12) are having very large volume. Bicluster (id2) contains the maximum number of conditions obtained in this method i.e. 91. Bicluster (ly2) is having row variance above 9000. As Federico Divina and Jesus S. Aguilar-Ruize have observed [37] there is no shifting and scaling patterns in the biclusters of Lymphoma dataset. But local shifting patterns are observed in some biclusters.



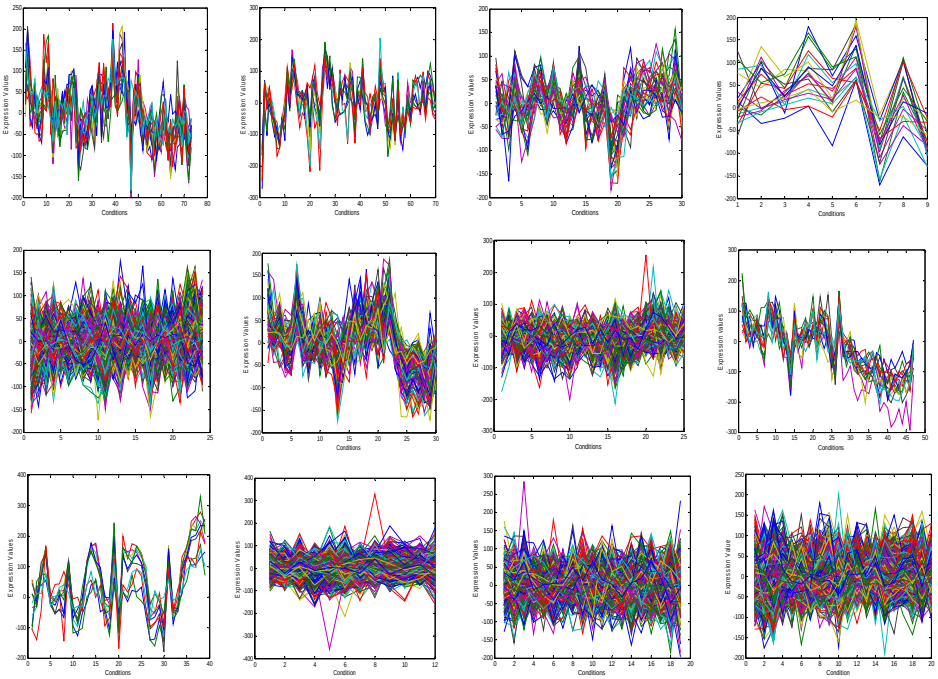


Figure 3.2 Twenty eight biclusters found for the Lymphoma dataset. Bicluster labels are (la2), (lb2), (lc2), (ld2), (le2), (lf2), (lg2), (lh2), (li2), (lj2), (lk2), (ll2), (lm2), (ln2) ,(lo2), (lp2), (lq2), (lr2), (ls2), (lt2), (lu2), (lv2), (lw2), (lx2), (ly2), (lz2), (la12) and (lb12) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 3.2 using bicluster label.

In the Table 3.2 given below, the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and the number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residues of the bicluster. The last column contains the row variance of the bicluster. Biclusters lp2, lu2, lz2 and lb12 are having very high volume. But the row variance is not

very high. More biclusters similar to lp2, lu2, lz2 and lb12 are also obtained from this dataset.

Table 3.2

Information about Biclusters of Figure 3. 2

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
la2	117	47	5499	1194.2	2173.2
lb2	164	49	8036	1188.7	1691.3
lc2	11	67	737	1186.7	6222.4
ld2	10	91	910	1190.5	5308.5
le2	890	21	18690	1198.6	1229.6
lf2	29	78	2262	1185.5	1557.3
lg2	10	79	790	1158.2	6100.0
lh2	18	64	1152	1191.0	2657.3
li2	677	22	14894	1199.2	1230.4
lj2	11	77	847	1189.5	4431.0
lk2	29	75	2175	1189.1	4043.5
ll2	120	15	1800	1196.5	2697.5
lm2	50	43	2150	1196.0	3180.3
ln2	52	44	2288	1199.5	3179.7
lo2	147	46	6762	1195.8	2097.6
lp2	702	10	7020	1199.8	1249.8
lq2	18	73	1314	1197.4	3907.1
lr2	10	70	700	1191.6	5122.2
ls2	33	30	990	1200.0	2258.1
lt2	20	9	180	1194.0	4786.2
lu2	614	24	14736	1197.7	1284.3
lv2	97	30	2910	1196.7	3077.4
lw2	338	25	8450	1198.6	1318.6
lx2	18	47	846	1197.2	7061.6
ly2	11	39	429	1199.2	9009.0
lz2	1311	12	15732	1199.0	1244.7
la12	779	19	14801	1197.7	1214.3
lb12	1136	20	22720	1194.0	1225.3

3.1.3 Advantages of MSRT Algorithm

As no other constraint is used for the identification of biclusters except MSR threshold, different seeds will result in different biclusters with a few exceptions. It is an advantage that the only one parameter required by the algorithm is the MSR threshold. It is noticed that some conditions which make significant change in the expression level is added to the bicluster, the MSR value will increase. Biclustering algorithms trying to minimize MSR will not identify such conditions which are relevant biologically. In this algorithm maximum possible variation is allowed for MSR. Hence it is possible to identify conditions with significant change as well as some of the shifting and scaling patterns [10] which make significant change in MSR. With the help of this algorithm some biclusters with very high row variance are identified from both Yeast and Lymphoma datasets (which are given in chapter 6).

3.1.4. Details of Significant Biclusters obtained by MSRT Algorithm

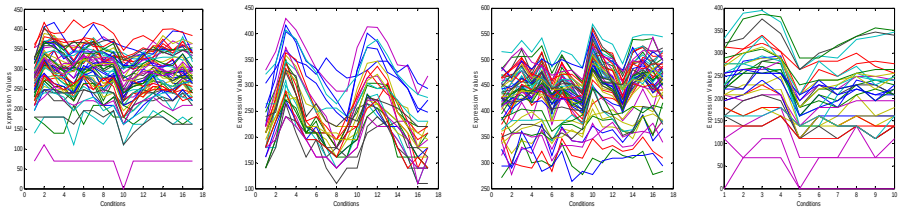


Figure 3.3 Four significant biclusters obtained by the MSRT algorithm on Yeast dataset. The bicluster labels are s21, s22, s23 and s24. The details about biclusters can be obtained from Table 3.3 using the bicluster label.

Table 3.3

Information about Biclusters of Figure 3.3

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
S21	6 1	17	198.9467	469.4058
S22	28	17	299.8488	1937.5000
S23	56	17	199.7856	587.8461
S24	34	10	277.0816	991.0000

Biological relevance of biclusters obtained using MSRT algorithm is verified using the four biclusters shown in Figure 3.3. GO annotation database [36] is used to verify the biological significance of biclusters. In the first bicluster s21 selected for testing the biological significance there are 61 genes. They are YAL007C, YAL011W, YAL035W, YBL024W, YBL083C, YCL031C, YCR059C, YCR087W, YDL008W, YDL150W, YDL153C, YDL166C, YDL167C, YDL231C, YDR017C, YDR057W, YDR060W, YDR083W, YDR120C, YDR121W,YDR170CYDR172W, YDR211W, YDR234W, YDR235W, YDR262W, YDR289C, YDR299WYDR312W, YDR324C, YDR339C, YDR352W, YDR361C, YDR365C, YDR392W, YDR444W, YDR478W, YDR518W, YGL214W, YGR042W, YGR200C, YGR216CYKR060W, YLL008W, YLR146C, YLR222C, YML066C, YNL132W, YNL199C, YNR003C, YOL080C, YOL124C, YOL140W, YOR061W, YOR098C, YOR145C, YOR252W, YOR272W, YPL126W, YPR053C, YPR112C.

In the second bicluster s22 there are 28 genes. They are YAL023C, YAR007C, YAR008W, YBL035C, YBR088C, YBR089W, YCR065W, YDL003W, YDL018C, YDL164C, YDR097C, YFL008W, YGR152C, YHR154W, YJL181W, YKL042W, YKL113C, YLL022C, YLR103C, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL312W, YOL090W, YOR074C, YPL208W.

In the third bicluster s23 there are 56 genes. They are YAL003W, YAL007C, YAL030W, YAL038W, YAR009C, YBL030C, YBL072C, YBL077W,

YBL092W, YBR009C, YBR031W, YBR035C, YBR048W, YBR084C-A, YBR106W, YBR111C, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL219W, YDL228C, YDL229W, YDR012W, YDR025W, YDR035W, YDR050C,YDR064W, YDR133C, YDR134C, YDR382W, YDR385W, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER117W, YGL102C, YMR048W, YNL067W, YOL127W, YOR234C,YOR312C, YPL037C, YPR102C.

In the fourth bicluster s24 there are 34 genes namely YBR038W, YBR138C, YCL012W, YGR108W, YHR151C, YIL106W, YJR092W, YKL129C, YKR021W, YKR056W, YLR190W, YLR353W, YLR453C, YML033W, YML034W, YML119W, YMR001C, YMR032W, YMR291W, YNL171C, YNL172W, YOL130W, YOR152C, YOR160W, YOR206W, YOR365C, YPL148C, YPL150W, YPL183C, YPL242C, YPL248C, YPR003C, YPR007C, YPR119W.

The Table 3.4 given below shows the significant GO terms used to describe genes of the biclusters of Figure 3.3 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 3.4 the first entry of the second column with the title process contains the term ribosome biogenesis (22, 8.41e-11) which means that 22out of the 61 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 8.41e -11. Second entry indicates that 22 out of 61 genes are involved in ribonucleoprotein complex biogenesis. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the MSRT algorithm used here is capable of identifying biologically significant biclusters from different GO categories.

Table 3.4

Significant Shared GO Terms (Process, Function, Component) of Biclusters Shown in Figure 3.3

Bicluster	Process	Function	Component
S21	Ribosome biogenesis (22, 8.41e-11) Ribonucleo-protein complex biogenesis (22,1.47e-09)cellular component biogenesis at cellular level (23, 1.01e-08) Gene expression (30, 0.00053)	27 out of 61 input genes are directly annotated to root term 'molecular function unknown':	Nucleolus (19, 2.91e-11) Preribosome (15, 8.40e-10) 90spreribosome (12, 7.50e-09) Nucleus (36, 0.00020)
S22	DNA repair (16, 4.82e-13) response to DNA damage stimulus (16, 5.57e-12) DNA metabolic process (17, 4.37e-11) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (21, 4.30e-06)	Double-stranded DNA binding (5,4. 13e-05) structure-specific DNA binding (5, 0.00103) DNA secondary structure binding (3, 0.00104) guanine/thymine mispair (2, 0.00335)	Chromosome (14, 2.01e-08) replication fork (8, 1.38e-07) chromosomal part (12, 1.53e-06) nucleus(22, 9.64e-06)
S23	Translation (34, 7.82e-25) cellular protein metabolic process (36, 3.25e-12) protein metabolic process (36, 8.24e-12) cellular macromolecule biosynthetic process (35, 5.82e-10) metabolic process (45, 0.00045)	Structural constituent of ribosome (28, 9.79e-24) structural molecule activity (28, 2.73e-18) translation elongation factor activity (4, 0.00015)	Cytosolic ribosome (29, 1.55e-26) cytosolic part (29, 2.95e-24) Ribosome (32, 8.24e-24) cytosolic large ribosomal subunit (18, 2.09e-17) cytoplasmic part (42, 2.50e-06)
S24	Cytokinesis (7, 0.00130) positive regulation of spindle pole body separation (3, 0.00195) cell cycle process (12, 0.00252) cell cycle (12, 0.00383) regulation of spindle pole body separation (3, 0.00387)	13 out of 34 input genes are directly annotated to root term 'molecular function unknown':	cellular bud (10, 3.48e-06) cellular bud neck(9, 3.81e-06) site of polarized growth(10, 1.63e-05) cellular bud neck contractile ring (4, 5.04e-05)

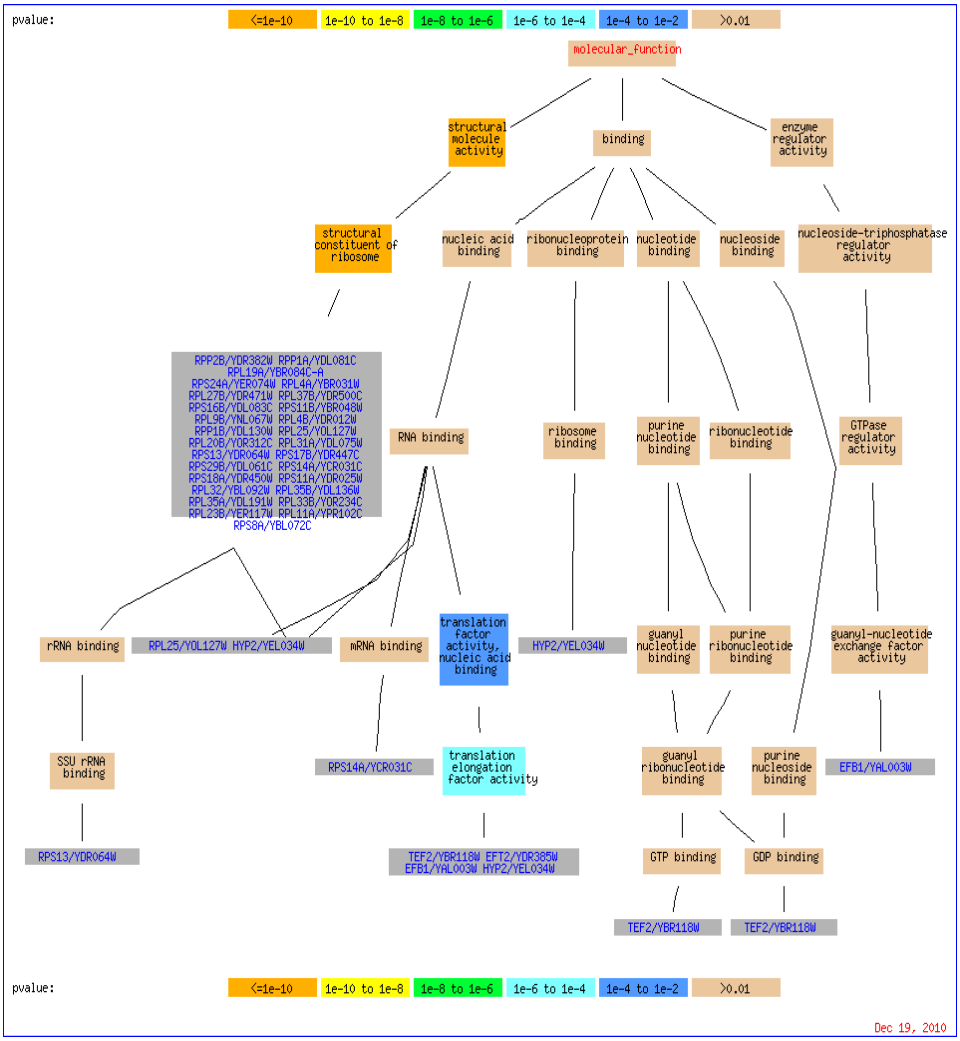


Figure 3.4 Sample of genes for the bicluster s23, with corresponding GO terms and their parents for function ontology

Figure 3.4 shows the significant GO terms for the set of genes in bicluster s23 along with their p values. It shows the branching of a generalized molecular function into sub-functions like structural molecule activity, binding and enzyme regular activity. These activities are

clustered using genes to produce the final result. Figure 3.4 is obtained when gene ontology database is searched by entering the names of genes and by selecting function ontology.

3.1.5 Comparison with other Biclustering Algorithms

3.1.5.1 Comparison based on Statistical and Biological Significance.

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. If the p-value is smaller, then the match will be better. In Table 3.5 the GO terms along with their p-values and percentage of genes associated with the GO terms in the bicluster for the MSRT algorithm is compared with that of MOGAB [75], SGAB [20], CC [29], RWB [9], Bimax [78], OPSM [16], ISA [56] and BiVisu [100]. This table is taken from [75]. From the Table 3.5 it is clear that in terms of the best p-value obtained by a bicluster which is used to denote statistical significance, MSRT algorithm is better than RWB, Bimax, OPSM and Bivisu. The percentage of genes involved in the first GO term is greater than that of RWB, OPSM and Bivisu. For the second GO term the p-value of MSRT algorithm is better than that of all the other algorithms except MOGAB and SGAB. The percentage of genes involved is greater than that of all the other algorithms. For the third GO term the p-value and the percentage of genes is greater than that of all the other algorithms except MOGAB. For the fourth GO term the p-value is better than that of all the other algorithms except MOGAB. But the percentage of genes involved is better than all the other methods. For the fifth GO term the p-value and the percentage of genes involved is better than all the other methods.

Table 3.5
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by MSRT and other Algorithms for Yeast Data

Terms	MSRT	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 51.8% 1.55e-26	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	translation 60.7% 7.82e-25	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Cytosolic Part 51.8% 2.95e-24	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 57.1% 8.24e -24	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Structural constituent of ribosome 50% 9.79e -24	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

3.1.5.2 Comparison based on Bicluster Size and MSR

The Table 3.6 given below provides a comparative summarization of the results of Yeast dataset involving the performance of related algorithms in terms of the average number of genes, conditions and the MSR value of the bicluster. The performance of MSRT algorithm in comparison with that of Cheng and Church's (CC) [29], FLOC by Yang et al. [106], DBF [109], SEBI [36] and SMOB [37] for the Yeast dataset are given. In the MSRT algorithm presented here the average mean squared residue is lower than that of CC, SEBI and SMOB. The average number of genes is greater than that of all other algorithms except CC, FLOC and DBF and average number of conditions is better than that of all other algorithms except SEBI and SMOB. The MSRT algorithm has highest value in the case of largest bicluster size compared to all other methods.

Table 3.6

Performance Comparison between MSRT and other Algorithms for the Yeast Dataset

Algorithm	AMR	ANG	ANC	AV	LB
MSRT	199.78	94.75	14.75	1422.87	8075
SEBI	205.18	13.61	15.25	209.92	1394
SMOB	206.17	27.28	15.46	453.48	697
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

AMR is Average mean squared Residue. ANG is Average Number of Genes. ANC is Average Number of Conditions. AV is Average Volume. LB is Largest Bicluster size. As 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.

Table 3.7 gives performance comparison for Human B-cell Lymphoma dataset. Value of δ is set to 1200 for Lymphoma dataset. In this dataset the average number of genes and average volume of the biclusters obtained are far better than that of SEBI and SMOB. Average number of conditions is greater than CC and SEBI.

Table 3.7

Performance Comparison between MSRT and other Algorithms for Human Lymphoma Dataset

Algorithm	AMR	ANG	ANC	AV
MSRT	1192.43	741.10	38.50	14455.30
CC	850.04	269.22	24.50	4595.98
SEBI	1028.84	14.07	43.57	615.84
SMOB	1019.16	11.60	78.47	709.13

AMR is the Average mean squared Residue. ANG is Average Number of Genes. ANC is the Average Number of Conditions. AV is Average Volume. As clear from the above table the average mean squared residue, the average number of genes and conditions and average volume

and are compared for various algorithms. For Lymphoma dataset AGN and AV are better than that of all other algorithms.

In multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 11 in Yeast dataset and 40 in Human B-cell Lymphoma dataset. But in MSRT algorithm there are biclusters with all 17 conditions for the Yeast dataset and 91 conditions for the Lymphoma datasets respectively. For the Yeast dataset the maximum number of genes obtained for this algorithm in all the 17 conditions is 117 with MSR value 199.9365. The maximum available in all the literature published so far is in multi-objective PSO [62]. They obtained 141 genes for 17 conditions with MSR value 203.25. Moreover as the MSRT algorithm uses simple sequential search rather than stochastic search the computation time required is very less compared to all the metahueristic and evolutionary algorithms.

Some of the biclusters are with high row variance (more than 2000 for the Yeast dataset and more than 9000 for Lymphoma dataset) even though no specific measures are taken to get biclusters of high row variance. A bicluster with 91 conditions is obtained for Lymphoma dataset. The row variance of the bicluster is 5308.5. This bicluster is shown in Figure 3.2 with label (ld2). This method is especially suitable for Lymphoma dataset for obtaining biclusters with large size. Even though the method used here is not multiobjective, the results obtained are better than such algorithms. This is faster than metaheuristic algorithms.

As no other method is used for reducing the MSR except MSR threshold, different seeds will result in different biclusters with a few exceptions.

3.2 MSRDT Algorithm

In this section, a novel algorithm for finding biclusters from gene expression data is described. This algorithm is developed using the newly introduced concept of MSR difference threshold. MSR difference threshold denotes the maximum variation that can be allowed for the MSR value when a gene or condition is added and still the added condition or gene remains coherent. In this algorithm node addition follows node deletion if necessary. In MSRT algorithm mentioned in the previous section the added node (gene or condition) is removed if the MSR value of the resulting bicluster exceeds the MSR threshold. The node thus added may not be optimal in terms of MSR value. In the case of biclustering problem the main objective is to reduce the MSR value of the bicluster. So the previous algorithm is modified by incorporating one more constraint i.e. the MSR difference threshold. Moreover when MSR threshold is used as the only constraint, the variation allowed for the MSR value goes on changing. But when the MSR difference threshold is applied as additional constraint, the variation allowed remains fixed. So in this algorithm before adding a node, the MSR X of the bicluster is calculated. After adding the node, again the MSR Y is calculated. The added node is deleted if Y minus X is greater than MSR difference threshold or if Y is greater than MSR threshold which depends on the dataset. MSR difference of a gene or condition is the incremental increase

in MSR after adding the same to the bicluster. It is found that the MSR difference threshold is different for gene list and condition list and it depends on the dataset also. Proper values should be identified through experimentation in order to obtain biclusters of high quality. The results obtained on Yeast and Lymphoma datasets clearly indicate that this algorithm is better than many of the existing biclustering algorithms.

It is observed that if MSR difference threshold for condition list is set to 30 it is possible to get biclusters with all 17 conditions for the Yeast dataset. For gene list the MSR difference threshold is set to 10. By properly adjusting the MSR difference threshold biclusters of high quality can be obtained. While experimenting it is found that reducing the MSR difference threshold for condition list eliminates the conditions which make significant change in the expression level from the bicluster, where as reducing the MSR difference threshold for gene list increases coherence. Hence difference threshold for conditions should be large and difference threshold for genes should be small (except for scaling patterns). In the case of MSRT algorithm, the added node is removed only when the MSR of the bicluster exceeds δ. But in the case of MSRDT algorithm an element which causes an incremental increase in MSR above MSR difference threshold is also removed from the bicluster. Hence this method can produce better biclusters compared with MSRT in terms of MSR value. A pseudo code description of the algorithm is given below.

```
Algorithm MSRdifferncethreshold(seed, δ, msrdiffgenethresh, msrdiffcondthresh )
bicluster := seed
previous=MSR(seed)
j:= 1;
While (j <= total _no_conditions)
If   condition[ j]  is not included in bicluster
Changed=1;
Add all elements of condition[ j]  corresponding to genes  already included to bicluster
present= MSR(bicluster)
if (present> δ) or (present-previous)>msrdiffcondthresh
remove elements of  condition[ j]  from bicluster
changed=0;
endif
if changed==1
previous=present
endif
endif
j:= j+1
end(while)
i := 1;
prev=MSR(bicluster)
While (i <= total _no_ genes)
If  gene[i]  is not included in bicluster
Changed=1;
Add all elements of gene[i] corresponding to conditions already included to bicluster
```

```
 present= MSR(bicluster)
        if (present> δ) or (present-previous)>msrdiffgenethresh
            remove elements of  gene[i] from bicluster
           changed=0
       endif
  if changed==1
     previous=present
 endif
endif
i:= i+1
end(while)
return bicluster
end(MSRdifferencethreshold)
```

3.2.1 Time Complexity of the MSRDT Algorithm

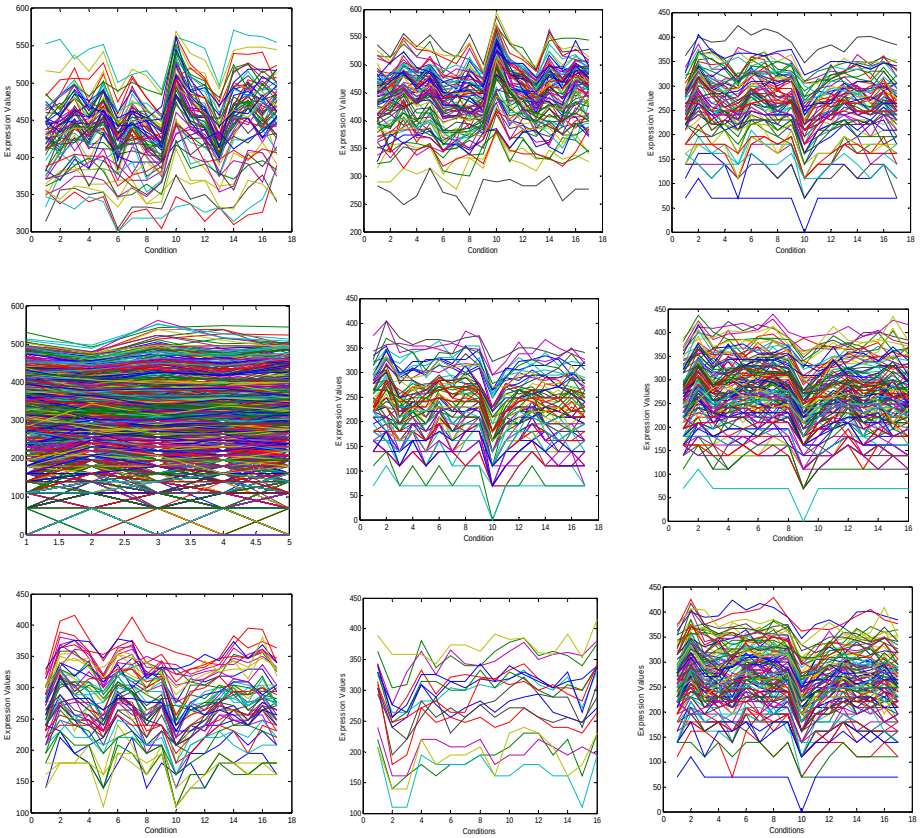
The basic operation for the identification of biclusters is the calculation of mean squared residue of a submatrix. Time complexity for calculating MSR is O(mn). In order to include a gene or condition MSR value is calculated once. There are m+n genes and conditions. Hence this calculation is performed atmost m+n times. That means the worst case time complexity of the algorithm is O((m+n)mn) where m and n are the number of genes and conditions respectively.

3.2.2 Experimental Results

3.2.2.1 Bicluster Plots for Yeast Dataset

Twenty one biclusters obtained by the algorithm on Yeast dataset are given below. Eight out of the twenty one biclusters contain all 17

conditions. All biclusters are with MSR less than 300 and row variance above 300. For Yeast dataset all conditions are obtained when the MSR difference threshold for condition lists is set to 30. For gene list MSR difference threshold is set to 10. All the means squared residues are lower than 300. Only biclusters with different shapes are selected. Biclusters containing more genes having similar shape as that of biclusters ys3 are obtained in this method.



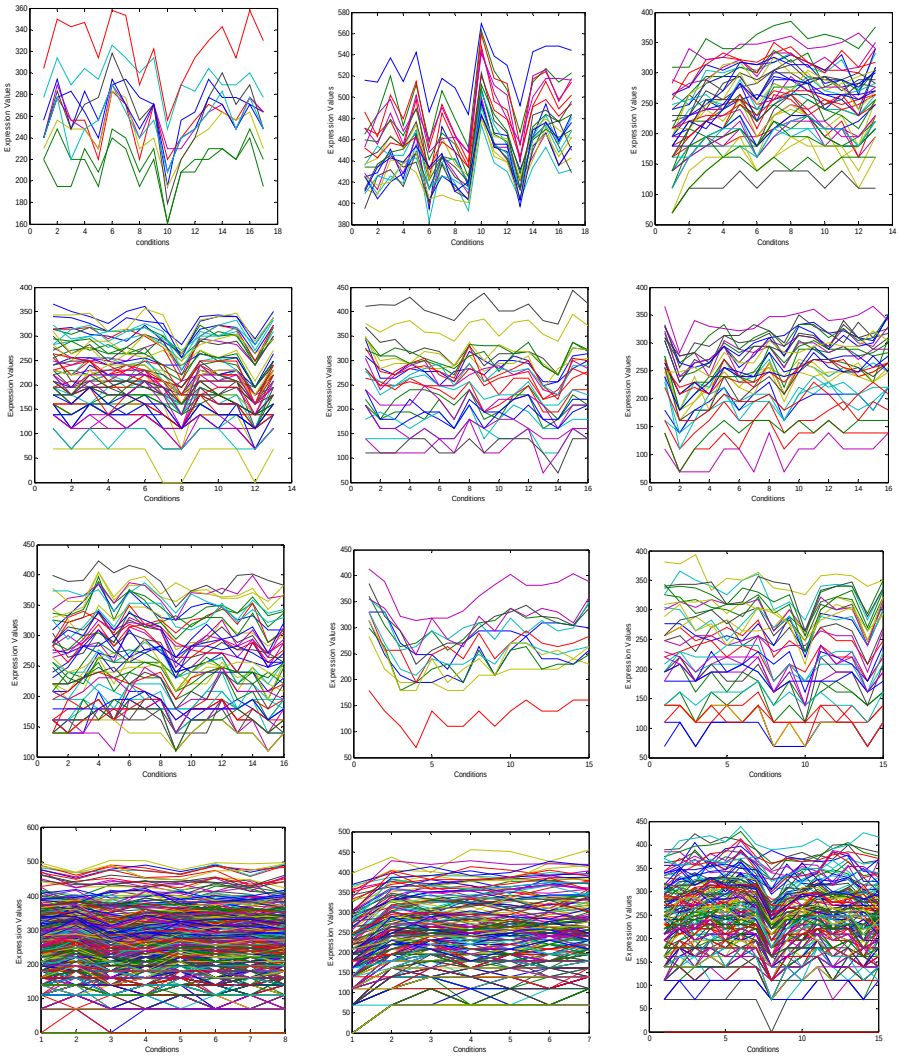


Figure 3.5 Twenty one biclusters found for the Yeast Dataset. Bicluster labels are (ya3), (yb3), (yc3), (yd3), (ye3), (yf3), (yg3), (yh3),(yi3), (yj3), (yk3), (yl3), (ym3), (yn3), (yo3), (yp3), (yq3), (yr3), (ys3) and (yt3), respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 3.8 using bicluster label.

Table 3.8

Information about Biclusters of Figure 3.5

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(ya3)	65	17	1105	198.8756	619.3479
(yb3)	86	17	1462	198.3953	526.8160
(yc3)	74	17	1258	199.6859	508.7565
(yd3)	1843	05	9215	299.8140	320.4440
(ye3)	81	17	1377	199.9548	551.3923
(yf3)	140	16	2240	199.6735	458.1247
(yg3)	55	17	935	199.4912	534.4627
(yh3)	17	16	272	199.3700	619.2619
(yi3)	119	17	2023	199.5356	518.8431
(yj3)	11	17	187	113.5428	501.8930
(yk3)	22	17	374	77.6240	641.7874
(yl3)	44	13	572	199.5335	695.5067
(ym3)	62	13	806	199.2022	531.5530
(yn3)	26	16	416	199.3954	455.0572
(yo3)	31	16	496	199.7230	625.6157
(yp3)	51	16	816	199.3443	385.4192
(yq3)	13	15	195	198.1322	959.9774
(yr3)	34	15	510	198.6582	489.9255
(ys3)	578	8	4624	198.5395	255.1215
(yt3)	444	7	3108	199.9317	514.9015
(yt4)	172	15	2580	199.8100	422.5933

In the Table 3.8 the first column contains the label of each bicluster. The second and third columns report the number of rows (genes) and number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residue or hscore of the bicluster. The last column contains the row variance.

3.2.2.2 Bicluster Plots for Human Lymphoma Dataset

In Figure 3.6 only nine biclusters obtained by the MSRDT algorithm are shown. The biclusters show similar up-regulation and down regulation. One bicluster (label lf3) is obtained with 91 conditions and row variance above 5700.

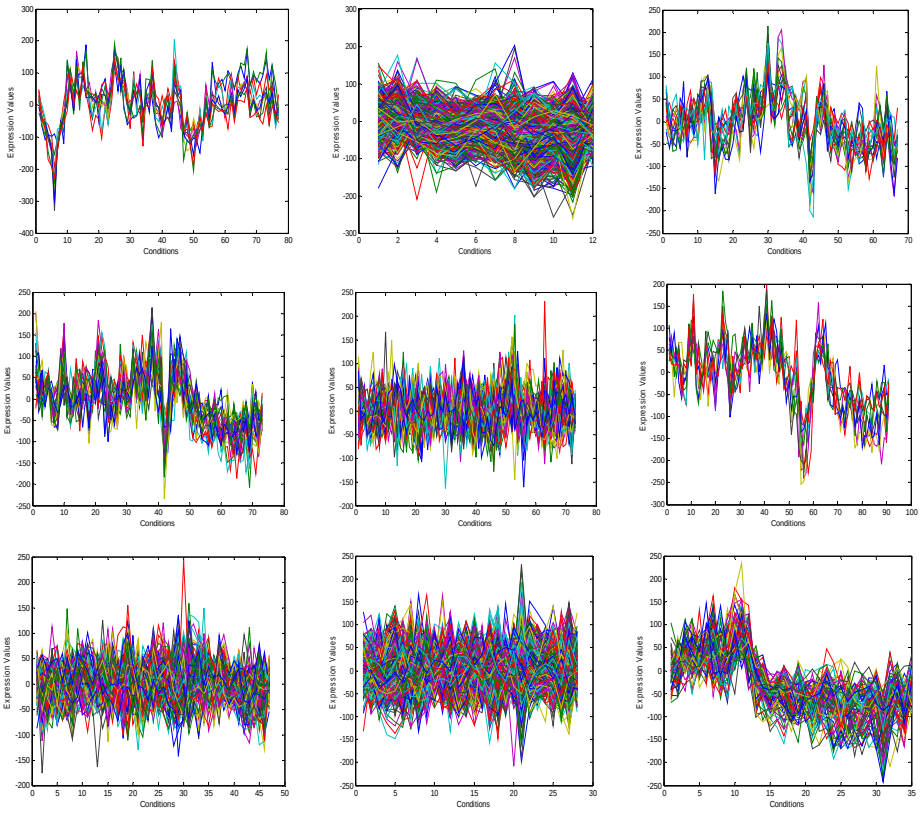


Figure 3.6 Nine biclusters found for the Lymphoma Dataset.Bicluster labels are (la3), (lb3), (lc3), (ld3), (le3), (lf3), (lg3) and (lh3) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 3.9 using bicluster label.

But for SEBI the maximum value of row variance for Lymphoma dataset is only 5691.07 and the maximum number of conditions obtained is only 72. All the means squared residues are lower than 1200. Experiments are conducted by setting the difference threshold for the condition list as 50, 60 etc and for gene list the values are 10, 20 etc.

Table 3.9

Information about Biclusters of Figure 3.6

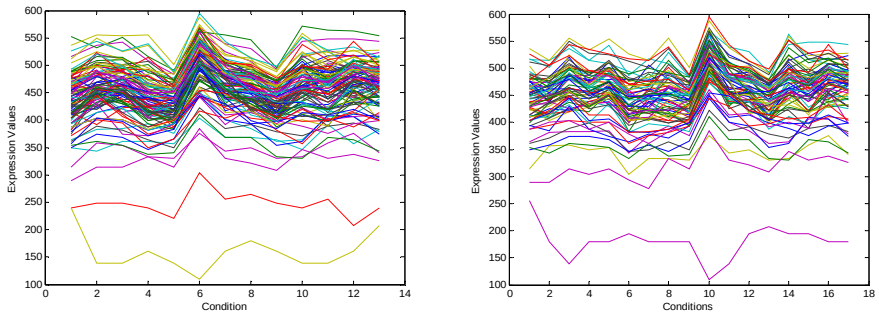
Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(la3)	10	77	770	1188.2	5439.2
(lb3)	910	12	10920	1199.0	1419.3
(lc3)	18	67	1206	1189.2	3430.8
(ld3)	30	73	2190	1197.4	3902.0
(le3)	64	73	4672	1199.6	1325.5
(lf3)	10	91	910	1183.1	5702.0
(lg3)	135	47	6345	1199.3	1321.1
(lh3)	690	28	19320	1199.7	1232.3
(li3)	72	35	2520	1183.1	3959.0

In the Table 3.9 the first column contains the label of each bicluster. The second column reports the number of rows (genes) of the bicluster. The third column reports the number of columns (conditions) of the bicluster. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residue or hscore of the bicluster. The last column contains the row variance of the bicluster.

3.2.3 Advantages of MSRDT Algorithm

This is the first algorithm to treat genes and conditions differently. This algorithm leads to the following research findings. The difference threshold created by genes is very small compared to that of conditions except for scaling patterns. This is one of the reasons by which metahueristic algorithms trying to minimize MSR will get biclusters with more genes. SEBI [36] is an exception to this problem because they are adjusting the fitness function to get more conditions. In this algorithm a bicluster (label lf3) with 91conditions is obtained for Lymphoma dataset and the MSR is less than that of the bicluster obtained by MSRT algorithm with 91 conditions. In MSRDT algorithm more genes and conditions can be accommodated compared to MSRT.

In MSRDT algorithm reducing the difference threshold for genes eliminates the possibility of adding inverted rows or mirror images [29] into the bicluster. This due to the fact that the genes which form mirror images will have high values for incremental increase in MSR. In Figure 3.7, two biclusters with inverted images are shown.



(m1) (m2)

Figure 3.7 Inverted images formed when MSR threshold alone is applied.

In the bicluster labelled (m1) there are 105 genes and 13 conditions with MSR value 215.2878. The gene which forms the mirror image is gene number 2581 and the incremental increase in MSR value when this gene is added is 16.8646. All other genes result in incremental increase in MSR less than 2. Similarly in the case of bicluster (m2) there are 73 genes and 17 conditions and MSR value is 182.74. The 520^{th} gene which causes the inverted image when added to the bicluster results in an incremental increase in MSR of 25.8368. But no other no other gene when added to the bicluster results in an incremental increase in MSR greater than 2.5. These biclusters can be obtained by any algorithm which makes use of MSR threshold alone. Even metaheuristic optimization algorithms with fitness function for minimizing MSR value will result in such biclusters.

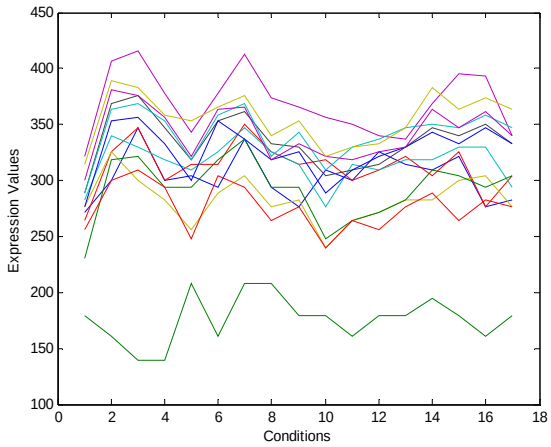


Figure 3.8 Another example of mirror image

If MSR difference threshold is applied with a difference threshold of value 10 for the gene list these genes will have to be removed. This proves that the newly introduced concept of MSR difference threshold

can eliminate the formation of mirror images in biclusters of gene expression data. The following figure shows the biclusters with the inverted images removed.

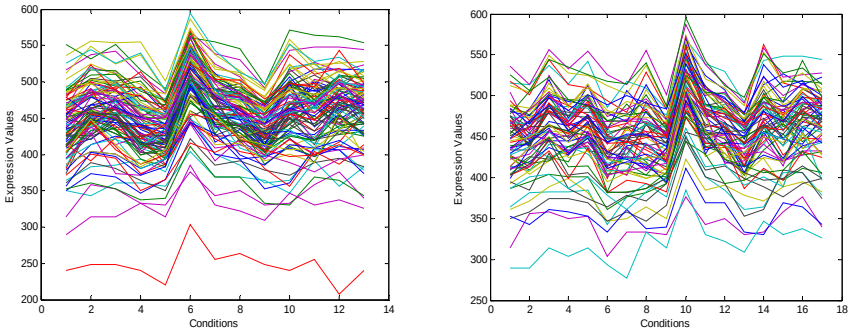


Figure 3.9 Inverted images removed when MSR difference threshold is applied.

It is found that decreasing the MSR difference threshold for condition list eliminates conditions which make significant change in the expression level from the bicluster whereas decreasing MSR difference threshold for gene list increases coherence.

3.2.4 Details of Significant Biclusters obtained by the MSRDT Algorithm

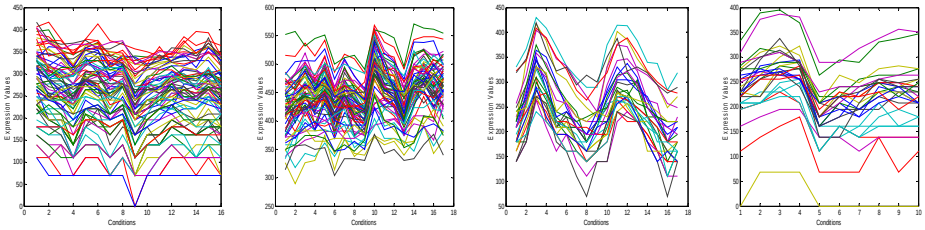


Figure 3.10 Four significant biclusters obtained by the algorithm on Yeast dataset. The bicluster labels are s31, s32, s33, s34. The details about biclusters can be obtained from Table 3.10 using bicluster label.

Table 3.10

Information about Biclusters of Figure 3.10

Bicluster label	Number of genes	Number of conditions	MSR	Row Variance
S31	77	16	199.5544	533.1660
S32	64	17	199.3198	654.5732
S33	28	17	286.3438	2034.1000
S34	28	10	235.4595	1186.3000

Biological relevance of biclusters obtained using MSRDT algorithm is verified using the four biclusters shown in Figure 3.10. GO annotation database is used to verify the biological significance of biclusters. In the first bicluster S31 selected for testing the biological significance there are 77 genes. They are YCL031C, YCR060W, YCR087W, YDL008W, YDL153C, YDL166C, YDL167C, YDL231C, YDL243C, YDR017C, YDR057W, YDR058C, YDR080W, YDR083W, YDR094W, YDR108W, YDR109C, YDR120C, YDR121W, YDR132C, YDR160W, YDR170C, YDR171W, YDR172W, YDR173C, YDR183W, YDR185C, YDR198C, YDR206W, YDR211W, YDR214W, YDR234W, YDR235W, YDR236C, YDR262W, YDR286C, YDR288W, YDR289C,YDR299W, YDR302W, YDR339C, YDR352W, YDR361C, YDR364C, YDR365C, YDR392W, YDR413C, YDR419W, YDR457W, YDR469W, YDR478W, YDR518W, YDR541C, YGL085W, YGL214W, YGR042W, YGR090W, YGR200C, YGR216C, YHR192W, YKR060W, YLR107W, YLR146CYML066C, YML096W, YNL132W, YNL199C, YOL031C, YOL080C, YOL124C, YOL140W, YOR061W, YOR091W, YOR098C, YOR145C, YOR272W, YPR053C.

In the second bicluster S32 there are 64 genes. They are YAL003W, YAL038W, YAR009C, YBL030C, YBL072C, YBL077W, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR106W, YBR118W, YBR181C, YBR189W, YBR191W, YCR012W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C,

YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL228C, YDL229W, YDR012W, YDR025W, YDR035W, YDR050C, YDR064W, YDR133C, YDR154C,YDR155C, YDR225W, YDR276C, YDR327W, YDR353W,YDR381W, YDR382W, YDR385W, YDR417C, YDR418W, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YDR529C, YDR545W, YEL034W, YER074W, YER117W, YGL102C, YKL152C, YMR202W, YOL127W, YOR234C, YPL037C, YPR102C.

In the third bicluster S33 there are 28 genes. They are YAR007C, YAR008W, YBL035C, YBR088C, YBR089W, YDL003W, YDL018C, YDL164C, YDR097C, YFL008W, YGR152C, YHR154W, YJL181W, YKL042W, YKL113C, YLR103C, YML021C, YML102W, YMR076C, YMR078C, YNL102W, YNL273W, YNL303W, YNL312W, YOL090W, YOR074C, YPL208W, YPR120C. In the fourth bicluster there are 28 genes. They are YBR038W, YBR138C, YCL012W, YDL039C, YGL021W, YGR035C, YGR092W, YGR108W, YHR023W, YHR151C, YIL106W, YIL162W, YJR092W, YKL129C, YKR021W, YLR190W, YLR353W, YML034W, YML119W, YMR001C, YMR032W, YNL053W, YNL171C, YOR152C, YPL148C, YPL242C, YPR007C, YPR119W.

The Table 3.11 given below shows the significant GO terms used to describe the set of genes of the biclusters of Figure 3.10 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 3.11 the first entry of the second column with the title 'Process' contains the term ribosome biogenesis (18, 4.78e-05) which means that 18 out of the 77 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 4.78e-05. This proves that the bicluster contains biologically similar genes and the MSRDT algorithm used here is capable of identifying biologically significant biclusters.

Table 3.11

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 3.10.

Bicluster	Process	Function	Component
S31	Ribosome biogenesis (18, 4.78e-05) ribonucleoprotein complex biogenesis 19, 7.68e-05), cellular component biogenesis at cellular level (20, 0.00039) RNA metabolic process (28, 0.00832)	32 out of 77 input genes are directly annotated to root term 'molecular function unknown':	Nucleolus (14, 8.24e-05) UTP-C complex (3, 0.00129) Preribosome (10, 0.00156) 90S preribosome (8, 0.00210)
S32	Translation (35, 2.26e-23) cellular protein metabolic process (38, 2.88e-11) protein metabolic process (38,7.49e-11) cellular metabolic process (51, 0.00015)	Structural constituent of ribosome (30, 2.58e-24) structural molecule activity (30, 1.79e-18) translation elongation factor activity (4, 0.00035)	Cytosolic ribosome (31, 2.71e-27) cytosolic part (31, 7.66e-25) ribosome (34, 6.60e-24) cytoplasmic part (47, 1.33e-06)
S33	DNA repair (17, 1.43e-14) DNA metabolic process (19, 7.23e-14) Response to DNA damage stimulus (17, 1.97e-13) Nucleobase, nucleoside, Nucleotide and nucleic acid metabolic process (21, 4.32e-06)	Double-stranded DNA binding (5, 4.58e-05) structure-specific DNA binding (5, 0.00115) DNA secondary structure binding (3, 0.00116)	Replication fork (9, 3.42e-09) chromosome (14, 1.85e-08) Nuclear replication fork (7, 1.01e-06) nucleus (22, 8.87e-06)
S34	Cytokinesis (8, 2.32e-05) Cell cycle process (13, 3.91e-05) Cell cycle (13, 6.36e-05) Cell division (8, 0.00014)	11 out of 28 input genes are directly annotated to root term 'molecular function unknown'	Cellular bud neck (11, 1.06e-09) Cellular bud (12, 1.12e-09) Site of polarized growth (12, 7.76e-09) Cytoskeletal part (10, 1.63e-06)

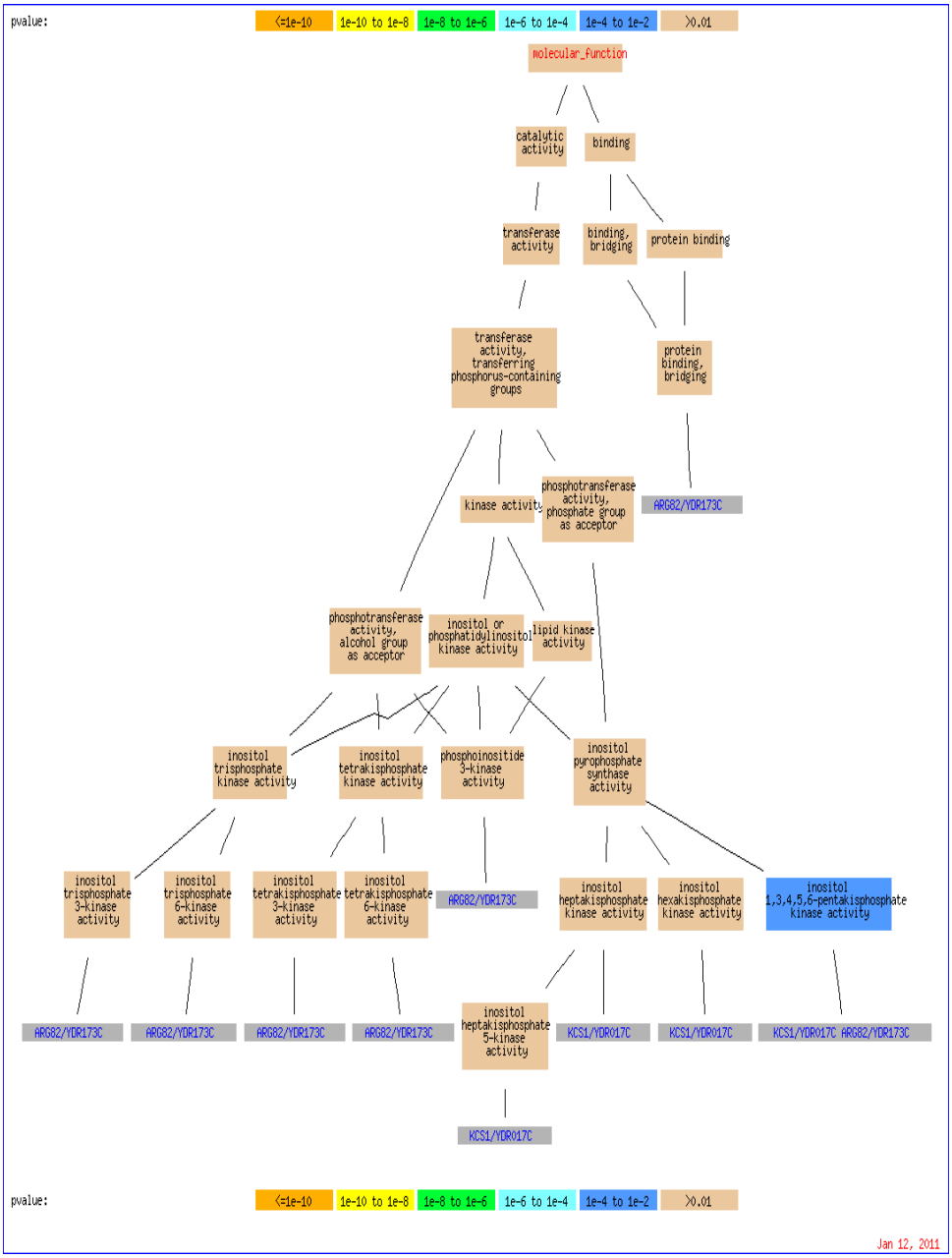


Figure 3.11 Sample of 98 genes for the bicluster s32 with corresponding GO terms and their parents for function ontology

Figure 3.11 shows the significant GO terms for the set of 98 genes in bicluster s32 along with their p values. It shows the branching of generalized molecular function into sub-functions like catalytic activity, and binding. These activities are clustered using genes to produce the final result. Figure 3.11 is obtained when gene ontology database is searched by entering the names of genes of S32 and by selecting function ontology.

3.2.5 Comparison with other Biclustering Algorithms

3.2.5.1 Comparison based on Statistical and Biological Significance

In Table 3.12 the GO terms along with their p-values and percentage of genes associated with the GO term in the biclusters for the MSRDT algorithm is compared with that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the Table it is clear that in terms of p.value obtained by a bicluster which is used to denote statistical significance MSRDT is better than RWB, Bimax, OPSM and Bivisu. The percentage of genes involved in the first GO term is better than that of RWB, OPSM and Bivisu. For the second and third GO terms the p-value and the percentage of genes are better than that of all the other algorithms except MOGAB. For the fourth GO term the p-value is better than all the other algorithms except MOGAB. Percentage of genes involved is better than all the other algorithms. For the fifth GO term p-value and percentage of genes involved is better than all the other algorithms.

Table 3.12
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by MSRDT and other Algorithms for Yeast Data

Terms	MSRDT	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 48.4% 2.71e-27	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Cytosolic Part 48.4% 7.66e-25)	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Structural constituent of ribosome 46.9% 2.58e-24)	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 53.1% 6.60e-24	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Structural Molecule Activity 46.9% 2.58e-24	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

3.2.5.2 Comparison on the basis of Bicluster Size and MSR

The Table 3.13 given below provides a comparative summarization of results of Yeast data involving the performance of related algorithms. The performance of MSRDT algorithm in comparison with that of Cheng and Church's (CC) [29], FLOC by Yang et al. [106], DBF [109], SEBI [36] and SMOB [37] for the Yeast dataset are given. In the MSRDT algorithm presented here the average mean squared residue is lower than that of CC, SEBI and SMOB. The average number of genes is greater than that of all other algorithms and the average number of conditions is better than that of all other algorithms except SEBI and SMOB. The MSRDT algorithm has highest value in the case of largest bicluster size compared to all other methods except CC.

In the case of MSRDT algorithm, MSR value is better than that of SEBI and CC. Largest bicluster size is better than that of all other algorithms. Average volume is better than that of all other algorithms. In multi-objective evolutionary biclustering [15] the maximum number of conditions obtained is only 11. In this method almost all biclusters are with 17 conditions. Moreover this algorithm provides better performance in terms of speed compared to all the metahueristic and evolutionary algorithms.

Table 3.13

Comparison between MSRDT and Other Algorithms for Yeast Dataset

Algorithm	AMR	ANG	ANC	AV	LB
MSRDT	199.63	170.16	14.83	2264.80	9215
SEBI	205.18	13.61	15.25	209.92	1394
SMOB	206.17	27.28	15.46	453.48	697
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	1627.00	4000

AMR is Average mean squared Residue. ANG is Average Gene Number of genes. ANC is Average Number of Conditions. AV is Average Volume. LB is Largest Bicluster. As 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.

Table 3.14 gives performance comparison for Human B-cell Lymphoma dataset. Value of δ is set to 1200 for Lymphoma dataset. In this dataset the average number of genes and average volume of the biclusters obtained are far better than that of SEBI and SMOB. Average number of conditions is greater than CC and SEBI.

In 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.

Table 3.14

Comparison between MSRDT and other Algorithms for Human Lymphoma Dataset

Algorithm	AMR	ANG	ANC	AV
MSRDT	1194.44	233.37	58.50	5791.63
CC	850.04	269.22	24.50	4595.98
SEBI	1028.84	14.07	43.57	615.84
SMOB	1019.16	11.60	78.47	709.13

AMR is Average mean squared Residue. ANG is Average Number Gene. ACN is Average Number of Conditions. AV is Average Volume.. As is clear from the above table the average mean squared residue, the average number of genes and conditions and average volume and are compared for various algorithms.

In multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 40 in Human B-cell Lymphoma dataset. But in this method there are biclusters with 91 conditions for Lymphoma dataset. Since the MSRDT algorithm uses simple sequential search rather than stochastic search the computation time required is very less compared to all the metahueristic and evolutionary algorithms. Some of the biclusters obtained are with high row variance (more than 2000 for the Yeast dataset and more than 7000 for Lymphoma dataset).

3.3 ISIMSRDT Algorithm

In this section a new algorithm developed using the concept of MSR difference threshold for finding biclusters from gene expression data is described. In MSRDT algorithm mentioned in the previous section it is difficult to find a suitable value for the MSR difference threshold. In the case of biclustering problem the main objective is to reduce the MSR value of the bicluster. Keeping this objective in mind the MSR difference threshold is initialised with a small value and it is incremented in each iteration until it reaches a final value. The results obtained on Yeast and Lymphoma datasets clearly indicate that this algorithm is better than many of the existing biclustering algorithms and also MSRDT, in terms of both bicluster size and MSR value. In this algorithm more genes and conditions are added to the seeds obtained from K-Means algorithm. After adding a gene or a condition if the incremental value of MSR is greater than MSR difference threshold, or if the MSR of the resulting bicluster is greater than δ, the added node is removed from the bicluster. In ISIMSRDT algorithm, MSR difference threshold is initialized with a small value and incremented after each iteration in fixed steps until it reaches a final value. So in ISIMSRDT algorithm there are three different parameters such as the initial value of MSR difference threshold, the amount by which it is incremented after each iteration and the final value of MSR difference threshold. These three parameters apply for both the gene list and condition list. By properly adjusting the MSR difference threshold parameters, biclusters of high quality can be obtained. A pseudo code description of the algorithm is given below.

Algorithm Iterative_MSRdifference (seed, δ, condthreshinitial, condthreshincrement, condthreshfinal, genethreshinitial, genethreshincrement, genethreshfinal)

```
bicluster := seed
previous=MSR(seed)
j:= 1;
msrdiffcondthresh=condthreshinitial;
while  (msrdiffcondthresh<condthreshfinal)
        While (j <= total _no_conditions)
    If   condition[ j]  is not included in bicluster
        Changed=1;
         Add all elements of condition[ j]  corresponding to genes
         already included to bicluster
         present= MSR(bicluster)
        if (present> δ) or (present-previous)>msrdiffcondthresh
            remove elements of  condition[ j]  from bicluster
            changed=0;
          endif
         if changed==1
              previous=present
        endif
  endif
 j:= j+1
 end(while)
 msrdiffcondthresh=msrdiffcondthresh+condthreshincrement
end(while)
 i := 1;
```

```
prev=MSR(bicluster)
msrdiffgenethresh=genethreshinitial
While(msrdiffgenethresh<=genethreshfinal)
   While (i <= total _no_ genes)
      If  gene[i]  is not included in bicluster
         Changed=1;
         Add all elements of gene[i] corresponding to conditions
         already included to bicluster
          present= MSR(bicluster)
         if (present> δ) or (present-previous)>msrdiffgenethresh
                  remove elements of  gene[i] from bicluster
                 changed=0
         endif
        if changed==1
           previous=present
       endif
    endif
 i:= i+1
end(while)
 msrdiffgenethresh=msrdiffgenethresh+genethreshincrement
end(while)
return bicluster
end(Iterative_MSRdifference)
```

3.3.1 Time Complexity of the Algorithm

The basic operation for the identification of biclusters is the calculation of MSR of a submatrix. Time complexity for calculating MSR is O(mn). This calculation is performed atmost m+n times for a single iteration. Hence the worst case time complexity is O(t((m+n)mn)) where m and n are the number of genes and conditions respectively and t is the total number of iterations.

3.3.2 Experimental Results

3.3.2.1 Bicluster Plots for Yeast Dataset

In Figure 3.12 nine biclusters obtained using ISIMSRDT algorithm are shown. Out of the nine biclusters, seven contain all 17 conditions and they differ in appearance. In short, the algorithm is ideal for identifying various biclusters with coherent values. All the biclusters are having mean squared residue less than 300. From the bicluster plots which show strikingly similar up-regulation and down-regulation it is concluded that this is an ideal method for identifying biclusters from gene expression data. For Yeast dataset biclusters are found by setting the initial value of MSR difference threshold for condition list as 5. It is incremented by 5 after each iteration and the final value of MSR difference threshold is set to 30. Initial value of MSR difference threshold for gene list is set to 1, and it is incremented by 1 and the final value is set to 10. All the means squared residues are lower than 300. Only biclusters with different shapes are selected.

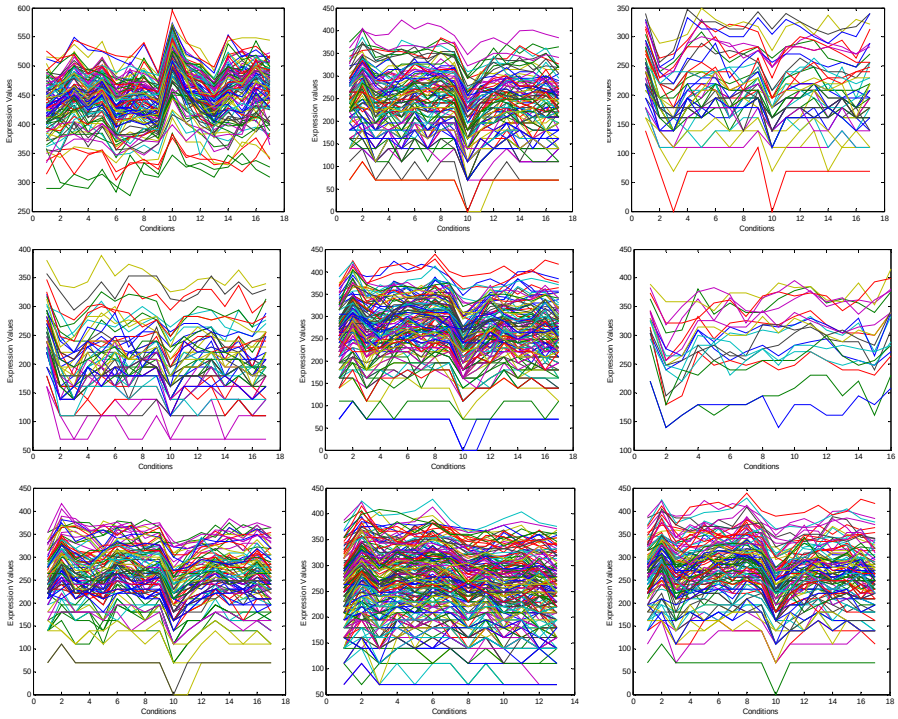


Figure 3.12 Nine biclusters found for the Yeast dataset. Bicluster labels are (ya4), (yb4), (yc4), (yd4), (ye4), (yf4), (yg4), (yh4) and (yi4) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 3.13 using bicluster label.

Table 3.15
Information about Biclusters of Figure 3.12

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(ya5)	98	17	1666	199.9381	591.9217
(yb5)	107	17	1819	199.9826	486.3663
(yc5)	43	17	731	199.8613	550.3640
(yd5)	50	17	850	199.5999	511.3709
(ye5)	127	17	2159	199.9656	471.1995
(yf5)	19	16	304	199.9141	564.4940
(yg5)	99	17	1683	199.9524	419.2172
(yh5)	188	13	2444	199.9713	353.8271
(yi5)	110	17	1870	199.9499	515.1427

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 number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residue or Hscore of the bicluster and the last column contains the row variance.

3.3.3.2 Bicluster plots for Human Lymphoma Dataset

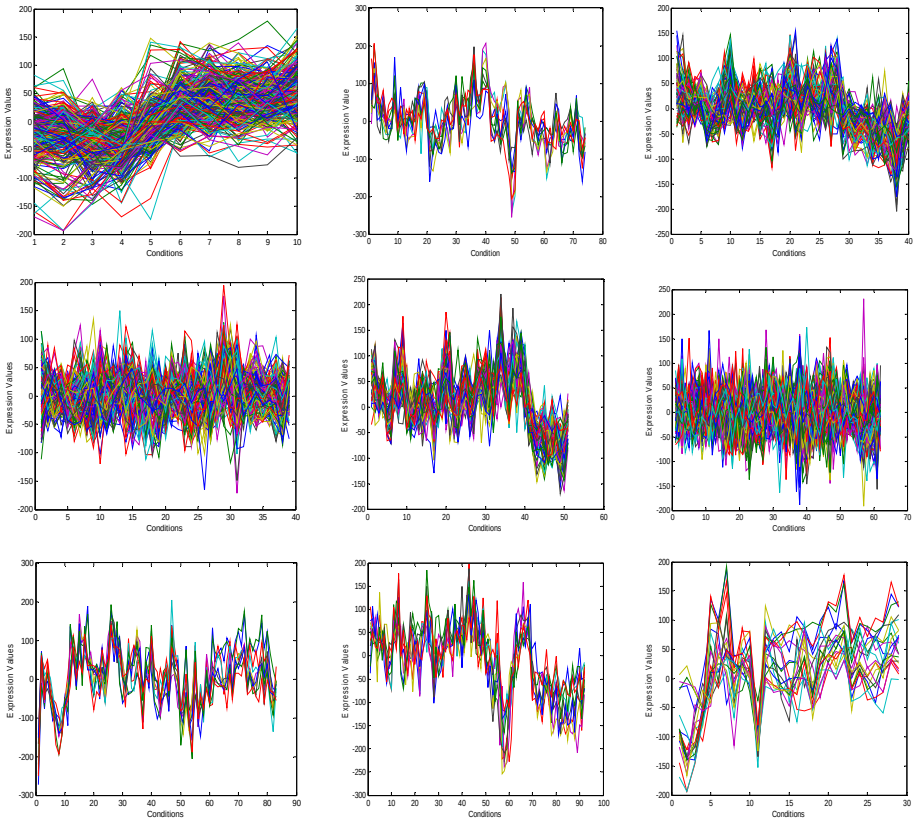


Figure 3.13 Nine biclusters found for the Lymphoma Dataset. The labels of biclusters are (la4), (lb4), (lc4), (ld4), (le4), (lf4), (lg4), (lh4) and (li4) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 3.16 using bicluster label.

Figure 3.13 shows nine biclusters obtained by ISIMSRDT algorithm on Human Lymphoma dataset. Here for condition list the initial value of MSR difference threshold is set to 30 and it is incremented by 30 after each iteration and the final value is set to 90. For the gene list the initial value of MSR difference threshold is set to 50 and it is incremented by 50 after each iteration and final value is set to 150. Experiments are conducted using other values also. All the bicluster plots show strikingly similar up-regulation and down-regulation. All the MSR are lower than 1200.

Table 3.16
Information about biclusters of Figure 3.13

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(la4)	280	10	2810	1001.40	2200.1
(lb4)	10	74	740	1199.00	4208.8
(lc4)	86	40	3440	999.88	2021.6
(ld4)	155	39	6045	999.92	1102.4
(le4)	51	51	2550	999.93	3139.5
(lf4)	172	62	10664	1199.80	1342.3
(lg4)	10	83	840	1194.90	5082.6
(lh4)	10	92	920	1197.40	5760.1
(li4)	20	29	580	987.80	4318.2

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 number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residue or hscore of the bicluster. The last column contains the row variance.

3.3.3 Advantages of ISIMSRDT Algorithm

This algorithm has various advantages over the MSRT and MSRDT algorithms. In the case of MSRT algorithm the added node is removed only when the MSR of the bicluster exceeds δ (MSR threshold). But when MSR difference threshold is applied in ISIMSRDT algorithm there is more restriction on the incremental value of MSR. This means that the elements in the biclusters are more tightly packed. This will result in biclusters of larger size and low MSR score. Hence ISIMSRDT method can produce better biclusters compared with other algorithms like MSRT. The ISIMSRDT algorithm gives the possibility of getting more genes and conditions compared to MSRDT algorithm. In MSRDT there is the disadvantage of finding a suitable value for MSR difference threshold. If a small value is assigned bicluster will be of small size. On the other hand if a big value is assigned the elements of the resulting bicluster will not be tightly co-regulated. This disadvantage of MSRDT can be overcome by using ISIMSRDT where MSR difference threshold is initialized with a small value and incrementing it after each iteration. There is another advantage of using iterative search in ISIMSRDT algorithm. The incremental increase in MSR of a gene or condition not included in a bicluster will vary as the size of the bicluster changes. For example in the case of bicluster labeled (lh4) in Figure 3.13 the MSR value of the bicluster when condition 95 is added is 1200.6. Since this is greater than MSR threshold for Lymphoma dataset (1200) condition 95 is removed from the bicluster. After adding condition 96, if condition 95 is added the MSR of the resulting bicluster is only 1191.9. This is less than 2000 and

hence after adding 96 if 95 is added it is not removed. Since conditions and genes are searched sequentially in all these algorithms, this is possible only if there is an iterative search. This is another option in iterative search for accommodating more genes and conditions. That means apart from finding a suitable value for MSR difference threshold iterative search has got another advantage of selecting the $(n-k)^{th}$ gene or condition whose incremental increase in MSR value reduces after adding the n^{th} gene or condition. Moreover in the case of ISIMSRDT algorithm also inverted rows are eliminated. In Lymphoma dataset a bicluster (label lh4) with 92 conditions is obtained.

3.3.4 Details of Significant Biclusters obtained by ISIMSRDT Algorithm

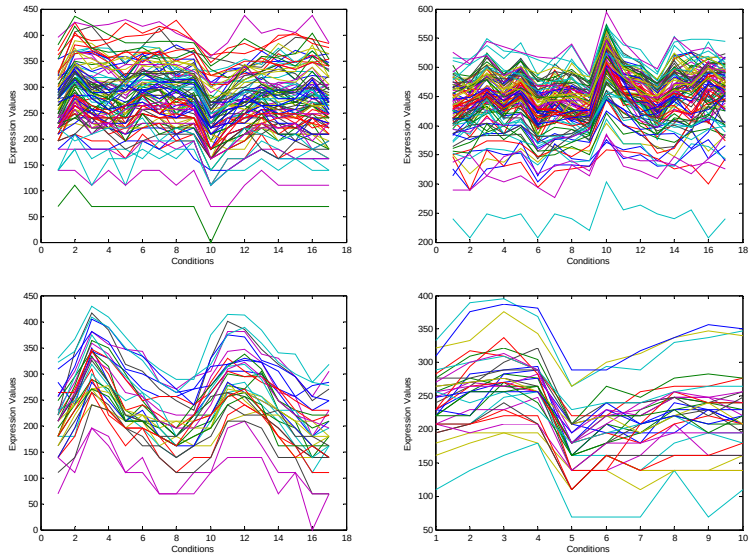


Figure 3.14 Four significant biclusters obtained by the ISIMSRDT algorithm on Yeast dataset. The bicluster labels are s41, s42, s43, s44. The details about biclusters can be obtained from Table 3.17 using bicluster label.

Table 3.17

Information about Biclusters of Figure 3.14

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
S41	98	17	199.9677	482.7704
S42	98	17	199.9924	600.9078
S43	33	17	299.2235	1970.1000
S44	33	10	242.2713	1125.8000

In the first bicluster S41 selected for testing the biological significance there are 98 genes namely YBL083C, YBR293W, YCL016C, YCL031C, YCL054W, YCR072C,YCR087W, YDL008W, YDL076C, YDL150W, YDL153C, YDL166C,YDL167C, YDL231C, YDL243C, YDR017C, YDR060W, YDR083W, YDR120C, YDR121W, YDR170C, YDR172W, YDR211W, YDR234W, YDR235W, YDR262W,YDR289C, YDR299W, YDR311W, YDR312W, YDR339C, YDR352W, YDR361C, YDR365C, YDR392W, YDR449C, YDR469W, YDR478W, YDR518W, YDR542W, YEL015W, YEL055C, YER005W, YER099C, YER107C, YER171W, YGL085W, YGL099W, YGL214W, YGR042W, YGR090W, YGR187C, YGR200C, YGR216C, YHR062C, YJL011C, YJL069C, YKR060W, YLL008W, YLL034C, YLR088W, YLR146C, YLR222C, YLR401C, YML066C, YML093W, YMR093W, YMR295C, YNL132W, YNL163C, YNL164C, YNL186W, YNL199C, YNR003C, YNR038W, YOL021C, YOL022C, YOL080C, YOL124C, YOL140W, YOL144W, YOR006C, YOR056C, YOR061W, YOR098C, YOR123C, YOR145C, YOR160W, YOR252W, YOR272W, YOR279C, YPL047W, YPL101W, YPL126W, YPL140C, YPL183C, YPR053C, YPR112C.

In the second bicluster S42 there are 98 genes. They are YAL003W, YAL038W, YAR009C, YAR020C, YBL027W, YBL030C, YBL072C, YBL077W, YBL092W, YBL113C, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR106W, YBR118W, YBR181C, YBR189W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W,

YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR382W, YDR385W, YDR433W, YDR447C, YDR450W, YDR471W, YEL034W, YER074W, YER117W, YGL102C, YGR118W, YJL188C, YJL190C, YJR009C, YJR123W, YKL056C,YKL060C,YKL096W-A, YKL152C, YKL153W, YKR057W, YKR094C, YLR029C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR340W, YLR406C, YLR441C, YLR467W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR045C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YNL339C, YOL039W, YOL040C, YOL083W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W, YPL090C, YPL142C, YPL143W, YPL283C, YPR043W, YPR102C, YPR204W.

In the third bicluster S43 there are 33 genes. They are YAR007C, YAR008W, YBL035C, YBR088C, YBR089W, YCR065W, YDL003W, YDL010W, YDL018C, YDL164C, YDR097C, YDR507C, YER095W, YFL008W, YGR151C, YGR152C, YHR154W, YIL026C, YJL074C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR236C, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL312W, YOR074C.

In the fourth bicluster S44 there are 33 genes namely YBR038W, YBR138C, YCL012W, YDL039C, YGL021W, YGR023W, YGR035C, YGR092W, YGR108W, YHR023W, YHR151C, YIL106W, YIL162W, YJL051W, YJR092W, YKL129C, YKR021W, YLR190W, YLR353W, YML033W, YML034W, YML119W, YMR001C, YMR032W, YMR213W, YMR291W, YNL053W, YNL171C, YOL130W, YOR152C, YPL148C, YPL242C, YPR119W

The Table 3.18 given below shows the significant GO terms used to describe the set of genes of the biclusters of Figure 3.14 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 3.18 the entry of the second column with the title process for the

bicluster S42 contains the term Translation(62, 2.03e-49) which means that 62 out of the 98 genes of the bicluster are involved in the process of translation and their p-value is 2.03e-49. Second and third entries indicate that 65 out of 98 genes are involved in cellular protein metabolic process and protein metabolic process. This proves that the bicluster contains biologically similar genes and ISIMSRDT algorithm used here is capable of identifying biologically significant biclusters.

Table 3.18
Significant Shared GO Terms (Process, Function and Component) of the Biclusters shown in Figure 3.14

Bicluster	Process	Function	Component
S41	Ribosome biogenesis (39, 3.08e-22) ribonucleoprotien complex biogenesis (40, 6.25e-21) cellular component biogenesis at cellular level((41, 1.68e-18) cellular nitrogen compound metabolic process(55, 1.86e-06)	snoRNA binding (4, 0.00480)	Nucleolus(31, 2.56e-19) Preribosome (23, 4.26e-15) nuclear lumen(43, 1.59e-13) Intracellular (90, 0.00018)
S42	Translation (62, 2.03e-49) cellular protein metabolic process (65, 3.08e-24) Protein metabolic process (65, 1.77e-23) Cellular metabolic process (77, 9.97e-07)	Structural constituent of ribosome (55, 6.05e-53) Structural molecule activity (56, 3.97e-42) translation elongation factor activity(5, 7.16e-05) RNA binding(15, 0.00208)	Cytosolic ribosome (57, 1.51e-60) Cytosolic part (57, 3.92e-55) Ribosome (61, 3.60e-51) Cytoplasmic part (74, 7.23e-12)

S43	DNA metabolic process(18, 1.11e-10) DNA repair(15, 3.25e-10) cell cycle(19, 1.13e-09) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (22, 6.05e-05)	structure-specific DNA binding (5, 0.00172) double-stranded DNA binding (4, 0.00202)	Chromosome (16, 2.01e-09) chromosomal part(14, 1.03e-07) Nuclear chromosome (13, 4.50e-07) replication fork (8, 6.19e-07) nucleus (24, 3.39e-05)
S44	Cytokinesis (8, 7.07e-05) cell division (8, 0.00043) cell cycle cytokinesis (6, 0.00130) cell cycle process (12, 0.00171)	12 out of 33 input genes are directly annotated to root term 'molecular function unknown	Cellular bud(13, 3.90e-10) cellular bud neck (11, 6.47e-09) Site of polarized growth(12, 5.50e-08) cellular bud neck contractile ring (5, 3.23e-07)

Figure 3.15 shows the significant GO terms for the set of 98 genes in bicluster S42 along with their p values. It shows the branching of a generalized molecular function into sub-functions like structural molecule activity, binding, protein tag, enzyme regulator activity and catalytic activity. These activities are clustered using genes to produce the final result. Figure 3.15 is obtained when gene ontology database is searched by entering the names of genes in bicluster S42 and by selecting function ontology.

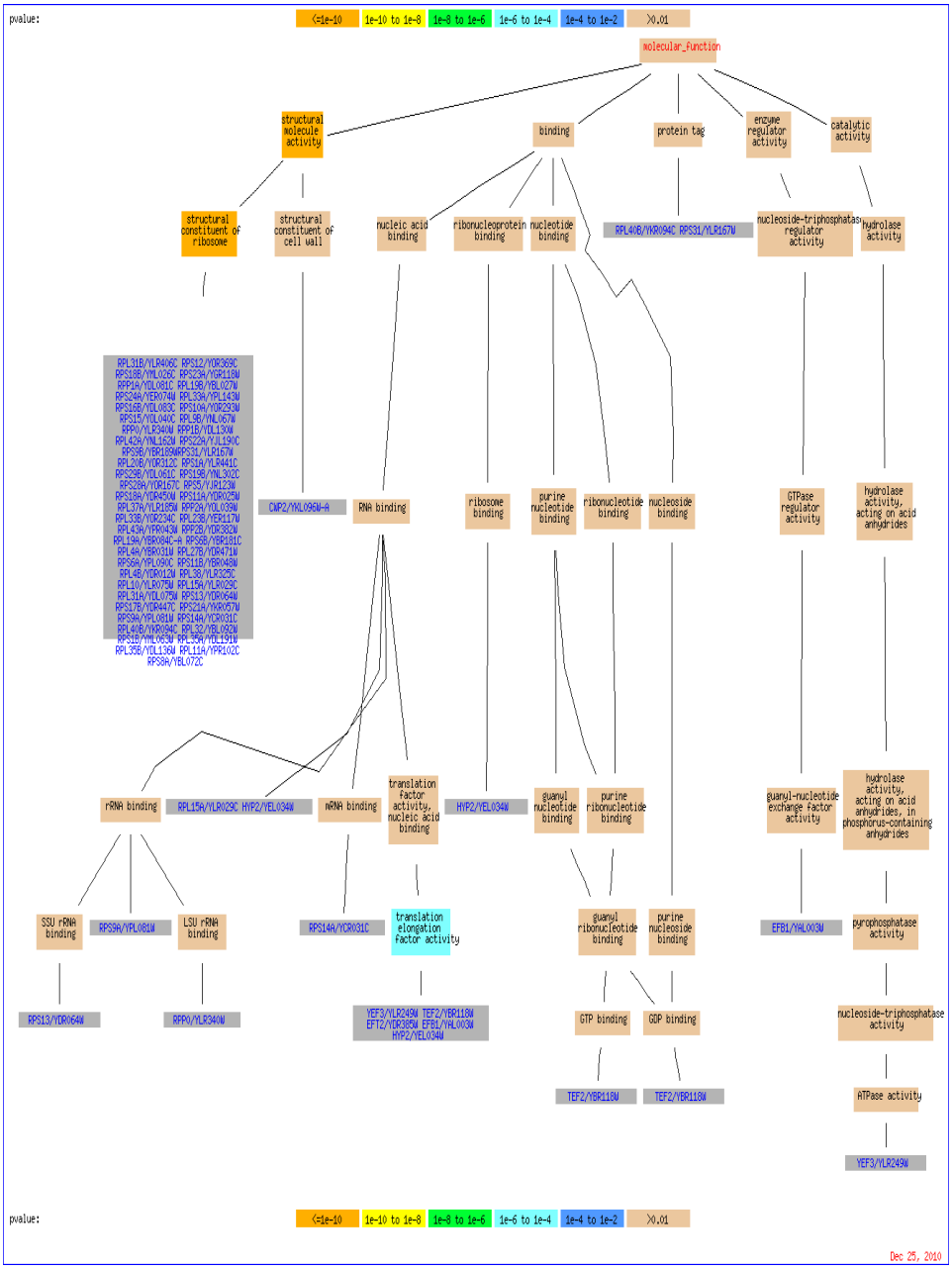


Figure 3.15 Sample of 98 genes for the bicluster S42, with corresponding GO terms and their parents for Function Ontology

3.3.5 Comparison with other Algorithms

3.3.5.1 Comparison of based on Statistical and Biological Significance

In Table 3.19 the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the ISIMSRDT is compared with MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the Table it is clear that in terms of p-value obtained by a bicluster which is used to denote statistical significance ISIMSRDT is better than all the other algorithms namely MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu for all the five GO terms. For the first GO term the percentage of genes involved is better than that of CC, RWB, OPSM, ISA and BiVisu. For the second, fourth and fifth GO terms the percentage of genes involved is better than that of all the other algorithms. For the third GO term the percentage of genes involved is better than that of all the other algorithms except MOGAB.

Table 3.19
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by ISIMSRDT and other Algorithms for Yeast Data

Terms	ISIMSRDT	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 58.2% 1.51e-60	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Cytosolic Part 58.2% 3.92e-55	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Structural constituent of ribosome 56.1% 6.05e-53	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 62.2% 3.60e-51	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Translation 62% 2.03e-49	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

3.3.5.2 Comparison of Biclusters Produced by MSRT, MSRDT and ISIMSRDT Algorithms using the Same Seed

A comparison of these three algorithms is given on the basis of size of biclusters obtained and their MSR value starting with the same seed and the result is given in Table 3.20. In terms of bicluster size ISIMSRDT is always better than the other two algorithms. MSRDT is better than MSRT for all seeds except for seed 3. In the case of biclustering using MSRDT algorithm there is a single but different MSR difference threshold value for the gene list and condition list. In this case the parameters for the MSRDT algorithm for Yeast dataset are condition difference threshold=30 and gene difference threshold=10. The parameters for ISIMSRDT are initial value of condition difference threshold=5, increment=5 and the final value of condition difference threshold=30. Similarly the parameters for gene list are initial value of gene difference threshold=1, increment=1 and the final value of gene difference threshold=10. From Table 3.17 it is clear that ISIMSRDT produces large size biclusters compared to MSRDT. Hence iterative search with incremental MSR difference threshold is always better than assigning a single value for MSR difference threshold.

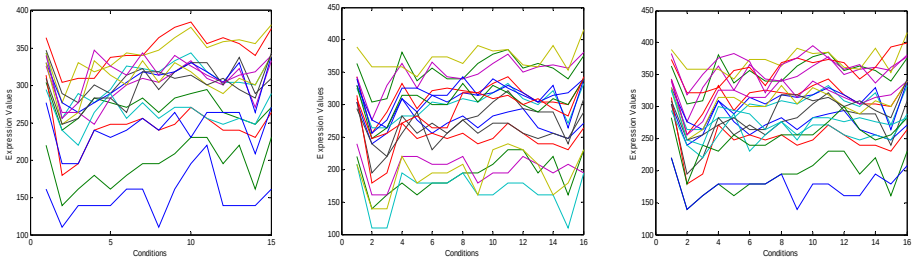
In the above Table the first column reports the seed number. The second column reports the size and MSR score of the bicluster generated by the ISIMSRDT algorithm. The third column reports the size and MSR score of the bicluster generated by the MSRDT algorithm. The fourth column reports the size and MSR score of the bicluster generated by the

MSRT algorithm. Figure 3.16 displays three biclusters obtained by the three algorithms from the same seed.

Table 3.20

Difference between Biclusters obtained by the Three Algorithms Starting from the Same Seed

S. No	ISIMSRDT		MSRDT		MSRT	
	Bicluster Size	MSR	Bicluster Size	MSR	Bicluster Size	MSR
1	110*17	199.95	78*16	199.96	75*17	199.95
2	93*17	199.79	65*17	198.88	57*17	199.09
3	99*17	199.95	74*17	199.69	92*17	199.71
4	96*17	199.69	86*17	198.39	79*17	198.96
5	125*17	199.91	119*17	199.54	117*17	199.94



(m) (n) (o)

Figure 3.16: Biclusters from same seed for the three algorithms.

The details about the biclusters shown in Figure 3.16 are given in Table 3.21. From the bicluster plots it is clear that in this case MSRDT algorithm gives the best bicluster in terms of row variance in this case.

Table 3.21
Information about Biclusters given in Figure 3.16.

Bicluster Label	Algorithm	Size	MSR	Row Variance
(m)	MSRT	15*15	198.58	610.64
(n)	MSRDT	17*16	199.37	619.26
(o)	ISIMSRDT	19*16	199.91	564.49

In the above table the first column reports the label of the bicluster. The second column reports the algorithm from which the bicluster is generated. The third column reports the size of the bicluster. The fourth column reports the MSR and the last column reports the row variance of the biclusters.

In the case of Lymphoma dataset starting from the same seed MSRT and MSRDT algorithms produced biclusters with 91 conditions. Even though both are of size 10*91 the bicluster produced by MSRDT is better in terms of MSR value (low) and row variance (high). But in the case of ISIMSRDT it should be noticed that this algorithm produced bicluster with 92 conditions and higher row variance from the same seed (Bicluster with label lh4 of Table 3.16).

3.3.5.3 Comparison based on Bicluster Size and MSR

The Table 3.22 given below provides a comparative summarization of the results of the performance of related algorithms in Yeast dataset. All the algorithms listed in Table 3.22 are having MSR value more or less equal to 200, even though the maximum limit of δ is 300. The

performance of ISIMSRDT algorithms in comparison with that of SEBI [36], Cheng and Church's algorithm (CC) [29], and the algorithm FLOC by Yang et al. [106] and DBF [109] for the Yeast dataset are given. For ISIMSRDT average number of conditions is better than that of all the other algorithms. In the case of ISIMSRDT algorithm presented here average number of genes is greater than that of SEBI whereas the average number of conditions is better than that of all other algorithms. Average volume is greater than that of SEBI and CC. Average residue is lower than that of CC and SEBI. The ISIMSRDT algorithm has high value for the largest bicluster size compared to SEBI and FLOC.

Table 3.22

Performance Comparison between ISIMSRDT and Other Algorithms for Yeast Dataset

Algorithm	AMR	ANG	ANC	AV	LB
ISIMSRDT	199.96	123.80	16.20	1954.20	2444
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	1627.00	4000

AMR is Average mean squared Residue. ANG is Average Gene Number of genes. ANC is Average Number of Conditions. AV is Average Volume. LB is Largest Bicluster. As 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.

Table 3.23 given below provides a performance comparison for Human B-cell Lymphoma dataset. Value of δ is set to 1200 for Lymphoma dataset. For ISIMSRDT average MSR is better than that of all the other algorithms except CC. Here the average gene number is greater than SEBI. Average value of condition is better than all other algorithms. Average volume is better than that of SEBI.

Table 3.23

Comparison between ISIMSRDT and other Algorithms for Human Lymphoma Dataset

Algorithm	ANG	ANC	AV	AMR
ISIMSRDT	98.00	48.63	3458.62	923.47
SEBI	14.07	43.57	615.84	1028.84
CC	269.22	24.50	4595.98	850.04

AMR is Average mean squared Residue. ANG is Average Gene Number of genes. ANC is Average Number of Conditions. AV is Average Volume. As clear from the above table the average mean squared residue, the average number of genes and conditions and average volume 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.

Usually multi-objective algorithms will produce biclusters of larger size. But in the case of multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 11 in the case of Yeast dataset and 40 in the case of Human B-cell Lymphoma dataset. But in this method there are biclusters with all 17 and 92 conditions for Yeast and Lymphoma datasets respectively. For the Yeast dataset the maximum number of genes obtained for this algorithm in all the 17 conditions is 127 with MSR value 199.9656. The maximum available in all the literature published so far is in multi-objective PSO [62]. They obtained 141 genes for 17 conditions with MSR value 203.25. Moreover the ISIMSRDT algorithm provides better performance in terms of speed compared to all the metaheuristic and evolutionary algorithms. Hence ISIMSRDT algorithm has a definite comparative differential advantage over the previous algorithms. In the multi-objective PSO, the maximum number of conditions obtained for Lymphoma dataset is 84. But for ISIMSRDT algorithm a bicluster with 92 conditions is obtained.

3.4 SGSC Algorithm

In this section a new algorithm is developed for biclustering gene expression data. Seeds obtained from K-Means are enlarged using a method in which the constraints used for genes and conditions are set separately to identify biclusters. Results obtained here are better than some of the metaheuristic and multi-objective evolutionary methods. The expansion for SGSC is Seed Growing using Separate Constraints for genes and conditions. In the seed growing phase after adding a gene or

condition the MSR value reduces or increases. Experiments are conducted by calculating MSR difference threshold. While experiments are done it is found that, reducing the difference threshold for conditions removes the conditions which make significant change in the expression level, whereas reducing the difference threshold for the genes increases coherence. A highly coherent gene which shows similar fluctuation will produce very small change in MSR value when added to the bicluster except in the case of scaling patterns. A negatively correlated gene will make a large variation in the MSR value. Moreover, the incremental increases in MSR caused by adding genes are small compared to that of conditions. But when a gene is added to the bicluster the pattern will not change. When a condition is added to the bicluster, the pattern of the bicluster will change. Conditions which cause a large variation in the expression level of genes will make a large incremental increase in MSR value also. The mean squared residue is a popular measure used to evaluate the quality of a bicluster. One drawback however is that it is biased towards flat biclusters with low row variance [24]. The row variance in a bicluster is increased by adding certain conditions in which the expression level of the gene is very high. Such conditions when added will increase the mean squared residue also. So optimization problems which try to add conditions by reducing MSR will rarely find biclusters with high row variance.

All these observations lead to the conclusion that to include those conditions which cause a large variation which in turn helps to get biclusters with high row variance the MSR difference threshold for the

condition should be large and to increase coherence the MSR difference threshold for the gene should be small. So the constraint for conditions is set to the maximum allowable limit that is the MSR threshold and the allowable incremental increase in genes is set to a very small value. Hence after adding a condition the MSR value of the resulting submatrix is calculated in order to verifying whether it exceeds the given MSR threshold. If it exceeds the given MSR threshold it is removed from the submatrix. After adding the gene MSR value of the resulting submatrix is calculated in order to verifying whether it exceeds the MSR difference threshold or the MSR threshold. If so the gene is removed from the bicluster. This process is continued till the last gene or condition is verified for inclusion in the bicluster. MSR difference threshold is set to very small value. In this study the MSR difference threshold is relevant for genes only and it is in the range of 0.1 to 10. Usually increasing this value increases the number of genes and reduces row variance. In this method some of the seeds will result in biclusters with large row variance. The results obtained here are superior compared to that of other algorithms which use multi-objective optimization methods. It is easy to get biclusters of different shapes since different seeds will result in different biclusters almost all the time with a few exceptions. This algorithm is deterministic in the sense that for a given threshold value of MSR, the MSR difference threshold and for a given seed, the repeated executions, will produce the same result. A pseudo code description of the algorithm is given below.

```
Algorithm Separateconstraintsgenecond(seed, δ,x)
// δ denotes the MSR threshold
//x denotes the MSR difference threshold for genes //which is set to a small value
bicluster := seed;    j := 1;
While (j <= total _no_conditions)
if condition[ j] is not included in the bicluster
Add all elements of condition[j] corresponding to genes already included to the bicluster
Msrbicluster=MSR(bicluster)
  if (Msrbicluster> δ) remove elements of condition[ j]
from the bicluster  and  restore previous MSR value
endif
endif
j:= j+1
end(while)
i=1;
While (i <= total _no_ genes)
If gene[i] is not included in the bicluster
Add all elements of gene[i] corresponding to conditions already included to the bicluster
Msrbicluster=MSR(bicluster)
MSRDifference=Incremental_Increasein_MSR(bicluster)
   if (Msrbicluster > δ or MSRDifference>x)
        remove elements of  gene[i] from the bicluster
        restore the previous MSR value
endif
endif
i=i+1; end(while)
 return bicluster
end(Separateconstraintsgenecond)
```

3.4.1 Time Complexity of the Algorithm

The basic operation for the identification of biclusters is the calculation of mean squared residue of a submatrix. Time complexity for calculating MSR is O(mn). In order to include a gene or a condition, the MSR value is calculated once. There are m+n genes and conditions. Hence this calculation is performed atmost m+n times. That means the worst case time complexity of the algorithm is O((m+n)mn) where m and n are the number of genes and conditions respectively. This algorithm is very fast compared to evolutionary or metahueristic algorithms. The main operation for finding bicluster is the calculation of the MSR value of a submatrix. In this algorithm, the number of submatrices whose MSR is to be calculated is at most m+n, where m and n are the number of genes and conditions respectively. Usually m+n will be less than 4200. In the case of evolutionary algorithms the number of submatrices whose MSR is to be calculated is P*I where P is the number of populations and I is the number of iterations. For SEBI and SMOB the value of P*I is 20000.

3.4.2 Experimental Results

Experiments are conducted on the Yeast Saccharomyces cerevisiae cell cycle expression dataset and Human Lymphoma dataset in order to evaluate the quality of the proposed algorithm.

3.4.2.1 Bicluster Plots for Yeast Dataset

In Figure 3.17, out of the many biclusters found by the algorithm only 12 biclusters with different shapes are shown. From the bicluster plots it is clear that highly coherent biclusters are obtained using this method. When

this algorithm is used some of the seeds produce biclusters with row variance above 2000. In SEBI the attempt was to identify biclusters with high row variance by adjusting the fitness function. The minimum value of row variance they obtained for the biclusters in Yeast dataset was 317.23. In this study, all biclusters obtained are with row variance above 317.23. Biclusters with all 17 conditions are obtained using this method.

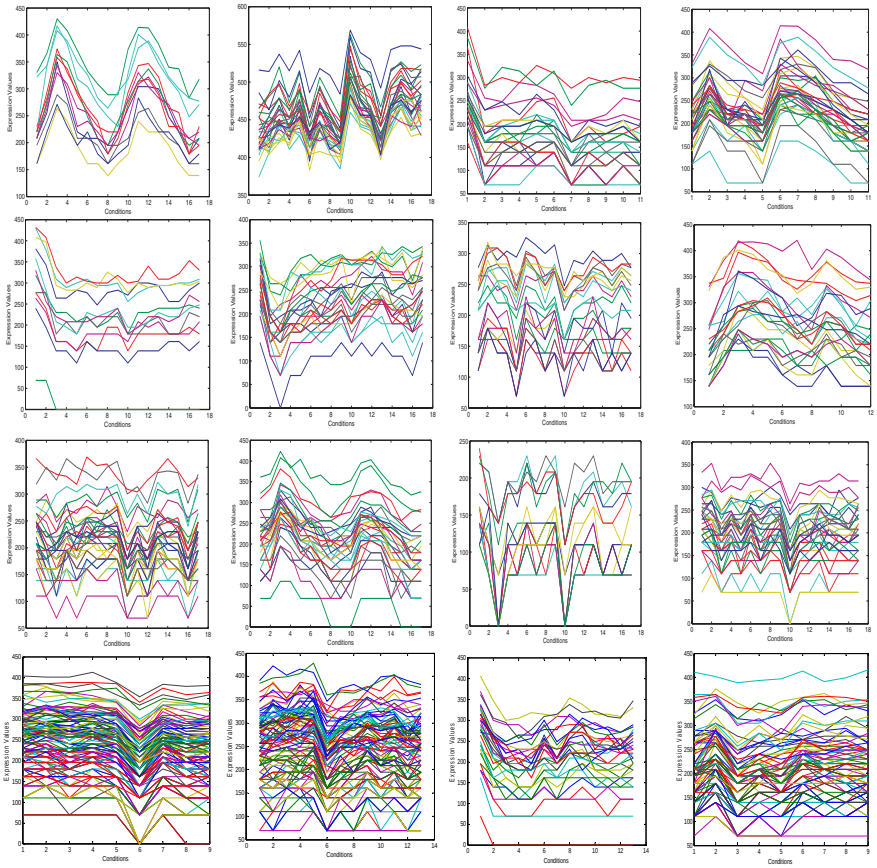


Figure 3.17 Sixteen biclusters obtained using SGSC algorithm on Yeast dataset. From left to right and from top to bottom the bicluster labels are (ya5), (yb5), (yc5), (yd5), (ye5), (yf5), (yg5), (yh5), (yi5), (yj5), (yk5), (yl5) ym5), (yn5), (yo5) and (yp5) respectively. The details of the biclusters can be obtined from Table 3.24 using bicluster label.

Table 3.24

Information about Biclusters of Figure 3.17

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
Ya5	12	17	204	197.66	2211.80
Yb5	29	17	493	95.45	643.59
Yc5	29	11	319	226.03	1301.50
Yd5	37	11	407	259.14	1454.90
Ye5	12	17	204	182.31	1092.30
Yf5	25	17	425	266.87	1079.50
Yg5	22	17	374	183.04	545.83
Yh5	24	12	288	244.92	917.74
Yi5	36	17	612	229.25	643.17
Yj5	32	17	544	298.21	1444.70
Yk5	17	17	289	294.08	1253.90
Yl5	36	17	612	194.12	592.80
Ym5	125	9	1125	163.35	720.08
Yn5	107	13	1391	166.99	405.97
Yo5	32	13	416	225.69	743.69
Yp5	87	9	783	189.52	447.71

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 fifth column contains the mean squared residue of the bicluster and the last column contains the row variance of the bicluster.

3.4.2.2 Bicluster Plots for Lymphoma Dataset

In Figure 3.18, out of the many biclusters found by the algorithm, only 12 biclusters are shown. The genes in the bicluster present a similar behavior under a set of conditions. Bicluster (la5) contains the maximum number of conditions obtained in this method i.e. 91. Bicluster (le5) is having row variance above 7000. The MSR value of the bicluster (ll5) is only 797.3 but the row variance is above 5000. As Federico Divina and Jesus S. Aguilar-Ruize has observed [37], even though there are no shifting and scaling patterns [10] in the biclusters of Lymphoma dataset, local shifting patterns are obtained in some biclusters.

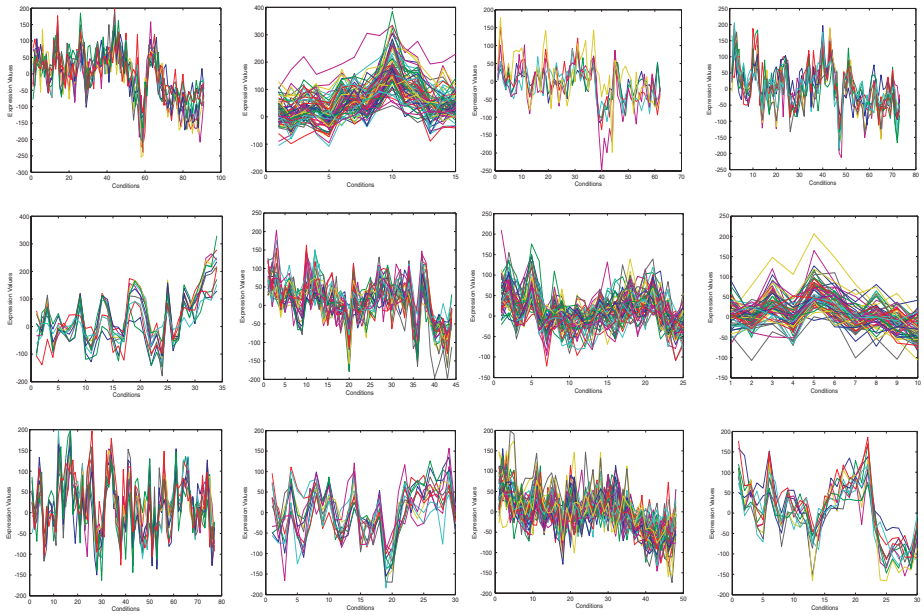


Figure 3.18 Twelve biclusters obtained using SGSC algorithm on Lymphoma dataset. From left to right and from top to bottom the bicluster labels are (la5), (lb5), (lc5), (ld5), (le5), (lf5), (lg5), (lh5), (li5), (lj5), (lk5) and (ll5) respectively. The details of the biclusters can be obtained from Table 3. 25 using bicluster label

Table 3.25

Information about Biclusters of Figure 3. 18

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
la5	10	91	910	1190.5	5308.5
lb5	68	15	1020	1085.0	3350.6
lc5	6	62	372	1048.8	2886.7
ld5	11	73	803	1150.9	4234.6
le5	11	34	374	1142.0	7936.9
lf5	26	44	1144	1032.7	3195.6
lg5	54	25	1350	894.3	1621.5
lh5	61	10	610	604.1	1307.9
li5	10	77	770	1140.5	4630.4
lj5	12	30	360	1118.6	3572.9
lk5	48	48	2304	946.8	2168.7
ll5	11	30	330	797.3	5314.8

3.4.3. Advantages of SGSC Algorithm

This algorithm identifies biclusters with very high coherence. With the help of bicluster plot it can identify biclusters with very high row variance and MSR above the threshold. Some of the shifting and scaling patterns can be identified by this algorithm.

3.4.4 Details of Significant Biclusters obtained by SGSC Algorithm

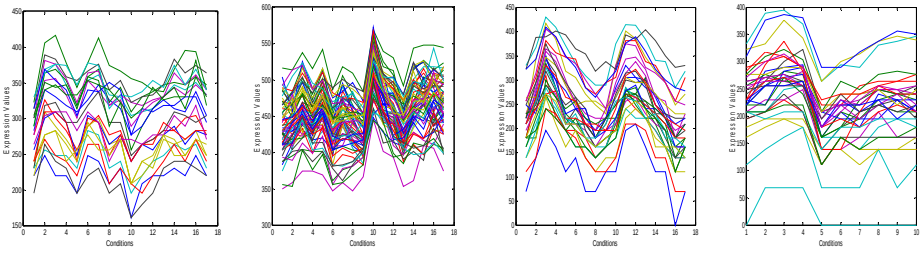


Figure 3.19 Four significant biclusters obtained by the SGSC algorithm on Yeast dataset. The bicluster labels are s51, s52, s53, s54. The details about biclusters can be obtained from Table 3.26 using bicluster label.

Table 3.26

Information about Biclusters of Figure 3.19

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
S51	23	17	131.3915	506.7582
S52	63	17	167.4308	615.9798
S53	31	17	297.1918	2036.0000
S54	33	10	243.6711	1135.7000

The biological relevance of biclusters obtained using SGSC algorithm is verified using the four biclusters shown in Figure 3.19. GO annotation database is used to verify the biological significance of biclusters.

In the first bicluster S51 selected for testing the biological significance there are 23 genes. They are YCL031C, YCR087W, YDL008W, YDL153C, YDL166C, YDL167C, YDR083W, YDR121W, YDR172W, YDR211W, YDR289C, YDR339C, YDR352W, YDR365C, YDR392W, YDR469W, YDR478W, YDR518W, YDR542W, YGR200C, YOL140W, YOR272W, YPR053C.

In the second bicluster S52 there are 63 genes. They are YAL003W, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR106W, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W,YDL136W, YDL191W, YDL192W, YDL208W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR382W, YDR385W, YDR433W, YDR447C, YDR450W, YDR471W, YGL102C, YKL152C, YKL153W, YLR029C, YLR167W, YLR325C, YLR406C, YLR441C, YML026C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W, YPL090C, YPL142C, YPL143W, YPL283C, YPR043W,YPR102C, YPR204W.

In the third bicluster S53 there are 31 genes. They are YAR007C, YAR008W, YBL035C, YBR088C, YBR089W, YDL003W, YDL018C, YDL164C YDR097C, YDR507C, YFL008W, YGR152C, YHR154W, YIL026C, YJL074C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR383W, YLR386W, YML021C, YML102W, YMR076C, YMR078C, YMR305C, YNL312W, YOL090W, YOR074C.

In the fourth bicluster S54 there are 33 genes. They are YBR038W, YBR138C, YCL012W, YDL039C, YGL021W, YGR023W, YGR035C, YGR092W, YGR108W, YHR023W, YHR151C, YIL106W, YIL162W, YJL051W, YJR092W, YKL129C, YKR021W, YLR190W, YLR353W, YML033W, YML034W, YML119W, YMR001C, YMR032W, YMR291W, YNL053W, YNL171C, YOL130W, YOR152C, YPL148C, YPL242C, YPR007C, YPR119W.

The Table 3.27 given below shows the significant GO terms used to describe genes of the biclusters of Figure 3.19 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 3.27 the first entry of the second column with the title process contains the term rRNA processing (7, 0.00144) which means that 7 out of the 23 genes of the bicluster are involved in the process of rRNA processing and their p-value is 0.00144. Second entry indicates that 8 out of the 23 genes are involved in ncRNA processing. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the SGSC algorithm used here is capable of identifying biologically significant biclusters from different GO categories.

Table 3.27

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 3.19

Bicluster	Process	Function	Component
S51	rRNA processing (7, 0.00144) ncRNA processing (8, 0.00171) RNA metabolic process(13, 0.00184) gene expression (14, 0.00678)	10 out of 23 input genes are directly annotated to root term 'molecular function unknown':	Nucleolus (6, 0.00622)
S52	Translation (46,3.12e-40) cellular protein metabolic process (49, 9.61e-24) protein metabolic process(49, 3.96e-23) cellular metabolic process (55, 5.85e-08)	Structural constituent of ribosome (42, 4.42e-44) structural molecule activity (42, 3.48e-35)	Cytosolic ribosome (42, 1.30e-46) cytosolic part (42, 5.45e-43) ribosome(45, 6.05e-41) organelle (54, 0.00081)
S53	DNA repair (16,4.68e-12) DNA metabolic process (18, 2.50e-11) response to DNA damage stimulus (16, 5.29e-11) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (21, 7.56e-05)	Double-stranded DNA binding(5, 8.01e-05) DNA secondary structure binding (3, 0.00162) structure-specific DNA binding (5, .00198) guanine/thymine mispair binding (2, 0.00469)	Chromosome (15, 8.04e-09) chromosomal part(13, 5.29e-07) mitotic cohesin complex (4, 5.95e-07) nucleus (23, 3.19e-05)
S54	Cytokinesis (8, 6.87e-05) cell cycle process (13, 0.00024) cell cycle (13, 0.00039) cell division (8, 0.00042)	13 out of 33 input genes are directly annotated to root term 'molecular function unknown':	Cellular bud (13, 3.41e-10) cellular bud neck(11, 6.47e-09) site of polarized growth (12, 4.96e-08) cellular bud neck contractile ring (5, 3.23e-07)

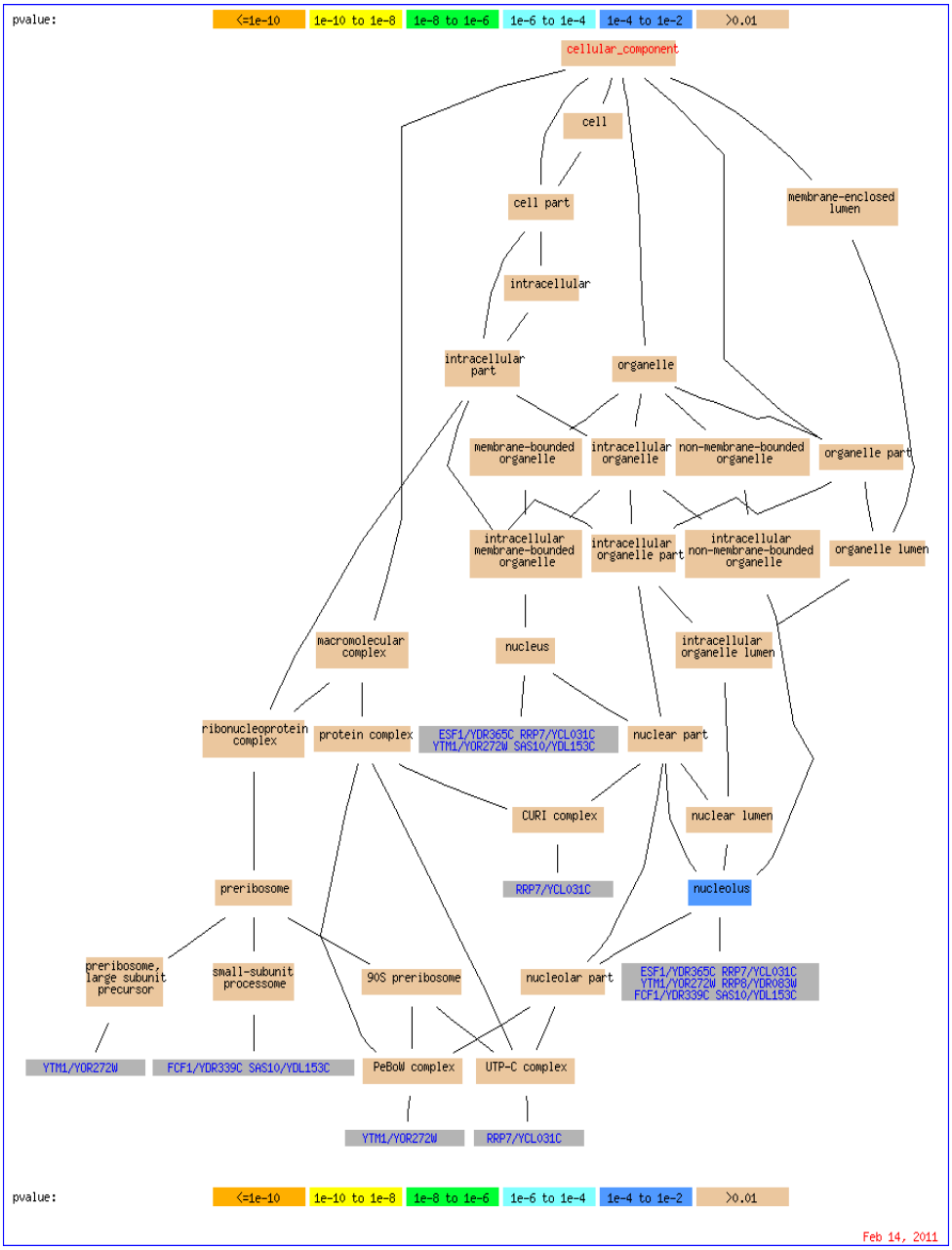


Figure 3.20 Sample of genes for the bicluster S51, with corresponding GO terms and their parents for Component Ontology

Figure 3.20 shows the significant GO terms for the set of genes in bicluster S51 along with their p values. It shows the branching of cellular component into sub-components like cell, cell part, membrane-enclosed lumen etc. These components are clustered using genes to produce the final result. Figure 3.20 is obtained when gene ontology database is searched by entering the names of genes of bicluster S51 and by selecting component ontology.

3.4.5 Comparison with other Algorithms

3.4.5.1 Comparison on the basis of Statistical and Biological Significance

In Table 3.28, the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the SGSC algorithm is compared with that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the table it is clear that in terms of the best p-value obtained by a bicluster which is used to denote statistical significance, SGSC algorithm is better than MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu for all the first, third, fourth and fifth GO terms. For the second GO term the p-value obtained is better than that of all the other algorithms except MOGAB. The percentage of genes involved is better than that of all the other algorithms for all the five GO terms.

Table 3.28
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by SGSC and other Algorithms for Yeast Data

Terms	SGSC	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 66.7% 1.30e-46	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Structural constituent of ribosome 66.7% 4.42e-44	Ribosomal subunit 53.46% 1.6e-45	Ribosome 46.21% 1.5e-25	Translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Cytosolic Part 66.7% 5.45e-43	Translation 57.14% 3.8e-41	Translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	Ribosome 71.4% 6.05e-41	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Translation 73% 3.12e-40	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

3.4.5.2 Comparison based on Bicluster Size and MSR

The Table 3.29 given below provides comparative summarization of the results of Yeast data involving the performance of related algorithms. The performance of the SGSC algorithm, in comparison with the performance of SEBI [36], SMOB [37], CC [29] and FLOC [106] are given for the Yeast dataset. In the SGSC algorithm presented here, only biclusters with row variance above 400 are taken into account, while calculating the average of mean squared residue, number of genes and conditions. For SGSC algorithm, the average MSR, average number of genes and conditions and the average volume, are better than that of SEBI an SMOB.

Table 3.29

Comparison between SGSC Algorithm and other Algorithms for Yeast Dataset

Algorithm	AMR	ANG	ANC	AV
SGSC	200.77	37.35	15.55	537.75
SEBI	205.18	13.61	15.25	209.92
SMOB	206.17	27.28	15.46	453.48
CC	204.29	166.71	12.09	1576.98
FLOC	187.54	195.00	12.80	1825.78

AMR is Average mean squared Residue. ANG is Average Gene Number of Genes. ANC is Average Number of Conditions. AV is Average Volume. As clear from the above Table the average MSR, the average number of genes and conditions, average volume are compared

for various algorithms. For the average MSR field lower values are better where as higher values are better for all other fields.

Table 3.30

Performance Comparison between SGSC and other Algorithms for the Human Lymphoma Dataset

Algorithm	AMR	ANG	ANC	AV
SGSC	1053.98	27.89	52.26	1169.63
SEBI	1028.84	14.07	43.57	615.84
SMOB	1019.60	11.60	78.47	709.13
CC	850.04	269.22	24.50	4595.98

AMR is Average mean squared Residue. ANG is Average Gene Number of Genes. ANC is Average Number of Conditions. AV is Average Volume. As is clear from the above Table the average mean squared residue, the average number of genes and conditions, average volume 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.

Table 3.30 gives performance comparison for Human B-cell Lymphoma dataset. Value of δ is set to 1200 for Lymphoma dataset. In this dataset the average number of genes and average volume of the biclusters obtained are better than that of SEBI and SMOB. Average number of conditions is greater than CC and SEBI.

In multi-objective evolutionary computation [4] the maximum number of conditions obtained is only 11 in Yeast dataset and 40 in

Human B-cell Lymphoma dataset. But in this method there are biclusters with all 17 and 91 conditions for Yeast and Lymphoma datasets respectively. Moreover as the SGSC algorithm uses simple sequential search rather than stochastic search the computation time required is very less compared to all the metahueristic and evolutionary algorithms.

This algorithm is capable of detecting some of the shifting and scaling patterns present in Yeast dataset. Some of the biclusters are with high row variance (more than 2000 for the Yeast dataset and more than 7000 for Lymphoma dataset.

3.5 Comparison of Constraint based Algorithms

3.5.1 Comparison based on p-value of GO terms for Biclusters Generated from Same Seeds

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. P-value indicates statistical significance of a bicluster. Four different seeds which on enlargement result in biologically significant biclusters were selected. These seeds are enlarged by all the constraint based algorithms and the p-values of the GO terms of these biclusters are compared for all these algorithms.

Table 3.31

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from First Seed and the Corresponding P-value Obtained for each Algorithm for Process Ontology

Go Terms	p-value and Percentage of Genes			
	MSRT	MSRDT	ISIMSRDT	SGSC
Ribosome biogenesis	8.41e-11 36.1%	4.78e-05 23.4%	3.08e-22 39.8%	0.00248 34.8%
Ribonucleoprotein complex biogenesis	1.47e-09 36.1%	7.68e-05 24.7%	6.25e-21 40.8%	0.00622 34.8%
Cellular component biogenesis at cellular level	1.08e-08 37.7%	0.00039 26.0%	1.68e-18 41.8%	---
ncRNA processing	2.95e-08 31.1%	0.00067 20.8%	1.86e-15 32.7%	0.00171 34.8%
ncRNA metabolic process	1.66e-07 31.1%	0.00247 16.9%	4.05e-14 32.7%	0.00352 34.8%
rRNA processing	5.98e-07 24.6%	0.00116 16.9%	5.80e-15 27.6%	0.00144(highest) 30.4%
RNA processing	8.40e-07 32.8%	0.00209 23.4%	2.74e-12 33.7%	-------
rRNA metabolic process	1.14e-06 24.6%	0.00194 16.9%	2.06e-14 27.6%	0.00194 30.4%
RNA metabolic process	2.09e-05 45.9%	0.00832 36.4%	3.08e-14 53.1%	0.00184 56.5%

In this case the order of algorithm based on p-value is ISIMSRDT, MSRT, MSRDT and SGSC for all GO terms. The percentage of genes

involved is the highest for the ISIMSRDT algorithm for the first three GO terms. But the percentage of genes involved for SGSC is better for GO terms starting from the fourth entry of the Table 3.31, that is, from ncRNA processing to RNA metabolic process. The p-values obtained for SGSC is very low. Because the difference threshold value assigned for the genes is very low, there are only 23 genes in the bicluster. By increasing this value more genes will be included and this will increase the p-value of GO terms for SGSC algorithm.

Table 3.32

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the First Seed and the Corresponding p-value Obtained for each Algorithm for the Function Ontology

GO Terms	MSRT	MSRDT	ISIMSRDT	SGSC
Number of genes annotated to the term molecular function unknown	27 genes	32 genes	0.00480 (p-value) snoRNA binding	10 out of 23genes

From the table it is clear that, for function ontology a fixed number of genes are annotated to the term molecular function unknown for all algorithms except ISIMSRDT. For ISIMSRDT algorithm 4 genes from the bicluster are annotated to the term snoRNA binding and the p-value is 0.0048.

Table 3.33

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the First Seed and the Corresponding p-value Obtained for each Algorithm for the Component Ontology

GO terms	p-value and Percentage of Genes			
	MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**
Nucleolus	2.91e-11 31.1%	8.24e-05 18.2%	2.56e-19 31.6%	0.00622 26.1%
Preribosome	8.40e-10 24.6%	0.00156 13%	4.26e-15 23.5%	--
90S Preribosome	7.50e-09 19.7%	0.00210 10.4%	1.56e-09 15.3%	--
Nuclear part	1.79e-06 47.5%	--	2.46e-12 50.0%	--
Nuclear lumen	3.56e-06 39.3%	--	1.59e-13 43.9%	--
Organelle lumen	1.04e-05 42.6%	--	6.41e-11 44.9%	--
Intracellular organelle lumen	1.04e-05 42.6%	--	6.41e-11 44.9%	--
Ribonucleoprotein complex	9.71e-05 32.8%	--	4.24e-07 31.6%	--
Nucleus	0.00020 59.0%	--	1.62e-08 61.2%	--
Nucleolar part	0.00071 11.5%	--	2.43e-06 11.2%	--
Macromoleular complex	0.00179 54.1%	--	9.23e-07 56.1%	--
Smallsubunit processome	--	--	1.80e-05 9.2%	--
Organelle part	--	--	0.00081 58.2%	--
Intracellular organelle part	--	--	0.00081 58.2%	--

In this case the order of algorithms based on best p-value is ISIMSRDT, MSRT, MSRDT, and SGSC. Since there are only 23 genes in the SGSC algorithm there is only one GO term associated with it for the component ontology. Even though the p-value is less for SGSC, the

percentage of genes involved is greater than MSRDT for the first GO term. The percentage of genes involved for the MSRT algorithm is greater than that of ISIMSRDT for GO the terms preribosome, 90S preribosome, ribonucleoprotein complex and nucleolar part.

Table 3.34

Comparison of Constraint based algorithms based on GO terms for biclusters generated from second seed and the corresponding p-value obtained for each algorithm for the Process Ontology

GO terms	p-value and the Percentage of Genes			
	MSRT	MSRDT	ISIMSRDT	SGSC
Translation	7.82e-25 60.7%	2.26e-23 54.7%	2.03e-49 63.3%	3.12e-40 73%
Cellularprotein metabolic process	3.25e-12 64.3%	2.88e-11 59.4%	3.08e-24 66.3%	9.61e-24 17.7%
Protein metabolic process	8.24e-12 64.3%	7.49e-11 59.4%	1.77e-23 66.3%	3.96e-23 77.8%
Cellular macromolecule biosynthetic process	5.82e-10 62.5%	1.42e-08 56.2%	6.74e-19 63.3%	8.59e-18 73.0%
Macromolecule biosynthetic process	6.47e-10 62.5%	1.58e-08 56.2	8.19e-19 63.3	1.00e-17 73%
Gene expression	1.09e-08 62.5%	5.29e-08 57.8%	2.12e-17 64.3%	5.78e-17 74.6%
Translational elongation	2.35e-08 16.1%	2.66e-09 15.6%	4.60e-17 16.3%	1.78e-09 15.9%
Cellular biosynthetic process	3.41e-07 64.3%	1.15e-07 62.5%	1.39e-15 67.3%	--
Biosynthetic process	6.64e-07 64.3%	2.42e-07 62.5%	5.05e-15 67.3%	3.10e-14 76.2%
Ribosome biogenesis	4.34e-05 26.8%	1.26e-06 28.1	6.01e-15 32.7%	1.25e-11 36.5%
rRNA processing	0.00010 21.4%	9.86e-06 21.9%	5.61e-10 22.4%	7.25e-08 25.4%
rRNA metabolic process	0.00017 21.4%	1.77e-05 21.9%	1.48e-09 22.4%	1.45e-07 25.4
Cellular macromolecule metabolic process	0.00025 67.9%	0.00024 65.6%	9.83e-10 70.4%	2.90e-11 81.0%

In this case the best p-values are obtained in the order ISIMSRDT, SGSC, MSRT and MSRDT respectively. But the order of algorithms based on the percentage of genes for the first GO terms is SGSC, ISIMSRDT, MSRT and MSRDT. For all GO terms, except cellular protein metabolic process and translational elongation, the percentage of genes involved in SGSC algorithm is better than that of all the other algorithms.

Table 3.35

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Second Seed and the Corresponding p-value Obtained for each Algorithm for the Function Ontology

GO Terms	p-value and Percentage of Genes			
	MSRT	MSRDT	ISIMSRDT	SGSC
Structural constituent of ribosome	9.79e-24 50%	2.58e-24 46.9%	6.05e-53 56.1%	4.42e-44 66.7%
Structural molecule activity	2.73e-18 50%	1.79e-18 46.9%	3.97e-42 57.1%	3.48e-35 66.7%
Translation elongation factor activity	0.00015 7.1%	0.00035 6.2%	7.16e-05 5.1%	--
RNA-directed DNA polymerase activity	--	--	--	--
RNA binding	-	-	0.00208	--
Translation elongation factor activity	-	-	-	-
DNA-directed DNA polymerase activity	-	-	-	-
DNA polymerase activity	-	-	-	-

In this case the best p-values are obtained in the order ISIMSRDT, SGSC, MSRDT and MSRT respectively. But the order of algorithms

based on percentage of genes is SGSC, ISIMSRDT, MSRT, and MSRDT for the first two GO terms. For the third GO term the order of algorithms based on the percentage of genes is MSRT, MSRDT and ISIMSRDT.

Table 3.36

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Second Seed and the Corresponding p-value Obtained for each Algorithm for the Component Ontology

GO Terms	p-value and the Percentage of Genes			
	MSRT	MSRDT	ISIMSRDT	SGSC
Cytosolic ribosome	1.55e-26 51.8%	2.71e-27 48.4%	1.51e-60 58.2%	1.30e-46 66.7%
Cytosolic part	2.95e-24 51.8%	7.66e-25 48.4%	3.92e-55 58.2%	5.45e-43 66.7%
Ribosome	8.24e-24 57.1%	6.60e-24 53.1%	3.60e-51 62.2%	6.05e-41 71.4%
Cytosol	1.36e-20 55.4%	3.91e-23 54.7%	1.42e-48 63.3%	1.45e-36 69.8%
Ribonucleoprotein complex	1.11e-18 60.7%	3.21e-18 56.2%	3.31e-38 64.3%	1.04e-32 74.6%
Cytosolic small ribosomal subunit	-	-	9.31e-28 27.6%	2.44e-19 30.2%
Cytosolic large ribosomal subunit	2.09e-17 32.1	3.59e-16 28.1%	4.42e-27 28.6%	1.84e-24 36.5%
Large ribosomal subunit	7.99e-15 32.1	1.30e-13 28.1%	1.45e-22 28.6%	6.11e-21 36.5%
Non-membrane-bounded organelle	1.23e-10 62.5%	1.50e-10 59.4%	1.12e-21 65.3%	1.12e-20 76.2%
Intracellular non-membrane-bounded organelle	1.23e-10 62.5%	1.50e-10 59.4%	1.12e-21 65.3%	1.12e-20 76.2%

In this case the best p-values are obtained in the order ISIMSRDT, SGSC, MSRDT and MSRT respectively, for the first five GO terms. But based on the percentage of genes involved, the order of algorithms are SGSC, ISIMSRDT, MSRT and MSRDT for the first five GO terms.

Percentage of genes involved is highest for SGSC for all GO terms. P-value obtained is the best for ISIMSRDT for all GO terms.

Table 3.37

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Third Seed and the Corresponding p-value Obtained for each Algorithm for the Process Ontology

GO Terms	p-value and the Percentage of Genes			
	MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**
DNA repair	4.82e-13 57.1%	1.43e-14 60.7%	3.25e-10 45.5%	4.68e-12 51.6%
Response to DNA damage stimulus	5.57e-12 57.1%	1.97e-13 60.7%	3.04e-09 45.5%	5.29e-11 51.6%
DNA metabolic process	4.37e-11 60.7%	7.23e-14 67.9%	1.11e-10 (highest) 54.5%	2.50e-11 58.1%
Cell cycle	8.19e-07 53.6%	-	1.13e-09 57.6%	4.87e-08 54.8%
Cell cycle process	5.15e-06 50%	-	7.20e-09 54.5%	2.99e-07 51.6%
Double-strand break repair	2.91e-07 32.1%	-	1.53e-06 27.3%	8.98e-07 29%
Cellular response to stress	5.99e-10 60.7%	6.03e-10 60.7%	2.60e-07 48.5%	8.51e-08 51.6%
Response to stress	2.28e-08 60.7	2.30e-08 60.7%	6.89e-06 48.5%	2.35e-06 51.6%
Mitotic sister chromatid cohesion	-	3.76e-05 21.4%	8.67e-08 24.2%	2.31e-06 22.6%
Cellular response to stimulus	2.29e-08 64.3%	2.31e-08 64.3%	8.61e-06 51.5%	2.58e-06 54.8%
Cell cycle phase	1.66e-05 42.9%	1.67e-05 42.9%	1.36e-07 45.5%	7.11e-06 41.9%
M phase	0.00021 35.7%	1.85e-05 39.3%	1.14e-06 39.4%	6.26e-06 38.7%
Chromosome organization	0.00018 39.3%	0.00158 35.7%	1.68e-06 42.4%	7.07e-05 38.7%

In this case, the order of algorithms based on p-value is MSRDT, MSRT, SGSC, and ISIMSRDT for most of the GO terms. The order of algorithms based on percentage of genes involved is MSRDT, MSRT, SGSC and ISIMSRDT for most of the GO terms.

Table 3.38

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Third Seed and the Corresponding p-value Obtained for each Algorithm for the Function Ontology

GO Terms	p-values and the Percentage of Genes			
	MSRT	MSRDT	ISIMSRDT	SGSC
Double-stranded DNA binding	4.13e-05 17.9%	4.58e-05 17.9%	0.00202 12.1%	8.01e-05 16.1%
Structure-specific DNA binding	0.00103 17.9%	0.00115 17.9%	0.00172 15.2%	0.00198 16.1%
DNA secondary structure binding	0.00104 10.7%	0.00116 10.7%	-	0.00162 9.7%
Guanine/thymine mispair binding	0.00335 7.1%	0.00372 7.1%	-	0.00469 6.5%
Single base insertion or deletion binding	0.00335 7.1%	0.00372 7.1%	-	0.00469 6.5%
Four-way junction DNA binding	0.00999 7.1%	-	-	-

In this case also the order of algorithms based on p-value is MSRT, MSRDT, SGSC, and ISIMSRDT for the first GO term. For the second Go term, the order of algorithms based on p-value is MSRT, MSRDT, ISIMSRDT and SGSC. The order of algorithms based on the percentage of genes involved is MSRT, MSRDT, SGSC and ISIMSRDT for the first two GO terms.

Table 3.39

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Third Seed and the Corresponding p-value Obtained for each Algorithm for the Component Ontology

GO Terms	p-value and the Percentage of Genes			
	MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**
Replication fork	1.38e-07 28.6%	3.42e-09 32.1%	6.19e-07 24.2%	9.43e-06 22.6%
Chromosome	2.01e-08 (highest) 50.0%	1.85e-08 50%	2.01e-09 (highest) 48.5%	8.04e-09 (highest) 48.4%
Chromosomal part	1.53e-06 42.9%	1.41e-06 42.9%	4.2e-07 42.4%	5.29e-07 41.9%
Nuclear chromosome	6.59e-06 39.3%	6.07e-06 39.3%	4.50e-07 39.4%	2.18e-05 35.5%
Nuclear replication fork	3.39e-05 21%	1.10e-06 25%	0.00010 18.2%	0.00146 16.1%
Nuclear chromosome part |	0.00036 32.1%	0.00033 32.1	2.23e-05 33.3%	0.00089 29%
Condensed nuclear chromosome	0.00594 17.9%	0.00546 17.9%	3.65e-06 24.2%	4.34e-05 22.6%
Mitotic cohesin complex	-	-	7.89e-07 12.1%	5.95e-07 12.9%
Nuclear mitotic cohesin complex	-	-	7.89e-07 12.1%	5.95e-07 12.9%
Nucleus	-	8.87e-06 78.6%	3.39e-05 72.2%	3.19e-05 74.2%
Condensed chromosome	-	.00925 17.9%	8.86e-06 24.2%	5.10e-06 25.8%
Nuclear cohesin complex	-	-	3.91e-06 12.1%	2.95e-06 12.9%
Cohesin complex	-	-	3.91e-06 12.1%	2.95e-06 12.9%

In this case the order of algorithms based on best p-value is ISIMSRDT, MSRDT, SGSC and MSRT.

Table 3.40

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Fourth Seed and the Corresponding p-value Obtained for each Algorithm for Process Ontology

GO Terms, p-value and Percentage of Genes of GO Terms for each Algorithm			
MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**
Cytokinesis 0.00130 20.6%	Cytokinesis 2.32e-05 28.6%	Cytokinesis 7.07e-05 24.2%	Cytokinesis 6.87e-05 24.2%
Positive regulation of spindle pole body separation 0.00195 8.8%	Cell cycle process 3.91e-05 46.4%	Cell division 0.00043 24.2	Cell cycle process 0.00024 39.4%
Cell cycle process 0.00252 35.3%	Cell cycle 6.36e-05 46.4%	Cell cycle cytokinesis 0.00130 18.2%	Cell cycle 0.00039 39.4%
Cell cycle 0.00383 35.3%	Cell division 0.00014 28.6%	Cell cycle process 0.00171 36.4%	Cell division 0.00042 24.2%
Regulation of spindle pole body separation 0.00387 8.8%	Cell cycle cytokinesis 0.00058 21.8%	Positive regulation of spindle pole body separation 0.00173 9.1%	Cell cycle cytokinesis 0.00126 18.2%
Cell division 0.00607 20.6%	Protein phosphorylation 0.00077 25.0%	Protein phosphorylation 0.00196 21.2%	Positive regulation of spindle pole body separation 0.00168 9.1%
--	Positive regulation of spindle pole body separation 0.00118 10.7%	Cell cycle 0.00261 36.4%	Cytokinetc process | 0.0025718.2%
--	Cytokinetic process 0.00120 21.4%	Cytokinetic process 0.00265 18.2%	Regulation of spindle pole body separation 0.00334 9.1%

--	Regulation of spindle pole body separation 0.00235 10.7%	Regulation of spindle pole body separation 0.00344 9.1%	Phosphorylation 0.00946 21.2%
--	Phosphorylation 0.00422 25%	-	-
--	Spindle pole body separation 0.00967 10.7%	-	-

In the biclusters obtained by the fourth seed, since the conditions selected are different for each algorithm, the genes selected are also different. The GO terms are different for biclusters obtained by each algorithm. Hence GO terms along with the p-values are given in the order of p-values. Here the order of algorithms in terms of best p-value and percentage of genes is MSRDT, SGSC, ISIMSRDT and MSRT for the first GO term cytokinesis.

Table 3.41

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Fourth Seed and the Corresponding p-value Obtained for each Algorithm for Function Ontology

GO Terms	MSRT	MSRDT	ISIMSRDT	SGSC
'molecular function unknown'	13 out of 34 input genes	11 out of 28 genes	12 out of 33 genes	13 out of 33 genes

From the table it is clear that for function ontology a fixed number of genes are annotated to the term molecular function unknown for all algorithms.

Table 3.42

Comparison of Constraint based Algorithms based on GO Terms for Biclusters Generated from the Fourth Seed and the Corresponding p-value Obtained for each Algorithm for Component Ontology

GO Terms, p-value and Percentage of Genes of GO Terms for each Algorithm			
MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**
Cellular bud 3.48e-06 29.4%	Cellular bud neck 1.06e-09 39.3%	Cellular bud 3.90e-10 39.4%	Cellular bud 3.41e-10 39.4%
Cellular bud neck 3.81e-06 26.5%	Cellular bud 1.12e-09 42.9%	Cellular bud neck 6.47e-09 33.3%	Cellular bud neck 6.47e-09 33.3%
Site of polarized growth 1.63e-05 29.4%	Site of polarized growth 7.76e-09 42.9%	Site of polarized growth 5.50e-08 36.4%	Site of polarized growth 4.96e-08 36.4%
Cellular bud neck contractile ring 5.04e-05 11.8%	Cellular bud neck contractile ring 1.44e-07 17.9%	Cellular bud neck contractile ring 3.23e-07 15.2%	Cellular bud neck contractile ring 3.23e-07 15.2%
Actomyosin contractile ring 5.04e-05 11.8%	Actomyosin contractile ring 1.44e-07 17.9%	Actomyosin contractile ring 3.23e-07 15.2%	Actomyosin contractile ring 3.23e-07 15.2%
Contractile ring 5.04e-05 11.8%	Contractile ring 1.44e-07 17.9%	Contractile ring 3.23e-07 15.2%	Contractile ring 3.23e-07 15.2%
Cell division site 0.00069 11.8%	Cytoskeletal part 1.63e-06 35.7%	Cytoskeleton part 7.71e-06 30.3%	Cytoskeleton part 7.71e-06 30.3%

Cell division site part 0.00069 11.8%	Cytoskeleton 1.78e-06 35.7%	Cytoskeleton 8.40e-06 30.3%	Cytoskeleton 8.40e-06 30.3%
Cytoskeleton part 0.00126 23.5%	Cell division site 4.97e-06 17.9%	Cell division site 1.10e-05 15.2%	Cell division site 1.10e-05 15.2%
Cytoskeleton 0.00135 23.5%	Cell division site part 4.97e-06 17.9%	Cell division site part 1.10e-05 15.2%	Cell division site part 1.10e-05 15.2%
Actin cytoskeleton 0.00222 14.7%	Actin cytoskeleton 3.40e-05 21.4%	Actin cytoskeleton 8.63e-05 18.2%	Actin cytoskeleton 8.63e-05 18.2%
-	Cell cortex part .00050 21.4%	Cell Cortex part .00123 18.2%	Cell Cortex part .00099 18.2%
-	Cell cortex 0.00182 21.4%	Cell cortex .00444 18.2%	Cell cortex .00099 18.2%

In this case the order of algorithms based on best p-value is SGSC, ISIMSRDT, MSRDT and MSRT.

In short, from these results it is not possible to conclude that a single algorithm is best in terms of p-value. The order is changing for each bicluster and in some situation for a particular ontology. But in most cases ISIMSRDT algorithm is best among the four constraint based algorithms in terms of p-

value. And in most cases SGSC algorithm is the best among the four constraint based algorithms in terms of the percentage of the genes involved.

3.5.2 Comparison based on the best five GO terms

Table 3.43

Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by Constraint Based Algorithms for the Yeast Dataset

Terms	MSRT	MSRDT	ISIMSRDT	SGSC
1	Cytosolic ribosome 51.8% 1.55e-26	Cytosolic ribosome 48.4% 2.71e-27	Cytosolic ribosome 58.2% 1.51e-60	Cytosolic ribosome 66.7% 1.30e-46
2	Translation 60.7% 7.82e-25	Cytosolic Part 48.4% 7.66e-25)	Cytosolic Part 58.2% 3.92e-55	Structural constituent of ribosome 66.7% 4.42e-44
3	Cytosolic Part 51.8% 2.95e-24	Structural constituent of ribosome 46.9% 2.58e-24)	Structural constituent of ribosome 56.1% 6.05e-53	Cytosolic Part 66.7% 5.45e-43
4	Ribosome 57.1% 8.24e -24	Ribosome 53.1% 6.60e-24	Ribosome 62.2% 3.60e-51	Ribosome 71.4% 6.05e-41
5	Structural constituent of ribosome 50% 9.79e -24	Structural Molecule Activity 46.9% 2.58e-24	Translation 62% 2.03e-49	Translation 73% 3.12e-40

Here all the algorithms are compared on the basis of the best 5 p-values obtained from all four biclusters. In this case the order of

algorithms based on p-value is ISIMSRDT, SGSC, MSRDT and MSRT for all GO terms. But the order of algorithms based on the percentage of genes for the first GO term is SGSC, ISIMSRDT, MSRT and MSRDT.

3.5.3 Comparison based on Size and MSR for Biclusters Generated from the Same Seed

For this comparison three different seeds are selected. These seeds are enlarged by all the constraint based algorithms. The size and MSR are compared for biclusters obtained from all these algorithms.

Analysing the algorithms based on the biclusters obtained from the same seed it should be noted that among the algorithms SGSC produces biclusters of low size but coherence is high since MSR value is very low. ISIMSRDT is the best among the four constraint based algorithms in terms of bicluster size. Reducing the incrementing factor in ISIMSRDT can improve the bicluster size further.

For the first seed, the order of algorithms in terms of bicluster size is ISIMSRDT, MSRDT, MSRT and SGSC. For the second seed the order of algorithms in terms of bicluster size is ISIMSRDT, MSRDT, SGSC, MSRT. For the third seed the order is, ISIMSRDT, MSRDT, SGSC, MSRT. In short ISIMSRDT, MSRDT, MSRT, SGSC are the order of algorithms in terms of bicluster size. The order of MSRT and SGSC changes for different biclusters depending on the value selected for the difference threshold. The row variance of biclusters obtained by MSRDT is greater than that of MSRT and ISIMSRDT in all the three cases.

Table 3.44

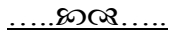
Comparison of Size and MSR of Three Biclusters obtained by Enlarging Three Different Seeds by each one of the Constraint based Algorithms

Sl.No	MSRT			MSRDT			ISIMSRDT			SGSC		
	Size	MSR	Row variance	size	MSR	Row variance	size	MSR	Row variance	size	MSR	Row variance
1	61*17	198.95	469.4058	77*16	199.54	533.1660	98*17	199.97	482.77	23*17	131.39	506.7582
2	56*17	199.78	587.8461	64*17	199.32	654.5732	98*17	199.99	600.91	63*17	167.43	615.9798
3	28*17	299.85	1937.5	28*17	286.34	2034.1	33*17	299.22	1970.1	31*17	297.19	2036

3.6 Summary

In this chapter, four constraint based algorithms developed for finding the biclusters from gene expression data for enlarging the seeds, are described. More genes and conditions are added to the seeds in which node addition follows node deletion, if necessary. Nodes are searched sequentially. The algorithms are implemented on both the Yeast Sacharomyces cerevisiae cell cycle expression dataset and Human Lymphoma dataset. A comparative assessment of the results is provided on both the above mentioned benchmark gene expression datasets in order to demonstrate the effectiveness of the proposed methods. The quality of biclusters obtained can be inspected visually by using bicluster plots. The expression values of genes in the bicluster show strikingly similar up-regulation and down-regulation under a set of experimental conditions. These algorithms are able to identify interesting biclusters from gene expression data. In the Yeast dataset MSRT and SGSC algorithms can identify some biclusters with shifting and scaling patterns; and some of the biclusters are with high very high row variance. Statistical significance and biological relevance of the biclusters obtained by each algorithm are also verified using gene ontology database. In terms of the best p-value obtained by biclusters, these algorithms are better than algorithms like SGAB, CC, RWB, Bimax, OPSM**,** ISA and Bivisu. A bicluster with the highest number of conditions (92) is obtained for Lymphoma dataset for ISIMSRDT algorithm. The row variance of this bicluster is also very high (above 5000). Another major research finding is in the case of iterative search. Iterative search has got the advantage of

selecting the (n-k)th gene or condition whose incremental increase in MSR value got reduced after adding the nth gene or condition. Comparisons of all the constrained based algorithms with other algorithms on the basis of statistical significance, size and MSR value of the biclusters are given in this chapter. Constrained based algorithms are also compared among themselves based on the quality of the biclusters obtained from the same seed.



Chapter 4

Greedy Algorithm

Chapter 4 describes the Greedy algorithm. The description of algorithm, its complexity, different biclusters obtained from Yeast and Lymphoma datasets, significant biclusters obtained (biological validation), and the comparison of the algorithm with other algorithms are given in this chapter.

4.1 Description of the Algorithm

A greedy algorithm is any algorithm that follows the problem solving strategy of making the locally optimal choice at each stage [31] with the hope of finding the global optimum. In general greedy algorithms are used for optimization problems. Biclustering is an optimization problem in which the objective is to maximize the volume and minimize the MSR. The seeds obtained from K-Means clustering algorithm are thus enlarged using greedy approach. In the seed growing phase a separate list is maintained for conditions and genes not included in the bicluster. Each seed is enlarged separately by adding more genes and conditions. Initially conditions are added followed by genes. In greedy search algorithm, the best element is selected from the gene list or condition list and added to the bicluster. The quality of the element is determined by the Hscore or MSR value of the bicluster after including the element in the bicluster. The element which results in minimum MSR value when added to the bicluster is considered as the best element. It cannot be specified as an element with smallest incremental cost of Hscore because adding some elements reduces the Hscore value. Seed growing starts from condition list followed by gene list until the MSR value reaches the given threshold. This is a greedy method since our aim is to select the next element which produces bicluster with minimum Hscore value. A pseudo-code description of the greedy search algorithm is given below.

```
Algorithm  greedysearch(seed, δ)
bicluster := seed
Calculate Column_List the list of conditions not included in the bicluster
While (MSR(bicluster) <= δ)
  No_elem_Col=size(Column_List)
          for  i:=1: No_elem_Col
                bicluster=bicluster+ Column_List [i]
                Column_List_msr[i]= MSR(bicluster)
                Remove Column_List[i] from bicluster
          end(for)
find minimum value in Column_List_msr  and corresponding index K
bicluster=bicluster+ Column_List [K]
delete Column_List [K] from Column_List
end(while)

Calculate Row_List the list of genes not included in the bicluster
While (MSR(bicluster) <= δ)
  No_elem_Row=size(Row_List)

             for  i:=1: No_elem_Row
                   bicluster=bicluster+ Row_List [i]
                   Row_List_msr[i]= MSR(bicluster)
                   Remove Row_List[i] from bicluster
              end(for)
find minimum value in Row_List_msr  and corresponding index J
bicluster=bicluster+ Row_List [J]
delete Row_List [J] from Row_List
end(while)
end(greedysearch)
```

4.2 Time Complexity

The basic operation for the identification of biclusters is the calculation of mean squared residue of a submatrix. Time complexity for calculating MSR is O(mn). In this algorithm conditions are added first followed by genes. In order to include a condition MSR value of all submatrices that result from adding a single condition is to be calculated for all conditions. This number decreases by one, after each iteration. That means the complexity can be calculated by the formula (n+(n-1)+(n-2)+ ...1). This is equal to n(n+1)/2 which is equivalent to $O(n^2)$. Hence for adding conditions the worst case complexity is $O(mn)(n^2)$. Similarly for adding genes the worst case complexity is $O(mn)(m^2)$. Hence the worst case complexity for adding genes and conditions is $O(mn)(m^2+n^2)$ where m and n are the number of genes and conditions respectively.

4.3 Experimental Results

4.3.1 Bicluster Plots for Yeast Dataset

In Figure 4.1, nine biclusters identified by the greedy algorithm on the Yeast dataset are shown. From the bicluster plots it can be noticed that genes present a similar behaviour under a set of conditions. Many of the biclusters found on the Yeast dataset contain all 17 conditions. Out of the nine biclusters shown in Figure 4.1, seven contain all 17 conditions and they differ in appearance. In short, greedy algorithm is ideal for identifying various biclusters with coherent values. Information about these biclusters is given in Table 4.1. All the biclusters are having mean squared residue less than 300.

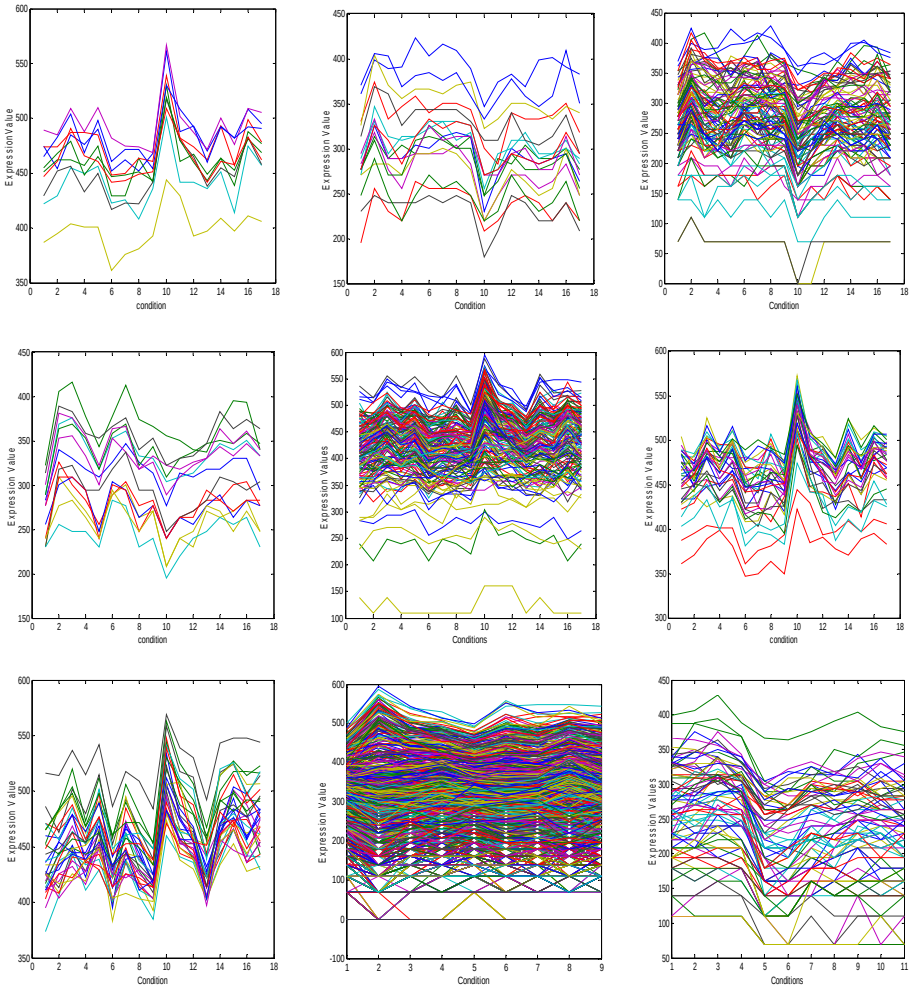


Figure 4.1 Nine biclusters obtained from the Yeast dataset using greedy algorithm. Bicluster labels are (ya6), (yb6), (yc6), (yd6), (ye6), (yf6), (yg6), (yh6) and (yi6) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about the biclusters can be obtained from Table 4.1 using bicluster label. Here only biclusters with different shapes are selected.

Table 4.1

Information about Biclusters of Figure 4.1

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(ya6)	10	17	170	66.4403	522.23
(yb6)	17	17	289	99.3497	407.47
(yc6)	108	17	1836	194.5204	472.34
(yd6)	14	17	238	97.8389	507.63
(ye6)	147	17	2499	200.2474	396.04
(yf6)	33	17	561	99.9639	506.14
(yg6)	31	17	527	97.9121	613.89
(yh6)	1405	9	12645	299.8968	348.07
(yi6)	79	11	869	241.3371	760.91

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 number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residues of the biclusters. The last column contains the row variance of the biclusters.

4.3.2 Bicluster Plots for Lymphoma Dataset

Eight biclusters obtained from Human Lymphoma dataset are shown in Figure 4.2. All the biclusters show strikingly similar up-regulation and down-regulation. All the means squared residues are lower than 1200. The first bicluster in Figure 4.2 contains 94 conditions. Number of genes in this bicluster is 11. The row variance of the bicluster is also very high (5317.5).

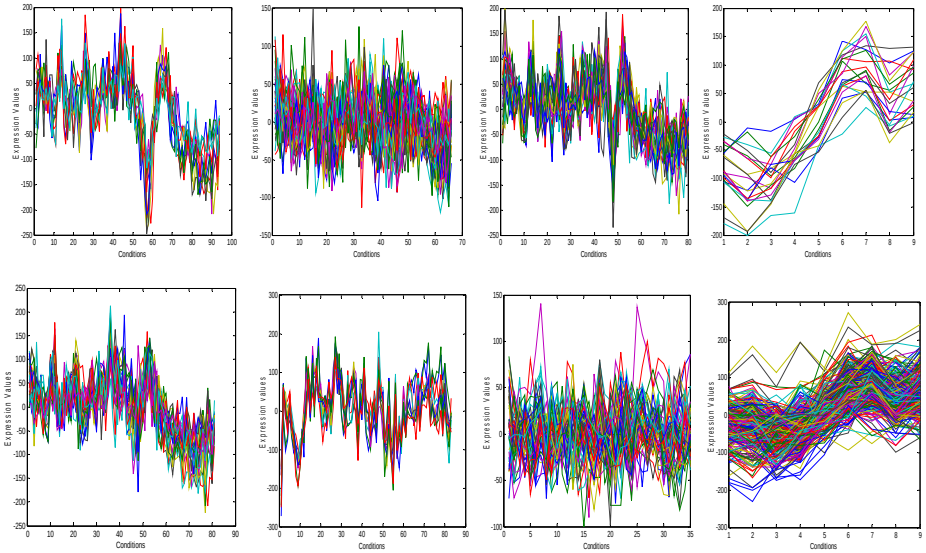


Figure 4.2 Eight biclusters found for the Lymphoma Dataset using greedy algorithm. Bicluster labels are (la6), (lb6), (lc6), (ld6), (le6), (lf6), (lg6) and (lh6) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 4.2 using bicluster label.

Table 4.2

Information about Biclusters of Figure 4.2

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR	Row Variance
(la6)	11	94	1034	1194.40	5317.5
(lb6)	40	66	2640	918.25	1156.4
(lc6)	30	80	2400	1175.90	3466.3
(ld6)	21	9	189	476.12	6183.5
(le6)	26	81	2106	1196.80	3906.0
(lf6)	10	83	830	1182.10	5070.1
(lg6)	53	35	1855	723.41	788.7
(lh6)	292	9	2628	1196.90	3359.1

4.4 Advantages of Greedy Algorithm

The advantage of this Greedy approach over the previous greedy approach of Cheng and Church [29] is that it avoids random interference. In the greedy method of Cheng and Church the program starts with the entire gene expression data matrix and deletes those rows or columns whose removal creates the greatest variation in MSR. This method is deterministic. So in order to identify different biclusters, the identified ones are replaced by random values. These random values will interfere with the discovery of future biclusters. This problem is known as random interference. This has the obvious effect of precluding the identification of biclusters with significant overlaps. Moreover mean squared residue is biased towards biclusters of low row variance [24]. Since seeds from K-Means are used, it can identify biclusters with high row variance without using row variance as a measure for optimization. Biclustering is a combinatorial optimization problem. Seeds from K-Means reduce the number of combinations.

4.5 Details of Significant Biclusters obtained by Greedy Algorithm

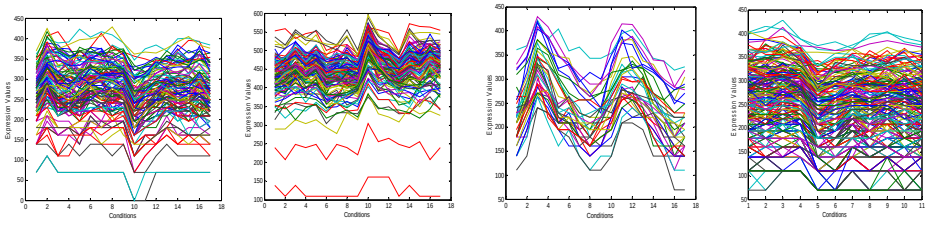


Figure 4.3 Four significant biclusters obtained by the greedy algorithm on Yeast dataset. The bicluster labels are s61, s62, s63, and s64. The details about biclusters can be obtained from Table 4.3 using bicluster label.

Table 4.3

Information about Biclusters of Figure 4.3

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
S61	121	17	199.9395	483.2784
S62	107	17	199.4776	568.0833
S63	36	17	297.6071	1806.9000
S64	224	11	209. 6618	455.5141

In the first bicluster s61 there are 121 genes. They are YBL014C, YBL083C, YBL084C, YBR293W, YCL016C, YCL031C, YCL053C, YCL054W, YCR072C, YCR087W, YDL008W, YDL030W, YDL076C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL215C, YDL231C, YDR017C, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR120C, YDR121W, YDR170C, YDR172W, YDR211W, YDR234W, YDR262W, YDR289C, YDR299W, YDR312W, YDR321W, YDR339C, YDR352W, YDR361C, YDR365C, YDR392W, YDR416W, YDR449C, YDR469W, YDR477W, YDR478W, YDR518W, YDR524C, YDR542W, YEL015W, YEL055C, YER005W, YER075C, YER099C, YER107C, YER166W, YER168C, YER171W, YFL001W, YGL085W, YGL099W, YGL214W, YGR042W, YGR090W, YGR187C, YGR200C, YGR216C, YHR062C, YJL011C, YJL069C, YJR017C, YJR066W, YKR056W, YKR060W, YLL008W, YLL034C, YLR051C, YLR088W, YLR107W, YLR146C, YLR215C, YLR222C, YLR227C, YLR401C, YML066C, YML080W, YML093W, YMR093W, YMR211W, YMR235C, YNL041C, YNL132W, YNL163C, YNL164C, YNL199C, YNL227C, YNL299W, YNR003C, YNR038W, YOL021C, YOL022C, YOL036W, YOL080C, YOL124C, YOL140W, YOL144W, YOR006C, YOR056C, YOR061W, YOR098C, YOR145C, YOR160W, YOR252W, YOR272W, YPL126W, YPL268W, YPR053C, YPR112C.

In bicluster s62 there are 107 genes namely YAL003W, YAL038W, YAR020C, YBL030C, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A,YBR106W, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR154C, YDR353W, YDR382W, YDR385W, YDR417C, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER117W, YGL102C, YGR118W, YHR141C, YJL136C, YJL188C, YJL189W, YJL190C, YJR009C, YJR094W-A, YJR123W, YKL056C, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL066C, YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR333C, YLR340W, YLR388W, YLR406C, YLR441C, YLR467W, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR045C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W,YPL090C,YPL143W, YPL283C, YPR102C,YPR204W.

In the third bicluster s63 there are 36 genes. They are YAR007C, YAR008W, YBL035C, YBR073W, YBR088C, YBR089W, YCR065W, YDL003W, YDL010W, YDL018C, YDL164C, YDR097C, YDR507C, YER095W, YFL008W, YGR151C, YGR152C, YHR154W, YIL026C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR386W, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL303W, YNL312W, YOR074C, YPL208W, YPR120C

In the fourth bicluster s64 only 224 genes are selected. They are YAL041W, YAL059W, YAR015W, YAR061W, YBL004W, YBL005W, YBL014C, YBL018C, YBL024W, YBL026W, YBL042C, YBL049W, YBL083C, YBL084C, YBR021W, YBR032W, YBR038W, YBR050C,YBR076W,YBR084W, YBR123C,

YBR133C, YBR138C, YBR155W, YBR228W, YBR257W, YBR267W, YBR293W, YCL012W, YCL016C, YCL031C, YCL054W, YCR036W, YCR043C, YCR051W, YCR062W, YCR063W, YCR072C,YCR081W, YCRX16C, YDL030W, YDL043C, YDL076C, YDL113C, YDL150W, YDL160C, YDL167C, YDL215C,YDL247W, YDR011W, YDR017C, YDR038C, YDR060W, YDR080W, YDR091C, YDR108W, YDR120C, YDR150W, YDR151C, YDR170C, YDR184C, YDR198C, YDR207C, YDR214W, YDR234W, YDR272W, YDR275W, YDR282C, YDR299W, YDR311W, YDR324C, YDR361C, YDR363W, YDR364C, YDR449C, YEL015W, YEL043W, YEL053C, YEL055C, YEL057C, YER005W, YER034W, YER064C, YER099C, YER107C,YER128W, YER137C, YER171W, YFL036W, YFL058W, YGL021W, YGL085W, YGL099W, YGL128C, YGL155W, YGL166W, YGL214W, YGL234W, YGL248W, YGR023W, YGR108W, YGR129W, YGR187C, YGR216C, YHR023W, YHR062C, YHR151C, YIL007C, YIL011W, YIL097W, YIL106W, YIL117C,YIL158W, YIL171W, YJL011C, YJL051W, YJL053W, YJR002W, YJR092W, YJR127C, YKL057C, YKL129C, YKL143W,YKL173W, YKL205W, YKL222C, YKR031C, YKR056W, YKR060W, YKR097W, YLL008W, YLL018C, YLL043W, YLR014C, YLR023C, YLR051C, YLR068W, YLR086W, YLR088W, YLR107W, YLR131C, YLR146C, YLR190W, YLR215C, YLR222C, YLR227C, YLR277C, YLR353W, YLR420W, YLR430W, YLR434C, YLR438W, YLR453C, YML033W, YML034W, YML080W, YML082W, YML093W, YML094W, YML096W, YML103C, YML104C, YML130C, YMR001C, YMR021C, YMR032W, YMR033W, YMR034C, YMR059W, YMR093W, YMR112C, YMR131C, YMR132C, YMR185W, YMR211W, YMR212C, YMR265C, YMR281W, YMR291W, YNL053W, YNL124W, YNL132W, YNL163C, YNL171C, YNL193W, YNL199C, YNL227C, YNL299W, YNR002C, YNR003C, YNR038W, YNR039C, YOL021C, YOL022C, YOL031C, YOL060C, YOL080C, YOL113W, YOL124C, YOL130W,YOL144W, YOR006C, YOR056C, YOR058C, YOR061W, YOR098C, YOR145C, YOR160W, YOR272W, YOR364W, YPL126W, YPL148C, YPL150W, YPL183C, YPL192C, YPL231W, YPL242C, YPL248C,YPR026W,YPR046W, YPR079W, YPR084W, YPR112C, YPR119W.

The Table 4.4 given below shows the significant GO terms used to describe genes of the biclusters 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 4.4 the first entry of the second column with the title process contains the term ribosome biogenesis (44, 3.45e-22) which means that 44 out of the 121 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 3.45e-22. Second entry indicates that 46 out of 121 genes are involved in ribonucleoprotein complex biogenesis. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the method used here is capable of identifying biologically significant biclusters from different GO categories.

Table 4.4

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 4.3

Bicluster	Process	Function	Component
S61	Ribosome Biogenesis (44, 1.45e-23) ribonucleoprotein complex biogenesis(46, 6.13e-23) cellular component biogenesis at cellular level (47,6.18e-20) ncRNA processing (39, 3.68e-19) nitrogen compound metabolic process (64, 4.38e-06)	44 genes annotated to the term molecular function unknown.	Nucleolus (35, 8.74e-21) preribosome (23, 5.33e-13) nuclear part (53, 1.28e-10) cell part (112, 0.00189)
S62	Translation (69, 1.52e-56) cellular protein metabolic process (72, 1.13e-27) protein metabolic process (72, 8.11e-27) metabolic process (84, 9.28e-07)	Structural constituent of ribosome(62, 5.81e-62) structural molecule activity (63, 4.33e-49) translation elongation factor activity (5, 0.00011) RNA binding (15, 0.00603	cytosolic ribosome (64, 1.42e-70) cytosolic part (64, 3.93e-64) ribosome (68, 1.10e-58) intracellular organelle (86, 0.00076)
S63	DNA metabolic process (19, 5.44e-11) DNA repair (16, 9.53e-11) cell cycle (20, 8.42e-10) nucleobase, nucleoside, nucleotide and nucleic acid (23, 0.00011)	Structure-specific DNA binding (5,0.00315) double-stranded DNA binding(4,0.00134)	Chromosome (15,1.21e-07) replication fork (8, 1.40e-06) Chromosomal part(13,4.93e-06) Nucleus (26, 1.52e-05)
S64	Ribonucleoprotein complex biogenesis (51, 1.55e-14) ribosome biogenesis (45, 9.55e-13) cellular component biogenesis at cellular level(52, 3.72e-11) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (86, 0.00060)	Endonuclease activity(9, 0.00591)	Nucleolus (36, 3.68e-12) nucleus(110, 8.02e-08) preribosome (24, 8.27e-08) nuclear part (72,7.92e-07) 90s preribosome (16, 3.88 e-05)

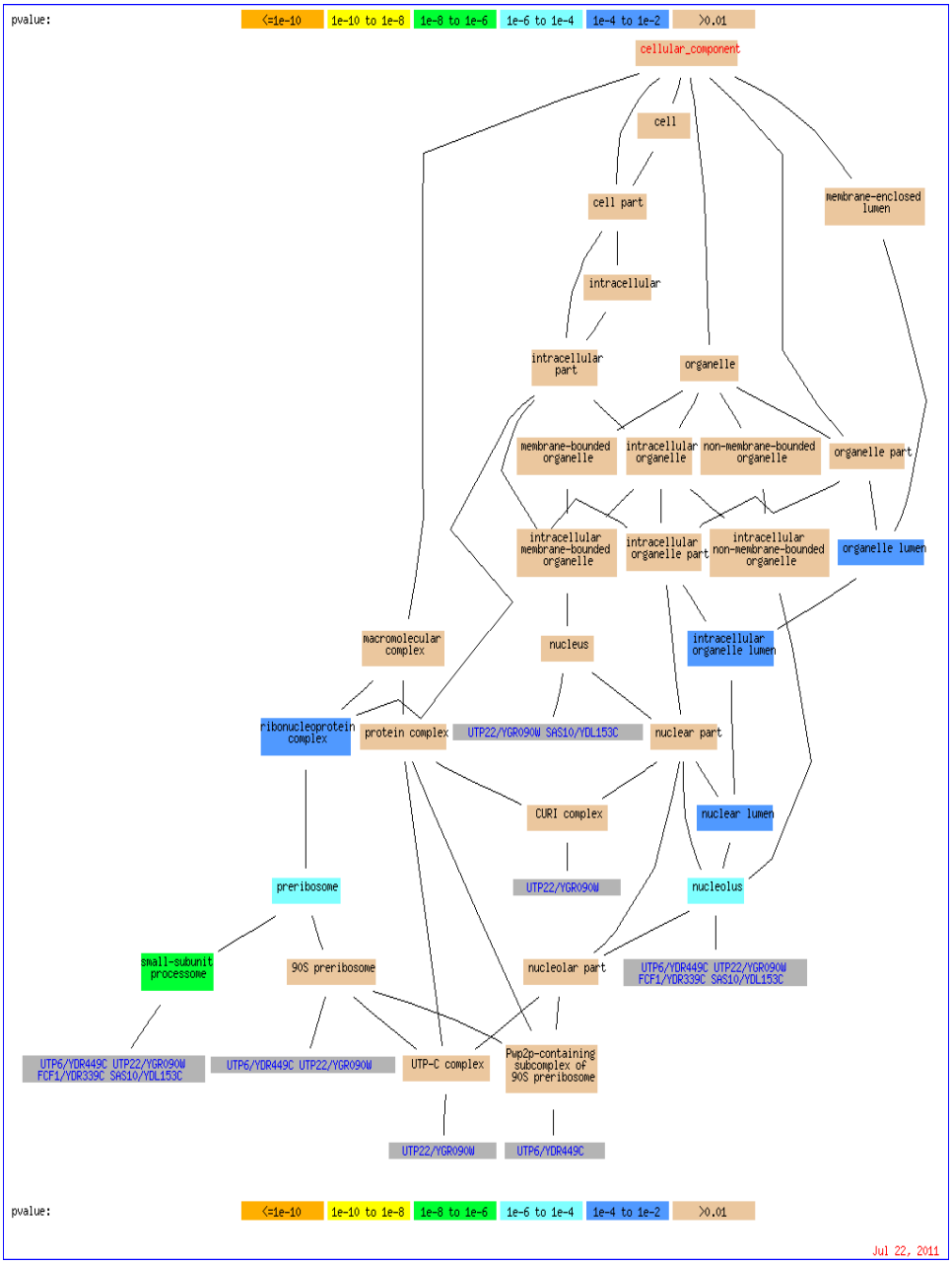


Figure 4.4: Sample of genes for bicluster s61, with corresponding GO terms and their parents for Component Ontology

Figure 4.4 shows the significant GO terms for the set of 121 genes in bicluster s61 along with their p-values. It shows the branching of cellular component into sub-components. These subcomponents are clustered using genes to produce the final result. Figure 4.4 is obtained when gene ontology database is searched by entering the names of genes and by selecting component ontology. Only four genes namely YDL153C, YDR339C, YDR449C, YGR090W are searched to reduce the size of the Figure.

4.6 Comparison with Other Algorithms

4.6.1 Comparison based on Statistical and Biological Significance

In Table 4.5 the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the greedy is compared with MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the table it is clear that in terms of p-value obtained by a bicluster which is used to denote statistical significance greedy is better than all the other algorithms namely MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu for all the five GO terms. The percentage of genes involved in the first GO term is better than that of all the other algorithms except MOGAB, SGAB, CC and Bimax. The percentage of genes involved in the second, third, fourth and fifth GO terms are better than that of all the other algorithms.

Table 4.5
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by Greedy and other Algorithms for Yeast Data

Terms	Greedy	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 59.8% (64, 1.42e-70)	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Cytosolic Part 59.8% (64, 3.93e-64)	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	structural constituent of ribosome 57.9% (62, 5.81e-62)	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromole cule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 63.6% (68, 1.10e-58)	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Translation 64.5% (69, 1.52e-56)	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishme nt of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

4.6.2 Comparison with other Algorithms based on Bicluster Size and MSR

A comparative summarization of results of Yeast data involving the performance of related algorithms are given in Table 4.6. The performance of greedy algorithm in comparison with that of SEBI [36], Cheng and Church's algorithm (CC) [29], and the algorithm FLOC by Yang et al. [106] and DBF [109] for the Yeast dataset are given. For the greedy algorithm presented here the average number of conditions is better than that of CC, FLOC and DBF. Average number of genes, average volume and the largest bicluster size is greater than that of all other algorithms. Average mean squared residue score is better than that of all other algorithms listed in the Table 4.6, except DBF.

In multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 11 for the Yeast dataset. But, in this method there are biclusters with all 17 conditions. For the Yeast dataset the maximum number of genes obtained for this algorithm in all the 17 conditions is 147 with MSR value 200.2474. The maximum available in all the literature published so far is in the case of multi-objective PSO [62]. They obtained 141 genes for 17 conditions with MSR value 203.25.

Table 4.6

Performance Comparison between Greedy and other Algorithms for Yeast Dataset

Algorithm	AMR	ANG	ANC	AV	LB
Greedy	185.88	515.21	13.36	4684.29	12645
CC	204.29	166.71	12.09	1576.98	4485
SEBI	205.18	13.61	15.25	209.92	1394
FLOC	187.54	195.00	12.80	1825.78	2000
DBF	114.70	188.00	11.00	1627.20	4000

AMR is average mean squared residue. ANG is average number of genes. ANC is the average number of conditions. AV is average volume. LB is largest bicluster. As 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.

Table 4.7 gives a performance comparison for Human B-cell Lymphoma dataset. Value of δ is set to 1200 for Lymphoma dataset. Here the average number of genes is greater than SEBI. Average number of conditions is better than all other algorithms. Average volume is better than SEBI. Average MSR is lower than SEBI. Usually multi-objective algorithms will produce biclusters of larger size compared to greedy algorithms. But in the case of multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 40 in the case of Human B-cell Lymphoma dataset. Here biclusters with 94 conditions is obtained where as maximum obtained in the case of multi-objective PSO

is 84 [62]. In the case of SEBI the maximum number of conditions obtained is 72 and the number of genes for this bicluster is only 3. But for greedy algorithm the bicluster with 94 conditions contains 11 genes. The row variance of this bicluster is also above 5000.

Table 4.7

Performance Comparison between Greedy Algorithm and other Algorithms for Human Lymphoma Dataset

Algorithm	AMR	ANG	ANC	AV
Greedy	1007.99	60.38	57.13	1710.25
SEBI	1028.84	14.07	43.57	615.84
CC	850.04	269.22	24.50	4595.98

AMR is average mean squared residue. ANG is average number of genes. ANC is the average number of conditions. AV is average volume. LB is largest bicluster. In the above table the average mean squared residue, the average number of genes and conditions and average volume and 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.

4.7 Summary

In this chapter a new algorithm is developed for identifying biclusters from the gene expression data. This greedy algorithm is implemented on both benchmark datasets. In the first step K-Means clustering algorithm is used to produce bicluster seeds. Then these seeds

are enlarged by greedy method in which the node with minimum incremental increase in MSR score is selected and added to the bicluster in each iteration. Hence it is possible to get bicluster having more genes and conditions with high coherence. Some of the biclusters have very high row variance also. The statistical significance and biological relevance of biclusters obtained in this method are verified using gene ontology database. In this study the maximum number of genes (147) is obtained in all the 17 conditions with the minimum MSR value (200.2474) for the Yeast dataset. A bicluster with the maximum number of conditions (94) is obtained for the Lymphoma dataset. The biclusters obtained here show similar up-regulation and down-regulation under a set of conditions. In terms of size and MSR value the biclusters obtained in this method are far better than the biclusters obtained in many of the metaheuristic algorithms. This algorithm has the best p-value compared to that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu.

…..ஜ۞ஜ…..

Chapter 5

Metaheuristic Algorithms

Chapter 5 describes the metaheuristic algorithms namely basic GRASP, CGRASP, RGRASP, PSO and Greedy-PSO hybrid. For finding biclusters from gene expression data, the seeds obtained from K-Means clustering are enlarged using these algorithms. The description of the algorithms, their time complexity, different biclusters obtained from Yeast and Lymphoma datasets, significant biclusters obtained (biological validation), comparison of the algorithms with other algorithms are also given in this chapter. The greedy and metaheuristic algorithms are compared based on the quality of bicluster.

5. 1 Greedy Randomized Adaptive Search Procedure

GRASP was developed by Feo and Resende in 1995 [42]. GRASP incorporates randomization in order to eliminate local minima problem existing in greedy approaches. GRASP is an iterative randomized sampling method in which each iteration consists of two phases: construction and local search. The construction phase generates a feasible solution, whose neighbourhood is investigated until a local minimum is identified during the process of local search phase. The best overall solution is reserved as the result. In this work biclusters were identified using three variants of Greedy Randomized Adaptive Search Procedure (GRASP) namely basic GRASP, Cardinality based GRASP and Reactive GRASP. In this work the objective is to identify biclusters with maximum size and low MSR. Biclusters with more genes and conditions and low MSR are obtained in this work. Moreover in this study GRASP variants are applied for the first time to Lymphoma dataset.

5.1.1 Review of Grasp Metaheuristics

5.1.1.1 Construction Phase

GRASP is a multi-start metaheuristics for solving combinatorial optimization problems. Metaheuristics is a computational method which optimizes a problem iteratively by improving a solution with regard to a particular measure of quality. In the construction phase a feasible solution is generated by adding one element at a time. In the local search phase the neighborhood of the feasible solution is investigated until a local

minimum is found. The best overall solution is retained as the result. During each iteration of the construction phase a set of candidate elements are formed by all the elements that can be incorporated to the partial solution under construction without eliminating feasibility. The selection of the next element for incorporation is resolved by the evaluation of all candidate elements in accordance with a greedy evaluation function [42].

This greedy function stands for the incremental increase in the cost function because of the incorporation of this element into the solution under construction. The evaluation of the elements by this function results in the creation of a restricted candidate list (RCL) produced by the best elements. That is, those elements whose incorporation to the current partial solution results in the smallest incremental costs. This is the greedy aspect of the algorithm. The element which is to be incorporated into the partial solution is randomly chosen from those in the RCL. This is the probabilistic aspect of the heuristic algorithm. Once the chosen element is included in the partial solution, the candidate list is restructured and the incremental costs are recalculated. This is the adaptive aspect of the heuristic algorithm. The restricted candidate list RCL is constituted of elements with the best (i.e., the smallest) incremental costs. This list can be limited by different factors. That is, either by the number of elements (cardinality-based) or by their quality (value-based) [42].

5.1.1.2 Local Search Phase

The solutions produced by the greedy randomized construction are not always optimal even with respect to simple neighbourhoods. The local search phase makes 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 neighbourhood of the existing solution. It finishes when no better solution is identified in the neighbourhood. Local search can be implemented by using the first improving or best improving strategy. In the case of best improving strategy all neighbours are investigated and the current solution is replaced by the best neighbour. In the case of a first improving strategy the current solution moves to the first neighbour whose cost function value is smaller than that of the current solution. In the first improving strategy the search stops as soon as a better solution is found [76].

5.1.2 Three variants of GRASP – Basic GRASP, Cardinality based GRASP (CGRASP) and Reactive GRASP (RGRASP)

The restricted candidate list RCL is made up of elements with the best incremental costs. This list can be limited by the number of elements (cardinality) or by their quality (value based or Basic GRASP). In the first case it is made up of the P elements with the best incremental cost where P is a parameter. In the second case all the elements less than RCL threshold will form the RCL. Hence this list will be variable in each iteration.

In the calculation of RCL threshold a parameter α is used. In basic GRASP α is assigned a single value for all iterations. The value of α can range from 0 to 1. The amount of greediness and randomness are controlled by the parameter α. The algorithm is purely greedy when α=0. But when α=1 it is equivalent to random construction. But in reactive GRASP at each iteration the value of α is chosen from a discrete set of values {α1, α2, α3, … αn} depending on the probability Pi associated with each αi. Initially all αi will have the same probability and each one is selected once. Depending on the quality of solution the probability is updated. Then after each iteration the αi with highest probability is selected. The probability is updated depending on the quality of solution obtained when αi is used so as to favour values that produce good solution. In this algorithm the quality of solution obtained is evaluated based on the size of the bicluster as well as the MSR value. If αi is the value of α selected in a particular iteration then after obtaining the result the difference between the solutions obtained in the previous iteration and present iteration D_i is calculated. Assume Avi as the average obtained for all D_is with αi as the probability. Then for updating the probability Pi after an iteration with αi the following formula can be used.

$$Pi = \frac{mi}{\sum_{i=1}^{n} mi} \text{ where } mi = Avi \text{ for } i=1\ldots n.$$

Larger values for probability is obtained for αi with better solutions when this formula is used.

5.1.3 Grasp Algorithms for Seed Growing Phase

5.1.3.1 Algorithm for the Construction Phase

Algorithm Greedy_Randomized_Construct (Seed)

```
bicluster←seed;
While solution construction  not done
    cand←construct _candidatelist (bicluster, δ)
     RCL←BuildRCL(bicluster,cand)
     Select an element S from RCL at random
     bicluster=bicluster U{S}
     Update Genelist or Conditionlist
End(while)
End(Greedy_Randomized_Construct)
```

5.1.3.2 Algorithm for Constructing Candidate list

Algorithm 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
          msrlist[i]=MSR(Bicluster1 U notinlist[i])
End(for)
 Candidatelist={}
For  i=1:notinlistcount
   If  msrlist[i]< δ
   Candidatelist=candidatelist U Notinlist[i]
End(for)
end(construct_candidatelist)
```

5.1.3.3 Algorithm for Building RCL from Candidate list

Algorithm BuildRCL(bicluster,CAND)

```
// CAND  is the candidate list
  SminMSR = inf
 SmaxMSR = -inf
 nocan=noofelements(CAND)
 for  I =1: nocan  do
    calculate H[i]← MSR{ bicluster  U CAND[i]}
      if H[i ]<SminMSR
        SminMSR=H[i]
      Endif
        if  H[i ]>SmaxMSR
          SmaxMSR=H[i]
          Endif
  Endfor
  RCLthresh=SminMSR+α*(SmaxMSR-
                              SminMSR
  RCL={}
   For i=1:nocan
        If  H[i]<RCLthresh
          RCL=RCL U{CAND[i]}
       Endif
   end(for)
 end BuildRCL
```

5.1.3.4 Algorithm for the Local Search phase

Algorithm Local_Search(bicluster)

```
    //local search
While there exists s ϵ genelist or conditionlist
  If  MSR(biclusterU s)<MSR(bicluster)
  bicluster={bicluster U s}
  endif
 end(while)
end(Local_Search)
```

5.1.4 Time Complexity of the Algorithm

The basic operation for the identification of biclusters is the calculation of MSR of a submatrix. Time complexity for calculating MSR is O(mn). In this algorithm conditions are added first followed by genes. There is construction phase and local search phase for both genes and conditions. In both these phases, for including a condition, the MSR value of all submatrices which result from adding a single condition, is to be calculated for all conditions. This number decreases by one after each iteration. That means the complexity can be calculated by the formula (n + (n-1) + (n-2) + ...1). This is equal to n(n+1)/2 which is equivalent to $O(n^2)$. Hence for adding conditions the worst case complexity is $O(mn)(n^2)$. Similarly for adding genes the worst case complexity is $O(mn)(m^2)$. Hence the worst case complexity for adding genes and conditions is $O(mn)(m^2+n^2)$ where m and n are the number of genes and conditions respectively.

5.1.5 Biclusters obtained Using GRASP (Basic GRASP)

In seed growing phase more conditions and genes are added to the seed. For this purpose list of conditions and genes not included in the bicluster is maintained. Thus a separate gene list and condition list is formed. From this list the candidate gene list and candidate condition list is formed by those elements whose incorporation into the seed will not exceed the MSR value above the MSR threshold. From this candidate list RCL list is formed by selecting the best elements. The best elements will have an MSR value less than RCL threshold where RCL threshold=

MSRmin+ α (MSRmax-MSRmin). The maximum of the MSR value obtained when a single gene or condition is added from the candidate list is MSRmax. The minimum value of MSR when a gene or condition is added from the candidate list for a given iteration is MSRmin. The value of α can range from 0 to 1. The amounts of greediness and randomness are controlled by the parameter α. The RCL list thus obtained is called value based. The number of elements in the RCL list will vary in each iteration. In seed growing phase, the next element to be added to the bicluster is selected randomly from the RCL. After adding the node the candidate list and RCL are updated. The process of adding the node is continued till the MSR value of the bicluster reaches the given MSR threshold.

5.1.5.1 Bicluster Plots for Yeast Dataset

In Figure 5.1 the eight biclusters obtained using GRASP are shown. Biclusters with all 17 conditions are obtained using this method. From the bicluster plots which show strikingly similar up-regulation and down-regulation we can conclude that GRASP is an ideal method for identifying coherent biclusters from gene expression data. All the means squared residues are lower than 215.

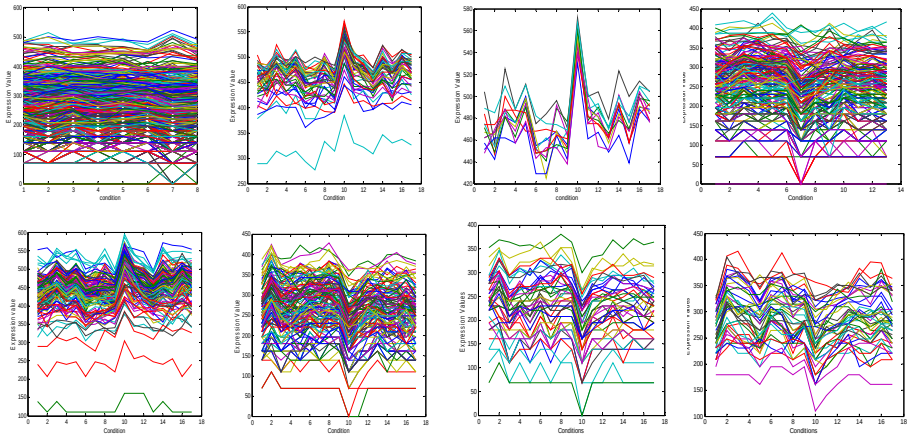


Figure 5.1 Eight biclusters found for the Yeast Dataset by GRASP. Bicluster labels are (yva7), (yvb7), (yvc7), (yvd7), (yve7), (yvf7), (yvg7) and (yvh7) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 5.1 using bicluster label.

Table 5.1

Information about Biclusters of Figure 5.1

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(yva7)	783	8	6264	215.0790
(yvb7)	42	17	714	121.6900
(yvc7)	12	17	204	69.9591
(yvd7)	208	13	2704	193.6400
(yve7)	108	17	1836	200.7372
(yvf7)	140	17	2380	200.0088
(yvg7)	47	17	799	145.3612
(yvh7)	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 the number 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.

5.1.5.2 Bicluster Plots for Human Lymphoma Dataset

In Figure 5.2 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. All the means squared residues of the biclusters are lower than 1200.

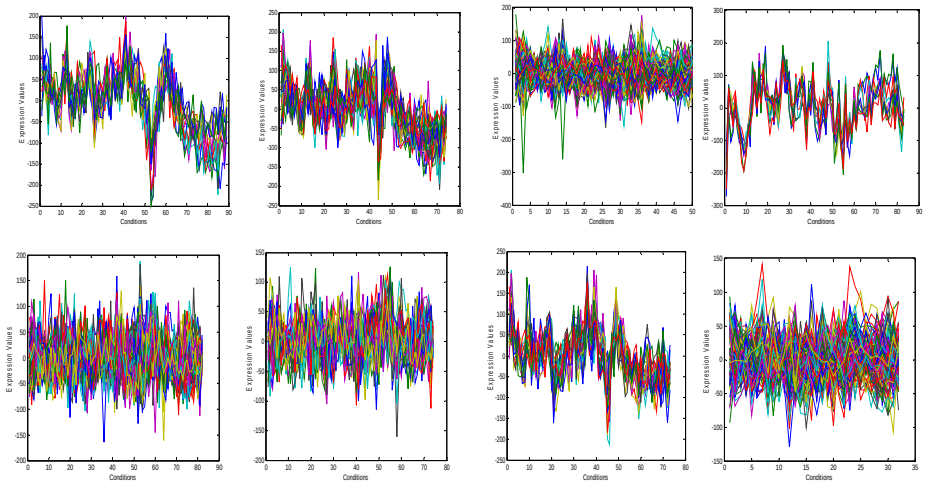


Figure 5. 2 Eight biclusters found for the Lymphoma Dataset by GRASP. Bicluster labels are (lva7), (lvb7), (lvc7), (lvd7), (lve7), (lvf7), (lvg7) and (lvh7) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 5.2 using bicluster label.

Table 5.2

Information about Biclusters of Figure 5.2

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(lva7)	16	89	1424	1196.9
(lvb7)	38	74	2812	1189.8
(lvc7)	175	50	8750	1075.2
(lvd7)	10	83	830	1182.1
(lve7)	62	82	5084	1197.3
(lvf7)	34	74	2516	1019.5
(lvg7)	24	73	1752	1197.9
(lvh7)	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.

5.1.5.3 Details of Significant Biclusters obtained by GRASP

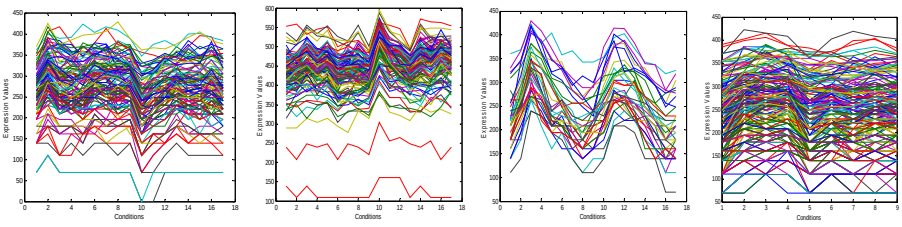


Figure 5.3 Four significant biclusters obtained by the GRASP algorithm on Yeast dataset. The bicluster labels are sv71, sv72, sv73 and sv74. The details about the biclusters can be obtained from Table 5.3 using bicluster label.

Table 5.3

Information about Biclusters of Figure 5.3

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
Sv71	121	17	199.9395	483.2784
Sv72	107	17	199.4776	568.0833
Sv73	36	17	297.6071	1806.9000
Sv74	224	9	228.1477	403.6127

Biological relevance of biclusters obtained using GRASP algorithm is verified using the four biclusters shown in Figure 5.3. GO annotation database is used to verify the biological significance of biclusters. In the bicluster Sv71 there are 121 genes. They are YBL014C, YBL083C, YBL084C, YBR293W, YCL016C, YCL031C, YCL053C, YCL054W, YCR072C, YCR087W, YDL008W, YDL030W, YDL076C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL215C, YDL231C, YDR017C, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR120C, YDR121W, YDR170C, YDR172W, YDR211W, YDR234W, YDR262W, YDR289C, YDR299W, YDR312W, YDR321W, YDR339C, YDR352W, YDR361C, YDR365C, YDR392W, YDR416W, YDR449C, YDR469W, YDR477W, YDR478W, YDR518W, YDR524C, YDR542W, YEL015W, YEL055C, YER005W, YER075C, YER099C, YER107C, YER166W, YER168C, YER171W, YFL001W, YGL085W, YGL099W, YGL214W, YGR042W, YGR090W, YGR187C, YGR200C, YGR216C, YHR062C, YJL011C, YJL069C, YJR017C, YJR066W, YKR056W, YKR060W, YLL008W, YLL034C, YLR051C, YLR088W, YLR107W, YLR146C, YLR215C, YLR222C, YLR227C, YLR401C, YML066C, YML080W, YML093W, YMR093W, YMR211W, YMR235C, YNL041C, YNL132W, YNL163C, YNL164C, YNL199C, YNL227C, YNL299W, YNR003C, YNR038W, YOL021C, YOL022C, YOL036W, YOL080C, YOL124C, YOL140W, YOL144W, YOR006C, YOR056C, YOR061W, YOR098C, YOR145C, YOR160W, YOR252W, YOR272W, YPL126W, YPL268W, YPR053C, YPR112C.

In bicluster sv72 there are 107 genes namely YAL003W, YAL038W, YAR020C, YBL030C, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A,YBR106W, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR154C, YDR353W, YDR382W, YDR385W, YDR417C, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER117W, YGL102C, YGR118W, YHR141C, YJL136C, YJL188C, YJL189W, YJL190C, YJR009C, YJR094W-A, YJR123W, YKL056C, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL066C, YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR333C, YLR340W, YLR388W, YLR406C, YLR441C, YLR467W, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR045C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W,YPL090C,YPL143W, YPL283C, YPR102C,YPR204W.

In the bicluster Sv73 there are 36 genes. They are YAR007C, YAR008W, YBL035C, YBR073W, YBR088C, YBR089W, YCR065W, YDL003W, YDL010W, YDL018C, YDL164C, YDR097C, YDR507C, YER095W, YFL008W, YGR151C, YGR152C, YHR154W, YIL026C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR386W, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL303W, YNL312W, YOR074C, YPL208W, YPR120C. The fourth seed results in a bicluster with more than 400 genes. Since there are a large number of genes, there is no search result. Hence the algorithm is executed in such a way to get only 224 genes. These genes are YAL028W, YAL035W, YAL041W, YAL059W, YBL004W, YBL014C, YBL024W, YBL026W, YBL032W, YBL037W, YBL042C, YBL049W, YBL052C, YBL056W, YBL068W,

YBL075C, YBL083C, YBL088C, YBR021W, YBR038W, YBR060C, YBR075W, YBR079C, YBR094W, YBR123C, YBR133C, YBR138C, YBR140C, YBR155W, YBR257W, YBR270C, YBR295W, YCL012W, YCL031C, YCL054W, YCL059C, YCR014C, YCR024C, YCR036W, YCR043C, YCR060W, YCR062W, YCR063W, YDL008W, YDL030W, YDL043C, YDL058W, YDL069C, YDL076C, YDL079C, YDL142C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL202W, YDL215C, YDL230W, YDL231C, YDL243C, YDR011W, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR109C, YDR120C, YDR150W, YDR170C, YDR172W, YDR185C, YDR197W, YDR198C, YDR211W, YDR214W, YDR235W, YDR236C, YDR262W, YDR272W, YDR282C, YDR286C, YDR288W, YDR312W, YDR313C, YDR324C, YDR352W, YDR363W, YDR375C, YDR391C, YDR392W, YDR419W, YDR456W, YDR466W, YDR524C, YEL043W, YEL053C, YEL055C, YER005W, YER034W, YER064C,YER107C, YFL006W, YFL036W, YGL021W, YGL248W, YGL255W, YGR035C, YGR092W, YGR108W, YGR129W, YGR187C, YGR200C, YGR216C, YHR023W, YHR062C, YHR073W, YHR151C, YIL007C, YIL097W, YIL106W, YIL117C, YIL158W, YIL162W, YJL096W, YJL192C, YJR002W, YJR092W, YKL057C, YKL118W, YKL129C, YKL143W, YKL173W, YKL205W, YKR021W, YKR031C, YKR060W, YKR079C, YLL008W, YLR014C, YLR051C, YLR068W, YLR107W, YLR131C, YLR190W, YLR215C, YLR222C, YLR227C, YLR277C, YLR320W, YLR353W, YLR420W, YLR434C, YLR438W, YML033W, YML034W, YML052W, YML064C, YML082W,YML093W, YML094W, YML103C, YML119W, YML130C, YMR001C, YMR025W, YMR032W, YMR033W, YMR059W, YMR072W, YMR093W, YMR132C, YMR156C, YMR211W, YMR212C, YMR225C, YMR278W, YMR291W, YNL041C, YNL051W, YNL053W, YNL132W, YNL163C, YNL171C, YNL172W, YNL196C, YNL201C, YNL223W, YNL227C, YNL299W, YNR003C, YOL021C, YOL022C, YOL042W, YOL060C, YOL070C, YOL080C, YOL113W, YOR006C, YOR049C, YOR061W, YOR098C, YOR104W, YOR127W, YOR145C, YOR152C, YOR160W, YOR205C, YPL029W, YPL150W, YPL173W, YPL183C, YPL198W, YPL205C, YPL242C, YPR003C, YPR026W, YPR079W, YPR112C, YPR119W.

The Table 5.4 given below shows the significant GO terms used to describe the genes of the biclusters of Figure 5.3 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 5.3 the first entry of the second column with the title process contains the term ribosome biogenesis (44, 1.46e-23) which means that 44 out of the 121 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 1.46e-23. Second entry indicates that 46 out of 121 genes are involved in ribonucleoprotein complex biogenesis. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the GRASP method used here is capable of identifying biologically significant biclusters from different GO categories.

Table 5.4

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 5.3

Bicluster	Process	Function	Component
Sv71	Ribosome Biogenesis (44, 1.46e-23) Ribonucleoprotien complex biogenesis (46, 6.18e-23) Cellular component biogenesis at celluar Level (47, 6.22e-20) Nitrogen compound metabolic process (64, 4.18e-06)	44 out of 121 genes are directly annotated to the term molecular function unknown	Nucleolus (35,8.74e-21) Preribosome (23,5.33e-13) Nuclear part (53, 1.17e-10) cell part (112, 0.00189)

Sv72	Translation(69, 1.52e-56) cellular protein metabolic process (72, 1.13e-27) protein metabolic process (72, 8.11e-27) metabolic process (84, 9.28e-07)	Structural Constituent of ribosome (62, 5.81e-62) structural molecule activity (63, 4.33e-49) translation elongation factor activity (5, 0.00011) RNA binding (15, 0.00603	Cytosolic ribosome (55, 3.68e-55) cytosolic part (55, 4.85e-50) Ribosome (59, 3.190e-46) cytoplasm (74, 0.00569
Sv73	DNA metabolic process(19, 5.44e-11) DNA repair (16, 9.53e-11) cell cycle(20, 8.42e-10) nucleobase, nucleoside, nucleotide and nucleic acid (23, 0.00011)	Structure-specific DNA binding (5,0.00315) double-stranded DNA binding (4,0.00134)	Chromosome(15,1.21e-07) replication fork (8, 1.40e-06) Chromosomal part (13,4.93e-06) Nucleus (26, 1.52e-05)
Sv74	RNA processing (42,1.63e-06) Ribosome biogenesis (35, 3.86e-06) ncRNA processing (34,6.13e-06) ribonucleoprotien complex biogenesis (37,1.50e-05) ncRNA metabolic process(35, 2.40 e-05) Cellular component organization or biogenesis(100, 0 .00291)	84 genes are annotated to the term molecular function unknown.	Nucleolus(32, 4.08e-09) Preribosome (19, 0.00034) Intracellular organelle (168, 0.00039) Organelle (168, 0.00041)

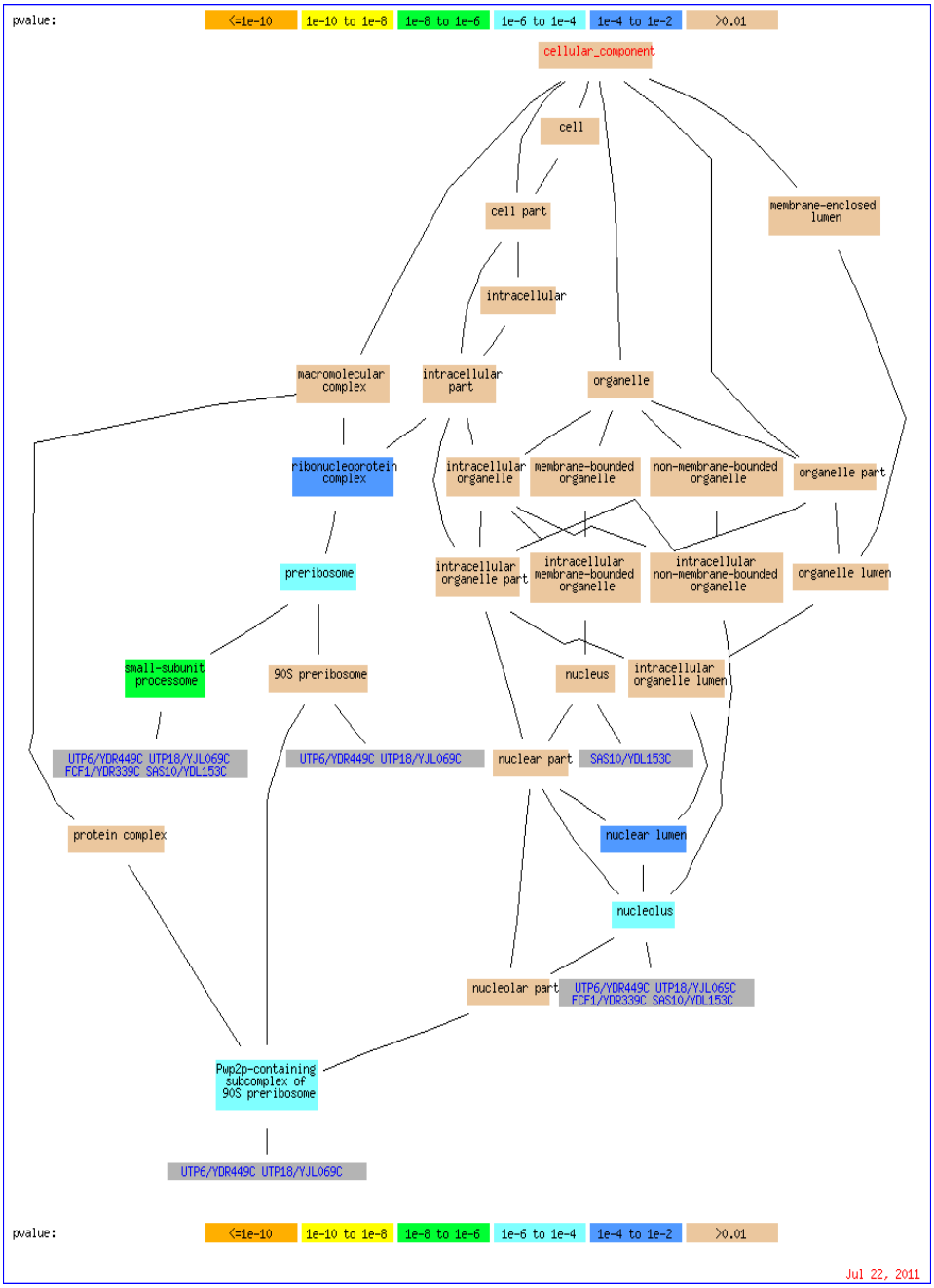


Figure 5.4 Sample of Genes for the bicluster sv71, with corresponding GO terms and their parents for Component ontology

Figure 5.4 shows the significant GO terms for the set of genes in bicluster sv71 along with their p-values. It shows the branching of cellular component into sub-components like cell and membrane-enclosed lumen. These components are clustered using genes to produce the final result. Figure 5.4 is obtained when gene ontology database is searched by entering the names of genes of bicluster sv71 and by selecting component ontology. Only 4 genes (YDL153C, YDR339C, YDR449C, YJL069C) are searched to reduce the size of the Figure.

5.1.6 Biclusters obtained Using CGRASP

In seed growing phase more conditions and genes are added to the seed. A separate list is maintained for genes and conditions not included in the bicluster. From this list, the candidate gene list and candidate condition list are formed by those elements whose incorporation into the seed will not exceed the MSR score above the MSR threshold. From this candidate list, RCL is formed by selecting the best elements. The best elements will have an MSR value less than RCL threshold where RCL threshold= MSRmin+ α (MSRmax-MSRmin). When this formula is used the RCL is called value based. For cardinality based GRASP P best elements are selected from the RCL. So the number elements which can be considered for inclusion in the bicluster will be fixed for each iteration.

5.1.6.1 Bicluster Plots for Yeast Dataset

In Figure 5.5 nine biclusters obtained using CGRASP are shown. Biclusters with all 17 conditions are obtained using this method. From the bicluster plots which show strikingly similar upregulation and down

regulation we can conclude that CGRASP is an ideal method for identifying coherent biclusters from gene expression data. All the means squared residues are lower than 215.

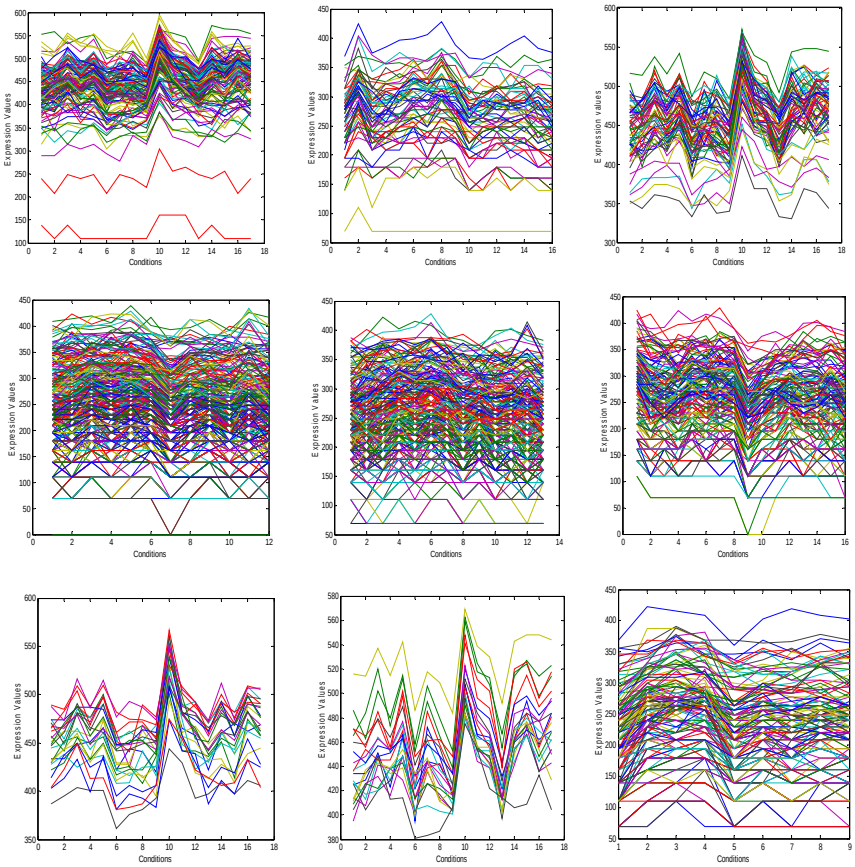


Figure 5.5 Nine biclusters found for the Yeast Dataset by CGRASP. Bicluster labels are (yac7), (ybc7), (ycc7), (ydc7), (yec7), (yfc7), (ygc7), (yhc7) and (yic7) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 5.5 using bicluster label.

Table 5.5

Information about Biclusters of Figure 5.5

Bicluster label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(yac7)	107	17	1819	199.1857
(ybc7)	63	17	1071	148.1866
(ycc7)	64	16	1024	149.6244
(ydc7)	324	12	3888	193.7751
(yec7)	256	13	3328	199.7194
(yfc7)	164	16	2624	199.7293
(ygc7)	24	17	408	104.5418
(yhc7)	21	17	357	94.4589
(yic7)	146	9	1314	250.1285

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 or hscore of the bicluster.

5.1.6.2. Bicluster Plots for Lymphoma Dataset

This is the first time CGRASP metaheuristics is applied to find biclusters from Lymphoma dataset. Eight biclusters obtained by applying CGRASP to lymphoma dataset are shown in Figure 5.6.

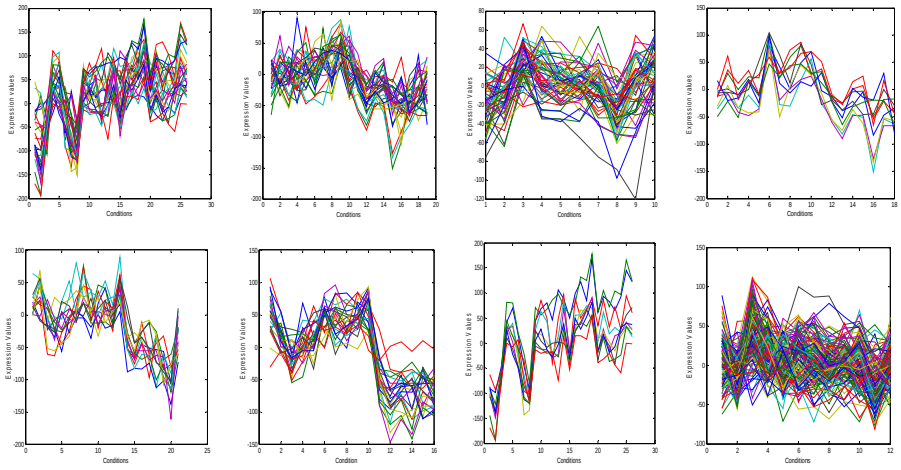


Figure 5.6 Eight biclusters found for the Lymphoma Dataset by CGRASP. The bicluster labels are (lac7), (lbc7), (lcc7), (ldc7), (lec7), (lfc7), (lgc7) and (lhc7) respectively. The details of the biclusters can be obtained from Table 5.6 using bicluster label.

Table 5.6

Information about Biclusters of Figure 5.6

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(lac7)	26	26	676	883.6869
(lbc7)	30	19	570	441.5052
(lcc7)	52	10	520	307.5545
(ldc7)	18	10	180	368.0541
(lec7)	14	21	294	409.6572
(lfc7)	24	16	384	542.8357
(lgc7)	10	26	260	388.4876
(lhc7)	112	12	1344	492.4187

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.

5.1.6.3 Details of Significant Biclusters obtained by CGRASP

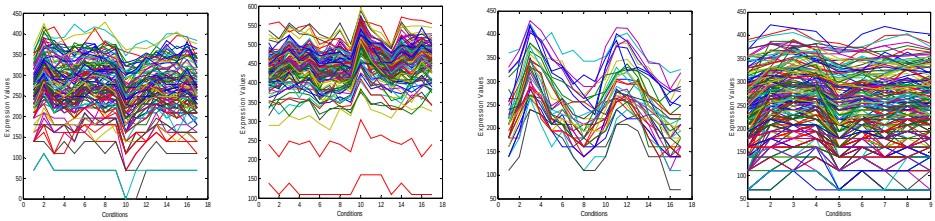


Figure 5.7 Four significant biclusters obtained by the CGRASP algorithm on Yeast dataset. The bicluster labels are sc71, sc72, sc73 and sc74. The details about the biclusters can be obtained from Table 5.7 using bicluster label.

Table 5.7

Information about Biclusters of Figure 5.7

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
Sc71	121	17	199.9395	483.2784
Sc72	107	17	199.4776	568.0833
Sc73	36	17	297.6071	1806.9000
Sc74	224	9	228.4546	403.1319

In the first bicluster sc71 there are 121 genes. They are YBL014C, YBL083C, YBL084C, YBR293W, YCL016C, YCL031C, YCL053C, YCL054W, YCR072C, YCR087W, YDL008W, YDL030W, YDL076C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL215C, YDL231C, YDR017C, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR120C, YDR121W, YDR170C, YDR172W, YDR211W, YDR234W, YDR262W, YDR289C, YDR299W, YDR312W, YDR321W, YDR339C, YDR352W, YDR361C, YDR365C,

YDR392W, YDR416W, YDR449C, YDR469W, YDR477W, YDR478W, YDR518W, YDR524C, YDR542W, YEL015W, YEL055C, YER005W, YER075C, YER099C, YER107C, YER166W, YER168C, YER171W, YFL001W, YGL085W, YGL099W, YGL214W, YGR042W, YGR090W, YGR187C, YGR200C, YGR216C, YHR062C, YJL011C, YJL069C, YJR017C, YJR066W, YKR056W, YKR060W, YLL008W, YLL034C, YLR051C, YLR088W, YLR107W, YLR146C, YLR215C, YLR222C, YLR227C, YLR401C, YML066C, YML080W, YML093W, YMR093W, YMR211W, YMR235C, YNL041C, YNL132W, YNL163C, YNL164C, YNL199C, YNL227C, YNL299W, YNR003C, YNR038W, YOL021C, YOL022C, YOL036W, YOL080C, YOL124C, YOL140W, YOL144W, YOR006C, YOR056C, YOR061W, YOR098C, YOR145C, YOR160W, YOR252W, YOR272W, YPL126W, YPL268W, YPR053C, YPR112C.

In the second bicluster sc72, there are 107 genes namely YAL003W, YAL038W, YAR020C, YBL030C, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A,YBR106W, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR154C, YDR353W, YDR382W, YDR385W, YDR417C, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER117W, YGL102C, YGR118W, YHR141C, YJL136C, YJL188C, YJL189W, YJL190C, YJR009C, YJR094W-A, YJR123W, YKL056C, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL066C, YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR333C, YLR340W, YLR388W, YLR406C, YLR441C, YLR467W, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR045C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C,

YOR369C, YPL037C, YPL081W, YPL090C, YPL143W, YPL283C, YPR102C, YPR204W.

In the third bicluster sc73, there are 36 genes. They are YAR007C, YAR008W, YBL035C, YBR073W, YBR088C, YBR089W, YCR065W, YDL003W, YDL010W, YDL018C, YDL164C, YDR097C, YDR507C, YER095W, YFL008W, YGR151C, YGR152C, YHR154W, YIL026C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR386W, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL303W, YNL312W, YOR074C, YPL208W, YPR120C

The fourth seed results in a bicluster with more than 400 genes. Since there are a large number of genes, there is no search result. Hence the algorithm is executed in such a way to get only 224 genes. These genes are YAL028W, YAL035W, YAL041W, YAL059W, YBL004W, YBL014C, YBL024W, YBL026W, YBL032W,YBL037W, YBL042C, YBL049W, YBL052C, YBL056W, YBL068W, YBL075C, YBL083C, YBL088C, YBR021W, YBR038W, YBR060C, YBR075W, YBR079C, YBR094W, YBR123C, YBR133C, YBR138C, YBR140C, YBR155W, YBR257W, YBR270C, YBR295W, YCL012W, YCL031C, YCL054W, YCL059C, YCR014C, YCR024C, YCR036W, YCR043C, YCR060W, YCR062W, YCR063W, YDL008W, YDL030W, YDL043C, YDL058W, YDL069C, YDL076C, YDL079C, YDL142C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL202W, YDL215C, YDL230W, YDL231C, YDL243C, YDR011W, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR109C, YDR120C, YDR150W, YDR170C, YDR172W, YDR185C, YDR197W, YDR198C, YDR211W, YDR214W, YDR235W, YDR236C, YDR262W, YDR272W, YDR282C, YDR286C, YDR288W, YDR312W, YDR313C, YDR324C, YDR352W, YDR363W, YDR375C, YDR391C, YDR392W, YDR419W, YDR456W, YDR466W, YDR477W, YDR524C, YEL043W, YEL053C, YEL055C, YER005W, YER034W, YER064C, YER107C, YFL006W, YFL036W, YGL021W, YGL248W, YGL255W, YGR035C, YGR092W, YGR108W, YGR129W, YGR187C, YGR200C, YGR216C, YHR023W, YHR062C, YHR073W, YHR151C, YIL007C, YIL097W, YIL106W, YIL117C,

YIL158W, YIL162W, YJL096W, YJL192C, YJR002W, YJR092W, YKL057C, YKL118W, YKL129C, YKL143W, YKL173W, YKL205W, YKR021W, YKR031C,YKR060W, YKR079C, YLL008W, YLR014C, YLR051C, YLR068W, YLR107W, YLR131C, YLR190W, YLR215C, YLR222C, YLR227C, YLR277C, YLR320W, YLR353W, YLR420W, YLR434C, YLR438W, YML033W, YML034W, YML052W, YML064C, YML082W, YML093W, YML094W, YML103C, YML119W, YML130C, YMR001C, YMR025W, YMR032W, YMR033W, YMR059W, YMR072W, YMR093W, YMR132C, YMR156C, YMR211W, YMR212C, YMR225C, YMR278W, YMR291W, YNL041C, YNL051W, YNL053W, YNL132W, YNL163C, YNL171C, YNL172W, YNL196C, YNL201C, YNL223W, YNL227C, YNL299W, YNR003C, YOL021C, YOL022C, YOL042W, YOL060C, YOL070C, YOL080C, YOL113W, YOR006C, YOR049C, YOR061W, YOR098C, YOR104W, YOR127W, YOR145C, YOR152C, YOR160W,YOR205C, YPL029W, YPL150W, YPL173W, YPL183C, YPL198W, YPL205C, YPL242C, YPR003C, YPR026W, YPR079W, YPR112C,YPR119W,

The Table 5.8 given below shows the significant GO terms used to describe the genes of the biclusters of Figure 5.7 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 5.8 the first entry of the second column with the title ‘process’ contains the term ribosome biogenesis (44, 1.46e-23) which means that 44 out of the 121 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 1.46e-23. Second entry indicates that 46 out of 121 genes are involved in ribonucleoprotein complex biogenesis. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the CGRASP method used here is capable of identifying biologically significant biclusters from different GO categories.

Table 5.8

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 5.7

Bicluster	Process	Function	Component
SC71	Ribosome Biogenesis (44, 1.46e-23) Ribonucleoprotien complex biogenesis (46, 6.18e-23) Cellular component biogenesis at celluar Level(47, 6.22e-20) Nitrogen compound metabolic process (64, 4.18e-06)	44 out of 121 genes are directly annotated to the term molecular function unknown	Nucleolus (35,8.74e-21) Preribosome (23,5.33e-13) Nuclear part(53, 1.17e-10) cell part (112, 0.00189)
SC72	Translation(69, 1.52e-56) cellular protein metabolic process (72, 1.13e-27) protein metabolic process(72, 8.11e-27) metabolic process (84, 9.28e-07)	Structural constituent of ribosome(62, 5.81e-62) structural molecule activity (63, 4.33e-49) translation elongation factor activity (5, 0.00011) RNA binding (15, 0.00603	Cytosolic ribosome (55, 3.68e-55) cytosolic part (55, 4.85e-50) Ribosome (59, 3.190e-46) cytoplasm (74, 0.00569
SC73	DNA metabolic process (19, 5.44e-11) DNA repair (16, 9.53e-11) cell cycle(20, 8.42e-10) nucleobase, nucleoside, nucleotide and nucleic acid (23, 0.00011)	Structure-specific DNA binding (5,0.00315) double-stranded DNA binding(4,0.00134)	Chromosome(15,1.21 e-07) replication fork (8, 1.40e-06) Chromosomal part(13,4.93e-06) Nucleus (26, 1.52e-05)
SC74	RNA processing (42, 1.92e-06) ribosome biogenesis (35, 4.45e-06) ncRNA processing (34, 7.03e-06) cellular Component Organization or biogenesis (101,0.00194)	84out of 224input genes are directly annotated to root term 'molecular function unknown'	Nucleolus (32,4.65e-09) Intracellular Organelle (169,0.00031) Organelle (169,0.00033) Intracellular(189,0.00 105)

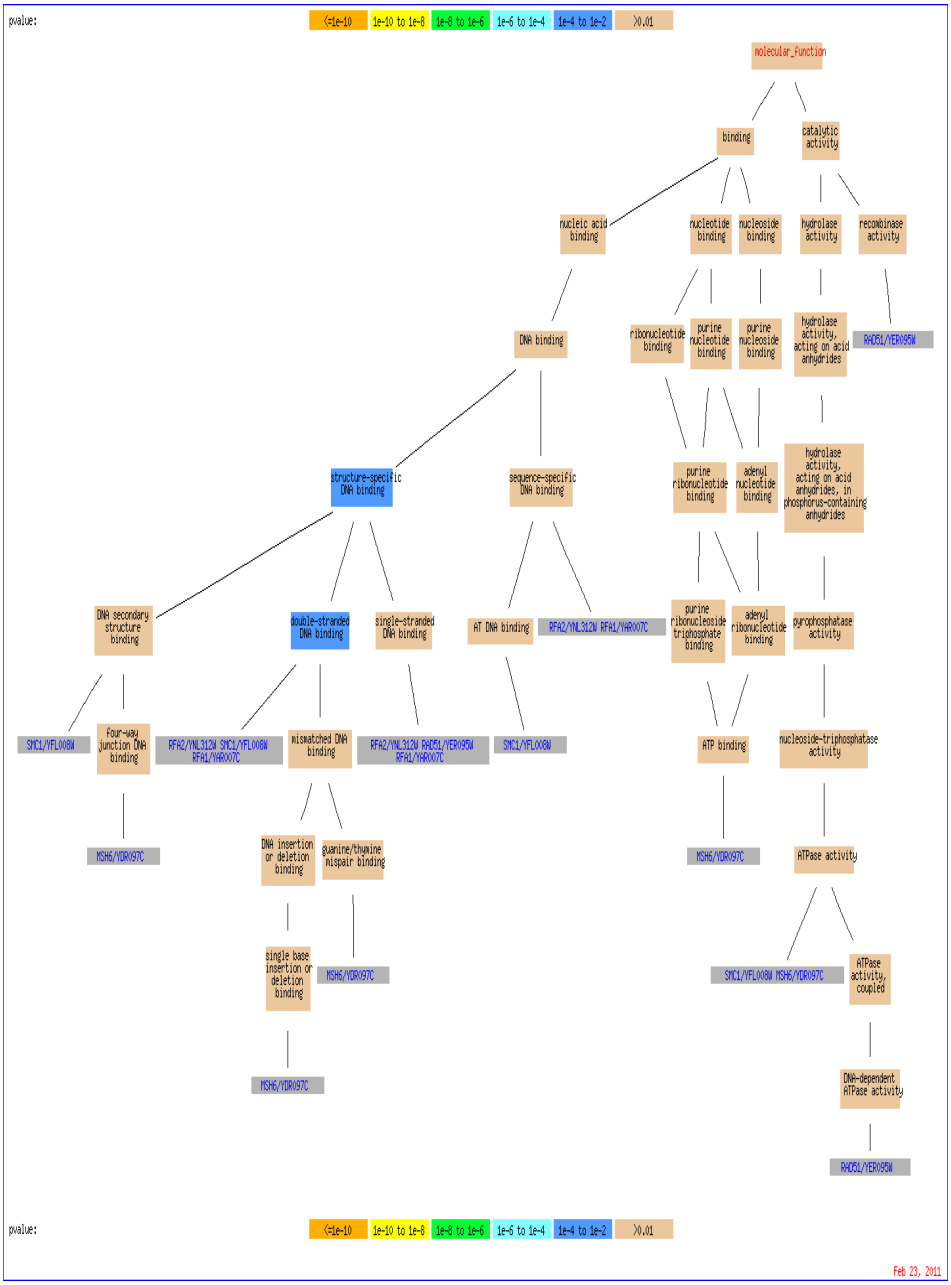


Figure 5.8: Sample of Genes for the Bicluster sc73, with corresponding GO terms and their parents for Function Ontology

Figure 5.8 shows the significant GO terms for the set of genes in bicluster SC73 along with their p-values. It shows the branching of molecular function into sub-functions binding and catalytic activity. These fuctions are subdivided further and clustered using genes to produce the final result. Figure 5.8 is obtained when gene ontology database is searched by entering the names of genes in bicluster sc73 and by selecting the function ontology.

5.1.7 Biclusters obtained Using RGRASP

In seed growing phase more conditions and genes are added to the seed from the Restricted Controlled List (RCL). RCL is formed by selecting best elements from candidate list. Candidate list is formed by those elements which can be added to the bicluster without incrementing the MSR value above the MSR threshold. From the candidate list RCL list is formed by selecting the best elements. The best elements will have an MSR value less than RCL threshold where RCL threshold= MSRmin+ α (MSRmax-MSRmin). For value based GRASP, the value of α is fixed. For reactive GRASP the value of α is selected from a discrete set of possible values. Initially all these values are given equal probability. Then the probability of αi is updated based on the quality of solution obtained. This updation will be such that, the αi with good solution will have higher probability of being selected. In this study the set of values assigned for α for condition list is {0.01, 0.02, 0.03, 0.04, 0.05, 0.06} and the set of values assigned for α for gene list is {0.0001, 0.0002, 0.0003, 0.0004, 0.0005, 0.0006}.

5.1.7.1. Bicluster Plots for Yeast Dataset

In Figure 5.9 eight biclusters obtained using RGRASP are shown. Biclusters with all 17 conditions are obtained using this method. From the bicluster plots which show strikingly similar up-regulation and down-regulation it is concluded that RGRASP is an ideal method for identifying coherent biclusters from gene expression data. All the means squared residues are lower than 205.

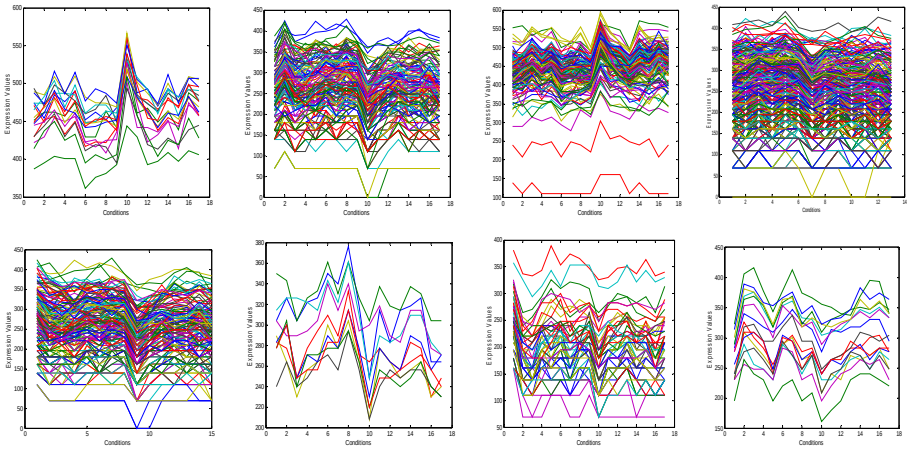


Figure 5.9 Eight biclusters found for the Yeast Dataset by RGRASP. Bicluster labels are (yar7), (ybr7), (ycr7), (ydr7), (yer7), (yfr7), (ygr7) and (yhr7) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about biclusters can be obtained from Table 5.9 using bicluster label.

Table 5.9

Information about Biclusters of Figure 5.9

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(yar7)	17	17	289	75.2721
(ybr7)	145	17	2465	202.0707
(ycr7)	107	17	1819	199.1857
(ydr7)	336	13	4368	199.8158
(yer7)	169	15	2535	199.7847
(yfr7)	10	17	170	115.9704
(ygr7)	55	17	935	199.3416
(yhr7)	16	17	272	104.2135

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 the number of columns (conditions) of the bicluster respectively. The fourth column reports the volume of the bicluster and the fifth column contains the mean squared residue or hscore of the bicluster.

5.1.7.2 Bicluster Plots for Lymphoma Dataset

This is the first time RGRASP metaheuristics is applied to find biclusters from Lymphoma dataset. Eight biclusters obtained by applying RGRASP to Lymphoma dataset are shown in Figure 5.10. Biclusters (ldr7) and (lhr7) are having very large volume.

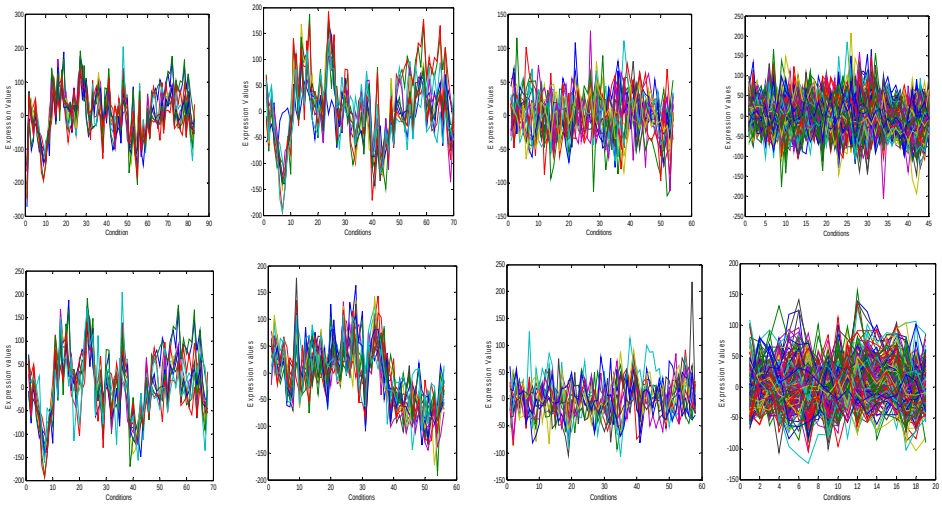


Figure 5.10: Eight biclusters found for the Lymphoma Dataset by RGRASP. The bicluster labels are (lar7), (lbr7), (lcr7), (ldr7), (ler7), (lfr7), (lgr7) and (lhr7) respectively. The details of the biclusters are given in Table 5.10

Table 5.10

Information about Biclusters of Figure 5.10

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(lar7)	10	83	830	1182.10
(lbr7)	11	70	770	1106.70
(lcr7)	20	54	1080	874.59
(ldr7)	261	45	11745	1197.70
(ler7)	11	68	748	1117.30
(lfr7)	18	56	1008	904.40
(lgr7)	15	58	870	952.83
(lhr7)	220	19	4180	961.05

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 the number 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.

5.1.7.3 Details of Significant Biclusters obtained by RGRASP

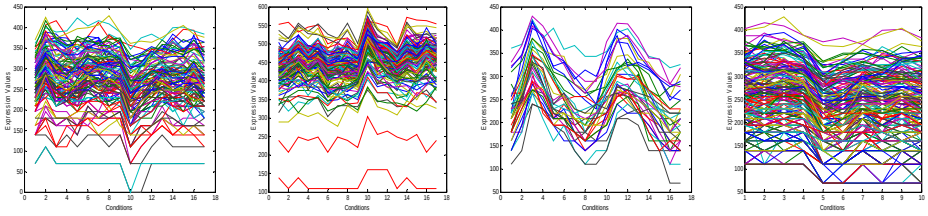


Figure 5.11 Four significant biclusters obtained by the RGRASP algorithm on Yeast dataset. The bicluster labels are sr71, sr72, sr73 and sr74. The details about the biclusters can be obtained from Table 5.11 using the bicluster label.

Table 5.11

Information about Biclusters of Figure 5.11

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
Sr71	121	17	199.9395	483.2784
Sr72	107	17	199.4776	568.0833
Sr73	36	17	297.6071	1806.9000
Sr74	224	10	204.2154	500.7598

In the first bicluster Sr71 there are 121 genes. They are YBL014C, YBL083C, YBL084C, YBR293W, YCL016C, YCL031C, YCL053C, YCL054W, YCR072C, YCR087W, YDL008W, YDL030W, YDL076C, YDL150W, YDL153C, YDL166C, YDL167C, YDL189W, YDL215C, YDL231C, YDR017C, YDR020C, YDR038C, YDR057W, YDR060W, YDR080W, YDR083W, YDR108W, YDR120C, YDR121W, YDR170C, YDR172W, YDR211W, YDR234W, YDR262W, YDR289C, YDR299W, YDR312W, YDR321W, YDR339C, YDR352W, YDR361C, YDR365C, YDR392W, YDR416W, YDR449C, YDR469W, YDR477W, YDR478W, YDR518W,

YDR524C, YDR542W, YEL015W, YEL055C, YER005W, YER075C, YER099C, YER107C, YER166W, YER168C, YER171W, YFL001W, YGL085W, YGL099W, YGL214W, YGR042W, YGR090W, YGR187C, YGR200C, YGR216C, YHR062C, YJL011C, YJL069C, YJR017C, YJR066W, YKR056W, YKR060W, YLL008W, YLL034C, YLR051C, YLR088W, YLR107W, YLR146C, YLR215C, YLR222C, YLR227C, YLR401C, YML066C, YML080W, YML093W, YMR093W, YMR211W, YMR235C, YNL041C, YNL132W, YNL163C, YNL164C, YNL199C, YNL227C, YNL299W, YNR003C, YNR038W, YOL021C, YOL022C, YOL036W, YOL080C, YOL124C, YOL140W, YOL144W, YOR006C, YOR056C, YOR061W, YOR098C, YOR145C, YOR160W, YOR252W, YOR272W, YPL126W, YPL268W, YPR053C, YPR112C.

In the second bicluster sr72 there are 107 genes namely YAL003W, YAL038W, YAR020C, YBL030C, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A,YBR106W, YBR118W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W, YDR154C, YDR353W, YDR382W, YDR385W, YDR417C, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER117W, YGL102C, YGR118W, YHR141C, YJL136C, YJL188C, YJL189W, YJL190C, YJR009C, YJR094W-A, YJR123W, YKL056C, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL066C, YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR333C, YLR340W, YLR388W, YLR406C, YLR441C, YLR467W, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR045C, YMR202W, YNL030W, YNL067W, YNL162W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W,YPL090C,YPL143W, YPL283C, YPR102C, YPR204W.

In the third bicluster sr73 there are 36 genes. They are YAR007C, YAR008W, YBL035C, YBR073W, YBR088C, YBR089W, YCR065W, YDL003W, YDL010W, YDL018C, YDL164C, YDR097C, YDR507C, YER095W, YFL008W, YGR151C, YGR152C, YHR154W, YIL026C, YJL181W, YJL187C, YKL042W, YKL113C, YLL022C, YLR103C, YLR386W, YML021C, YML102W, YMR076C, YMR078C, YNL273W, YNL303W, YNL312W, YOR074C, YPL208W, YPR120C.

The fourth seed results in a bicluster with more than 400 genes. Since there are a large number of genes, there is no search result. Hence the algorithm is executed in such a way to get only 224 genes. These genes are YAL035W, YAL059W, YAR015W, YBL004W, YBL005W, YBL014C, YBL018C, YBL024W, YBL026W, YBL037W, YBL049W, YBL054W, YBL083C, YBL084C,YBL088C, YBR002C, YBR021W, YBR032W, YBR038W, YBR060C, YBR075W, YBR076W, YBR084W, YBR094W, YBR123C, YBR133C, YBR138C, YBR155W, YBR228W, YBR257W, YBR266C, YBR267W, YBR270C, YBR293W, YBR295W, YCL012W, YCL016C, YCL054W, YCR036W, YCR043C, YCR051W, YCR062W, YCR063W, YCR072C, YCRX16C, YDL030W, YDL043C, YDL058W, YDL063C, YDL076C, YDL150W, YDL153C, YDL160C, YDL167C, YDL215C, YDL231C, YDL247W,YDR011W, YDR020C, YDR038C, YDR060W, YDR080W, YDR091C, YDR108W, YDR120C, YDR150W, YDR151C, YDR170C, YDR198C, YDR213W, YDR234W, YDR249C, YDR275W, YDR282C, YDR299W, YDR311W, YDR324C, YDR361C, YDR363W, YDR364C, YDR374C, YDR449C, YEL015W, YEL043W, YEL053C, YEL055C, YEL057C, YER005W, YER034W, YER064C, YER081W, YER107C, YER128W, YER137C, YER171W, YFL036W, YFL058W, YGL021W, YGL099W, YGL128C, YGL155W, YGL214W, YGL234W, YGR023W, YGR108W, YGR129W, YGR169C, YGR187C, YGR200C, YGR216C, YHR023W, YHR062C, YHR151C, YIL007C, YIL097W, YIL106W, YIL158W, YIL171W, YIL172C, YJL011C, YJL039C, YJL051W, YJL053W, YJR002W, YJR092W, YJR127C,

YKL057C,YKL129C, YKL173W, YKL205W, YKL222C, YKR056W, YKR060W, YLL008W, YLR014C, YLR023C, YLR068W, YLR107W, YLR131C, YLR146C, YLR190W, YLR215C, YLR222C, YLR227C, YLR277C, YLR353W, YLR430W, YLR453C, YML033W, YML034W, YML080W, YML082W, YML093W, YML103C, YMR001C, YMR021C, YMR032W, YMR033W, YMR093W, YMR132C, YMR211W, YMR235C, YMR265C, YMR278W, YMR281W, YMR291W, YNL041C, YNL049C, YNL053W, YNL124W, YNL132W, YNL163C, YNL164C, YNL171C, YNL172W, YNL196C, YNL227C, YNL299W, YNR002C, YNR003C, YNR038W, YNR039C, YOL021C, YOL028C, YOL041C, YOL042W, YOL060C, YOL080C, YOL081W, YOL113W, YOL124C, YOL130W, YOL144W, YOR006C, YOR012W, YOR061W, YOR098C, YOR145C, YOR152C, YOR160W, YOR205C, YOR206W, YOR272W, YOR315W, YOR318C, YOR364W, YPL002C, YPL126W, YPL148C, YPL150W, YPL174C, YPL183C, YPL192C, YPL205C, YPL242C, YPL248C, YPR026W, YPR040W, YPR046W, YPR079W, YPR084W, YPR112C, YPR119W, YPR129W

The Table 5.12 given below shows the significant GO terms used to describe the genes of the biclusters of Figure 5.11 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 5.12 the first entry of the second column with the title process contains the term ribosome biogenesis (44, 1.46e-23) which means that 44 out of the 121 genes of the bicluster are involved in the process of ribosome biogenesis and their p-value is 1.46e-23. Second entry indicates that 46 out of 121 genes are involved in ribonucleoprotein complex biogenesis. Also from the table it is clear that the biclusters are distinct along each category. This proves that the bicluster contains biologically similar genes and the RGRASP method used here is

capable of identifying biologically significant biclusters from different GO categories.

Table 5.12

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 5.11

Bicluster	Process	Function	Component
Sr71	Ribosome Biogenesis (44, 1.46e-23) Ribonucleoprotien complex biogenesis (46, 6.18e-23) Cellular component biogenesis at celluar Level (47, 6.22e-20) Nitrogen compound metabolic process (64, 4.18e-06)	44 out of 121 genes are directly annotated to the term molecular function unknown	Nucleolus (35,8.74e-21) Preribosome (23,5.33e-13) Nuclear part(53, 1.17e-10) cell part(112, 0.00189)
Sr72	Translation(69, 1.52e-56) cellular protein metabolic process (72, 1.13e-27) protein metabolic process(72, 8.11e-27) metabolic process (84, 9.28e-07)	Structural constituent of ribosome(62, 5.81e-62) structural molecule activity (63, 4.33e-49) translation elongation factor activity (5, 0.00011) RNA binding (15, 0.00603	Cytosolic ribosome (64, 1.42e-70) cytosolic part (64, 3.93e-64) ribosome (68, 1.10e-58) intracellular organelle (86, 0.00076)
Sr73	DNA metabolic process(19, 5.44e-11) DNA repair (16, 9.53e-11) cell cycle (20, 8.42e-10) nucleobase, nucleoside, nucleotide and nucleic acid (23, 0.00011)	Structure-specific DNA binding (5,0.00315) double-stranded DNA binding(4,0.00134)	Chromosome (15,1.21e-07) replication fork (8, 1.40e-06) Chromosomal part(13,4.93e-06) Nucleus (26, 1.52e-05)
Sr74	Ribonucleoprotein complex biogenesis(52, 4.56e-15) ribosome biogenesis (45, 1.63e-12) cellular component biogenesis at cellular level (53, 9.69e-12) nucleobase, ucleoside, nucleotide and nucleic acid metabolic process (86, 0.00032)	85 out of 224 input genes are directly annotated to root term 'molecular function unknown':	Nucleolus (36, 3.34e-12) nucleus (110, 5.96e-08) preribosome (24, 7.91e-08) nuclear part (69, 8.88e-06)

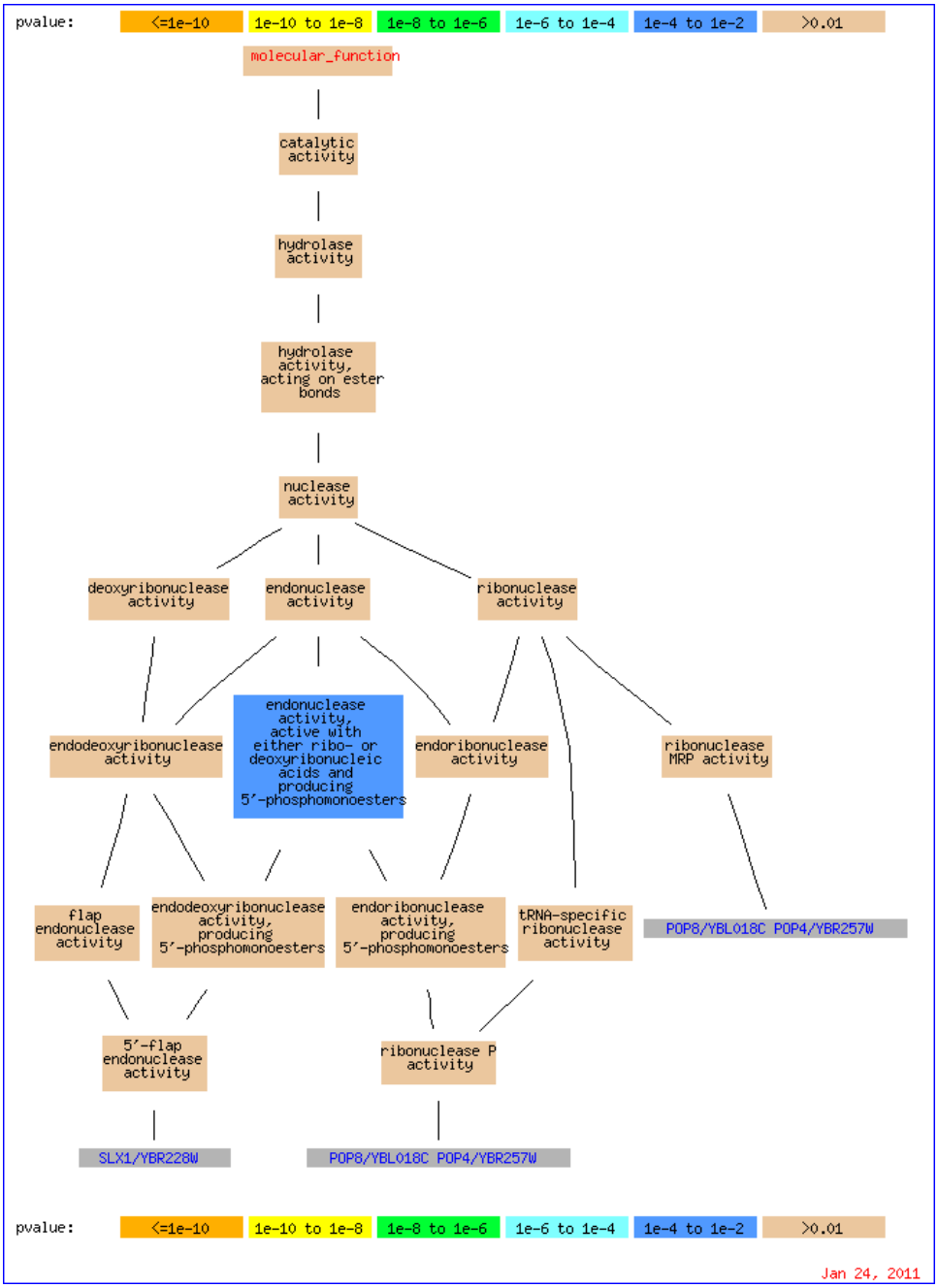


Figure 5.12 Sample of Genes for Bicluster sr74, with corresponding GO terms and their parents for the Fuction Ontology.

Figure 5.12 shows the significant GO terms for the set of genes in bicluster sr74 along with their p-values. Only 36 out of 224genes are used to search the gene ontology database to reduce the size of the Figure. It shows the branching of molecular function into sub-fuctions like catalytic activity which are further divided are divided into sub-fuctionss and clustered using genes to produce the final result. Figure 5.12 is obtained when gene ontology database is searched by entering the names of genes in bicluster sr74and by selecting function ontology.

5.1.8 Comparison with other Algorithms

5.1.8.1 Comparison on the basis of Statistical and Biological Significance

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. If the p-value is smaller, then the match will be better. In Table 5.11 the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the GRASP, CGRASP and RGRASP algorithms are compared with that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the table it is clear that in terms of best p-value obtained by a bicluster which is used to denote statistical significance, GRASP, CGRASP and RGRASP algorithms are better than all the other algorithms namely MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu for all the five GO terms. The percentage of genes involved in the first GO term for GRASP variants is better than that of all the other algorithms except MOGAB, SGAB and Bimax. The percentage of genes involved in the second, third, fourth and fifth GO terms are better than that of all the other algorithms.

Table 5.13
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by GRASP, CGRASP, RGRASP and other Algorithms for Yeast Dataset

Terms	GRASP	CGRASP	RGRASP	MOGAB	SGAB	CC	RWB	OPSM	Bimax	ISA	BiVisu
1	Cytosolic Ribosome 59.8% 1.42e-70	Cytosolic Ribosome 59.8% 1.42e-70	Cytosolic ribosome 59.8% 1.42e-70	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Ribonucleo protein complex 60.00% 9.4e-11	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 59.8% 3.93e-64	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Protein modification process 9.38% 2.8e-08	Cytosolic Part 44.44% 1.3e-10	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Structural constituent of ribosome 57.9% 5.81e-62	Structural constituent of ribosome 57.9% 5.81e-62	Structural constituent of ribosome 57.9% 5.81e-62	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Biopolymer modification 6.26% 3.1e-07	Sulfar metabolic process 16.66% 4.2e-10	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 63.6% 1.10e-58	ribosome 63.6% 1.10e-58	ribosome 63.6% 1.10e-58	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Carbohydrate metabolic process 5.93% 1.4e-06	Chromosome 19.2% 1.1e-09	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Translation 64.5% 1.52e-56	Translation 64.5% 1.52e-56	Translation 64.5% 1.52e-56	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	M phase of meiotic cell Cycle 2.44% 3.2e-05	Cellular bud 23.21% 2.4e-09	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

5.1.8.2 Comparison in Terms of Bicluster Size and MSR

The performance of GRASP algorithms in comparison with that of SEBI [36], Cheng and Church's algorithm (CC) [29], and the algorithm FLOC by Yang et al. [106] and DBF [109] are given in Table 5.14. With regard to the GRASP algorithm presented in this study, all the fields in it are better than that of SEBI, CC, FLOC and DBF. But DBF is having a lower value for the average residue score and for SEBI the average number of conditions is better than that of GRASP. In terms of average number of conditions the RGRASP in this study is better than all other algorithms listed in Table 5.12. For CC the average number of genes is better than RGRASP and CGRASP in this study. But this is due to the fact that the average number of conditions in RGRASP and CGRASP are greater than that of CC.

In this study there are biclusters with all 17 conditions for Yeast dataset. But in metaheuristic methods like multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 11 in Yeast dataset. For the Yeast dataset the maximum number of genes obtained by RGRASP in this study in all the 17 conditions is 145 with MSR value 202.0707 (label of bicluster is (ybr7) in Table 5.9). The result in this study is superior because the maximum number of genes obtained so far in a bicluster with all 17 conditions is only 141 genes for multi-objective PSO [62]. Moreover the MSR value of the bicluster (ybr7) is better (202.0707) than that of the bicluster obtained by multi-objective PSO (203.25).

Table 5.14

Performance Comparison between GRASP Variants and other Algorithms for the Yeast Dataset

Algorithm	ANG	ANC	AV	AMR	LB
GRASP	215.50	14.83	2350.33	166.85	6264
CGRASP	163.00	15.17	2292.33	181.70	3888
RGRASP	106.88	16.25	1606.63	161.96	4368
SEBI	13.61	15.25	209.92	205.18	1394
CC	166.71	12.09	1576.98	204.29	4485
FLOC	195.00	12.80	1825.78	187.54	1200
DBF	188.00	11.00	1627.20	114.70	4000

ANG is average number of genes. ANC is the average number of conditions. AV is average volume. AMR is average mean squared residue. LB is largest bicluster. As 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 5.15 given below provides a summary of results obtained by related algorithms on Lymphoma dataset. GRASP variants is not applied for finding biclusters from Lymphoma data so far. Only SEBI and CC are used for comparison in the Lymphoma dataset. Here RGRASP is better than all other algorithms in terms of average number of genes except CC. This is due to the fact that in CC average number of conditions is very low compared to RGRASP. In GRASP average number

of conditions is greater than all other algorithms. Average volume is better for CC than all other algorithms. This is due to the fact that reducing the MSR by removing one condition can result in the addition of more than 20 genes. Average MSR is better for CGRASP than all other algorithms. In metaheuristic methods like multi-objective evolutionary computation [15] the maximum number of conditions obtained is only 40 in Lymphoma dataset. In this study, using GRASP a bicluster with 89 conditions is obtained (label lva7 Table 5.2). Maximum value of conditions obtained in multi-objective PSO is only 84 for Lymphoma dataset

Table 5.15

Performance Comparison between GRASP Variants and other Algorithms for Human Lymphoma Dataset

Algorithm	ANG	ANC	A V	AMR
GRASP	61.38	69.63	3424.00	1101.35
CGRASP	35.75	17.50	528.50	479.27
RGRASP	70.75	56.25	2653.87	1037.08
SEBI	14.07	43.57	615.84	1028.84
CC	269.22	24.50	4595.98	850.04

ANG is average number of genes. ANC is the average number of conditions. AV is average volume. AMR is average mean squared residue. As clear from the above table the average mean squared residue, the average number of genes and conditions and average volume 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.

5.2 Particle Swarm Optimization (PSO)

Particle swarm optimization is a biologically inspired computing technique. In this section a PSO based algorithm developed for biclustering gene expression data is described. This algorithm has three steps. In the first step high quality bicluster seeds are generated using K-Means clustering algorithm. From these seeds biclusters are generated using particle swarm optimization. In the third stage an iterative search is performed to check the possibility of adding more genes and conditions within the given threshold value of mean squared residue score. Experimental results on real datasets show that our approach can effectively find high quality biclusters.

5.2.1 Initial Population for PSO

PSO is a population based optimization technique like genetic algorithm. Usually PSO is initialized with a population of random solutions. Here the seeds obtained from K-Means are used to initialize PSO. The advantage of initializing with seeds is that of faster convergence compared to random initialization. Another advantage of it is that it maintains diversity in the population.

5.2.2 PSO based Biclustering

The particle swarm optimization proposed by Kennedy and Eberhart [66] is a heuristics based optimization approach simulating the movements of a bird flock trying to find food. Particle swarm optimization (PSO) is a population based evolutionary computation method and the members of the whole population are maintained

throughout the search procedure. It is different from all other evolutionary-type methods in that it does not use the filtering operation such as crossover or mutation. What makes PSO extremely suitable for solving the optimization problems is its convergence speed and relative simplicity. Biclustering is an optimization problem with an objective to search for biclusters with low mean squared residue and high volume. Hence PSO is extremely suitable for solving it. Each potential solution of PSO is named as particle and is initialized with random velocity. Each particle is flown to the optimal solution in the solution space.

In the solution space of PSO each particle keeps track of its best position achieved hitherto. This is denoted by pbest (personal best). The optimal solution attained by the entire swarm is gBest (global best). PSO iteratively converts the velocity of each particle towards its pBest and gBest positions efficiently. For finding an optimal or near-optimal solution to the problem, PSO keeps updating the current generation of particles. Each particle is a candidate for the solution of the problem. The whole function is accomplished by using the information about the best solution obtained by each particle and the entire population. Each particle has got a set of attributes such as current velocity, current position, the best position discovered by the particle so far and, the best position discovered by the particle and its neighbours so far. Each particle begins with an initial velocity and position. Thereafter the n^{th} component of the new velocity and the new position for the i^{th} particle are updated in accordance with the following equations:

$$V_{i,n}(t+1)=w*V_{i,n}(t)+c1*r1[G_i(t)-X_{i,n}(t)]+c2*r2[G_i(t)-L_{i,n}(t)] \quad \text{.....(1)}$$

$$X_{i,n}(t+1)=X_{i,n}(t)+V_{i,n}(t+1) \quad \text{.....(2)}$$

In equation (1), *w* is the inertia weight; r1 and r2 are random numbers, Gi is the best particle found so far within the neighbors and Li,n is the best position discovered so far by the corresponding particle [30]. Vi,n(t+1) is the new velocity and Xi,n(t+1) is the new position of the i^{th} particle. In binary PSO [67], Vi,n that is velocity of the i^{th} particle is a probability, and it must be constrained to the interval [0, 1]. A logistic transformation S(Vi,n) can be used to attain this modification. The consequent change in the position is defined by the rule: If (rand() < S(vi,n)) then Xi,n = 1;else Xi,n = 0 where the function S(v) is a sigmoid limiting transformation and rand() is a random number selected from a uniform distribution in [0,1].

5.2.3 Fitness Function

As an optimization problem the main objective here is to search for biclusters with low mean squared residue and maximum size. (Given the value of δ (δ>0), the following fitness function can be used to assess the quality of each bicluster B [12]. G(B)= |I|.|J| if MSR(B) less than or equal to δ otherwise G(B) =δ/MSR(B). Here size of the bicluster B is IxJ.

Algorithm PSObiclustering(seeds, δ, noofpar,maxiter)
For i=1 to noofpar
Initialize particle i using seed i generated by K_Means
Initialize velocity of particle i
END (for)
While iterno<=maxiter
For each particle do
Calculate fitness value
If the fitness value is better than the best fitness value (pBest) in history set current value as the new pBest
End (for)
Choose the particle with the best fitness value of all the particles as the gBest
For each particle do
Calculate particle velocity according equation (1)
Update particle position according equation (2)
End (for)
End (while)

5.2.4 Time Complexity

The basic operation for the identification of biclusters is the calculation of mean squared residue of a submatrix. Time complexity for calculating MSR is O(mn). In order to include a gene or condition MSR value is calculated once. Hence this calculation is performed atmost P*I times where P is the number of particles and I is the number of iterations.

That means the worst case time complexity of the algorithm is O(P*I*(mn)) where m and n are the number of genes and conditions respectively.

5.2.5 Biclusters obtained Using PSO

5.2.5.1 Bicluster Plots for Yeast Dataset

Figure 5.13 shows eight biclusters obtained by Binary PSO algorithm on Yeast dataset. Some of the biclusters contain all 17 conditions. All the biclusters show strikingly similar up-regulation and down-regulation.

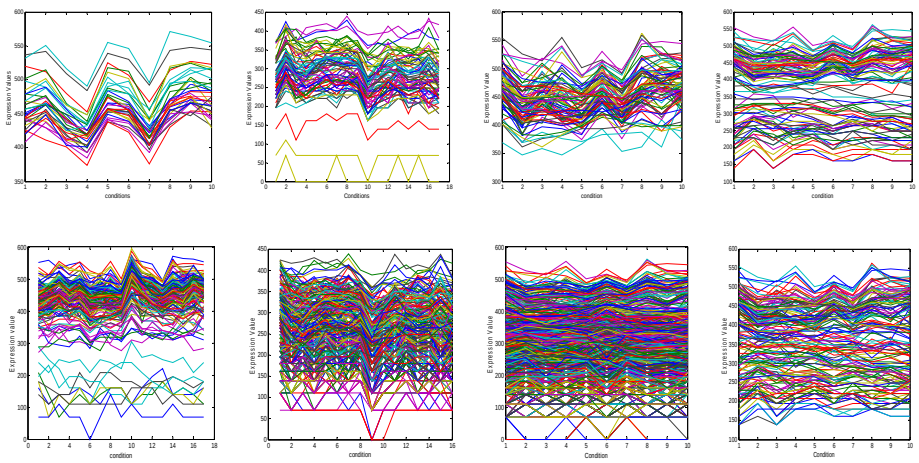


Figure 5.13 Eight biclusters found for the Yeast dataset by binary PSO. Bicluster labels are (ya8), (yb8), (yc8), (yd8), (ye8), (yf8), (yg8) and (yh8) respectively. The details about the biclusters can be obtained from Table 5.16 using bicluster label.

Table 5.16

Information about Biclusters of Figure 5.13

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(ya8)	32	10	320	63.0642
(yb8)	75	17	1275	199.5888
(yc8)	80	10	800	190.3379
(yd8)	100	10	1000	298.6600
(ye8)	136	17	2312	297.9888
(yf8)	323	16	5168	286.1017
(yg8)	1030	10	1030	299.9427
(yh8)	150	10	1500	298.8481
(yi8)	882	11	9702	299.7275
(yj8)	1399	8	11192	299.9149
(yk8)	656	12	7872	299.8829
(yl8)	848	11	9328	299.8653
(ym8)	145	17	2465	299.6139
(yn8)	318	16	5088	281.5787

In Table 5.16 given above the first column reports the label of each bicluster, the second column contains the number of rows (genes), third column contains the number of columns (conditions), fourth column contains the volume or size of the bicluster and the last column reports the mean squared residue score. Table 5.16 contains the details of some more biclusters which are not shown in Figure 5.13. The labels of these biclusters are (yi8), (yj8), (yk8), (yl8), (ym8) and (yn8).

5.2.5.2 Bicluster Plots for Human Lymphoma Dataset

In Figure 5.14 eight biclusters obtained from Human Lymphoma dataset using Binary PSO algorithm are shown. The algorithm is better for

identifying more genes than conditions where as some other metaheuristic methods like GRASP can identify more number of conditions. The maximum number of conditions obtained here is only 27. The maximum number of genes obtained is 1180.

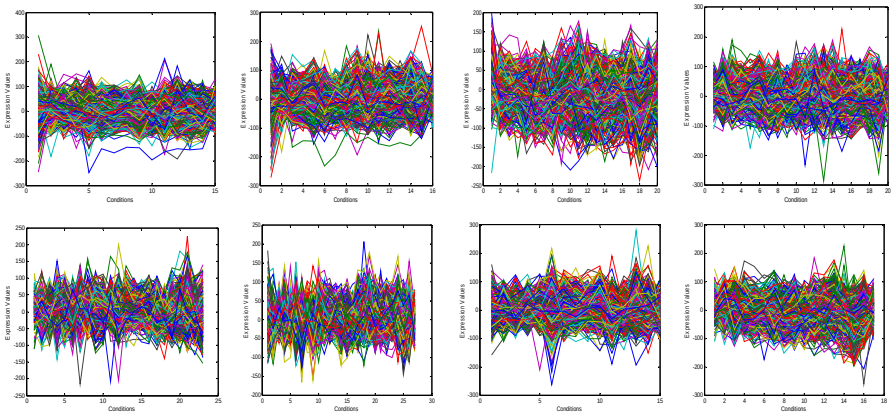


Figure 5.14 Eight biclusters found for the Lymphoma dataset by binary PSO. Bicluster labels are (la8), (lb8), (lc8), (ld8), (le8), (lf8), (lg8) and (lh8) respectively. The details about the biclusters can be obtained from Table 5.17 using bicluster label.

Table 5.17

Information about Biclusters of Figure 5.14

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(la8)	1180	15	17700	1198.8
(lb8)	1060	16	16960	1192.2
(lc8)	747	20	14940	1198.0
(ld8)	974	20	19480	1199.7
(le8)	505	23	11615	1199.9
(lf8)	339	27	9153	1199.4
(lg8)	967	15	14505	1197.7
(lh8)	836	17	14212	1199.0

5.2.6 Advantages of PSO based Biclustering

The method identifies biclusters with large number of genes from both Yeast and Lymphoma datasets. Number of iterations required for convergence is less than 100. In this method biclusters with best p-value is obtained which is better than some of the algorithms like MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu.

5.2.7 Details of Significant Biclusters obtained by PSO

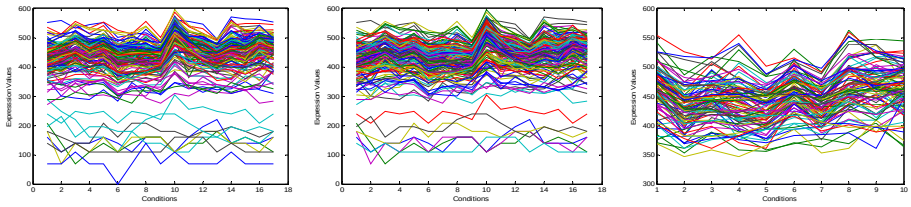


Figure 5.15 Three significant biclusters obtained by the binary PSO algorithm on Yeast dataset. The bicluster labels are s81, s82 and s83. The details about biclusters can be obtained from Table 5.18 using bicluster label.

Table 5.18

Information about Biclusters of Figure 5.15

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
S81	136	17	297.9888	587.8266
S82	145	17	299.6139	585.0625
S83	92	10	198.7709	450.1407

These biclutsters are overlapping in the sense that some genes are common. As a population based technique it is more difficult to obtain biclusters of distinct category compared to seed growing methods because in PSO all particles are flying towards the global best.

In the bicluster s81 there are 136 genes. They are YAL001C, YAL002W, YAL003W, YAL004W, YAL007C, YAL009W, YAL011W, YAL030W, YAL038W, YAR009C, YAR020C, YBL030C, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR106W, YBR111C, YBR118W, YBR181C, YBR189W, YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR064W,YDR133C, YDR154C, YDR353W, YDR382W, YDR385W, YDR418W, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER102W, YER117W, YER138C, YER160C, YGL102C, YGR118W, YHR141C, YJL136C, YJL158C, YJL177W, YJL188C,YJL189W, YJL190C, YJL225C, YJR009C, YJR094W-A, YJR123W, YJR145C, YKL006W, YKL056C, YKL060C,YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL045C, YLL066C, YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR184W, YLR185W, YLR249W, YLR325C, YLR333C, YLR340W, YLR388W, YLR406C, YLR441C, YLR467W, YML008C, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR007W, YMR045C, YMR050C, YMR074C, YMR202W, YMR230W, YNL030W, YNL067W, YNL162W, YNL209W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W, YPL090C, YPL143W, YPL283C, YPR043W, YPR102C, YPR204W.

Similarly in bicluster S82 there are 145 genes. They are: YAL001C, YAL002W, YAL003W, YAL007C, YAL009W,YAL011W, YAL030W, YAL038W, YAR009C, YAR020C, YBL030C, YBL072C, YBL077W, YBL092W, YBR009C, YBR010W, YBR031W, YBR048W, YBR078W, YBR084C-A, YBR106W, YBR118W, YBR181C, YBR206W, YCL018W, YCLX11W, YBR189W, YCR013C, YCR031C,YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W,YDL192W, YDL208W, YDL221W, YDL228C, YDL229W, YDR012W, YDR025W,YDR035W, YDR050C, YDR064W, YDR133C, YDR134C, YDR154C,

YDR158W,YDR225W, YDR276C, YDR353W, YDR382W, YDR385W, YDR417C, YDR418W, YDR433W, YDR447C, YDR450W, YDR471W, YDR500C, YEL034W, YER074W, YER102W, YER117W, YER138C, YER160C, YGL102C, YGR118W, YHR141C, YJL136C,YJL158C, YJL177W, YJL188C, YJL189W, YJL190C, YJL225C, YJR009C, YJR094W-A, YJR123W, YJR145C, YKL006W, YKL056C, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YKR094C, YLL045C, YLL066C,YLL067C, YLR029C, YLR048W, YLR062C, YLR075W, YLR076C, YLR110C, YLR167W, YLR185W, YLR249W, YLR294C, YLR325C, YLR333C, YLR340W,YLR388W, YLR406C, YLR441C, YLR467W, YML008C, YML024W, YML026C, YML039W, YML045W, YML063W, YML133C, YMR007W, YMR045C, YMR050C, YMR202W, YMR230W, YNL030W, YNL067W, YNL162W, YNL209W, YNL302C, YNL339C, YOL039W, YOL040C, YOL127W,YOR167C, YOR234C, YOR293W, YOR312C, YOR369C, YPL037C, YPL081W, YPL090C, YPL143W, YPL283C, YPR043W, YPR102C, YPR204W.

In the third bicluster s83 there are 92 genes namely YAL003W, YAL038W, YBL072C,YBL092W, YBR009C, YBR010W, YBR031W, YBR048W, YBR078W, YBR084C-A, YBR106W, YBR118W, YBR181C, YBR189W, YCL018W, YCLX11W,YCR013C, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL228C, YDL229W, YDR012W, YDR025W, YDR035W, YDR050C, YDR064W, YDR133C, YDR134C, YDR158W, YDR225W, YDR276C, YDR382W, YDR385W, YDR418W, YDR433W,YDR447C, YDR450W, YDR471W, YDR500C, YGL102C, YJL136C, YJL158C, YJL177W,YJL189W, YJL190C, YKL006W, YKL060C, YKL096W-A, YKL152C, YKL153W, YKL180W, YKR057W, YLR029C, YLR048W, YLR075W, YLR110C, YLR167W, YLR185W, YLR249W, YLR325C, YLR406C, YML024W, YML039W, YML063W, YNL030W, YNL067W, YNL162W, YNL209W, YNL302C, YNL339C,YOL039W, YOL040C, YOL127W, YOR167C, YOR234C, YOR293W, YOR312C,YPL037C, YPL081W, YPL090C, YPL143W, YPL283C, YPR043W, YPR102C.

The Table 5.19 given below shows the significant GO terms used to describe the genes of the biclusters of Figure 5.15 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 5.19, the first entry of the second column with the title process contains the term translation (80, 0.99e-62) which means that 80 out of the 136 genes of the bicluster are involved in the process of translation and their p-value is 0.99e-62. Second entry indicates that 83 out of 136 genes are involved in cellular protein metabolic process. This proves that the biclusters contain biologically similar genes and the binary PSO method used here is capable of identifying biologically significant biclusters.

Table 5.19

Significant Shared GO Terms (Process, Function, Component) of Biclusters shown in Figure 5.15

Bicluster	Process	Function	Component
S81	Translation (80, .99e-62) cellular protein metabolic Process (83, 1.29e-27) protein metabolic process (83, 1.18e-26) cellular macromolecule biosynthetic process (81, 7.36e-23) macromolecule biosynthetic process (81, 9.52e-23)	Structural constituent of ribosome (72, 7.00e-70) structural molecule activity (74, 3.27e-55) RNA-directed DNA polymerase activity (7, 2.21e-05) RNA binding (20, 0.00023)	Cytosolic ribosome (74, 1.01e-79) cytosolic part (74, 1.37e-71) ribosome (78, 2.59e-63) cytosol (80, 1.09e-60)

S82	Gene expression(84, 1.36e-19) Biosynthetic process (92, 3.52e-19) ribosome biogenesis (40, 3.21e-16) translational elongation (17, 7.81e-16)	RNA-directed DNA polymerase Activity (7, 3.47e-05) translation elongation factor activity (5, 0.00055) DNA-directed DNA polymerase activity (7, 0.00324) DNA polymerase activity (7, 0.00438)	Ribonucleoprotein complex (81, 1.59e-43) Cytosolic small ribosomal subunit (35, 3.99e-36) cytosolic large ribosomal subunit (37, 5.49e-36) small ribosomal subunit(35, 6.82e-31)
S83	Macromolecule metabolic process (69,7.47e-12) primary metabolic process (76, 4.37e-11) ribosomal small subunit biogenesis (19, 6.89e-11) cellular component biogenesis at cellular level (31, 1.03e-10)	Structural constituent of ribosome(55, 1.25e-55) structural molecule activity (57, 2.47e-46) translation elongation factor activity (4, 0.00148)	Macromolecular complex (67, 2.33e-17) cytoplasmic part (70, 7.06e-12) organelle part (62, 3.52e-08) intracellular organelle part (62, 3.52e-08) cytoplasm(72, 1.75e-05)
S84	Translation (52, 1.24e-23) ribosome biogenesis(44, 3.56e-19) ribonucleoprotien complex biogenesis (44, 1..38e-16) ncRNA metabolic process (29, 2.25e-06) rRNA transport (10, 9.55e-06)	Structural constituent of ribosome (45, 4..48e-27) structural molecule activity (52, 6.13 e-25)	Cell (140, 5.65e-05) 90S preribosome (12, 0.0004) Cytoplasmic part (84, 0.00126) Nucleolus (17, 0.00691)

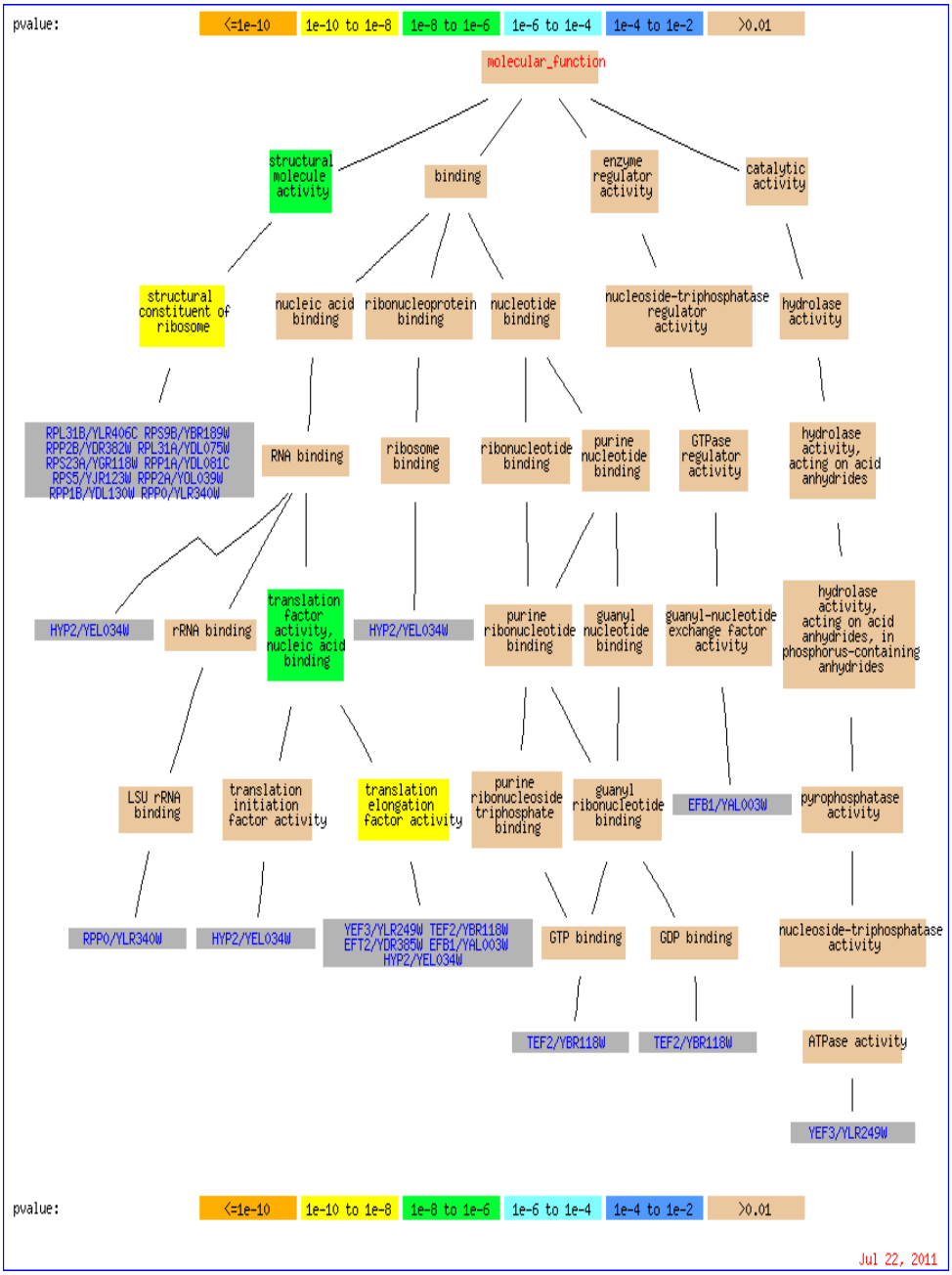


Figure 5.16 Sample of genes for the bicluster s81, with corresponding GO terms and their parents for Function Ontology

Figure 5.16 shows the significant GO terms for the set of genes in bicluster s81 along with their p-values. It shows the branching of molecular function into sub-components like structural molecule activity, binding, enzyme regulator activity and catalytic activity. These sub-functions are further divided into and are clustered using genes to produce the final result. Figure 5.16 is generated when gene ontology database is searched by entering the names of genes in bicluster s81 and by selecting function ontology. Only 17 genes (YAL003W, YBR118W, YBR189W, YDL075W, YDL081C, YDL130W, YDL229W, YDR382W,YDR385W, YEL034W, YGR118W, YJR123W, YLR249W, YLR340W, YLR406C, YNL209W, YOL039W) out of 136 genes are selected to search the GO database to reduce the size of the Figure.

5.2.8 Comparison with other Algorithms

5.2.8.1 Comparison on the basis of Statistical Significance

In Table 5.20 the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the binary PSO algorithm is compared with that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the Table 5.20 it is clear that in terms of the p-value obtained by a bicluster which is used to denote statistical significance, PSO algorithm is better than MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu for all GO terms. The percentage of genes involved for the first GO term is better than that of RWB, OPSM and BiVisu. The percentage of genes involved for the second, third and fifth GO terms are better than that of all the other algorithms mentioned in the Table 5.20. The percentage of genes involved for the third GO term is better than that of all the other algorithms except MOGAB.

Table 5.20
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by Binary PSO and other Algorithms for Yeast Data

Terms	Binary PSO	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic ribosome 54.4% 1.01e-79	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	Cytosolic Part 54.4% 1.37e-71	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	Structural constituent of ribosome 50% 7.00e-70	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromole cule biosyntheti c process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	ribosome 57.4% 2.59e-63	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	translation 58.8% 3.99e-62	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondr ial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishm ent of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

5.2.8.2 Comparison in terms of Bicluster Size and MSR

The performance of Binary PSO is compared with that of SEBI [36], Cheng and Church's algorithm (CC), and the algorithm FLOC by Yang et al. [106], DBF [109] and single objective GA [20] for the Yeast dataset are given in Table 5.21. Single objective GA (SGAB) [20] has been used with local search to generate overlapped biclusters. In terms of average number of genes, average volume and largest bicluster size Binary PSO is better than all other algorithms listed in Table 5.21. The MSR value is relatively high for Binary PSO. But for the Yeast dataset it can be within the maximum limit of 300.

Table 5.21

Performance Comparison between Binary PSO and other Algorithms for Yeast dataset

Algorithm	ANG	ANC	AMR	AV	LB
Bin. PSO	581.70	12.80	285.49	6422.70	11192
DBF	188.00	11.00	114.70	1627.20	4000
SEBI	13.61	15.25	205.18	209.92	1394
CC	166.71	12.09	204.29	1576.98	4485
FLOC	195.00	12.80	187.54	1825.78	2000
SGA	191.12	5.13	52.87	570.86	1408

ANG is average number of genes. ANC is the average number of conditions. AMR is average mean squared residue. AV is average volume. LB is the largest bicluster size. As clear from the above table the average mean squared residue, the average number of genes and conditions and average volume 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 5.22 given below lists the performance comparison of different algorithms for Human Lymphoma dataset. SEBI and CC algorithms are compared with Binary PSO. It is observed that the method is good in identifying large number of genes compared to the number of conditions. In Lymphoma dataset the biclusters obtained by Binary PSO is better than that of CC and SEBI in terms of average number of genes and average volume.

Table 5.22

Performance Comparison between Binary PSO and other Algorithms for Lymphoma Dataset

Algorithm	ANG	ANC	AMR	AV
Bin.PSO	826.00	19.13	1198.09	14820.63
SEBI	14.07	43.57	1028.84	615.84
CC	269.22	24.50	850.04	4595.98

ANG is average number of genes. ANC is the average number of conditions. AMR is average mean squared residue. AV is average volume. In the table given above the average number of genes and conditions, average volume and average mean squared residue 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.

5.3 Greedy Search-Binary PSO Hybrid

5.3.1 Initial Population for PSO

PSO is a population based evolutionary optimization algorithm. Usually PSO is initialized with a population of random solutions. In this algorithm the results obtained from greedy search algorithm mentioned in chapter 4 is used to initialize PSO. This will result in faster convergence compared to random initialization. Maintaining diversity in the population is another advantage of initializing with biclusters from greedy search method. Moreover greedy methods suffer from local minima problem which can be eliminated by methods like PSO.

5.3.2 PSO based Biclustering

Each particle of PSO explores a possible solution. It adjusts its flight according to its own and its companions flying experience. The personal best position is the best solution found by the particle during the course of flight. This is denoted by pbest (personal best). The optimal solution attained by the entire swarm is gBest (global best). PSO iteratively updates the velocity of each particle towards its pBest and gBest positions efficiently. For finding an optimal or near-optimal solution to the problem, PSO keeps updating the current generation of particles. Each particle is a candidate for the solution of the problem. The whole function is accomplished by using the information about the best solution obtained by each particle and the entire population. Each particle has got a set of attributes such as current velocity, current position, the best position discovered by the particle so far and, the best position discovered by the entire particle so far. Each particle begins with an initial

velocity and position. Thereafter a swarm particle-i will update its own speed in accordance with the following equations:

$$V(i+1)=w*Vi+ \{Cp*r1*(pBesti - Xi)\} +\{Cg*r2*(gBest-Xi)\} \text{ ---------- (3)}$$

$$X(i+1)=Xi +V(i+1) \text{ ---------- (4)}$$

In equation (1), *w* is the inertia weight; r1 and r2 are random numbers within the range {0,1}. Cp is the Cognitive learning rate and Cg is the Social learning rate. gBest is the best particle found so far and pBesti is the best position discovered so far by the corresponding particle.

In binary PSO, Vi is a probability, and it must be constrained to the interval {0, 1}. A logistic transformation S(Vi) is used to convert the value to this range. The consequent change in the position is defined by the following rule: If (rand() < S(Vi)) then Xi = 1;else Xi = 0. The function S(v) is a sigmoid limiting transformation and rand() is a random number selected from a uniform distribution in {0,1}.

5.3.3 Fitness Function

The main objective is to find maximal biclusters with low mean squared residue. Algorithm is used to maximize the objective function. Given the value of δ (δ>0), the following fitness function can be used to assess the quality of bicluster [12].

$$G(B(I,J) = |I|.|J| \quad \text{if MSR(I,J) less than or equal to } \delta$$
$$= \delta/MSR(i,j) \text{ otherwise}$$

5.3.4 Biclusters obtained Using Greedy-PSO Hybrid

5.3.4.1 Bicluster Plots for Yeast Dataset

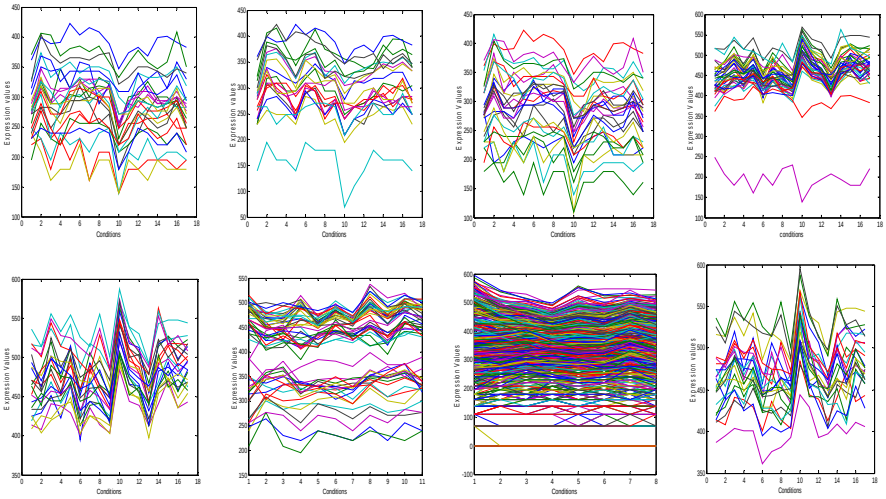


Figure 5.17 Eight biclusters obtained from the Yeast dataset by greedy-PSO. Bicluster labels are (yag8), (ybg8), (ycg8), (ydg8), (yeg8), (yfg8), (ygg8) and (yhg8) respectively. In the bicluster plots X axis contains conditions and Y axis contains expression values. The details about the biclusters can be obtained from Table 5.23 using bicluster label.

Table 5.23

Information of Biclusters of Figure 5.17

Bicluster Label	Number of Genes	Number of Conditions	Bicluster Volume	MSR
(yag8)	25	17	425	195.9666
(ybg8)	21	17	357	178.1294
(ycg8)	28	17	476	189.1636
(ydg8)	36	17	612	195.7957
(yeg8)	22	17	374	146.7061
(yfg8)	54	11	594	192.1012
(ygg8)	500	8	4000	199.4028
(yhg8)	23	17	391	150.2494

5.3.5 Details of Significant Bicluster obtained by Greedy-PSO Algorithm

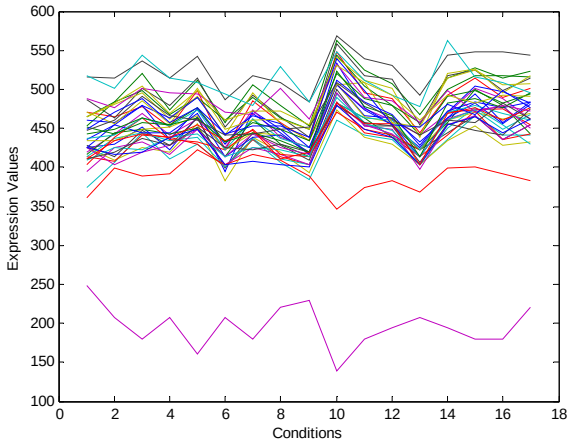


Figure 5.18 A significant bicluster obtained by the greedy-PSO algorithm on Yeast dataset. The bicluster label is sgp81. The size, MSR and row variance of the bicluster is (36*17, 195. 7957, 606.0198)

In the bicluster selected there are 36 genes namely YAL003W, YBL072C, YBL092W, YBR009C, YBR031W, YBR048W, YBR084C-A, YBR118W, YCR031C, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL192W, YDL208W, YDL228C, YDL229W, YDR012W, YDR025W, YDR050C, YDR060W, YDR064W, YDR369C, YDR382W, YDR385W, YDR447C, YDR450W, YDR471W, YJL177W, YKL180W, YOL127W, YPL037C, YPR102C.

The Table 5.24 given below shows the significant GO terms used to describe the genes of the bicluster of Figure 5.18 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 5.24 the first entry of the second column with the title 'process' contains the term translation (28, 5.00e-25) which means that 28 out of the 36 genes of the bicluster are involved in the process of translation and their p-value is 5.00e-25. Second entry indicates that 30 out of 36 genes are involved in cellular

protein metabolic Process. This proves that the biclusters contains biologically similar genes and the greedy search-binary PSO method used here is capable of identifying biologically significant biclusters.

Table 5.24

Significant Shared GO Terms (Process, Function, Component) of Bicluster shown in Figure 5.18

Bicluster	Process	Function	component
Sgp81	Translation (28, 5.00e-25) cellular protein metabolic process (30, 2.84e-15) protein metabolic process (30, 6.56e-15) cellular macromolecule biosynthetic process (28, 1.22e-11) macromolecule biosynthetic process (28, 1.34e-11)	Structural constituent of ribosome (24, 1.73e-24) structural molecule activity (24, 8.97e-20) translation elongation factor activity (.00149)	Cytosolic ribosome (24, 1.49e-25) ribosome (27, 7.35e-25) cytosolic part (24, 1.09e-23) cytosol (25, 4.68e-20)

5.3.6 Comparison with other Algorithms

5.3.6.1 Comparison on the basis of Statistical and Biological Significance

In Table 5.25 the GO terms along with their p-values and percentage of genes associated with the GO term in the bicluster for the Greedy-Binary PSO hybrid algorithm is compared with that of MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. From the Table it is clear that in terms of the p-value obtained by a bicluster which is used to

denote statistical significance, Greedy-PSO algorithm is better than RWB, Bimax, OPSM and BiVisu for the first GO term. In terms of p-value Greedy-PSO is better than all other algorithms mentioned in Table 5.25 except MOGAB and SGAB for the second GO term. In terms of p-value Greedy-PSO is better than all other algorithms mentioned in Table 5.25 except MOGAB for the third and fourth GO terms. It is better than all the other algorithms for the p-value obtained for the fifth GO term. In terms of percentage of genes involved in a GO term greedy-PSO algorithm is better than that of all the other algorithms for all the five GO terms.

Table 5.25
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by Greedy- PSO and other Algorithms for Yeast Dataset

Terms	Greedy PSO	MOGAB	SGAB	CC	RWB	Bimax	OPSM	ISA	BiVisu
1	Cytosolic Ribosome 66.7% 1.49e-25	Cytosolic Part 63.76% 1.4e-45	Cytosolic Part 60.21% 1.4e-45	Cytosolic Part 56.38% 4.2e-45	Ribosome Biogenesis & assembly 23.45% 9.3e-09	Ribonucleo protein complex 60.00% 9.4e-11	Intrecellular membrane-bound organelle 10.22% 2.8e-09	Cytosolic Part 57.27% 3.6e-44	Ribonucleo protein complex 20.63% 1.4e-20
2	translation 77.8% 5.00e-25	Ribosomal subunit 53.46% 1.6e-45	ribosome 46.21% 1.5e-25	translation 36.73% 1.5e-21	RNA metabolic process 37.82% 4.9e-08	Cytosolic Part 44.44% 1.3e-10	Protein modification process 9.38% 2.8e-08	Sulfar metabolic process 26.38% 6.9e-10	Ribosome Biogenesis & assembly 16.77% 9.5e-20
3	ribosome 75.0% 7.35e-25	translation 57.14% 3.8e-41	translation 41.45% 7.4e-24	Ribosome Biogenesis & assembly 27.33% 1.9e-15	MAPKKK cascade 15.28% 2.5 e-06	Sulfar metabolic process 16.66% 4.2e-10	Biopolymer modification 6.26% 3.1e-07	Macromolecule biosynthetic process 36.92% 2.9e-05	RNA metabolic process 18.36% 5.8e-18
4	Structural constituent of ribosome 66.7% 1.73e-24	RNA metabolic process 42.65% 8.4e-25	Chromosome 27.92% 2.3e-13	Ribonucleo protein complex Biogenesis & assembly 28.82% 2.5e-12	RNA processing 20.33% 2.6e-06	Chromosome 19.2% 1.1e-09	Carbohydrate metabolic process 5.93% 1.4e-06	Nucleic acid binding 22.54% 7.3e-04	RNA processing 13.48% 4.5e-16
5	Cytosolic Part 66.7% 1.09e-23	DNA metabolic process 38.33% 3.1e-21	RNA metabolic process 30.22% 1.3e-11	Mitochondrial part 12.52% 9.1e-12	Response to osmotic Stress 8.38% 3.9e-06	Cellular bud 23.21% 2.4e-09	M phase of meiotic cell Cycle 2.44% 3.2e-05	Establishment of cellular localization 16.28% 7.8e-04	Ribonucleo protein complex Biogenesis & assembly 10.27% 3.3e-15

5.3.6.2 Comparison based on MSR and Bicluster Size

Table 5.26 lists a comparison of results of various algorithms on Yeast data. Performance of Greedy Search- Binary PSO hybrid with that of SEBI [36], Cheng and Church's algorithm (CC) [29], and the algorithm FLOC by Yang et al. [106] and DBF [109] are given. Here biclusters with MSR less than 100, obtained from greedy search, is used as initial population of PSO. Computation time required is very less compared to greedy search running completely to attain the desired MSR. The average value of MSR for greedy binary PSO hybrid is better than all other algorithms except DBF. Average number of conditions is better than all other algorithms except SEBI. Average number of genes is better than SEBI. The largest Bicluster size is the same as DBF, and better than FLOC and SEBI.

Table 5.26

Performance Comparison between Greedy Search Binary PSO Hybrid and other Algorithms for the Yeast Dataset

Algorithm	ANG	ANC	AMR	AV	LB
GS Binary PSO	88.62	15.13	180.94	903.63	4000
DBF	188.00	11.00	114.70	1627.20	4000
SEBI	13.61	15.25	205.18	209.92	1394
Cheng-Church	166.71	12.09	204.29	1576.98	4485
FLOC	195.00	12.80	187.54	1825.78	2000
Greedy	515.57	13.36	185.86	4690.36	12645

ANG is average number of genes. ANC is the average number of conditions. AMR is average mean squared residue. AV is average volume. LB is the largest bicluster size. In the table given above the average number of genes and conditions, average volume, average mean squared residue 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.

5.4 Comparison of Greedy and Metaheuristic Algorithms

5.4.1 Comparison on the basis of Statistical Significance

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. Four different seeds, which on enlargement result in biologically significant biclusters, were selected. These seeds are enlarged by the greedy and GRASP variants and the p-values of the GO terms of these biclusters are compared for all these algorithms. Since PSO is a population based technique a significant bicluster similar to the enlargement of seed 2 is obtained for Binary-PSO and Greedy-PSO. Hence only in bicluster 2 such comparisons are given for these two algorithms.

5.4.1.1 Comparison based on p-values of GO Terms for Four Different Seeds

Table 5.27

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from First Seed and the Corresponding p-value obtained for each Algorithm for Process Ontology

GO Terms	p-value and Percentage of Genes			
	GREEDY	GRASP	CGRASP	RGRASP
Ribosome biogenesis	1.45e-23 36.7%	1.46e-23 36.7%	1.46e-23 36.7%	1.46e-23 36.7%
Ribonucleoprotein complex biogenesis	6.13e-23 38.3%	6.18e-23 38.3%	6.18e-23 38.3%	6.18e-23 38.3%
Cellular component biogenesis at cellular level	6.18e-20 39.2%	6.22e-20 39.2%	6.22e-20 39.2%	6.22e-20 39.2%
ncRNA processing	3.68e-19 32.5%	3.71e-19 32.5%	3.71e-19 32.5%	3.71e-19 32.5%
ncRNA metabolic process	1.80e-18 33.3%	1.81e-18 33.3%	1.81e-18 33.3%	1.81e-18 33.3%
rRNA processing	1.93e-15 25%	1.94e-15 25%	1.94e-15 25%	1.94e-15 25%
RNA processing	8.59e-17 35%	8.65e-17 35%	8.65e-17 35%	8.65e-17 35%
rRNA metabolic process	6.16e-17 26.7%	6.21e-17 26.7%	6.21e-17 26.7%	6.21e-17 26.7%
RNA metabolic process	3.59e-14 49.2%	3.25e-14 49.2%	3.25e-14 49.2%	3.25e-14 49.2%

In this case similar p-values and percentage of genes are obtained for greedy, GRASP CGRASP and RGRASP.

Table 5.28

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the First Seed and the Corresponding P-value obtained for each Algorithm for the Function Ontology

GO Terms	GREEDY	VGRASP	CGRASP	RGRASP
Number of genes annotated to the term molecular function unknown	44 genes	44 genes	44 genes	44 genes

From the above Table it is clear that for function ontology a fixed number of genes are annotated to the term 'molecular function unknown' for all the algorithms.

Table 5.29

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the First Seed and the Corresponding P-value obtained for each Algorithm for the Component Ontology

GO terms	p-vales and percentage of genes for each GO Term			
	GREEDY	GRASP	CGRASP	RGRASP
Nucleolus	8.74e-21 29.2%	8.74e-21 29.2%	8.74e-21 29.2%	8.74e-21 29.2%
Preribosome	5.33e-13 19.2%	5.33e-13 19.2%	5.33e-13 19.2%	5.33e-13 19.2%
90S preribosome	3.22e-08 12.5%	3.22e-08 12.5%	3.22e-08 12.5%	3.22e-08 12.5%
Nuclear part	1.28e-10 44.2%	1.17e-10 44.2%	1.17e-10 44.2%	1.17e-10 44.2%
Nuclear lumen	1.74e-10 36.7%	1.57e-10 36.7%	1.57e-10 36.7%	1.57e-10 36.7%
Organelle lumen	4.12e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%
Intracellular organelle lumen	4.12e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%
Ribonucleoprotein complex	2.32e-05 26.7%	2.32e-05 26.7%	2.32e-05 26.7%	2.32e-05 26.7%
Nucleus	8.71e-10 60%	8.71e-10 60%	8.71e-10 60%	8.71e-10 60%
Nucleolar part	1.89e-06 10%	1.89e-06 10%	1.89e-06 10%	1.89e-06 10%
Macromoleular complex	2.04e-05 50.8%	2.04e-05 50.8%	2.04e-05 50.8%	2.04e-05 50.8%
Smallsubunit processome	0.00011 7.5%	0.00011 7.5%	0.00011 7.5%	0.00011 7.5%
Organelle part	0.00016 57.5%	0.00015 57.5%	0.00015 57.5%	0.00015 57.5%
Intracellular organelle part	0.00016 57.5%	0.00015 57.5%	0.00015 57.5%	0.00015 57.5%

In this case similar p-values and percentage of genes are obtained for greedy, GRASP, CGRASP and RGRASP.

Table 5.30

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the Second Seed and the Corresponding p-value obtained for each Algorithm for the Process Ontology

GO terms	P-value and percentage of genes of GO Terms					
	GREEDY	GRASP	CGRASP	RGRASP	PSO	GREEDY -PSO
Translation	1.52e-56 64.5%	1.52e-56 64.5%	1.52e-56 64.5%	1.52e-56 64.5%	3.99e-62 58.8%	5.00e-25 77.8%
Cellularprotein metabolic process	1.13e-27 67.3%	1.13e-27 67.3%	1.13e-27 67.3%	1.13e-27 67.3%	1.29e-27 61%	2.84e-15 83.3%
Protein metabolic process	8.11e-27 67.3%	8.11e-27 67.3%	8.11e-27 67.3%	8.11e-27 67.3%	3.96e-23 61%	6.56e-15 83.3%
Cellular macromolecule biosynthetic process	6.21e-22 64.5%	6.21e-22 64.5%	6.21e-22 64.5%	6.21e-22 64.5%	7.36e-23 59.6%	1.22e-11 77.8%
Macromolecule biosynthetic process	7.76e-22 64.5%	7.76e-22 64.5%	7.76e-22 64.5%	7.76e-22 64.5%	9.52e-23 59.6%	1.34e-11 77.8%
Gene expression	3.70e-20 64.5%	3.70e-20 64.5%	3.70e-20 64.5%	3.70e-20 64.5%	1.90e-20 60.3%	1.34e-11 80.6%
Translational elongation	1.32e-14 14%	1.32e-14 14%	1.32e-14 14%	1.32e-14 14%	2.54e-16 12.5%	1.41e-08 22.2
Cellular biosynthetic process	7.02e-18 68.2%	7.02e-18 68.2%	7.02e-18 68.2%	7.02e-18 68.2%	9.12e-19 64%	1.55e-08 77.8%
Biosynthetic process	2.96e-17 68.2%	2.96e-17 68.2%	2.96e-17 68.2%	2.96e-17 68.2%	3.99e-18 64%	2.58e-08 77.8%
Ribosome biogenesis	1.49e-15 31.8%	1.49e-15 31.8%	1.49e-15 31.8%	1.49e-15 31.8%	2.85e-17 29.4%	3.01e-08 41.7%
rRNA processing	4.00e-09 20.6%	4.00e-09 20.6%	4.00e-09 20.6%	4.00e-09 20.6%	2.81e-10 19.1%	4.40e-06 30.6%
rRNA metabolic process	1.04e-08 20.6%	1.04e-08 20.6%	1.04e-08 20.6%	1.04e-08 20.6%	8.79e-10 19.1%	7.11e-06 30.6%
Cellular macromolecule metabolic process	3.82e-11 71.0%	3.82e-11 71.0%	3.82e-11 71.0%	3.82e-11 71.0%	2.03e-09 64.7%	9.46e-09 88.9%

In this case the best p-values are obtained in the order PSO, Greedy, GRASP, CGRASP, RGRASP and Greedy-PSO respectively. Similar p-values are obtained for greedy, GRASP, CGRASP and RGRASP. The order of algorithms based on the percentage of genes for the first GO term is Greedy-PSO, Greedy and GRASP Variants and PSO.

Table 5.31

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the Second Seed and the corresponding p-value obtained for each Algorithm for the Function Ontology

GO Terms	p-values and percentage of genes for GO Terms					
	GREEDY	**GRASP**	**CGRASP**	**RGRASP**	**PSO**	**GREEDY-PSO**
Structural constituent of ribosome	5.81e-62 57.9%	5.81e-62 57.9%	5.81e-62 57.9%	5.81e-62 57.9%	7.00e-70 52.9%	1.73e-24 66.7%
Structural molecule activity	4.33e-49 58.9%	4.33e-49 58.9%	4.33e-49 58.9%	4.33e-49 58.9%	3.27e-55 54.4%	8.97e-20 66.7%
Translation elongation factor activity	0.00011 4.7%	0.00011 4.7%	0.00011 4.7%	0.00011 4.7%	0.00039 3.7%	0.00149 8.3%
RNA-directed DNA polymerase activity	-	-	-	-	2.21e-05 5.1%	-
RNA binding	0.00603 14%	-	-	-	0.00023 14.7%	-
Translation elongation factor activity	-	-	-	-	0.00039 3.7%	-
DNA-directed DNA polymerase activity	-	-	-	-	0.00211 5.1%	-
DNA polymerase activity	-	-	-	-	0.00287 5.1%	-

In this case the p-values are obtained in the order PSO, Greedy GRASP, CGRASP, RGRASP, and Greedy-PSO for the first two GO terms. The order of algorithms based on the percentage of genes is Greedy-PSO, Greedy, GRASP, CGRASP, RGRASP, and PSO for the first two GO terms. Similar p-values and percentage of genes are obtained for greedy, GRASP, CGRASP and RGRASP.

Table 5.32

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the Second Seed and the corresponding p-value obtained for each Algorithm for the Component Ontology

GO Term	p-values and Percentage of Genes for GO Terms					
	GREEDY	GRASP	CGRASP	RGRASP	PSO	GREEDY -PSO
Cytosolic ribosome	1.42e-70 59.8%	1.42e-70 59.8%	1.42e-70 59.8%	1.42e-70 59.8%	1.01e-79 54.4%	1.49e-25 66.7%
Cytosolic part	3.93e-64 59.8%	3.93e-64 59.8%	3.93e-64 59.8%	3.93e-64 59.8%	1.37e-71 54.4%	1.09e-23 66.7%
Ribosome	1.10e-58 63.6%	1.10e-58 63.6%	1.10e-58 63.6%	1.10e-58 63.6%	2.59e-63 57.4%	7.35e-25 75%
Cytosol	4.32e-57 65.4%	4.32e-57 65.4%	4.32e-57 65.4%	4.32e-57 65.4%	1.09e-60 58.8%	4.68e-20 69.4%
Ribonucleoprotein complex	1.36e-43 65.4%	1.36e-43 65.4%	1.36e-43 65.4%	1.36e-43 65.4%	3.01e-46 59.6%	4.34e-23 83.3%
Cytosolic small ribosomal subunit	7.82e-32 28%	7.82e-32 28%	7.82e-32 28%	7.82e-32 28%	3.95e-37 25.7%	9.12e-08 25%
Cytosolic large ribosomal subunit	1.63e-32 29.9%	1.63e-32 29.9%	1.63e-32 29.9%	1.63e-32 29.9%	4.73e-37 27.2%	2.21e-16 41.7%
Large ribosomal subunit	5.06e-27 29.9%	5.06e-27 29.9%	5.06e-27 29.9%	5.06e-27 29.9%	3.54e-30 27.2%	2.86e-14 41.7%
Non-membrane-bounded organelle	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	2.09e-25 61%	1.13e-14 83.3%
Intracellular non-membrane-bounded organelle	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	2.09e-25 61%	1.13e-14 83.3%

In this case the p-values are obtained in the order PSO, Greedy GRASP, CGRASP, RGRASP, and Greedy-PSO for all the GO terms. The order of algorithms based on the percentage of genes is Greedy-PSO, Greedy, GRASP, CGRASP, RGRASP, and PSO for all GO terms, except for the sixth GO term. Similar p-values and percentage of genes are obtained for greedy, GRASP, CGRASP and RGRASP for all GO terms.

Table 5.33
Comparison of greedy and GRASP variants on GO terms for biclusters generated from the third seed and the corresponding P-value obtained for each algorithm for the Process ontology

GO terms	p-value and percentage of genes for GO terms			
	GREEDY	GRASP	CGRASP	RGRASP
DNA repair	9.53e-11 44.4%	9.53e-11 44.4%	9.53e-11 44.4%	9.53e-11 44.4%
Response to DNA damage stimulus	1.03e-09 44.4%	1.03e-09 44.4%	1.03e-09 44.4%	1.03e-09 44.4%
DNA metabolic process	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%
Cell cycle	8.42e-10 55.6%	8.42e-10 55.6%	8.42e-10 55.6%	8.42e-10 55.6%
cell cycle process	4.80e-09 52.8%	4.80e-09 52.8%	4.80e-09 52.8%	4.80e-09 52.8%
double-strand break repair	1.86e-07 27.8%	1.86e-07 27.8%	1.86e-07 27.8%	1.86e-07 27.8%
cellular response to stress	1.44e-07 47.2%	1.44e-07 47.2%	1.44e-07 47.2%	1.44e-07 47.2%
response to stress	4.62e-06 47.2%	4.62e-06 47.2%	4.62e-06 47.2%	4.62e-06 47.2%
mitotic sister chromatid cohesion	7.19e-06 19.4%	7.19e-06 19.4%	7.19e-06 19.4%	7.19e-06 19.4%
cellular response to stimulus	6.59e-06 50%	6.59e-06 50%	6.59e-06 50%	6.59e-06 50%
cell cycle phase	5.98e-08 44.4%	5.98e-08 44.4%	5.98e-08 44.4%	5.98e-08 44.4%
M phase	3.89e-07 38.9%	3.89e-07 38.9%	3.89e-07 38.9%	3.89e-07 38.9%
chromosome organization	3.89e-07 38.9%	7.15e-06 38.9%	7.15e-06 38.9%	7.15e-06 38.9%

In this case similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all GO terms.

Table 5.34

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the Third Seed and the Corresponding p-value obtained for each Algorithm for the Function Ontology

GO Terms	p-value and Percentage of Genes for GO Terms			
	GREEDY	GRASP	CGRASP	RGRASP
Double-stranded DNA binding	0.00341 11.1%	0.00341 11.1%	0.00341 11.1%	0.00341 11.1%
structure-specific DNA binding	0.00315 13.9%	0.00315 13.9%	0.00315 13.9%	0.00315 13.9%

In this case similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all GO terms.

Table 5.35

Comparison of greedy and GRASP variants based on GO terms for biclusters generated from the third seed and the corresponding P-value obtained for each algorithm for the Component ontology

GO Terms	p-value and percentage of genes for GO Terms			
	GREEDY	**GRASP**	**CGRASP**	**RGRASP**
Replication fork	1.40e-06 22.2%	1.40e-06 22.2%	1.40e-06 22.2%	1.40e-06 22.2%
Chromosome	1.21e-07 44.7%	1.21e-07 44.7%	1.21e-07(highest) 44.7%	1.21e-07 44.7%
Chromosomal part	4.93e-06 36.1%	4.93e-06 36.1%	4.93e-06 36.1%	4.93e-06 36.1%
Nuclear chromosome	1.53e-05 33.3%	1.53e-05 33.3%	1.53e-05 33.3%	1.53e-05 33.3%
Nuclear replication fork	0.00019 16.7%	0.00019 16.7%	0.00019 16.7%	0.00019 16.7%
Nuclear chromosome part	0.00050 27.8%	0.00050 27.8%	0.00050 27.8%	0.00050 27.8%
Condensed nuclear chromosome	0.00014 19.4%	0.00014 19.4%	0.00014 19.4%	0.00014 19.4%
Mitotic cohesin complex	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%
Nuclear mitotic cohesin complex	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%
Nucleus	1.52e-05 72.2%	1.52e-05 72.2%	1.52e-05 72.2%	1.52e-05 72.2%
Condensed chromosome	0.00030 19.4%	0.00030 19.4%	0.00030 19.4%	0.00030
Nuclear cohesin complex	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%
Cohesin complex	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%

In this case similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all GO terms.

Table 5.36

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters Generated from the Fourth Seed and the Corresponding p-value Obtained for each Algorithm for Process Ontology

GREEDY	GRASP	CGRASP	RGRASP
Ribonucleoprotein complex biogenesis 1.55e-14 22.9%	RNA processing 1.63e-06 18.9%	RNA processing 1.92 e-06 18.8%	Ribonucleoprotein complex biogenesis 4.56e-15 23.4%
Ribosome biogenesis 9.55e-13 20.2%	Ribosome biogenesis 3.86e-06 15.8%	Ribosome biogenesis 4.45e-06 15.7%	Ribosome biogenesis 1.63e-12 20.3%
Cellular component biogenesis at cellular level 3.72e-11 23.3%	ncRNA processing 6.13e-06 15.3%	ncRNA processing 7.03e-06 15.2%	Cellular component biogenesis at cellular level 9.69e-12 23.9%
RNA processing 8.10e-09 20.6%	Ribonucleoprotei n complex biogenesis 1.50e-05 16.7%	Ribonucleoprotei n complex biogenesis 1.74e-05 16.6%	RNA processing 3.04e-08 20.3%
ncRNA processing 2.31e-08 17.0%	ncRNA metabolic process 2.40e-05 15.8%	ncRNA metabolic process 2.76e-05 15.7%	ncRNA processing 8.36e-08 16.7%

In the biclusters obtained by the fourth seed, there are more than 400 genes. Hence the algorithms are executed to get only 224 genes and only these genes are used to search for GO terms of process, function and

component ontologies. Since the conditions in the biclusters are different for each algorithm, the genes selected are different, and hence the order of the GO terms is also different. Hence in Table 5.36, the GO term, p-value and percentage of genes are included in each entry. Here the order of algorithms in terms of p-value for the first and third GO terms is RGRASP, Greedy, GRASP and CGRASP. But for the second, fourth and fifth GO terms, the order of algorithms based on p-value is Greedy, RGRASP, GRASP and CGRASP. The variation in p-value for GRASP and CGRASP is very less.

Table 5.37

Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters generated from the Fourth Seed and the corresponding p-value obtained for each Algorithm for Function Ontology

GO TERM	GREEDY	GRASP	CGRASP	RGRASP
'Molecular function unknown'	Endonucleas e activity (9, 0.00591)	84 genes	84 genes	85 genes

From the Table 5.37 it is clear that for function ontology a fixed number of genes are annotated to the term 'molecular function unknown' for all algorithms except for the greedy approach.

Table 5.38
Comparison of Greedy and GRASP Variants based on GO Terms for Biclusters generated from the Fourth Seed and the corresponding p-value obtained for each Algorithm for Component Ontology

GREEDY	GRASP	CGRASP	RGRASP
Nucleolus 3.68e-12 16.1%	Nucleolus 4.08e-09 14.4%	Nucleolus 4.65e-09 14.3%	Nucleolus 3.34e-12 16.2%
Nucleus 8.02e-08 49.3%	Preribosome 0.00034 8.6%	Intracellular Organelle 0.00031 75.8%	Nucleus 5.96e-08 49.5%
Preribosome 8.27e-08 10.8%	Intracellular Organelle 0.00039 75.7%	Organelle 0.00033 75.8%	Preribosome 7.91e-08 10.8%
Nuclear part 7.92e-07 32.3%	Organelle 0.00041 75.7%	Preribosome 0.00037 8.5%	Nuclear part 8.88e-06 31.1%
90s Preribosome 3.88e-05 7.2%	Intracellular Part 0.00066 84.7%	Intracellular Part 0.00055 84.8%	90s preribosome 3.83e-05 7.2%
Nucleolar part 5.49e-05 6.3%	Intracellular 0.00123 84.7%	Membrane bounded organelle 0.00105 68.6%	nucleolar part 5.46e-05 6.3%
Nuclear Lumen 7.74 e-05 23.8%	Membrane bounded organelle 0.00134 68.5%	Intracellular Membrane bounded organelle 0.00105 68.6%	Nuclear Lumen .00035 23.0%

In this case the order of algorithms based on p-value is RGRASP, GREEDY, GRASP and CGRASP in most of the GO terms. But for the fourth and seventh GO terms, the p-vale of Greedy is better than RGRASP. For the first three seeds, these methods results in the same bicluster. For the fourth seed Greedy and RGRASP are better than CGRASP and GRASP. RGRASP is better than Greedy for some GO terms. There are also GO terms for which Greedy is better than RGRASP.

Table 5.39

Result of Biological Significance Test: The Top Five Functionally Tnriched Significant GO Terms Produced by Greedy and Metaheuristic Algorithms for Yeast Dataset

Terms	GREEDY	GRASP	CGRASP	RGRASP	PSO	GREEDY-PSO
1	Cytosolic ribosome 59.8% 1.42e-70)	Cytosolic ribosome 59.8% 1.42e-70)	Cytosolic ribosome 59.8% 1.42e-70)	Cytosolic ribosome 59.8% 1.42e-70)	Cytosolic ribosome 54.4% 1.01e-79	Cytosolic Ribosome 66.7% 1.49e-25
2	Cytosolic Part 59.8% 3.93e-64)	Cytosolic Part 59.8% 3.93e-64)	Cytosolic Part 59.8% 3.93e-64)	Cytosolic Part 59.8% 3.93e-64)	Cytosolic Part 54.4% 1.37e-71	translation 77.8% 5.00e-25
3	structural constituent of ribosome 57.9% 5.81e-62)	Structural constituent of ribosome 57.9% 5.81e-62)	Structural constituent of ribosome 57.9% 5.81e-62)	structural constituent of ribosome 57.9% 5.81e-62)	Structural constituent of ribosome 50% 7.00e-70	ribosome 75.0% 7.35e-25
4	ribosome 63.6% 1.10e-58)	ribosome 63.6% 1.10e-58)	ribosome 63.6% 1.10e-58)	ribosome 63.6% 1.10e-58)	ribosome 57.4% 2.59e-63	Structural constituent of ribosome 66.7% 1.73e-24
5	Translation 64.5% 1.52e-56)	Translation 64.5% 1.52e-56)	Translation 64.5% 1.52e-56)	Translation 64.5% 1.52e-56)	translation 58.8% 3.99e-62	Cytosolic Part 66.7% 1.09e-23

5.4.1.2 Comparison based on best Five GO Terms

Here in Table 5.39 all the algorithms are compared on the basis of the best 5 p-values obtained from all the four biclusters. In this case the order of algorithms based on p-value is PSO, RGRASP, GREEDY, CGRASP, GRASP and GREEDY-PSO for all GO terms. The p-value is the same for GREEDY, GRASP, CGRASP and RGRASP for all GO terms. In terms of percentage of genes involved Greedy-PSO is better than all the other methods for all the five GO terms.

5.4.2 Comparison of Algorithms based on Size and MSR

Three different seeds are selected. These seeds are enlarged by Greedy and GRASP variants. The bicluster size and MSR are compared for biclusters obtained from all these algorithms. From Table 5.40 it is clear that for the three seeds, same bicluster is obtained by greedy, RGRASP, GRASP and CGRASP. But due to randomization in GRASP, if the conditions selected are different, these algorithms result in different biclusters. Analysing the algorithms based on the biclusters obtained from the same seed, it should be noted that among the algorithms Greedy, GRASP, CGRASP and RGRASP result in different biclusters, only if the conditions selected are different. These algorithms then differ in the order in which the genes are added. The local search phase also results in the addition of similar genes.

Table 5.40

Comparison of Size and MSR of 3 Biclusters obtained by Enlarging 3 Different Seeds by Greedy and GRASP Variants

Biclusters	Greedy			GRASP			RGRASP			CGRASP		
	Size	MSR	Row Variance	Size	MSR	Row variance	Size	MSR	Row variance	Size	MSR	Row Variance
1	121*17	199.94	483.2784	121*17	199.94	483.2784	121*17	199.94	483.2784	121*17	199.94	483.2784
2	107*17	199.48	568.0833	107*17	199.48	568.0833	107*17	199.48	568.0833	107*17	199.48	568.0833
3	36*17	297.61	1806.9	36*17	297.61	1806.9	36*17	297.61	1806.9	36*17	297.61	1806.9

5.5 Summary

In this chapter algorithms based on the metaheuristic methods GRASP and its variants, PSO and greedy-PSO hybrid are used for finding biclusters in gene expression data. The algorithms are implemented on the Yeast dataset and also the Human Lymphoma dataset. This is the first time that GRASP metaheuristics and its variants are applied for identifying biclusters from Human Lymphoma dataset. The biologically significant biclusters obtained from these algorithms are compared with other algorithms. In terms of the best p-value obtained GRASP, CGRASP, RGRASP and PSO algorithms are better than that of MOGAB, SGAB, CC, RWB, OPSM, Bimax, ISA and Bivisu. The metaheuristic algorithms and Greedy approach are compared based on p-value, bicluster size and MSR. It is found that when the conditions selected are different these algorithms result in different biclusters.

Chapter 6

Performance Evaluation of MSR Based Algorithms

In this chapter the performance of all MSR based algorithms are evaluated based on the quality of biclusters obtained. High row variance is an important quality of the bicluster. MSR has a problem in the detection of biclusters with highly significant change in the expression level. This problem is clearly illustrated in this chapter. Constraint based algorithms SGSC and MSRT solve this problem to a certain extent compared to all other algorithms which are trying to minimize MSR, including the metaheuristic and greedy approach developed in this study. The performance of all these algorithms are also evaluated and compared based on the other qualities of bicluster namely bicluster size, MSR and p-values obtained for different GO terms.

6.1 A Critical Problem of MSR in the Identification of Biclusters with High Row Variance

The mean squared residue introduced by Cheng and Church has become one of the most popular measures to identify biclusters in most of the biclustering algorithms. In this section a critical problem with MSR in detecting highly significant biclusters is discussed by giving examples from both Yeast and Lymphoma datasets. These biclusters are highly significant because the row variance of some of these biclusters is far greater than the row variance of all the biclusters detected so far by any other algorithm using MSR. These biclusters are also coherent even though their MSR value exceeds the predefined MSR threshold. Cheng and Church [29] defined a bicluster as a uniform submatrix having low Mean Squared Residue (MSR). MSR is used to compute the coherence among the group of genes. There is a threshold value denoted by δ for MSR which depends on the dataset. Many biclustering algorithms were developed using MSR.

6.1.1 Relationship between Row Variance and MSR

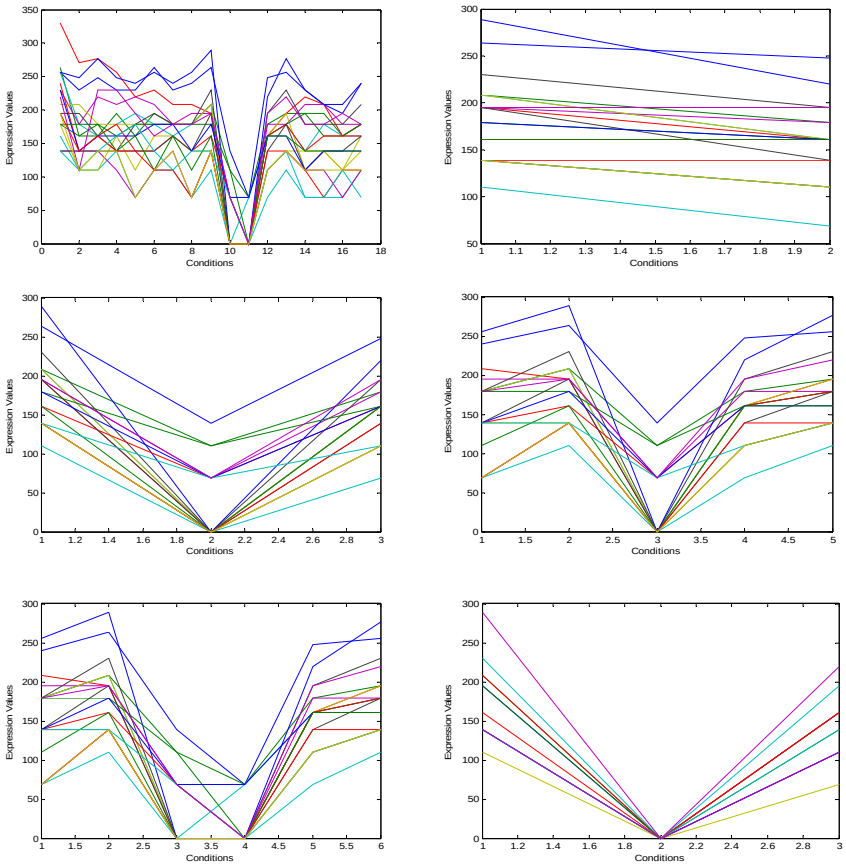
Algorithms using mean squared residue uses row variance as an accompanying score to eliminate trivial biclusters. Biclusters with high row variance are more interesting because they make significant changes in the expression level of the genes. Hence they are biologically more relevant. According to the general notion it is assumed that the biclusters should have low MSR and high row variance. Now the question is how MSR and row variance are related. The MSR is used for measuring the

variance of the set of all elements in the bicluster, plus the mean row variance plus the mean column variance [10]. From this statement it is clear that row variance forms an incremental factor in the calculation of MSR. It is observed that genes having low row variance fills the MSR value by small amounts so that such biclusters can accommodate more genes whereas biclusters with high row variance fill the MSR value by larger amounts, so that only few genes can be accommodated within the given MSR threshold. MSR depends on row variance, column variance and the variance of the set of all elements in the bicluster. Hence it is found that when one more condition in which the genes are expressing similarly is added to a bicluster and the row variance is making only a slight variation, then sometimes both MSR and row variance are increasing and sometimes one is increasing and the other is decreasing, and in some other situations both are decreasing. The problematic situation is when both are increasing because biclustering algorithms are trying to maximize row variance and minimize the MSR. MSR is minimized thinking that the increase in the value is due to the lack of coherence. This may not be true always because some of the conditions which make a significant increase in the row variance will make a significant increase in the MSR value also. So in the optimization methods which are trying to minimize MSR, there is least chance of identifying such biclusters. Sometimes this increase in MSR will be above the predefined MSR threshold value of the dataset so that the biclustering algorithm using the MSR will never identify such biclusters with highly significant variation in the expression level. These facts are

established by giving example biclusters from Yeast and Lymphoma datasets.

6.1.2 Biclusters from Yeast dataset

Some biclusters which can clearly illustrate the problem of MSR are given in Figure 6.1. From one of these biclusters (yc9) it can be noticed that how the MSR increases above the predefined MSR threshold and how row variance increases abruptly just by adding a single condition.



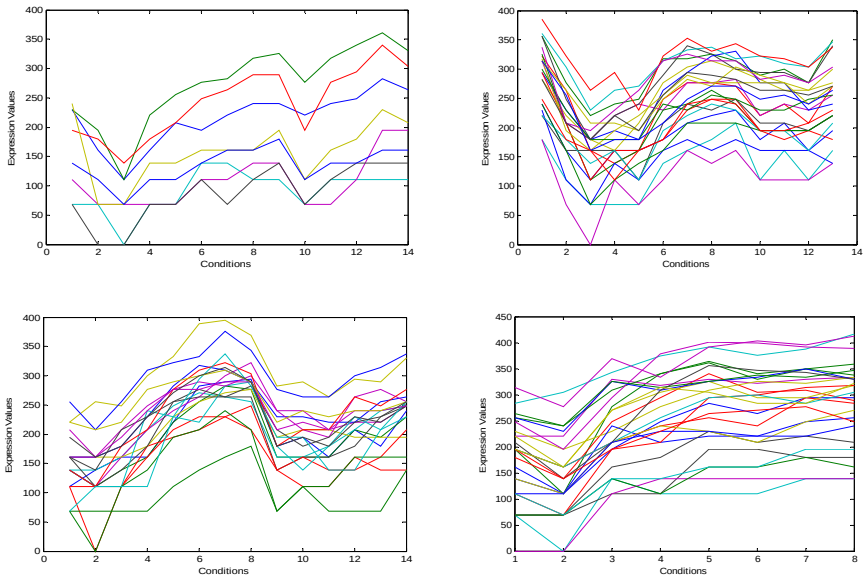


Figure 6.1 Ten biclusters from the Yeast dataset. From left to right and top to bottom the bicluster labels are: ya9, yb9, yc9, yd9, ye9, yf9, yg9, yh9, yi9 and yj9 respectively. The details about the biclusters can be obtained from Table 6.1 using bicluster label.

Table 6.1

Information about Biclusters shown in Figure 6.1

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
ya9	20	17	598.59	2930.10
yb9	20	2	88.17	271.77
y c9	20	3	608.22	4771.60
yd9	20	5	555.58	3528.90
ye9	20	6	658.84	5401.50
yf9	12	3	477.11	6509.40
yg9	8	14	510.67	2263.60
yh9	24	13	322.83	1782.00
yi9	21	14	416.47	2609.20
yj9	27	8	458.39	3085.80

The set of genes and conditions shown with the label ya9 is obtained by expanding a seed from K-Means clustering algorithm under 17 conditions without imposing any constraints on conditions. The set of 20 genes in all 17 conditions are shown to clarify how significantly the expression level changes from its normal level in two conditions. In ya9 genes are not expressing similarly under all 17 conditions.

6.1.3 MSR and Row Variance Increase Significantly by the Addition of a Single Condition

A bicluster is shown with label yb9 which contains the same 20 genes in the bicluster plot labelled ya9. In the bicluster labelled yb9 there are two conditions. The genes in yb9 present similar behaviour under these two conditions. When one more condition with a significant change in the expression level is included to the bicluster yb9, then the bicluster yc9 is obtained. The row variance of yc9 is 4771.60 whereas the row variance of yb9 is only 271.77. Similarly the MSR value of yc9 is 608.22 where as for yb9 the MSR is only 88.17. Thus the addition of a single condition increases the MSR above the predefined MSR threshold and row variance increases from 271.77 to 4771.60. This high variation in MSR is not due to the lack of coherence but because of the significant change made in the expression level of the genes which is denoted by row variance. Biclusters with labels yd9 and ye9 are also obtained from the same set of genes and has high row variance and MSR.

6.1.4 Row Variance and MSR are very high even for Genes Converging to a Single Point

From the bicluster labeled yc9 it is clear that the 20 genes in this bicluster are divided into four groups depending on the point to which the expression value reaches. In one of the 20 genes, the expression value changes from 264 to 139. Two of them reach the value 110. Five of them reach the value 69. Twelve of them reach the value 0. In order to see the effect of genes converging to a single point, only such genes are selected from bicluster yc9 whose expression level reaches 0. The bicluster with such genes are shown as bicluster labeled yf9. Even though the genes are converging to the same point, the MSR value is above the threshold which is 477.11 and row variance is 6509.40. When the condition which makes the significant change is removed from this bicluster the MSR value is only 93.30 and the row variance is only 393.75.

6.1.5 A Bicluster with the Highest Row Variance Identified

Biclusters with labels yg9 to yj9 contain another set of genes with high row variance and MSR above the threshold. The maximum row variance among different biclustering algorithms for the Yeast Dataset is obtained by Cheng and Church [29] and the value they obtained is 4162. But in this study the row variance is above 6000 for bicluster yf9. For the Yeast dataset MSR threshold is only 300. Hence no algorithm using MSR can identify biclusters from yc9 to yj9. From the bicluster plots it is clear that the genes present a similar behavior in the biclusters from yc9 to yj9 even though their MSR value is above the threshold. The row variance of these biclusters is also very high.

6.1.6 Biclusters from Human Lymphoma Dataset

Some biclusters from Human Lymphoma dataset which can illustrate the problem of MSR are shown in Figure 6.2.

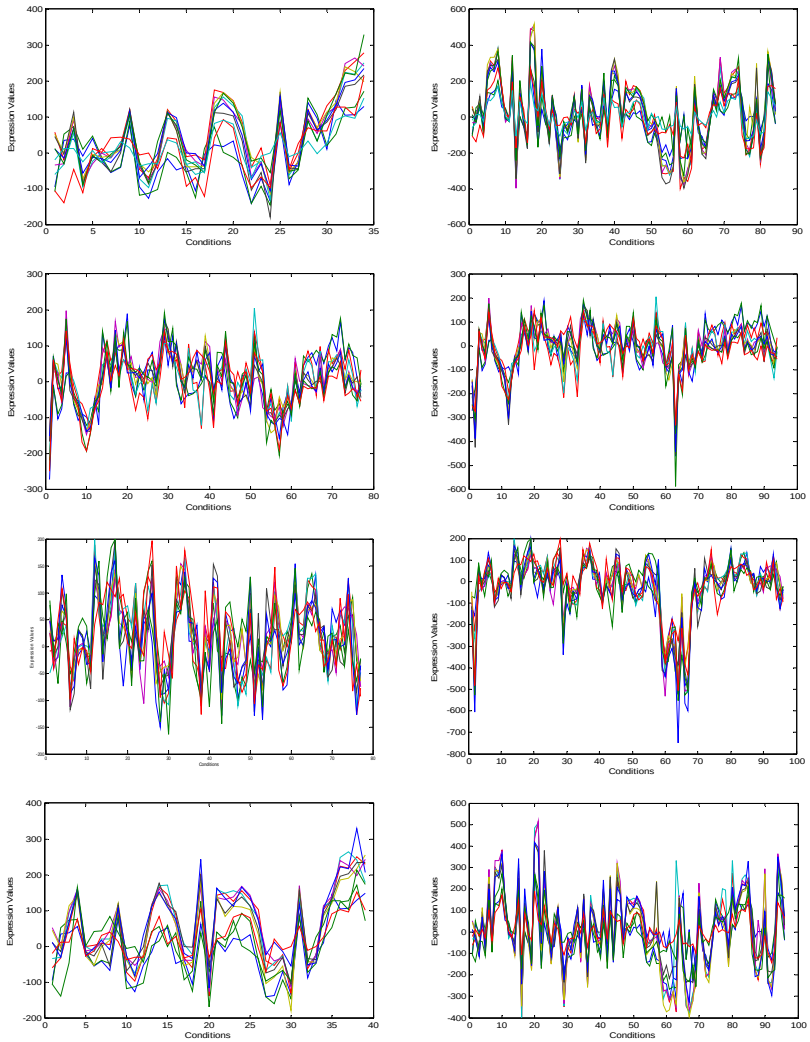


Figure 6.2 Eight Biclusters from the Lymphoma Dataset. From left to right and top to bottom Bicluster labels are la9, lb9, lc9, ld9, le9, lf9, lg9 and lh9 respectively. The details about the biclusters can be obtained from Table 6. 2 using bicluster label. Biclusters lb9, ld9, lf9 and lh9 are obtained from la9, lc9, le9 and lg9 respectively.

Their row variance is very high showing that there is significant change in the expression level. They are coherent even though their MSR value is above the predefined MSR threshold.

Table 6.2

Information about Biclusters of Figure 6.2

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
la9	11	34	1142.0	7936.9
lb9	11	84	3927.8	27674.0
lc9	10	77	1187.7	5428.8
ld9	10	94	1562 .0	8813.4
le9	10	77	1140.5	4630.4
lf9	10	96	2092.5	19160.0
lg9	10	39	1173.3	8691.7
lh9	10	96	4522.8	25431.0

Biclusters la9, lc9, le9 and lg9 are identified by enlarging seeds from K-Means by adding more conditions. The seed bicluster will contain some conditions. All other conditions are verified for inclusion in the bicluster. An added condition is removed if the MSR value of the resulting bicluster exceeds the MSR threshold as in SGSC and MSRT algorithms. Bicluster lb9 is obtained from la9 by adding more conditions and by checking visually using bicluster plot whether the increase in MSR above the threshold value is due to the lack of coherence or significant change. In the same way Biclusters ld9, lf9 and lh9 are obtained from lc9, le9 and lg9 respectively. In the bicluster plot la9 which is obtained by

enforcing MSR threshold, the Y axis varies from -200 to 400. But in lb9 which is obtained from la9, Y axis varies from -600 to 600. Similar difference in the range of Y axis can be observed in biclusters lc9 and ld9, le9 and lf9, lg9 and lh9 respectively. This clearly indicates that the conditions which make significant variations are eliminated by enforcing MSR threshold. The row variance of bicluster lb9 is 27674. This is far above the row variance obtained so far by algorithms using MSR. The previous instance of maximum value of row variance for Lymphoma data is obtained by ISA [56] and the value is only 14682.47 [75]. The significant changes in the expression levels of the genes which result in high row variance can be verified from the bicluster plots in Figure 6.2.

The MSR threshold value for the Lymphoma dataset is 1200. This threshold value prevents the identification of highly coherent biclusters such as lb9, ld9, lf9 and lh9 shown above. It is difficult for other biclustering algorithms to identify even the biclusters la9, lc9, le9 and lg9 even though their value is less than the MSR threshold. This is because these biclusters are having significant variation in the expression level denoted by their row variance. These biclusters were identified by algorithms MSRT and SGSC which allowed maximum possible variation for MSR. In these algorithms an added condition is removed if it exceeds the MSR threshold. Hence it allows maximum variation for MSR. The objective of other biclustering algorithms is to minimize MSR. When the objective is to minimize MSR, conditions which do not make significant change will get more preference than the conditions which make significant change. It is because in the latter case incremental increase in the MSR will be greater than the former.

When the MSR exceeds the predefined threshold it prevents the inclusion of other conditions and genes which are coherent. For biclustering algorithms which enlarge seeds by adding more genes and conditions, incremental increase in MSR above the MSR threshold could be a situation in the intermediate stage due to the addition of some conditions with significant change. In such cases the conditions will have to be removed. If those conditions are retained, even though the MSR of the bicluster is greater than the threshold, after adding more genes to the bicluster the MSR value will get reduced. For example, in the case of bicluster shown in Figure 6.1 with label yh9, MSR value is 322.83 and row variance is 1782. There are 24 genes in this bicluster. But when there were only 10 genes in the bicluster, the MSR value was 367.5 and the row variance was 2040.5. This means that both MSR and row variance got reduced after adding more genes. Sometimes such additions will reduce the MSR below the threshold.

In short, some conditions which make significant changes in the expression level are not included in the bicluster due to the value of MSR threshold. So the knowledge discovered by the algorithm is that the genes are exhibiting similar expression levels only under X conditions. In fact the genes are coherent under Y conditions. Here Y is greater than X. In this context SGSC and MSRT algorithms are better than all other algorithms mentioned in this study because in these algorithms maximum possible variation is allowed for MSR.

6.2 Comparison of Biclusters Generated from Four Different Seeds by MSR based Algorithms

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. P-value indicates the statistical significance of a bicluster. Four different seeds, which in the event of enlargement results in biologically significant biclusters, were selected. These seeds are enlarged by all the eight algorithms and the p-values of the GO terms of these biclusters are compared for all these algorithms. Since PSO is a population based technique a significant bicluster similar to the enlargement of seed 2 is obtained for Binary-PSO and Greedy-PSO. Hence only in bicluster 2 such comparisons are given for these two algorithms. All the eight seed growing algorithms are also compared based on bicluster size and MSR by enlarging the three different seeds. These comparisons are given in this chapter.

6.2.1 Comparison based on p-values obtained for GO Terms

The first seed was enlarged by all the algorithms developed. The names of the genes in each bicluster are found out. Then the names of the genes are entered into the gene ontology database and GO terms for process, function and component ontology are searched. Terms for each ontology, the corresponding p-value and the percentage of genes involved in a particular ontology are given in the following tables. The findings derived from each Table are given after the Table and the final conclusion is summarized at the end of the chapter.

From Table 6.3 it is clear that similar p-values are obtained for greedy, RGRASP, GRASP AND CGRASP. The order of algorithms based on p-value is greedy, RGRASP, GRASP, CGRASP, ISIMSRDT, MSRT, MSRDT and SGSC respectively based on the first GO term. The p-values obtained for SGSC is very low. The reason is that since the difference threshold value assigned for genes is very low, there are only 23 genes in the bicluster. By increasing this value more genes will be included, and this will increase the p-value of GO terms for SGSC algorithm. The p-value and the percentage of genes involved for greedy and GRASP variants are the same for all the GO terms. For all GO terms the p-value obtained by greedy and GRASP variants are better than all the other algorithms except for the last GO term. For the last GO term, the p-value obtained by ISIMSRDT is the best. The order of algorithms based on the percentage of genes involved for the first GO term is ISIMSRDT, greedy, GRASP variants, MSRT, SGSC and MSRDT. For the first three GO terms, the percentage of genes involved is the highest for ISIMSRDT algorithm. For all other GO terms except the third and the seventh, the percentage of genes involved in SGSC is the highest.

Table 6.3

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from First Seed and the Corresponding P-value obtained for each Algorithm for Process Ontology

GO TERMS	p-value							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	VGRASP	CGRASP	RGRASP
ribosome biogenesis	8.41e-11 36.1%	4.78e-05 23.4%	3.08e-22 39.8%	0.00248 34.8%	1.45e-23 36.7%	1.46e-23 36.7%	1.46e-23 36.7%	1.46e-23 36.7%
ribonucleoprotein complex biogenesis	1.47e-09 36.1%	7.68e-05 24.7%	6.25e-21 40.8%	0.00622 34.8%	6.13e-23 38.3%	6.18e-23 38.3%	6.18e-23 38.3%	6.18e-23 38.3%
cellular component biogenesis at cellular level	1.08e-08 37.7%	0.00039 26.0%	1.68e-18 41.8%	---	6.18e-20 39.2%	6.22e-20 39.2%	6.22e-20 39.2%	6.22e-20 39.2%
ncRNA processing	2.95e-08 31.1%	0.00067 20.8%	1.86e-15 32.7%	0.00171 34.8%	3.68e-19 32.5%	3.71e-19 32.5%	3.71e-19 32.5%	3.71e-19 32.5%
ncRNA metabolic process	1.66e-07 31.1%	0.00247 16.9%	4.05e-14 32.7%	0.00352 34.8%	1.80e-18 33.3%	1.81e-18 33.3%	1.81e-18 33.3%	1.81e-18 33.3%
rRNA processing	5.98e-07 24.6%	0.00116 16.9%	5.80e-15 27.6%	0.00144(highest) 30.4%	1.93e-15 25%	1.94e-15 25%	1.94e-15 25%	1.94e-15 25%
RNA processing	8.40e-07 32.8%	0.00209 23.4%	2.74e-12 33.7%	-----	8.59e-17 35%	8.65e-17 35%	8.65e-17 35%	8.65e-17 35%
rRNA metabolic process	1.14e-06 24.6%	0.00194 16.9%	2.06e-14 27.6%	0.00194 30.4%	6.16e-17 26.7%	6.21e-17 26.7%	6.21e-17 26.7%	6.21e-17 26.7%
RNA metabolic process	2.09e-05 45.9%	0.00832 36.4%	3.08e-14 53.1%	0.00184 56.5%	3.59e-14 49.2%	3.25e-14 49.2%	3.25e-14 49.2%	3.25e-14 49.2%

Table 6.4

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from First Seed and the Corresponding p-value obtained for each Algorithm for the Function Ontology

GO Terms	p-value							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	VGRASP	CGRASP	RGRASP
Number of genes annotated to the term molecular function unknown	27 genes Out of 61 genes	32 genes Out 0f 77 genes	-	10 out of 23genes	44 genes	44 genes	44 genes	44 genes
snoRNA binding	-	-	0.00480	-	-	-	-	-

From the Table 6.4 it is clear that for function ontology a fixed number of genes are annotated to the term 'molecular function unknown' for all the algorithms except ISIMSRDT. For ISIMSRDT algorithm, 4 genes from the bicluster are annotated to the term snoRNA binding and its p-value is 0.0048.

Table 6.5

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from First Seed and the Corresponding P-value obtained for each Algorithm for the Component Ontology

GO TERMS	p-value							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	VGRASP	CGRASP	RGRASP
nucleolus	2.91e-11 31.1%	8.24e-05 18.2%	2.56e-19 31.6%	0.00622 26.1%	8.74e-21 29.2%	8.74e-21 29.2%	8.74e-21 29.2%	8.74e-21 29.2%
preribosome	8.40e-10 24.6%	0.00156 13%	4.26e-15 23.5%	--	5.33e-13 19.2%	5.33e-13 19.2%	5.33e-13 19.2%	5.33e-13 19.2%
90S preribosome	7.50e-09 19.7%	0.00210 10.4%	1.56e-09 15.3%	--	3.22e-08 12.5%	3.22e-08 12.5%	3.22e-08 12.5%	3.22e-08 12.5%
nuclear part	1.79e-06 45.7%	--	2.46e-12 50.0%	--	1.28e-10 44.2%	1.17e-10 44.2%	1.17e-10 44.2%	1.17e-10 44.2%
nuclear lumen	3.56e-06 39.3%	--	1.59e-13 43.9%	--	1.74e-10 36.7%	1.57e-10 36.7%	1.57e-10 36.7%	1.57e-10 36.7%
organelle lumen	1.04e-05 42.6%	--	6.41e-11 44.9%	--	4.12e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%
intracellular organelle lumen	1.04e-05 42.6%	--	6.41e-11 44.9%	--	4.12e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%	3.47e-09 39.2%
ribonucleoprotein complex	9.71e-05 32.8%	--	4.24e-07 31.6%	--	2.32e-05 26.7%	2.32e-05 26.7%	2.32e-05 26.7%	2.32e-05 26.7%
nucleus	0.00020 59.0%	--	1.62e-08 61.2%	--	8.71e-10 60%	8.71e-10 60%	8.71e-10 60%	8.71e-10 60%
nucleolar part	0.00071 11.5%	--	2.43e-06 11.2%	--	1.89e-06 10%	1.89e-06 10%	1.89e-06 10%	1.89e-06 10%
macromoleular complex	0.00179 54.1%		9.23e-07 56.1%	--	2.04e-05 50.8%	2.04e-05 50.8%	2.04e-05 50.8%	2.04e-05 50.8%
smallsubunit processome	--		1.80e-05 9.2%	--	0.00011 7.5%	0.00011 7.5%	0.00011 7.5%	0.00011 7.5%
organelle part	--		0.00081 58.2%	--	0.00016 57.5%	0.00015 57.5%	0.00015 57.5%	0.00015 57.5%
intracellular organelle part	--		0.00081 58.2%	--	0.00016 57.5%	0.00015 57.5%	0.00015 57.5%	0.00015 57.5%

Table 6.6

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Second Seed and the Corresponding P-value Obtained for each Algorithm for the Process Ontology

GO TERMS	p-value									
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	VGRASP	CGRASP	RGRASP	PSO	GREEDY-PSO
translation	7.82e-25 60.7%	2.26e-23 54.7%	2.03e-49 63.3%	3.12e-40 73%	1.52e-56 64.5%	1.60e-44 58.3%	1.60e-44 58.3%	1.52e-56 64.5%	3.99e-62 58.8%	5.00e-25 77.8%
cellularprotein metabolic process	3.25e-12 64.3%	2.88e-11 59.4%	3.08e-24 66.3%	9.61e-24 17.7%	1.13e-27 67.3%	3.38e-23 63.1%	3.38e-23 63.1%	1.13e-27 67.3%	1.29e-27 61%	2.84e-15 83.3%
protein metabolic process	8.24e-12 64.3%	7.49e-11 59.4%	1.77e-23 66.3%	3.96e-23 77.8%	8.11e-27 67.3%	2.01e-22 63.1%	2.01e-22 63.1%	8.11e-27 67.3%	3.96e-23 61%	6.56e-15 83.3%
cellular macromolecule biosynthetic process	5.82e-10 62.5%	1.42e-08 56.2%	6.74e-19 63.3%	8.59e-18 73.0%	6.21e-22 64.5%	1.30e-16 59.2%	1.30e-16 59.2%	6.21e-22 64.5%	7.36e-23 59.6%	1.22e-11 77.8%
macromolecule biosynthetic process	6.47e-10 62.5%	1.58e-08 56.2%	8.19e-19 63.3%	1.00e-17 73%	7.76e-22 64.5%	1.57e-16 59.2%	1.57e-16 59.2%	7.76e-22 64.5%	9.52e-23 59.6%	1.34e-11 77.8%
gene expression	1.09e-08 62.5%	5.29e-08 57.8%	2.12e-17 64.3%	5.78e-17 74.6%	3.70e-20 64.5%	8.51e-16 61.2%	8.51e-16 61.2%	3.70e-20 64.5%	1.90e-20 60.3%	1.34e-11 80.6%
translational elongation	2.35e-08 16.1%	2.66e-09 15.6%	4.60e-17 16.3%	1.78e-09 15.9%	1.32e-14 14%	3.56e-13 13.6%	3.56e-13 13.6%	1.32e-14 14%	2.54e-16 12.5%	1.41e-08 22.2
cellular biosynthetic process	3.41e-07 64.3%	1.15e-07 62.5%	1.39e-15 67.3%	--	7.02e-18 68.2%	6.10e-14 64.1%	6.10e-14 64.1%	7.02e-18 68.2%	9.12e-19 64%	1.55e-08 77.8%
biosynthetic process	6.64e-07 64.3%	2.42e-07 62.5%	5.05e-15 67.3%	3.10e-14 76.2%	2.96e-17 68.2%	1.83e-13 64.1%	1.83e-13 64.1%	2.96e-17 68.2%	3.99e-18 64%	2.58e-08 77.8%
ribosome biogenesis	4.34e-05 26.8%	1.26e-06 28.1	6.01e-15 32.7%	1.25e-11 36.5%	1.49e-15 31.8%	2.41e-12 29.1%	2.41e-12 29.1%	1.49e-15 31.8%	2.85e-17 29.4%	3.01e-08 41.7%
rRNA processing	0.00010 21.4%	9.86e-06 21.9%	5.61e-10 22.4%	7.25e-08 25.4%	4.00e-09 20.6%	1.08e-07 19.4%	1.08e-07 19.4%	4.00e-09 20.6%	2.81e-10 19.1%	4.40e-06 30.6%
rRNA metabolic process	0.00017 21.4%	1.77e-05 21.9%	1.48e-09 22.4%	1.45e-07 25.4	1.04e-08 20.6%	2.53e-07 19.4%	2.53e-07 19.4%	1.04e-08 20.6%	8.79e-10 19.1%	7.11e-06 30.6%
cellular macromolecule metabolic process	0.00025 67.9%	0.00024 65.6%	9.83e-10 70.4%	2.90e-11 81.0%	3.82e-11 71.0%	1.85e-08 67%	1.85e-08 67%	3.82e-11 71.0%	2.03e-09 64.7%	9.46e-09 88.9%

From Table 6.5 it is clear that the order of algorithms based on p-value for the first GO term is greedy, GRASP, CGRASP, RGRASP, ISIMSRDT, MSRT, MSRDT, and SGSC for the first GO term. Similar p-values are obtained for greedy, GRASP, CGRASP and RGRASP for all GO terms. Since there are only 23 genes in the bicluster obtained by the SGSC algorithm there is only one GO term associated with it for the component ontology. The best value for the percentage of genes involved is obtained by ISIMSRDT algorithm for all GO terms. Out of the 14 GO terms, the best p-value is obtained by ISIMSRDT algorithm for all GO terms except in the case of GO terms 1, 9, 10, 13 and 14.

From Table 6.6, it is clear that the p-values are obtained in the order PSO, Greedy, RGRASP, CGRASP, GRASP, ISIMSRDT, SGSC, greedy-PSO, MSRT and MSRDT respectively for the first GO term. Out of the 13 GO terms, p-value obtained by binary PSO is better than all other algorithms, except for the GO terms 2, 3, 7 and 13. For GO terms 2 and 3, the greedy and GRASP variants obtained the best p-value. For the seventh GO term, the best p-value is obtained by ISIMSRDT and for the 13^{th} GO term the best p-value is obtained by SGSC. In terms of percentage of genes involved the greedy-PSO hybrid is better than all the other algorithms for all the GO terms.

Table 6.7

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Second Seed and the Corresponding p-value and Percentage of Genes obtained for each Algorithm for the Function Ontology

GO TERMS	p-value and percentage of genes									
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	GRASP	CGRASP	RGRASP	PSO	GREEDY-PSO
structural constituent of ribosome	9.79e-24 50%	2.58e-24 46.9%	6.05e-53 56.1%	4.42e-44 66.7%	5.81e-62 57.9%	5.81e-62 57.9%	5.81e-62 57.9%	5.81e-62 57.9%	7.00e-70 52.9%	1.73e-24 66.7%
structural molecule activity	2.73e-18 50%	1.79e-18 46.9%	3.97e-42 57.1%	3.48e-35 66.7%	4.33e-49 58.9%	4.33e-49 58.9%	4.33e-49 58.9%	4.33e-49 58.9%	3.27e-55 54.4%	8.97e-20 66.7%
translation elongation factor activity	0.00015 7.1%	0.00035 6.2%	7.16e-05 5.1%	--	0.00011 4.7%	0.00011 4.7%	0.00011 4.7%	0.00011 4.7%	0.00039 3.7%	0.00149 8.3%
RNA-directed DNA polymerase activity	--	--	--	--	-	-	-	-	2.21e-05 5.1%	-
RNA binding	-	-	0.00208 15.3%	--	0.00603 14%	0.00603 14%-	0.00603 14%-	0.00603 14%-	0.00023 14.7%	-
translation elongation factor activity	-	-	-	-	-	-	-	-	0.00039 3.7%	-
DNA-directed DNA polymerase activity	-	-	-	-	-	-	-	-	0.00211 5.1%	-
DNA polymerase activity	-	-	-	-	-	-	-	-	0.00287 5.1%	-

Table 6.8
Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Second Seed and the Corresponding p-value and Percentage of Genes obtained for each Algorithm for the Component Ontology

GO TERMS	p-value and percentage of genes									
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	GRASP	CGRASP	RGRASP	PSO	GREEDY-PSO
cytosolic ribosome	1.55e-26 51.8%	2.71e-27 48.4%	1.51e-60 58.2%	1.30e-46 66.7%	1.42e-70 59.8%	1.42e-70 59.8%	1.42e-70 59.8%	1.42e-70 59.8%	1.01e-79 54.4%	1.49e-25 66.7%
cytosolic part	2.95e-24 51.8%	7.66e-25 48.4%	3.92e-55 58.2%	5.45e-43 66.7%	3.93e-64 59.8%	3.93e-64 59.8%	3.93e-64 59.8%	3.93e-64 59.8%	1.37e-71 54.4%	1.09e-23 66.7%
ribosome	8.24e-24 57.1%	6.60e-24 53.1%	3.60e-51 62.2%	6.05e-41 71.4%	1.10e-58 63.6%	1.10e-58 63.6%	1.10e-58 63.6%	1.10e-58 63.6%	2.59e-63 57.4%	7.35e-25 75%
cytosol	1.36e-20 55.4%	3.91e-23 54.7%	1.42e-48 63.3%	1.45e-36 69.8%	4.32e-57 65.4%	4.32e-57 65.4%	4.32e-57 65.4%	4.32e-57 65.4%	1.09e-60 58.8%	4.68e-20 69.4%
ribonucleoprotein complex	1.11e-18 60.7%	3.21e-18 56.2%	3.31e-38 64.3%	1.04e-32 74.6%	1.36e-43 65.4%	1.36e-43 65.4%	1.36e-43 65.4%	1.36e-43 65.4%	3.01e-46 59.6%	4.34e-23 83.3%
cytosolic small ribosomal subunit			9.31e-28 27.6%	2.44e-19 30.2%	7.82e-32 28%	7.82e-32 28%	7.82e-32 28%	7.82e-32 28%	3.95e-37 25.7%	9.12e-08 25%
cytosolic large ribosomal subunit	2.09e-17 32.1	3.59e-16 28.1%	4.42e-27 28.6%	1.84e-24 36.5%	1.63e-32 29.9%	1.63e-32 29.9%	1.63e-32 29.9%	1.63e-32 29.9%	4.73e-37 27.2%	2.21e-16 41.7%
large ribosomal subunit	7.99e-15 32.1	1.30e-13 28.1%	1.45e-22 28.6%	6.11e-21 36.5%	5.06e-27 29.9%	5.06e-27 29.9%	5.06e-27 29.9%	5.06e-27 29.9%	3.54e-30 27.2%	2.86e-14 41.7%
non-membrane-bounded organelle	1.23e-10 62.5%	1.50e-10 59.4%	1.12e-21 65.3%	1.12e-20 76.2%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	2.09e-25 61%	1.13e-14 83.3%
intracellular non-membrane-bounded organelle	1.23e-10 62.5%	1.50e-10 59.4%	1.12e-21 65.3%	1.12e-20 76.2%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	7.56e-25 66.4%	2.09e-25 61%	1.13e-14 83.3%

From Table 6.7, it is clear that in terms of p-value the order of algorithms are PSO, Greedy, RGRASP, CGRASP, GRASP, ISIMSRDT, SGSC, greedy-PSO, MSRDT and MSRT respectively for the first and second GO terms. Similar p-values and percentage of genes are obtained for greedy, GRASP, CGRASP and RGRASP for all GO terms. The percentage of genes involved is the highest for greedy-PSO for the first three GO terms. More GO terms are obtained for PSO compared to that of all the other algorithms.

From Table 6.8, it is clear that the best p-values are obtained in the order PSO, Greedy, RGRASP, CGRASP, GRASP, ISIMSRDT, SGSC, MSRDT, MSRT and greedy-PSO for all the GO terms. The order of MSRDT and MSRT is changing for a few GO terms. The percentage of genes involved is the highest for greedy-PSO in all cases except for the GO term cytosolic small ribosomal subunit.

Table 6.9
Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Third Seed and the Corresponding p-value, and Percentage of Genes obtained for each Algorithm for the Process Ontology

GO TERMS	p-value and percentage of genes							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	GRASP	CGRASP	RGRASP
DNA repair	4.82e-13 57.1%	1.43e-14 60.7%	3.25e-10 45.5%	4.68e-12 51.6%	9.53e-11 44.4%	9.53e-11 44.4%	9.53e-11 44.4%	9.53e-11 44.4%
response to DNA damage stimulus	5.57e-12 57.1%	1.97e-13 60.7%	3.04e-09 45.5%	5.29e-11 51.6%	1.03e-09 44.4%	1.03e-09 44.4%	1.03e-09 44.4%	1.03e-09 44.4%
DNA metabolic process	4.37e-11 60.7%	7.23e-14 67.9%	1.11e-10(highest) 54.5%	2.50e-11 58.1%	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%	5.44e-11(high) 52.8%
cell cycle	8.19e-07 53.6%	-	1.13e-09 57.6%	4.87e-08 54.8%	8.42e-10 55.6%	8.42e-10 55.6%	8.42e-10 55.6%	8.42e-10
cell cycle process	5.15e-06 50%	-	7.20e-09 54.5%	2.99e-07 51.6%	4.80e-09 52.8%	4.80e-09 52.8%	4.80e-09 52.8%	4.80e-09 52.8%
double-strand break repair	2.91e-07 32.1%	-	1.53e-06 27.3%	8.98e-07 29%	1.86e-07 27.8%	1.86e-07 27.8%	1.86e-07 27.8%	1.86e-07 27.8%
cellular response to stress	5.99e-10 60.7%	6.03e-10 60.7%	2.60e-07 48.5%	8.51e-08 51.6%	1.44e-07 47.2%	1.44e-07 47.2%	1.44e-07 47.2%	1.44e-07 47.2%
response to stress	2.28e-08 60.7%	2.30e-08 60.7%	6.89e-06 48.5%	2.35e-06 51.6%	4.62e-06 47.2%	4.62e-06 47.2%	4.62e-06 47.2%	4.62e-06 47.2%
mitotic sister chromatid cohesion	-	3.76e-05 21.4%	8.67e-08 24.2%	2.31e-06 22.6%	7.19e-06 19.4%	7.19e-06 19.4%	7.19e-06 19.4%	7.19e-06 19.4%
cellular response to stimulus	2.29e-08 64.3%	2.31e-08 64.3%	8.61e-06 51.5%	2.58e-06 54.8%	6.59e-06 50%	6.59e-06 50%	6.59e-06 50%	6.59e-06 50%
cell cycle phase	1.66e-05 42.9%	1.67e-05 42.9%	1.36e-07 45.5%	7.11e-06 41.9%	6.59e-06 44.4%	6.59e-06 44.4%	6.59e-06 44.4%	6.59e-06 44.4%
M phase	0.00021 35.7%	1.85e-05 39.3%	1.14e-06 39.4%	6.26e-06 38.7%	3.89e-07 38.9%	3.89e-07 38.9%	3.89e-07 38.9%	3.89e-07 38.9%
chromosome organization	0.00018 39.3%	0.00158 35.7%	1.68e-06 42.4%	7.07e-05 38.7%	7.15e-06 38.9%	7.15e-06 38.9%	7.15e-06 38.9%	7.15e-06 38.9%

From Table 6.9, it is clear that for most of the GO terms the order of algorithms based on p-values are MSRDT, MSRT, SGSC, Greedy, GRASP, CGRASP, RGRASP and ISIMSRDT. The percentage of genes involved is the highest for MSRDT in most of the GO terms. Similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all the GO terms.

From Table 6.10, the order of algorithms based on p-value and percentage of genes involved is MSRT, MSRDT, SGSC, ISIMSRDT, Greedy, GRASP, CGRASP and RGRASP for the first GO term. Similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all the GO terms. Only two GO terms are obtained for all algorithms except MSRT, MSRDT and SGSC.

Table 6.10
Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Third Seed and the Corresponding p-value, and Percentage of Genes obtained for each Algorithm for the Function Ontology

GO TERMS	p-value and percentage of genes							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	VGRASP	CGRASP	RGRASP
double-stranded DNA binding	4.13e-05 17.9%	4.58e-05 17.9%	0.00202 12.1%	8.01e-05 16.1%	0.00341 11.1%	0.00341 11.1%	0.00341 11.1%	0.00341 11.1%
structure-specific DNA binding	0.00103 17.9%	0.00115 17.9%	0.00172 15.2%	0.00198 16.1%	0.00315 13.9%	0.00315 13.9%	0.00315 13.9%	0.00315 13.9%
DNA secondary structure binding	0.00104 10.7%	0.00116 10.7%	--	0.00162 9.7%	--		--	
guanine/thymine mispair binding	0.00335 7.1%	0.00372 7.1%	--	0.00469 6.5%	--		--	
single base insertion or deletion binding	0.00335 7.1%	0.00372 7.1%	--	0.00469 6.5%	--		--	
four-way junction DNA binding	0.00999 7.1%		--	--	--		--	

Table 6.11
Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Third Seed and the Corresponding p-value and Percentage of Genes obtained for each Algorithm for the Component Ontology

GO TERMS	p-value and percentage of genes							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	GRASP	CGRASP	RGRASP
chromosome	2.01e-08 50.0%	1.85e-08 50%	2.01e-09 (highest) 48.5%	8.04e- 09(highest) 48.4%	1.21e-07 44.7%	1.21e-07 44.7%	1.21e-07(highest) 44.7%	1.21e-07 44.7%
replication fork	1.38e-07 28.6%	3.42e-09 32.1%	6.19e-07 24.2%	9.43e-06 22.6%	1.40e-06 22.2%	1.40e-06 22.2%	1.40e-06 22.2%	1.40e-06 22.2%
chromosomal part	1.53e-06 42.9%	1.41e-06 42.9%	4.2e-07 42.4%	5.29e-07 41.9%	4.93e-06 36.1%	4.93e-06 36.1%	4.93e-06 36.1%	4.93e-06 36.1%
nuclear chromosome	6.59e-06 39.3%	6.07e-06 39.3%	4.50e-07 39.4%	2.18e-05 35.5%	1.53e-05 33.3%	1.53e-05 33.3%	1.53e-05 33.3%	1.53e-05 33.3%
nuclear replication fork	3.39e-05 21%	1.01e-06 25%	0.00010 18.2%	0.00146 16.1%	0.00019 16.7%	0.00019 16.7%	0.00019 16.7%	0.00019 16.7%
nuclear chromosome part |	0.00036 32.1%	0.00033 32.1	2.23e-05 33.3%	0.00089 29%	0.00050 27.8%	0.00050 27.8%	0.00050 27.8%	0.00050 27.8%
condensed nuclear chromosome	0.00594 17.9%	0.00546 17.9%	3.65e-06 24.2%	4.34e-05 22.6%	0.00014 19.4%	0.00014 19.4%	0.00014 19.4%	0.00014 19.4%
mitotic cohesin complex		--	7.89e-07 12.1%	5.95e-07 12.9%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%
nuclear mitotic cohesin complex			7.89e-07 12.1%	5.95e-07 12.9%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%	0.00042 8.3%
nucleus		8.87e-06 78.6%	3.39e-05 72.2%	3.19e-05 74.2%	1.52e-05 72.2%	1.52e-05 72.2%	1.52e-05 72.2%	1.52e-05 72.2%
condensed chromosome		.00925 17.9%	8.86e-06 24.2%	5.10e-06 25.8%	0.00030 19.4%	0.00030 19.4%	0.00030 19.4%	0.00030
nuclear cohesin complex			3.91e-06 12.1%	2.95e-06 12.9%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%
cohesin complex			3.91e-06 12.1%	2.95e-06 12.9%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%	0.00104 8.3%

From Table 6.11, the order of algorithms based on p-value is ISIMSRDT, SGSC, MSRDT, MSRT, Greedy, GRASP, CGRASP and RGRASP for the first and third GO terms. Similar p-values and percentage of genes are obtained for Greedy, GRASP, CGRASP and RGRASP for all the GO terms. In this case constraint based algorithms are better than Greedy and GRASP variants. For the second GO term, the order of algorithms based on p-value is MSRDT, ISIMSRDT, MSRT, Greedy, GRASP variants and SGSC. The percentage of genes involved is the best for the MSRDT algorithm for the first, second and third GO terms.

In the significant biclusters obtained from the fourth seed, since the conditions selected are different for each algorithm, the genes selected are also different. Hence the GO terms are different for biclusters obtained by each algorithm. Hence GO terms along with the p-values are given in the order of p-values. From Table 6.12 the order of algorithms based on p-value is RGRASP, Greedy, GRASP, CGRASP, MSRDT, SGSC, ISIMSRDT and MSRT for the first and third GO terms. For the second, fourth and fifth GO terms, the order of algorithms based on p-value is Greedy, RGRASP, GRASP, CGRASP, MSRDT, SGSC, ISIMSRDT and MSRT. The percentage of genes involved in a GO term, is the highest for MSRDT algorithm in the case of the first, second, third and fifth GO terms. For the fourth GO term, the percentage of genes involved is the highest for MSRT.

Table 6.12

Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Fourth Seed and the Corresponding p-value and Percentage of Genes obtained for each Algorithm for Process Ontology

p-value and percentage of genes							
MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**	**GREEDY**	**GRASP**	**CGRASP**	**RGRASP**
cytokinesis 0.00130 20.6%	cytokinesis 2.32e-05 28.6%	cytokinesis 7.07e-05 24.2%	cytokinesis 6.87e-05 24.2%	ribonucleoprotein complex biogenesis 1.55e-14 22.9%	RNA processing 1.63e-06 18.9%	RNA processing 1.92 e-06 18.8%	ribonucleoprotein complex biogenesis 4.56e-15 23.4%
positive regulation of spindle pole body separation 0.00195 8.8%	cell cycle process 3.91e-05 46.4%	cell division 0.00043 24.2%	cell cycle process 0.00024 39.4%	ribosome biogenesis 9.55e-13 20.2%	ribosome biogenesis 3.86e-06 15.8%	ribosome biogenesis 4.45e-06 15.7%	ribosome biogenesis 1.63e-12 20.3%
cell cycle process 0.00252 35.3%	cell cycle 6.36e-05 46.4%	cell cycle cytokinesis 0.00 130 18.2%	cell cycle 0.00039 39.4%	cellular component biogenesis at cellular level 3.72e-11 23.3%	ncRNA processing 6.13e-06 15.3%	ncRNA processing 7.03e-06 15.2%	cellular component biogenesis at cellular level 9.69e-12 23.9%
cell cycle 0.00383 35.3%	cell division 0.00014 28.6%	cell cycle process 0.00171 36.4%	cell division 0.00042 24.2%	RNA processing 8.10e-09 20.6%	ribonucleoprote in complex biogenesis 1.50e-05 16.7%	ribonucleoprote in complex biogenesis 1.74e-05 16.6%	RNA processing 3.04e-08 20.3%
regulation of spindle pole body separation 0.00387 8.8%	cell cycle cytokinesis 0.00058 21.8%	positive regulation of spindle pole body separation 0.00173 9.1%	cell cycle cytokinesis 0.00126 18.2%	ncRNA processing 2.31e-08 17.0%	ncRNA metabolic process 2.40e-05 15.8%	ncRNA metabolic process 2.76e-05 15.7%	ncRNA processing 8.36e-08 16.7%

Table 6.13

Comparison of MSR based Algorithms based on GO Terms for Biclusters generated from Fourth Seed and the Corresponding P-value obtained for each Algorithm for Function Ontology

GO Terms	p-value							
	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	GRASP	CGRASP	RGRASP
'molecular function unknown'	13 out of 34 input genes	11 out of 28 genes	12out of 33 genes	13 out 33 genes	Endonuclease activity (9, 0.00591)	84genes Out of 224	84 genes Out of 224	85 genes Out of 224

From the Table 6.13, it is clear that for function ontology a fixed number of genes are annotated to the term 'molecular function unknown' for all algorithms except for the Greedy algorithm. For Greedy algorithm, 9 out of the 224 genes are annotated to the term Endonuclease activity and the corresponding p-value is 0.00591.

Table 6.14
Comparison of MSR based Algorithms based on GO Terms for Biclusters Generated from Fourth Seed and the Corresponding p-value and Percentage of Genes obtained for each Algorithm for Component Ontology

p-value							
MSRT	**MSRDT**	**ISIMSRDT**	**SGSC**	**GREEDY**	**VGRASP**	**CGRASP**	**RGRASP**
cellular bud 3.48e-06 29.4%	cellular bud neck 39.3% 1.06e-09	cellular bud 39.4% 3.90e-10	cellular bud 39.4% 3.41e-10	cellular bud neck contractile ring 5.70e-05 6.4%	--	nucleolus 0.00032 13.1%	nucleolus 3.34e-12 16.2%
cellular bud neck 3.81e-06 26.5%	cellular bud 1.12e-09 42.9%	cellular bud neck 33.3% 6.47e-09	cellular bud neck 33.3% 6.47e-09	actomyosin contractile ring 5.70e-05 6.4%	--	cellular bud neck 0.00033 10.3%	nucleus 5.96e-08 49.5%
site of polarized growth 1.63e-05 29.4%	site of polarized growth 7.76e-09 42.9%	site of polarized growth 36.4% 5.50e-08	site of polarized growth 36.4% 4.96e-08	contractile ring 5.70e-05 6.4%	--	cellular bud 0.00066 11.7%	Preribosome 7.91e-08 10.8%
cellular bud neck contractile ring 5.04e-05 11.8%	cellular bud neck contractile ring 1.44e-07 17.9%	cellular bud neck contractile ring 3.23e-07 15.2%	3.23e-07 15.2% cellular bud neck contractile ring	cellular bud 0.00011 16.7%	--	site of polarized growth 0.00154 12.4%	Nuclear part 8.88e-06 31.1%
Actomyosin contractile ring 5.04e-05 11.8%	Actomyosin contractile ring 1.44e-07 17.9%	Actomyosin contractile ring 15.2% 3.23e-07	Actomyosin contractile ring 3.23e-07 15.2%	preribosome 0.00028 14.1%	--	cellular bud neck contractile ring .00210 3.4%	90s preribosome 3.83e-05 7.2%
contractile ring 5.04e-05 11.8%	contractile ring 1.44e-07 17.9%	contractile ring 15.2% 3.23e-07	contractile ring 3.23e-07 15.2%	cellular bud neck .00175 12.8%	--	actomyosin contractile ring .00210 3.4%	nucleolar part 5.46e-05 6.3%
cell division site 0.00069 11.8%	Cytoskeletal part 1.63e-06 35.7%	cytoskeleton part 30.3% 7.71e-06	cytoskeleton part 7.71e-06 30.3%	cell division site .00180 6.4%	--	contractile ring 0.00210 3.4%	Nuclear Lumen .00035 23.0%

From Table 6.14, it is clear that the order of algorithms based on p-value is RGRASP, Greedy, SGSC, ISIMSRDT, MSRDT, GRASP CGRASP and MSRT for the first GO term. The order of algorithms based on p-value is MSRDT, SGSC, ISIMSRDT, Greedy, RGRASP, MSRT, CGRASP and GRASP and for the second GO term. The order of algorithms based on p-value is MSRDT, SGSC, ISIMSRDT, RGRASP, Greedy, MSRT, CGRASP and GRASP for the third GO term. For the first GO term the percentage of genes involved is the best for SGSC and ISIMSRDT. In short, from these results we cannot conclude that a single algorithm is best in terms of p-value. The order is changing for each bicluster and in some situation for a particular ontology.

6.2.2 Comparison based on best 5 p-values obtained for the MSR based Algorithms

In Table 6.15 all the MSR based algorithms are compared on the basis of the best 5 p-values obtained from all the four biclusters. In this case, the order of the algorithms is PSO, RGRASP, Greedy, CGRASP, GRASP, ISIMSRDT, SGSC, MSRDT, MSRT and Greedy-PSO for the first GO term. PSO is the best in terms of p-value for all GO terms. The percentage of genes involved is the best for SGSC and Greedy-PSO for the first GO term. The p-value obtained by Greedy, GRASP, CGRASP and RGRASP are the same for all GO terms.

Table 6.15
Result of Biological Significance Test: The Top Five Functionally Enriched Significant GO Terms Produced by MSR based Algorithms

Terms	MSRT	MSRDT	ISIMSRDT	SGSC	GREEDY	CGRASP	GRASP	RGRASP	PSO	GREEDY-PSO
1	Cytosolic ribosome 51.8% 1.55e-26	Cytosolic ribosome 48.4% 2.71e-27	Cytosolic ribosome 58.2% 1.51e-60	Cytosolic ribosome 66.7% 1.30e-46	Cytosolic ribosome 59.8% 1.42e-70	Cytosolic ribosome 59.8% 1.42e-70	Cytosolic ribosome 59.8% 1.42e-70	Cytosolic ribosome 59.8% 1.42e-70	Cytosolic ribosome 54.4% 1.01e-79	Cytosolic Ribosome 66.7% 1.49e-25
2	translation 60.7% 7.82e-25	Cytosolic Part 48.4% 7.66e-25	Cytosolic Part 58.2% 3.92e-55	Structural constituent of ribosome 66.7% 4.42e-44	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 59.8% 3.93e-64	Cytosolic Part 54.4% 1.37e-71	translation 77.8% 5.00e-25
3	Cytosolic Part 51.8% 2.95e-24	Structural constituent of ribosome 46.9% 2.58e-24	Structural constituent of ribosome 56.1% 6.05e-53	Cytosolic Part 66.7% 5.45e-43	structural constituent of ribosome 57.9% 5.81e-62	structural constituent of ribosome 57.9% 5.81e-62	structural constituent of ribosome 57.9% 5.81e-62	structural constituent of ribosome 57.9% 5.81e-62	Structural constituent of ribosome 50% 7.00e-70	ribosome 75.0% 7.35e-25
4	ribosome 57.1% 8.24e -24	ribosome 53.1% 6.60e-24	ribosome 62.2% 3.60e-51	ribosome 71.4% 6.05e-41	ribosome 63.6% 1.10e-58	ribosome 63.6% 1.10e-58	ribosome 63.6% 1.10e-58	ribosome 63.6% 1.10e-58	ribosome 57.4% 2.59e-63	Structural constituent of ribosome 66.7% 1.73e-24
5	Structural constituent of ribosome 50% 9.79e -24	Structural Molecule Activity 46.9% 2.58e-24	Translation 62% 2.03e-49	translation 73% 3.12e-40	Translation 64.5% 1.52e-56	Translation 64.5% 1.52e-56	Translation 64.5% 1.52e-56	Translation 64.5% 1.52e-56	Translation 58.8% 3.99e-62	Cytosolic Part 66.7% 1.09e-23

6.2.3 Comparison of Algorithms based on Bicluster Size and MSR

Three different seeds are selected. These seeds are enlarged by all the MSR based algorithms. The bicluster size and MSR are compared for biclusters obtained from all these algorithms.

Table 6.16.a

Comparison of Size and MSR of 3 Biclusters by Enlarging Three Different seeds by each one of the MSR based Algorithms

Biclusters	MSRT			MSRDT			ISIMSRDT		
	Size	MSR	Row variance	size	MSR	Row variance	size	MSR	Row variance
1	61*17	198.95	469.4058	77*16	199.54	533.1660	98*17	199.97	482.8
2	56*17	199.78	587.8461	64*17	199.32	654.5732	98*17	199.99	600.9
3	28*17	299.85	1937.5000	28*17	286.34	2034.1000	33*17	299.22	1970.1

Table 6.16.b

Comparison of Size and MSR of 3 Biclusters by Enlarging 3 Different seeds by each one of the MSR based Algorithms

SGSC			Greedy		
Size	MSR	Row Variance	Size	MSR	Row Variance
23*17	131.39	506.7582	121*17	199.94	483.2784
63*17	167.43	615.9798	107*17	199.48	568.0833
31*17	297.19	2036.0000	36*17	297.61	1806.9000

Table 6.16.c

Comparison of Size and MSR of Three Biclusters by Enlarging Three Different seeds by each one of the MSR based Algorithms

Biclusters	GRASP			RGRASP			CGRASP		
	Size	MSR	Row variance	Size	MSR	Row variance	Size	MSR	Row variance
1	121*17	199.94	483.278	121*17	199.94	483.278	121*17	199.94	483.278
2	107*17	199.48	568.083	107*17	199.48	568.083	107*17	199.48	568.083
3	36*17	297.61	1806.900	36*17	297.61	1806.900	36*17	297.61	1806.900

Analysing the algorithms based on the biclusters obtained from the same seed, it is noted that among the algorithms SGSC produced biclusters of low size but coherence is high since the MSR value is very low. ISIMSRDT is the best among the four constraint based algorithms in terms of bicluster size. Reducing the increment factor in ISIMSRDT can improve the bicluster size further. But Greedy and GRASP algorithms identify better biclusters than the four constraint based algorithms in terms of bicluster size and MSR. CGRASP, RGRASP and GRASP algorithms result in different biclusters only if the conditions selected are different. These algorithms differ in the order in which the genes are added. The local search phase also results in the addition of similar genes.

From Tables 6.16 (a), 6.16 (b) and 6.16 (c), it is clear that for the first seed, same bicluster is obtained by Greedy, RGRASP, GRASP and CGRASP. This is the largest in terms of bicluster size. The second highest is in the order ISIMSRDT, MSRDT, MSRT and SGSC. For the second

seed the order of algorithms in terms of bicluster size is Greedy, RGRASP, CGRASP, GRASP, ISIMSRDT, MSRDT, SGSC, and MSRT. For the third seed the order is Greedy, RGRASP, CGRASP, GRASP, ISIMSRDT, MSRDT, SGSC, and MSRT. In short Greedy, RGASP, CGRASP, GRASP, ISIMSRDT, MSRDT, MSRT and SGSC are the order of algorithms in terms of bicluster size. The order of MSRT and SGSC changes for different biclusters depending on the value selected for the difference threshold.

6.3 Summary

Mean Squared Residue (MSR) is used as a measure of coherence in many of the biclustering algorithms developed so far. In this chapter a problem with the MSR in the identification of biclusters with large row variance is presented. The problem is that most often the large incremental increase in MSR may be due to the lack of coherence. But sometimes it may be due to the significant change in the expression level of the genes indicated by the high value of the row variance. When the row variance increases significantly, the MSR value also increases. But sometimes this increase in MSR will be above the predefined MSR threshold. The visual inspection of the bicluster plot can help towards differentiating between lack of coherence and significant change. Genes with such highly significant change in the expression level are of great biological significance. In this context SGSC and MSRT algorithms are better than all other algorithms mentioned in this study, because in these algorithms maximum possible variation is allowed for MSR. In all other

algorithms which are trying to minimize MSR, it is difficult to identify biclusters from Yeast and Lymphoma datasets with high row variance and coherence (even though the MSR value exceeds the predefined threshold) as shown in Figure 6.1 and 6.2.

In terms of bicluster size (high) and MSR (low) Greedy and GRASP variants are better than the constraint based algorithms. In terms of p-value Greedy and GRASP variants are better than the constraint based algorithms for the biclusters from first two seeds. But in the case of the constraint based algorithms, the p-value is better than Greedy and GRASP variants for the bicluster generated from the third seed. For the fourth seed it is difficult to make a final conclusion. In terms of time complexity the constraint based algorithms are better than Greedy and metaheuristic approaches.

Biclustering problem is NP-Hard [29, 36]. Heuristic based search methods are used to solve the biclustering problem in polynomial time [27, 91]. Similar problem solving methods are used in this study for the identification of biclusters.

Chapter 7

Conclusion and Future Work

Different types of algorithms namely constraint based, greedy and metaheuristic algorithms are developed in this work for the identification of coherent biclusters from high dimensional gene expression data. Some of the constraint based algorithms are able to identify biclusters with significant change in the expression level of the genes. The row variance of some of such biclusters is higher than that of any other algorithm using MSR. In terms of the best p-value obtained these algorithms are better than some of the well known biclustering algorithms namely MOGAB, SGAB, CC, RWB, Bimax, OPSM, ISA and BiVisu. The algorithms developed in this work overcome some of the disadvantages associated with the already existing biclustering algorithms. The results obtained and the performance analysis, show that these algorithms are suitable for the identification of coherent biclusters. Suggestions for the further work in this area of research are also given.

7.1 Conclusion

In recent years large amounts of high-dimensional data in gene expression profiles are generated. Analyzing such high-dimensional gene expression data have become an issue of significant research interest. Elucidating the patterns hidden in high-dimensional gene expression data is a highly relevant and challenging research endeavour.

Biclustering identifies local patterns from high dimensional data. Biclustering is simultaneous clustering of both the rows and columns of a data matrix. In this thesis algorithms are developed for the identification of coherent biclusters from gene expression data using different algorithm design techniques. All these algorithms are using a measure called mean squared residue to search for biclusters. Biclustering is an optimization problem with the objective of maximizing the volume and minimizing the mean squared residue of the bicluster. All these algorithms are enlarging the seeds obtained from K-Means clustering algorithm.

Different types of algorithms, namely constraint based, greedy and metaheuristic algorithms are developed in this work for the identification of coherent biclusters from high dimensional gene expression data. There are four constraint based algorithms, one greedy approach, four metaheuristic algorithms and the last one is a combination of greedy and metahueristic approach. The different algorithms are:

1) Mean Squared Residue Threshold (MSRT) algorithm
2) Mean Squared Residue Difference Threshold (MSRDT) algorithm

3) Iterative Search with Incremental MSR Difference Threshold (ISIMSRDT) algorithm

4) Seed Growing using Separate Constraints (SGSC) algorithm

5) Algorithm based on Greedy approach

6) Algorithm based on Greedy Randomized Adaptive Search Procedure (GRASP)

7) Algorithm based on Cardinality based Greedy Randomized Adaptive Search Procedure (CGRASP)

8) Algorithm based on Reactive Greedy Randomized Adaptive Search Procedure (RGRASP)

9) Algorithm based on Binary Particle Swarm Optimization (PSO)

10) Algorithm based on Greedy - Binary Particle Swarm Optimization hybrid

In all the ***constraint based algorithms*** node (gene or condition) addition follows node deletion if necessary. The added node is deleted depending on the constraints used by the algorithm. **The MSRT algorithm** uses the only constraint namely the MSR threshold. This method allows maximum variation possible for MSR. It is advantageous for including conditions which make significant change in the bicluster. But the disadvantage is that the added node may not be optimal in terms of MSR value. Hence one more constraint called the **MSR Difference Threshold (MSRDT)** is introduced with the objective of minimizing MSR. This constraint resulted in different research findings in connection

with the gene expression data. It is found that this threshold value is different for genes and conditions. Reducing the difference threshold value for genes increases coherence and reducing the threshold value for conditions eliminates conditions which make significant change in the expression level. It is also found that the difference threshold for the negatively correlated genes is higher than that of other genes. It is difficult to find a suitable value of difference threshold for each bicluster. Hence in **ISIMSRDT algorithm,** the MSR difference threshold is initialized with a small value and it is incremented after each iteration. Iterative search has the advantage of including the n^{th} condition whose MSR value got reduced after adding the $(n-k)^{th}$ condition. After experimenting with MSRDT algorithm it is found that reducing the MSR difference threshold for genes increases coherence and reducing the MSR difference threshold for conditions eliminates conditions which make significant change in the expression level. Thus it is concluded that separate constraints should be used for genes and conditions. Hence the algorithm **Seed Growing using Separate Constraints (SGSC)** for genes and conditions is developed. Highly coherent biclusters can be identified with this algorithm. Moreover, with the help of bicluster plot this algorithm can identify some biclusters with very high row variance from both Yeast and Lymphoma datasets. These biclusters are coherent even though there MSR value is above the predefined MSR threshold.

As optimization problem the main objective of biclustering is to identify highly coherent biclusters. With this objective in mind a ***Greedy approach*** is used to enlarge the seeds obtained by K-Means clustering

algorithm. The greedy approach used by Cheng and Church has random interference problem. When seeds from K-Means are used this problem can be eliminated. Moreover MSR is biased towards the flat biclusters. Seeds from K-Means help the identification of biclusters with high row variance.

Greedy approach usually suffers from local minima problem. ***Metahueristic methods*** like Greedy Randomized Adaptive Search procedure incorporate randomization for eliminating the local minima problem. **Three variants of GRASP namely (basic) GRASP, Cardinality based GRASP (CGRASP) and Reactive GRASP (RGRASP)** are used for the identification of biclusters. The GRASP variants implemented in this approach is able to find biclusters with more size and low MSR. Moreover, in this study GRASP variants are applied for the first time to Lymphoma dataset. Another metahueristic method which can eliminate local minima problem namely the **Particle Swarm Optimization (PSO)** is used for the identification of coherent biclusters. This is the only technique which is population based, whereas all other methods are enlarging a single seed at a time. One more approach which is a **combination of Greedy and Binary PSO** is used for the identification of biclusters in which the biclusters obtained by the greedy approach is used as initial population for PSO.

These algorithms identified biclusters from both Yeast and Human Lymphoma datasets. These algorithms are compared with other biclustering algorithms based on bicluster size and MSR. Biologically

relevant and statistically significant biclusters are identified by all these algorithms. The algorithms are also compared based on p-value which denotes the statistical significance. All the algorithms developed in this work are better than some of the well known biclustering algorithms namely RWB, Bimax, OPSM and Bivisu, in terms of the best p-value obtained. The best p-value obtained by binary-PSO, Greedy, GRASP variants, SGSC and ISIMSRDT, which are developed in this work, are even better than that of MOGAB, SGAB, CC and ISA.

In short the following limitations of already developed biclustering algorithms can be overcome by using one or more of the algorithms, which are developed in this work.

1) The maximum limit for the number of conditions that can be identified for a bicluster. For example the multi-objective evolutionary approach the maximum number of conditions obtained for the Yeast dataset is only 11 and Human Lymphoma dataset is only 40 [15].
2) The maximum limit for the number of genes that can be identified for a bicluster. For example in SEBI [36] the maximum number of genes obtained for the Yeast dataset is only 82.
3) The difficulty in identifying biclusters with different shapes.
4) Random interference problem in the Greedy approach of Cheng and Church.

5) Difficulty in finding genes with overlap. For example in the Greedy approach of Cheng and Church, for identifying different biclusters, the identified biclusters are replaced with random values. This affects the identification of genes with overlap.

6) Inability to identify biclusters with very low row variance. The genes in such biclusters are useful for marker gene identification. Since row variance is not given as a measure for optimization, biclusters with low row variance as well as high row variance will be obtained in the methods developed in this work. In this work biclusters obtained from unfiltered data will contain biclusters of low row variance compared to that of filtered data.

7) Inability to identify biclusters with very high row variance and mean squared residue above the predefined threshold.

Finally the MSR based algorithms are compared based on the quality of biclusters. In terms of best p-value and bicluster size, the binary PSO, Greedy and GRASP variants are better than constraint based algorithms. But for some biclusters, the p-value obtained by constraint based algorithms is better than Greedy and Metaheuristic algorithms. Some biclusters with very high row variance are identified from both Yeast and Lymphoma datasets with the help of constraint based algorithms SGSC and MSRT. In terms of time complexity, the constraint based algorithms are better than Greedy and Metaheuristic algorithms.

Biclustering is a multi-objective optimization problem and some of the objectives of biclustering like low MSR and high row variance are conflicting. Hence no single algorithm can be considered as the best in terms of different parameters. Based on the experiments and the analysis of the results of all the algorithms in this study, the following recommendations can be made. The recommendations are presented in the Table 7.1 given below.

Table 7.1

Recommendations for the Selection of an Algorithm Based on different Bicluster Qualities

Bicluster Quality	Recommendations
Bicluster Size	Greedy and Metaheuristic approaches are better than Constraint based algorithms.
Conflicting nature of MSR and ROW variance	MSRT, SGSC algorithms are better than Greedy and Metaheuristic algorithms
p-value	Greedy and Metaheuristic approaches obtained best p-value, but for biclusters of some category the Constraint based algorithms are better than Greedy and Metaheuristic approaches.
Percentage of genes involved	SGSC, Greedy-Binary PSO hybrid
Biclusters with different shape	Seed growing approaches are better than population based techniques like PSO.
Time Complexity	Constraint based algorithms are better than Greedy and Metaheuristic approaches.

7.2 Suggestions for Future Work

1) There are many metaheuristic approaches available. But only some of these methods are applied for the identification of biclusters from gene expression data so far. The remaining methods can also be used for the identification of biclusters.

2) Mean squared residue is used as a measure of coherence in many of the biclustering algorithms developed so far. There is a problem with the mean squared residue in the identification of biclusters with large row variance. The problem is that most often the large incremental increase in MSR may be due to the lack of coherence. But sometimes it may be due to the significant changes in the expression levels indicated by the high value of the row variance. When the row variance increases significantly the MSR value also increases. But sometimes this increase in MSR will be above the predefined MSR threshold. At present only visual inspection of the bicluster plot can help towards differentiating between lack of coherence and significant change. Further research can be directed towards developing new measures and methods to solve this problem.

…..ൟ…..

References

[1]. Achuthsankar S. Nair, "Computational Biology & Bioinformatics: A Gentle Overview", Communications of the Computer Society of India, January 2007, pp.1-12.

[2]. Akutsu, T., Miyano, S. and Kuhara, S. "Inferring qualitative relations in genetic networks and metabolic pathways", Bioinformatics 16: 2000, pp. 727-734.

[3]. Alan W.-C. Liew, Hong Yan, Mengsu Yang and Y.-P. Phoebe Chen, "Microarray Data Analysis", *Bioinformatics Technologies,* Edited by Phoebe Chen, Springer, 2005.

[4]. Alberts B, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, "*Molecular Biology of the Cell"*, 3rd Ed., Garland Publishing, New York, 1994.

[5]. Alizadeh, A. et al. "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling", Nature, 403, 2000, pp. 503-511.

[6]. Alter O, P.O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling", *Proc. Natl. Acad. Sci. USA*, 97:10101.10106, 2000.

[7]. Alter O. et al. "Singular value decomposition for genome wide expression data processing and modeling", *Proceedings of the National Academy of Science, USA* 97(18): (2000), pp. 10101–10106.

[8]. Anderson A, T. Olofsson, D. et al. "Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations", *Proceedings National Academy of Science USA*, 102(52), pp. 19069-19074, 2005.

[9]. Angiulli. E. Cesario and C. Pizzuti, "Gene expression biclustering using random walk strategies", 7th International Conference Data Warehousing Knowledge Discovery (DAWAK 2005), Copenhagen, Denmark.

[10]. Aquilar-Ruiz, J. S., "Shifting and Scaling Patterns from Gene Expression Data", Bioinformatics. Vol. 21, 2005, pp. 3840-3845.

[11]. Arkin A, J.Ross, and H.H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage lambda-Infected Escherichia coli Cells", Genetics, 1998, 149 (4): pp. 1633-1648.

[12]. Baiyi Xie, Shihong Chen, Feng Liu, "Biclustering of Gene Expression data using PSO-GA hybrid", *Proc. of the First International Conference on Bioinformatics and Biomedical Engineering*, 2007, pp.302-305.

[13]. Balaji Krishnapuram, Lawrence Carin, and Alexander J. Hartemink, "Joint Classifier and Feature Optimization for Cancer Diagnosis using Gene Expression Data", *Recomb*, 2003.

[14]. Baldi P and G.W. Hatfield, DNA *Microarrays and Gene Expression: Experiments to Data Analysis and Modelling*. Cambridge, University Press, 2002.

[15]. Banka H. and Mitra S., "Multi-objective Evolutionary Biclustering of Gene Expression data", *Journal of Pattern Recognition*, Vol.39 2006, pp. 2464-2477.

[16]. Ben-Dor A, B.Chor, R. Karp, and Z.Yakhini, "Discovering Local Structure in Gene Expression Data: The Order Preserving Sub-matrix Problem," *Proc. 6th Annual International Conference Computational Biology*, 2002, vol. 1-58113-498-3, pp. 49-57.

[17]. Benjamin Good, Jeremy Peay, Satish Pillai, and Jacques Corbeil, "Class prediction based on gene expression: Applying neural networks via a genetic algorithm wrapper", *2001 Genetic and Evolutionary Computation Conference Late Breaking Papers*, July 2001, pp. 122–130.

[18]. Bing Liu, Wynne Hsu, and Yiming Ma. "Integrating Classification and Association Rule Mining", *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining*, 1998.

[19]. Bittner M, P. Meltzer, et al, "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling", *Nature*, 406:536.540, 2000.

[20]. Bleuler S, A. Prelic, and E.Zitzler, "An EA framework for biclustering of gene expression data", *Proceedings of Congress on Evolutionary Computation*, 2004, pp.166-173.

[21]. Bolsover SR, Hyams JS, Jones S, Shepard EA, White HA, *From Genes to Cells*. New York: Wiley, 1997.

[22]. Brazma, A. and Vilo, J. Minireview: Gene expression data analysis. European Molecular Biology Laboratory, Outstation Hinxton – the European Bioinformatics institute, Cambridge CB10 ISD UK, 2000.

[23]. Brown M.P.S, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr,and D. Haussler. Knowledge-based Analysis of Microarray Gene Expression data by using Support Vector Machines. *Proc. Natl. Acad. Sci. USA*, 2000.

[24]. Bryan, K., Cunningham, P. and Bolshakova, N, "Bottom-UP biclustering of Expression Data", *Proc. of Computational Intelligence in Bioinformatics and Computational Biology*, *2006.*

[25]. Bryan K., Cunningham P. and Bolshakova N, "Application of Simulated Annealing to the Biclustering of Gene Expression Data", *IEEE Transactions on Information Technology in Biomedicine*, 2006, Vol.10, No. 3, pp 519-525.

[26]. Chad Creighton and Samir Hanash. "Mining gene expression databases for association rules", *Bioinformatics*, 19, 2003.

[27]. Chakraborty A. "Biclustering of gene expression data by simulated annealing", *Proceedings of the 8th International Conference on High-Performance Computing in Asia-Pacific Region* (HPCASIA 05), 2005, pp.627-632.

[28]. Chakraborty A. and Hitashyam Maka "Biclustering of Gene Expression Data Using Genetic Algorithm" *Proceedings of Computation Intelligence in Bioinformatics and Computational Biology*, CIBCB, 2005, pp. 1-8.

[29]. Cheng Y. and Church G.M. "Biclustering of Expression Data", *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, (ISMB 2000) La Jolla, CA, 20-23 August, 2000, pp. 93-103.

[30]. Cios K.J, W. Pedrycz, and R. Swiniarski, *Data Mining Methods for Knowledge Discovery*, Dordrecht: Kluwer, 1998.

[31]. Cormen, Leiserson, and Rivest, *Introduction to Algorithms*, second edition, McGraw-Hill book Company, Cambridge, 2002, p.329.

[32]. De Jong H, "Modelling and Simulation of Genetic Regulatory Systems: A Literature Review", *Journal of computational Biology*, 2002, 9(1): 67-103.

[33]. DeRisi, J.L., Lyer, V.R. and Brown, P.O. "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science* 278: 1997, pp. 680-686.

[34]. Dharan S and Nair AS, "Biclustering of Gene expression Data using Greedy Randomized Adaptive Search Procedure", *Proceedings of IEEE TENCON*, 2008, pp. 1-5.

[35]. Diggle P, Liang KY, Zeger SL, *Analysis of Longitudinal Data*, Oxford: Oxford University Press, 1994.

[36]. Divina, F. and Aguilar-Ruize, J.S, "Biclustering of Expression Data with Evolutionary Computation", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, 2006, pp. 590- 602.

[37]. Divina, F. and Aguilar-Ruiz J.S. "A Multi-Objective Approach to Discover Biclusters in Microarray Data", *Proceedings of the ACM Int. Conference GECCO'07*, 2007, pp. 385-392.

[38]. Doddi S, A. Marathe, S.S. Ravi, and D.C. Torney. "Discovery of association rules in medical data", *Med. Inform. Internet. Med.*, 2001, 26:25–33.

[39]. Eisen M.B, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci. USA*, 1998.

[40]. Eytan Domany, **"**Cluster Analysis of Gene Expression Data**"** *Journal of Statistical Physics*, Vol. 110, Nos. 3–6, March 2003.

[41]. Fayyad U.M, G. Piatetsky-Shapiro, P. Smyth, and R.Uthurusamy,eds., *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA:AAAI/MIT Press, 1996.

[42]. Feo TA and Resende MGC "Greedy Randomized Adaptive Search Procedures", *Journal of Global Optimization* Vol. 6, 1995. pp. 109-133.

[43]. Furey T.S, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler "Support vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data. *Bioinformatics*, 16, 2000, pp. 906-914.

[44]. Garrett RH, Grisham CM, "Principles of Biochemistry" Pacific Grove, CA: Brooks/Cole, 2002.

[45]. Getz G, E. Levine and E. Domany, "Coupled two way clustering analysis of gene microarray data" *Proc. National Academy of Science*, vol. 94, 2000, pp. 12079-12084.

[46]. Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, Scott L.Zeger, *The Analysis of Gene Expression Data: An Overview of Methods and Software*, Springer, 2003.

[47]. Glass, L. "Combinatorial and Topological Methods in Nonlinear Chemical Kinetics", J. Chem. Phys. 63 (4), 1975, pp. 1325-1335.

[48]. Glass, L. and Pasternack, J.S., "Stable oscillations in mathematical models of biological control systems", J. Math. Biol. 6: 1978, pp. 207-223.

[49]. T.R. Golub, D. K. Slonim, P. et al, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, 1999.

[50]. Granucci F, Vizzardelli C, et al, "Inducible IL-2 Production by Dendritic Cells Revealed by Global Gene Expression Analysis", *Nature Immunology*, 2001, 2:882–888.

[51]. Gupta G.K, *Introduction to Data Mining with case studies*, PHI, New Delhi, 2006.

[52]. Hartigan, J.A, "Direct Clustering of a Data Matrix", *J. Am. Stat. Assoc*, 67, 1972, pp. 123-129.

[53]. Hartigan JA, Wong MA, "A K-means Clustering Algorithm", *Applied Statistics*, 28: 1979, pp. 100–108.

[54]. Hedenfalk I, D. Duggan, Y. Chen, et al, "Gene Expression profiles in Hereditary Breast Cancer", *The New England Journal of Medicine*, February, 2001.

[55]. Huang, S., "Gene Expression profiling, Genetic networks, and Cellular states: An Integrating Concept for Tumori-genesis and Drug discovery*", J. Mol. Med*., 1999, vol. 77, pp. 469-480.

[56]. Ihmels, J., Bergmann, S. and Barkai, N., "Defining Transcription Modules Using Large-Scale Gene Expression Data", *Bioinformatics*, 20, 2004, pp. 1993-2003.

[57]. Ivan Gesteira Costa Filho, *Mixture Models for the Analysis of Gene Expression: Integration of Multiple Experiments and Cluster Validation*, Ph.D Thesis, Freie University, Berlin, 2008.

[58]. Jens Nilsson, *Nonlinear dimensionality reduction of gene expression data*, Ph.D Thesis, Center for mathematical Sciences, Lund University, Sweden, 2006.

[59]. Jiawei Han, Jian Pei, and Yiwen Yin, "Mining frequent patterns without candidate generation", *Proc. 2000 ACM-SIGMOD International Conferenece Management of Data*, 2000.

[60]. Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong, "Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients", *Bioinformatics*, 2003, vol. 19, pp. 71–78.

[61]. John L. Pfaltz and Christopher M. Taylor. "Closed set mining of biological data", *Workshop on Data Mining in Bioinformatics*, 2002, pp. 43–48.

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

[63]. Kauffman, S.A., "Metabolic stability and epigenesist in randomly connected nets". J. Theor. Biol, 1969, vol. 22, pp. 437-467.

[64]. Kauffman, S.A., *The Origin of Order: Self-organization and Selection in Evolution*, Oxford University Press, New York, 1993.

[65]. Kauffman S. A., "Homeostasis and differentiation in random genetic control networks", *Nature*, 224:177-178, 1969.

[66]. Kennedy J, R. Eberhart, "Particle Swarm Optimization," *Proc. Of IEEE international Conference on Neural Networks* (ICW), Australia, 4, 1995, pp. 1942-1948.

[67]. Kennedy J. and Eberhart R.C., "A Discrete Binary Version of the Particle Swarm Optimization", *Proc. of the conference on Systems, Man, and Cybernetics* SMC97, 1997, pp.4104-4109.

[68]. Khan J, J.S. Wei, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", *Nature Medicine*, 2001.

[69]. J. Khan, R. Simon, et al., "Gene Expression Profiling of Alveolar Rhabdomyosarcoma with cDNA Microarrays", *Cancer Research*, November 1998.

[70]. Y. Kluger, R. Barsi, JT. Cheng, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions", *Genome Res.*, 13 (4), 2003, pp. 703–16.

[71]. Kruskal J B, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika*, vol. 29, pp. 1-27.

[72]. L. Lazzeroni and A. Owen, "Plaid models for gene expression data", *Statistica Sinica*, 2002, vol. 12, pp. 61–86.

[73]. Liu J, Wang W. "OP-cluster: Clustering by Tendency in High Dimensional Space", *Proceedings of the third IEEE International Conference on Data Mining*, 2003, pp. 187–94.

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

[75]. Maulik, U., Mukhopadhyay, A.and Bandyopadhyay, S., "Finding multiple coherent biclusters in Microarray data using variable string length multiobjective Genetic Algorithm", IEEE Transactions on information technology in Biomedicine, 2009, Vol.13, NO.6, pp. 969-975.

[76]. Mauricio G.C Resende and Celso C, Rebeiro, "Greedy Randomized Adaptive Search Procedures", *Handbook of Metaheuristics*, Edited by Fred Glover and Gary A Kochenberger, Kluwer Academic Publishers, New York, 2003. p. 221.

[77]. Pan F.Wang B., Hu X., and PerrizoW, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis", *J Biomed Inform*, 2004, Aug, 37 (4), 240-8.

[78]. A. Prelic, S. Bleuler, P. Zimmermann, et al., "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data", Bioinformatics, 2006, vol.22, no.9, pp.1122-1129.

[79]. Quackenbush J, "Computational analysis of microarray data", *Nature Reviews Genetics*, 2001, vol. 2, pp.418–427.

[80]. J.R. Quinlan, "Programs for machine learning", *Morgan Kaufmann, San Mateo, CA*, 1993.

[81]. Rakesh Agarwal and Ramakrishnan Srikanth, "Fast algorithms for mining association rules", Proc. Int. Conf. Very Large Data Bases (VLDB'94) pages 487-499, Sept. 1994.

[82]. Raychaudhuri S, Stuart JM, Altman RB, "Principal components analysis to summarize microarray experiments: Application to sporulation time series", RB Altman, AK Dunker, L Hunter, K Lauderdale, TE Klein (eds.), *Fifth Pacific Symposium on Biocomputing*, 2000, pp. 455–466.

[83]. Roberto J. Bayardo, Rakesh Agrawal, and Dimitrios Gunopulos, "Constraint based Rule mining in large, dense databases", *Proc. 15th International Conference on Data Engineering*, 1999.

[84]. S. L. Salzberg, D.B. Searls, and S. Kasif, eds, *Computational Methods in Molecular Biology*, Amsterdam: Elsevier Sciences B.V., 1998.

[85]. SGD GO Termfinder *[http://db.yeastgenome.org/cgi-bin/GO/ go TermFinder*.

[86]. Shi Y. and R. Eberhart, "A Modified Particle Swarm Optimizer", *Proc. IEEE Int. Conf. on Evolutionary Computation,* 1999, pp. 69-73.

[87]. Shmulevich, I., Dougherty, E.R. and Zhang, W, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks", *Proc. IEEE*, 2002, 90 (11), pp. 1778-1792.

[88]. Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W, "Probablistic Boolean networks: a rule-based uncertainty model for gene regulatory networks", *Bioinformatics*, 2002, 18, pp. 261-274.

[89]. Simone Mocellin and Carlo Riccardo Rossi "Principles of Gene Microarray Data Analysis", *Microarray Technology and Cancer Gene Profiling*, 2007.

[90]. 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.

[91]. 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.

[92]. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, BrownPO, Botstein D, Futcher B, "Comprehensive identification of cell cycleregulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization", *Molecular Biology of the Cell*, 1998, 9:3273–3297.

[93]. Stanislav Busygin, Oleg Prokopyev, Panos M. Pardalos, "Biclustering in datamining", *Computers and Operations Research*, 2008, 35, pp. 2964 – 2987.

[94]. Sungroh Yoon, Christine Nardini, Luca Benini, and Giovanni De Micheli "Discovering Coherent Biclusters from Gene Expression Data Using Zero-Suppressed Binary Decision Diagrams", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.2. NO.4, October-December 2005.

[95]. Sushmita Mitra and Tinku Acharya, "Data Mining : Multimedia, Soft Computing and Bioinformatics",

[96]. Szallasi, Z. and Liang, S, "Modeling the normal and neoplastic cell cycle with realistic Boolean genetic networks: their application for understanding carcinogenesis and assessing therapeutic strategies", *Pacific Symposium on Biocomputing*, 3: 66-76.

[97]. Tanay, A. et al., "Discovering Statistically Significant Biclusters in Gene Expression Data", *Bioinformatics*, 2002, 19 (Suppl 2), pp.196-205.

[98]. Amos Tanay, Roded Sharan, Ron Shamir, "Biclustering algorithms : A survey", *Handbook of Computational Molecular Biology*, Edited by Aluru S Chapman, Hall/CRC Computer and Information Science Series, 2005.

[99]. Tavazoie, S. et al., "Systematic Determination of Genetic Network Architecture", *Proc. Natl. Acad.Sci. USA*, 1999, vol. 22, pp. 281-285.

[100]. L. Teng and L-W. Chan, "Biclustering gene expression profiles by alternately sorting with weighted correlated coefficient," *Proc. IEEE Int. Workshop Mach. Learning Signal Process.*, 2006, pp. 289-294.

[101]. Thieffry, D. and Thomas, R., "Qualitative analysis of gene networks", *Pacific Symposium on Biocomputing*, 1998, vol. 3, pp. 77-88.

[102]. Thomas, R., "Regulatory networks seen as asynchronous automata: a logical description", J. Theor. Biol., 1991, vol. 153, pp. 1-23.

[103]. Wenmin Li, Jiawei Han, and Jian Pei. "CMAR: Accurate and efficient classification based on multiple class-association rules", *Proc.of 2001 IEEE Int. Conf. on Data Mining*, pp. 369–376, 2001.http://citeseer.nj.nec.com/li01cmar.html.

[104]. Wuensche, A., "Classifying cellular automata automatically: Finding gliders,filtering, and relating space-time patterns, attractor basins, and the Z parameter", *Complexity*, 1999, **4** (3), pp. 47-66.

[105]. Xin Xu "Data Mining Techniques in Gene Expression Data Analysis", Ph.D thesis, School of Computing, National University, Singapore, July 2006.

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

[107]. *Yeast Saccharomyces cerevisiae cell cycle expression dataset and Human Lymphoma Dataset* [http://arep.med. harvard.edu/ biclustering.]

[108]. Yeung KY, Haynor DR, Ruzzo WL, "Validating clustering for gene expression data", *Bioinformatics,* 2001, vol. 4, pp. 309–318.

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

[110]. Zhao LP, Prentice R, Breeden L, "Statistical modeling of large microarray data sets to identify stimulus-response profiles", *Proceedings of the National Academy of Science, USA,* 2001, *vol.* 98, pp. 5631–5636.

List of Publications

Edited Book Chapters

1. Shyama Das and Sumam Mary Idicula, "K-Means Greedy Search Hybrid Algorithm for Biclustering Gene Expression Data", ***Advances in Computational Biology,*** Edited by Hamid R Arabnia, Springer, 2010, pp. 180-188.
2. Shyama Das and Sumam Mary Idicula, "Comparative Advantages of Novel Algorithms using MSR Threshold and MSR difference Threshold for Biclustering Gene Expression Data", ***Software Tools and Algorithms for Biological Systems***, Edited by Hamid R Arabnia & Quoc Nam Tran, Springer, 2011, pp.123-134.

International Journal Papers

1. Shyama Das and Sumam Mary Idicula, "Application of Cardinality Based GRASP to the Biclustering of Gene Expression Data", ***International Journal of Computer Applications,*** Vol.1, No.18, 2010, pp. 44-51.
2. Shyama Das and Sumam Mary Idicula, "Iterative Search with Incremental MSR difference threshold", ***International Journal of Computer Applications***, Vol.1, No.18, 2010, pp. 35-43.
3. Shyama Das and Sumam Mary Idicula, "Greedy Search-Binary PSO Hybrid for Biclustering Gene Expression Data", ***International Journal of Computer Applications***, Vol.2, No.3, 2010, pp. 1-5.
4. Shyama Das and Sumam Mary Idicula, "Application of Greedy Randomized Adaptive Search Procedure to the Biclustering of Gene Expression Data", ***International Journal of Computer Applications***, Vol.2, No.3, 2010, pp. 6-13.

International Conference Papers

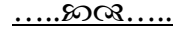
1. Shyama Das and Sumam Mary Idicula, "A novel approach in Greedy Search Algorithm for Biclustering Gene Expression Data", ***Proc. International Conference on Computational and Systems Biology (ICBCSB),*** Singapore, August 26-28, 2009, pp. 144-147.

2. Shyama Das and Sumam Mary Idicula, "Modified Greedy Search Algorithm for Biclustering Gene Expression Data", ***17th International Conference on Advanced Computing & Communication (ADCOM 2009),*** ACS, 14-17 December, 2009, IISc Banglore, India, pp. 83-88.

3. Shyama Das and Sumam Mary Idicula, "Biclustering Gene Expression Data using MSR difference Threshold", ***IEEE INDICON 2009,*** 18-20 December, 2009, Ahmedabad, India, pp. 1-4, ISBN: 0-7803-8909-3 DOI : 10.1109/INDICON.2009.5409489.

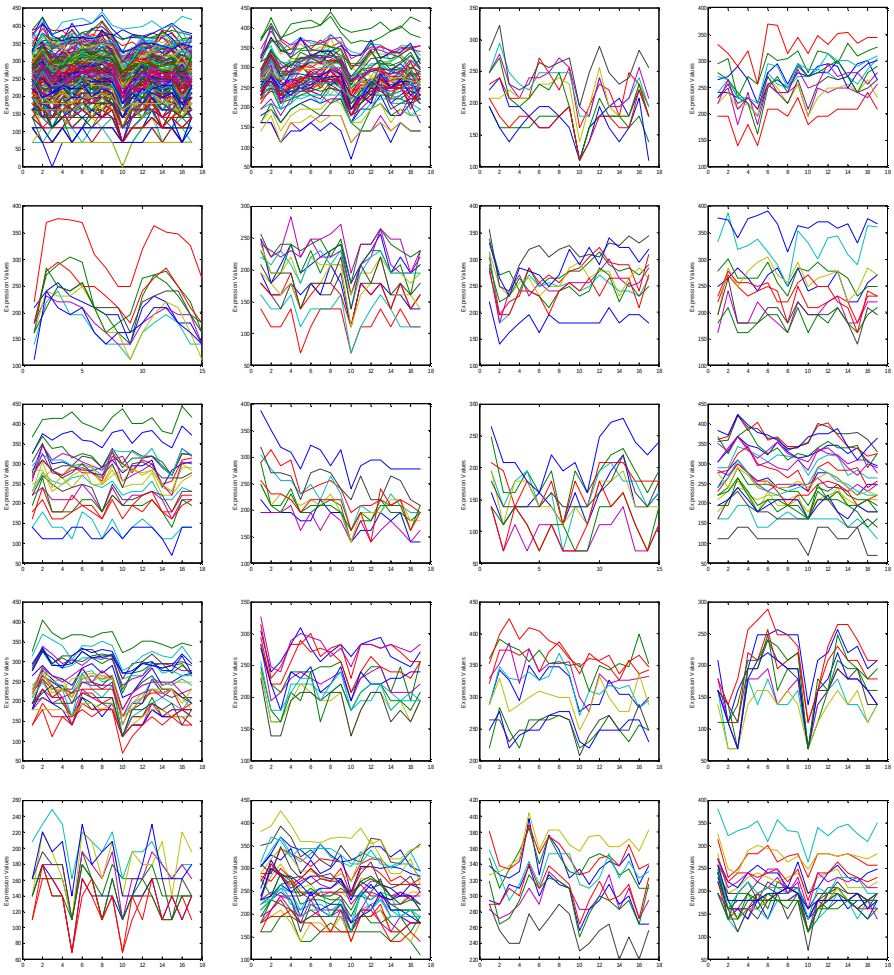
4. Shyama Das and Sumam Mary Idicula, "K-Means Binary PSO Hybrid Algorithm for Biclustering Gene Expression Data", ***ACM International Symposium on Biocomputing,*** 15-17 Feb 2010, jointly organized by NIT, Calicut, India, Indiana University Purdue University Indianapolise, USA in cooperation with ACM, (available in ACM digital Library).

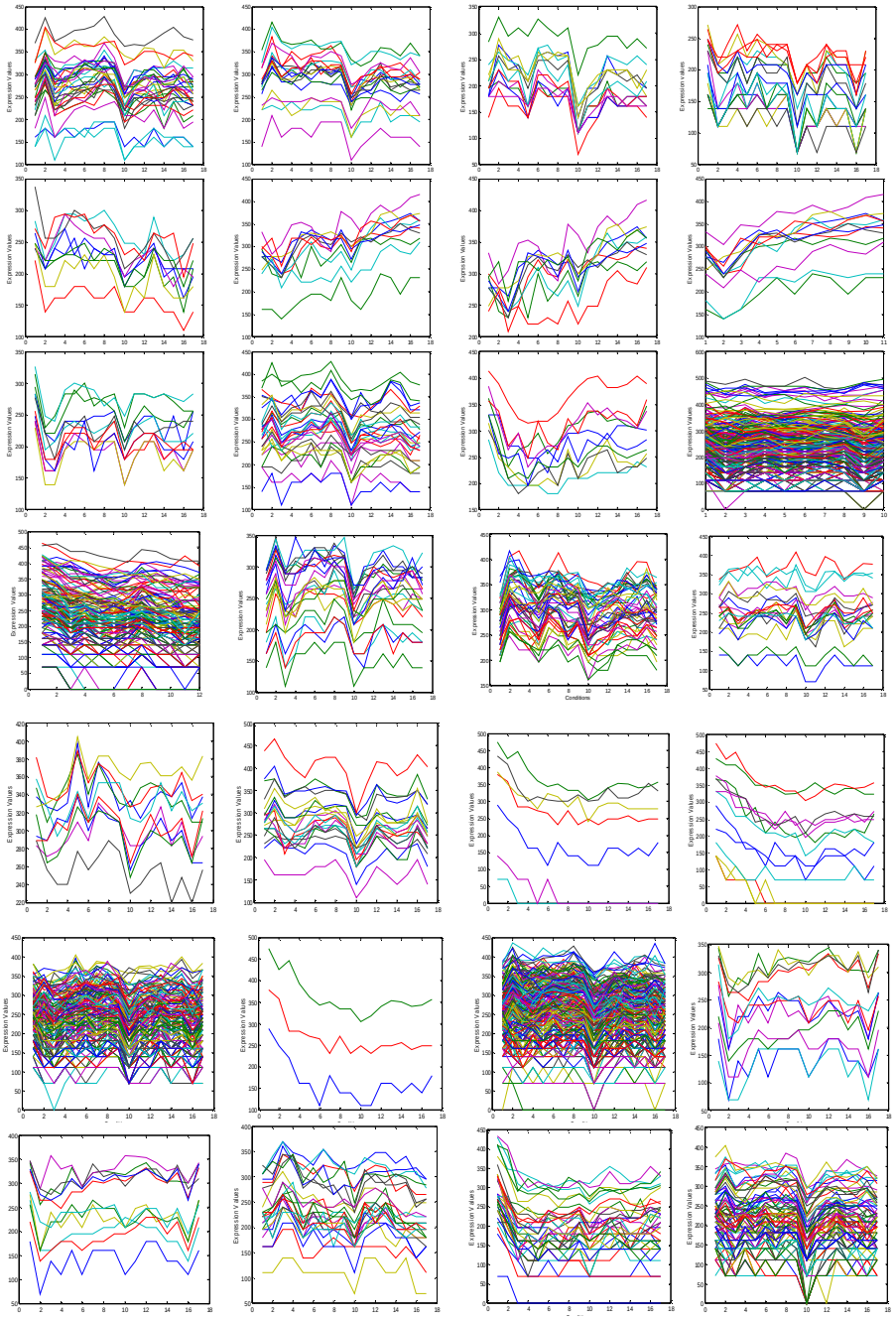
5. Shyama Das and Sumam Mary Idicula, "Application of Reactive GRASP to the Biclustering of Gene Expression Data", ***ACM International Symposium on Biocomputing,*** 15-17 Feb 2010, Jointly organized by NIT, Calicut, India, Indiana University Purdue University Indianapolise, USA in cooperation with ACM (available in ACM digital Library).

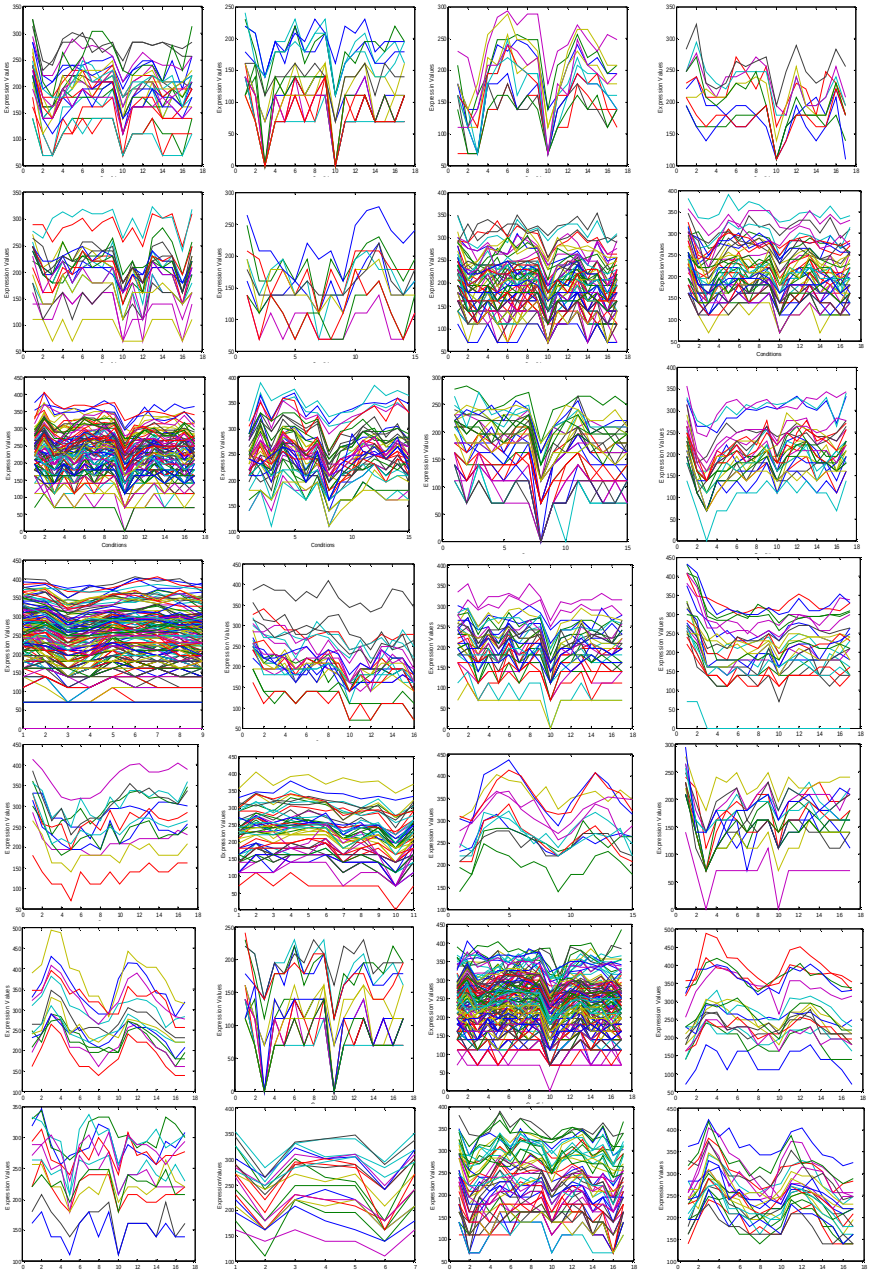


Appendix

Appendix I: Some more Biclusters obtained from the MSR based Algorithms for Yeast Dataset. The Bicluster Labels are from Ap1 to Ap76, from left to right and top to bottom. The details of the Biclusters can be obtained from the Table in Appendix 2 using Bicluster Label.




Expression Values
Conditions

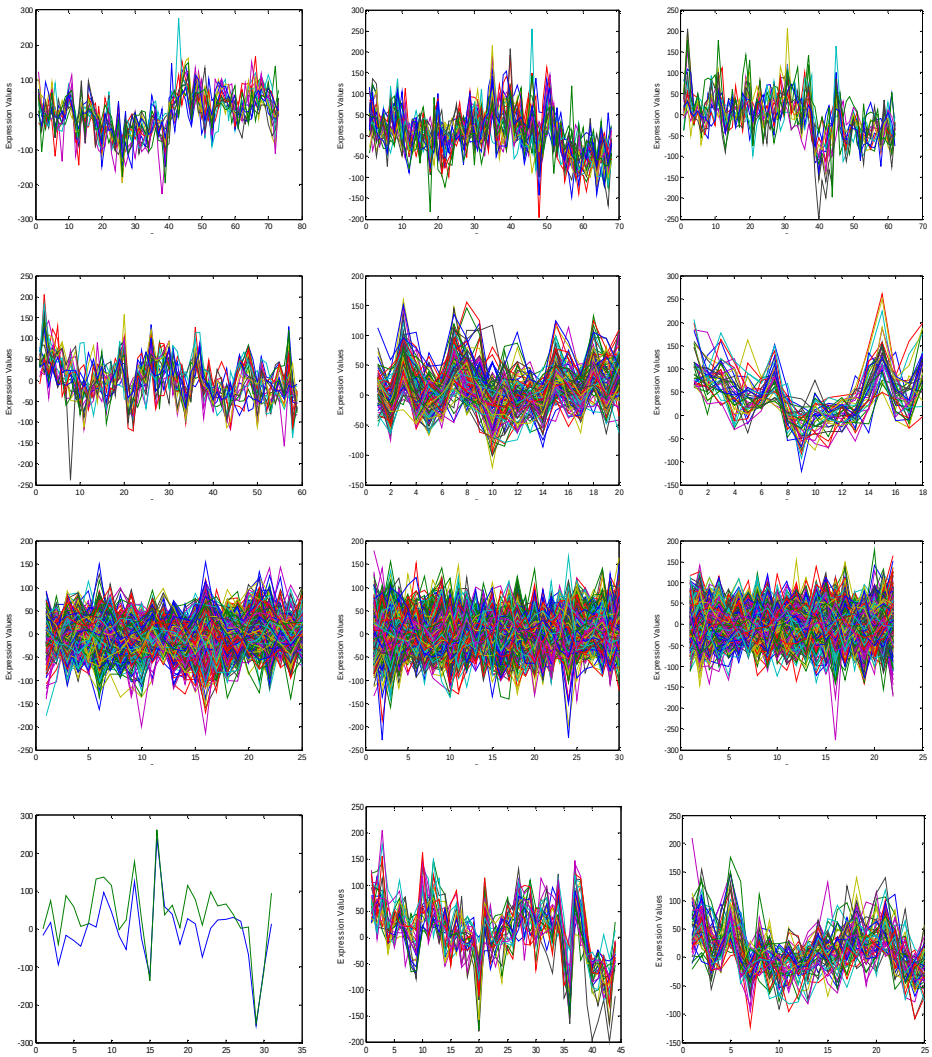

Expression Values
Conditions

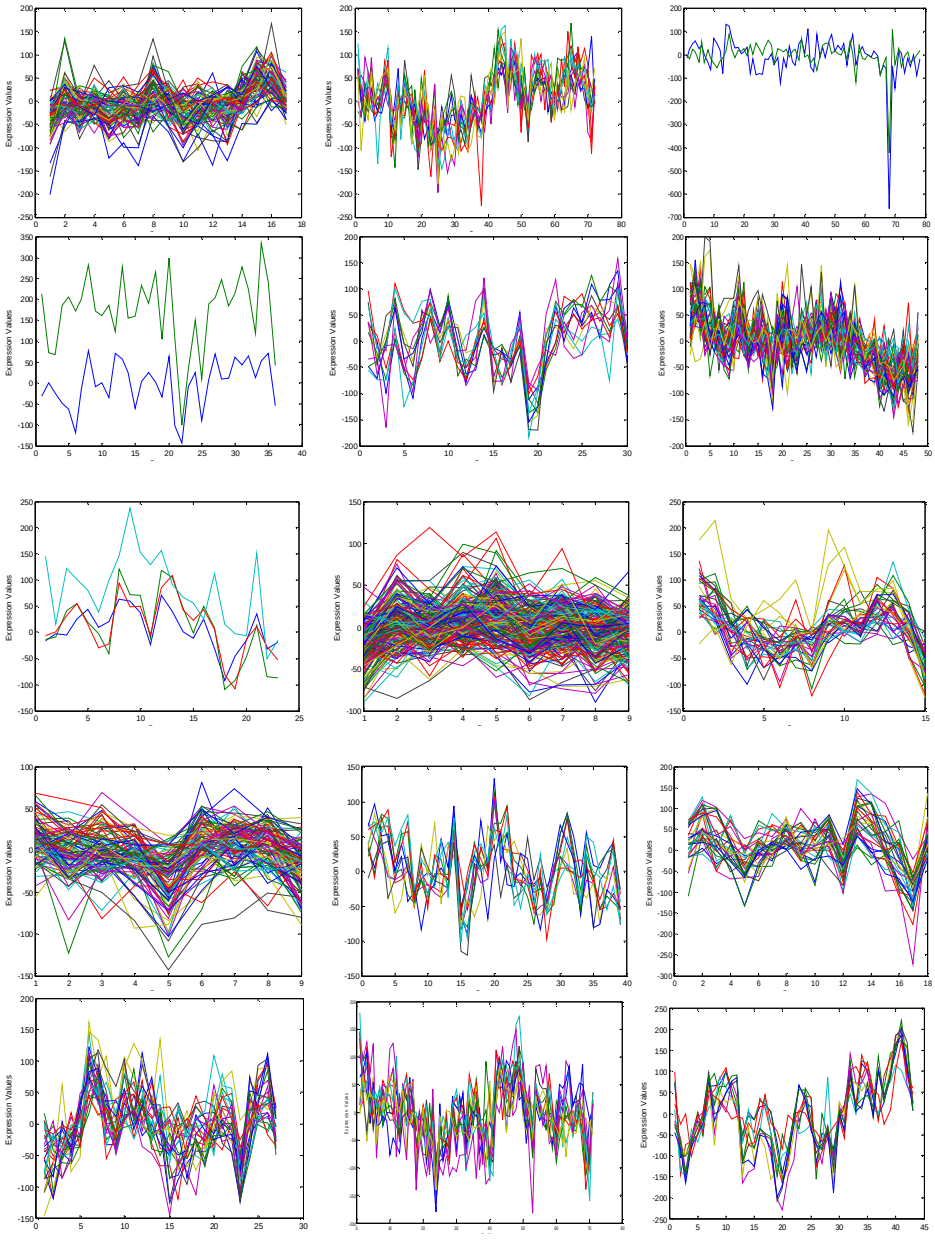
Appendix II: Details of the Biclusters of the MSR based Algorithms shown in Appendix 1

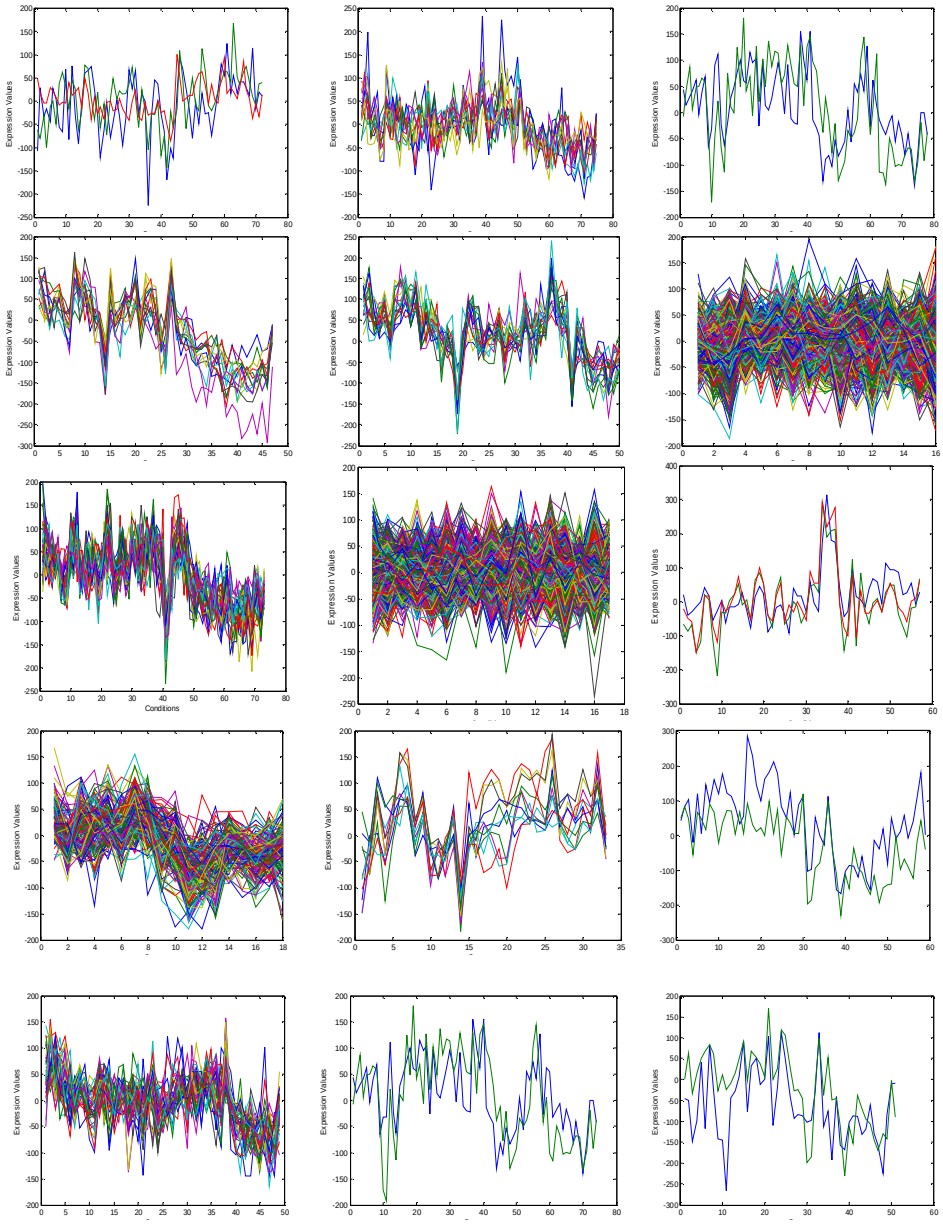
Bicluster Label	Number of Genes	Number of Conditions	MSR	Row Variance
Ap1	229	17	289.0000	412.2021
Ap2	68	17	199.2974	496.7250
Ap3	10	17	255.2242	792.5869
Ap4	12	17	186.9630	663.1355
Ap5	10	15	294.6495	1538.9000
Ap6	14	17	130.1876	501.2056
Ap7	10	17	261.8143	695.2803
Ap8	10	17	258.5586	568.9377
Ap9	21	17	198.4656	465.4810
Ap10	10	17	215.9523	833.7370
Ap11	10	15	299.4252	943.3236
Ap12	23	17	198.3296	556.8663
Ap13	29	17	168.4380	701.6955
Ap14	12	17	113.1148	505.9383
Ap15	10	17	237.8181	467.6422
Ap16	10	17	266.2193	1763.5000
Ap17	10	17	212.1967	703.9460
Ap18	32	17	212.6103	492.3886
Ap19	10	17	148.6853	409.3723
Ap20	23	17	202.9663	607.4846
Ap21	33	17	138.7529	479.2220
Ap22	19	17	107.2792	482.5351
Ap23	13	17	133.5914	921.8440
Ap24	18	17	165.9538	756.5682
Ap25	10	17	185.6726	705.5066
Ap26	11	17	265.9692	1074.0000
Ap27	10	17	271.2748	1091.5000
Ap28	12	11	187.3307	1160.1000
Ap29	11	17	110.2305	519.7502
Ap30	29	17	147.6002	433.5421
Ap31	10	17	213.4028	961.9329
Ap32	358	10	299.9277	458.4218
Ap33	158	12	299.8289	736.6941
Ap34	23	17	133.9864	449.2968
Ap35	49	17	184.8044	487.7615
Ap36	20	17	187.6572	429.4796

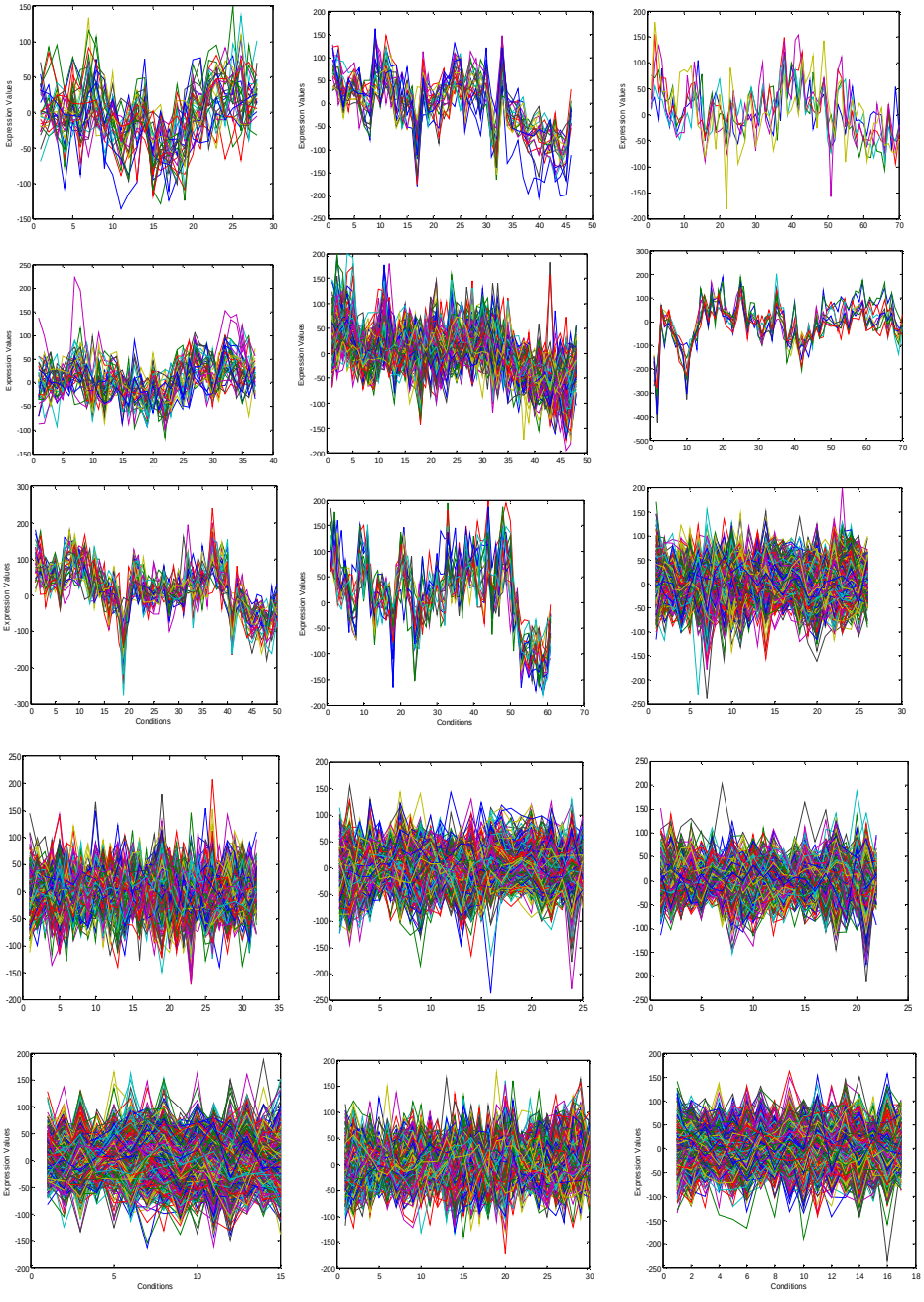
Ap37	10	17	409.3723	148.6853
Ap38	21	17	151.8050	603.8359
Ap39	7	17	297.8726	1538.7000
Ap40	13	17	299.2638	1835.8000
Ap41	172	17	299.9067	507.5788
AP42	3	17	184.2414	1952.5000
AP43	229	17	299.5817	402.5612
AP44	12	17	195.3356	725.5283
AP45	11	17	198.4083	642.3693
AP46	25	17	205.2872	532.9982
AP47	38	17	274.4020	1043.1000
AP48	93	17	629.6392	244.5466
Ap49	26	17	221.4451	781.5637
AP50	18	17	293.4744	1215.2000
AP51	13	17	267.3184	1664.2000
AP52	10	17	255.2242	792.5869
AP53	21	17	199.7015	687.8534
AP54	10	15	299.4252	943.3236
AP55	80	17	243.7829	538.3563
AP56	108	17	217.3164	521.2705
AP57	44	15	177.3988	547.1125
AP58	35	15	234.9981	855.1088
AP59	27	17	266.6634	975.0855
AP60	181	9	144.6135	240.0470
AP61	26	16	241.6395	765.3514
AP62	36	17	194.1223	592.8078
AP63	25	17	251.0314	1086.1000
AP64	12	17	207.8144	926.1125
AP65	54	11	186.1359	532.1206
AP66	12	15	288.9307	1094.4000
AP67	16	17	226.4725	939.5225
AP68	13	17	222.4583	1369.9000
AP69	17	17	294.0819	1253.9000
AP70	149	17	255.2447	479.9828
AP71	17	17	255.9226	970.2308
AP72	12	17	185.0830	528.7670
AP73	19	7	199.1124	882.3008
AP74	49	17	270.3296	606.0990
AP75	23	17	214.2337	1045.1000
AP76	12	17	207.8144	926.1125

Appendix III: More Biclusters obtained by MSR based algorithms from Human Lymphoma Dataset. The Bicluster Labels are from APL1 to APL57, from left to right and top to bottom. The details of the Biclusters can be obtained from the Table in Appendix 4 using Bicluster Label.




Expression Values


Expression Values
Conditions

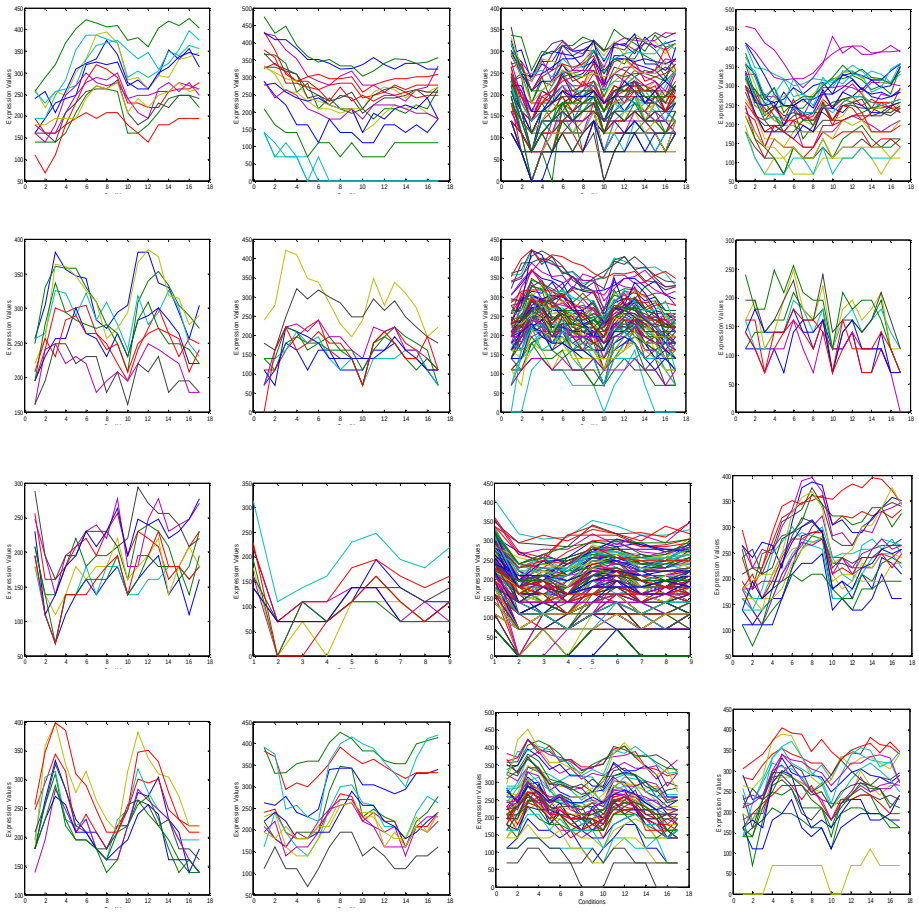

Expression Values
Conditions

Appendix IV: Details of the Biclusters of the MSR based Algorithms shown in Appendix 3.

Biclusters Label	Number of Genes	Number of Conditions	MSR	Row Variance
APL1	16	73	1198.1	3378.7
APL2	30	68	1199.6	2699.2
APL3	16	62	1195.7	2863.3
APL4	22	59	1198.5	2314.2
APL5	105	20	774.5	1248.6
APL5	37	18	2714.4	953.0
APL6	368	25	1198.7	1272.9
APL7	537	30	1199.5	1239.4
APL8	663	22	1199.3	1230.9
APL9	2	31	364.9	7713.3
APL10	26	44	1032.7	3195.6
APL11	54	25	894.2	1621.5
APL12	73	17	738.5	1248.5
APL13	13	72	1099.8	3486.8
APL14	2	78	1091.5	6163.2
APL15	13	73	1098.9	3497.0
APL16	2	36	1164.1	5508.8
APL17	12	30	1118.6	3572.9
APL18	48	48	946.8	2168.7
APL19	4	23	742.6	3293.1
APL20	215	9	375. 5	476.9
APL21	34	15	737.1	1944.2
APL22	96	9	415.6	639.8
APL23	11	39	738.1	1862.4
APL24	30	18	978.8	2674.0
APL25	26	27	1003.5	2283.3
APL26	13	71	1159.7	2843.6

APL27	10	43	1172.9	6865.4
APL28	3	72	1149.8	2853.8
APL29	13	75	1070.6	2311.6
APL30	2	78	1090.3	5797.3
APL31	14	47	1096.4	6839.7
APL32	18	50	1071.3	4628.7
APL33	896	16	1199.1	1239.4
APL34	33	73	1197.4	3822.9
APL35	997	17	1198.6	1213.4
APL36	3	57	962.5	7029.5
APL37	140	18	1198.3	2301.2
APL38	11	33	1138.2	3650.6
APL39	2	59	1155.0	9515.6
APL40	40	49	973.1	2285.4
APL41	2	67	1073.5	5975.9
APL42	2	51	1193.0	7497.1
APL43	31	28	984.2	1721.2
APL44	22	46	1057.5	4062.9
APL45	6	70	999.8	2733.4
APL46	36	37	956.1	1540.3
APL47	126	48	1200.0	2317.5
APL48	10	70	1190.5	5308.5
APL49	35	50	1195.8	4365.1
APL50	18	61	1195.2	4842.6
APL51	582	26	1197.5	1250.6
APL52	285	32	1199.3	1241.9
APL53	468	25	1198.9	1237.6
APL54	358	22	1199.3	1240.6
APL55	936	15	1198.9	1221.7
APL56	440	30	1200.0	1252.0
APL57	997	17	1198.6	1213.4

Appendix V: Some more Biclusters with High Row Variance & MSR above the Pre-defined Threshold obtained from Yeast Dataset. The Bicluster Labels are from APL1 to APL57, from left to right and top to bottom. The details of the Biclusters can be obtained from the Table in Appendix 4 using Bicluster Label.



Appendix VI: Details about the Biclusters with High Row Variance & MSR above Pre-defined Threshold shown above in Appendix 5

Bicluster Label	Number of Genes	Number of Conditions	MSR	Row variance
APR1	13	17	340.7195	2150.9
APR2	18	17	350.0203	1732.5
APR3	72	17	359.5201	969.2
APR4	34	17	318.1225	1002.0
APR5	10	17	383.3512	1178.8
APR6	10	17	680.2878	2086.2
APR7	89	17	414.1173	979.9
APR8	10	17	440.5057	1303.2
APR9	10	17	322.1075	1133.8
APR10	10	9	515.0200	2323.8
APR11	178	9	306.4028	916.7
APR12	22	17	426.8087	2117.8
APR13	10	17	365.2835	2375.0
APR14	13	17	494.7380	1590.2
APR15	56	17	300.5922	1002.0
APR16	21	17	463.2928	1366.4

Appendix VII: Sample Yeast Dataset (Only 31 rows out of 2884 rows are displayed here).

161 110 139 139 161 139 110 161 161 110 161 195 220 139 139 139 161
208 139 69 110 139 110 139 161 161 110 139 139 179 139 161 139 110
425 429 451 423 465 395 472 448 416 507 466 464 432 463 494 458 484
289 248 220 161 161 110 179 139 139 110 110 161 161 139 161 139 179
366 364 340 256 283 208 240 208 195 208 179 208 208 195 220 208 264
271 300 347 300 304 294 337 294 277 309 300 326 314 309 322 277 283
179 69 139 110 161 110 161 161 161 110 139 179 179 139 139 110 161
240 179 139 161 179 139 195 208 195 139 161 179 195 161 179 179 179
179 161 139 139 208 161 208 208 179 179 161 179 179 195 179 161 179
337 326 322 330 397 326 376 358 322 333 314 333 343 337 353 314 337
195 220 179 161 264 179 248 264 230 220 220 220 230 220 208 179 230
294 289 264 230 264 248 283 283 264 256 240 256 277 240 264 248 256
271 300 300 462 300 300 340 309 289 294 289 300 304 304 314 277 277
264 110 110 139 208 161 179 179 161 161 208 220 230 179 161 139 195
139 69 69 69 0 0 69 69 69 0 69 0 69 69 0 69 0
264 248 264 230 264 230 271 264 240 208 230 240 240 195 208 179 220
179 110 139 110 110 69 110 110 69 69 69 110 110 110 69 69 110
277 264 277 264 283 248 300 277 283 248 271 289 294 256 277 248 248
264 256 277 240 230 248 340 343 343 304 304 294 283 294 330 330 340
304 333 369 347 340 322 353 330 314 318 330 347 347 326 340 304 294
277 353 353 309 340 330 381 347 314 294 294 309 322 326 322 277 322
161 161 161 378 179 179 220 179 195 179 161 179 208 179 179 161 161
240 161 179 179 195 161 220 208 195 139 179 195 195 179 195 179 195
161 0 0 69 110 69 110 69 69 0 69 69 69 69 69 0 69
350 353 314 318 361 330 376 369 350 330 356 361 350 343 353 350 350
358 300 294 289 322 283 347 314 309 347 326 333 322 326 350 322 309
220 139 161 179 195 161 195 179 179 179 179 179 208 179 195 195 179
271 283 300 283 289 256 314 304 283 240 248 271 271 264 294 248 289
110 139 195 161 161 139 161 110 69 110 139 179 179 110 110 69 69
0 110 110 69 110 69 110 69 69 69 69 110 110 69 69 0 69
264 326 347 300 314 314 350 326 314 318 300 309 322 304 326 277 304