Prior Information Based Bayesian Infinite Mixture Model

Zhen Hu[1], Siva Sivaganesan[2] and Mario Medvedovic[3]

[1.]Department of Computer Science, University of Cincinnati; [2.]Department of Mathematical Sciences; [3.]Department of Environmental Health, University of Cincinnati
Kettering Room 318, 3223 Eden Ave., Cincinnati, OH, 45267
huze@mail.uc.edu; sivagas@ucmail.uc.edu;Mario.Medvedovic@uc.edu

Abstract

Unsupervised learning methods have been tremendously successful in extracting knowledge from genomics data generated by high throughput experimental assays. However, analysis of each dataset in isolation without incorporating potentially informative prior knowledge is limiting the utility of such procedures. Here we present a novel probabilistic model and computational algorithm for semi-supervised learning from genomics data. The probabilistic model is an extension of the Bayesian semi-parametric Gaussian Infinite Mixture Model (GIMM) and training of model parameters is performed using Markov Chain Monte Carl algorithm. The utility of the procedure in improving precision of cluster analysis by incorporating prior information is demonstrated in a simulation study and the analysis of the real world genomics data.

1. Introduction

Identifying groups of co-expressed genes through cluster analysis has been successfully used to elucidate affected biological pathways and postulate transcriptional regulatory mechanisms [1,21]. Virtually all classical clustering algorithms [4,24,26,30], as well a multitude of brand new procedures [9,27] have been applied in the context of clustering genomics data. The integration of biological knowledge in such analyses has been most commonly facilitated by assessing the enrichment of clusters with genes from pre-defined functionally coherent gene lists ("functional categories") [12]. A systematically different approach is to use semi-supervised learning to incorporate information as the prior knowledge into the clustering algorithm itself [1,4,10,11,22,25]

Here we describe a novel probabilistic framework, Prior Gaussian Infinite Mixture Model (PGIMM), for semi-supervised learning of clusters of co-expressed genes. PGIMM is an extension of Bayesian semi-parametric Gaussian Infinite Mixture Model (GIMM). The use of GIMM [18] helps us avoid specifying the number of clusters [3,19] before the initiation of clustering procedure. For PGIMM, we postulate a generative probabilistic model for the observed genomics data in terms of a Bayesian network. The leaning is performed by estimating the parameters of the model via Gibbs sampler.

The paper is organized as follows: we define the probabilistic model and the computational algorithm in section 2; describe different sources of prior information in section 3; present results of a simulation study in section 4 and describe the application in clustering cancer genomics data in section 5.

2. Probabilistic Model And Computational Algorithm

2.1 Probabilistic model

Suppose we have a dataset for T data points and M features, $\mathbf{X}$ is the $T \times M$ matrix where x_{ij} is the values of data point i in feature j. Accordingly, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$. In our applications x_{ij} is the expression level of the i^{th} gene for the j^{th} experimental condition, then $\mathbf{x}_i=(x_{i1}, x_{i2},\ldots, x_{iM})$ denotes the complete expression profile for the i^{th} gene. Each data point can be viewed as being generated by one of the Q underlying patterns represented by probability distributions. Data points generated by the same pattern form a **cluster** of similar objects. If c_i is the classification variable indicating the probability pattern that generates the i^{th} data point ($c_i=q$ means that the i^{th} data point was generated by the q^{th} pattern), then a **clustering** is defined by a set of classification variables for all data points $\mathbf{C}=(c_1, c_2,\ldots, c_T)$. The probability distributions generating data points in a cluster are assumed to be multivariate M-dimensional Gaussian distributions. That is, $c_i=q$ implies that $\mathbf{x}_i \sim N_M(\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q)$, where $\boldsymbol{\mu}_q$ is the mean and $\boldsymbol{\Sigma}_q$ is the variance-covariance matrix of the M-dimensional multivariate Gaussian distribution.

The probabilistic model describing the distribution of the data (i.e. observed expression profiles $\boldsymbol{x}_i$) is given in the form of a Bayesian network. Dependencies between various model parameters and the data are defined by the Directed Acyclic Graph (DAG) in Figure 1. Nodes in the network represent random variables and arcs define the independence structure of the joint probability distribution function. An arc drawn between a node and a dotted rectangle containing multiple nodes implies that it is the parent node for all nodes within the rectangle.

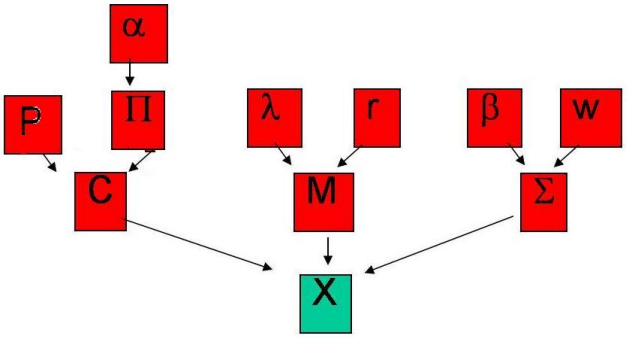


Figure 1

Figure 1 shows the Bayesian networks of PGIMM: M denotes the means of each cluster, Σ denotes the variance and C denotes the cluster indication for each data point. Variables α, Π, C represent relative preference between clusters that is based on the size of each cluster [19]. The variable at the outmost level which contains β, w, λ, r are hyper parameters. The prior information incorporated is represented by variable P. As we can conclude from Figure 1, P is only related with cluster labels C.

Assuming that the probability distribution of any node is independent of its non-descendants if values of the parent nodes are given (Directed Markov Assumption), the joint probability distribution of all parameters and data is given by the product of the local probability distributions of individual random variables given their parents.

$$p(X,C,M,S,\alpha,\lambda,\tau,\beta,\phi,P) = p(X \mid C,M,\Sigma)p(C \mid \alpha,P)p(S \mid \beta,\phi) \\ p(M \mid \lambda,\tau)p(\alpha)p(\lambda)p(\tau)p(\beta)p(\phi)p(P)$$

where $\boldsymbol{M}=(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_Q)$ and $\boldsymbol{S}=(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_Q)$ are the set of all mean vectors and variance-covariance matrices defining expression patterns, **P** is the prior clustering structure. As specified above,

$$p(X_i \mid c_i = q, M, S) = f_N(x_i \mid \mu_q, \Sigma_q)$$

where $f_N(.|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the multivariate Gaussian probability distribution function with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

The learning of the clustering structure is achieved by estimating the marginal posterior distribution of the cluster labels **C** given data and prior knowledge

$$p(C \mid X,P) = \int p(M,S,\alpha,\lambda,\tau,\beta,\phi \mid X,P)d(M,S,\alpha,\lambda,\tau,\beta,\phi \mid X,P)$$

Our new model deviates from the previously described GIMM model in the way that prior probability distribution of the clustering variable C is specified. Suppose that prior information consists of a clustering structure $P=(p_1, p_2,\ldots, p_T)$, where p_i is the prior clustering label for data point i (ie $p_i=z$ if i^{th} data point was clustered in the prior cluster z). The prior probability of C given α and P is then given by:

$$p(c_i = j \mid c_{-i},\alpha,P) = b\frac{n_{-i,j}}{T-1+\alpha} \times e^{PC_{i,j}}$$

$$p(c_i \neq c_j, i \neq j \mid c_{-i},\alpha,P) = b\frac{\alpha}{T-1+\alpha} \qquad (1)$$

Where $n_{-i,j}$ denotes the number of data points in cluster j without counting x_i. ***PC*** is defined as:

$$PC_{i,j} = \frac{n_{jz} \times T}{n_z \times n_j} \qquad (2)$$

where n_{jz} is number of data points in j^{th} cluster which has the same prior clustering assignment with i^{th} data point (p_i), n_j is total number of data points in j^{th} cluster and n_z is total number of data points whose prior clustering assignments are the same with i^{th} data point. Intuitively, ***PC*** responses to measure the relative consistency between potential clusters and prior clustering assignments and new clusters are generated without considering prior clustering.

2.2 Learning Algorithm

The semi-supervised learning of the clustering structure in the data proceeds by estimating the posterior distribution of the parameters in the model given data and prior knowledge $p(M,S,\alpha,\lambda,\tau,\beta,\phi \mid X,P)$ using the Gibbs sample algorithm. The Gibbs sampler iteratively draws values from the conditional posterior probability distributions for each random variable in the model given all other variables and the data. The resulting Markov Chain converges to the joint posterior distribution. The posterior conditional distributions for most variables in the model remain the same as previously described [18]. However, the conditional posterior probabilities of assigning data points to clusters are modified based on incorporating the prior information and follows from Equation (1).

$$p(\mathrm{c_i} = \mathrm{j} \mid \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_j, \sigma_j^2) = \mathrm{b}\frac{\mathrm{n}_{-i,j}}{\mathrm{T}-1+\alpha}\ f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \sigma_j^2\mathbf{I}) \times \mathbf{e}^{\mathbf{PC}_{i,j}}$$

$$p(\mathrm{c_i} \neq \mathrm{c_j}, \mathrm{j} \neq \mathrm{i} \mid \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_x, \sigma_x^2) \qquad (3)$$

$$= \mathrm{b}\frac{\alpha}{\mathrm{T}-1+\alpha}\int f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \sigma_j^2\mathbf{I})\ p(\boldsymbol{\mu}_j, \sigma_j^2 \mid \boldsymbol{\lambda}, \mathrm{r}^{-1})\, \mathrm{d}\boldsymbol{\mu}_j \mathrm{d}\sigma_j^2$$

Initialization: The learning procedure started with setting every data points assigned into one same cluster. That is:

$$C_1^0 = C_2^0 = ... = C_T^0 = 1$$

Corresponding clustering parameters (μ, σ) are calculated and sampled. The number of cluster (q) is set to one. Base on the values of prior variables P , the values of PC are set as:

$$PC_{1,1}^0 = PC_{2,1}^0 = ... = PC_{T,1}^0 = 1$$

The global mean and variance (μ_x, σ_x^2) are calculated from the data and all other model parameters are initialized by sampling from prior distributions.

Iterations:

1. Update cluster allocation variables

Given all parameters' value after k^{th} iterations $(C^k, M^k, \Sigma^k, \lambda^k, r^k, \beta^k, w^k)$ the Gibbs sampler updates each parameter in the $(k+1)^{th}$ iteration by drawing values from the posterior conditional distributions functions of each parameters given all current values of other parameters and prior knowledge. Cluster allocation variables C^{k+1} are updated first accordingly based on Equation 3, conditional on the $(M^k, \Sigma^k, \lambda^k, r^k, \beta^k, w^k, P, X)$.

2. Update means and variances for all clusters

(M^{k+1}, Σ^{k+1}) are drawn from their posterior conditional distributions, given $(C^{k+1}, \lambda^k, r^k, \beta^k, w^k)$ Rest of the variables in the model $(\lambda^{k+1}, r^{k+1}, \beta^{k+1}, w^{k+1})$ are then drawn from their respective posterior distributions given the current values of all other variables and the data $(M^{k+1}, \Sigma^{k+1}, P, X)$ [18]. The value of α is set to 1[19].

Estimated posterior marginal distributions of clustering allocations are summarized by calculating posterior pair-wise probabilities (PPP) of co-groupings as the proportion of Gibbs sampler cycles in which two data points were grouped together after enough 'burn-in' steps. Hierarchical clusterings of genes and samples were created by using PPPs as the similarity measure and applying the complete linkage agglomeration method.

3. Prior Information Sources

The prior information to the PGIMM algorithm is provided in the form of a partition of the data point into a certain number of clusters. In principle, any informative source of meaningful biological information which can be used to create such partition can be used as source of the prior information. Here we demonstrate the use of prior clustering structure based on, (i) independent gene expression datasets; (ii) computationally predicted transcription factor motif binding information; and (iii) Gene Onologies (GO)[17].

3.1 Clustering of independent expression dataset

One obvious source of prior knowledge that could be used in clustering gene expression data is the clustering results obtained in a different, already analyzed dataset. For example, numerous data sets have been generated that profile transcriptomes of primary breast cancer tumors. One way to integrate information in these independent, but related datasets is to incorporate the clustering structure uncovered in the unsupervised analysis in one of the datasets to as the prior information into the semi-supervised analysis of the other dataset. Unfortunately, evaluating benefits of adding such prior information is difficult since we do not know the correct answer to the problem. Here we use one such dataset [15] to demonstrate the utility of such approach when the prior clustering is perfectly representative of the unknown "truth". We first cluster the complete dataset and designate the results of the cluster analysis as both the "gold standard" and the prior information. Then we select a smaller subset of the samples and assess our ability to re-create the "correct" clustering by using incorporating the prior adopting PGIMM methodology.

3.2 Motif Matching Measurements

It is generally assumed that the co-expression patterns revealed in a cluster analysis are reflective of the underlying gene expression regulatory mechanisms. Binding of transcription factors (TF) to their cognate DNA motifs within gene promoter regions is one of the most important mechanisms employed by a cell in regulating gene expression levels. Here we demonstrate the use of computationally derived information about the putative existence of TF binding motifs within gene promoters as the prior knowledge which is, further, integrated into our PGIMM framework.

For each of the 304 human transcription factors with at least one PWM in the *Transfac* [6] version 12.1 database, we scored genes as to how likely they were to have such a motif within 10kb of their TSS. Suppose that θ_{jl} is the l^{th} PWM defining the DNA binding motif of length $L(\theta_{jl})$ associated with the j^{th} TF and S_{ijx} is any DNA fragment of length $L(\theta_{jl})$ in the 10 kbp DNA region around the TSS for the i^{th} gene. We first score all such fragments for all genes based on the simple string matching algorithm using the IUPAC summaries of PWMs. For fragments having the string matching score higher than the median, we calculate the score measuring the likelihood of S_{ijx} being the binding site for the j^{th} TF as

$$R_{ijlx} = \log 2\left(\frac{p(S_{ijx} \mid \theta_{jl})}{p(S_{ijx} \mid \theta_0)}\right) * \frac{1}{\sqrt{L(\theta_{jl})}}$$

where $p(S_{ijx}|\theta_{jl})$ is the probability of S_{ijx} being generated by the product multinomial model with the PWM θ_{jl} and $p(S_{ijx}|\theta_0)$ is the probability of S_{ijx} being generated by the background 3rd order Markov chain with the transition matrix θ_0 estimated using 10 kb fragments around TSS for all genes in the genome. The gene-specific scores for the i^{th}

gene and the j^{th} TF are then set to the maximum likelihood score among all DNA fragments for this gene. Genes without a single fragment in the top 50% based on the initial string matching algorithm are assigned a score of 0.

3.3 GO category

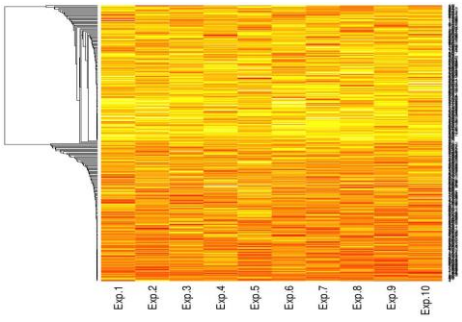
Gene Ontologies (GO) are prototypical groupings of functionally-related genes. They are commonly used in functionally annotating results of the cluster analysis [12]. Assuming that the functionally-related genes are more likely to be co-regulated and thus co-expressed than functionally-unrelated genes, we use co-occurrence in the same GO categories as the prior information within the PGIMM framework. Due to a complex structure of Gene Ontologies as a directed acyclic graph of categories such that parental nodes contain genes of all its descendants, clustering genes based on their co-membership in GO categories is a non-trivial problem [12, 29]. We use the information theory based similarity measurements.

$$S_{ij} = \max_{k \in GO} \left(1 - \frac{\log_2 n_k}{\log_2 N}\right) \quad (6)$$

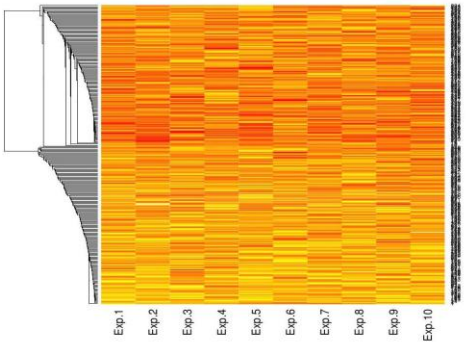
here n_k is number genes annotated in k^{th} GO categories, N is the total number of genes annotated in the GO database. and hierarchical clustering to construct clusterings of genes based on their GO annotations.

4. Simulation Study

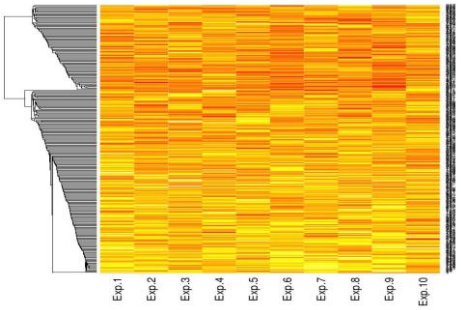
The performance of our computational framework was first tested in a simulation study. We simulated a series of datasets based on a simple 2-cluster clustering structure with 150 genes in each cluster (300 rows) measure across 10 samples (columns). All 150x10 values in the first cluster were randomly sampled from a Gaussian probability distribution with the mean 0 and the fixed variance, and all values in the second cluster were generated from a Gaussian distribution with the mean equal to one and the same fixed variance. By increasing the variance used to simulate gene expression levels, we increased the "difficulty" of the clustering problem. For each experimental scenario (ie each variance level, we generated 50 synthetic datasets



(a)



(b)



(c)

Figure 2

Figure 2 shows the example datasets and the resulting cluster analysis for three different simulation settings: (i) Figure 2(a) shows data with standard deviation 1.0 and (ii) Figure 2(b) with standard deviation 1.5 and (iii) Figure 2(c) with standard deviation 2.0. As expected, the higher the variance the more difficult to reconstruct the clustering structure correctly.

For each simulation scenario, we constructed average Receiver Operating Characteristic (ROC) curves. ROC for hierarchical clustering was constructed by cutting the tree at heights of each branching depths to create clusterings with every possible number of clusters. For a fixed number of clusters, a pair of genes belonging to the same cluster was assumed to be a "true positive" if they both belonged to the same original cluster under the simulation design, and it was considered to be a "false positive" if they both belonged to different original cluster under the simulation design. True and false positive rates were then obtained by dividing the number of true/false positives with the total number of gene pairs belonging to the same cluster and the total number of gene pairs not belonging to different clusters. When the number of clusters is equal to the number of genes and all genes are placed in their own individual clusters, both true and false positive rates are equal to zero. An ROC curve is defined when we reduce the number of clusters and both true and false positive rates increase. At the extreme when all genes are placed in the same cluster, both true and false positive rates are equal to one. By averaging FPRs and TPRs across 50 datasets for each number of clusters, we obtained the average ROCs for each scenario depicted in Figure 3. The higher the area

under the average ROC curve, the better is the performance of the clustering algorithm in reconstructing the original clustering structure.

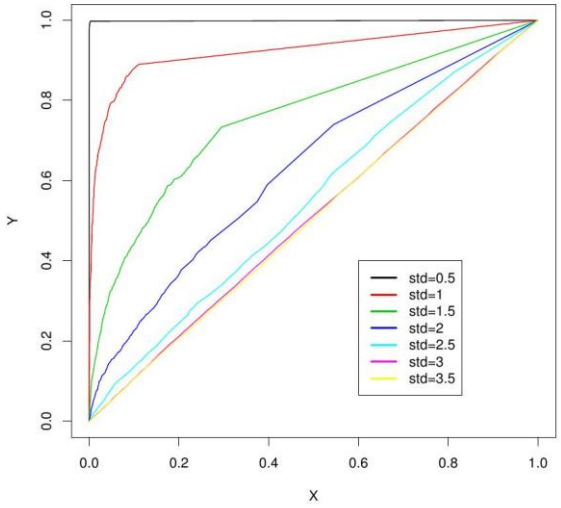
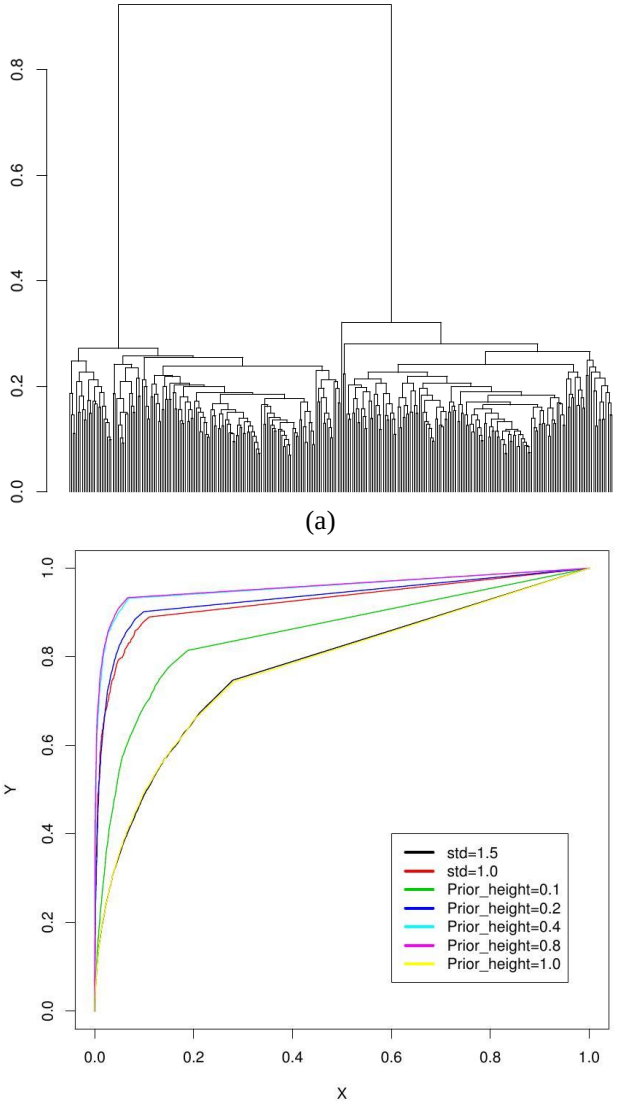


Figure 3

There are seven lines in the figure: black line represents the clustering performance when variance is smallest (std=0.5) and there is virtually no false positive pairs in any of the 50 simulated datasets. On the other hand the average ROC curve for the two highest noise levels is overlapping the 45 degree line indicating that the clustering algorithm is not better than a random assignment into different clusters. To assess the effects of incorporating the prior knowledge we use the variance structure that is in the middle of our "dynamic range" (std=1.5) allowing us to asses datasets both improvements and reduction in precision with designed cluster variance 1.0 and 2.0 as priors.

Figure 4(a) shows the hierarchical tree generated by simple GIMM algorithm for dataset with variance 1.0 (also referred to as the prior dataset). Prior clusterings with different number of clusters are obtained by cutting the tree at 0.1, 0.2, 0.4, 0.8 and 1.0 levels and examine the improvement in accuracy using the PGIMM algorithm on the same simulated datasets. The best ROC curve (Figure 4(b)) corresponds to the prior clusterings with two clusters obtained in this case by cutting the prior tree at 0.4 and 0.8 levels. This is somewhat intuitive given that the correct clustering structure has two clusters. The gains in accuracy are progressively diminished as the tree is cut at lower levels creating prior clusterings with too many clusters, but are still significant for levels 0.2 and 0.1. For example, with the prior clustering structure obtained by cutting the tree at 0.2 level, the average ROC curve is higher than for the clustering of the dataset simulated at the lower noise level (std=1) generated without the prior knowledge. When the prior clustering structure consists of putting all data points in the same cluster (cut-off=1), there was not additional information provided to the PGIMM algorithm and the clustering is the same as without the prior knowledge.

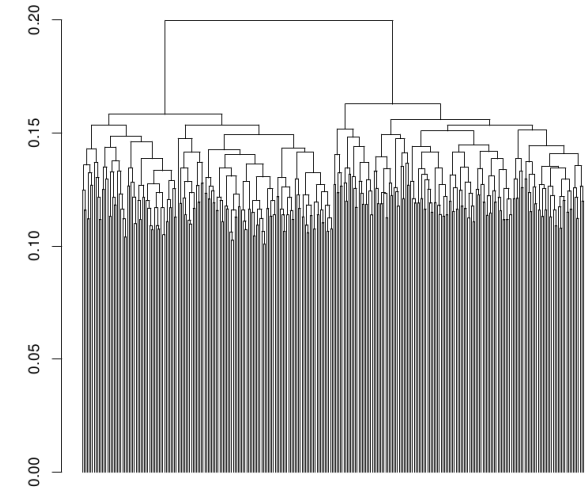


(a)

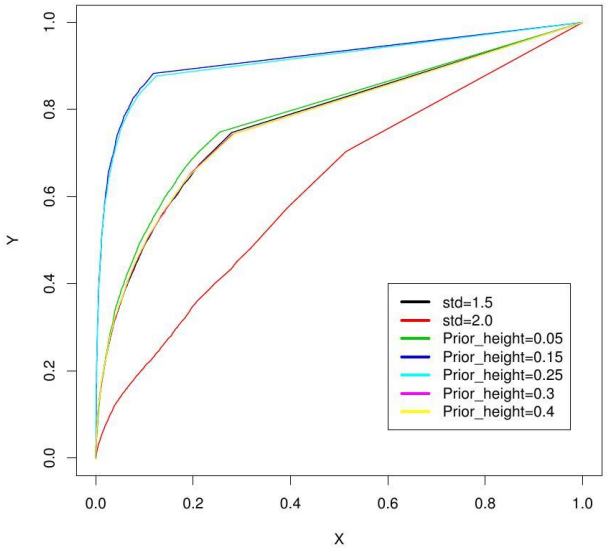
(b)

Figure 4

Figure 5(a) shows the hierarchical tree generated by GIMM for dataset with higher noise level (std=2.0). To generate prior clusterings, the tree is cut at 0.05, 0.15, 0.25, 0.3 and 0.4 levels. Similarly as in the previous example, adding prior information either improves the accuracy, or it does not decrease it (Figure 5(b)) even though the prior information is very noisy. In summary, PGIMM algorithm effectively incorporates the prior clustering knowledge that is consistent with the correct clustering structure. Even when such prior information is rather vague, it still improves the accuracy of the semi-supervised analysis over the simple unsupervised analysis. Cutting the prior hierarchical tree at different levels will affect the level of gain in precision, but the accuracy is never worse than in the baseline unsupervised analysis.



(a)



(b)

Figure 5

5. Cancer data application

We examined the problem of improving the accuracy of clustering genes based on their expression levels across independent primary breast cancer samples using the PGIMM algorithm and different source of prior knowledge. The use one of the commonly used microarray datasets [15] consisting of 251 primary breast tumors analyzed using Affymetrix human 133A gene chip. The gold standard clustering of genes was constructed by the unsupervised analysis of all samples. Then, we construct 20 random subsets by drawing randomly each time 10 samples without replacement from the set of 251 samples

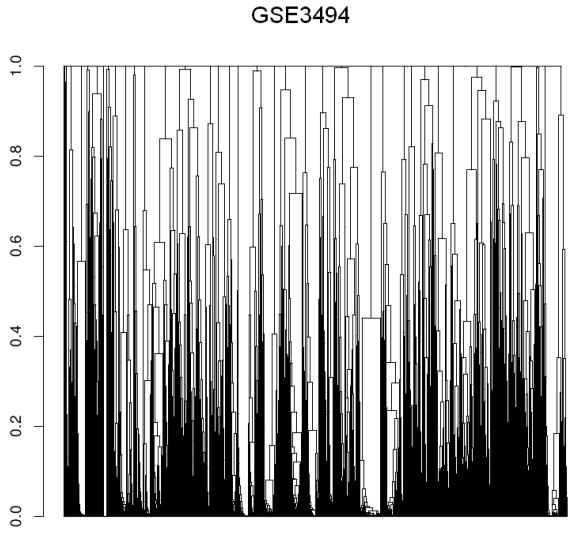


Figure 6

Figure 6 shows the gold-standard hierarchical clustering results on generated by GIMM using the whole dataset. The gold standard clusterings with different number of clusters were constructed by cutting this tree at 0.1, 0.2, 0.5, and 0.8 levels. Then we use three different types of prior information to overcome the small sample size in the 20 sub-sampled datasets and make the resulting clustering as close as possible to the gold-standard clusterings based on the whole dataset.

5.1 Perfect Priors

We first assess the performance of the PGIMM algorithm when using the gold standard clusterings as prior information. The purpose of these tests is to evaluate the appropriateness of our generative model for the real-world microarray data. In the simulation study we showed that when the data is generated in accordance to our model, the algorithm behaves as expected. Here we show that our model approximate the real world data sufficiently well for the algorithm to be effective. The number of clusters in prior clusterings for all three types of prior information are given in Table 1. The first column (perfect prior) corresponds to the gold standard clusterings in Figure 6

Height	Perfect Prior	Motif Priors	GO Priors
0.1	915	782	-
0.2	684	423	1154
0.5	237	116	440
0.8	82	61	79

Table 1

As in the simulation study each individual ROC curve is averaged over 20 random samples. To compare directly multiple priors for multiple gold standards, each average ROC curve was summarized by calculating the Area Under

the Curve (AUC). The random assignment corresponding to the 45 degree ROC line has AUC of 0.5.

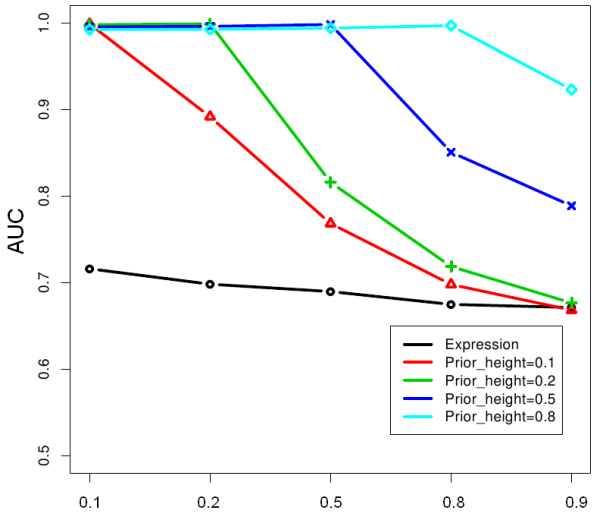


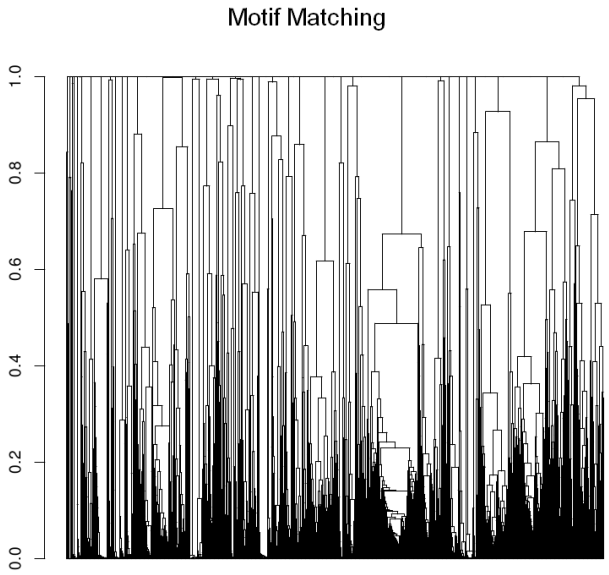
Figure 7

Figure 7 shows summarized AUCs for different cut-offs generating the gold standards (x-axis) and the use of these same gold standards as the prior information (different colored) lines. The black line shows the average performance of the simple unsupervised GIMM algorithm. Adding any kind of prior information within the PGIMM algorithm improves the performance over the unsupervised analysis. The performance is generally better when prior clusterings have fewer clusters with every situation in which the number of prior clusters is smaller than or equal to the gold standard, the accuracy is virtually perfect.

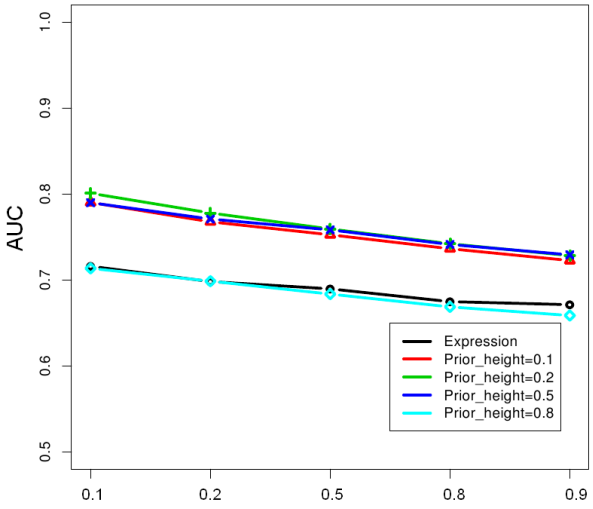
5.2 Motif Matching Priors.

We explored the use of computationally derived transcription factor binding scores described in section 3.2. instead of gold-standard priors. All datasets and the gold standard used to compare the produced clusterings are same as in the previous section. Binding scores for 5 transcription factors reported to be associated with the ERα binding and transcriptional regulation [13] were used to cluster genes. Figure 8(a) shows the hierarchical tree of the clustering generated by GIMM. Again, prior clusterins were created by cutting this tree at different levels. The distribution of the number of clusters in such clustering was similar to the distribution of the number of clusters in the gold-standard (second column in Table1).

Figure 8(b) shows the clustering performance of the PGIMM algorithm for different prior clustering and different gold standards (x-axis). In this case cutting the prior tree at intermediated levels worked better than cutting at the high level. Still, none of the prior clusterings have detrimentally affected the accuracy in any significant way.



(a)

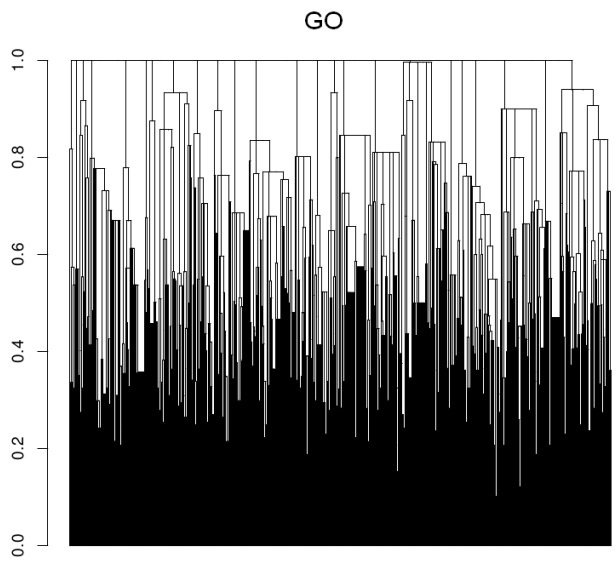


(b)

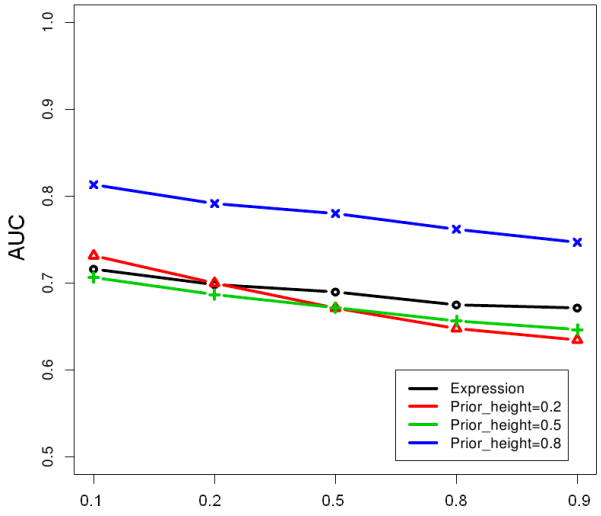
Figure 8

5.3 GO Priors.

We repeat the tests from the previous section using now the prior tree constructed based on the Gene Ontology annotations to construct the prior tree in Figure 9(a).



(a)



(b)

Figure 9

In this situation, the number of clusters obtained by cutting the prior tree were significantly higher for the lower cut-off levels (third column in Table 1), and the PGIMM performed slightly worse than the simple unsupervised in a couple of scenario (Figure 9(b)). However, while the improvements when using the prior clustering with the fewest clusters are significant, the slight deteriorations for the prior clusterings with large number of clusters seem to be well within the margin of error.

5. Conclusion

We demonstrate the utility of the PGIMM framework in performing semi-supervised analysis of genomics data. Unlike previously developed procedures, PGIMM does not require prior specification of the number of clusters. In the simulation study, we showed that even the very noisy prior information that is concordant with the correct clustering structure can significantly improve the accuracy of cluster analysis. These results were extended to the real-world data analysis utilizing the gold-standard clustering as the prior information. In the situation when the prior information comes from completely independent sources, and its correlation with the gold standard in uncertain, the PGIMM still offered significant improvements, but it was less robust with respect to selecting the number prior clusters. The future research work is needed to develop optimal ways to use the prior information when the gold standard is not known.

Acknowledgments

This work was funded by grants from National Human Genome Research Institute and National Library of Medicine (R01HG003749, R21LM009662).

References

[1]. Do JH and Choi DK, *Clustering approaches to identifying gene expression patterns from DNA microarray data.* Mol.Cells **25**: 279-288, 2008.

[2]. Dotan-Cohen D, Melkman AA, and Kasif S, *Hierarchical tree snipping: clustering guided by prior knowledge.* Bioinformatics **23**: 3335-3342, 2007

[3]. E.A. Gelfand and F.M.A. Smith, *Sampling-Based Approaches to Calculating Marginal Densities*, Journal of the American Statistical Association **85**(410): 398–409, 1990

[4]. Eisen MB, Spellman PT, Brown PO, and Botstein D, *Cluster analysis and display of genome-wide expression patterns.* Proc.Natl.Acad.Sci.U.S.A **95**: 14863-14868, 1998.

[5]. E.M. Conlon, X.S. Liu, J.D. Lieb and J.S. Liu, *Integrating regulatory motif discovery and genome-wide expression analysis*, Proc. Natl. Acad. Sci. U.S.A.**100**(6): 3339-3344, 2003

[6]. E. Wingender, et al, *The TRANSFAC system on gene expression regulation. Nucleic Acids Research*, **29**(1): 281-283, 2001

[7]. *Gene Expression Omnibus*: http://www.ncbi.nlm.nih.gov/geo/

[8]. *Genomic Portal: Integrative platform for acessing and analyzing genomics data*: http://www.eh3.uc.edu:8080/GenomicsPortals/

[9]. Herrero J, Valencia A, and Dopazo J, *A hierarchical unsupervised growing neural network for clustering gene expression patterns*. Bioinformatics. **17**: 126-136, 2001.

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

[11]. Huang D, Wei P, and Pan W, *Combining gene annotations and gene expression data in model-based clustering: weighted method*, OMICS. **10**: 28-39, 2006

[12]. J. M. Freudenberg, V. K. Joshi, Z. Hu and M. Medvedovic, *CLEAN: CLustering Enrichment Analysis*, BMC Bioinformatics **10**(234): 2009

[13]. J. S. Carroll, C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu and M. Brown, *Genome-wide analysis of estrogen receptor binding sites*, Nature Genetics, **38**: 1289-1297, 2006

[14]. Lee SI and Batzoglou S, *Application of independent component analysis to microarrays*. Genome Biol. **4**: R76, 2003.

[15]. LD. Miller, J. Smeds, J. George, VB. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, ET. Liu, et al., *From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival*, PNAS, **102**: 13550-13555, 2005

[16]. Liu X, Jessen W, Sivaganesan S, Aronow BJ, and Medvedovic M: *Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data*. BMC Bioinformatics, **8**(1): 283. 2007.

[17]. M. Ashburner, CA. Ball, JA. Blake, D. Botstein, H. Butler, JM. Cherry, AP. Davis, K. Dolinski, SS. Dwight, JT. Eppig, et al.: *Gene ontology: tool for the unification of biology*, Nature Genetics, **25**(1): 25-29, 2000

[18]. M. Medvedovic and S. Sivaganesan, *Bayesian infinite mixture model based clustering of gene expression profiles*. Bioinformatics **18:** 1194-1206, 2002

[19]. Neal. R. M, *Markov chain sampling methods for Dirichlet process mixture models*, Journal of Computational and Graphical Statistics, **9**: 249-265, 2000.

[20].P.Baldi and S.Brinak, *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inference of gene changes*, Bioinformatics **17:** 509-519, 2001

[21]. Slonim DK, *From patterns to pathways: gene expression data analysis comes of age*, Nat.Genet. **32** Suppl: 502-508, 2002

[22]. Sugato Basu , Mikhail Bilenko and Raymond J. Mooney, *A probabilistic framework for semi-supervised clustering*, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004

[23]. S. Srivastava, L. Zhang, R. Jin and C. Chan, *A novel method incorporating gene ontology information for unsupervised clustering and feature selection*, PLoS One, **3**(12): 2008

[24]. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, and Golub TR, *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc.Natl.Acad.Sci.U.S.A **96**: 2907-2912, 1999.

[25]. Tan MP, Smith EN, Broach JR, and Floudas CA, *Microarray data mining: a novel optimization-based approach to uncover biologically coherent structures.* BMC.Bioinformatics. **9**: 268, 2008.

[26]. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, and Church GM, *Systematic determination of genetic network architecture*. Nat.Genet. **22**: 281-285, 1999.

[27]. Tseng GC and Wong WH, *Tight clustering: a resampling-based approach for identifying stable and tight patterns in data*. Biometrics **61**: 10-16, 2005.

[28]. W. Pan, *Incorporating gene functions as priors in model-based clustering of microarray gene expression data*, Bioinformatics, **22**(7):795-801, 2006

[29]. Y. Lu, R. Rosenfeld, I. Simon, G. J. Nau and Z. Bar-Joseph, *A probabilistic generative model for GO enrichment analysis*, Nucleic Acids Research, **36**(17), 2008

[30]. Yeung KY, Fraley C, Murua A, Raftery AE, and Ruzzo WL, *Model-based clustering and data transformations for gene expression data*. Bioinformatics. **17**: 977-987, 2001.