2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

Multiview clustering on PPI network for gene selection and enrichment from microarray data

Tripti Swarnkar*†, Sérgio Nery Simões‡§, David Correa Martins-Jr¶
Anji Anura‖, Helena Brentani**, Ronaldo Fumio Hashimoto‡ and Pabitra Mitra*
*Depatment of Computer Science & Engineering – Indian Institue of Technology, Kharagpur -721302, India
†Institute of Technical Education & Research – SOA University, Bhubaneswar-751030, India
Email: swarnkar.tripti@gmail.com
‡Institute of Mathematics and Statistics – University of São Paulo, São Paulo-SP, Brazil
§Federal Institute of Espírito Santo – Serra-ES, Brazil
¶Center of Mathematics, Computation and Cognition – Federal University of ABC, Santo André-SP, Brazil
‖School of Medical Science & Technology – Indian Institue of Technology, Kharagpur -721302, India
**Institute of Psychiatry – University of São Paulo, São Paulo-SP, Brazil

***Abstract*—Various statistical and machine learning based algorithms have been proposed in literature for selecting an informative subset of genes from microarray data sets. The recent trend is to use functional knowledge to aid the gene selection process. In this paper we propose a clustering algorithm which generates multiple views (clusters) from the microarray expression profiles, each representing a particular facet of the data. Such multiple clusters are found to represent strongly connected regions of the known protein–protein interaction (PPI) networks, perhaps corresponding to those responsible for certain biological processes. Thus we integrate microarray data clustering with PPI knowledge to obtain enriched gene sets. Results on benchmark microarray data sets demonstrate the competitiveness of our method compared to gene selection techniques.**

I. Introduction

Gene selection is the process of filtering the most significant and differentially expressed genes associated with disease conditions [1]. Selecting informative genes from microarray facilitates disease classification, as well as, in understanding the embedded biological processes, which further helps in diagnosis, prognosis and treatment of the disease [2].

In early literature gene selection was primarily done by statistical methods on microarray data sets [2]. The work done by [2], [3], [4] proposes gene selection models for different microarray gene expression datasets that selects biologically significant genes in mediating certain cancers.

The gene ontology (GO) based similarities between genes carry significant information of the functional relationships among the set of genes [2]. The integration of such domain knowledge helps to biologically enrich the selected genes and thus overcome the drawback of traditional statistical gene selection algorithms [2], [5], [6]. In large scale genomic data functional analysis, GO annotations are largely used to find the enriched genes [7]. Subramanian et al.[6] studied the different existing tools for gene enrichment.

In order to study the structural and functional characteristics of biological processes, large and complex networks of molecular interactions are being rapidly generated for humans and other organisms [8], [9]. The network of mechanistic physical interactions between proteins is referred as protein-protein interaction (PPI) network [10]. The integration of information from gene expression and PPI network data is a promising approach to prioritize disease-associated genes [8], [9], [11]. Barabasi et al.[12] summarized a series of hypothesis and principles (*Network Medicine Hypotheses*) which link topological properties of PPI networks to biological functionalities. Some of these hypothesis are often used to prioritize candidate genes related to a given disease. We highlight three hypothesis:

- *disease module hypothesis*: gene products associated with the same disease phenotype tend to form a cluster in the PPI network;
- *network parsimony*: shortest paths between known disease genes often coincide with disease pathways;
- *local hypothesis*: gene products associated with similar diseases are likely to strongly interact with each other.

However, the PPI data are often incomplete and noisy. Such problems can be alleviated considering the observation based on indirect molecular interactions [13]. The integration of the gene expression profiles and the PPI network improves the performance of prognostic prediction [14], [15], [16].

Traditional methods for gene expression analysis focus on identifying gene sets that help to discriminate between two states of interest. Although these sets of genes are useful, they fail to detect biological relevance and the processes involved with these states [15]. Most of the ranking based gene selection methods select informative genes in isolation and ignore their plausible co-relation or interaction[17]. The goal of our work is to integrate cluster structure among co-expressed genes in microarray experiments with graph structure of gene interactions in PPI networks. Not just discriminatory genes but genes which are responsible for concerned biological processes are likely to be selected in this process. In most biological applications with high dimensional data, there may exist multiple groupings of the data that are all reasonable in some perspective. In such cases, different feature subspaces with each feature subspace presenting the user a different view of the data structure [18] can often warrant different ways

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.33

to partition the data. Further, gene clusters as expressed in microarray data may be of various representations and views depending on the regulatory process being considered. Hence, we use a multiview clustering algorithm to cluster the microarray data and obtain several alternate views that correspond to the biological processes. These views are then mapped to the PPI network to extract the subnetworks that correspond to these hypothetical processes. The degree of interactions in these subnetworks are significantly higher than a random network exhibiting identical degree distribution. We propose a significance measure based on this principle. The genes in these subnetworks are ranked based on the significance measure which points to their centrality and importance in the processes. The top ranked genes are finally selected.

We demonstrate our algorithm on three benchmark data sets. The selected genes are found to closely correspond to known biological processes associated with the diseases in literature. A high classification accuracy is also obtained.

II. Proposed Methodology

In this section we discuss the proposed methodology for gene selection and enrichment. The statistical significance and biological relevance of genes should be considered together to identify the differentially expressed genes from high-throughput gene expression data [19]. In the preprocessing step we have filtered the genes based on their variance across the samples and thus considered the genes with variance less than tenth percentile for further processing [20]. Further, the $10,000$ permutations t-test [21] is used and the genes with p-value cutoff of 0.05 are considered to have statistical significance [7], [19] and are used for subsequent analysis. This set of data are normalized using the mean column intesity and the raw intensities are transformed to the range of 0 to 1 values for each sample. The Multiview Clustering (MVC) [22] based gene selection algorithm is used on these datasets to obtain gene subsets resulting in multiview clusters. The most efficient gene clusters in terms of performance are selected. The genes of selected clusters are further ranked/scored using PPI network.

Figure 1 represents the schematic workflow of the proposed Multiview Clustering and PPI Network (MVC-PPI) based gene selection model.

A. Multiview clustering (MVC) of microarray data

We describe below a MVC based gene selection algorithm on microarray dataset. The network representations and clusterings used in the methodology are first defined as follows.

(i) *Gene profile network*: Based on the expression level of a single gene we may construct a graph with each microarray sample as the vertices and similarity of expression of that particular gene determining the existence of an edge between two samples. We denote this graph as the *gene profile network*. There are as many networks as the number of genes for a particular experiment.
In summary, a graph is constructed for each feature (gene), with each sample as a vertex. There is an edge between two vertices in the graph if the difference in the values of the normalized expression for the gene for those two samples is less than a threshold of 0.5.

(ii) *Gene correlation network*: The distance between two gene profile networks is measured as symmetric difference between the edge sets of the networks, i.e., the number of edges present in one network but absent in the other. The symmetric difference between two sets A and B is defined as $A\Delta B = (A - B) \cup (B - A)$. The symmetric difference may be computed by computing the exclusive OR (XOR) between the corresponding adjacency matrices [23]. Though other measures of graph similarity may be considered, we have used the symmetric difference measure for its computational simplicity and effectiveness. By using this distance, a gene correlation network is built with genes as vertices and the above distance as the edge weight between them. The edge weights in the gene profile network represent the degree of co-expression between two genes.

(iii) *Gene network clustering*: These resulting gene correlation networks are now clustered into k partitions, where k is previously defined. The choice of k takes into consideration the cluster/view quality (the quality was measured by intracluster correlation), as well as the size of the clusters. For the three datasets considered here we got best result for $k = 10$. The proposed model here uses hierarchical agglomerative clustering with edge weight of the gene correlation network as the distance measure.

(iv) *Gene subset profile network*: Each cluster obtained above represents a group of genes, expectedly having some biological functional similarity. We denote this as a *view*. Thus, we have k clusters of the data each corresponding to a set of related genes.
Now, gene subset profile networks are constructed, one for each of these k clusters, with samples as vertices and edge weights being the Euclidean distance between two samples considering only the expression of the gene subset belonging to that particular gene network cluster or view.

(v) *Gene subset clustering*: Each of these gene subset profile networks has samples as nodes. Now we partition each of these k networks into l sample clusters. Thus, each of these k graphs gives rise to a separate clustering set of l clusters each. If the true class label is known as in case of our data then l is the same as that of the true class label count, otherwise it can be dependent on the cluster quality or domain knowledge. This step of sample clustering is performed to measure the class performance of each of the views obtained at step (iv).

Each of these clusterings can be considered to provide a separate cluster of the microarray expression data, leading to multiview interpretation of the data set. The genes in these clusters have similar expression patterns and are expected to be involved in a specific biochemical pathway showing the biological relevance of the clusters obtained from MVC. To further biologically enrich the genes in this multiview cluster set $\mathcal{V}$ we ranked the genes in top two views obtained from step (iii) using PPI information as described in Section II-B.

B. Enrichment of clusters using PPI networks

To enrich the cluster views obtained from previous step we use PPI graph structures considering the network medicine hypothesis. Based on disease module hypothesis and network

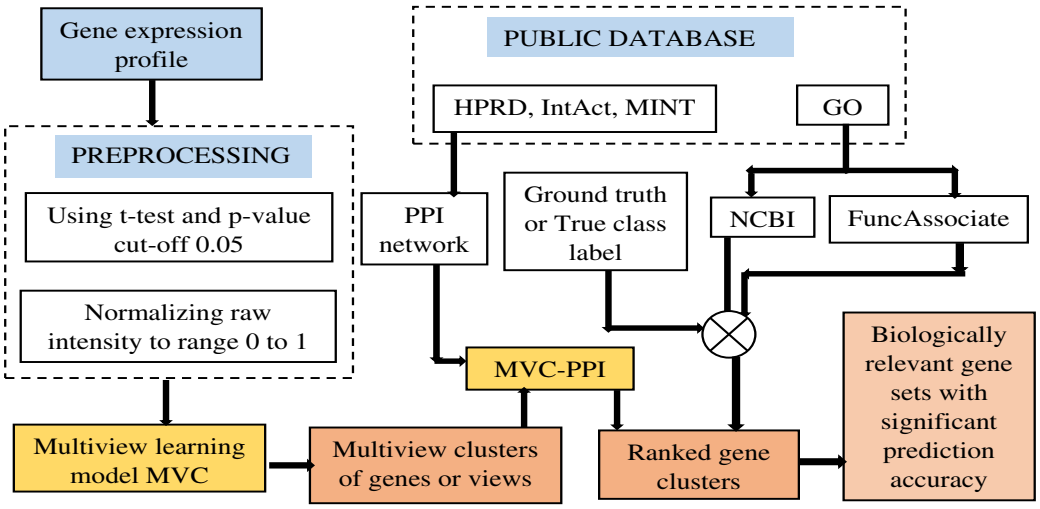

Fig. 1. Steps of the Multiview Clustering and PPI Network Gene selection method based on MVC enriched with PPI information

parsimony principle, the idea is to evaluate how close genes from a given view/cluster (obtained from Section II-A) are to each other in the resulting PPI network. Such a PPI network is composed by known PPI databases (see Section III-B) and is represented by a graph $G = (P, I)$, where P is the set of proteins/genes and I is the set of interactions. Each gene g in a view V is mapped to its respective protein in PPI network. Based on the assumption that a given view obtained above contains strongly connected genes in a PPI network [12], the goal is to assess the significance of these genes, as compared to randomly generated gene sets that preserve the same degree distribution of the original cluster. This is done to avoid bias towards highly connected genes that tend to be favored by some prioritization methods [13]. The genes are ranked based on this significance measure and the most significant ones are finally selected.

The algorithm for significance computation that we propose is as follows :

(i) For each gene g from a cluster view V, compute the average distance of g to all other genes in V:

$$\bar{d}_V(g) = \frac{1}{|V|} \sum_{v \in V \setminus \{g\}} d(g, v)$$

(ii) To estimate the significance of $\bar{d}_V(g)$ for a given gene g in a cluster view V, randomly generate another set of genes V' with the same degree distribution of V and compute the average distance $\bar{d}_{V'}(g)$ by applying the previous equation with V' in place of V.

(iii) Repeat step (ii) n times (e.g., $n = 1,000$), generating n random cluster views V'_j, for $j = 1, \ldots, n$. Then, estimate the mean $\hat{\mu}_V(g)$ and standard deviation $\hat{\sigma}_V(g)$ of average distances $\bar{d}(g, V'_j)$, for $j = 1, \ldots, n$.

(iv) For each gene g from the original cluster view V, compute its normalized score $s_V(g)$ as defined below:

$$s_V(g) = \frac{\bar{d}_V(g) - \hat{\mu}_V(g)}{\hat{\sigma}_V(g)} \quad (1)$$

Better interactions are represented by more negative scores $s_V(g)$, since this indicates that gene g tends to have smaller average distances in the original view V than in other randomly generate views V' with the same degree distribution. Thus, it is expected that most genes from the original cluster view will present negative scores, showing strong connection in a PPI network. This score is used to rank the genes.

C. Performance Evaluation

In our experiment, predictive accuracy of the multiview clusters is measured in terms of overall accuracy in comparison to the known true classes [17]. Samples are considered to be divided in two categories, namely, positive samples and negative samples. Accuracy is defined by equation 2;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where TP is the number of true positive samples, TN is the count for true negative samples, FP is the number of false positive samples and FN is the number of false negative samples. To biologically validate these clusters/views we have summarized the functional association of the genes present in these clusters and their biological significance in the progression of the disease. For this we revisited these clusters in the light of biological domain knowledge available in literature (see Section III-C).

The percentage of disease related genes (true positive) present in these clusters is used as a measure for showing the biological significance. The p-value cutoff of 5×10^{-5} is being considered in our study to find the enriched functional attributes (EA) for a set of genes.

III. Datasets

A. Gene expression datasets

The study has been applied to three different NCBI's GEO[1] datasets.

[1]Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/)

(i) B-CLL chronic lymphocytic leukemia is a heterogeneous disease with a pronounced variation in the clinical course. The dataset used after preprocessing consists of intensities of 6,572 genes in 11 B-CLL stable patients and 10 patients with clinically progressive disease [24].
(ii) Interstitial lung disease (ILD) represents a broad category of restrictive lung disorders [25]. The dataset containing intensities of 4,739 genes in 12 normal and 23 ILD is further used for analysis.
(iii) The third dataset used (called Leukemia) is Waldenstrom's macro globulinemia (B lymphocytes and plasmacells). More details about this dataset can be found in [3], [26]. The dataset after preprocessing consists of 3,384 genes with 13 normal and 43 diseased samples.

B. PPI datasets

The PPI network was built by the union of three PPI databases: (i) Human Protein Reference Database – HPRD[2] [27]; (ii) IntAct[3] [28]; (iii) Molecular Interaction database – MINT[4] [29]. The resulting PPI network has 12,373 proteins and 74,337 interactions. The transcripts from NCBI expression data (mentioned in Section III-A) were mapped to their respective original PPI network genes/proteins.

C. Biological databases adopted for validation

The NCBI Gene[5] database is used to collect disease related genes. The results are also validated using biochemical pathways, p-value statistics of Gene Ontology (GO) attributes which is a consolidated high-throughput strategy that helps the researchers to identify biological processes most pertinent to their study [7]. FuncAssociate[6], a web based application that discovers properties enriched in lists of genes or proteins that emerge from large-scale experimentation, is used for biological significance measure [30].

IV. Results and Discussion

We discuss the effectiveness of the proposed multiview gene selection methodology for human gene expression data collected from NCBI's GEO. We study two aspects of the selected genes - (i) their predictive classification accuracy, and (ii) the strength of their functional association based on known biological processes related to these diseases as obtained from the literature.

We have also shown that our proposed method outperforms widely used feature selection methods, namely Relief [31] and Principal Components Analysis (PCA) [32], [22]. A comparative analysis with a few existing models has also been reported, focusing on the relative advantages of multiview clustering over the single view gene selection.

[2]http://www.hprd.org/
[3]http://www.ebi.ac.uk/intact/
[4]http:// mint.bio.uniroma2.it/mint/
[5]http://www.ncbi.nlm.nih.gov/gene/
[6]http://llama.med.harvard.edu/funcassociate

A. Comparison of classification accuracy

We study the effectiveness of each view or cluster in terms of its predictive accuracy in comparison to the known true classes. The average predictive accuracy of all the views are found to be 0.66, 0.77 and 0.78 for B-CLL, Lung and Leukemia datasets respectively. Further from our preliminary classification performance analyses, we observed that three to five clusters out of k gene clusters generated by the proposed method described in Section II-A select small number of genes with high predictive accuracy.

Table I shows the performance of our proposed multiview gene selection technique individually and integrated with pathway knowledge from human PPI network (we report only one cluster from each dataset with best predictive accuracy). Our method is compared with single view gene selection algorithms, namely, PCA and Relief considering the same number of gene/feature sets. To evaluate the accuracy of our proposed model, PCA and Relief, we used hierarchical clustering and K nearest neighbors (kNN) classifier as our learning models on the selected genes by these methods. The comparison also includes other state of art single view methods from [24], [25], [26]. As seen in Table I, MVC-PPI gene selection with kNN was the only method that presented 100% accuracy for B-CLL. For ILD dataset, all compared methods presented 100% accuracy, except Relief with HC. Similarly, for Leukemia dataset MVC based methods with kNN outperformed all single view feature selection methods compared. This shows that MVC integrated with PPI by using gene scoring (described in Sections II-A and II-B) obtains either comparable or better accuracy with a reduced number of genes/features. We observe that the proposed MVC based model integrated with PPI enrichment on average selects relatively small gene subsets with competitive prediction accuracy.

B. Functional association of selected genes

In this section we study the biological relevance in terms of functional association of the genes in a cluster obtained using MVC integrated with PPI network knowledge. The biological significance of the genes belonging to an enriched functional GO category can be measured in terms of p-value [3]. The enrichment of each GO category or enriched attributes (EA) [30] in these clusters has been calculated in terms of p-value as well. Here we have considered the p-value cutoff equal to 5×10^{-5}. The enrichment of the functional association of the gene sets was evaluated with different ranges of gene ranks/scores.

Fält et al.[24] reports the functional association of 11 genes with p-value statistics 0.063 using WV (Weighted Voting) and 6 genes with p-value statistics 0.01 using LDA (Linear Discriminant Analysis) for B-CLL data. Cho et al.[25] identifies 30 biologically enriched genes with p-value statistics in the range of 0.1–0.0001 for ILD data. The paper studies the enrichement of each differentially expressed gene seperately without considering the functional association of these set of genes. Table II reports the top two enriched gene sets from Table I with their respective gene count, gene score range and the number of functionally EA's (enriched attributes). The gene score (GS) cutoffs were set in a way that the number of EA's recovered be in the range from 5 to 15. Interestingly, for the databases B-CLL and ILD the number of genes (GC: gene

TABLE I. ACCURACY OF GENE SELECTION ALGORITHMS: HC (HIERARCHICAL CLUSTERING), KNN (K NEAREST NEIGHBORS), WV (WEIGHTED VOTING), LDA (LINEAR DISCRIMINANT ANALYSIS), RBFS (RANKED BASED FEATURE SELECTION), NSC (NEAREST SHRUNKEN CENTROIDS), MVC (MULTIVIEW CLUSTERING), MVC-PPI (MULTIVIEW CLUSTERING INTEGRATED WITH PPI INFORMATION) – FOR DATASETS B-CLL (CHRONIC LYMPHOCYTIC LEUKEMIA), ILD (INTERSTITIAL LUNG DISEASE) AND LEUKEMIA (WALDENSTROM'S MACRO GLOBULINEMIA).

Test data	Accuracy											
	Single view								Multiview			
	PCA		Relief		Fält et al.[24]	Fält et al.[24]	Cho et al.[25]	Gutierrez et al.[26]	MVC		MVC-PPI	
	HC	kNN	HC	kNN	WV	LDA	RbFS	NSC	HC	kNN	HC	kNN
B-CLL	0.67	0.60	0.33	0.75	0.71	0.90	–	–	0.76	0.88	0.76	**1.00**
ILD	1.00	1.00	0.97	1.00	–	–	1.00	–	1.00	1.00	1.00	1.00
Leukemia	0.80	0.70	0.11	0.91	–	–	–	0.99	0.91	**1.00**	0.91	**1.00**

count) obtained were relatively small (between 5 and 13) while the number of EA's achieved were relatively large (between 5 and 15). Gutierrez et al.[26] identifies 171 differentially expressed genes for leukemia data, of which 29 are present in related pathways, where as from Table II it's seen that our method selects more biologically significant enriched genes in a smaller set of genes (view). This points to the effectiveness of multiview clustering integrated with PPI network.

Furthermore, considering only the most enriched gene set from Table II, the detailed functional description (EA Name) of its Enrichment Attributes ID (EA ID) are shown in Table III. The column (%) shows the percentage of genes from a gene set V sharing the EA's, related with the specified biological process shown in column named EA Name. As seen in Table III, more than half of the obtained gene set is functionally related in more than 30% of the enriched attributes for B-CLL and ILD. It is important to note that these results corroborate the local hypothesis, which states that gene products associated with a similar disease are likely to strongly interact with each other, and these genes tend to form a cluster in the PPI network [12]. Thus, from Tables II and III we can say that our proposed multiview clustering embedded with PPI network is able to select strongly connected regions of the known PPI networks, corresponding to those responsible for certain disease related biological processes.

C. Analysis of biological processes associated with selected genes

We have also analysed in details the biological processes that are associated with the enriched gene sets and tried to corroborate them with biological literature. It is noteworthy that, 5 out of 15 enriched attributes for B-CLL belong to immune system as seen in Table III. The immune system can play a vital role in promotion or elimination of tumours [33], [34]. Chronic lymphocytic leukemia (CLL) is the most common form of adult leukemia which is characterized by a progressive accumulation of functionally incompetent lymphocytes [24], [35]. The gene functional analysis of B-CLL data base (Table II), showed the strong interconnection among 13 genes. Most of these genes were involved in the process related with activation of immune response and cellular protein metabolic processes.

Interestingly, PAK, MAPK10, YES1, ERBB4 and CLU are the genes commonly present among all 13 processes recognized by our enrichment analysis. YES1 and ERBB4 are the retrovirus origin proto-oncogenes which functionally act as tyrosin kinase receptor and their over-expression is verified in numerous cancer conditions as they maintain cell proliferation [36]. Moreover, PAK and MAPK10 are involved in multiple biochemicals signaling at downstream signal of cell cycle regulator [37]. Similarly CLU is a gene found to be up-regulated during cancer progression.

In corroboration from literature its found that these genes play functionally important role in the development of CLL [38]. The interstitial lung diseases (ILD) is a restrictive lung disorder associated with cellular infiltration and distortion of the interstitium and alveolar gas units [25]. Pathologically the disease exhibits the activated process of wound healing, apoptosis, scarring and fibrosis and at molecular level it showed association with deregulation of pathways related to cytokines and signaling molecules [39]. The PPI analysis of the ILD database reveals that there are 5 genes which are strongly connected and are related to 9 different biological processes. Most of these processes are related to transcription, cellular response to dsRNA, cellular response to chemical stimulus and JAK-STAT signaling pathway. ILD is a disorder with numerous causative agents. Hence, the emergency of diverse biological process through microarray analysis is not an unexpected result. Thus, it indicates that ILD may be manifested due to the effect of viral infection or prolonged exposure to exogenous chemicals [40], [41]. STAT5B and SMAD1 are the genes commonly observed in all the enriched processes of Table II and III. These gene products act as signal transducers and transcriptional modulators that mediate multiple signaling pathways. Change in these molecular expressions has been witnessed during abnormalities in cytokines and growth factor signaling [42].

In case of Leukemia, more than 70% of the cluster is functionally associated in half of the enriched attributes. Waldenstroms macroglobulinemia (WM) is a C-cell monoclonal disorder that is characterized by infiltration of lymphoplasmacytic cells to bone marrow along with excessive presence of IgM in blood [43]. It is suggested that genetic factor plays significant role in pathogenesis of the cancer [44]. The functional analysis of the studied gene set reveals the presence of 46 closely related genes, most of which play important role in intercellular organization of cells. CHD4, EZH2, CCDC85B, WT1, BAX, TUBA1C, KLC1, RPS6KA6, CD9, PALLD, USP22, AGNF1, TCEB2, and CRIPT are the few genes in the gene set refered in Table II and III for Leukemia dataset, whose synchronized work regulates intercellular organization at transcription and proteomics levels. Among these genes, CDH4, EZH2, CCDC85B, WT1, RPS6KA6, USP22, TCEB2, and BANF1 maintain transcriptional repressive state of genes by

regulating chromatin organization, however genes TUBA1C, KLC1, CD9, and PALLD were related to the microtubular cytoskeleton organization which may play a significant role in over synthesis and secretion of IgM in WM effected patients [45].

D. Comparison of MVC with related gene selection techniques

Finally, Figure 2 illustrates a comparative study of the proposed multiview method and other existing gene selection methods from literature, namely,

1) Self organizing map (SOM) [46],
2) Random forest gene selection (RFGS) [47],
3) Support vector sampling technique (SVST) [47],
4) Neuro–fuzzy models NFM-1 and NFM-2 [3]

in terms of their ability to identify disease related genes. The disease related genes are obtained from the NCBI database, and matched with the enriched gene set obtained by various methods including multiview clustering. As seen in Figure 2 the MVC based method presents higher or same percentage of disease related genes in a cluster for all test datasets in comparison to other analysed methods, especially when combined with PPI information. This indicates the potential of the proposed multiview clustering based methods to achieve genes well known to be associated with the diseases in the three case studies presented.

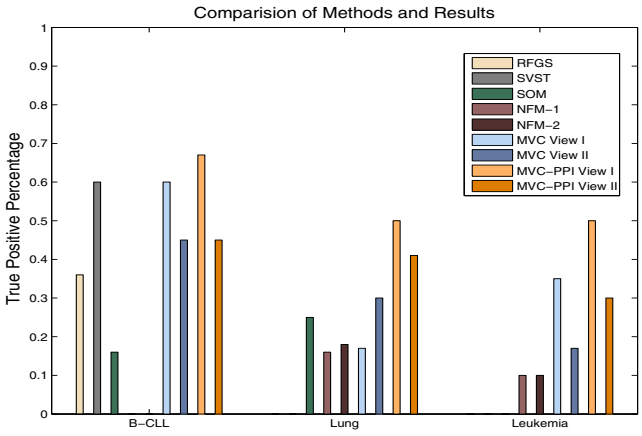

Fig. 2. Comparison in terms of biological significance obtained by the proposed multiview clustering based method (considering the two most enriched clusters separately) and other methods for datasets B-CLL (chronic lymphocytic leukemia), ILD (interstitial lung disease) and Leukemia (Waldenstrom's macro globulinemia). SOM: Self organizing map; RFGS: random forest gene selection; SVST: Support vector sampling technique; NFM-1 and NFM-2: neuro–fuzzy models; MVC: Multiview Clustering; MVC-PPI: Multiview Clustering with PPI information; View I and View II: first and second most enriched clusters obtained by MVC respectively. It is important to note that MVC-PPI performed better than MVC alone for all cases.

While performing the biological relevance analysis of the ranked gene sets, two important observations have been made:

(a) There exists a cluster having genes that exibit differences among various states of interest, but do not show any biologically enriched functional association (or any GO category association, or the count of functionally enriched attributes for this set of genes is zero).

(b) The gene clusters with high score value, or that are least significant (positive score) by MVC-PPI method, generates almost no GO attributes or functional association.

The first observation in (a) shows that our method is robust in clustering the genes, as it removes the outliers that show efficient class performance but no functional association. From the second observation (b) we can conclude that the incorporation of the topological properties of PPI network in ranking genes is an efficient approach in understanding the genetic conditions underlying the disease.

The observations made in Sections IV-A and IV-B lead to conclude that our proposed method of MVC and MVC after enrichment with PPI network show improved result in terms of gene selection performance. Our method identifies more biologically relevant genes in terms of their relevance with the disease, as well as, shows strongly connected regions in the PPI network.

V. Conclusion

The goal of explorative data analysis in bioinformatics is to find the underlying structure of the data, which may be multi-faceted by nature. We address the problem of gene selection with a multiview perspective. The proposed graph based multiview gene selection algorithm enriched with PPI network information attempts to address this problem by extracting multiple clustering views from high dimensional data having statistical, as well as biological relevance. PPI network knowledge is used to augment the views.

The effectiveness of the method is evaluated by comparing it with popular gene selection techniques. Through comparison of biological significance of the proposed method and some existing methods using selected genes, it has been found that our methodology provides improved performance in terms of classification accuracy and is able to identify gene subsets that are biologically relevant and functionally enriched. The results may facilitate to uncover the collective behavior of genes.

Acknowledgment

We would like to thank the financial support from CAPES, CNPq, FAPESP (grant 2011/50761-2), FAPESP-Microsoft (grant 2010/52138-8), eScience-PRP-USP and Indo-Brazil Collaborative Project, DST, Govt. of India and Govt. of Brazil.

References

[1] K. Yang, Z. Cai, J. Li, and G. Lin, "A stable gene selection in microarray data analysis," *BMC bioinformatics*, vol. 7, no. 1, p. 228, 2006.

[2] X. Xu and A. Zhang, "Selecting informative genes from microarray dataset by incorporating gene ontology," in *Bioinformatics and Bioengineering, 2005. BIBE 2005. Fifth IEEE Symposium on*. IEEE, 2005, pp. 241–245.

[3] A. Ghosh, B. C. Dhara, and R. K. De, "Selection of genes mediating certain cancers, using a neuro-fuzzy approach," *Neurocomputing*, vol. 133, no. 0, pp. 122 – 140, 2014.

[4] I. Trajkovski, N. Lavrač, and J. Tolar, "Segs: Search for enriched gene sets in microarray data," *Journal of biomedical informatics*, vol. 41, no. 4, pp. 588–601, 2008.

[5] B. H. Berg, C. Thanthiriwatte, P. Manda, and S. M. Bridges, "Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data," *BMC bioinformatics*, vol. 10, no. Suppl 11, p. S9, 2009.

TABLE II. PERCENTAGE OF ENRICHED ATTRIBUTES IN PROPOSED MVC-PPI MULTIVIEW GENE SELECTION: MULTIVIEW CLUSTERING VIEW (MVC-V), PPI NETWORK INTEGRATED MULTIVIEW CLUSTERING VIEW (MVCPPI-V), GENE COUNT (GC), GENE SCORE (GS), ENRICHED ATTRIBUTES (EA) – FOR DATASETS B-CLL (CHRONIC LYMPHOCYTIC LEUKEMIA), ILD (INTERSTITIAL LUNG DISEASE) AND LEUKEMIA (WALDENSTROM'S MACRO GLOBULINEMIA).

Test data	MVC-V GC	MVCPPI-V GS	MVCPPI-V GC	Enriched Attributes ID	No. of EA
B-CLL	42	< -3	13	GO:0038093, GO:0002764, GO:0050778, GO:0050776, GO:0002253, GO:0045087, GO:0048584, GO:004672, GO:0031401, GO:0032270, GO:0051247, GO:0006464, GO:0036211, GO:0043412, GO:0044267	15
		< -4	9	GO:0038093, GO:0002253, GO:0002764, GO:0050778, GO:0050776, GO:0045087, GO:0048584, GO:0002768, GO:0031401, GO:0044267, GO:0006464, GO:0036211	12
ILD	64	< -3	13	GO:0071359, GO:0000979, GO:0043331, GO:0007259, GO:0070887	5
		< -5	5	GO:0071359, GO:0000979, GO:0001046, GO:0043331, GO:0001047, GO:1901699, GO:0014070, GO:0070887	8
Leukemia	99	< 0	46	GO:0044424, GO:0005622, GO:0001654, GO:0043226, GO:0043229, GO:0044446, GO:0044422, GO:0071840	8
		< -2	32	GO:0044424, GO:0005622, GO:2000026, GO:0022008, GO:0071840	5

TABLE III. DESCRIPTION OF ENRICHED ATTRIBUTES PRESENTED BY THE MOST ENRICHED GENE CLUSTER BY THE PROPOSED MVC-PPI MULTIVIEW METHOD FOR DATASETS B-CLL (CHRONIC LYMPHOCYTIC LEUKEMIA), ILD (INTERSTITIAL LUNG DISEASE) AND LEUKEMIA (WALDENSTROM'S MACRO GLOBULINEMIA). THE COLUMN (%) SHOWS THE PERCENTAGE OF GENES FROM THE MOST ENRICHED CLUSTER SHARING THE ENRICHED ATTRIBUTES.

Test Data	EA ID	EA Name	%
B-CLL	GO:0044267	cellular protein metabolic process	77
	GO:0006464	cellular protein modification process	69.2
	GO:0036211	protein modification process	69.2
	GO:0043412	macromolecule modification	69.2
	GO:0048584	positive regulation of response to stimulus	54
	GO:0031401	positive regulation of protein modification process	54
	GO:0050776	regulation of immune response	46.2
	GO:0045087	innate immune response	46.2
	GO:0032270	positive regulation of cellular protein metabolic process	46.2
	GO:0051247	positive regulation of protein metabolic process	46.2
	GO:0002768	immune response-regulating cell surface receptor signaling pathway	44.4
	GO:0002764	immune response-regulating signaling pathway	38.5
	GO:0050778	positive regulation of immune response	38.5
	GO:0002253	activation of immune response	38.5
	GO:004672	protein kinase activity	38.5
	GO:0038093	Fc receptor signaling pathway	30.8
ILD	GO:0070887	cellular response to chemical stimulus	100
	GO:1901699	cellular response to nitrogen compound	60
	GO:0014070	response to organic cyclic compound	60
	GO:0071359	cellular response to dsRNA	40
	GO:0000979	RNA polymerase II core promoter sequence-specific DNA binding	40
	GO:0043331	response to dsRNA	40
	GO:0001046	core promoter sequence-specific DNA binding	40
	GO:0001047	core promoter binding	40
	GO:0007259	JAK-STAT cascade	23.1
Leukemia	GO:0044424	intracellular part	95.7
	GO:0005622	intracellular	95.7
	GO:0043229	intracellular organelle	89.2
	GO:0044446	intracellular organelle part	71.72
	GO:0044422	organelle part	71.72
	GO:0043226	organelle	31.5
	GO:0071840	cellular component organization or biogenesis	31.3
	GO:2000026	regulation of multicellular organismal development	31.3
	GO:0022008	neurogenesis	31.3
	GO:0001654	eye development	15.2

[6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[7] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[8] Y. Cun and H. Fröhlich, "Network and data integration for biomarker signature discovery via network smoothed t-statistics," *PloS one*, vol. 8, no. 9, p. e73074, 2013.

[9] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative ap-

proaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[10] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[11] C. Wu, J. Zhu, and X. Zhang, "Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes," *BMC bioinformatics*, vol. 13, no. 1, p. 182, 2012.

[12] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease." *Nature reviews. Genetics*, vol. 12, no. 1, pp. 56–68, jan 2011.

[13] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "Dada: Degree-aware algorithms for network-based disease gene prioritization." *BioData mining*, vol. 4, no. 1, 2011.

[14] H. Rahmani, H. Blockeel, and A. Bender, "Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks," in *Prooceedings of the Fifth International Workshop on Machine Learning in Systems Biology*, 2011, pp. 68–72.

[15] M. R. Sehhati, A. M. Dehnavi, H. Rabbani, and S. H. Javanmard, "Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence," *Journal of medical signals and sensors*, vol. 3, no. 2, p. 87, 2013.

[16] W. Zhang, F. Sun, and R. Jiang, "Integrating multiple protein-protein interaction networks to prioritize disease genes: a bayesian regression approach," *BMC bioinformatics*, vol. 12, no. Suppl 1, p. S11, 2011.

[17] X. Lu, Y. Deng, L. Huang, B. Feng, and B. Liao, "A co-expression modules based gene selection for cancer recognition," *Journal of theoretical biology*, 2014.

[18] Y. Cui, X. Z. Fern, and J. G. Dy, "Non-redundant multi-view clustering via orthogonalization," *Proc. 7th IEEE International Conference on Data Mining (ICDM'07)*, pp. 133 – 142, 2007.

[19] Y. Xiao, T.-H. Hsiao, U. Suresh, H.-I. H. Chen, X. Wu, S. E. Wolf, and Y. Chen, "A novel significance score for gene selection and ranking," *Bioinformatics*, vol. 30, no. 6, pp. 801–807, 2014.

[20] I. S. Kohane, A. J. Butte, and A. Kho, *Microarrays for an integrative genomics*. MIT press, 2002.

[21] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments," *Statistica sinica*, vol. 12, no. 1, pp. 111–140, 2002.

[22] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 129–136.

[23] T. Rama and P. Kolachina, "How good are typological distances for determining genealogical relationships among languages?" in *COLING (Posters)*. Citeseer, 2012, pp. 975–984.

[24] S. Fält, M. Merup, G. Gahrton, B. Lambert, and A. Wennborg, "Identification of progression markers in b-cll by gene expression profiling," *Experimental hematology*, vol. 33, no. 8, pp. 883–893, 2005.

[25] J.-H. Cho, R. Gelinas, K. Wang, A. Etheridge, M. G. Piper, K. Batte, D. Dakhlallah, J. Price, D. Bornman, S. Zhang *et al.*, "Systems biology of interstitial lung diseases: integration of mrna and microrna expression changes," *BMC medical genomics*, vol. 4, no. 1, p. 8, 2011.

[26] N. Gutierrez, E. Ocio, J. de Las Rivas, P. Maiso, M. Delgado, E. Ferminan, M. Arcos, M. Sanchez, J. Hernandez, and J. San Miguel, "Gene expression profiling of b lymphocytes and plasma cells from waldenström's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals," *Leukemia*, vol. 21, no. 3, pp. 541–549, 2007.

[27] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, and A. Venugopal, "Human Protein Reference Database - 2009 update," *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D767–D772, 2009.

[28] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R. C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger, P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob, "The IntAct molecular interaction database in 2012." *Nucleic acids research*, vol. 40, no. Database issue, pp. D841–6, Jan. 2012.

[29] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2012 update." *Nucleic acids research*, vol. 40, no. Database issue, pp. D857–61, Jan. 2012.

[30] G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth, "Next generation software for functional trend analysis," *Bioinformatics*, vol. 25, no. 22, pp. 3043–3044, 2009.

[31] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec 2004.

[32] C. Boutsidis *et* al., "Unsupervised feature selection for principal components analysis," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, Aug 2008, pp. 61–69.

[33] G. K. Alderton and Y. Bordon, "Tumour immunotherapyleukocytes take up the fight," *Nature Reviews Immunology*, vol. 12, no. 4, pp. 237–237, 2012.

[34] A. Palazón, J. Aragonés, A. Morales-Kastresana, M. O. de Landázuri, and I. Melero, "Molecular pathways: hypoxia response in immune cells fighting or promoting cancer," *Clinical Cancer Research*, vol. 18, no. 5, pp. 1207–1213, 2012.

[35] D. Mertens and S. Stilgenbauer, "Prognostic and predictive factors in patients with chronic lymphocytic leukemia: Relevant in the era of novel treatment approaches?" *Journal of Clinical Oncology*, vol. 32, no. 9, pp. 869–872, 2014.

[36] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, K. S. Wang, F. Mandelli, R. Foà, and J. Ritz, "Gene expression profiles of b-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation," *Clinical cancer research*, vol. 11, no. 20, pp. 7209–7219, 2005.

[37] A. L. Pritchard and N. K. Hayward, "Molecular pathways: mitogen-activated protein kinase pathway mutations and drug resistance," *Clinical Cancer Research*, vol. 19, no. 9, pp. 2301–2309, 2013.

[38] M. Braoudaki, G. I. Lambrou, K. Vougas, K. Karamolegou, G. T. Tsangaris, and F. Tzortzatou-Stathopoulou, "Protein biomarkers distinguish between high-and low-risk pediatric acute lymphoblastic leukemia in a tissue specific manner," *Journal of hematology & oncology*, vol. 6, no. 1, p. 52, 2013.

[39] R. P. Daniele, J. A. Elias, P. E. Epstein, and M. D. Rossman, "Bronchoalveolar lavage: role in the pathogenesis, diagnosis, and management of interstitial lung disease," *Annals of internal medicine*, vol. 102, no. 1, pp. 93–108, 1985.

[40] R. Vassallo, "Viral-induced inflammation in interstitial lung diseases." in *Seminars in respiratory infections*, vol. 18, no. 1, 2003, pp. 55–60.

[41] O. Matsuno, "Drug-induced interstitial lung disease: mechanisms and best diagnostic approaches," *Respir Res*, vol. 13, no. 1, p. 39, 2012.

[42] K. Kuwano, "Involvement of epithelial cell apoptosis in interstitial lung diseases." *Internal medicine (Tokyo, Japan)*, vol. 47, no. 5, pp. 345–353, 2007.

[43] E. Kimby, S. P. Treon, A. Anagnostopoulos, M. Dimopoulos, R. Garcia-Sanz, M. A. Gertz, S. Johnson, V. LeBlond, J.-P. Fermand, D. G. Maloney *et al.*, "Update on recommendations for assessing response from the third international workshop on waldenström's macroglobulinemia," *Clinical Lymphoma and Myeloma*, vol. 6, no. 5, pp. 380–383, 2006.

[44] S. P. Treon, L. Xu, G. Yang, Y. Zhou, X. Liu, Y. Cao, P. Sheehy, R. J. Manning, C. J. Patterson, C. Tripsas *et al.*, "Myd88 l265p somatic mutation in waldenström's macroglobulinemia," *New England Journal of Medicine*, vol. 367, no. 9, pp. 826–833, 2012.

[45] M. J. Stone and V. Pascual, "Pathophysiology of waldenströms macroglobulinemia," *haematologica*, vol. 95, no. 3, pp. 359–364, 2010.

[46] D. G. Covell, A. Wallqvist, A. A. Rabow, and N. Thanki, "Molecular classification of cancer: Unsupervised self-organizing map analysis of gene expression microarray data," *Molecular cancer therapeutics*, vol. 2, no. 3, pp. 317–332, 2003.

[47] A. H. Chen, Y.-W. Tsau, and C.-H. Lin, "Novel methods to identify biologically relevant genes for leukemia and prostate cancer from gene expression profiles," *BMC genomics*, vol. 11, no. 1, p. 274, 2010.