

Graphical model-based gene clustering and metagene expression analysis

Adrian Dobra, Quanli Wang and Mike West

Duke University, Durham, NC 27708, USA

ABSTRACT

Summary: We describe a novel gene expression analysis method for the creation of overlapping gene clusters and associated metagene signatures that aim to characterize the dominant common expression patterns within each cluster. The analysis is based on the use of statistical graphical models to identify and estimate patterns of association among gene subsets from gene expression data, and then clustering is based formal estimates of very sparse covariance matrices arising from these models. Metagene summaries, which are of interest as reduced dimensional summaries for phenotyping studies, are simply the resulting model-based estimates of dominant singular factors (principal components) of population variance matrices within resulting overlapping clusters. We describe connections between graph-theoretic approaches to exploring gene expression graphical models and exploration in biological contexts of gene subsets represented by identified metagenes, illustrating some aspects of the utility of this framework for summary representation of observational gene expression data.

Availability: The software implementing our method is called *MetageneCreator* and is available for download at <http://www.stat.duke.edu/~adobra/metagenecreator.htm>

Supplemental information: <http://www.stat.duke.edu/~adobra/ermeta.zip>

Contact: adobra@stat.duke.edu

INTRODUCTION

In a number of gene expression studies, the utility of multivariate statistical methods to define and estimate aggregate, common patterns underlying groups of genes has been demonstrated. Various clustering methods are in common use to define groups of genes, and data reduction methods to define weighted averages of expression of co-clustered genes can both reduce dimension and improve signal resolution in relation to predicting a phenotype that is inherently related to multiple co-expression or co-regulated genes. Singular value decomposition (principal component) methods are standard tools, and underlie methods for expression data summary, reduction and characterization as well as the use of such aggregate summaries as predictors of defined clinical or physiological phenotypes. Some key examples include the *eigengenes* of Alter *et al.* (2000, 2003), and the *metagenes* of West *et al.* (2001); Huang *et al.* (2003a,b); Pittman *et al.* (2004). The latter authors focus on the use of clustering methods, as pop-

ularized by Eisen *et al.* (1998) for example, to define multiple clusters with a view to reduce dimension while hopefully maintaining a representation of multiple common aspects of variation in gene expression across samples through weighted averages defined as the dominant singular factors within each cluster.

There are many possible variations on this kind of application of standard statistical clustering and dimension reduction. Our interest here lies in three aspects: first, the development of refined methods of clustering to ensure that the aggregate dominant singular factor does indeed represent a common pattern underlying a group of genes that show reasonable co-expression patterns; second, the enrichment of gene subsets defining clusters using estimated patterns of association between existing clusters and individual genes; and, third, improvement of the overall strategy utilizing improved estimates of covariance matrices of gene expression variables based on the use of Bayesian statistical graphical models.

We begin with discussion with a simple and effective heuristic algorithm that, beginning with k-means clustering, constructs subsets of genes whose variation can be summarized by the first singular factor (principal component) within the group. The idea is simply to iteratively refine larger clusters to focus on smaller subsets within which genes are more and more coherently co-expressed. This is followed with discussion of a method of enriching gene membership of clusters using a key but apparently novel measure of association - in covariance terms - between individual genes that are candidates to join a cluster and the existing group of genes within that cluster. This (and other) development of clustering and cluster enrichment of course relies on an estimate of the covariance matrix of expression of genes. The final contributions here focus on the use of sparse Bayesian graphical models for improving estimation of such high-dimensional covariance matrices. Here we discuss issues related to choosing model and parameter priors as well as distributed computational algorithms for model search. This is followed by details of how to derive model-based estimates of high-dimensional covariance matrices for use in clustering and other studies, and the broader use of such models in identifying candidate statistical association graphs - network representations of gene expression data that are of value in visualizing the empirical associations in new and sometimes insightful ways. The paper concludes with an example from breast cancer genomics and summary comments.

REFINING K-MEANS CLUSTERS

Throughout, we use *metagene* to refer explicitly to the dominant singular factor (principal component) of a defined subset of genes. Let $x^{(n)} = (x^1, \dots, x^n)$ denote the expression data on p genes (rows) and n samples (columns). The genes are identified by indices in $K = \{1, 2, \dots, p\}$. Consider a subset of genes $A \subset K$ of size $\#A$. The data $x_A^{(n)}$ associated with the genes in A has $\#A$ rows and n columns. The SVD of $x_A^{(n)}$ is given by:

$$x_A^{(n)} = UDV',$$

where U, V are orthogonal matrices of dimensions $\#A \times n$ and $n \times n$, and $D = \text{diag}(d_1, \dots, d_n)$ is the diagonal matrix of non-negative singular values of $x_A^{(n)}$ in decreasing order. The common dominant pattern within the cluster A (i.e., the *metagene*) represented by the first principal component f of $x_A^{(n)}$ is a linear combination of the $\#A$ variables in this cluster: $f = c'x_A^{(n)}$. Here f is the first column of V and c is the first column of UD^{-1} .

The score of the cluster A is defined as the total variation explained by the common dominant pattern f :

$$\text{Score}(A) = 100d_1^2 / \left(\sum_{i=1}^n d_i^2 \right). \quad (1)$$

A large value of $\text{Score}(A)$ indicates a tight cluster whose variation can be summarized by f .

A simple but effective algorithm now described aims to refine cluster membership to ensure coherent groupings of genes in the sense that the resulting metagene for each group is highly representative of the major common pattern within that group. The heuristic procedure constructs a clustering of the genes K such that the score of each cluster is above a threshold s_0 . Proceed as follows:

STEP 0. Initialize the set of genes that have not been clustered so far: $\mathcal{L} \leftarrow K$.

STEP 1. Select at random at most r genes from \mathcal{L} .

STEP 2. Cluster these genes using k-means.

STEP 3. **For** each cluster A **do**:

- If $\text{Score}(A) \geq s_0$ save the cluster A and delete all the genes in A from \mathcal{L} . Go to the next cluster.
- Otherwise delete from A the gene i_0 whose removal leads to a maximum increase in the score of the resulting cluster:

$$i_0 = \text{argmax}\{\text{Score}(A \setminus \{i\}) : i \in A\}.$$

- If the cluster contains less than l_0 genes, go to the next cluster.

STEP 4. Repeat steps 1, 2 and 3 until \mathcal{L} becomes empty or until no additional clusters are saved.

Ideally we would like to use k-means on the entire set of un-clustered genes \mathcal{L} but this might be computationally infeasible on most computing systems. The method performs much better if r is as large as possible (usually several thousand when dealing with 10-30 thousand genes).

The minimum allowed score s_0 should be set as large as possible to create tighter clusters especially if the sample size is small (less than a hundred). Larger values of s_0 lead to a larger number of clusters created. The minimum size of a cluster l_0 can be set to three or to larger values depending on each dataset. The genes that are still in \mathcal{L} when the procedure stops are taken to be clusters of size 1 and appended to the rest of the groups.

Assume that the procedure generates q clusters A_1, A_2, \dots, A_q with $q \ll p$. Then the new $q \times n$ data matrix that replaces $x^{(n)}$ is given by $F = [f'_1, f'_2, \dots, f'_q]'$ where f_j is the metagene associated with cluster A_j .

Related ideas of using SVD to construct overlapping clusters of genes with common expression patterns are present in the “gene shaving” method of Hastie *et al.* (2000). Their procedure sequentially “cleans” clusters by discarding genes that have low absolute correlation with the corresponding metagene (also called eigengene).

ENRICHING CLUSTERS OF GENES

Zhou *et al.* (2002) point out that genes with similar functions do not necessarily have highly correlated expression profiles. Such genes are very likely to end up in different clusters produced by the k-means algorithm and consequently by the heuristic procedure for producing more coherent clusters of genes. Zhou *et al.* (2002) introduce the notion of *transitive co-expression* to account for the situations when genes with similar function are only weakly correlated in their expression levels, but are strongly correlated in expression with some other group of genes. We have defined an approach to enriching gene clusters to address this, based on measuring association between non-cluster genes and all the genes in a cluster. This appears to be a quite novel application of a standard statistical measure of multiple association.

Consider a cluster of genes indexed by A and any gene $g \notin A$. Given the estimated covariance matrix of all genes, that for $B = \{g\} \cup A$ is

$$\Sigma_B = \begin{bmatrix} \sigma_{gg} & \Sigma_{gA} \\ \Sigma_{Ag} & \Sigma_{AA} \end{bmatrix}.$$

Here $\Sigma_{Ag} = \Sigma'_{gA}$, $\Sigma_{AA} = \Sigma_A$ and σ_{gg} is the variance of gene g . Then, the percent of the variance of gene g that is explained by the genes in A is given by:

$$\rho(g|A) = 100 * (\Sigma_{gA} \Sigma_{AA}^{-1} \Sigma'_{gA}) / \sigma_{gg}. \quad (2)$$

The score $\rho(g|A)$ is a generalized version of correlation coefficient between a gene g and a group of genes A and it is equivalent to the familiar R^2 statistic from simple linear regression models. Higher values of $\rho(g|A)$ indicate that the expression pattern of the gene g is closely related to the expression patterns of the genes in A . Therefore we can order the genes in $K \setminus A$ in decreasing order with respect to (2); the genes at the top of this list are the most likely genes to be functionally related with the genes in A . Therefore $\rho(g|A)$ represents the basis for making use of the transitive co-expression property of cellular processes to enrich clusters of genes as follows:

STEP 0. Start with a cluster of genes $A \subset K$.

STEP 1. Sort the genes $g \in K \setminus A$ in decreasing order of their scores $\rho(g|A)$ as in (2).

STEP 2. Start at the top of the list generated at STEP 1 and sequentially append genes to cluster A as long as $Score(A) \geq s_0$, where $Score(A)$ is defined in (1).

STEP 3. Repeat steps 1 and 2 until no additional genes are added to cluster A .

The resulting clusters are still coherent since their corresponding metagenes are required to explain at least s_0 percent of the total variation in expression within the groups. Moreover, the clusters contains genes with strongly correlated expression profiles as well as genes that are strongly correlated with sub-groups of genes in the same cluster. STEP 1 considers all the genes that are currently not in the cluster, and it follows that the method produces *overlapping* clusters as a gene can now belong to any number of groups. The resulting potentially overlapping subsets of genes could have a common pattern of expression but can also be related by *transitive co-expression* (Zhou et al., 2002). This construction considerably increases the likelihood that functionally related genes end up in the same cluster while genes with shared functions can belong to the groups that represent these functions.

It is important to point out that, if g_1 and g_2 are two probe sets for the same gene, then it is likely that the expression patterns for g_1 and g_2 are very similar. Hence, if g_2 belongs to A , the score $\rho(g_1|A)$ will be high irrespective of the other genes in A and hence it is likely that g_1 will also be included in A . Therefore the inherent collinearities that are strong in expression studies do not adversely affect our method for enriching clusters since duplicate probes should be examined together.

GAUSSIAN GRAPHICAL MODELS

Gaussian graphical models provide a formal, model-based framework for parametric inference on the high-dimensional covariance matrix of gene expression. Assume a suitably transformed version of the expression data is centered and scaled element-wise, and modeled as a zero-mean, multivariate normal random sample. That is, each sample of gene expression of the p genes is a p -vector $x \sim N_p(0, \Sigma)$ where $\Sigma = \{\sigma_{ij}\}$ is the positive definite covariance matrix; we often

work in terms of the precision matrix $\Omega = \Sigma^{-1} = \{\omega_{ij}\}$. In most gene expression datasets p is very large (tens of thousands) while n is relatively small (tens or possibly hundreds). Graphical models utilizing priors that encourage sparsity of the precision matrix Ω provide opportunities to substantially improve the precision of inferences on covariance patterns in this hugely challenging space.

The foundations of the approach lie in Dempster (1972), who introduced the idea of simplifying the structure of Σ by setting elements of Ω to zero. This leads to more robust estimates of Σ if Ω is required to have a substantial number of structural zeros. In addition, the dependency patterns among the genes in the dataset can be visually summarized by means of an *independence graph* $\mathcal{G} = (K, E)$ where K represents the vertices of \mathcal{G} (each variable/gene is associated with a vertex) while E is the set of edges E given by the off-diagonal elements of Ω that are not constrained to be zero:

$$E = \{(i, j) | \omega_{ij} \neq 0, \quad i \neq j\}.$$

Two genes that are connected by an edge in \mathcal{G} are believed to have a direct association. If the edge between two genes is missing in \mathcal{G} , the genes might still have a substantial association but this association is indirect. Then the genes are conditionally independent since they have no association given the information on some other subset of genes; see, for example, Lauritzen (1996).

The Gaussian distribution given by the covariance matrix Σ and the independence graph \mathcal{G} is a *graphical model* $M = (\Sigma, \mathcal{G})$. This model is undirected since the edges in E are lines that represent symmetric associations: any two variables (or genes) joined by an edge can be either response or explanatory for the other. The duality between predictor and response variables is the intuition behind the methodology for constructing large-scale graphical models proposed by Dobra et al. (2004) and further refined in Dobra and West (2004). The full joint distribution is modeled as a set of regressions for each of the p variables. These univariate models can be written in structural form as

$$x = \Gamma x + \varepsilon, \quad (3)$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)'$ are the error terms for the p regressions and $\Gamma = \{\gamma_{ij}\}$ is a $p \times p$ matrix of regression coefficients. The set of regressions (3) define a valid joint distribution as given by the chain rule in cases when (as a sufficient condition) Γ is upper triangular form with zero diagonal elements. In this case the regression error terms are independent, i.e. $\varepsilon \sim N_p(0, \Psi)$ with $\Psi = \text{diag}(\psi_1, \dots, \psi_p)$.

The computations involved are substantially simplified since we have immediate access to a Cholesky decomposition of $\Omega = LL'$ where $L = (I - \Gamma)'\Psi^{-1/2}$ is a lower-triangular matrix. If the predictor set of each regression model involves a relatively small number of regressors, the resulting matrix L is sparse which translates into a sparse precision matrix Ω .

The corresponding matrix $\Sigma = (L^{-1})'L^{-1}$ is not sparse enough to be fully calculated but variance-covariances of subsets of variables $A \subset K$ can readily be obtained via a simple matrix multiplication:

$$\Sigma_A = V_A'V_A \quad (4)$$

where V_A are the rows of L^{-1} corresponding to the variables in A .

Constructing *sparse* graphical models is key for gene expression data because sparsity reflects the view that patterns of variation for a given gene are well predicted by those of a relatively small subset of other genes. Sparse regression models are induced by a penalty term $\beta/(1-\beta)$ for the inclusion of an additional covariate in the model. We empirically observed that $\beta = 1/(p-1)$ gives good results in most applications (Dobra *et al.*, 2004).

An encompassing inverse Wishart prior for the full covariance matrix Σ implies consistent normal/inverse gamma priors for the regression parameters given by $\{\gamma_{ij}\}_{i>j}$ and $\{\psi_i\}$. This leads to exact formulas for calculating the marginal likelihood $p(x^{(n)}|M)$ of a graphical model $M = (\Sigma, \mathcal{G})$. It follows that the posterior of M is proportional with

$$p(M|x^{(n)}) \propto p(x^{(n)}|M)p(M).$$

The prior weight of the model is given by the product of prior weights of each regression model:

$$p(M) = [\beta/(1-\beta)]^{\#M},$$

where $\#M$ represents the total number of predictors in all p compositional regressions.

We are interested in structural learning on the space \mathcal{S} of graphical models that can be obtained from a triangular linear system (3). This means that we want to determine a set of models $\mathcal{M} = \{M_i = (\Sigma_i, \mathcal{G}_i) : i = 1, 2, \dots, m\}$ having large posterior probabilities $p(M_i|x^{(n)})$. Dobra and West (2004) present a three step model search procedure that is guaranteed to converge to local optima in \mathcal{S} . The first step of this method identifies candidate predictors for each variable. The second step consists of a heuristic for finding models in \mathcal{S} with high posterior weights. At the third step the models found at Step 2 are sequentially improved until convergence. Versions of the first two steps are also described in Dobra *et al.* (2004).

Once we have determined a set of good candidate models \mathcal{M} , we can account for model uncertainty by employing Bayesian model averaging (Raftery *et al.*, 1997) on \mathcal{M} . Consider the posterior normal distribution given by the average of the posterior distributions under each model weighted by the corresponding posterior probabilities:

$$\sum_{i=1}^m \pi_i N_p(0, \Sigma_i), \quad (5)$$

with $\pi_i = p(M_i|x^{(n)})/[\sum_{j=1}^m p(M_j|x^{(n)})]$. Quantities of interest can then be evaluated by repeatedly sampling from the

mixture (5) then averaging across the samples obtained. For example, if we are interested in the precision matrix Ω of the mixture (5) and/or in the variance-covariance matrix Σ_A associated with a subset of variables $A \subset K$, we need to repeat the following steps until the convergence of the mean of the estimates produced:

1. Sample a model M_{i_0} using the weights $\{\pi_i\}$.
2. Sample the regression parameters from the corresponding closed form posteriors (Dobra *et al.*, 2004) to obtain the matrices $(\Gamma_{i_0}, \Psi_{i_0})$.
3. Calculate $L_{i_0} = (I - \Gamma_{i_0})'\Psi_{i_0}^{-1/2}$. An estimate for Ω is $L_{i_0}L_{i_0}'$.
4. Take the inverse of L_{i_0} and obtain an estimate for Σ_A as in (4).

GRAPHICAL ASSOCIATION NETWORKS

We define a graphical association network for gene expression to be the independence graph \mathcal{G} associated with the precision matrix $\hat{\Omega}$ estimated from the mixture (5) across the statistically significant graphical models determined (Dobra *et al.*, 2004). Visualization of such an empirical network provides novel access to gene expression based information with potential to generate insights into biological relationships and gene function.

We contrast our definition with a more straightforward but very effective approach of defining association networks that assigns an edge between two genes if their absolute expression correlation is above a certain threshold; see, for example, Zhou *et al.* (2002). The length of an edge is taken to be a decreasing function of the absolute expression correlation, thus highly correlated genes are closer in the resulting graph. This type of networks can be successfully explored with graph-theoretic procedures such as Dijkstra's shortest-path algorithm: given any two genes g_1 and g_2 , the transitive genes that are potentially functionally related with g_1 and g_2 are found on paths of minimum length between g_1 and g_2 . Related ideas are presented in Rives and Galitski (2003) who identify components in protein interaction networks through shortest-path distances between any pair of vertices.

Our method of assigning links between genes if the corresponding entries in $\hat{\Omega}$ are not constrained to be zero is fundamentally different than the networks of Zhou *et al.* (2002). In our definition two genes g_1 and g_2 might be connected with an edge even if their expression levels are only weakly correlated. For example, this could happen if g_1 and g_2 are both predictors in the regression model associated with another gene g , but g_2 (g_1) is not a predictor in the regression model for g_1 (g_2). The edge between g_1 and g_2 is generated through the process of "moralization" (Lauritzen, 1996).

A direct consequence of this fact is that exploring networks derived from a precision matrix using paths of minimum length between genes does not have the same meaning as in the correlation-based networks. Jones and West (2004) have generated insights into this issue with theory of decompositions of the covariance between two genes g_1 and g_2 into weights that measure the strengths of the relationships between genes along paths that link g_1 and g_2 in \mathcal{G} . Their construction makes use of the underlying graphical models \mathcal{M} to determine the genes that are most relevant in mediating correlation between g_1 and g_2 . More directly in connection with our interests here are measures of association between groups of genes, and between one gene and a defined subset of more than one gene, in which context the general score metric earlier defined, in (2), is of obvious relevance.

IMPLEMENTATION

We have implemented the methods described in this paper in a MATLAB package called *MetageneCreator*. The program creates an initial clustering of the genes based on their expression profiles $x^{(n)}$ using our heuristic procedure based on the k-means algorithm. Then the program generates the corresponding metagenes and selects the clusters whose corresponding metagenes have the highest absolute correlation with a phenotype y . These will be the clusters that are further enriched based on the Gaussian graphical models. The user needs to specify the number of clusters that are selected, the minimum score of a cluster as well as the actual data $x^{(n)}$ and y . The program saves the final enriched clusters together with the metagenes associated with these clusters. The clusters that are subsets of other clusters are removed. We point out that our method for producing overlapping clusters is inherently un-supervised as the response vector y influences only the number of clusters that are ultimately produced not the structure of these clusters.

The constructive approach for generating large-scale sparse Gaussian graphical models described in Dobra and West (2004) is implemented in a C++ package called *HdBCS* (Dobra, 2004). The model search algorithms are specifically designed to exploit the architecture of a shared set of computers thus the implementation makes extensive use of MPI libraries.

The graphical association networks produced by *HdBCS* can be visualized with *GraphExplore* (Wang et al., 2004). This is a stand-alone multi-platform JAVA application that dynamically queries and renders complex networks of interactions. It can also retrieve relevant information about the objects in the network from the Internet.

We emphasize that *MetageneCreator*, *HdBCS* and *GraphExplore* are able to handle datasets with tens of thousands of genes/variables.

APPLICATION

We illustrate our methods through an example that involves expression data from 158 breast cancer samples at the Koo Foundation Sun Yat-Sen Cancer Center in Taipei. This dataset is publicly available as supplemental material in Pittman et al. (2004). Gene expression assays were performed on the Human U95Av2 GeneChip. The resulting MASS5.0 signal measures of expression were transformed on a log2 scale and quantile normalized after the removal of the 67 controls. Genes with small variation or genes expressed only at low levels were removed which leaves a total of 7,027 probe sets in the final dataset.

We employed *HdBCS* to search for statistically significant Gaussian graphical models. We chose to generate five starting models that were further improved in 1,000 iterations; one iteration represents about one thousand potentially different models. A number of 30 graphical models were obtained and were further used to estimate relevant quantities through Bayesian model averaging.

We use *MetageneCreator* to construct clusters and the metagenes associated with these groups with respect to estrogen receptor (ER, henceforth) status represented as 0,1,2,3 to reflect intensity of immunohistochemical staining for the ER protein. ER status is a key clinical factor in breast tumors (West et al., 2001; Huang et al., 2003a). A cluster is considered coherent if its score from (1) is at least 65%. The preliminary heuristic method created 801 such clusters of size three or more while leaving un-clustered a number of 3,689 genes. In this process k-means clustered blocks of genes of size at most 5,000 with a mean number of genes per cluster equal to 25 (this last parameter is needed since k-means has to know how many clusters to create). We extracted 500 clusters that exhibit the highest absolute correlation with ER status as given by their corresponding metagenes. These groups were further enriched to yield 495 overlapping clusters; five clusters were removed as they were contained within other groups.

We compare the clusters obtained using *MetageneCreator* with the clusters generated by directly applying k-means on the entire set of 7,027 genes. A number of 498 clusters were created when k-means was asked to create 500 groups; these are actually the same clusters associated with the metagenes in Pittman et al. (2004). Therefore the two clusterings of the genes contain about the same number of groups.

Figure 1 presents summaries of the two clusterings as bar plots of the frequencies of the cluster size, cluster score and absolute correlation with ER status of the metagenes associated with each cluster. The k-means clusters tend to be larger and have lower correlation values with ER status. Moreover, their corresponding metagenes does not seem to represent the overall expression patterns of the genes within the clusters as the total variation explained by the metagene (i.e., the cluster score) is in most cases below 60%. On the other hand, the all *MetageneCreator* clusters have scores above 65% as we

required; they also seem to exhibit significantly larger correlations with ER status. This is not surprising since the *MetageneCreator* clusters contain only 1,630 different genes and these genes might be more indicative of ER status. Figure 2 shows that there is a significant amount of overlap between the *MetageneCreator* clusters. Remark that a relatively large number of genes belong to 10 or more clusters.

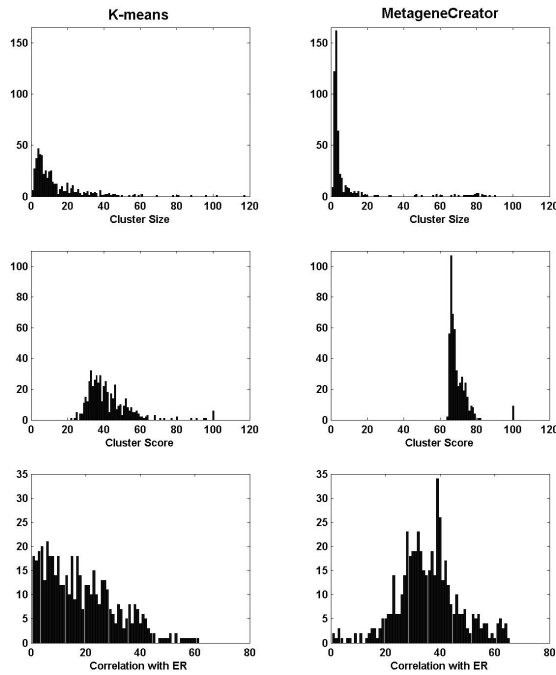


Figure 1: Comparison between k-means and *MetageneCreator* clusterings. The left column shows summaries associated with k-means clusters while the right column shows summaries associated with *MetageneCreator* clusters. These summaries are frequencies of cluster size, cluster score and absolute correlation with ER (as percentages) of the corresponding metagenes.

Therefore the resulting metagene dataset with 158 samples and 495 covariates is likely to lead to good predictive models for ER status; this represents a huge drop in the initial number of available covariates (7,027 in this example). However, the main quality of *MetageneCreator* is that it creates meaningful groups of genes that might potentially be functionally related.

The graphical association network induced by the 30

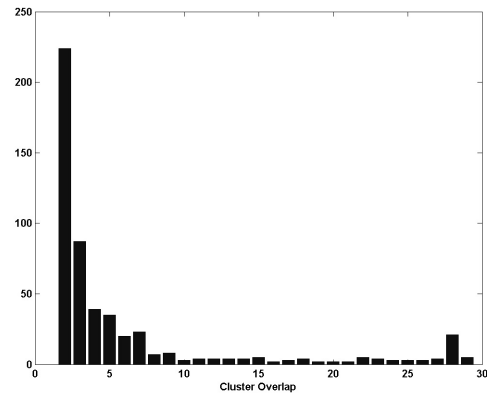


Figure 2: Overlap between the 495 clusters created by *MetageneCreator*. We counted the number of genes that belong to two clusters, three clusters and so on. The genes that belong to only one cluster were not considered.

Gaussian graphical models found by *HdBCS* is sparse since 90% of the genes have at most 23 neighbors – see Figure 3.

Figure 4 shows the sub-graph associated with the genes in cluster 438. This cluster is only one of the several *MetageneCreator* clusters that contains three genes that encode transcription factors that are known to have strong associations in expression with the estrogen receptor ESR1 gene (Lacroix and Leclercq, 2004). These genes are: GATA3, HNF3A (or FOXA1) and XBP1. The estrogen-inducible trefoil factor TFF1 gene that is a known ER target is also present in this cluster. Cluster 438 also includes TFF3, another member of the trefoil factor family that is closely related to TFF1.

The strength of the relationships among the genes in *MetageneCreator* clusters is evident from the fact that oligonucleotide sequences associated with the same genes are clustered together. For example, cluster 438 contains two probes associated with XBP1 and TFF3, respectively.

We reach similar conclusions by examining the cluster associated with the ESR1 gene—see Figure 5. This cluster contains two probes associated with ESR1 (namely 1893_s_at and ESR1 in Figure 5) and four probes associated with c-MYB (MYB_3, MYB_4, MYB_5 and 1474_s_at) which gives a clear indication of the strong association in expression between ESR1 and c-MYB.

The full list of clusters created by *MetageneCreator* together with the graphical association network containing all the 7,027 genes is available as supplemental information.

DISCUSSION

A key feature of the clusters found by *MetageneCreator* is that the gene-expression sub-graphs associated with them are

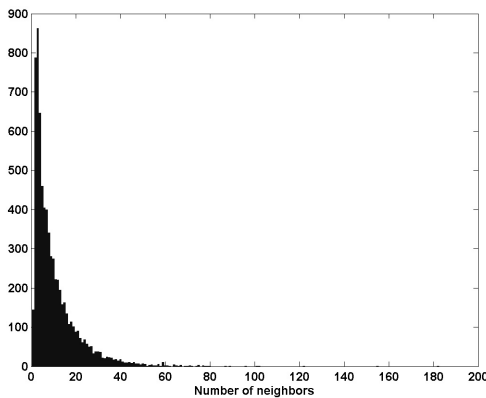


Figure 3: Number of direct neighbors of the 7,027 genes.

likely to be connected as illustrated by Figures 4 and 5. This means that the procedure for enriching a cluster A presented in this paper is in many ways equivalent to exploring the neighbors of the genes in A that belong to paths of length one, two, three, etc in the graphical association networks induced by the Gaussian graphical modes. Exploring gene expression association graphs using shortest-paths of varying length is valuable and leads to meaningful results (Dobra et al., 2004), but it does not offer a sound way for ranking the genes found on these paths. Therefore the procedure for enriching a cluster represents an alternative way for identifying genes associated in expression with any set of target genes one might be interested in. The clear advantage of this approach is that it is based on the statistical models that generated the graphical association network not on graph-related algorithms that might lead to less precise inferences.

Gaussian graphical models play a key role for the developments we have presented in this work. They provide an appropriate and statistically reliable approach to imposing sparsity on graphs underlying the precision matrix - and hence the structured covariance matrix - of many variables, and offer a way of reducing the dimensionality of a gene expression dataset by identifying groups of genes that are potentially functionally related. The underlying associations in expression among the genes in each cluster or among the genes in different clusters can then be visualized through the graphical association networks generated by these models. Evidently, much of the analysis of graphical models, and their access and visualization, is computationally demanding, and the software tools developed for the work here are provided for others to explore and use; this includes software for producing Gaussian graphical models (*HdBCS*) and to generate and explore resulting graphical displays (*GraphExplore*).

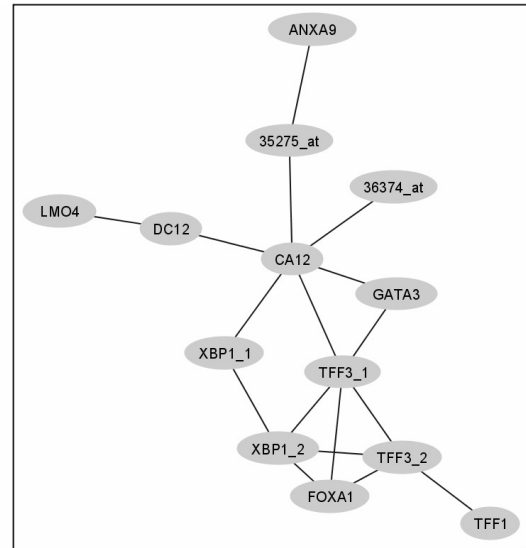


Figure 4: Genes in cluster 438 and the connection among them in the graphical association network obtained from Gaussian graphical models. The absolute correlation with ER status of the metagene associated with this cluster is 0.61. This image was produced with *GraphExplore* (Wang et al., 2004).

ACKNOWLEDGMENTS

This work has benefited from discussions with our colleagues Jennifer Pittman, Joseph Nevins and Guang Yao. Research was partially supported by grants NIH HL-073042, NSF DMS-0102227 and NSF DMS-0342172.

References

- Alter, O., Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modelling. *Proc. Natl. Acad. Sci.*, **97**, 10101–10106.
- Alter, O., Brown, P. O. and Botstein, D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci.*, **100**, 3351–3356.
- Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.
- Dobra, A. (2004) HdBCS: Bayesian covariance selection in high dimensions. Available for download at <http://www.stat.duke.edu/~adobra/hdbsc.html>.
- Dobra, A., Jones, B., Hans, C., Nevins, J. and West, M. (2004) Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**, 196–212.

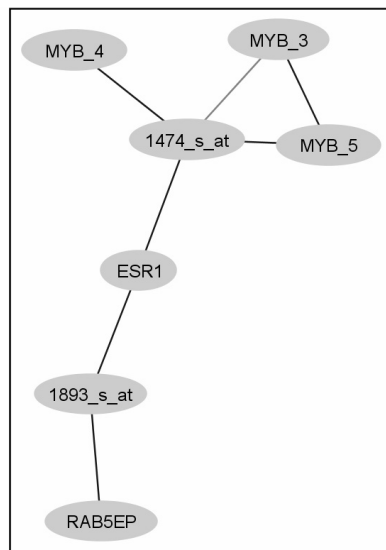


Figure 5: Gene expression sub-graph associated with cluster 398. The absolute correlation with ER status of the metagene associated with this cluster is 0.46. This image was produced with *GraphExplore* (Wang *et al.*, 2004).

- Pittman, J., Huang, E., Dressman, H., Horng, C. F., Cheng, S. H., Tsou, M. H., Chen, C. M., Bild, A., Iversen, E. S., Huang, A. T., Nevins, J. R. and West, M. (2004) Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl. Acad. Sci.*, **101**, 8431–8436.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**, 179–191.
- Rives, A. W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci.*, **100**, 1128–1133.
- Wang, Q., Dobra, A. and West, M. (2004) GraphExplore: a software tool for graph visualization. ISDS Discussion Paper #04-22.
- West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. and Nevins, J. R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.*, **98**, 11462–11467.
- Zhou, X., Kao, M. C. J. and Wong, W. H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci.*, **99**, 12783–12788.
- Dobra, A. and West, M. (2004) Bayesian covariance selection. ISDS Discussion Paper #04-23.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, **95**, 14863–14868.
- Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C. and Botstein, D. (2000) 'Gene shaving' as a method for identifying distinct set of genes with similar expression patterns. *Genome Biology*, **1**.
- Huang, E. H., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R. and Huang, A. T. (2003a) Gene expression predictors of breast cancer outcomes. *The Lancet*, **361**, 1590–1596.
- Huang, E. H., Ishida, S., Pittman, J., Dressman, H., Bild, A., D'Amico, M., Pestell, R., West, M. and Nevins, J. R. (2003b) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics*, **34**, 226–230.
- Jones, B. and West, M. (2004) Covariance decomposition in multivariate analysis. ISDS Discussion Paper #04-15.
- Lacroix, M. and Leclercq, G. (2004) About GATA3, HNF3A, and XBP1, three genes co-expressed with the oestrogen receptor- α gene (ESR1) in breast cancer. *Molecular and Cellular Endocrinology*, **219**, 1–7.
- Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon Press.