A Bayesian Approach to Modeling Uncertainty in Gene Expression Clusters

David L. Wild
Keck Graduate Institute of
Applied Life Sciences
535 Watson Drive
Claremont, CA 91171, USA
david_wild@kgi.edu

Carl Edward Rasmussen
Gatsby Computational
Neurosciene Unit
University College London
London, WC1N 3AR, UK
edward@gatsby.ucl.ac.uk

Zoubin Ghahramani
Gatsby Computational
Neurosciene Unit
University College London
London, WC1N 3AR, UK
zoubin@gatsby.ucl.ac.uk

The use of clustering methods has rapidly become one of the standard computational approaches to understanding microarray gene expression data [3, 1, 7]. In clustering, the patterns of expression of different genes across time, treatments, and tissues are grouped into distinct clusters (perhaps organized hierarchically) in which genes in the same cluster are assumed to be potentially functionally related or to be influenced by a common upstream factor. Such cluster structure can be used to aid in the elucidation of regulatory networks. For example, a compendium of gene expression profiles corresponding to mutants and chemical treatments can be used as a systematic tool to identify gene functions because mutants or drug targets that display similar profiles are likely to share cellular functions [5].

One commonly used computational method of non-hierarchical clustering based on measuring Euclidean distance between gene expression profiles is given by the k-means algorithm [4]. However, the k-means algorithm is inadequate for describing clusters of unequal size or shape [6]. A generalization of k-means can be derived from the theory of maximum likelihood estimation of Gaussian mixture models [8]. In a Gaussian mixture model, the data (e.g. gene expression profiles, which can be arranged into p-dimensional vectors $\mathbf{y}$) is assumed to have been generated from a finite number (k) of Gaussians,

$$P(\mathbf{y}) = \sum_{j=1}^{k} \phi_j P_j(\mathbf{y}) \qquad (1)$$

where ϕ_j is the mixing proportion for cluster j (fraction of population belonging to cluster j; $\sum_j \phi_j = 1$; $\phi_j \geq 0$) and $P_j(\mathbf{y})$ is a multivariate Gaussian distribution with mean μ_j and covariance matrix Σ_j. The clusters can be found by fitting the maximum likelihood Gaussian mixture model as a function of the set of parameters $\theta = \{\phi_j, \mu_j, \Sigma_j\}_{j=1}^{k}$ using the EM algorithm [8]. Euclidean distance corresponds to assuming that the Σ_j are all equal multiples of the identity matrix.

An important issue that must be addressed in any clustering method is the question of how many clusters to use. Bayesian statistics can provide a solution to model selection questions of this kind (e.g. [2]). An elegant alternative approach is to assume that the data was in fact generated from an *infinite* number of Gaussian clusters and to do Bayesian inference under this assumption. This is a sensible way to capture the fact we don't really believe that gene expression data is well modeled by a finite number of Gaussians. Also, infinite Gaussian mixture models can readily model a finite number of non-Gaussian clusters. Finally, in an infinite Gaussian mixture model there is no need to make arbitrary choices about how many clusters there are in the data; nevertheless, after modeling one can ask questions such as how probable it is that two genes belong to the same cluster?

The theory of infinite mixture models is laid out in [9]. Starting from a finite mixture model (1), we define a prior over the mixing proportion parameters ϕ. The natural conjugate prior for mixing proportions is the symmetric Dirichlet distribution: $P(\phi|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^{k} \phi_j^{\alpha/k-1}$. We then explicitly include indicator variables c_i for each data point (i.e. gene or condition) which can take on integer values $c_i = j$, $j \in \{1, \ldots, k\}$, corresponding to the hypothesis that data point i belongs to cluster j. Under the mixture model, by definition, the prior probability is proportional to the mixing proportion: $P(c_i = j|\phi) = \phi_j$. A key observation is that we can compute the conditional probability of one indicator variable given the setting of all the other indicator variables after *integrating over* all possible settings of the mixing proportion parameters: $P(c_i = j|\mathbf{c}_{-i}, \alpha) = \int P(c_i = j|\mathbf{c}_{-i}, \phi) P(\phi|\mathbf{c}_{-i}, \alpha)\ d\phi \ = \frac{n_{-i,j}+\alpha/k}{n-1+\alpha}$, where $\mathbf{c}_{-i}$ is the setting of all indicator variables except the i^{th}, n is the total number of data points, and $n_{-i,j}$ is the number of data points belonging to class j not including i. We now can take the limit of k going to infinity, obtaining a Dirichlet Process with differing conditional probabilities for clusters with and without data: for clusters where $n_{-i,j} > 0$: $p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha}$, for all other clusters combined: $p(c_i \neq c_{i'} \text{ for all } i' \neq i|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}$. This shows that the probabilites are proportional to the occupation numbers, $n_{-i,j}$. Using these conditional probabilities one can Gibbs sample from the indicator variables efficiently, even though the model has infinitely many Gaussian clusters. Having integrated out the mixing proportions one can also Gibbs sample from all of the remaining parameters of the model, i.e. $\{\mu, \Sigma\}_j$. The details of these procedures can be found in [9].

We have used infinite Gaussian mixtures to model experimentally measured microarray gene expression data with the intention of creating a general system which can answer queries of the kind: what is the probability that two genes belong to the same cluster (i.e. have similar functional roles or are influenced by a common upstream factor))? Unlike previous methods based on a single clustering of the data, this approach computes this probability while taking into account all sources of model uncertainty (including number of clusters and location of clusters). We use the probability p_{ij} that two genes i and j belong to the same cluster in the infinite mixture model as a measure of the similarity of these gene expression profiles. Conversely $1 - p_{ij}$ defines a dissimilarity measure which for the purposes of visualization can be input to one of the standard linkage algorithms used for hierarchical clustering. We compare the dendrograms thus obtained to the usual hierarchical clustering approach which computes distance metrics on directly on the gene expression profiles or correlation coefficients between profiles [3].

We illustrate our methods with application to two published data sets. The first is the Rosetta compendium of expression profiles corresponding to 300 diverse mutations and chemical treatments in *S. cerevisiaie* [5]. The second is the NCI-60 data set of expression profiles of cancer cell lines and drug treatments [11]. For the NCI-60 data set, which has labelled classes of cell lines and drug mechanisms of action, it is possible to compare the quality of hierarchical clusterings obtained from different methods to these known classes. However, as the literature is notably lacking in quantitative measures of dendrogram quality, in order to perform this comparison we have devised a quantitative measure **DendrogramPurity**, which takes as input a dendrogram tree structure $\mathcal{T}$ and a set of class labels $\mathcal{C}$ for the leaves of the tree and outputs a single number measuring how "pure" the subtrees of $\mathcal{T}$ are with respect to the class labels $\mathcal{C}$.

DendrogramPurity($\mathcal{T}$,$\mathcal{C}$): where T is a binary tree (dendrogram) with set of leaves $\mathcal{L} = \{1 \ldots, L\}$ and $\mathcal{C} = \{c_1, \ldots, c_L\}$ is the set of known class assignments for each leaf. The DendrogramPurity is defined to be the measure obtained from this random process: pick a leaf ℓ uniformly at random. Pick another leaf j in the same class, i.e. $c_\ell = c_j$. Find the smallest subtree containing ℓ and j. Measure the fraction of leaves in that subtree which are in the same class, i.e. c_ℓ. The expected value of this fraction is the DendrogramPurity. This measure can be computed efficiently using a bottom up recursion (without needing to resort to sampling). The overall tree purity is 1 if and only if all leaves in each class are contained within some pure subtree.

For each leaf of the tree it also useful to measure how well it fits in with the labels of the leaves in the surrounding subtree. Leaves which do not fit well contribute to decreasing the overall dendrogram purity. These may highlight unusual or misclassified genes, drugs or cell lines. We define the **LeafHarmony** of a leaf ℓ as a measure of how well that leaf fits in.

LeafHarmony(ℓ,$\mathcal{T}$,$\mathcal{C}$): Pick a random leaf j in same class as leaf ℓ, i.e. $c_j = c_\ell$, $j \neq \ell$. Find the smallest subtree containing ℓ and j. Measure the fraction of leaves in that subtree which are in class c_ℓ. The expected value of this fraction is the LeafHarmony for ℓ and it measures the contribution of that leaf to the DendrogramPurity.

For the case of the Rosetta compendium analysis where there are not clearly defined class labels these measures are not applicable so we have defined a measure, the **LeafDisparity**, which highlights differences between two hierarchical clusterings (i.e. dendrograms) of the same data. Intuitively, this measures for each leaf of one dedrogram how similar the surrounding subtree is to the corresponding subtree in the other dendrogram.

Define the correlation between two sets $\mathcal{S}$ and $\mathcal{R}$ to be $c(\mathcal{S},\mathcal{R}) = |\mathcal{S} \cap \mathcal{R}|/|\mathcal{S} \cup \mathcal{R}|$, where $|\cdot|$ denotes the number of elements in a set. $c(\mathcal{S},\mathcal{R}) = 1$ iff $\mathcal{S} = \mathcal{R}$ and $c(\mathcal{S},\mathcal{R}) = 0$ iff $|\mathcal{S} \cap \mathcal{R}| = \emptyset$. Note that a tree $\mathcal{T}$ can be converted into a set-of-sets representation $\mathcal{T} = \{\tau_1, \ldots, \tau_k\}$. For each node j in the tree, τ_j is the set of the leaves in the subtree descending from j. (Thus in a binary tree with n leaves contains $n - 1$ non-leaf internal nodes, so $k = 2n - 1$).

LeafDisparity(ℓ,$\mathcal{T}$,$\mathcal{T}'$): Convert each tree into a set-of-sets representation. Align the trees: For each set τ_j in $\mathcal{T}$, find the set ρ_k in $\mathcal{T}'$ such that the correlation is greatest: $r_j = \max_k c(\tau_j, \rho_k)$. For each leaf ℓ find the average of r_j over all sets that contain ℓ, calling this $\bar{r}(\ell)$. If the element ℓ appears in both $\mathcal{T}$ and $\mathcal{T}'$ let its disparity be the minimum of $1-\bar{r}(\ell)$ in either tree. Thus this measure will be symmetric and sensitive to disagreement between the hierarchical clustering given by each tree.

For both data sets the dimensionality of the data was first reduced by principal components analysis. Empirically, the first 10 principal components were used. The mixture model was started with a single component, and 330000 iterations of Gibbs sampling were performed: the first 33000 steps for initial "burn-in", with the remaining 297000 used to generate 100 roughly independent samples from the posterior distribution (spaced evenly 2970 steps apart).

Our results show some similarities and but also biologically significant differences to the published dendrograms of [5] and [11]. Whilst the aim of this paper is not to provide a detailed analysis of these data sets, we note that, as a validation of our method, known associations of tumor cell lines and drug mechanisms of action in [11] and experimentally validated associations between genes of known and unknown function in [5] are reproduced in our clustering.

In the case of the NCI-60 data set, our clustering of tumor cell lines on the basis of gene expression broadly reproduces that of [11], whilst the dendrogram produced from our clustering of cell lines on the basis of drug activity exhibits greater "purity" than that obtained with a distance metric of (1 - Pearson correlation coefficient) [11]. We observe that 7/8 melanoma, 6/6 leukaemia and 7/7 breast cell lines cluster together. In our clustering of drugs on the basis of acivity across the cell lines (A matrix), we note that fluorouracil(5-FU) clusters together with the platin compunds (tetraplatin, diaminocyclohexyl-Pt-II), rather than with the RNA synthesis inhibitors, as in [11]. This pattern is repeated in the clustering of drugs on the basis of correlation between drug activity and gene expression. Since the platin compounds and antimetabolites such as fluorouracil(5-FU) and floxuridine(FUdur) at low doses are also DNA synthesis inhibitors [10], this association appears plausible.

In our clustering of 127 experiments (diverse mutations and chemical treatments) from the Rosetta data set [5] on the basis of gene expression patterns, we observe that all of the uncharacterised ORFs whose function was experimentally validated by [5] cluster with genes of similar known function. Our results reveal 5 well defined clusters which correspond to particular cellular processes: a large group of mutations which appear to have no consistent phenotype, a cluster comprising mating and MAPK pathway related proteins, a cluster of ER proteins related to ergosterol biosynthesis, a cluster of ribosomal proteins and a cluster of DNA-repair related genes.

1. REFERENCES

[1] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci*, 96:6745–6750, 1999.

[2] Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. *J. Comput. Biol.*, 9(2):161–191, 2002.

[3] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression. *PNAS*, 95:14863–14868, 1998.

[4] J. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

[5] T. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2001.

[6] D. Mackay. *Information Theory, Inference and Learning Algorithms.* forthcoming.

[7] G. McLachlan, R. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

[8] G. McLachlan and D. Peel. *Finite Mixture Models.* Wiley, New York, 2000.

[9] C. E. Rasmussen. The infinite gaussian mixture model. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

[10] K. Scanlon, M. Kashani-Sabet, T. Tone, and T. Funato. Cisplatin resistance in human cancers. *Pharmac. Ther.*, 52:383–406, 1991.

[11] U. Scherf et al. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24(3):236–244, 2000.