Article can be accessed online at http://www.publishingindia.com

Clustering Techniques for Biological Sequence Analysis: A Review

Jyoti Lakhani*, Anupama Chowdhary, Dharmesh Harwani***

Abstract

In the present scenario there are a variety of technical tools for supporting and validating wet-lab experiments in the field of science and biotechnology. In order to analyze biological sequences it is necessary to group similar genes. Grouping of genes can be done by using various techniques like pattern matching, classification, clustering etc. In the present study clustering is used as a tool for analyzing biological data. Clustering of Biological sequences is a very interesting and fascinating area as various researchers are working on it. But simple clustering algorithms are not much suitable for sequence analysis problems. Most of the biological sequence analysis problems are NP-hard and some strong optimization algorithm are required for these types of problems.

The manuscript presented here is a survey of various clustering techniques useful for analysis of biological sequences. The 3+ stage review process is adopted for the review of literature. To prepare this report 98 papers have been reviewed from year 1997 to 2014 according to the year of publish. The papers reviewed have discussed various issues related to the analysis of biological sequences. The major issues discovered in the reviewed papers were prediction, sequence alignment, motif discovery, cluster boundary prediction etc. Various solution approaches used by researchers for the biological sequence analysis are evolutionary clustering, neural networks, hierarchical clustering, k-means, Go technologies, feature selection, incremental approach, bio-inspired methods, particle swarm optimization, fuzzy techniques, rough set theory and bi-clustering etc. Researchers have applied these solution approaches on various types of datasets. In this communication we have also discussed about these datasets and the parameters used with results mentioned in papers.

Keywords: Biological Sequences, Sequence Analysis, Clustering, Sequence Clustering

1. Introduction

Data clustering is a valuable tool for analyzing biological sequences. The goal is to analyze biological sequences using data clustering algorithms. Due to the biological mutation and evolution, there exist enormous complex sequence datasets. There is a number of research tasks related to sequence clustering such as pattern discovery, classification, clustering, and prediction, sequence alignment, motif discovery etc. Various approaches have been applied in the literature that is used by the researchers to analyze the biological sequences. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) (Jain et al, 1999). The presented communication has focused on clustering techniques used for biological sequence analysis. it also give a light on bi-clustering techniques used for sequence analysis problems. Bi-clustering technique was proposed by Cheng and Church (2000) and has been extensively used for analysis of microarray datasets. All reviewed papers in this manuscript have been categorized according to their solution approaches in clustering and bi-clustering approaches which are further classified according to the sequence analysis problems. Under Cluster techniques, we have categorized all papers which have solved the problems of biological sequence analysis using simple clustering techniques and

*Assistant Professor, Deaprtment of Computer Science, Maharaja Ganga Singh University, Bikaner, Rajasthan, India.
E-mail: jyotilakhanimgsu@gmail.com

experimented on sequential datasets whereas in second section, the categorized papers have applied bi-clustering techniques and worked on microarray datasets.

Problems solved by using various clustering techniques

The researchers have used various clustering techniques to solve problems related to sequence analysis. Following are some sequence analysis issues which have covered by researchers using clustering approaches-

1. **Pattern Matching-**(Huang et-al, 2002)
2. **Gene Searching-** (Tsai et-al, 2002)
3. **Gene clustering-** (Geng et-al, 2004), (Lee et-al, 2005), (Dutta et-al, 2012), (Kuo et-al., 2013)
4. **Grouping proteins-** (Krause et-al, 2005)
5. **Pattern Mining-** (Chezhian et-al, 2011)
6. **Gene Expression Analysis-** (Tjaden et-al, 2006), (Sharan et. al), (Shimokawa, et-al, 2007), (Ma et-al, 2009), (Swathi, 2010), (Sarmah, 2013)
7. **Evolutionary Computation based Approach-** (Leon et-al, 2006), (Sarafis et-al. 2006)
8. **Molecular Docking-** (Nesamalar et-al, 2012)
9. **Gene Selection- (**Mohamad et al., 2013)
10. **Feature Selection and Extraction-** (Sharma and Mukherjee, 2014)

Problems solved by various bi-clustering techniques

Using bi-clustering approach, following issues have been covered by researchers in our reviewed papers-

1. **Gene clustering –** (Zheng et-al, 2005), (Carmona-Saez et-al,2006), (Datta et-al, 2006), (Ceccarelli et. al., 2008), (Shon et-al, 2008), (Das et-al, 2014).
2. **Gene Expression Analysis-** (Toronen, et-al, 2004), (Lu et. al., 2004), (Yang, et-al, 2005), (Shamir, et-al, 2005), (Kanungo et. al. 2006), (Fu et al, 2007), (Yi et-al, 2008), (Sarmah et-al, 2010), (T.Chandrasekhar, et-al, 2011), (Roy et al, 2012),(Gao, et-al, 2012), (Parimala, 2013).
3. **Gene Selection-** (Sathishkumar, et-al, 2013).
4. **Stability of Cluster-** (Smolkin et al, 2003).
5. **Gene Prediction** – (Wang et al., 2003).
6. **Missing Value Prediction-** (Brevern et. al. 2004).
7. **Knowledge Discovery-** [Dawson, et-al, 2005], (Kléma, et-al, 2006).
8. **Grouping Genes-** (Reiss et-al, 2006), (Santamaría et al, 2008).
9. **Partitioning-** (Kim et-al, 2006)
10. **Cluster Validation-** (Preli´c et. al., 2006).
11. **Cluster Identification-** (Sen et-al, 2007).
12. **DNA Visualization-** (Zhu et. al., 2008).
13. **Stability of Clusters-** (Dresen et-al, 2008).
14. **Improving clusters-**(Hanczar, et-al, 2011), (Hussain, et-al, 2011).

2. Solution Approaches and Technologies Used

Several different techniques have been used by researcher for different sequence analysis problems using clustering and bi-clustering techniques.

A. Clustering

- **Parallel Pattern Matching-** Used by (Huang et-al, 2002) for pattern matching in biological sequences. It speeds up the matching process.
- **Multiple Searching Genetic Algorithm (MSGA)-** proposed by (Tsai et-al, 2002). He also hybridized it with k-means (MGSKA). This technique is useful for optimizing the clustering by facilitating with the combination of features of k-means algorithm and genetic algorithm.
- **Message Passing Clustering (MPC)-** proposed by (Geng et-al, 2004) for biological data.
- **Cluster Utility (CU) Metric-** (Lee et-al, 2005) has proposed a new Metric for clustering of biological sequences and get 95% hit rate.
- **Graph based SISTERS algorithm-** a hierarchical proposed by (Krause et-al, 2005) to group protein sequences. This technique can sometimes show weak results.
- **Super Paramagnetic Clustering (SPC)-** proposed by (Tetko et-al, 2005) for clustering protein sequences. These techniques can improve the specificity and sensitivity of clustering process but limited to the local scope. The extension of this, global SPC (gSPC) algorithm has also been proposed for considering global results.

- **Clustering Of Repeat Expression (CORE) -** proposed by (Tjaden et-al, 2006) for clustering of repeated expression.
- **Agglomerative algorithm -** presented by (Chezhian et-al, 2011) for sequential pattern mining and validated using statistical measures.
- **Vector Space Models -** (Semeiks, et-al, 2006) has constructed the vector space models for a pre-defined list of genes from the Gene Ontology (GO) terms and the Conserved Domain Database (CDD) protein domain terms assigned to the loci by the genecentered corpus LocusLink.
- **CLICK -** proposed by (Sharan et. al) for gene expression analysis.
- **ECSAGO** - a self adaptive clustering algorithm proposed by (Leon et-al, 2006). This algorithm performs clustering by finding distances between objects. This can be used for NP hard problems to get optimized results.
- **Rule-based Evolutionary Algorithm** - proposed by (Sarafis et-al. 2006) for high dimensional data. This algorithm can be used for analysis of microarray data.
- **A combination of hierarchical and non-hierarchical clustering -** used by (Shimokawa et.al, 2007) to cluster expression profiles of TSSs based on a large amount of CAGE data.
- **Iterative Approach –** proposed by (Ma et-al, 2009) for mining overlapping patterns in gene expression data.
- **Multi Objective Genetic Algorithm (MOGA) based K-clustering** - proposed by (Dutta et-al, 2012). It is a real coded method
- **Genetic Clustering Bee Colony Optimization (GCBCO) -** proposed by (Nesamalar et-al, 2012) to solve molecular docking problem.
- **Automatic Clustering using PSO (ACPSO)** has been proposed by (Kuo et-al., 2013).
- **Enhanced version of k-means algorithm-** has been proposed by (Shakti et al, 2013).
- **Enhanced Binary Particle Swarm Optimization (EBPSO) –** proposed by (Mohamad et al., 2013) to perform the selection of small subsets of informative genes which is significant for cancer classification.
- **Fuzzy Link Based Clustering (FLBC) -** has been proposed by **(Sarmah, 2013)** to perform gene expression data clustering.
- **Symbol Table -** (Baridam et-al, 2013) has presented a novel method for clustering of biological sequences by application of symbol table.
- **Grey level Co-occurrence Matrix (GLCM), ANFIS plus Genetic Algorithm and FCM-** (Sharma and Mukherjee, 2014) has proposed research work uses Grey level Co-occurrence Matrix (GLCM) for texture feature extraction, ANFIS (Adaptive Network Fuzzy inference System) plus Genetic Algorithm for feature selection and FCM (Fuzzy C-Means) for segmentation of Astrocytoma (Brain Tumor) with all four Grades.

B. Bi-clustering Techniques

- **Hierarchical Clustering** - has been used by (Smolkin et al, 2003) to find out the stability of clusters using for cancer microarray data. It is also used by (Toronen, et-al, 2004) for analysis of gene expression data. (Brevern et. al. 2004) have study the influence of microarrays missing values on stability of gene groups by hierarchical clustering. **[Boratyn et-al, 2006]** have introduced a novel biologically supervised hierarchical clustering algorithm for grouping genes based on gene expression data.
- **Multi-Source Clustering (MSC) -** has been proposed by (Yang, et-al, 2005). They investigated the problem of integrating gene expression data to produce more biologically significant clusters.
- **Isomap -** has been applied by (Dawson, et-al, 2005 for the interpretation of microarray data.
- **Fuzzy Partition Algorithm -** has been developed by (Kim et-al, 2006) for normalized DNA microarray data.
- **Bipartite Graph** - (Kanungo et. al. 2006) has introduced an efficient framework for biclustering of gene expression data using bipartite graph.
- **Fuzzy clustering by Local Approximation of MEmbership (FLAME) -** has been developed by (Fu et al, 2007) to extract information from gene expressions obtained with DNA microarrays.
- **Semi Supervised Fuzzy C-means Clustering (SSFC) -** proposed by (Ceccarelli et. al., 2008) for biological data.
- **Markov Cluster Algorithm -** has been presented by (Shon et-al, 2008)

- **VIsual Statistical Data Analyzer (VISDA) -** (Zhu et. al., 2008) has developed the VIsual Statistical Data Analyzer (VISDA) for cluster modeling, visualization, and discovery in genomic data.
- **Response Projected Clustering (RPC) -** introduced by (Yi et-al, 2008) which have used a high-dimensional geometrical projection of response data to the gene expression space.
- **Resampling Method -** proposed by (Dresen et-al, 2008) based on continuous weights to assess the stability of clusters in hierarchical clustering.
- **Imputation Methods -** (Tuikkala et al, 2008) compared a number of advanced imputation methods on recent microarray datasets.
- **GenClus and InGenClus -** (GenClus) was proposed by (Sarmah et-al, 2010) for gene expression data and he has also introduced InGenClus which can handle incremental data.
- **Ensemble Methods -** (Hanczar, et-al, 2011) has proposed a new approach for improving the performance of bi-clustering algorithms by using ensemble methods.
- **Field Programmable Gate Arrays (FPGA) method -** has been proposed by (Hussain, et-al, 2011), a highly parallel hardware design to accelerate the K-means clustering of Microarray data by implementing the K-means algorithm in FPGA.
- **biClusters with Linear Patterns (CLiP) -** introduced by (Gao, et-al, 2012) for bi-clustering of linear patterns in gene expression data.
- **Enhanced Bi-clustering Algorithm** - has been presented by (Parimala, 2013) for gene expression data and to optimize the performance of bi-clustering process.
- **Two Phase Algorithm** - has been introduced by (Das et-al, 2014)

3. Data Used and Results

A Clustering Techniques

Parallel Clustering

(Haung et-al., 2002) have implemented parallel algorithm on 1-32 processors. The parallel algorithm was applied on EMBL Nucleotide Sequence Database. 35K to 0.5M long sequences have been used and performance of the algorithm was measured on the basis of running time. There was a clear cut decrease in running time with the increase in parallelism.

Gene Expression Analysis

(Saharan et-al.,2003) use CLICK and EXPANDER for clustering and visualization of gene expression. Simulated data sets of various sizes containing 10 equal size clusters with parameters- standard deviations 0.75,1,1.5, 2.0,2.5 and Cluster no./ Size ratio -5/100, 10/50, 6/50,60/100. Jaccard coefficient was used as the performance measure which results 94.4% accurate classifications. (Gang et-al., 2004) has introduced Message Passing Clustering (MPC) algorithm for clustering gene expression data. They have applied MPC on 35 sets of simulated dynamic gene expression data total 674 genes with 10 to 100 genes time points for 20-40 dimensions as parameter. The performance measure used was the hit rate for correct clustering which was resulted in a 95% hit rate in which 639 genes out of total 674 genes were correctly clustered. (Brain Tjaden, 2006) proposed CORE for clustering gene expression with error information. He used synthetic expression data as well as real gene expression data from Escherichia coli and Saccharomyces cerevisiae. The parameters applied were n = 1000 genes over m = 50, degrees of freedom=14 and scaling parameter = 0.1. Error weighted similarity and Rand Index were used as performance measure. To reduce the cluster cost of gene expression data, (Shimokawa et al., 2007) have made efforts to combine the hierarchical and non-hierarchical clustering and perform large scale clustering on genome-wide expression data, including 159,075 TSSs derived from 127 RNA samples of various organs of mouse. The parameters inputted for algorithm was number of clusters=0 to 200 and p-value was used as the performance measure. (Swathi, 2010) has proposed a hybrid Hierarchical k-Means algorithm and bi clustering algorithm for local and global clustering of gene expression data. Yeast cell cycle dataset with 317 genes over two cell cycles ranges from 10 min to 24 hours has been used. Pearson's Correlation Coefficient has been used as the performance measure. The combination of clustering and bi-clustering gave better results but using both on the same data and at the same time was an overhead. (Sharma et al., 2010) have proposed a fast algorithm GenClus for incremental gene expression data. The algorithm was implemented on three different datasets-Yeast CDC28-

13 with 6218 genes and 17 conditions, Yeast Diauxic Shift with 6089 genes and 7 conditions and a Subset of Human Fibroblasts Serum with 517 genes and 13 conditions. Number of clusters and z-score was used as performance measure. Average 61 clusters and z-score = 7.39 has been obtained as a result. (Das et al., 2010) have proposed FINN and DGC with a new dissimilarity measure for clustering gene expression data to find finer and embedded clusters. They have experimented on 6 different datasets: Yeast diauxic shift database with 6,089 genes and 7 conditions, Subset of yeast cell cycle with 384 genes and 17 conditions, Rat CNS with 112 genes and 9 conditions, Arabidopsis thaliana with 138 genes and 8 conditions, Subset of human fibroblasts serum with 517 genes and 13 conditions, Yeast cell cycle with 698 genes and 72 conditions. As parameters σ=2, cutoff=46 and θ=0.7 have been used. Total number of 7 clusters have been recognized with z-score=12.6. (Dutta et al, 2012) have used MOGA based k-clustering which was effective for local and global clustering. Acute Inflammations and Liver Disorders datasets for number of runs (k) =10 have been used. The performance measure used was DB Index. This algorithm combined global optimum of Genetic Algorithm and Local Optimum of k-means. (Sarmah et al, 2010) have proposed a fuzzy link based approach and implemented on three datasets: Subset of Yeast Cell Cycle dataset with 384 genes 17 conditions, Rat CNS dataset with 112 genes 9 conditions and Yeast Diauxic Shift dataset with 6089 genes 7 conditions. The excellent results with 0.94 homogeneity, 0.09 separation and 2.59 z-score have been found with δ=0.8-0.95. In (Baridam et al, 2013), symbol table have been used for fuzzy and hard clustering of gene expression data. This technique was implemented with breast and colon cancer gene expression data. Using fuzzy clustering data 5 clusters have been detected where as with hard clustering only 3 clusters have been found.

Cluster Utility Matrix

(Lee et-al., 2005) have proposed the use of cluster utility matrix to cluster biological sequences and show the relation of input data with cluster utility. The data set used for the experiment include four classes COG0002, COG0160, COG1959, COG3000 with 29, 79, 42, 17 sequences respectively. The parameter used was cut-off value set to 50. A clear indication of 68.19% reduction in incorrect clusters has been seen.

Clustering Protein Sequences

(Krouse et al, 2005) have been proposed SYSTERS algorithm for hierarchical clustering of protein sequences to analyse evolutionary relations of sequences in clusters. The algorithm was applied on protein sequences from the Swiss-Prot Rel. 41 and TrEMBL Rel. 23 databases and Jaccard Coefficient was used as performance measure which showed best results for sub clusters. This algorithm did not perform well for Pfm dataset. Tetko et-al. introduced SPC and gSPC algorithms to cluster protein sequences. Researchers used SwissProt and SCOP databases. Authors used Specificity and Number of Errors as performance measure and results indicated that the proposed algorithms improve specificity and sensitivity up to 30%. Results of experimentation were evidence of occurrence of 7.7 % to 14.8% number of errors.

Evolutionary Clustering

(Leon et-al, 2006) have proposed ECSAGO which could detect clusters without knowing number of clusters in advance. Authors have used 12 two dimensional datasets and set the parameter population size = 100, generations = 50, weight threshold = 0.3, fitness crossover = 0.6 Refinement MDE iterations = 10 and Extraction fitness threshold = 25 which results in 94-100% success rate of finding clusters.

(Semekis, 2006) had proposed ensemble attribute profile clustering NOCEA which is a rule based clustering. He experimented it on Conserved Domain Database (CDD) -52 genes represented as 43-dimensional vectors (the LUMINAL collection), myoepithelial cell markers consisted of 89 96-dimensional vectors (the MYOEPITHELIAL collection) with parameters population size=50, generations=300, mutation probability=0.01. As an experimental result, algorithm found four consensus clusters for 52 number of genes.

Pattern Matching

A new similarity measure CLUSS was proposed by (Kelil et al., 2007) for clustering protein sequences. The COG database and the Gproteins of GH2 and ROK families were used for empirical study. The approach used substitute the matching similarity measure. The performance measure used was Q-measure with standard deviation 3.57% which show 92% accuracy.

Overlapping Clusters

(Ma and Chan, 2009) have proposed an iterative algorithm to discover overlapping clusters. They have implemented this algorithm on an artificial dataset containing 450 genes and set the number of clusters (k)=9 as parameter. Adjusted Rand Index was used as the performance measure. The algorithm run in two phases and the major flaw of this algorithm was that the overall results of this algorithm were depended on only first phase of the algorithm.

Outlier Detection

(Thom et-al, 2009) proposed a new fuzzy algorithm for outlier detection. A synthetic dataset with outlier point 12, 400 was used with threshold α=0.5. Membership value was used as the performance measure and the algorithm generate 0 or 1 for value of membership. Zero membership denoted for outliers. This algorithm was not effective for large datasets.

Hierarchical Clustering

(Chezhian et al., 2011) have used UPGMA for hierarchical sequence clustering. Synthetic datasets have been used for number of clusters N= {10, 50, 100, 200, 300, 500, 1000}. Computation times for similarity and dis-similarity matrix have been computed. This approach acquire O(n3) computing time.

Protein Docking

(Nasamlar et al, 2012) have proposed GCBCO a Bee Colony Optimization based algorithm for flexible protein docking. They have used Zinc Code database with binding energy as performance measure. The results show that GCBCO find lower binding energy by proposed algorithm.

Visualization

(Campello et al, 2012) have used Lloyd's k-means and progressive greedy k-means, density based clustering techniques, for clustering and visualization of DNA sequences. A randomly generated dataset have been used with k(number of clusters)=2,3,4 and 5. The performance measure used was running time and results clearly indicated that Lloyd's k-means is 13% faster than other techniques.

Co-expression Analysis

FLBCI was used to find out co expression of genes and empirical study with Synthetic dataset find 91.20% co-expression of genes with the presented technique. Z-score, p-value and Q-value have been used as performance measure.

Gene Selection

In (Mohamad et al, 2013) BPSO and EBPSO have been used for gene selection problem. The empirical study was performed on 10 various datasets. These were- i) 11_Tumors dataset with 174 samples 12,533 genes and 11 classes; ii) 9_Tumors dataset with 60 samples 5,726 genes and 09 classes iii) Brain_Tumor1 dataset with 90 samples 5,920 genes and 5 classes; iv) Brain_Tumor2 dataset with 50 samples 10,367 genes and 4 classes; v) Leukemia1 dataset with 72 samples 5,327 genes and 3 classes vi) Leukemia2 dataset with 72 samples 11,225 genes and 3 classes vii) Lung_Cancer dataset with 203 samples 12,600 genes and 5 classes; viii) SRBCT dataset with 83 samples 2,308 genes and 4 classes; ix) Prostate_Tumor dataset with 102 samples 10,509 genes and 2 classes x) DLBCL dataset with 77 samples 5,469 genes and 2 classes. Parameters used for this study were number of particles=100 and number of generations=500. Accuracy and fitness function were used as performance measure. The average accuracy of the genes selection was 85.56%. EBPSO show better results than BPSO.

Gene clustering

A rough subspace based clustering was applied on Cancer gene dataset with fuzzification coefficient m=1.1 to 2.0 as parameter. Approximately 68% accurate results have been generated by this study (Gao et al). A. Hatamlou has proposed Black Hole algorithm (Hatamlou, 2013) for gene clustering and applied it on breast cancer dataset with level of confidence =0.05. As a result of experimental study 96.65% intra cluster distance, 10.02% error rate has been found. Automatic clustering has been performed by (Kuo et al, 2013) by implementing ACPSO algorithm which is a combination of PSO and k-means. The parameters used for empirical study were iteration=200 pop size=20 learning factor=1.49 Inertia weight=0.9-0.4. As result avg 5.15 clusters have been found with standard deviation 0.62 A combination of fuzzy c means, ANFIS and Genetic algorithm has been used for gene segmentation

and the experimental study has been performed on Brain tumour database. The parameters used for experiment were Population Size= 10 and Maximum Chromosome Length= 7. Total 96.6% sensitivity, 95.3% specificity and 98.67% accuracy have been found in results.

B. Bi-clustering Techniques

Gene clustering

In (Getz et al., 2000), Couple Two Way Clustering (CTWC) algorithm has been proposed for gene clustering analysis of microarray data. The proposed algorithm has been implemented on Colon Cancer and Leukemia microarray datasets with expression level greater than 20. The proposed algorithm could detect masked and hidden partitions and correlations. Number of genes identified was used as performance measure. The proposed algorithm has discovered stable clusters and sub clusters from the datasets. (Wang et al, 2002) has performed a two level analysis by a combination of k-means and SOM algorithm and implemented on public data from diffuse large B-cell lymphomas. The SOM map size used was 22 x 14. P-value was used as performance measure. This study found interesting clusters of mutually exclusive expressed genes.

(Preli´c et al., 2006) applied hierarchical clustering HCL and GO tools on SAGE profiles of differentially expressed tags between normal and ductal carcinoma in situ samples of breast cancer patients and data set contains the expression profiles over time of positively expressed genes (ORF's) during sporulation of budding yeast. Parameters used were classes=11, iterations =500, probability<5% and k=6 to 9. The performance measure was BHI and BSI. Results indicated that bi-clustering techniques were easy to interpret and easy to implement than other algorithms.

Gene Expression Analysis

IGKA, a combination of k-means and Genetic algorithm have been proposed in (Lu et al., 2004) for gene expression analysis of microarray data. Datasets used were serum data contains expression data for 517 genes and yeast data *chodata* composed of expression data for 2907 genes. The parameters used were population = 50, generation= 100 threshold (0.005 for fig2data, and 0.0005 for chodata). The running time of IGKA was less than other algorithms. In (Yin et al, 2006) Cluster Diff was used and tested on Yeast Saccharomyces cerevisiae gene expression data and efficiently determine the closest match. Fuzzy nearest clusters have been used with 5 fold cross validation in. The TNA was used as performance measure. The TNA value for proposed algo was 65.27%, which is 4.55%, 23.17%, and 4.76% higher than KNN, L-SVM, RBF-SVM respectively. Yeast dataset was used in (Corchado et al., 2006) for implementing Similar Nearest Neighbour with threshold=14. The final number of clusters was low and genes within show a significant level of cohesion. A compendium of time series gene expression data measuring the responses of human cells to infectious agents have been applied with GO tool which results in significantly enriched GO categories (50%) or KEGG pathways (59%) (Li et al, 2012). Two-stage clustering algorithm (Bandyopadhyay et al, 2007), SiMM-TS CLUSTERING TECHNIQUE have been applied on Yeast Sporulation (Chu et al., 1998), Human Fibroblasts Serum (Iyer et al., 1999) and AD400_10_10 datasets with parameters generations=100, pop size=50, cross prob=0.8, mutation prob= 0.01 and iterations= 200. In results, SiMM-TS clustering scheme outperforms all the other clustering methods significantly. In (Yi et al, 2008), Response Projected Clustering was proposed and it was applied on Docetaxel chemosensitivity and microarray gene expression data and RPC Analysis on NCI-60 Data. False discovery rate(FDR)=0.2 or less was applied and results show that no gene was found with FDR < 0.05. SCC-based hierarchical clustering (Yao et al, 2008) was applied on synthetic expression data as well as real gene expression data from Saccharomyces cerevisiae with low noise level = 0.5 and high noise level=2.5 as parameters. In this method Shrinkage correlation coefficient has been used which outperform the Pearson correlation. (Das et al., 2009) have proposed Regulation based density clustering and applied on 5 datasets- Yeast Diauxic Shift; no of genes=6089; conditions=7, Subset of Yeast Cell Cycle; no of genes=384; conditions=17, Arabidopsis Thaliana, genes=138, conditions=8, Subset of Human Fibroblasts Serum; genes=517; conditions=13, Yeast Cell Cycle; genes=698; conditions= 72. Parameters set were σ=2, k=2 to 30, and increments=2. RDCLUST perform better than its counterparts. In (Xu et al., 2009) CeaGO has been applied on two acute leukemia (ALL and ALL/AML) microarray expression datasets, with parameters d<2 and p<0.5 to perform cluster enrichment analysis. This method identified groups that are significantly differentially expressed in microarray data. R/BHC (Savage et al, 2009) has been applied on expression profiles of A. Thaliana for fast clustering. The proposed algorithm could handle uncertinity in data effectively but the limitation of this method was that this method show overlaps between clusters. GRASP and Reactive GRASP (Dharan et al, 2009) has been applied on Yeast Saccharomyces

Cerevisiae cell cycle expression dataset with δ=200 as parameter and HScore and p-value as performance measure. Reactive-GRASP outperforms GRASP and Bi-clustering with p-value ranging from 2.9778 to 6.0545. In (Dharan et al, 2009) CMOPSOB –SOM was tested with yeast and human B-cells expression datasets with p-value<0.01 and achieved a good diversity in the obtained Pareto front, and rapid convergence. For clustering incremental gene expression data, Density based clustering has been applied on – i) Yeast CDC28-13 with genes=6218 and conditions=1; ii) Yeast Diauxic Shift with genes= 6089 and conditions=7; iii) Subset of Human Fibroblasts Serum with genes= 517 and conditions=13. The parameters used were Mean =0 and Std Deviation =1. Results show that k-means find out 62 clusters in total 614 genes with z-score 5.57; SOM findout 42/614 z-score 5.78; GenClus 61/614 z-score =7.39. In (Zhang et al, 2010), Binary Matrix Factorization have been applied on four datasets – i) ALL/AML with samples=38, genes=5000 and class =2; ii) ALL/AML with samples =38, genes =5000 and class =3; iii) Lung Cancer with samples =32, genes =5000 and classes=2; iv) CNS with samples=34, genes=5597 and classes=4 with parameters tg = 2.0, tc = 2.0 and seeds = 50 for analysis of gene expression. The algorithm finds the global maxima. (Das et al., 2010) has proposed DGC and FINN for clustering gene expression and applied it on seven different datasets- i) Yeast diauxic shift with genes= 6,089 and conditions=7; ii) Subset of yeast cell cycle having genes=384 and conditions=17; iii) Rat CNS having genes=112 and conditions=9; iv) Arabidopsis thaliana with genes =138 and conditions=8; v) Subset of human fibroblasts serum with genes = 517 and conditions =13; vi) Yeast cell cycle having genes= 698 and conditions= 72. The parameters were σ=2, θ=2, k = {16, 20, 30, 40, 48} and cutoff=46. Results indicate that FINN outperforms k-means, SOM, DCCA and FINN has identify finer clusters and may detect embedded clusters. In (Krishna et al., 2010), temporal precedence based clustering has been proposed and applied on synthesized datasets and real biological datasets obtained for Arabidopsis thaliana with parameter α = 0.05 for clustering gene expression data. This algorithm produce sets of functionally related genes. AutoSOME has been proposed in (Newman and Cooper, 2010) and implemented with Gene ontology enrichment Database for Annotation Visualization and Integrated Discovery (DAVID) with parameters k=3 and p<0.05. F-measure was used as performance measure. The proposed method identify >3400 up-regulated genes and effectively extracting important information without prior knowledge or the need for data filtration. MGC (Zeng et al., 2010), have been applied on four yeast expression time series datasets with parameter w=0.5. This method successfully combined multiple constraints from different nature for gene clustering. Multi-core Parallelization algorithm has been applied on SNP dataset with kÎ = {2, 3... 10}. Parallelization makes it possible to perform laborious tasks as cluster sensitivity and cluster number estimation easy. QPCR arrays (Lock et al., 2010) have been used for efficiency clustering and applied on synthetic datasets with parameters N = 96-102, n < 10, and p < 0.0001. Using n < 10, cluster-based mean efficiencies was comparable to using individually determined efficiency adjustments for each primer pair (N = 96-1024). The major limitations of proposed method were the imprecision in assuming uniform efficiency, danger of over fitting and Computational complexity. In [54] Simulated expression data have been used to implement Merged consensus clustering (Simpson et al., 2010). Consensus matrices can be used as corrected distance matrices, to calculate cluster and membership. FPGA & K-means have been used for real Yeast Microarray data. This approach speed up the data analysis process but re-sampling was required for this method which was an overhead. (Oghabian et al, 2014) compare clustering algorithms for gene expression analysis. they implemented this algorithm on 228 samples from 5 distinct healthy human tissues from the GeneSapiens database [10]: 59 blood t-cell, 95 cerebral cortex, 13 liver, 41 striated muscle, and 20 testis samples with α=0.05%. Results show that k-means over performs all other algorithms. Two phase clustering (Das et. al., 2014) was performed on yeast Saccharomyces cerevisiae cell cycle expression data from Cho et al., 1998. (Roy et al, 2013) have applied GeCON, to find out gene co-expression and experimented with real and synthetic gene expression data consisting of publicly available seven benchmark gene expression datasets and thirteen in-silico datasets from the DREAM. This method found out local expression patterns and extract functionally enriched network from gene expression. Requirement of pre-processing was the major drawback in this method.

Comparison of Clustering Algorithms

In (Bagyamani et al., 2014) a comparison of bi-clustering algorithms has been performed by applying it on 228 samples from 5 distinct healthy human tissues from the GeneSapiens database with parameters δ=50 and α=1.5.In results bi-clustering methods such as Plaid and SAMBA were useful for discovering relevant subsets of genes. SIMBIC+ was implemented on Yeast Saccharomyces dataset with parameters i=288; j=14 for performance measure MSR. SIMBIC+ outperforms SIMBIC and MSB. A Comparison of clustering algorithms has been performed on 52 gene expression microarray datasets

having Distance measure and p-value as performance measure (Roy et al, 2013). PE, EUC, and SP, provided better results than other distances in 72%, 70% and 65% in this research.

Gene Selection

In (Tritchler et al, 2009) filtering genes method has been used for selection of genes for clustering. This method was applied on E. coli and S. cerevisiae regulatory network dataset with parameters SUMCOV= 95% and VARCOV=95% . Results show that SUMCOV performed well for all models. In (Sathishkumar, et-al, 2013) Gene Extraction has been performed using Bi-clustering and applied on yeast genome (Saccharomyces cerevisiae), a human cancer dataset and random data with parameters- overlapping cluster=50 each cluster extends over 5 genes and 15 conditions. The algorithm identify medium percentage (> 30 percent and < 80 percent) of the embedded bi-clusters at different noise levels.

Gene Cluster Analysis

(Medeiros et al, 2005) has performed text mining on data related to Saccaromyces cerevisiae genes with SGB links for gene cluster analysis and detect error rates as performance measure. The errors rates with and without the ontology was very small and the error rates have been reduced 35% after the use of GO. In (Scharl et al, 2009) gcExplorer has been used for gene cluster analysis. it has been applied on E .coli cultivation data with different cutoff values. This method found clusters with similarity>10%-30%.

Overlapping Bi-Clusters

In order to find out overlapping clusters, CLAG has been used in (Dib and Carbone, 2012) which was implemented on i)Breast tumormiRNA expression data; ii)Brain cancer gene expression data; iii)Coevolved residues in protein families data; with score threshold=0.25 and δ=0. Results indicate that CLAG did not perform better than mixtur-model based method.

Gene Association Analysis

In (Riess et al, 2006), hierarchical clustering algorithm was used for analysis of gene association. Haplotype, lung cancer and lymphoma datasets were used with parameters m=100,000 and steps=0 to 20. The p-value was used as the performance measure which was 0.5263, 0.0321 and 0.0364 respectively for datasets.

Gene Prediction

(Wang et al., 2003) has proposed a hybrid technique with SOM and Fuzzy c-means for gene prediction in microarray data. The datasets used for experimentation were :(i) Leukemia dataset consisted of 38 bone marrow samples (ii) Colon cancerdataset containing intensities of 2000 genes in 22 normal and 40 tumor colon tissues (iii) Brain tumors dataset consisting of 42 brain tumor samples (iv) NCI60 data set contained 61 cell lines derived from human cancers from a variety of tissues and organs. Test errors and misclassification rates were the performance measure and the respective results for these datasets were: 2.62% to 5.88%, 9.68%, 13.53%, 20% misclassification.

Missing Value Prediction

(Brevern et al, 2004) have used hierarchical clustering for missing value analysis and implemented it on Yeast data set with parameters: number of genes= 552-16523; range=4 to 178; % MVs= 0.8 - 10.6% and τ = 1%. As a result 0.8% to 10.6% missing values have been detected. Imputation algorithm in combination with GO tools has been applied on 8 datasets- Brauer05, Ronen05, Spahira04A, pahira04B, Hirao03, Yoshimoto02, Wyrick99 with parameters missing values=0.5%, 1%, 5%, 10%, 15% and 20%, and k = 2, 3,..., 10. Results suggested that knn is 10 times faster than imputation methods .

Knowledge Discovery

In (Parimala, 2013), bi-clustering has been enhanced by using gene expression matrix resulted in 14% performance improvement. Unsupervised gene clustering has been performed on simulated data sets for integration of biological knowledge. This results in highest proportion of good candidates in bi-clusters. But the external biological validation was not performed in this method to validate results (Verbanck et al., 2013).

Grouping Genes

(Petri Toronen, 2004) have performed gene classification on two Yeast datasets-MIPS and SDG. The number of

clusters detected (size > 5 genes) was the performance measure and resulting 175 clusters with SGD classes and 116 clusters with MIPS classes from an average cluster tree has been detected. This technique could not be used with big dataset.

Partitioning

Reduced Pheripheral Blood Monocytes (RPBM) dataset, yeast cell cycle (YCC) expression dataset, hypoxia response (HR) dataset and mouse tissues (MT) dataset have been used for FLAME algorithm. Local/Neighborhood Approximation Error was used as performance measure which provides better partioning of biological functions (Sarmah, 2013). In (Hartsperger et al, 2010) k-partite graph partitioning algorithm was applied on CORUM database with parameters $\mu >= 0.9$ and p-value < 10-5. This algorithm was able to structure large cluster into functionally correlated and biologically meaningful small clusters.

Bi-Cluster Validation

In (Okada, 2005), Dunn's index, Devis-Bouldin index and Silhouette index have been used on MIPS database contains expression data of 1609 genes for five time points taken at 0.25, 0.5, 2, 4 and 8h after cold shock to yeast cells to validate bi-clusters boundaries. The Gene Function Vector (GVF) has been used as performance measure with reference to MIPS functionalCatalogue data. This technique obtained novel annotations. In (Garge et-al, 2005) validity of clustering been performed by detecting reproducible clusters. 37 microarray datasets have been used with sample sizes ranging from 12 to 172. The performance measure was Cramer's v2 from a k x k table which found avg. stability < 0.55 for the four clustering routines, even at a sample size of 50. George et al have used the concept of reproducible clusters for identify complex relationship in clusters and verify the validity of clusters of sample size ranging from 2 to 172. In (Kléma, et-al, 2006), distance was used as validation measure and applied on SAGE data of breast cancer patients and time course cDNA microarray data on yeast to evaluate various clustering algorithms. Results are calculated at 5% significance level and it was shown that k-means and Diana are amongst best algorithms.

Bi-Cluster Identification

In (Tetko et. al., 2005) ISOMAP has been introduced for clustering microarray data. The experiments have been performed on soft tissue tumors, S. cerevisiae cell cycle, and serum-induced fibroblast, 170 rat U34A high density oligonucleotide arrays with 8,799 genes on each array. Isomap can successfully detect biologically relevant structures even in the background noise. Fuzzy c-means algorithm has been proposed in (Thomas et-al) from expression data. Serum, Sporulation and Yeast datasets have been used with Silhouette index, Rand index, Similarity as the performance measures. FCM shows 85.2%, 63.7% and 97.8% similarity for Serum, Sporulation and Yeast datasets whereas PAM shows 81.7%, 62.9% and 93.8%. This technique could not found co-express genes. In (Carmona-Saez et-al, 2006) bi-clustering has been performed by using Non-smooth Non-Negative Matrix Factorization (nsNMF) and applied it on three datasets- synthetic dataset, human transcriptome dataset, soft-tissue tumour dataset. The parameters set for these datasets are- for synthetic data-100 × 20 noisy matrix and rank= 2 or 3; for Human transcriptome dataset-rank value =8, soft tissue dataset rank value=4. P value and rank value have been used as performance measures. The value of factorizations in results were - for synthetic data=60%; human transcriptome=80% and for soft tissue dataset=90%. Results indicated that cMonkey clusters were more parsimonious. In (Li et al.), hierarchical clustering and GO tools have been used and applied on synthetic and real data sets for Saccharomyces cerevisiae and Arabidopsis thaliana. The parameters used were n = 100,m = 50 for scenario 1 and n = 100,m = 100, . . . , 108. The rate of identification of clusters was >90% with this technique.

Exact, Randomized, Extension algorithms (RMSBE) have been applied on cDNA microarray data with parameters D= 40, N1=4, N2=4, k=20, L=10 to find maximum similarity bi-clusters. RMSBE method could find high quality bi-clusters and 98% bi-clusters Recognized were statistically significant (Sathishkumar, et-al, 2013). In (Bagyamani et al., 2014), BiModule method have been applied on Human Tissue data with parameters L=7, Mg=40 and Mc=5. In results the number of significant bi-clusters found was 65 for GO, 50 for PPI, 56 for Motif and 12 for Pathway, respectively. A leukaemia data set was used in (Shon et al, 2008) for Stochastic method with parameters genes =7129 and samples= 72. Muscular dystrophy data and muscle regeneration data have been used for VIsual Statistical Data Analyzer (VISDA) for parameters n=9 or 10. This technique obtained the correct cluster number across the

cross-validation trials with an average frequency of 97%. Yeast Saccharomyces cerevisiae microarray dataset have been used for BicOverlapper with 2467 × 79 matrix as parameter. This technique could detect super bi-clusters and hub nodes very easily (Santamaria et al).

A new measure NIFTI has been introduced in (Jonnalagada et al, 2009) to identify bi-clusters and applied on Yeast cell-cycle, Serum, Lymphoma, Pancreas datasets. NIFTI outperforms Silhouette, Dunns, Davis Bouldin but the limitation of this index was that it could produce incorrect results if cluster do not have spherical shape. In (Marttinen et al, 2009), Bayesian clustering and feature selection have been used and applied on simulated data and a large database of DNA copy number amplifications in human neoplasms with parameter α=.875, .9, .95 and .975. This method automatically recognized the chromosomal areas that were relevant for the clustering. In (Kim et al, 2009), MULTI-K algorithm, an ensemble clustering method, has been applied on Leukemia, Lymphoma, Colon tumor, Thyroid tumor I, Thyroid tumor II, St. Jude, Normal tissue I, Normal tissue II with parameters α=0.5, 0, -0.5 for classification of microarray subtypes. Results indicate that the MULTI-K outperformed other methods. Ensemble Methods implemented on Brain Cancer, Brain 2 and Lung Cancer in (Hanczar, et-al, 2011) and produced better bi-clusters. The major limitation of this method was the computation time required which was very high. In (Chandrasekhar, et-al, 2011), ISODATA- AGMFI and ISODATA- EAGMFI have been experimented on – i)Serum dataset with 517 genes, ii) Yeast dataset with 2884 genes and 17 conditions; iii) Simulated data with 300 Genes. During the experimentation parameters set upon were- for serum dataset: k=10, 5 times running and clusters=7; for yeast data set: k=34, 5 times running and clusters = 18. The performance measure used was Silhouette Coefficient and the average initial and final clusters found were – Serum dataset: initial clusters =10 and final cluster=7; Yeast Dataset: l cluster=34, final cluster=18; Simulated dataset: initial 10, final cluster=6. Non-parametric density estimation technique has been implemented on simulated data with n<=20. The error rates acquired for pdfClust=0.08 and Mclust=0.77. The major limitation was that Mclust perform badly for moderate size clusters. Pattern driven neighbourhood search(PDNS) have been applied on microarray datasets (Yeast cell cycle and Saccharomyces cerevisiae) with parameters α=0.8, β=0.8, threshold_ASR=0.7, Y=100, Z=50, p-values (p = 5%,1%, 0.5%, 0.1% and 0.001%). Best results were found with p=0.0001(Ayadi et al., 2012). A Bi-partite graph algorithm **(**Kanungo et al**,** 2012**)**, BiRange have been applied on Yeast Saccharomyces Cerevisiae dataset with p=0.33. This method found highly accurate bi-clusters and was effective for noisy data. Lee et al [62], have compared various bi-cluster methods by applying on two expression datasets (GDS1620 and pathway) with different dimensions in Arabidopsis thaliana (A. thaliana). In results SAMBA show less data dependency and QUBIC & FABIA has poorly performed. Greedy search and genetic optimization methods applied on 3 different synthetic datasets and two well-studied yeast (S. cerevisiae) cell cycle datasets: the Tavazoie et al. (1999) dataset and Spellman et al. (1998). The parameters set were Initial population size = 100 and total generations = 200. Bi clustering of linear patterns uncovered a set of similarly expressed genes (Gao et al, 2012). Bi-clustering of co-regulated genes have been performed by one-pass association mining technique(OPAM)(Roy et al, 2012) and implemented on Yeast Microarray dataset with parameters θ= 50% and τ= 15. This technique identified biologically significant gene bi-clusters. In (Gusenleitner et al, 2012) iBBiG algorithm has been used and applied on Single sample GSA data; GSEAlm data (Breast Cancer Datasets) with parameter k=8. This method show high specificity (99%) and precision (92%) in finding out bi-clusters. The major limitation of this approach is that the results lack sensitivity in results by 56%. (Henriques et al., 2014) has applied a flexible biclustering**,** BicSPAM, on synthetic data and discover flexible sequential structures of order preserving biclusters for noise upto 10%, θ=8%, α=5, p value=0.01 and 0.05 as parameters. BicSPAM could surpass the drawbacks identified for existing order-preserving approaches. The drawback of this method was that it could not search lengthy sequential patterns.

Stability of Clusters

In (Wang et al., 2002), a hierarchical sub sampling technique has been proposed which check the stability of clusters. The datasets used for the empirical study were childhood cancer study, lymphoma study and cutaneous melanoma study datasets. The parameters used for experimentation were Cluster Stability Score d = {0.85 p, 0.75 p, 0.5 p and 0.25 p}, where p is the number of genes. R Index and cluster stability scores were treated as performance measure. The limitation of this technique is that it is suitable for gene expression data only. (Dresen et al., 2008) have proposed Resampling Method for the stability of clusters. The method was applied on uveal

melanoma dataset of Tschentscher et al. Simulated data with Number of genes as parameter and Rand Index as performance measure. Results indicate that imputation always gave better results than ignoring missing data points or replacing them with zeros or average.

Improving clusters

Exploratory analysis technique has been used for cluster improvement. This technique have been applied on dataset consisted of 5473 genes and 250 terms with parameter k=2 to 100. Z-score was treated as a performance measure and the z-scores were plotted against the number of k. In the outcome z-scores decay as k grows over 70 (Scharl et al, 2009). In (Li et al, 2009) QUBIC was used for qualitative bi-clustering on simulated data sets with parameters r=1, q=0.06, c=0.95, o=100, f=1. Empirical study reveal that the accuracy of QUBIC with r=1 was 69%. CLEAN (67. Freudenberg et al., 2009) was applied on human and mouse datasets, selecting the top 5,000 genes with the overlapping genes in the two datasets with parameter q-value=0.01 and performance measure coherence score (CLEAN score). Coherence score simplified the comparisons of clustering.

Cluster Boundary Analysis

(Dettling et al., 2002) have performed supervised clustering for clustering margin analysis. The proposed algorithm was implemented on eight datasets: (i) Leukemia dataset with n = 72 and p= 3,571 genes (ii) Breast cancer dataset with n =49 breast tumor samples and p = 7,129 genes (iii) Colon cancer dataset with 40 tumor and 22 normal colon tissues for 6,500 human genes and p= 2,000 genes (iv) Prostate cancer dataset with n= 52 prostate tumors and 50 non-tumor prostate samples and p = 6,033 genes (v) SRBCT dataset with 20 SRBCT and 5 non-SRBCT samples and p = 2,308 genes (vi) Lymphoma dataset with n = 62 and p = 4,026 (vii) Brain tumor dataset with n = 42 and p = 5,597 genes (viii) National Cancer Institute (NCI) dataset with n = 61 human tumor cell lines and p = 5,244 genes where n was number of observations and p was number of genes. The misclassification rate was the performance measure and the respective misclassification rate found for these datasets were 4.415714%, 1.201429%, 19%, 9.328571%, 1.225714%, 1.43%, 28.70429% and 32.66857%. In (Boratyn et-al, 2006), researchers proposed Biologically Supervised Clustering technique and detect overlapping clusters from Yeast time course cDNA data set and Normal versus breast carcinoma, SAGE dataset. Distance from model profiles has been used as performance measure. In (Reiss et al., 2006) have implemented integrated biclustering and gene association networks to find overlapped clusters. Helicobacter pylori, Saccharomyces cerevisiae, and Escherichia coli data have been used with number of clusters k=10 to 300. Results indicated that largest 100 cluster with a volume-overlap (genes × conditions) of < 0.5, excluding any with > 200 genes.

Phylogenetic Analysis

(Zheng et al., 2005) perform operon prediction on a total of 345 operons with multiple genes extracted from the RegulonDB dataset for phylogenetic detection in conserved genes. This was a scalable framework for discovering conserved gene clusters.

Protein Analysis

(Kléma, et-al, 2006) have used GO ontologies on SAGE dataset and find mine plausible patterns to encode proteins that display an RNA binding activity.

Identify Co-expressed genes

In (Maulik et al., 2009), MOGA-SVM has been proposed and applied on microarray data sets, viz., Yeast Sporulation, Yeast Cell Cycle, Arabidopsis Thaliana, Human Fibroblasts Serum and Rat Central Nervous System with parameters generation=100; pop =50; cross probability=0.8; mutation probability=0.1 for identifying co-expressed genes. Silhouette index scores and p value has been used as performance measure. Results indicated that MOGA-SVM is able to produce biologically relevant and functionally enriched clusters.

Evolutionary Bi-clustering

In (Joung et al., 2012), PCOBA has been applied on synthetic dataset and yeast expression profiles with parameters δ = 20, mutation rate= 0.01, Pop. Size for genes =1000. Results were- Avg. Residue= 219.15, Avg. Variance= 412.11, Avg. Volume= 1369.18, Avg. Num. (Genes) = 98.54, Avg. Num. (Conditions) = 12.18. PCOBA outperform other evolutionary bi-clustering algorithms like GA, CGA and EDA but bi clusters found by this algorithm could not be larger than 200 size.

4. Strengths, Weaknesses and Gaps in Published Research

Some of the clustering techniques discussed in previous sections of the paper, are more advantageous than others. Like way, there are some techniques showing limitations for some dataset or in some empirical conditions. Under this heading, we have discussed such issues.

A. Strengths

- ECSAGO (Leon et-al, 2006) reduces the number of parameters used. ECSAGO does not require setting GA rates. ECSAGO Automatically adapts the GA operates values. ECSAGO determining which GA operator works better for solving the clustering problem. ECSAGO uses real genetic operators allow to get rid of mating restriction
- FAST (Song et al) obtains the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification accuracy
- NOCEA (Kelil et al., 2007) give effective treatment to high dimensionality. NOCEA provide precise discrimination of clusters
- Biclustering algorithms show 14% performance improvement (Parimala, 2013). Biclusters with scaling pattern are more biologically significant than the biclusters with shifting pattern. Advantage using bi-clustering in the analysis of gene expression data are similar genes may exhibit similar behaviors only under a subset of conditions, not all conditions In Bi-clustering genes may participate in more than one function, resulting in one regulation pattern in one context and a different pattern in another.
- Density based clustering algorithm retains the regulation information. No proximity measures are used with density based clustering. Density based clustering handle incremental data. Density based algorithms can work even with noise (Dharan et al, 2009).
- Automatic clustering (Kuo et al., 2013) using PSO did not require clusters apriori.
- Overlapping clusters can be generated using biologically supervised hierarchical clustering. Hierarchical sequence clustering can detect outliers. Query executed much quicker when only one comparison to an entire cluster has to be performed rather than one comparison per database sequence with large scale hierarchical clustering. With large scale hierarchical clustering there is possibility of analyzing evolutionary relationships among the sequences in a cluster. Hierarchical clustering tree based approach relying on one step of clustering may furnish a formally significant result while the overall experiment is not significant(Thom et-al, 2009).
- Symbol table method optimizes the performance of the clustering task (Baridam et al, 2013).
- IGKA algorithm has a convergence pattern similar to FGKA, it has a better time performance when the mutation probability decreases to some point (Lu et al., 2004).
- Overlapping clustering allows items to belong to multiple clusters (Sathishkumar, et-al, 2013), (Tritchler et al, 2009).
- The ensemble approach produces better biclusters than single approach. Ensemble attribute profile clustering is a simple method for collating and sythesizing the annotation data. (Kim et al, 2009), (Gao et al., 2013)
- Cluster Validity Framework could help in predicting the correct clusters and no. of clusters. (Okada, 2005), (Garge et-al, 2005), (Kléma, et-al, 2006).
- Isomap can successfully detect biologically relevant structures even in the background noise of tens of thousands of genes present (Tetko et. al., 2005).
- Phylogenetic assumption serves as a platform for many other functional genomic analyses in microorganisms, such as operon prediction, regulatory site prediction, functional annotation of genes, evolutionary origin and development of gene clusters. (Zheng et al., 2005)
- Matrix factorization method is able to identify complex relationships among genes and conditions that are difficult to identify by standard clustering algorithms (Zhang et al, 2010).
- Enhanced clustering algorithm did not depend upon the any choice of the number of cluster (Sakthi and Thanamani, 2013).
- cMonkey biclusters are more parsimonious (Carmona-Saez et-al, 2006).
- Bi-Range method has compact representation. Bi-

Range has accuracy and effectiveness with respect to noise handling (Kanungo et al, 2012).

- EXPANDER has very powerful analytical capabilities, built in support for genome-wide analysis (Shamir et. al., 2005).
- FLAME is easily extendable; FLAME clusters are more internally homogeneous and more diverse from each other, and provide better partitioning of biological functions (Fu et al, 2007).
- VISDA has complementary building blocks and has flexible functionality (Zhu et. al., 2008).
- CLICK is fast, allowing high accuracy clustering of large data sets of size over 100 000. CLICK performs better in terms of average homogeneity (Saharan et al, 2003).
- The detection of super-bi-clusters and hub nodes is easy and useful with BicOverlapper (Santamaría et al, 2008).
- Large Scale clustering has reduced the calculation cost, and is an appropriate solution for analyzing large-scale TSS usage data (Shimokawa, et-al, 2007).
- Response Projected clustering can identify many known and novel gene factors(Yi et-al, 2008)
- Imputation method is useful for ignoring the missing values or replacing them with zeros or average values(Tuikkala et al, 2008)
- CARPENTER is especially designed to mine frequent closed associated patterns in database Containing large number of columns and small number of rows(Mishra et al, 2011)
- iBBiG show robustness in the presence of noise and its tolerance of missing data, does not require a gene set to be associated with all phenotypes in a bi-cluster(Guselenitner et al, 2012)
- Maximum similarity bi-cluster require no discretization procedure. Maximum similarity bicluster performs well for overlapping bi-clusters and works well for additive bi-clusters(Liu et al, 2007)
- Two way clustering method could discover more relevant genes comparing to one-way clustering methods(Rathipriya et-al, 2011)
- The major advantage with incremental clustering algorithms is that it is not necessary to store the entire dataset in the memory. Therefore, the space and time requirements of incremental algorithms are very small.
- The reference gene method (RMSBE) do not need to mask discovered bi-clusters. RMSBE require no discretization procedure.(Liu et al, 2007) .

B. Weaknesses

- FAST does not limit to some specific types of data. FAST algorithm employs clustering based method to choose features (Song et-al, 2013).
- Bio inspired meta-heuristics need to know the number of clusters and the experiment were not performed using a fixed computational cost (Colanzi et al., 2011).
- EPSO outperformed the other methods in terms of the LOOCV accuracy (Kuo et-al., 2013).
- The ensemble bi-clustering takes much more times than single biclustering.(Semeiks, et-al, 2006)
- A problem with the Rand index is that its expected value of two random clustering does not take a constant value (such as zero). Adjusted Rand index overcomes this disadvantage.
- In text mining approach, when applied to unknown genes clusters, the rules must be analyzed by an expert.(Medeiros et-al, 2005)
- It is very difficult to derive the threshold value, which is dataset dependant with IGKA. (Lu et. al., 2004)
- Fuzzy clustering cannot find co-expressed genes under all situations(Shakti et al, 2013)
- EXPANDER has limited visualization capabilities(Sharan et. al)
- caBIG VISDA require minor algorithm modifications required for gene, phenotype and sample clustering (Zhu et al, 2008)
- iBBiG lacked some sensitivity (56%)(Nilamadhab Mishra, 2011)
- Hierarchical clustering tree based approach did not facilitate the exploratory analysis of big data sets Misinterpretations (Shimokawa, et-al, 2007)
- The main disadvantages of the hierarchical methods are:
- Inability to scale well.
- The time complexity of hierarchical algorithms is at

least $O(m2)$ (where m is the total number of instances), which is non-linear with the number of objects.

- Clustering a large number of objects using a hierarchical algorithm is also characterized by huge I/O costs.
- Hierarchical methods can never undo what was done previously. There is no back-tracking capability.
- *K*-means algorithm is sensitive to initial seed selection and even in the best case, it can produce only hyper spherical clusters. For K-means, the clustering is usually unidirectional, which means the cluster of a point is determined by the distance to the cluster centre. It might be very sensitive to the outliers in the data. Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome) for k-means. Fixed number of clusters can make it difficult to predict what K should be for k-means. it Does not work well with non-globular clusters for k-means
- Maximum cut algorithm is very slow, and potentially must be run many times
- Phylogenetic assumption method made several simplifying assumptions become a problem for those "hitchhiking" genes that may be transferred with the cluster(Zheng et al., 2005)
- The concept of distance between points in high dimensional spaces is ill-defined for partition algorithm (Kim, et-al, 2006). Partition algorithm is poor cluster descriptor. Partition algorithm has reliance on the user to specify the number of clusters in advance. Partition algorithm has high sensitivity to initialization phase, noise and outliers. Partition algorithm frequently entrapments into local optima. Partition algorithm has inability to deal with non-convex clusters of varying size and density. Partition algorithm has inability to make corrections once the splitting/merging decision is made.
- Lack of interpretability regarding the cluster descriptors.
- Vagueness of termination criterion.
- Prohibitively expensive for high dimensional and massive datasets.
- Severe effectiveness degradation in high dimensional spaces due to the curse of dimensionality phenomenon.
- Density-based clustering has high sensitivity to the setting of input parameters. Density-based clustering has poor cluster descriptors. Density-based clustering is unsuitable for high-dimensional datasets because of the curse of dimensionality phenomenon. (Campello et al, 2012)
- OPSM algorithm takes long time for high dimensional datasets.

C. Gaps in Published Research

The limitations of various methodologies used in the reviewed paper are pointing towards the gaps in the process/techniques. Following gaps have been found in the reviewed papers-

- Some good algorithms such as FAST, are useful only with a particular type. These algorithms cannot be used for other dataset. A further research should be done to extend the capability of such algorithm to other datasets also.
- Most of the clustering algorithms require number of clusters in advanced. A research is required to develop such algorithms without the need of clusters apriori.
- A number of algorithms outperformed the other techniques in terms of a particular performance measure only. For example EPSO outperformed the other methods in terms of the LOOCV accuracy only. Other than LOOCV accuracy, its performance is not better than others. A field of research is open here to find out one such performance measure which can be used universally in further research.
- Very few algorithms can work on non-globular clusters. A research should be done to find out non-globular clusters
- Ensemble bi-clustering is useful for gene expression analysis but its time complexity is very much higher than other single bi-clustering.
- In some cases few points that form a bridge between two clusters causes the single-link clustering to unify these two clusters into one. A further research should be done to cope up with this problem.
- Stability based method is useful for gene expression data only. A further research should be done to extend this method to other type of data also.

- The threshold value in IGKA is dependent on data sets. A research should be done to overcome this. The threshold value should be chosen independent of datasets.
- Evolutionary algorithms can be useful for clustering and bi clustering purpose. But still there is not much research work has been done with evolutionary algorithm. Out of 150 reviewed papers only 8 papers used evolutionary algorithm. A further research should be done to apply evolutionary algorithms.
- EXPANDER has limited visualization capabilities. Good visualization is very much required to show gene-gene interaction or functional analysis of various genes. A further research should be done to improve the visualization capabilities of such type of algorithms and tools.
- Misinterpretations of data by bi-algorithms should be handled by further research.
- *K*-means algorithm is sensitive to initial seed selection and even in the best case. Controlling the sensitivity of clustering algorithms towards the initial seed values should be incorporated in further research.

Conclusion

With the increase in availability of biological data and emergence of techniques to automate the wet-lab experiments, there is a need to analyze biological data using computational techniques. This paper is an attempt to survey various problems and solution approaches for biological sequence analysis. There are several different methods to solve sequence analysis problems but this paper is emphasizing on Clustering techniques for solving such problems. (Cheng and Church, 2000) have proposed bi-clustering, another approach for clustering which has been appropriate for microarray datasets. Present communication has consider both of these approaches and tried to find the problems and their solutions with respective to both clustering and bi-clustering techniques.

References

Ayadi, W. (2012). Pattern-driven Neighbourhood Search for Bi-clustering of Microarray Data. BMC Bioinformatics. Retrieved from http://www.biomed-central.com/ 1471-2105/13/S7/S11

Bagyamani, J. (2014). Biological significance of gene expression data using similarity based bi-clustering algorithm. *International Journal of Biometrics and Bioinformatics*, 4(6), 201-216.

Bandyopadhyay, S. (2007). An improved algorithm for clustering gene expression data. *Bioinformatics,* 23(21), 2859-2865.

Baridam, B. (2013). *Alternate Biological Sequence Clustering using Symbol Table.* Science and Information Conference.

Boratyn, G. M. (2006). *Biologically Supervised Hierarchical Clustering Algorithms for Gene Expression Data*. Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, pp. 5515

Brevern, D. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*.

Campello, R. J. G. B. (2012). *A Simpler and More Accurate AUTO-HDS Framework for Clustering and Visualization of Biological Data*. IEEE/ ACM Transactions on Computational Biology and Bioinformatics, 6(6), 1850-1852.

Cheng, Y., & Church, G. M. (2000). *Bi-clustering of Expression Data. Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology,* pp. 93-103.

Chezhian, V. (2011). *Hierarchical Sequence Clustering Algorithm for Data Mining.* Proceedings of the World Congress on Engineering.

Corchado, E. (2006). *Neighbourhood-based Clustering of Gene-Gene Interactions*. LNCS 4224, (pp. 1111-1120), Springer-Verlag Berlin Heidelberg.

Das, R. (2010). Clustering gene expression data using an effective dissimilarity measure. *International Journal of Computational Bioscience*, 1(1), 55-68.

Das, R. (2014). Bi-clustering of gene expression data using a two-phase method. *International Journal of Computer Applications,* 103(13), 6-10.

Dawson, K. (2005). Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, A nonlinear algorithm. *BMC Bioinformatics*, 6, 1-17.

Dettling, M. (2002). Supervised clustering of genes. *Genome Biology*, 3(12), 1-15.

Dib, L., & Carbone, A. (2012). CLAG: An unsupervised non hierarchical clustering algorithm handling biological data. BMC Bioinformatics. Retrieved from http://www.biomedcentral.com/1471-2105/13/194.

Dresen, G. (2008). New re-sampling method for evaluating stability of clusters. *BMC Bioinformatics*.

Dutta, D. (2012). *Data Clustering with Mixed Features by Multi Objective Genetic Algorithm*. 12th International Conference on Hybrid Intelligent Systems (HIS), (pp. 336-341).

Datta, D., & Dutta, P. (2006). Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, 7(4), S4-S17.

Freudenberg, J. M. (2009). CLEAN: Clustering enrichment analysis. *BMC Bioinformatics.*

Fu, L. (2007). FLAME- A novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics*, pp-1-15.

Gao, C. (2013). Rough subspace-based clustering ensemble for categorical data. *Soft Computing*, 17(9), 1643-1658.

Geng, H. (2004). *A new approach to clustering biological data using message passing.* Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004).

Getz, G. (2000). *Coupled Two-way Clustering Analysis of Gene Microarray Data*. Proceedings of National Academy of Science, 97(22), 12079-12084.

Gusenleitner, D. (2012). iBBiG: Iterative binary bi-clustering of gene sets. *Bioinformatics*, 28(19), 2484-2492. Oxford University Press

Hanczar, B. (2011). *Improving the biological relevance of bi-clustering for microarray data in using ensemble methods*. 22nd International Workshop on Database and Expert Systems Applications, (pp. 413-417).

Hartsperger, M. L. (2010). Structuring heterogeneous biological information using fuzzy clustering of k-partite graphs. *BMC Bioinformatics*, 11(522), 1471-2105.

Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information Sciences*, 222, 175-184.

Henriques, R. (2014). BicSPAM: Flexible bi-clustering using sequential Patterns. *BMC Bioinformatics,* 2014, 15, 130.

Huang, C. H. (2002). *Parallel Pattern Identification in Biological Sequences on Clusters*. Proceedings of the IEEE International Conference on Cluster Computing, (pp. 1- 8).

Jain, A. (1999). Data Clustering: A Review. *ACM Computing Surveys*, September, 31(3), 264-323.

Jaskowiak, P. A. (2014). *On the Selection of Appropriate Distances for Gene Expression Data Clustering*. 12th Asia Pacific Bioinformatics Conference, (pp. 17-19).

Jonnalagadda, S. (2009). NIFTI: An evolutionary approach for finding number of clusters in microarray data. *BMC Bioinformatics*.

Joung, J. G. (2012). A probabilistic co-evolutionary biclustering algorithm for discovering coherent patterns in gene expression dataset. *BMC Bioinformatics*, 13(17), S12. Retrieved from http://www.biomedcentral.com/1471-2105/13/S17/S12

Kelil, A. (2007). CLUSS: Clustering of protein sequences based on a new similarity Measure. *BMC Bioinformatics*, (pp. 1-19).

Kim, S. Y. (2006). Effect of data normalization on fuzzy clustering of DNA microarray Data. *BMC Bioinformatics*, (pp-1-14).

Kim, E. Y. (2009). MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering.

Kléma, J. (2006). *Mining Plausible Patterns from Genomic Data*. 19th IEEE Symposium on Computer-based Medical Systems.

Kraus, J. M. & Kestler, H. A. (2010). A highly efficient multi-core algorithm for clustering extremely large datasets. *BMC Bioinformatics.*

Krause, J. M. (2005). Large Scale Hierarchical Clustering of Protein Sequences. *BMC Bioinformatics*, (pp.1-12).

Krishna, R. (2010). A temporal precedence based clustering method for gene expression microarray data. *BMC Bioinformatics*, 11-68

Kuo, R. J. (2013). Automatic Clustering using an improved particle swarm optimization. J*ournal of Industrial and Intelligent Information*, 1(1), pp- 46-51.

Lee, A. J. T. (2005). *Cluster Utility: A New Metric for Clustering Biological Sequences*. Proceedings of the IEEE Computational Systems Bioinformatics Conference Workshops (CSBW'05).

Leon, E. (2006). *ECSAGO: Evolutionary Clustering using Self Adaptive Genetic Operators.* IEEE Conference on Evolutionary Computation, pp. 1768.

Levenstien, M. A. (2003). Statistical significance for hierarchical clustering in genetic association and microarray expression studies. *BMC Bioinformatics*, (pp-1-9).

Li, X. L. (2006). Systematic gene function prediction from gene expression data by using a fuzzy nearest-cluster method. *BMC Bioinformatics*, 7(4), S23

Li, X. L. (2009). QUBIC: A qualitative bi-clustering algorithm for analyses of gene expression data. *Nucleic Acids Research, The Authors*, 37, 1-10.

Li, X. L. (2012). *A comparison and evaluation of five bi-clustering algorithms by quantifying goodness of bi-clusters for gene expression data*. Bio Data Mining, (pp-15).

Liu, X. (2007). Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics*, 23(1), 50-56.

Lock, E. F. (2010). Clustering for low-density microarrays and its application to QPCR. *BMC Bioinformatics*, 11(386), 1471-2105.

Lu, L. (2004). Incremental genetic K-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics*, pp-1-10.

Ma, P. C., & Chan, K. C. (2009). *An Iterative Data Mining Approach for Mining Overlapping Co-expression Patterns in Noisy Gene Expression Data*. IEEE Transactions on Nano-bioscience, 8(3), 252.

Marttinen, P. (2009). Bayesian clustering and feature selection for cancer tissue samples. *BMC Bioinformatics*.

Maulik, U. (2009). Combining Pareto-optimal clusters using supervised learning for identifying co-expressed genes. BMC Bioinformatics.

Medeiros, D. M. R. (2005). *Applying Text Mining and Machine Learning Techniques to Gene Clusters Analysis*. 6th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05), (pp. 1-6).

Mohamad, M. S. (2013). An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes, *Algorithms for Molecular Biology*, 8(1), 1-15.

Nesamalar, E. K. (2012). *Genetic Clustering with Bee Colony Optimization for Flexible Protein-ligand Docking*. International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), pp. 82-87.

Newman, A. M. & Cooper, J. (2010). Auto SOME: A clustering method for identifying gene expression modules without prior knowledge of cluster number", BMC Bioinformatics, 11:117

Mishra, N. (2011). A framework for associated pattern mining over microarray database. *Journal of Global Research in Computer Science*, February, 2(2), 8-11.

Oghabian, A. (2014). Bi-clustering methods: Biological relevance and application in gene expression analysis. *PLoS One*, 9(3), 1-10.

Okada, Y. (2005). Detection of Cluster Boundary in Microarray Data by Reference to MIPS Functional Catalogue Database.

Preli´c, A. (2006). A systematic comparison and evaluation of bi-clustering methods for gene expression data. *Bioinformatics Advance Access*. Oxford University Press.

Quest, D. & Ali, H. (2005). *A Grammar Based Approach for Mining Bioinformatics Databases*. Proceedings of the 38th Hawaii International Conference on System Sciences.

Dharan, S. (2009). *Randomized adaptive search procedure*. BMC Bioinformatics.

Rathipriya, R. (2011). Evolutionary Bi-clustering of Click-stream Data. *International Journal of Computer Science Issues*, 8(3), 341-347.

Reiss, D. J. (2006). Integrated bi-clustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 22(9), 1-22.

Robinson, M. D. (2002). FunSpec: A web-based cluster interpreter for yeast. *BMC Bioinformatics*, 1-5.

Roy, S. (2013). *Reconstruction of Gene Co-expression Network from Microarray Data using Local Expression Patterns.* 10th Annual Biotechnology and Bioinformatics Symposium (BIOT 2013) Provo, UT, USA.

Sakthi, M. & Thanamani, A. S. (2013). An enhanced K means clustering using improved HOP field artificial neural network and genetic algorithm. *International Journal of Recent Technology and Engineering*, 2(3), 16-21.

Santamaría, R. (2008). A visual analytics approach for understanding bi-clustering results from microarray data. *BMC Bioinformatics*, 1-19.

Sarmah, S. & Bhattacharyya, D. K. (2010). An effective technique for clustering incremental gene expression data. *International Journal of Computer Science Issues*, 7(3), 31-41.

Sathishkumar, K. (2008). Identification of Bi-clustering Algorithms for Gene Extraction. *International Journal of Engineering Sciences & Research Technology*, 2(10), 13.

Scharl, T. (2009). *Exploratory and Inferential Analysis of Gene Cluster Neighbourhood Graphs*. BMC Bioinformatics.

Semeiks, J. R. (2006). Ensemble attribute profile clustering: discovering and characterizing groups of genes with similar patterns of biological features. *BMC Bioinformatics*, (pp. 1-11).

Shamir, R. (2005). EXPANDER - An integrative program suite for microarray data analysis. *BMC Bioinformatics*, (pp. 1-12).

Sharan, R. (2003). CLICK and EXPANDER: A system for clustering and visualizing gene expression data. *Bioinformatics*, 19(14), 1787-1799. Oxford University Press,

Shimokawa, K. (2007). Large-scale clustering of CAGE tag expression data. *BMC Bioinformatics*.

Shon, H. S. (2008). *Clustering Microarray Data by Using a Stochastic Algorithm*. IEEE 8th International Conference on Computer and Information Technology Workshops, (pp. 456-461).

Simpson, T. I. (2010). Merged consensus clustering to assess and improve class discovery with microarray data. *BMC Bioinformatics*, 11,590.

Smolkin, M. (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 1-7.

Song, Q. (2013). *A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data*. IEEE Transactions on Knowledge and Data Engineering, 25(1).

Swathi, H. (2010). Gene expression data knowledge discovery using global and local clustering. *Journal of Computing*, 2(3), 116-133.

Tetko, I. V. (2005). Super paramagnetic clustering of protein sequences. *BMC Bioinformatics*, 1-13.

Thomas, J. (2011). A novel fuzzy clustering method for outlier detection in data mining. *International Journal of Recent Trends in Engineering*, 1(2), 161-165.

Tjaden, B. (2006). An approach for clustering gene expression data with error Information. *BMC Bioinformatics*, 1-15.

Toronen, P. (2004). Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics*, 5(32), 1-19.

Tritchler, D. (2009). Filtering genes for cluster and network analysis. *BMC Bioinformatics*.

Verbanck, M. (2013). A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, 14(1), 42.

Wang, J. (2002). Clustering of the SOM easily reveals distinct gene expression patterns: Results of a reanalysis of lymphoma study. *BMC Bioinformatics*, 1-9.

Wang, J. (2003). Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinformatics*, 1-12.

Xu, T. (2009). Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology. *BMC Bioinformatics*.

Yao, J. (2008). Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. *BMC Bioinformatics*.

Yi, S. G. (2008). Response projected clustering for direct association with physiological and clinical response data. *BMC Bioinformatics*, 9, 76.

Yin, L. (2006). Clustering of gene expression data: Performance and similarity analysis. *BMC Bioinformatics*, 7(4), S4-S19.

Zhang, Z. Y. (2010). Binary matrix factorization for analyzing gene expression data. *Data Mining and Knowledge Discovery*, 28-52.

Zheng, Y. (2005). Phylo-genetic detection of conserved gene clusters in microbial Genomes. *BMC Bioinformatics*, 1-14.

Zhu, Y. (2008). caBIG™ VISDA: Modelling, visualization, and discovery for cluster analysis of genomic data. *BMC Bioinformatics*.