Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge

Sushmita Mitra, *Fellow, IEEE*, and Sampreeti Ghosh

***Abstract*—In this paper, a novel feature selection algorithm, which is governed by biological knowledge, is developed. Gene expression data being high dimensional and redundant, dimensionality reduction is of prime concern. We employ the algorithm clustering large applications based on RAN-domized search (CLARANS) for attribute clustering and dimensionality reduction based on gene ontology (GO) study. Feature selection with unsupervised learning is a difficult problem, with neither class labels present nor any guidance available to the search. Determination of the optimal number of clusters is another major issue, and has an impact on the resulting output. The use of GO analysis helps in the automated selection of biologically meaningful partitions. Tools such as Eisen plot and cluster profiles of these clusters help establish their coherence. Important representative features (or genes) are extracted from each correlated set of genes in such partitions. The algorithm is implemented on high-dimensional Yeast cell-cycle, Human Multiple Tissues, and Leukemia microarray data. In the second pass, clustering on the reduced gene space validates preservation of the inherent behavior of the original high-dimensional expression profiles. While the reduced gene set forms a biologically meaningful gene space, it simultaneously leads to a decrease in computational burden. External validation of the reduced subspace, using various well-known classifiers, establishes the effectiveness of the proposed methodology.**

***Index Terms*—Attribute clustering, clustering large applications based on RAN-domized search (CLARANS), feature selection, gene ontology (GO) medoid.**

I. Introduction

ONE of the important problems in extracting and analyzing information from large databases is the associated high complexity. Feature selection is helpful as a preprocessing step to reduce dimensionality, remove irrelevant data, improve resultant learning accuracy, and enhance output comprehensibility. Unlike other dimensionality reduction methods (involving feature extraction), feature selection preserves a subset of the original features. Search is a key issue in feature selection, involving search starting point, search direction, and search strategy. One also needs to measure the goodness of the generated feature subset. Feature selection can be supervised as well as unsupervised, depending on the class information availability in data. The algorithms are typically categorized under filter and wrapper models [1], with different emphasis on dimensionality reduction or accuracy enhancement.

Feature selection has been widely used in the supervised framework to achieve better generalization on unseen data [2], [3]. Genetic search [4] and boosting [5] have also been used for efficient feature selection. However, there has been comparatively less attention toward feature selection with unsupervised learning [6], [7]. The main reasons include 1) a lack of understanding about assessing the relevance of a subset of features, without resorting to class labels; and 2) the choice of the optimal number of clusters, which can in turn have a major impact on the partitioning and feature selection.

Gene expression data being typically high dimensional, they require appropriate data mining strategies such as clustering, feature selection, and biclustering for further analysis [8]–[10]. Clustering is prevalent in any discipline that involves analysis of multivariate data. Clustering large applications based on RANdomized search (CLARANS) [11] efficiently partitions large data in terms of medoid-based partitive clustering. It is found to outperform algorithms such as partitioning around medoids (PAM) and clustering large applications (CLARA) in terms of accuracy and computational complexity, and can handle outliers.

Biological knowledge about coexpressed genes has also been incorporated in clustering [12], [13], to determine quality-based partitions. Gene ontology (GO) annotations have been used to extend the k-medoids algorithm such that genes with known function get clustered together [14]. The use of GO in hierarchical clustering is reported in [15]. Fuzzy c-means clustering has been enhanced with GO annotations as prior knowledge while guiding the process of grouping functionally related genes in terms of fuzzy membership evaluation [16]. The incorporation of biological knowledge provides a direction toward the extraction of meaningful groups of genes [17], [18]. Computation of pairwise distances between gene annotation similarities has been used [19] to develop a fast software.

In this paper, we focus on feature selection from microarray data by attribute (or gene) clustering while utilizing the biological relevance of these genes. We employ CLARANS to cluster the attributes, and select the representative medoids (or genes) from each biologically enriched cluster. For this purpose, we compute the biological relevance of these clusters in terms of the statistically significant GO annotation database. The clusters are qualitatively evaluated in terms of Eisen plot and gene profiles. The reduced gene space leads to a decrease in computational burden. In the second stage, CLARANS is used to cluster the expression profiles to evaluate the significance of the reduced subset of genes. External validation of the clustering

The authors are with Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: sushmita@isical.ac.in; sampreeti_t@isical.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

is done in terms of classification accuracy (%) with standard classifiers [20]. We used the publicly available WEKA [21] and DTREG [22] software implementations of k-nearest neighbors (k-NN), decision tree C4.5, random forest, multilayer perceptron (MLP), support vector machine (SVM), and naive Bayes (NB). It is observed that the performance is, in general, better in the reduced subspace. This establishes the effectiveness of feature selection algorithm using biological knowledge.

The rest of this paper is organized as follows. Section II describes algorithm CLARANS. The proposed algorithm for feature selection, incorporating biological knowledge in terms of GO study, is described in Section III. It results in the automated extraction of biologically meaningful subspaces. The experimental results on *Yeast* cell-cycle, *Multiple Tissues,* and *Leukemia* data, along with the validation, are presented in Section IV. Finally, Section V concludes this paper.

II. Clustering Large Applications Based on RAN-Domized Search

Large datasets require the application of scalable algorithms. CLARANS [11] draws a sample of the large data, with some randomness, at each stage of the search. Each cluster is represented by its medoid. Multiple scans of the database are required by the algorithm. Here, the clustering process searches through a graph G, where node v^j is represented by a set of c medoids (or centroids) $\{m_1^j, \ldots, m_c^j\}$. Two nodes are termed as neighbors if they differ by only one medoid. More formally, two nodes $v^1 = \{m_1^1, \ldots, m_c^1\}$ and $v^2 = \{m_1^2, \ldots, m_c^2\}$ are termed neighbors if and only if the cardinality of the intersection of v^1 and v^2 is given as $\text{card}\,(v^1 \bigcap v^2) = c - 1$. Hence, each node in the graph has $c * (N - c)$ neighbors. For each node v^j, we assign a cost function

$$J_c^j = \sum_{x_k \varepsilon U_i} \sum_{i=1}^{c} d_{ki}^j \tag{1}$$

where d_{ki}^j denotes the dissimilarity measure (Euclidean distance) of the kth object x_k from the ith cluster medoid m_i^j in the jth node. The aim is to determine that set of c-medoids $\{m_1^0, \ldots, m_c^0\}$ at node v^0, for which the corresponding cost is the minimum as compared with all other nodes in the tree.

The algorithm considers two parameters $numlocal$, representing the number of iterations (or runs) for the algorithm, and $maxneighbor$, the number of adjacent nodes (set of medoids) in the graph G that need to be searched up to convergence. These parameters are provided as input at the beginning. The main steps, thereafter, are outlined as follows.

1) Set iteration counter $l \leftarrow 1$, and set the minimum cost $mincost$ to an arbitrarily large value. A pointer $bestnode$ refers to the solution set.
2) Start randomly from any node $v^{current}$ in graph G, consisting of c medoids. Compute cost $J_c^{current}$ by (1).
3) Set node counter $j \leftarrow 1$.
4) Select randomly a neighbor v^j of node $v^{current}$. Compute the cost J_c^j by (1).
5) **If** the criterion function improves as $J_c^j < J_c^{current}$
   Then set the current node to be this neighbor node by $current \leftarrow j$, and **go to Step 3** to search among the neighbors of the new $v^{current}$
   Else increment j by 1.
6) **If** $j \leq maxneighbor$
   Then go to Step 4 to search among the remaining allowed neighbors of $v^{current}$
   Else calculate the average distance of patterns from medoids for this node; this requires one scan of the database.
7) **If** $J_c^{current} < mincost$
   Then set $mincost \leftarrow J_c^{current}$ and choose as a solution this set of medoids given by $bestnode \leftarrow current$.
8) Increment the number of iterations l by 1.
   If $l > numlocal$
   Then output $bestnode$ as the solution set of medoids and halt
   Else go to Step 2 for the next iteration.

Note that *maxneighbor* can be computed as

$$maxneighbor = p\% \text{ of } \{c * (N - c)\} \tag{2}$$

with p being provided as input by the user. Typically, $1.25 \leq p \leq 1.5$ [11].

The clustering algorithm described here is partitive, requiring prespecification of the number of clusters. The result is, therefore, dependent on the choice of c. There exist validity indices to evaluate the goodness of clustering, corresponding to a given value of c. A commonly used measure is the Davies–Bouldin (DB) index [23].

III. Feature Selection and Clustering Large Applications Based on RAN-Domized Search

We employ attribute clustering, in terms of CLARANS, for feature selection. This results in dimensionality reduction, with particular emphasis on high-dimensional gene expression data, thereby helping one to focus the search for meaningful partitions within a reduced attribute space. While most clustering algorithms require user-specified input parameters, it is often difficult for biologists to manually determine suitable values for these. The use of clustering validity indices for an automated determination of optimal clustering has been reported in the literature [24], [25]. In this paper, we incorporate biological knowledge, in terms of GO, along with the DB index, to automatically extract the biologically relevant cluster prototypes.

A. Gene Ontology

We determined the biological relevance of the gene clusters for the *Yeast cell-cycle*[1] and *Human*[2] (*Multiple Tissues, Leukemia*) data, in terms of the statistically significant GO annotation database. Here, genes are assigned to three structured, controlled vocabularies (ontologies) that describe gene products in terms of associated biological processes (BP), cellular

components (CC), and molecular functions (MF) in a species-independent manner. Such incorporation of knowledge enables the selection of biologically meaningful groups, consisting of biologically similar genes.

We have measured the degree of enrichment, i.e., p-values[3] using a cumulative hypergeometric distribution, which involves the probability to observe the number of genes from a particular functional GO category (i.e., MF, BP, and CC) within each feature (or gene) subset. The probability p to find at least k genes, from a particular category within a cluster of size n, is expressed as

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}} \tag{3}$$

where f is the total number of genes within a category, and g is the total number of genes within the genome [26]. The p-values are calculated for each functional category in each cluster. Statistical significance is evaluated for the genes in each of these partitions by computing p-values that signify how well they match with the different GO categories. Note that a smaller p-value, close to zero, is indicative of a better match. A p-value closer to zero indicates high confidence that the associated category represents the correct annotation for a cluster, whereas a value closer to 1 implies low confidence in the annotation. Our algorithm uses the p-value as a selection criterion to extract "biologically meaningful" clusters of genes from a large collection.

B. Algorithm

The proposed algorithm focused on the feature selection from microarray data by attribute clustering, while utilizing the biological relevance of the genes. The main objective is to demonstrate the utility of incorporating biological knowledge in feature selection procedure for improved classification performance as well as reduced computational complexity. The concept to select the medoid (the most representative gene) of each biologically enriched gene cluster to constitute a reduced set of features, on the basis of GO-based biological knowledge, is new.

In the first-pass clustering toward feature selection, the conventional algorithm CLARANS is employed as a tool to handle the large gene space. The GO study has been used to find significant clusters that are generated in terms of enrichment of the functional categories in individual clusters. The p-value has been used to filter out the significant clusters. Clusters with at least one significantly enriched GO term are considered as biologically significant clusters. The percentages of GO knowledge have also been incorporated to evaluate the biological relevance of the clusters. Medoids are selected only from such biologically significant clusters. These are subsequently used as the most representative genes for the reduced feature set. Thus, the first-pass clustering involves attribute clustering for feature selection using biological knowledge. The second-pass clustering on the reduced gene or feature set, over the time points or samples, mainly validates the claim for preservation of the class information over the reduced gene space. Thus, working on this reduced set one can attain improved visualization and better interpretability of the results.

First, we perform clustering for feature selection, with $c = \sqrt{g}$ [27]. The prototype (medoid) of each biologically "good" gene cluster (measured in terms of GO) is selected as the representative gene (feature) for that cluster, and the remaining genes in that cluster are eliminated. Thereafter, the remaining set of genes (in the "not-so-good" clusters) are again partitioned with CLARANS, for $c = c_0$ which minimizes the validity index. Finally, the goodness of the generated partitions is biologically evaluated in terms of GO, and the representative genes are selected.

Upon completion of gene selection, the gene expression dataset is transposed and reclustered over the reduced gene space. The cluster validity index is used to evaluate the generated partitions. The gene expression patterns are studied to biologically justify the generated partitions. This leads to dimensionality reduction, followed by partitioning into biological relevant subspaces. The steps of the algorithm are outlined next.

1) Initialize $g \leftarrow$ number of genes, $N \leftarrow$ number of samples
   Initialize number of medoids $n_m \leftarrow 0$.
2) Transpose the gene expression array.
3) Cluster set of genes using CLARANS for $c = \sqrt{g}$.
4) Use GO to detect coregulated genes in terms of process-, component-, and function-related p-values $<= e^{-05}$.
5) **If** any biologically meaningful cluster is detected
   Then perform **Step 6** for each such cluster
   Else go to Step 8.
6) Replace each set of coregulated genes g_c by its medoid, increment n_m, and decrement $g \leftarrow g - g_c$.
7) **Repeat Steps 3–6 until** no more good clusters can be found.
8) a) Cluster the remaining set of genes g with CLARANS while minimizing validity index DB for $c = c_0$ over $c = 2, \ldots, \sqrt{g}$.
   b) Test the p-values for these c_0 clusters.
   c) Compress each of the biologically meaningful clusters by its medoid, such that $g \leftarrow g - g_c$ and $n_m \leftarrow n_m + 1$.
9) Biologically validate the selected genes (medoids of enriched clusters) using publicly available databases and annotations. Visualize the selected clusters using relevant tools such as Eisen plot and cluster profile.
10) Retranspose the gene expression array to cluster the gene expression data in the reduced space of g genes corresponding to n_m medoids.
11) Use cluster validity index to evaluate optimal partition.
12) Validate the generated subspaces qualitatively, in terms of the original gene expression data.

The grouping of genes, which is based on GO analysis, helps to capture different aspects of gene association patterns in terms of associated BP, CC, and MF. The mean of a cluster (which need not coincide with any gene) is replaced by the medoid (or most representative gene), and deemed significant in terms of

[3]The p-value of a statistical significance test represents the probability to obtain values of the test statistic that are equal to or greater in magnitude than the observed test statistic.

TABLE I
MICROARRAY DATASETS

Data	# Clusters for minimum index	No. of genes	
		original	reduced
Yeast	2	2884	15
Multiple Tissues	4	5565	42
Leukemia	2	7129	44

ontology study. The set of medoids, selected from the partitions, contain useful information for subsequent processing of the profiles. The smaller number of such significant genes leads to a reduction in the search space as well as improved performance for clustering.

IV. EXPERIMENTAL RESULTS

We have implemented the proposed feature selection algorithm on microarray data consisting of three benchmark gene expression datasets for 1) *Yeast*[4] [26]; 2) *Multiple Tissues*[5] [28]; and 3) *Leukemia*.[6] External validation was done using the publicly available WEKA[7] and DTREG[8] software packages to evaluate the qualitative subspaces.

Yeast data are a collection of 2884 genes (attributes) under 17 conditions (time points), having 34 null entries with -1 indicating the missing values. These are categorized as two broad phases, each involving G1, S, G2, and M subphases [29]. All entries are integers lying in the range of 0 to 600. The missing values are replaced by random numbers between 0 to 800. *Multiple Tissues* data comprise 103 samples with 5565 genes from four normal tissue types of humans, viz., breast, prostate, lung, and colon. The *Leukemia* dataset is a collection of 7129 gene expression measurements (attributes) of 72 leukemia samples, from high density oligonucleotide microarrays, belonging to two types of the disease, viz., acute lymphoblastic leukemia (ALL) and acute myeloblastic leukemia (AML).

Table I presents a summary of the feature selection process for the three datasets. The optimal numbers of partitions, in terms of DB index, are indicated in column 2 of the table.

Tables II, V, and VII illustrate the first-pass clustering, for the three datasets, respectively, where genes with similar expression patterns are grouped together across all the conditions. The selected clusters are marked in bold.

The GO study has been employed to find significant clusters that are generated in terms of enrichment of functional categories within the individual clusters. Thus, to find biologically enriched clusters, the genes in each cluster are mapped to the functional categories that are present in Saccharomyces genome database (SGD), Munich Information Center for Protein Sequences (MIPS), and FatiGO genome databases. Tables III, VI, and VIII demonstrate the biological validation of sample clusters that are selected from the respective datasets. Table III summarizes the biological validation results for some of the selected clusters of Table II in terms of enrichment of functional categories in the MIPS comprehensive yeast genome database[9] [30]. It was determined, from the SGD[10] [31], that the selected open reading frame (ORF) in the first column of the tables have more meaningful biological annotation in terms of their belonging to a larger number of GO categories (viz., MF, BP, and CC). Tables VI and VIII present the biological validation of sample clusters of *Multiple Tissues* and *Leukemia* datasets using FatiGO[11] [32] which is a functional enrichment tool in Babelomics. Here also the selected clusters (and their medoids in the first column) were found to have more biological significance, as estimated from the rat genome database[12] [33], in terms of more meaningful annotations with reference to a larger number of GO categories. We then referred to the *Universal Protein Resource* (Uniprot)[13] database for cross-validating the significance of the selected gene subset.

Sample clusters are pictorially depicted in Figs. 1, 2, 4, and 5 in terms of Eisen plot and gene profiles of *Yeast* and *Multiple Tissue* datasets. The dots in the cluster profiles correspond to the mean expression values of genes along different time points.

A. Yeast

The ORFs[14] of the genes (or medoids) corresponding to the attribute clusters that are selected (with cardinality marked in bold) in Table II, by Steps 3–6 of the algorithm for *Yeast*, are as follows:

YDR165W, YDL164C, YDR385W, YKL190W, YGR092W, YMR076C, YER018C, YOR234C, YGR152C, YLR325C, YFL008W.

We were able to determine the role of the clusters that are presented in the tables from the yeast cell-cycle context, and correlate this to the average profiles of the clusters. The percentages of the GO knowledge of the clusters were also studied. At 60% GO threshold, selected clusters were found to have more number of significant terms in different GO categories. Here, the GO% are mostly found to lie between 70% and 100%. This highlights the biological relevance of these clusters, as well as their corresponding medoids (ORFs/genes). Fig. 3 depicts the bar chart of the percentages of the GO knowledge of a cluster.

Next the medoids (or representative genes) of the selected clusters were analyzed using SGD, to evaluate the biological significance of these genes. It was found that all medoids (genes) selected to constitute the reduced feature set are annotated. Let us discuss the functionality of some of the genes with reference to column 1 of Table III.

The gene YDL164C is a DNA ligase found in the nucleus and mitochondria. It acts as an essential enzyme that takes part in DNA replication as well as takes part in nucleotide excision repair, base excision repair, and recombination. YDR385W is an elongation factor 2 (EF-2) that catalyzes ribosomal translocation

TABLE II
FIRST-PASS CLUSTERING FOR DIMENSIONALITY REDUCTION, BASED ON GENE ONTOLOGY STUDY, ON *Yeast* DATA

Iteration	No. of clusters	Genes in each cluster	# Compressed clusters n_{cl}
1	$\sqrt{2879} = 54$	62,40,84,71,27,14,47,55,32,49,**87**,32,**15**,25,80,79,45,109,71,92,50,81,55,**56**,**65**,**35**,**17**, 74,64,22,**37**,**42**,49,60,62,84,30,**39**,78,86,39,45,121,56,46,44,54,32,60,**32**,24,27,43,**22**	11
2	$\sqrt{2879 - 417} = 49$	49,24,27,**47**,45,72,**36**,17,58,29,29,35,29,79,38,62,30,30,43,65,69,-3,**12**, 45,80,51,82,41,31,98,78,27,25,29,67,74,30,59,42,56,30,69,44,83,83,91,55,44,50	3

TABLE III
BIOLOGICAL VALIDATION OF SAMPLE CLUSTERS FOR *Yeast* DATA

Medoid & # ORFs	MIPS functional category / subcategory			
	category	subcategory	# ORFs	p value
YDL164C	CELL CYCLE AND		7	2.74e-03
	DNA PROCESSING	DNA synthesis and replication	5	5.99e-06
15	(biological process)	extension/polymerization activity	3	5.55e-05
YDR385W	PROTEIN SYNTHESIS		31	8.60e-21
	(biological process)	ribosome biogenesis	27	3.54e-21
56		ribosomal proteins	27	6.75e-24
		translation elongation	5	8.81e-07

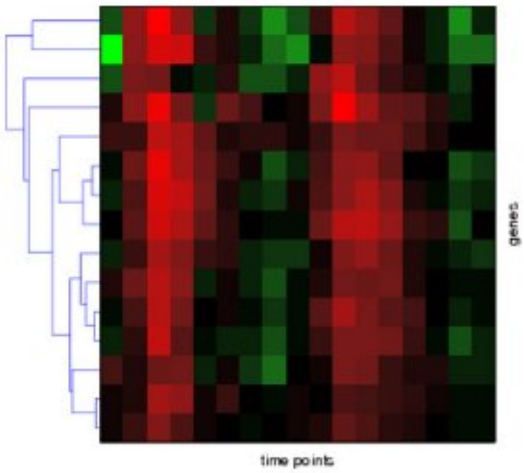

Fig. 1. *Yeast* data Eisen plot having 15 genes.

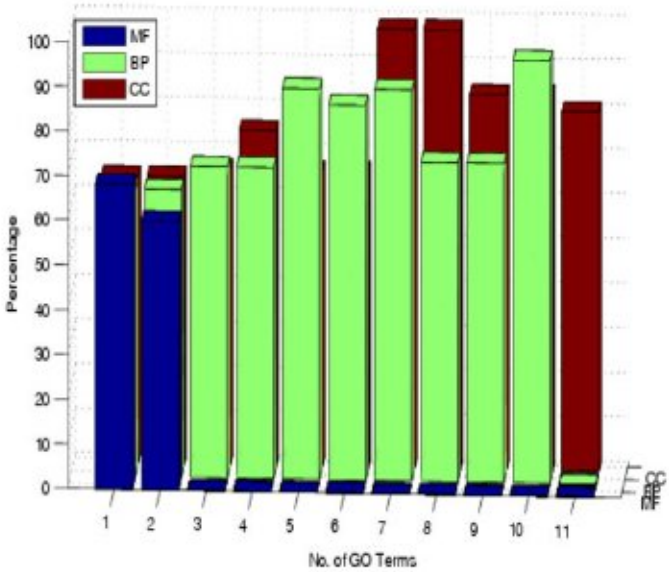

Fig. 3. Percentages of GO knowledge of sample cluster of *Yeast* cell-cycle, having 56 genes.

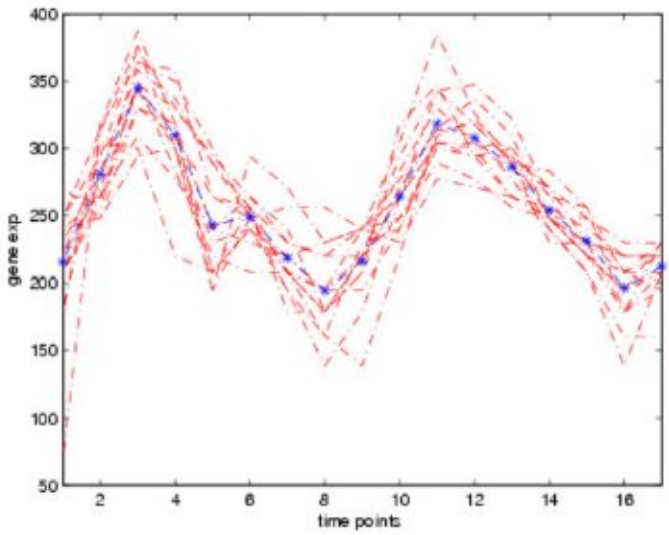

Fig. 2. *Yeast* data cluster profile plot having 15 genes.

during protein synthesis. YER018C is a component of the evolutionarily conserved kinetochore-associated Ndc80 complex. It is involved in chromosome segregation, spindle checkpoint activity and kinetochore clustering. YOR234C is a ribosomal protein L37 of the large (60S) ribosomal subunit, and is involved in structural constituent of ribosome and translation.

The medoids, i.e, genes, of the selected clusters are then taken as the reduced attribute set for the second stage of clustering (by CLARANS) in Table IV. Here, the time points are shown being grouped together to identify the existing biological classes over the reduced feature set. The original biological class information of the dataset is compared for the purpose of validation. It is observed that the optimal partitioning corresponding to two clusters (time points 1–8, 9–16), at minimum value of DB, is biologically meaningful as evident from the original biological class information of these cell-cycle data. Note that these

TABLE IV
SECOND-PASS CLUSTERING USING VALIDITY INDEX ON *YEAST*

No. of clusters	Time points in each cluster	Index value
2	1-8,9-17	0.09
3	1, 2-8, 9, 10-12, 13, 14-17	1.27
4	1,2-8, 9, 10-11, 12-13, 14-17	1.88
5	1-2,3-5,6-8,9, 10-11, 12-13,3, 14-17	1.09
6	1 2,3 5,6 8,9, 12 15, {10 11,16 17}	1.21
7	1-2,3-5,6-8,9, 10, 12-13, {11,14-17}	1.19
8	1,2,3-5,6-8,9, 10, 11-12, 13, 14-15, 16-17	2.41

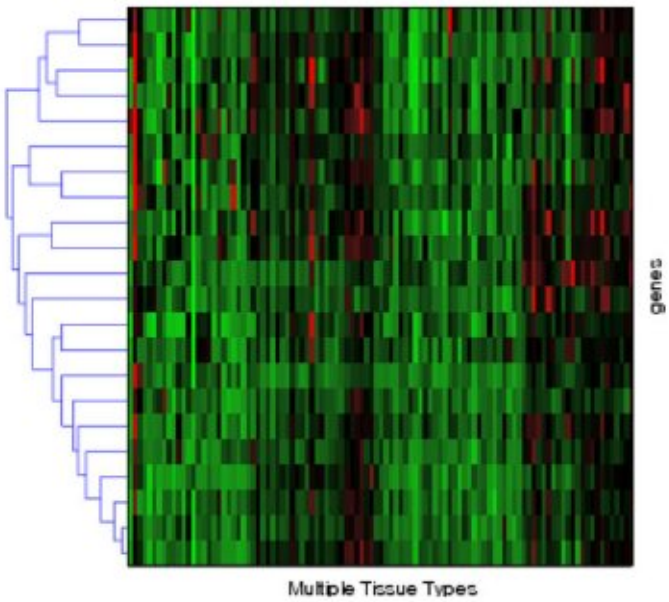

Fig. 4. Multiple Tissues data Eisen plot having 22 genes.

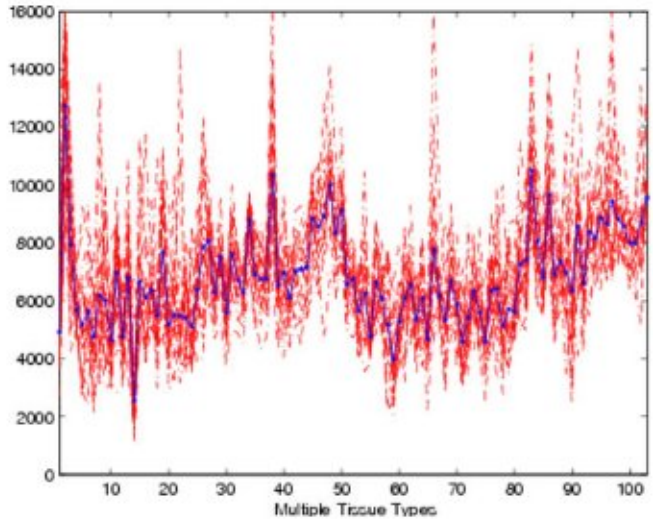

Fig. 5. Multiple Tissues data cluster profile plot having 22 genes.

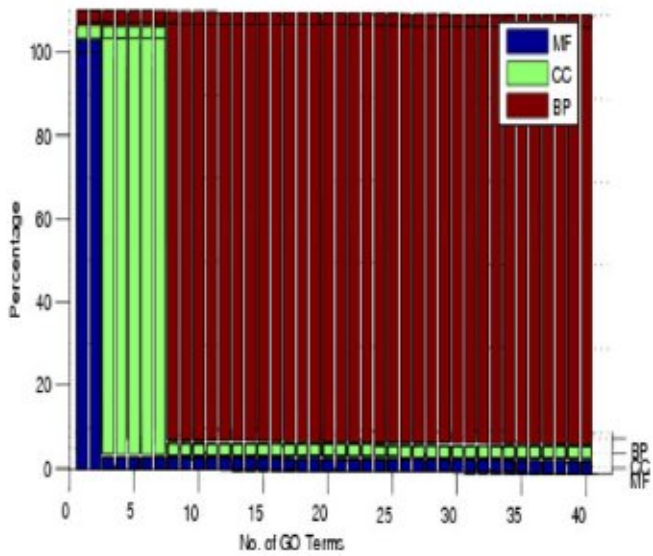

Fig. 6. Percentages of GO knowledge of sample clusters of Multiple Tissues, having nine genes.

partitions correspond to a reduction in the number of genes from 2884 to 15.

B. Multiple Tissues

The medoids (ORFs) of the selected clusters are indicated next:

RPL3, PRPF31, RPL13A, RPS3, RNP24, DLST, PUM1, TRIM33, FBXO7, HIPK3, CCT7, NACA, S100A2, RPL31, NPPA, OAZ1, FBXW11, MT1H, ZWINTAS, KHDRBS1, MFN2, EEF1G, PRSS11, TNFRSF25, MYLK, PFDN5, ZFR, SFTPB, ENO1, TMEM1, PCBP2, CUL1, LEREPO4, HLA-DPA1, RECK, HNRPC, IL7R, COL6A3, UBC.

The results are found to match with the corresponding average profiles depicted in Figs. 4 and 5. The percentages of GO knowledge of the different GO categories in the significant clusters were found to be lying between 80% and 100%. The sample cluster is illustrated in Fig. 6.

All the selected genes found to be annotated. We present here the findings that are related to some of the genes of Table VI in detail. The gene EEF1G is defined as a eukaryotic translation elongation factor 1 gamma which takes part in response to virus, protein biosynthesis, and translation elongation factor activity. It was found that EEF1G mRNA was expressed at higher levels in 7 of 9 pancreatic tumors than in the corresponding normal tissues. RPL31, which is defined as ribosomal protein L31, takes part in structural constituent of ribosome and translation regulator activity.

C. Leukemia

The medoids (ORFs) of the selected are as follows:

M26311_s_at, HG2887-HT3031_at, L07633_at, M63438_s_at, L06499_at,

AFFX-HSAC07/X00351_5_at, HG3597-HT3800_f_at, U83239_s_at, U87459_at,

AF006084_at, J00105_s_at, X00437_s_at, X06617_at, Z84721_cds1_at, D26598_at,

X56997_rna1_at, Z48950_at, X62691_at, U23852_s_at, J04130_s_at, M11147_at,

M33600_f_at, M19507_at, M26602_at, U80987_s_at, Z19554_s_at, M14483_rna1_s_at,

M27783_s_at, U10323_at, M95627_at, M28130_rna1_s_at, M15395_at.

The percentages of the GO knowledge of all the selected clusters were found to lie between 60% and 90%, thereby exhibiting the biological relevance of the clusters. Fig. 7 demonstrates the results of such a cluster. The study of the selected medoids (genes), using Uniprot, also demonstrates their significance in terms of meaningful annotation. Gene U87459_at is defined as an L-antigen family member and takes part in protein binding. Moreover, it is found to be expressed in a wide variety of cancers. The gene X00437_s_at associates with IL12B to form the IL-23 interleukin, which again is an heterodimeric cytokine functioning in situations of innate and adaptive immunity.

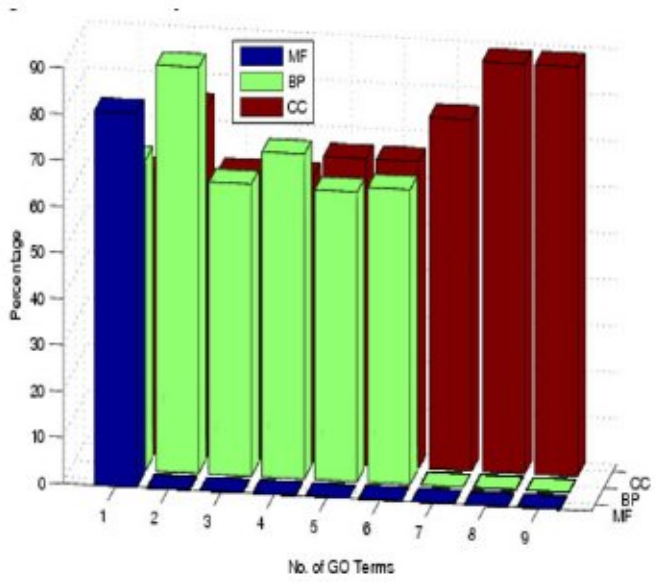

Fig. 7. Percentages of GO knowledge of sample clusters of Leukemia, having 1072 genes.

D. External Validation

The biologically meaningful subspaces, which are generated qualitatively in terms of GO, were next validated using various standard classifiers. These included the k-NN, C4.5, random forest, MLP, SVM, and NB (using WEKA and DTREG implementations). The gene sets in the original and reduced subspaces were examined for all the three microarray datasets of Sections IV-A–IV-C on the basis of their classification perfromance. Ten-fold cross validation was done in all cases. The k-NN classifier was implemented for $k = 1, 3, 5, 7$, the MLP was used with one hidden layer having three hidden nodes, and the SVM employed the radial basis function kernel.

Next, we made a comparative study with the performance of attribute clustering algorithm (ACA) [34]. The compared algorithm ACA is able to group genes based on their interdependence, so as to mine meaningful patterns from the gene expression data and select significant attributes for subsequent classification. No biological knowledge is involved during attribute selection. Table IX depicts the comparative study for the *Leukemia* dataset, where the selected gene subsets (by ACA and our algorithm) were validated using some well-known classifiers. For each classifier, the first column refers to ACA while the second corresponds to our algorithm. In the case of ACA, the five gene subsets were of sizes 10, 20, 30, 40, and 50, selected based on a ranking threshold used. Our algorithm, on the other hand, automatically selects a particular subset of genes using ontology-based biological knowledge. It is evident from the results that our algorithm is able to select a better subset of genes from the *Leukemia* data, due to the incorporation of biological knowledge, and resulted in an improved classifier performance (entries marked in bold in the table).

Finally, we present an experimental comparison with the case of not including biological knowledge in the first pass of our algorithm. Table X depicts the results on the three microarray datasets. In the case of each data, the first row refers to the situation where we do not use the p-value to select only the biologically good clusters (marked in bold in Tables II, V, and VII). Instead, we have taken the medoids (genes) of all the clusters to constitute the reduced feature set. The third row corresponds to the gene subset after reduction using biological knowledge by our algorithm. This, therefore, serves to highlight the utility of incorporating biological knowledge (here in terms of GO) in Steps 4–6 of the algorithm. The second row refers to a random selection of genes, such that this number equals the number of medoids (involving biological knowledge) in the corresponding third row. Again, a set of well-known classifiers are used for the external validation.

We observe that the incorporation of biological knowledge resulted in an improved performance for classifier NB over all three datasets. We have marked in bold all the cases where ever this reduced subspace yielded better classification performance. Overall, the performance of the algorithm is found to be comparable and often better after the incorporation of biological information during reduced feature subset selection.

Moreover, the performance of the proposed algorithm is compared with those of some biomarker extraction methods,[15] such as the empirical ebayes t-statistic (eBayes) [35], [36], ensemble [37], partial least-squares cross validation (PLSCV) [38], random forest feature selection (RFMDA) [39], and the significance analysis in microarrays (SAM) [40]. Results of the NB and decision tree are presented in Table XI for the three datasets. Here, the classification accuracy of our algorithm is found to be better in most cases. In this connection, it may be noted that the biomarker extraction methods use statistical measures to select differentially expressed genes. As such, the computational complexity increases. Our algorithm is simple, yet comparable in performance.

[15] http://www.arraymining.net.

TABLE V
FIRST-PASS CLUSTERING FOR DIMENSIONALITY REDUCTION ON MULTIPLE TISSUES DATA

Step	No. of clusters	Genes in each cluster	# Compressed clusters n_{cl}
1	$\sqrt{5565} = 75$	**3.199**,4.2.3,1.**15**,2.1.4.1.**25.418.215.259.135.482.45**,10,4,11,1.5.9,1.**94**,4.**318**,4,1.**74**,**106**. 1,1,1,1,**133**,11.**22**,1.**19.281. 10. 52** . 2.**415**,2,45,2.83,2.**15.506.48.273**,2,1.**160**,1,2.1.6, 1,126.1,**8**,1.3,**653**,**115**.2,**43**,**10**.4,1,6	39
2	$\sqrt{5565 - 5193} = 19$	8.9.4,2.1.39.5.3,7.6.14.7,**147**,13,37,12.8,**42**,8	3

TABLE VI
BIOLOGICAL VALIDATION OF SAMPLE CLUSTERS FOR MULTIPLE TISSUES DATA

Medoid & # ORFs	FatiGo functional category / subcategory			
	category	level	subcategory	p-value
EEF1G 22	biological process	3	biosynthetic process	2.10e-20
		4	cellular biosynthetic process	4.22e-21
			cellular macromolecule metabolic process	2.86e-12
			protein metabolic process	1.30e-12
		5	macromolecule biosynthetic process	1.86e-24
			cellular protein metabolic process	2.97e-12
		6	translation	2.16e-27
	molecular function	3	structural constituent of ribosome	9.20e-34
		4	RNA binding	2.33e-18
		5	rRNA binding	4.62e-10

TABLE VII
FIRST-PASS CLUSTERING FOR DIMENSIONALITY REDUCTION ON LEUKEMIA DATA

Step	No. of clusters	Genes in each cluster	# Compressed clusters n_{cl}
1	$\sqrt{7129} = 84$	9.1,2,2.1,**319**,1.**29.3.3**,2,**2.1**,4,**76**,**378.1072**,56,1.**603**,29.1,**788**,7.**11**.1.1,**73**,**10.16**. 3,**9**.13.**119**.3,7,**83**,5.1.22...6,**316**.1.**5**,1.7.2,**32**,10.1,14.12.1,**865**.1.**191**,**218**.1.1, 2.1,1,3.6.56.1.1,1.**420.5**,1.3.1,**4.39**,**171**.**212**,9,2.**530**.21.**184**,2.	32
2	$\sqrt{7129 - 6783} = 19$	**36**,**4**.**17**,**32**,1.13,17,13,46.20.30.10,15,6.18,**56**,6.**9.5**	12

TABLE VIII
BIOLOGICAL VALIDATION OF SAMPLE CLUSTERS FOR LEUKEMIA DATA

Medoid & # ORFs	FatiGo functional category / subcategory			
	category	level	subcategory	p-value
X00437_s_at	biological process	3	response to external stimulus	2.10e-08
			cellular developmental process	4.78e-07
			response to stress	1.24e-06
788			regulation of biological process	1.28e-06
			cell proliferation	2.51e-06
			regulation of body fluids	2.69e-06
		4	positive regulation of biological process	1.86e-08
			negative regulation of biological process	7.94e-08
			cell differentiation	6.84e-07
			cell-cell signaling	6.85e-06
			system development	6.58e-06
	molecular function	3	protein binding	4.14e-10
			transcriptional activator activity	4.02e-07

TABLE IX
COMPARATIVE STUDY OF VALIDATION BY DIFFERENT CLASSIFICATION ALGORITHMS ON THE REDUCED GENE SET FOR LEUKEMIA DATA

# Genes Selected		Classification Accuracy(%)					
		ANN		k-NN		NB	
ACA	Proposed method	ACA	Proposed method	ACA	Proposed method	ACA	Proposed method
10	44.0	94.1	96.0	91.2	**92.0**	82.4	**98.6**
20		94.1		91.2		61.8	
30		97.1		91.2		61.8	
40		97.1		91.2		61.8	
50		97.1		91.2		61.8	

TABLE X
COMPARATIVE STUDY WITH AND WITHOUT BIOLOGICAL KNOWLEDGE

Dataset	Method	# Genes	Accuracy (%)				
			k-NN	C4.5	RF	MLP	NB
Yeast (Cho) N=17 c=2	Without	54	98.6	94.11	94.1	98.0	91.11
		15	83.82	82.35	88.23	94.11	99.0
	With	15	86.78	**94.12**	**94.12**	94.12	**94.12**
Multi Tissue N=103 c=4	Without	75	74.05	63.21	83.0	89.0	78.2
		42	77.79	64.46	75.45	89.32	89.32
	With	42	**81.03**	**65.0**	76.0	**89.32**	**92.23**
Leukemia N=72 c=2	Without	84	88.53	83.33	83.33	94.44	88.88
		44	84.37	87.50	84.72	97.22	93.05
	With	44	**90.83**	83.33	**90.3**	**96.0**	**97.22**

TABLE XI
COMPARATIVE STUDY WITH BIOMARKER EXTRACTION METHODS

Data set	# Genes	Methods/ Algorithms	Accuracy(%)	
			NB	Decision Tree
Yeast (Cho)	15	EBayes	**100**	**100**
		ENSEMBLE	94.11	82.35
		PLSCV	88.23	70.58
		RFMDA	**100**	**100**
		SAM	**100**	**100**
		Proposed	94.12	94.12
Multi Tissue	42	EBayes	89.32	70.87
		ENSEMBLE	90.29	70.72
		PLSCV	93.20	70.87
		RFMDA	92.05	**80.43**
		SAM	89.32	70.87
		Proposed	**92.33**	71.16
Leukemia	44	EBayes	95.83	90.27
		ENSEMBLE	95.83	90.27
		PLSCV	93.05	92.83
		RFMDA	95.83	90.27
		SAM	89.32	93.05
		Proposed	**97.22**	**93.68**

V. CONCLUSION

In this paper, we have proposed an algorithm for feature selection using biological knowledge. Here, biological knowledge implies the statistical significance of the enrichment of GO terms in the clusters as represented by the constituent genes. The concept to select the medoid (of most representative gene) as a feature from each biologically enriched gene cluster, incorporating biological information, was unique. The algorithm CLARANS was employed as a tool for attribute clustering in the large gene space. The clusters were evaluated in terms of GO, and pictorially depicted in terms of Eisen plot and gene profiles. Subsequent partitioning of the reduced gene space over the conditions validated the significance of this reduced subset of genes.

The main objective was to demonstrate the utility of incorporating biological knowledge, in the form of annotation from the GO study, for enhanced dimensionality reduction by attribute clustering. Since the selection of genes was based on statistically significant biological meaning, therefore some potentially useful clusters could have been overlooked in the process. However, the algorithm improves the interpretability of the results by providing the biologist with clusters that are based on genes selected using background knowledge. Extraction of subsets of genes, from the high-dimensional gene space, lead to reduced computational complexity. External validation demonstrated that the classification accuracy in the reduced subspace was, in general, better.

Extensive comparison was provided with related algorithms, including biomarker extraction methods [35]–[40]. Other pieces of literature, involving set-level techniques, include [41] and [42]. The use of differentially expressed genes is also being currently investigated by us.

It is to be noted that a computationally efficient globally optimum measure of evaluation does not necessarily converge to a biologically meaningful solution. The objective was, hence, not a faster or more efficient clustering. Rather, the algorithm was expected to serve as an automated aid to biologists by enhancing visualization, freeing them of the need for manual intervention, and allowing them scope for refinement according to their focus of interest. A biologically meaningful subspace can also lead to the discovery of valuable pathways of genes. This, therefore, holds promise for biologists to interpret and analyze various subspaces according to their individual requirements.

REFERENCES

[1] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowledge Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[2] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, pp. 245–271, 1997.

[3] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.

[4] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 44–49, Mar./Apr. 1998.

[5] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Graph. Statist.*, vol. 55, pp. 119–139, 1997.

[6] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[7] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 373–378, Mar. 2003.

[8] S. Mitra and T. Acharya, *Data Mining: Multimedia, Soft Computing, and Bioinformatics*. New York: Wiley, 2003.

[9] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary-rough feature selection in gene expression data," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 4, pp. 622–632, Jul. 2007.

[10] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 1, no. 1, pp. 24–45, Jan.–Mar. 2004.

[11] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE Trans. Knowledge Data Eng.*, vol. 14, no. 5, pp. 1003–1016, Sep./Oct. 2002.

[12] D. R. Bickel, "Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically," *Bioinformatics*, vol. 19, pp. 818–824, 2003.

[13] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau, "Adaptive quality-based clustering of gene expression profiles," *Bioinformatics*, vol. 18, pp. 735–746, 2002.

[14] D. Huang and W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," *Bioinformatics*, vol. 22, pp. 1259–1268, 2006.

[15] Z. Fang, J. Yang, Y. Li, Q. Luo, and L. Liu, "Knowledge guided analysis of microarray data," *J. Biomed. Informat.*, vol. 39, pp. 401–411, 2006.

[16] L. Tari, C. Baral, and S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *J. Biomed. Informat.*, vol. 42, pp. 74–81, 2009.

[17] I. Gat-Viks, R. Sharan, and R. Shamir, "Scoring clustering solutions by their biological relevance," *Bioinformatics*, vol. 19, pp. 2381–2389, 2003.

[18] F. Gibbons and F. Roth, "Judging the quality of gene expression-based clustering methods using gene annotation," *Genome Res.*, vol. 12, pp. 1574–1581, 2002.
[19] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast gene ontology based clustering for microarray experiments," *BioData Mining*, vol. 1, pp. 1–11, 2008.
[20] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, NJ: Wiley, 2001.
[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, pp. 10–18, 2009.
[22] P. H. Sherrod. (2008). DTREG, software for predictive modeling and forecasting. [Online]. Available: http://www.dtreg.com
[23] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[24] S. Mitra, H. Banka, and W. Pedrycz, "Rough-fuzzy collaborative clustering," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 36, no. 4, pp. 795–805, Aug. 2006.
[25] V. S. Tseng and C.-P. Kao, "Efficiently mining gene expression data via a novel parameterless clustering method," *IEEE Trans. Comput. Biol. Bioinformat.*, vol. 2, no. 4, pp. 355–365, Oct./Dec. 2005.
[26] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat. Genet.*, vol. 22, pp. 281–285, 1999.
[27] J. Bezdek and N. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 28, no. 3, pp. 301–315, Jun. 1998.
[28] A. I. Su, M. P. Cooke, K. A. Ching, and Y. Hakak, J. R. Walker, "Large-scale analysis of the human and mouse transcriptomes," *Proc. Nat. Acad. Sci.*, vol. 99, pp. 4465–4470, 2002.
[29] S. B. Cho and S. H. Yoo, "Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data," *Pattern Recog.*, vol. 39, pp. 2405–2414, 2006.
[30] H. W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil, "MIPS: A database for genomes and protein sequences," *Nucleic Acids Res.*, vol. 28, pp. 37–40, 2000.
[31] E. L. Hong, R. Balakrishnan, Q. Dong, K. R. Christie, J. Park, G. Binkley, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, C. J. Krieger, M. S. Livstone, S. R. Miyasato, R. S. Nash, R. Oughtred, M. S. Skrzypek, S. Weng, E. D. Wong, K. K. Zhu, K. Dolinski, D. Botstein, and J. M. Cherry, "Gene Ontology annotations at SGD: New data sources and annotation methods," *Nucleic Acids Res.*, vol. 36, pp. D577–D581, 2008.
[32] F. Al-Shahrour and R. Daz-Uriarte, J. Dopazo, "FatiGO: A web tool for finding significant associations of gene ontology terms with groups of genes," *Bioinformatics*, vol. 20, pp. 578–580, 2004.
[33] S. N. Twigger, M. Shimoyama, S. Bromberg, A. E. Kwitek, and H. J. Jacob, "The Rat Genome Database, update 2007—Easing the path from disease to data and back again," *Nucleic Acids Res.*, vol. 35, pp. D658–D662, 2007.
[34] W. H. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," *IEEE Trans. Comput. Biol. Bioinformat.*, vol. 2, no. 2, pp. 83–101, 2005.
[35] I. Lönnstedt and T. Speed, "Replicated microarray data," *Stat. Sinica*, vol. 12, no. 1, pp. 31–46, 2002.
[36] G. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genetics Mol. Biol.*, vol. 3, no. 1, 2004.
[37] E. Glaab, J. Garibaldi, and N. Krasnogor, "Arraymining: A modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization," *BMC Bioinformat.*, vol. 10, no. 1, p. 358, 2009.
[38] A. Boulesteix and K. Strimmer, "Partial least squares: A versatile tool for the analysis of high-dimensional genomic data," *Briefings Bioinformat.*, vol. 8, no. 1, pp. 32–44, 2007.
[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[40] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci.*, vol. 98, no. 9, p. 5116, 2001.
[41] J. Klema, M. Holec, F. Zelezny, and J. Tolar, "Comparative evaluation of set-level techniques in microarray classification," *Bioinformat. Res. Appl.*, pp. 274–285, 2011..
[42] M. Mramor, M. Toplak, G. Leban, T. Curk, J. Demšar, and B. Zupan, "On utility of gene set signatures in gene expression-based cancer class prediction," *Mach. Learn. Syst. Biol.*, vol. 8, pp. 65–74, 2009.

Sushmita Mitra (S'91–M'92–SM'00–F'12) received the Ph.D. degree in computer science from the Indian Statistical Institute, Kolkata, India, in 1995.

She is the Head and Full Professor with the Machine Intelligence Unit, Indian Statistical Institute. From 1992 to 1994, she was a German Academic Exchange Service (DAAD) Fellow with the RWTH, Aachen, Germany. She was a Visiting Professor in the Department of Computer Science, University of Alberta, Edmonton, Canada, in 2004 and 2007; Meiji University, Tokyo, Japan, in 1999, 2004, 2005, and 2007; and Aalborg University Esbjerg, Denmark, in 2002 and 2003. She is the author of books *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing* (Wiley), *Data Mining: Multimedia, Soft Computing, and Bioinformatics* (Wiley), and *Introduction to Machine Learning and Bioinformatics* (Chapman & Hall/CRC Press), besides a host of other edited books. She has more than 100 research publications in referred international journals. According to the Science Citation Index, two of her papers have been ranked 3rd and 15th in the list of top-cited papers in engineering science from India during 1992–2001. Her current research interests include data mining, pattern recognition, soft computing, image processing, and bioinformatics.

Dr. Mitra received the National Talent Search Scholarship (1978–1983) from NCERT, India, the University Gold Medal in 1988, the IEEE TNN Outstanding Paper Award in 1994 for her pioneering work in neuro-fuzzy computing, and the CIMPA-INRIA-UNESCO Fellowship in 1996. She has guest edited special issues of several journals. She is an Associate Editor of the IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS and *Neurocomputing*. She is also a Founding Associate Editor of *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. She is a Fellow of the Indian National Academy of Engineering and The National Academy of Sciences, India.

Sampreeti Ghosh received the B.C.A. degree from the University of Burdwan, West Bengal, India, and the M.C.A. degree from the West Bengal University of Technology, West Bengal, in 2003 and 2006, respectively. She is currently working toward the Ph.D. degree at the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India.

During 2007–2010, she was a Research Scholar and during 2010–2011 a Visiting Scientist with the Center for Soft Computing Research, Indian Statistical Institute. Her current research interests include data mining, bioinformatics, and soft computing.