Hierarchical Generative Biclustering for MicroRNA Expression Analysis

José Caldas and Samuel Kaski

Aalto University School of Science and Technology
Department of Information and Computer Science
Helsinki Institute for Information Technology
P.O. Box 15400, FI-00076 Aalto, Finland
jose.caldas@tkk.fi, samuel.kaski@tkk.fi

Abstract. Clustering methods are a useful and common first step in gene expression studies, but the results may be hard to interpret. We bring in explicitly an indicator of which genes tie each cluster, changing the setup to biclustering. Furthermore, we make the indicators hierarchical, resulting in a hierarchy of progressively more specific biclusters. A non-parametric Bayesian formulation makes the model rigorous and yet flexible, and computations feasible. The formulation additionally offers a natural information retrieval relevance measure that allows relating samples in a principled manner. We show that the model outperforms other four biclustering procedures in a large miRNA data set. We also demonstrate the model's added interpretability and information retrieval capability in a case study that highlights the potential and novel role of miR-224 in the association between melanoma and non-Hodgkin lymphoma. Software is publicly available.[1]

Keywords: Biclustering, graphical model, information retrieval, nested Chinese restaurant process, miRNA, melanoma, non-Hodgkin lymphoma.

1 Introduction

Unsupervised learning methods are often used as a first step in biological gene expression studies [1]. The fact that most methods do not provide interpretable structures as to why the data was grouped as such hinders the subsequent analysis. Biclustering, where objects are both grouped and associated with feature subsets, is a natural framework for improving interpretability [2]. Although several biclustering approaches exist, few are capable of handling the uncertainty that necessarily arises for a large enough number of biclusters. We recur to the probabilistic modelling framework [3] in order to develop a biclustering method that is interpretable, has flexibility and expressive power, and is efficiently computable. Probabilistic approaches to biclustering in the biological sciences have already been successfully used in the analysis of chemogenomic studies [4] and gene expression data [5], although the corresponding models differ significantly

[1] http://www.cis.hut.fi/projects/mi/software/treebic/

B. Berger (Ed.): RECOMB 2010, LNBI 6044, pp. 65–79, 2010.

from ours. In particular, we propose a method to jointly group microarray samples hierarchically and assign genes to nodes in the hierarchy, with the node assignments implying that samples under the scope of a node in the hierarchy are homogeneous with respect to the genes assigned to it.[2] This enables the method to both provide a tree-structured clustering and explicitly state which features in the data were responsible for the groupings.

We show how the model yields a natural information retrieval relevance measure that allows relating samples in a principled manner. We apply the model to a large miRNA data set [6], compare it to other biclustering approaches, and illustrate the model's advantages with a case study about the role of miR-224 on the relation between melanoma and non-Hodgkin lymphoma.

The paper is organized as follows: We first describe the model, its inference procedure, and an information retrieval relevance measure. We then compare our model to four other biclustering approaches in a miRNA data set, quantify the model's information retrieval performance, and elaborate on a case study. Finally, we summarize our work and describe potential future directions.

2 Generative Model

2.1 Specification

The research problem is to find a hierarchy of clusters such that the objects (microarray samples) associated with a cluster are homogeneous for a subset of features (genes). Child clusters are to be associated with less objects but wider feature subsets than their corresponding parent clusters.

The proposed model can be seen as a particular instance of a biclustering method [2], where each bicluster corresponds to a group of samples that behave like replicates for a subset of genes. Biclusters are arranged as nodes in a tree hierarchy, with nodes closer to the root corresponding to broad sample groups tied by a low number of genes, and with nodes closer to the bottom of the hierarchy corresponding to limited but highly homogeneous sample groups. The generative process for our model consists of three parts: First, samples are partitioned into a tree structure. Second, genes are positioned along nodes in the tree. Third, the expression data is generated accordingly.

In order to partition samples into a tree structure, we use a probability distribution over infinitely-branched trees called the nested Chinese restaurant process (nCRP) [7]. This process may be defined over infinite-depth or finite-depth trees. We opt for specifying a maximum depth parameter in advance. Running the nCRP with a set of samples results in each sample being assigned a unique path from the root to a leaf node. The tree is initialized with a single node (the root), to which all samples are assigned. The samples are then probabilistically partitioned into groups according to the Chinese restaurant process (CRP)[8].[3]

[2] Alternatively, it may hierarchically group genes and assign samples to nodes, although we did not explore that option in the present work.

[3] Using the standard gastronomic metaphor associated with the CRP, we will interchangeably refer to groups as tables and samples as clients.

Formally, assume n clients are partitioned into k different tables ($k \leq n$), making each of those tables j contain m_j clients. The assignment probabilities for the $(n+1)$-th client are given as follows:

$$P(c_{n+1} = j | \boldsymbol{c}_{1,\ldots,n}) = \begin{cases} \dfrac{m_j}{n+\gamma}, & j \leq k, \\ \dfrac{\gamma}{n+\gamma}, & j = k+1. \end{cases} \tag{1}$$

The joint distribution for all clients is exchangeable (i.e. invariant to client order permutation), with γ controlling the final number of tables. We consider γ to be a random variable with a vague prior distribution,

$$\gamma \sim \text{Gamma}\left(a_\gamma = 1, b_\gamma = 1\right). \tag{2}$$

The obtained tables become the child nodes of the root. The CRP is again run for each of the child nodes and corresponding clients. This recursion continues until the maximum tree depth has been reached. See Fig. 1 for an example.

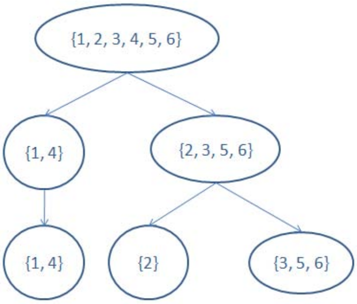

Fig. 1. Running the nCRP for a set of 6 clients (numbered from 1 to 6) in an infinitely-branched tree of maximum depth 3. The clients assigned to each node are between braces.

Given the assignment of samples to paths in the tree, we represent genes as binary features and provide a feature activation model. First, for each directed edge (u,v) in the tree, we sample an edge length from a uniform Beta distribution,

$$l_{(u,v)} \sim Beta(\alpha = 1, \beta = 1). \tag{3}$$

All features (i.e. all genes) are set to 0 at the root node. For each directed edge from a node u to one of its child nodes v, each feature may switch to 1 with probability equal to the corresponding edge length. Finally, whenever a feature switches to 1, it stays at 1 for the remainder of the directed path. More formally, let $z_{j,u}$ denote the value of feature j at node u. The activation of feature j at child node v is determined by the following conditional probabilities:

$$P\left(z_{j,v} = 1 | z_{j,u} = 0, l_{(u,v)}\right) = l_{(u,v)}, \tag{4}$$

$$P\left(z_{j,v} = 1 | z_{j,u} = 1, l_{(u,v)}\right) = 1. \tag{5}$$

This models the notion that genes may be indicative of either broad or specific phenotypes. By allowing genes to be activated along different paths, the model also encompasses the idea that two sample groups may be homogeneous with regard to the same gene, albeit in different ways, as we shall see below in more detail. Notice that the above probability rules are defined without recurring to assignments of samples to paths, that is, they can be formally defined as being applied on the entire infinite tree. The probability rule in (5) is also a component of the phylogenetic Indian buffet process (pIBP) model [9]. The scope of the two models is however disparate, as in the pIBP the authors present a non-exchangeable prior for representing objects as infinite feature vectors, where object relations are given in the form of a pre-specified tree.

The path assignment and feature activation patterns determine the distribution of the expression data. Assume that feature j switches from 0 to 1 at node u. Denote the set of samples in the subtree that has u as its root by S_u, and the expression data for those samples restricted to feature j as $\boldsymbol{Y}_{j,S_u}$. Then,

$$\begin{aligned} \boldsymbol{Y}_{j,S_u} &\sim N\left(\mu_{j,u}\mathbf{1}, \sigma^2_{j,u}\boldsymbol{I}\right), && (6)\\ \mu_{j,u} &\sim N(\mu = 0, \sigma^2 = \sigma^2_{j,u}), && (7)\\ \sigma^2_{j,u} &\sim \text{Inv-Gamma}\,(a = 1, b = 1)\,. && (8) \end{aligned}$$

where $\mu_{j,u}$ and $\sigma^2_{j,u}$ are respectively scalar mean and variance parameters, specific to the group induced by feature j at node u. The prior distribution for each $\mu_{j,u}$ assumes adequately normalized data; the random variable $\sigma^2_{j,u}$ is given a vague prior distribution. Our choice of prior probability density functions allows us to analytically integrate out $\mu_{j,u}$ and $\sigma^2_{j,u}$, obtaining a multivariate Student-t distribution for $\boldsymbol{Y}_{j,S_u}$ [10]. This increases the efficiency of the sampler, although normality assumptions are in practice only an approximation whose usefulness is ultimately only validated by the results. If, for a given path ending in a leaf node u, a feature j never becomes activated, then, for every sample $s \in S_u$, we draw the corresponding scalar expression value $Y_{j,s}$ from a baseline Gaussian distribution, assuming standardized data,

$$Y_{j,s} \sim N(\mu_0 = 0, \sigma_0^2 = 1). \tag{9}$$

Notice that the same baseline distribution is used when required, regardless of the actual path or feature.

Figure 2 provides an example of an artificial data set with 3 genes and 4 samples generated using this approach.

2.2 Inference

We are interested in analyzing the joint posterior distribution of the path assignment and feature activation variables (respectively, $\boldsymbol{c}$ and $\boldsymbol{z}$), as well as γ, given the input expression data, which according to Bayes' rule is

$$P(\boldsymbol{c}, \boldsymbol{z}, \gamma|\boldsymbol{Y}) = \frac{P(\gamma)P(\boldsymbol{c}|\gamma)P(\boldsymbol{z})P(\boldsymbol{Y}|\boldsymbol{c}, \boldsymbol{z})}{P(\boldsymbol{Y})} \propto P(\gamma)P(\boldsymbol{c}|\gamma)P(\boldsymbol{z})P(\boldsymbol{Y}|\boldsymbol{c}, \boldsymbol{z}). \tag{10}$$

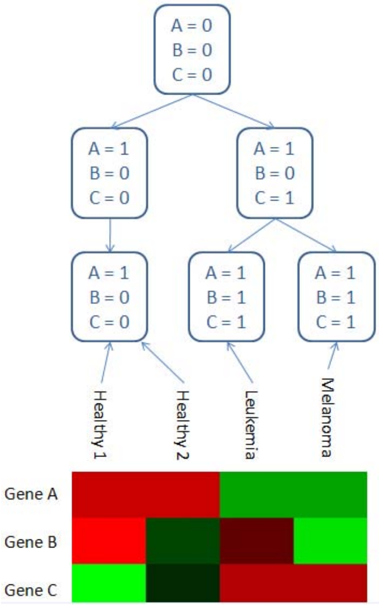

Fig. 2. Illustration of the generative process for a fictitious noise-free data set with 3 genes (A, B, and C) and 4 samples (”healthy 1”, ”healthy 2”, ”leukemia”, and ”melanoma”). The healthy samples share the same path assignment, while each of the cancer samples has its own unique path. The rounded rectangles represent nodes and indicate the current feature activation state. Gene A becomes active at both of the root’s child nodes, leading to homogeneous expression for the healthy samples as well as for the cancer samples, although the between-group difference in expression is significant. Gene C exhibits homogeneous expression under both cancer samples, but not under the healthy samples. Gene B has a specific expression pattern for each of the samples.

The term $P(\boldsymbol{z})$ results from integrating out all edge length variables, and the term $P(\boldsymbol{Y}|\boldsymbol{c}, \boldsymbol{z})$ results from integrating out all mean and variance variables. The posterior distribution is intractable and we approximate it by means of a collapsed Gibbs sampler [11,12].

Sampling path assignments. The posterior distribution for the path assignment of client i is given by

$$P(c_i|\boldsymbol{c}_{-i}, \boldsymbol{z}, \boldsymbol{Y}) \propto P(c_i|\boldsymbol{c}_{-i})P(\boldsymbol{Y}_{\cdot,i}|\boldsymbol{c}, \boldsymbol{z}, \boldsymbol{Y}_{\cdot,-i}), \tag{11}$$

where $\boldsymbol{c}_{-i}$ is the collection of path assignments for all clients except i, $\boldsymbol{Y}_{\cdot,-i}$ is the expression data for all features and all clients but i, and dependency on γ has been dropped from the notation for succinctness. See Fig. 3 for an illustration of path assignments. The number of available paths to choose from is equal to the total number of nodes in the current tree (discarding any previous path

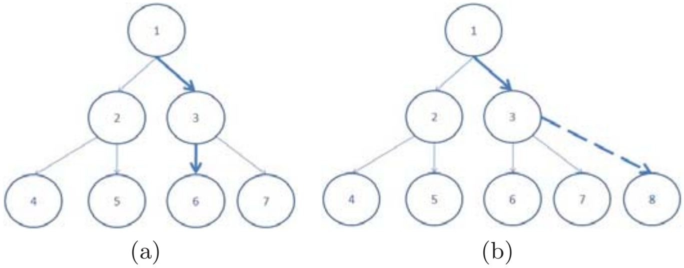

Fig. 3. Two possible paths for a new client, given a tree with 7 nodes. In 3(a), the path 1→3→6 does not involve the creation of new nodes. In 3(b), the path 1→3→8 implies adding a new node to the tree.

assignments of client i). The first term in (11) can be computed with (1). The second term can be decomposed into the following product:

$$P(\boldsymbol{Y}_{\cdot,i}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{Y}_{\cdot,-i}) = \left(\prod_{l=1}^{L}\prod_{j=1}^{G} P\left(Y_{j,i}|\boldsymbol{Y}_{j,S_{u_l}}\right)^{z_{j,u_l}(1-z_{j,p(u_l)})}\right)\prod_{j=1}^{G} P(Y_{j,i})^{1-z_{j,u_L}}. \tag{12}$$

Each node u_l corresponds to the node at the l-th level on the given path. We denote the parent of u by $p(u)$. The first term in (12) is interpretable as follows: For every node u_l in the path, we take the features that switch to 1 in that node. For each of those features j, we consider the clients assigned to paths that include u_l, and compute the predictive probability of the corresponding induced group generating the observed $Y_{j,i}$. It is straightforward to derive that the predictive distribution for each induced group is a univariate Student-t distribution [10]. The second term in (12) involves the features that are never activated in the path. For each of those, we must compute the probability that $Y_{j,i}$ was generated from a baseline Gaussian distribution, as described in (9). For every previously unpopulated section of a path, there needs to be an instantiation of the corresponding feature activation variables. We choose to draw them from their prior distribution. Since feature values are formally generated throughout the entire infinite tree, our approach conceptually corresponds to a type of *lazy loading*, where feature values are instantiated from their prior distribution as required. This implies that, although we are effectively bringing in novel feature variables into the model, their specific values do not contribute to the probability computations in (11). Alternative approaches involving simultaneously sampling path assignments and novel feature values are however possible.

Sampling feature values. The posterior odds for the value of feature j at node u are given by

$$\frac{P(z_{j,u}=1|\boldsymbol{c},\boldsymbol{z}_{-(j,u)},\boldsymbol{Y})}{P(z_{j,u}=0|\boldsymbol{c},\boldsymbol{z}_{-(j,u)},\boldsymbol{Y})} = \frac{P(z_{j,u}=1|\boldsymbol{z}_{-(j,u)})}{P(z_{j,u}=0|\boldsymbol{z}_{-(j,u)})}\frac{P(\boldsymbol{Y}_{j,\cdot}|z_{j,u}=1,\boldsymbol{z}_{-(j,u)},\boldsymbol{c})}{P(\boldsymbol{Y}_{j,\cdot}|z_{j,u}=0,\boldsymbol{z}_{-(j,u)},\boldsymbol{c})}, \tag{13}$$

where $\boldsymbol{z}_{-(j,u)}$ is the set of feature values excluding feature j at node u, $\boldsymbol{Y}_{j,\cdot}$ is the expression data restricted to feature j, and $\boldsymbol{z}_{j,\cdot}$ is the set of feature values for all nodes but restricted to feature j. Due to the conditional probability distributions specified in (4) and (5), some feature values are deterministic and thus do not require sampling. Namely, if a feature is set to 1 at a node u, then all values for that feature at any node v descendant from u must be equal to 1. This entails that the process of sampling a feature value corresponds to incrementing or decrementing the feature's generality level for a specific path.

The first term in (13) is given by

$$\frac{P(z_{j,u}=1|\boldsymbol{z}_{-(j,u)})}{P(z_{j,u}=0|\boldsymbol{z}_{-(j,u)})}=\frac{\alpha+n_{u+}^{-j}}{\beta+n_{u-}^{-j}}, \tag{14}$$

where n_{u+}^{-j} is the number of features that switched from 0 to 1 when traversing the edge (w,u) (w being the parent node of u) and n_{u-}^{-j} is the number of features that were kept at 0 when traversing that same edge, with both parameters disregarding feature j. The second term in (13) is given by

$$\frac{P(\boldsymbol{Y}_{j,\cdot}|z_{j,u}=1,\boldsymbol{z}_{-(j,u)},\boldsymbol{c})}{P(\boldsymbol{Y}_{j,\cdot}|z_{j,u}=0,\boldsymbol{z}_{-(j,u)},\boldsymbol{c})}=\frac{P(\boldsymbol{Y}_{j,S_u}|z_{j,u}=1,\boldsymbol{c})}{\prod_{i=1}^{d_u}P(\boldsymbol{Y}_{j,S_{v_i}}|z_{j,v_i}=1,z_{j,u}=0,\boldsymbol{c})}, \tag{15}$$

where v_i is the i-th child node of u, and d_u is the total number of child nodes of u. Both the numerator and the terms in the denominator correspond to multivariate Student-t distributions. In the numerator, all samples under node u are assumed to form a group with respect to feature j. In the denominator, samples instead form subgroups, each of them homogeneous with respect to feature j, but without assuming between-group homogeneity.

Sampling $\boldsymbol{\gamma}$. We sample the variable γ by use of an auxiliary variable scheme developed for Dirichlet process mixture models [13,14]. The procedure presented here is identical to the one in the hierarchical Dirichlet process model [14]. For a given node u in the tree, let d_u be the number of its child nodes and n_u be the number of samples assigned to it. It can be shown [15] that d_u is distributed as

$$P(d_u|\gamma,n_u)\propto\gamma^{d_u}\frac{\Gamma(\gamma)}{\Gamma(\gamma+n_u)}, \tag{16}$$

where terms that do not depend on γ have been discarded. Multiplying the above over all nodes in the tree yields

$$P(d_1,\ldots,d_V|\gamma,n_1,\ldots,n_V)\propto\prod_{u=1}^{V}\gamma^{d_u}\frac{\Gamma(\gamma)}{\Gamma(\gamma+n_u)}, \tag{17}$$

where the product is taken across all nodes u that are not leaf nodes, and V designates the number of those nodes. The posterior distribution for γ depends exclusively on its prior distribution from (2) and the above product. The main idea behind this sampling scheme is to represent each fraction of Gamma functions as

$$\frac{\Gamma(\gamma)}{\Gamma(\gamma+n_u)} = \frac{1}{\Gamma(n_u)} \int_0^1 w_u^\gamma (1-w_u)^{n_u-1} \left(1+\frac{n_u}{\gamma}\right) dw_u, \tag{18}$$

where $w_u \in [0,1]$ is an auxiliary variable. Define $\boldsymbol{w} = (w_u)_{u=1}^V$, introduce an extra vector of binary auxiliary variables $\boldsymbol{b} = (b_u)_{u=1}^V$, and specify the joint distribution of γ, $\boldsymbol{w}$, and $\boldsymbol{b}$ as

$$q(\gamma, \boldsymbol{w}, \boldsymbol{b}) \propto \gamma^{a_\gamma - 1 + d.} e^{-\gamma b_\gamma} \prod_{u=1}^V w_u^\gamma (1-w_u)^{n_u-1} \left(\frac{n_u}{\gamma}\right)^{b_u}, \tag{19}$$

where we have used dot $(\cdot)$ notation for vector summation. Marginalizing the auxiliary variables from the above joint distribution yields the original posterior distribution for γ [14,13]. Gibbs sampling updates are then given by

$$q(\gamma|\boldsymbol{w}, \boldsymbol{b}) \propto \gamma^{a_\gamma - 1 + d. - b.} e^{-\gamma(b_\gamma - \sum_{u=1}^V \log w_u)}, \tag{20}$$

$$q(w_u|\gamma) \propto w_u^\gamma (1-w_u)^{n_u-1}, \tag{21}$$

$$q(b_u|\gamma) \propto \left(\frac{n_u}{\gamma}\right)^{b_u}. \tag{22}$$

Visual inspection of the sampled values shows that the sampler converges under 50 iterations.

2.3 Information Retrieval

Generative models offer a natural measure of pairwise object relevance. Consider an arbitrary probabilistic model parameterized by $\boldsymbol{\theta}$ with input data $\boldsymbol{X}$. Assume a query object q, corresponding to the data point x_q, and a potentially relevant object r. Denote the parameters relating to r as $\boldsymbol{\theta}_r$. The relevance of r to q can be defined as

$$rel(q,r) \overset{\text{def}}{=} \int_{\boldsymbol{\theta}} P(x_q|\boldsymbol{\theta}_r) P(\boldsymbol{\theta}|X) d\boldsymbol{\theta} \tag{23}$$

[16]. This measure can be interpreted as the expected probability that the data point corresponding to object q was generated with the parameters from object r. A standard approximation is to obtain an estimate $\hat{\boldsymbol{\theta}}$ and compute $P(x_q|\hat{\boldsymbol{\theta}}_r)$. Notice that this measure is not symmetric.

In our context, the relevance of a sample r to another sample q can be defined as the expected probability that the expression data $\boldsymbol{Y}_{\cdot,q}$ was generated with the path variable c_r. This implies that any two samples r_1 and r_2 with equal path assignments ($c_{r_1} = c_{r_2}$) are equally relevant to a query sample q. Thus, in this model the proposed relevance measure works at node granularity. Averaging (23) over samples yields an estimate of between-node relevance, although we have not explored this possibility in the present work. We approximate (23) by using only the sample with the highest posterior probability, generated via the described Gibbs sampler.

3 Results

We tested our model on a collection of 199 miRNAs profiled in 218 human healthy tissues, tumors, and cell lines. We pre-processed the data set and standardized the resulting expression data in a gene-wise fashion, as originally described [6]; this makes the data set coherent with the parameter choices stipulated in the previous section. We ran the Gibbs sampler for 2500 burn-in iterations and further 2500 iterations, collecting the sample with the highest posterior probability. The path and feature variables were initialized with a draw from their prior. The maximum tree depth was fixed at 3, which is the lowest number that allows the model to form a sample hierarchy. The method took about 14 hours to run on an AMD Opteron Dual Core Processor with 2.8GHZ.[4] This procedure was repeated 30 times. In the following analysis, we considered the sample with the overall highest posterior probability.

3.1 Comparison to Previous Work

We compared the performance of our method to that of 4 well-established biclustering approaches [17,18,19,20] with default parameterizations. The results are presented in table 1. As miRNAs are known to have tissue-specific expression profiles [21], we first tested for the enrichment of specific tissues in the obtained biclusters. Significance was computed by means of Bonferroni-corrected hypergeometric tests with an original p-value of 0.01. Our method, named TreeBic, had the highest fraction of biclusters enriched for at least one tissue; at the other extreme, the CC method failed to significantly cluster samples from the same tissues in any bicluster. Our method, along with Samba, also managed to obtain the highest number of tissues enriched in at least one bicluster. Next, we assessed the functional homogeneity of each bicluster. We extracted a collection of confirmed miRNA targets from the TarBase database [22]. For each bicluster, we took the corresponding miRNAs and obtained the union of their targets. We then computed the functional enrichment of Gene Ontology (GO) [23] biological process terms in each target set, again using a Bonferroni-corrected hypergeometric test with an original p-value of 0.01 (terms with 5 or less genes were discarded). Our method outperforms all others with respect to the number of enriched GO categories. The biclusters found by our method also appear to be overall more functionally homogeneous, as shown by the percentage of biclusters enriched for at least one GO category. The overall low number of enriched GO categories is possibly due to the current sparsity of confirmed microRNA targets. Despite these results, our method has the second-lowest number of biclusters.

[4] Preliminary experiments on an artificial data set with 218 samples and 5970 features indicate that the same simulation takes approximately 250 hours, with the average number of nodes in the inferred tree being 145. Path variable sampling takes approximately 95% of inference time, indicating that a combination of heterogeneous features and high sample size leading to a large tree is the main bottleneck in the method.

Table 1. Method comparison with regard to tissue and miRNA target gene functional enrichment. Our model is named TreeBic; it outperforms 4 standard biclustering methods both in the fraction of biclusters enriched for at least one tissue/GO category and in the total number of enriched tissues/GO categories. See text for details on the meaning of each performance measure.

	TreeBic	Samba[17]	Plaid[18]	CC[19]	OPSM[20]
# Biclusters	16	54	29	20	10
% Tissue-Enriched Biclusters	**63%**	50%	41%	0%	40%
% GO Term-Enriched Biclusters	**63%**	46%	0%	18%	60%
# Enriched Tissues	**14**	**14**	8	0	2
# Enriched GO Terms	**12**	11	0	4	9

This suggests that the inferred hierarchical structure allows for a more efficient representation of the signal in the data set. Overall, by performing best both in terms of the fraction of enriched biclusters and the total number of enriched tissue and GO categories, our method appears to dominate over the other tested approaches.

3.2 Information Retrieval

In previous work we have shown that graphical models are useful in deriving object relevance measures that allow performing information retrieval in gene expression data in both an efficient and interpretable manner [24]. Here, we performed a feasibility study on the ability of the model to retrieve samples from the same tissue as a query sample. For a query taken from one of the 218 samples, we defined as positive the samples with the same tissue as the query. We computed true-positive and false-positive rates at each point in the relevance-ranked list of leaf nodes, and summarized the measures with the area under the corresponding ROC curve (AUC) [25]. For each tissue or cell line class, we computed the median of the AUC. Out of 20 classes, 13 (65%) led to a median AUC higher than the 0.5 baseline. The list of ranked results can provide important biological insight. As a case study, we queried the system with a follicular cleaved lymphoma sample. The method considers a sequence of 2 melanoma samples, 7 follicular cleaved lymphoma samples, and 6 large B-cell lymphoma samples as the most relevant. Although melanoma is a malignancy of a different cell type than non-Hodgkin lymphoma, there is epidemiological evidence for their association [26], a relation which is highlighted in our results and which we further investigate below. The practical usefulness of this model for information retrieval remains to be further assessed.

3.3 Biological Analysis

Figure 4 portrays the inferred sample tree. The method separates samples into organs from the reproductive system (node 1, with the exception of ovary, which falls under node 4), malignancies (nodes 2 and 3), and organs from the gastrointestinal tract (node 6). The method isolates the only two brain tissue samples in the data set, with a potential explanation being that they are the only healthy samples of ectodermal origin in the data set, in contrast with e.g. organs from node 6, which are of endodermal origin. On the other hand, node 4 appears to contain a more heterogeneous set of enriched tissues and pathological entities,

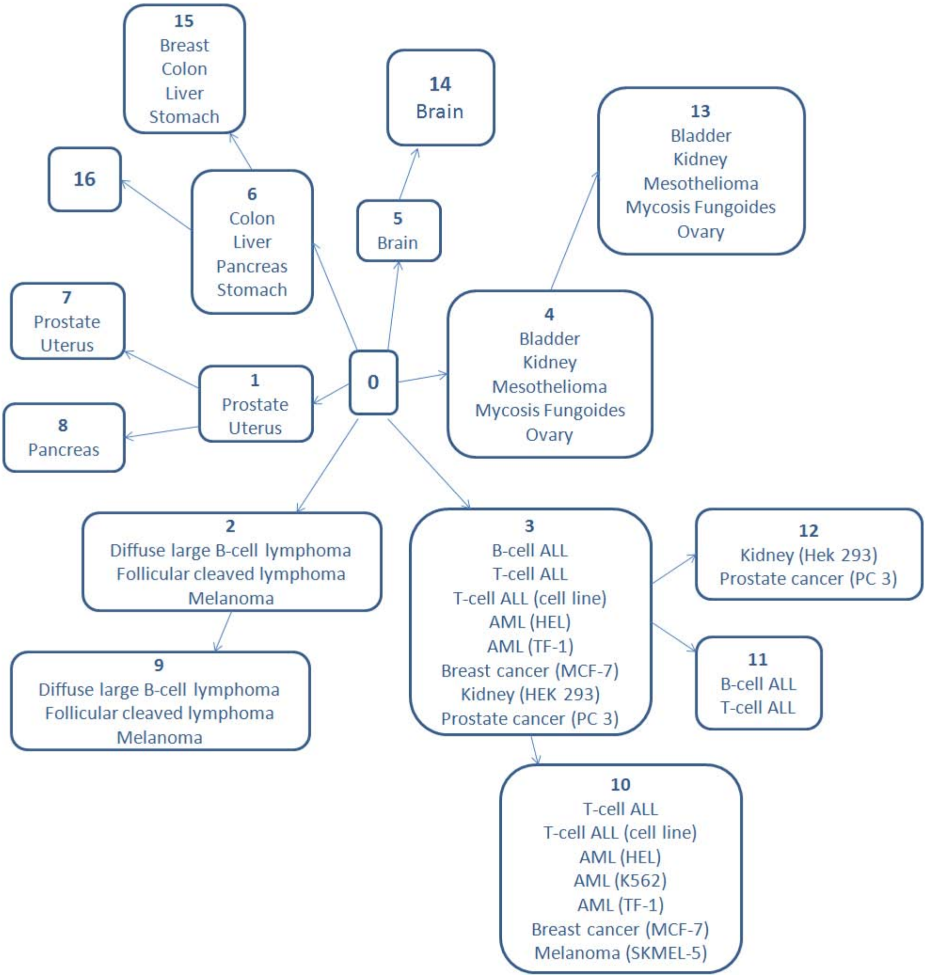

Fig. 4. Inferred tree structure. Nodes are numbered in breadth-first order and labelled with overrepresented tissues or cell lines (FDR q-value < 0.25). The non-stringent q-value enables richer node annotations. Some of the tissue types are overrepresented in more than one leaf node (e.g. T-cell ALL in nodes 10 and 11). Notice that this annotation approach does not guarantee that significant tissues in a parent node are also significant in the corresponding child nodes (e.g. nodes 6 and 15). Node 16 did not have any significantly overrepresented tissues.

Table 2. Genes differentially over-expressed between two melanoma sample groups (designated as types A and B) [28]. Genes predicted to be miR-224 targets are in bold text.

Gene Function	Over-Expressed in Type A	Over-Expressed in Type B
Pro-Apoptotic	**APAF1**, BAD, BNIP1, **BNIP3L**, CASP1, **CASP7**, CYCS, **VDAC1**	BAK1, **CASP2**, CASP4, **ENDOG**, HTRA2, PDCD5, PRODH, SEPT4, TNFSF10
Anti-Apoptotic	**BCL2**, BCL2A1, PPARD, RAF1	**API5**, **FIS1**, PPP2CA, PPP2R1A, **PPP2R1B**, **PSEN1**
Antioxidant	GLRX2, GPX4, GSR, MT3, **PRDX3**, PRDX5	ATOX1, **CAT**, GSS, HSPD1, **SOD1**

including a combination of healthy (bladder, kidney, and ovary) and cancerous (mesothelioma, mycosis fungoides) tissues. The method is also able to further decompose leukemias (node 3) into leukemia cell lines (node 10) and leukemic tissue (node 11).

The previously mentioned relation between melanoma and non-Hodgkin lymphoma is also hinted at by the contents of node 2. In order to find miRNAs with a role specifically in both melanoma and lymphoma, we computed the set difference between miRNAs that are activated in the melanoma and lymphoma nodes and those which are activated in any of the other haematological malignancy nodes. The single resulting miRNA, miR-224, is known to have a dual function, conditionally inducing both apoptosis and cell proliferation, and it was found to be either over or under-expressed in several tumor types [27]. In order to grasp potential mechanisms by which miR-224 may have a common role in melanoma and lymphoma, we first analyzed a collection of 38 genes that were found to be differentially over-expressed between two subsets of melanoma samples in an independent study (designated as type A and B) [28]. We used a recent miRNA target prediction algorithm [29] to compute which of those genes are potential miR-224 targets (Table 2). The prediction that 50% of type-A pro-apoptotic genes and 67% of type-B anti-apoptotic genes are regulated by miR-224 is evidence of its dual role in cell proliferation and apoptosis, and indicative that it may have an important post-transcriptional regulatory effect in melanoma. The role of miR-224 in stimulating proliferation is not well understood [27]. We hypothesize that it may enhance proliferation by targeting some of the predicted type-A pro-apoptotic genes. The anti-apoptotic gene API-5, recently proposed as a target for cancer treatment [36], is known to be targeted by miR-224 [30], and its protein product interacts with FGF-2 [31], which has in turn been observed to have increased levels of expression in patients with haematological malignancies, including lymphoma [32]. There is also evidence that miR-224 directly binds CD40 [33], which is known to have an important role both in lymphoma [34] and melanoma [35]. Together, these results indicate miR-224 may be an important element in explaining the association between melanoma and non-Hodgkin lymphoma. Although this analysis is speculative, it brings out the model's ability to generate hypotheses and drive the biological analysis.

4 Conclusions

We have introduced a graphical model which allows grouping microarray samples and providing an interpretation basis for that grouping. The model makes the assumption that samples are grouped in a tree structure, where nodes correspond to hierarchical subgroups, and where each node is associated with a subset of genes for which the corresponding samples are highly homogeneous. We applied the model to a large miRNA data set, where it was shown to outperform other biclustering approaches. We then provided a case study that depicts how the model variables and information retrieval formulation can be used to direct the biological analysis. The case study highlighted the potential role of miR-224 in the association between melanoma and non-Hodgkin lymphoma.

The current model may be extended in several ways. While in the present work we fixed the maximum tree depth at a specific level, selection of the appropriate depth may be conducted by recurring to cross-validation measures or by enhancing the model with an automatic depth selection capability. The assumption that each sample chooses a single path allows for the use of a flexible prior over trees that also makes computations feasible. This assumption can be relaxed, although it may lead to slower mixing during inference. Finally, alternative feature activation models may be devised, incorporating notions such as e.g. pathway enrichment among genes activated throughout the same edges.

Acknowledgments. We thank Leo Lahti for helpful comments. This work was supported by the Finnish Funding Agency for Technology and Innovation (TEKES, grant no. 40101/07). J.C. and S.K. belong to the Finnish Centre of Excellence on Adaptive Informatics Research, and were additionally supported by the Pattern Analysis, Statistical Modelling and Computational Learning Network of Excellence (PASCAL 2 EU Network of Excellence, grant no. ICT 216886). J.C. is additionally supported by a doctoral grant from the Portuguese Science and Technology Foundation (FCT, grant no. SFRH/BD/35974/2007).

References

1. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster Analysis and Display of Genome-Wide Expression Patterns. P. Natl. Acad. Sci. U.S.A. 95, 14863–14868 (1998)
2. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Trans. Comput. Biol. Bioinform. 1, 24–45 (2004)
3. Jordan, M.I. (ed.): Learning in Graphical Models. MIT Press, Cambridge (1999)
4. Flaherty, P., et al.: A Latent Variable Model for Chemogenomic Profiling. Bioinformatics 21, 3286–3293 (2005)
5. Gerber, G.K., et al.: Automated Discovery of Functional Generality of Human Gene Expression Programs. PLoS Comput. Biol. 3, 1426–1440 (2007)
6. Lu, J., et al.: MicroRNA Expression Profiles Classify Human Cancers. Nature 435, 834–838 (2005)
7. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The Nested Chinese Restaurant Process and Bayesian Inference of Topic Hierarchies. J. ACM (to appear)

8. Aldous, D.: Exchangeability and Related Topics. In: École d'été de probabilités de Saint-Flour, XIII, pp. 1–198. Springer, Berlin (1985)
9. Miller, K.T., Griffiths, T.L., Jordan, M.I.: The Phylogenetic Indian Buffet Process: A Non-Exchangeable Nonparametric Prior for Latent Features. In: McAllester, D., Myllymaki, P. (eds.) Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, pp. 403–410. AUAI Press, Corvallis (2008)
10. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. Chapman & Hall/CRC, Boca Raton (2004)
11. Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, Boca Raton (1996)
12. Liu, J.S.: The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. J. Am. Stat. Assoc. 89, 958–966 (1994)
13. Escobar, M.D., West, M.: Bayesian Density Estimation and Inference Using Mixtures. J. Am. Stat. Assoc. 90, 577–588 (1995)
14. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. J. Am. Stat. Assoc. 101, 1566–1581 (2006)
15. Antoniak, C.E.: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Ann. Stat. 2, 1152–1174 (1974)
16. Buntine, W., et al.: A Scalable Topic-Based Open Source Search Engine. In: Zhong, N., et al. (eds.) Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence, pp. 228–234. IEEE Computer Society, Los Alamitos (2004)
17. Tanay, A., Sharan, R., Shamir, R.: Discovering Statistically Significant Biclusters in Gene Expression Data. Bioinformatics 18, S136–S144 (2002)
18. Lazzeroni, L., Owen, A.: Plaid Models for Gene Expression Data. Stat. Sinica 12, 61–86 (2002)
19. Cheng, Y., Church, G.M.: Biclustering of Expression Data. In: Bourne, P., et al. (eds.) Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 93–103. AAAI Press, Menlo Park (2000)
20. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem. In: Istrail, S., Waterman, M.S., Clark, A.G. (eds.) Proceedings of the Sixth Annual International Conference on Computational Biology, pp. 49–57. ACM, New York (2002)
21. Landgraf, P., et al.: A Mammalian MicroRNA Expression Atlas Based on Small RNA Library Sequencing. Cell 129, 1401–1414 (2007)
22. Papadopoulos, G.L., Reczko, M., Simossis, V.A., Sethupathy, P., Hatzigeorgiou, A.G.: The Database of Experimentally Supported Targets: A Functional Update of TarBase. Nucleic Acids Res. 37, D155–D158 (2008)
23. Ashburner, M., et al.: Gene Ontology: Tool for the Unification of Biology. Nat. Genet. 25, 25–29 (2000)
24. Caldas, J., et al.: Probabilistic Retrieval and Visualization of Biologically Relevant Microarray Experiments. Bioinformatics 25, i145–i153 (2009)
25. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
26. Lens, M.B., Newton-Bishop, J.A.: An Association Between Cutaneous Melanoma and Non-Hodgkin's Lymphoma: Pooled Analysis of Published Data with a Review. Ann. Oncol. 16, 460–465 (2004)
27. Wang, Y., Lee, C.G.L.: MicroRNA and Cancer: Focus on Apoptosis. J. Cell. Mol. Med. 13, 12–23 (2009)
28. Su, D.M., et al.: Two Types of Human Malignant Melanoma Cell Lines Revealed by Expression Patterns of Mitochondrial and Survival-Apoptosis Genes: Implications for Malignant Melanoma Therapy. Mol. Cancer Ther. 8, 1292–1304 (2009)

29. Kertesz, M., et al.: The Role of Site Accessibility in MicroRNA Target Recognition. Nat. Genet. 39, 1278–1284 (2007)
30. Wang, Y., et al.: Profiling MicroRNA Expression in Hepatocellular Carcinoma Reveals MicroRNA-224 Up-regulation and Apoptosis Inhibitor-5 as a MicroRNA-224-specific Target. J. Biol. Chem. 283, 13205–13215 (2008)
31. Van den Berghe, L., et al.: FIF [Fibroblast Growth Factor-2 (FGF-2)-Interacting-Factor], a Nuclear Putatively Antiapoptotic Factor, Interacts Specifically with FGF-2. Mol. Endochrinol. 14, 1709–1724 (2000)
32. Krejci, P., et al.: FGF-2 Expression and its Action in Human Leukemia and Lymphoma Cell Lines. Leukemia 17, 817–819 (2002)
33. Mees, S.T., et al.: Involvement of CD40 Targeting Mir-224 and Mir-486 on the Progression of Pancreatic Ductal Adenocarcinomas. Ann. Surg. Oncol. 16, 2339–2350 (2009)
34. French, R.R., et al.: CD40 Antibody Evokes a Cytotoxic T-Cell Response that Eradicates Lymphoma and Bypasses T-Cell Help. Nat. Med. 5, 548–553 (1999)
35. Pirozzi, G., et al.: CD40 Expressed on Human Melanoma Cells Mediates T Cell Co-Stimulation and Tumor Cell Growth. Int. Immunol. 12, 787–795 (2000)
36. Rigou, P., et al.: The Antiapoptotic Protein AAC-11 Interacts with and Regulates Acinus-Mediated DNA Fragmentation. EMBO J. 28, 1576–1588 (2009)