Mathematical Biosciences 268 (2015) 80–91

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Multimodal probabilistic generative models for time-course gene expression data and Gene Ontology (GO) tags

Prasad Gabbur[a,*], James Hoying[b], Kobus Barnard[a]

[a] University of Arizona, United States
[b] Cardiovascular Innovation Institute, University of Louisville, United States

ARTICLE INFO

Article history:
Received 28 August 2014
Revised 28 July 2015
Accepted 10 August 2015
Available online 17 August 2015

Keywords:
Gene expression
Ontology
Microarray
Model
Cluster

ABSTRACT

We propose four probabilistic generative models for simultaneously modeling gene expression levels and Gene Ontology (GO) tags. Unlike previous approaches for using GO tags, the joint modeling framework allows the two sources of information to complement and reinforce each other. We fit our models to three time-course datasets collected to study biological processes, specifically blood vessel growth (angiogenesis) and mitotic cell cycles. The proposed models result in a joint clustering of genes and GO annotations. Different models group genes based on GO tags and their behavior over the entire time-course, within biological stages, or even individual time points. We show how such models can be used for biological stage boundary estimation de novo. We also evaluate our models on biological stage prediction accuracy of held out samples. Our results suggest that the models usually perform better when GO tag information is included.

1. Introduction

The Gene Ontology (GO) Consortium (http://www.geneontology.org) maintains annotations of genes and gene products of eukaryotic cells using a standardized controlled vocabulary [1,2]. The GO is organized as three sub-ontologies, each being a directed acyclic graph (DAG), describing a particular gene attribute: Biological Process (BP), Molecular Function (MF), or Cellular Component (CC). As new knowledge is gained about the specific roles of genes or their products, the ontology is updated to reflect the new findings. Since eukaryotic organisms share a significant number of genes, i.e., similar nucleotide sequences with similar functions, the ontology helps biologists working on different model organisms share useful knowledge between them. The GO annotations are also useful to statisticians and computational biologists working with high throughput gene expression data (e.g., microarrays) to evaluate their methods and draw biological conclusions from them. Uses of the GO tags include biological analysis of differentially expressed genes and enrichment analysis of gene clusters [3] using various methods [4–6].

In general, previous research combining expression data and GO annotations has used the annotations in a second stage of analysis after a quantitative analysis of expression data. Specifically, genes are clustered using continuous expression data and then the clusters are examined for GO tag enrichment. We propose instead that using the two sources of evidence simultaneously in a probabilistic generative modeling framework offers substantive benefit. For example, the information in GO tags can help deal with biological and experimental noise in expression data. The reverse can be true as high throughput expression measurements can alleviate uncertainty in GO annotations arising from missing or incorrect annotations. Note that we assume a relevant subset of genes are already selected using a suitable prior analysis such as differentially expressed gene set selection [7,8].

Probabilistic models for multimodal data provide a way to combine information from multiple modalities. They have been used extensively in other areas of research such as linking words and image regions [9–12], to fuse information from different sources and draw useful inferences from them. In this work, we use similar probabilistic generative models to cluster GO terms associated with genes in conjunction with their expression profiles measured using microarrays. This helps to exploit the information available in GO annotations to account for some of the noise in the expression measurements [13]. For example, the inclusion of GO terms in clustering increases the likelihood of genes with similar GO annotations to lie within the same cluster even though their expression profiles may differ significantly due to noise.

We focus on time-course microarray experiments, where snapshots of gene expression are taken at predetermined instants of time. The observation interval is of biological interest such as a mitotic cell cycle [14–16] or the response of a tissue to an injury [17]. The main thrust of this paper is models that capture the joint probability distributions of temporal gene expression patterns and

* Corresponding author.Tel.: 5202476726.
E-mail address: pgabbur@email.arizona.edu (P. Gabbur).

http://dx.doi.org/10.1016/j.mbs.2015.08.007
0025-5564/

GO tags. These models are generative statistical models in that they describe how the data can be sampled. The first model we describe is the multimodal mixture model (MMM) which clusters genes simultaneously based on GO tags and expression patterns, which are considered conditionally independent given the cluster. The second model (Pooled-MMM), modifies MMM under the assumptions that the stages are known, and carry the bulk of the variation of interest in the underlying expression patterns. Under these two assumptions, it makes sense to assume that variation within a stage is due to noise processes, and thus we treat the temporal observations within a stage as independent measurements of the same underlying latent expression level. Although our approach has resemblances to biclustering [18], we adopt a statistical generative modeling framework and use expression data and tags jointly in a multimodal approach, which is different from clustering rows and columns together based on reducing squared residues. Our work is also similar to data integration approaches [19–21] but the modalities being integrated are different.

Notice that in MMM and Pooled-MMM we implicitly assume that GO tags are best linked to patterns over all stages within the experiment time frame. An alternative hypothesis is that the GO tags are best linked to a specific stage. Hence we introduce the stage specific multimodal mixture model (SS-MMM) where GO tags are assumed to be emitted from one of the stages. This complicates inference because the data is not informative regarding which stage is responsible for which GO tags, as GO tags are only associated with gene activity taken as a whole. This situation is similar to related work on associating words with images [9], where we assume that words come from image regions, but in training we only have labels (e.g., keywords) at the image level. Nonetheless, in analogy with that work, if we have multiple genes with overlap in their GO tags, and similar expressions at a particular stage, then we could learn relevant relationships between GO tags and stages. Finally, if the stage time windows are not known, but we wish to experiment with the temporal locality of the association between GO tags and expression levels, then we can simply use each time point as a stage, which leads to the time specific multimodal mixture model (TS-MMM).

*Evaluation:*We use biological stage (henceforth referred to as "stage") prediction performance on held out data to measure how well the proposed models capture the underlying biological process. A number of analyses of gene expression data aim at predicting the stage of a tissue, e.g., diseased or not, and it is thus sensible to use phenotype prediction performance to evaluate our models. For the SS-MMM and Pooled-MMM models we adopt a Bayesian approach that uses stage information learned as part of model training. To compute stages for the two models that do not explicitly represent stages (MMM and TS-MMM), we estimate the stage likelihood using time based marginal likelihoods. We study the effect of including GO tags by evaluating the models both with GO tags and without. Our results suggest that the use of GO tags is helpful for biological stage prediction.

Apart from biological stage prediction, when the number of stages are known for the training data, our generative models can be used to estimate stage boundaries de novo. We demonstrate this by applying our SS-MMM model to estimate stage boundaries first in the data sets for which we have ground truth information. The results suggest that our best estimates are within one time point of the true boundaries. This implies that our models could also be useful for estimating which time points can be grouped into stages where such information is not available a priori. We apply the same method to determine the best estimate of stage boundaries in a data set where this information is not available.

2. Methods

Definitions: A sample refers to the collection of measurements on a particular microarray. A microarray dataset is a collection of samples, where each sample is obtained by measuring a tissue under a certain biological condition. The vector of expression measurements for a gene g across all the arrays in the dataset is denoted by $\boldsymbol{e}^g$ and the set of GO tags associated with it is denoted by O^g. A particular sample in the data is indexed by t, which corresponds to a time point in the case of time-course experiments. The gene's expression value for the sample at that time is denoted by $\boldsymbol{e}_t^g$. In addition, the expression vector of all genes at the time point t is denoted by $\boldsymbol{e}_t$. While our methodology nicely supports multiple observations per gene for a time point, this does not occur in the data we experimented with, and hence this simplification is well defined. A *stage* s represents a particular phenotypic condition or biological stage. The time course datasets have distinct biological stages, each one containing a number of samples for the time points falling within the stage, referred to as a *pool* of samples within that stage.

2.1. Proposed models

In what follows, we first describe the MMM model that clusters genes based on their expression behavior across the entire time course and GO tags associated with them. We then describe the Pooled-MMM model, which assumes that observations at different time points within a stage are simply replicated measurements. In particular, the Pooled-MMM model uses this notion to constrain the expression levels within a biological stage to have the same average value, i.e., they come from the same stage-specific cluster.

Both the MMM and the Pooled-MMM associate GO tags with the expression behavior over the time series. However, it is possible that GO tags are better linked to a specific biological stage or even to a specific time point. The SS-MMM model addresses the former possibility by allowing GO tags to arise from any one of the biological stages. Finally the TS-MMM model allows for the GO tags to be associated with gene expression levels at any one of the individual time points addressing the latter possibility.

2.2. Multimodal mixture model (MMM)

In the multimodal mixture model (MMM), a gene g's expression vector $\boldsymbol{e}^g$ and its associated GO tags O^g jointly arise from a certain cluster c^g, which is one of a set of possible clusters of genes with similar expression profiles and GO annotations. The sampling of a particular cluster c^g is governed by a prior over clusters $P(c^g)$. The joint distribution $p(\boldsymbol{e}^g, O^g)$ is modeled as

$$p_{MMM}(\boldsymbol{e}^g, O^g) \propto \sum_{c^g} P(c^g) \left(\prod_t p(\boldsymbol{e}_t^g | c^g) \right) \left(\prod_{o \in O^g} (P(o|c^g))^{\frac{1}{|O^g|}} \right), \quad (1)$$

where $\boldsymbol{e}_t^g$ denotes the expression value of gene g at time point t. The choice of the number of clusters is discussed in Section 5.1. The above likelihood is defined up to a scaling factor but all the probabilities are appropriately normalized during training and inference, which applies to all the models proposed in this work. The GO tags O^g are generated independently of the expression values conditioned on the clusters. We use o to denote a particular GO term and $|O^g|$ for the total number of GO terms associated with gene g, which includes all ancestors in the GO DAG. The form of $P(o|c^g)$ is a multinomial with as many bins as the number of unique ontology terms associated with all genes. The exponent $\frac{1}{|O^g|}$ adjusts for different numbers of GO tags associated with different genes and lets each gene contribute the same towards the likelihood function over all the genes (Section 3). The form of $p(\boldsymbol{e}_t^g | c^g)$ is a univariate Gaussian. This model leads to a clustering of genes into groups where the genes within a group have similar expression profiles across all the time points and have similar GO tags. The discrete distributions $P(c^g)$, $P(o|c^g)$ and the Gaussian means and variances are parameters of the model that are learned during model training (Section 3).

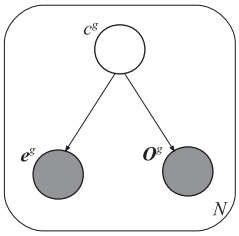


Fig. 1. Graphical representation of the multimodal mixture model (MMM) and its pooled version (Pooled-MMM). The shaded nodes represent random vectors that are observed, which in this case are the gene expression profile ($\boldsymbol{e}^g$) and the GO tags (O^g) of each gene g. It is assumed that there are N such genes and that their data is independently generated. The hidden node represents the latent cluster variable c^g responsible for generating the expression profile and the GO tags. The only difference between the MMM and Pooled-MMM models is that the parameters of Gaussians corresponding to pooled time points within a stage are shared, i.e. constrained to be the same in the latter case.

A graphical representation of the above model is shown in Fig. 1. The graph depicts a generative model for the expression profile $\boldsymbol{e}^g$ of a gene g and its associated GO tags O^g through the latent variable c^g corresponding to a cluster.

2.3. Pooled multimodal mixture model (Pooled-MMM)

The pooled multimodal mixture model (Pooled-MMM) is a special case of the MMM model of Section 2.2 and allows for pooling of gene expression levels across the time points corresponding to a biological stage. This is achieved by generating the expression levels at all time points t within a stage s from the same shared cluster parameters. This leads to the following expression for the joint distribution $p(\boldsymbol{e}^g, O^g)$:

$$p_{Pooled\text{-}MMM}(\boldsymbol{e}^g, O^g) \propto \sum_{c^g} P(c^g)\left(\prod_s p(\boldsymbol{e}_s^g|c^g)\right)\left(\prod_{o\in O^g}(P(o|c^g))^{\frac{1}{|O^g|}}\right), \tag{2}$$

where s corresponds to a stage and $\boldsymbol{e}_s^g$ is the gene expression (sub) vector, i.e. a pool of samples encompassing time points $t \in s$. The form of $p(\boldsymbol{e}_s^g|c^g)$ is a multivariate Gaussian with a diagonal covariance matrix with shared parameters, i.e. the means and variances are constrained to be the same value across all the time points of a stage. Note that similar to the MMM model, the GO tags are generated at the entire time-course level. The graphical model for the Pooled-MMM is the same as that of MMM except that the parameters across the pooled time points of a stage are shared as shown in Fig. 1.

2.4. Stage-specific multimodal mixture model (SS-MMM)

The stage-specific multimodal mixture model assumes that genes cluster independently at different biological stages, a gene's expression values at time points within a stage being noisy versions of the same latent value. Thus we pool the samples for each stage, and assume that the observations are conditionally independent, given the stage. The joint distribution $p(\boldsymbol{e}^g, O^g)$ is modeled as

$$\begin{aligned} p_{SS\text{-}MMM}(\mathbf{e}^g, O^g) &\propto p(\mathbf{e}^g)P(O^g|\mathbf{e}^g) \\ &\propto \prod_s \sum_{c_s^g} P\left(c_s^g\right) p\left(\mathbf{e}_s^g|c_s^g\right) \\ &\times \left(\prod_{o\in O^g}(P(o|\mathbf{e}^g))^{\frac{1}{|O^g|}}\right), \end{aligned} \tag{3}$$

where s refers to a biological stage or phenotype, consisting of a pooled set of time points $t \in s$. c_s^g represents the latent cluster variable for stage s that generates all the samples within the gene expression (sub) vector $\boldsymbol{e}_s^g$ corresponding to the pool of s. The form of $p(\boldsymbol{e}_s^g|c_s^g)$ is assumed to be a multivariate Gaussian with a diagonal covariance matrix. The means and variances of this multivariate Gaussian are shared across the pooled time points similar to the Pooled-MMM model. $P(o|\boldsymbol{e}^g)$ is a discrete distribution given by

$$P(o|\boldsymbol{e}^g) = \frac{1}{M}\sum_s \sum_{c_s^g} P(c_s^g|\boldsymbol{e}_s^g)P(o|c_s^g), \tag{4}$$

where the assumption is that each tag o is generated by the same set of stage specific clusters c_s^g within each stage s. However, the tag is randomly sampled from only one of the clusters within one of the stages. This sampling is with equal pool priors as indicated by the weighting factor $\frac{1}{M}$ in the above equation (M is the total number of stages) but with cluster priors conditioned on the pooled expression level(s) within the corresponding stage. The form of the conditional cluster priors can be deduced using Bayes' rule

$$P(c_s^g|\boldsymbol{e}_s^g) \propto p(\boldsymbol{e}_s^g, c_s^g) \tag{5}$$

$$= p(\boldsymbol{e}_s^g|c_s^g)P(c_s^g) \tag{6}$$

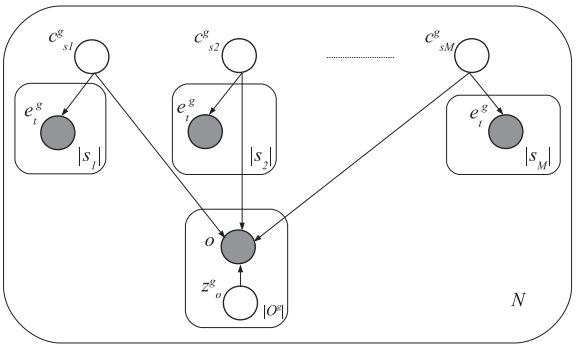


Fig. 2. Graphical representation of the stage-specific multimodal mixture model (SS-MMM) and time-specific multimodal mixture model (TS-MMM). In the case of SS-MMM, the gene expression measurements $\boldsymbol{e}^g_{t\in s_i}$ for all samples t within the pool belonging to a particular stage s_i are generated from the stage specific clusters denoted by $c^g_{s_i}$. The number of such pooled samples within a stage s_i is denoted by $|s_i|$. The GO tags O^g are generated independently, conditioned on the cluster, of the expression levels by the same process. However, each GO tag o comes from only one of the clusters within one of the stages. This choice is with a uniform prior on all the stages as indicated by the hidden variable z_o^g followed by a cluster prior conditioned on the expression level(s) within the chosen stage. It is assumed that there are M such stages and N genes in the microarray dataset. TS-MMM is a special case of SS-MMM with each stage s_i consisting of a single time point t.

Fig. 2 shows a graphical model representation of the SS-MMM model. In the figure, s_i corresponds to a particular stage that spans a subset of the available samples and $|s_i|$ denotes the total number of such samples. The clusters for the stage s_i are represented by the latent variable $c^g_{s_i}$. The shaded nodes $\boldsymbol{e}^g_{t\in s_i}$ represent the observed gene expression value for a gene g at any of the time points $t \in s_i$. The shaded node o represents a single GO tag within the set O^g of GO tags for gene g, and z_o^g is the hidden variable indicating the choice of a stage, among all stages, from one of whose clusters the tag o is randomly generated.

2.5. Time specific multimodal mixture model (TS-MMM)

Our fourth model (TS-MMM) is similar to the SS-MMM, but treats each time point as a stage. It thus addresses the possibility that suitable grouping of time points into biological stages is not available, or is not helpful. The joint distribution $p(\boldsymbol{e}^g, O^g)$ is modeled as

$$\begin{aligned} p_{TS\text{-}MMM}(\mathbf{e}^g, O^g) &\propto \prod_t \sum_{c_t^g} P\left(c_t^g\right) p\left(\mathbf{e}_t^g|c_t^g\right) \\ &\times \left(\prod_{o\in O^g}(P(o|\mathbf{e}^g))^{\frac{1}{|O^g|}}\right), \end{aligned} \tag{7}$$

which is a special case of the SS-MMM model of Section 2.4 with each stage s representing a single time point t. The expression for the GO tags distribution $P(o|\boldsymbol{e}^g)$ takes a similar form as Eq. (4) and is given by

$$P(o|\boldsymbol{e}^g) = \frac{1}{T}\sum_t \sum_{c_t^g} P(c_t^g|\boldsymbol{e}_t^g)P(o|c_t^g), \tag{8}$$

where the contributions from GO tags are averaged across all the time points t, as each observed tag is associated with each of the time points with equal prior probability. The graphical model representation of the TS-MMM is the same as that of SS-MMM except that each stage now corresponds to a single time point as shown in Fig. 2.

3. Model training and inference

Each of the above proposed models is trained using a suitable version of the expectation–maximization (EM) algorithm [22–24]. EM finds parameters that locally maximize the likelihood for all the genes assuming that the data for each gene is independently generated. That is it maximizes the likelihood function $L_I = \prod_g p(\boldsymbol{e}^g, O^g)$. The estimated parameters include the Gaussian mixture means, variances, cluster priors and cluster conditional ontology term distributions. Note that the cluster priors and cluster conditional ontology term distributions are multinomials with their bin probabilities estimated directly during the learning process.

3.1. Bayesian stage prediction for pooled sample models

The pooled sample models of Sections 2.3 and 2.4 encode stages, and learn a distribution of stages and observations during training. Given a new observation, $\boldsymbol{e}_U$, whose stage is unknown (U), and the set of GO tags O, the posterior distribution over stages $P(s|\boldsymbol{e}_U, O)$ can be written using Bayes' rule as

$$P(s|\boldsymbol{e}_U, O) \propto p(s, \boldsymbol{e}_U, O). \tag{9}$$

The form of $p(s, \boldsymbol{e}_U, O)$ depends on the particular model. For the Pooled-MMM model, we have

$$p(s, \boldsymbol{e}_U, O|\boldsymbol{\theta}) \propto \prod_g \sum_c p(\boldsymbol{e}_U^g|s, c, \boldsymbol{\theta})P(O^g|c, \boldsymbol{\theta})P(c|\boldsymbol{\theta}), \tag{10}$$

where $\boldsymbol{\theta}$ are the model parameters learned from training data using EM. We emphasize that only the (shared) Gaussian parameters of the cluster c within the stage s are used for computing the likelihood $p(\boldsymbol{e}_U^g|s, c, \boldsymbol{\theta})$ above. For the SS-MMM model of Section 2.4, it is proportional to the stage specific likelihood, i.e.,

$$\begin{aligned} p(s, \boldsymbol{e}_U, O|\boldsymbol{\theta}) &\propto p(\boldsymbol{e}_U, O|s, \boldsymbol{\theta}) \\ &\propto \prod_g \left(\sum_{c_s} P(c_s|\boldsymbol{\theta})p(\boldsymbol{e}_U^g|c_s, \boldsymbol{\theta})\right) \\ &\times \left(\prod_{o\in O^g}\left(\sum_{c_s} P(c_s|\boldsymbol{e}_U^g, \boldsymbol{\theta})P(o|c_s, \boldsymbol{\theta})\right)^{\frac{1}{|O^g|}}\right) \end{aligned} \tag{11}$$

3.2. Time-based stage prediction

The MMM and TS-MMM models do not represent biological stages and hence cannot be used directly to make inferences about the stage of a new sample. We propose a framework for these models to do stage prediction based on marginalizing over the time points of training samples using the estimated model parameters. The posterior distribution over stages s given the new measurement sample $\boldsymbol{e}_U$, whose stage is unknown (U), and the set of GO tag assignments O for all genes can be written as

$$P(s|\boldsymbol{e}_U, O) \propto p(s, \boldsymbol{e}_U, O) \tag{12}$$

$$= \sum_t p(t, s, \boldsymbol{e}_U, O) \tag{13}$$

$$= \sum_t P(s|t, \boldsymbol{e}_U, O)\, p(\boldsymbol{e}_t = \boldsymbol{e}_U, O) \tag{14}$$

$$= \sum_t P(s|t)\, p(\boldsymbol{e}_t = \boldsymbol{e}_U, O), \tag{15}$$

marginalizing over time. The term $p(\boldsymbol{e}_t = \boldsymbol{e}_U, O)$ is the density of the new data being observed at time corresponding to sample t of the training data, and is expanded further below. Note that we use the fact that the stage s is conditionally independent of both gene expressions and GO tags given the time point. The quantity $P(s|t)$ can be set to 1 or 0 depending on whether the sample t in the training data arises from stage s or not followed by appropriate normalization of the posterior. Marginalization implies that any of the possible time points within a stage might have contributed to the observed gene expression levels and their tags. The computation of $p(\boldsymbol{e}_t = \boldsymbol{e}_U, O)$ for each of the two models (Sections 2.2 and 2.5) is described next.

3.2.1. Marginal time densities

The maximum likelihood parameters $\boldsymbol{\theta}$, estimated by the EM procedure, imply a joint probability density function over the space of gene expression levels of different time points and GO terms. Given T time points, we denote this joint density by

$$p(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_T, O|\boldsymbol{\theta}), \tag{16}$$

where $\boldsymbol{e}_t$ is the vector of measurement of all genes for time point t and O is the set of GO term assignments to genes. We express the marginal joint density $p(\boldsymbol{e}_t = \boldsymbol{e}_U, O|\boldsymbol{\theta})$ for a new sample $\boldsymbol{e}_U$ by using that value in the slot for t in (16), and marginalizing out the other times:

$$\begin{aligned} &p(\boldsymbol{e}_t = \boldsymbol{e}_U, O|\boldsymbol{\theta}) \\ &= \int_{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{t-1},\boldsymbol{e}_{t+1},\ldots\boldsymbol{e}_T} p(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{t-1}, \boldsymbol{e}_U, \boldsymbol{e}_{t+1}, \ldots, \boldsymbol{e}_T, O|\boldsymbol{\theta}) \\ &\quad \times d\boldsymbol{e}_1 \ldots d\boldsymbol{e}_{t-1} d\boldsymbol{e}_{t+1} \ldots d\boldsymbol{e}_T \end{aligned} \tag{17}$$

In the general case this is complex, but recall that in MMM and TS-MMM, the genes are independent. If we rearrange the $d\boldsymbol{e}_i$ into blocks by gene, we see that (17) reduces to the product of the marginals for each gene. Specifically,

$$p(\boldsymbol{e}_t = \boldsymbol{e}_U, O|\boldsymbol{\theta}) = \prod_{g=1}^{N} p(\boldsymbol{e}_t^g = \boldsymbol{e}_U^g, O^g|\boldsymbol{\theta}), \tag{18}$$

where

$$\begin{aligned} &p(\boldsymbol{e}_t^g = \boldsymbol{e}_U^g, O^g|\boldsymbol{\theta}) \\ &= \int_{\boldsymbol{e}_1^g,\ldots,\boldsymbol{e}_{t-1}^g,\boldsymbol{e}_{t+1}^g,\ldots\boldsymbol{e}_T^g} p(\boldsymbol{e}_1^g, \ldots, \boldsymbol{e}_{t-1}^g, \boldsymbol{e}_U^g, \boldsymbol{e}_{t+1}^g, \ldots, \boldsymbol{e}_T^g, O^g|\boldsymbol{\theta}) \\ &\quad \times d\boldsymbol{e}_1^g \ldots d\boldsymbol{e}_{t-1}^g d\boldsymbol{e}_{t+1}^g \ldots d\boldsymbol{e}_T^g \end{aligned} \tag{19}$$

and the form of $p(\boldsymbol{e}_1^g, \ldots, \boldsymbol{e}_{t-1}^g, \boldsymbol{e}_U^g, \boldsymbol{e}_{t+1}^g, \ldots, \boldsymbol{e}_T^g, O^g|\boldsymbol{\theta})$ is given by the expression for the joint distribution of the model of interest. We derive expressions for MMM and TS-MMM next.

3.2.2. MMM

In the case of MMM we substitute the corresponding joint distribution (Eq. (1)) for $p(\boldsymbol{e}_1^g, \ldots, \boldsymbol{e}_{t-1}^g, \boldsymbol{e}_U^g, \boldsymbol{e}_{t+1}^g, \ldots, \boldsymbol{e}_T^g, O^g|\boldsymbol{\theta})$ in Eq. (19) and move the integral within the sum over clusters

$$\begin{aligned} &p(\mathbf{e}_t^g = \mathbf{e}_U^g, O^g|\boldsymbol{\theta}) \propto \sum_c P(c|\boldsymbol{\theta})P(O^g|c, \boldsymbol{\theta}) \\ &\times \int_{\mathbf{e}_1^g,\ldots,\mathbf{e}_{t-1}^g,\mathbf{e}_{t+1}^g,\ldots\mathbf{e}_T^g} p(\mathbf{e}_1^g, \ldots, \mathbf{e}_{t-1}^g, \mathbf{e}_U^g, \mathbf{e}_{t+1}^g, \ldots, \mathbf{e}_T^g|c, \boldsymbol{\theta}) \\ &\times d\mathbf{e}_1^g \ldots d\mathbf{e}_{t-1}^g d\mathbf{e}_{t+1}^g \ldots d\mathbf{e}_T^g. \end{aligned} \tag{20}$$

Using the cluster conditional independence assumption among gene expression levels across individual time points and GO tags (Eq. (1)) and also the fact that cluster conditional densities of expressions at each time point integrate to one, this becomes

$$p(\boldsymbol{e}_t^g = \boldsymbol{e}_U^g, O^g|\boldsymbol{\theta}) \propto \sum_c P(c|\boldsymbol{\theta})\, p(\boldsymbol{e}_U^g|c, \boldsymbol{\theta}) P(O^g|c, \boldsymbol{\theta}). \quad (21)$$

3.2.3. TS-MMM

Following an approach similar to the MMM model, the marginal $p(\boldsymbol{e}_t^g = \boldsymbol{e}_U^g, O^g|\boldsymbol{\theta})$ is given by

$$p(\boldsymbol{e}_t^g = \boldsymbol{e}_U^g, O^g|\boldsymbol{\theta}) \propto \sum_{c_t} P(c_t|\boldsymbol{\theta})\, p(\boldsymbol{e}_U^g|c_t, \boldsymbol{\theta}) P(O^g|c_t, \boldsymbol{\theta}), \quad (22)$$

where the GO tag probabilities $P(O^g|c_t, \boldsymbol{\theta})$ are evaluated using only the cluster parameters of the tth time point in Eq. (8) because the hidden variable z^o for every tag o is now assigned to time point t (Fig. 2).

3.3. Stage boundary estimation

The learned parameters of the pooled sample models can be used to estimate stage boundaries in a time course data set. This can be done by examining the training likelihood L_I given a grouping of the time points into stages. If the total number of time points is small and the number of stages is assumed, then it is feasible to perform a brute force search over all possible groupings as we do for the experiments described below. The grouping that leads to the highest likelihood is taken to be the best estimate of the stage boundaries given the assumed number of stages.

4. Datasets and experimental protocol

We used data from three time-course experiments to train our models and then predict the biological stage labels using the estimated model parameters: angiogenesis data of Hoying et al. [25,26], yeast cell cycle data of Cho et al. [15], and human cell cycle data of Whitfield et al. [16]. The predicted label is the most likely stage label obtained from the posterior distribution $P(s|\boldsymbol{e}_U, O, \boldsymbol{\theta})$ using the learned model parameters with ties resolved randomly.

4.1. Hoying angiogenesis data

The Hoying angiogenesis data came from an experimental model (in vivo) of tissue vascularization in SCID mice with the implants obtained from tie2:GFP mouse [25]. Tissue samples were extracted from the implanted constructs at discrete time points – days 3, 7, 14, 21 and 28. In a unique experimental design these samples and a day 0 sample (implant source) were hybridized using two channel microarrays to obtain measurements of gene expression. Care was taken so that biological variations were averaged out in the measurements. There were 4 measurements per gene per time point. The intensity measurements were background subtracted and lin-log transformed [26] for variance stabilization and only those measurements that were consistently well above the background level were retained. The transformed measurements were corrected for spatial location and intensity variations using *lowess* regression in a custom statistical software called CARMA [26]. Then a gene-by-gene ANOVA was performed to adjust for the gene specific variations introduced by the experimental factors namely the Array (A), Dye (D) and Variety (V) effects. The V effect corresponded to time and was used as the time point specific level of gene expression. Data from two such hybridization runs (run-1 and run-2) was used for the experiments.

To evaluate prediction accuracies, we trained models using measurements from each run separately. Held-out accuracies were computed by predicting labels for samples from the other run. In other words, since we do not have any notion of ground truth for this data, we validated the models on the assumption that the stages for the two runs were the same. Only differentially expressed genes selected by CARMA were used. Since the differentially expressed genes that were selected were different for different runs we included only the ones that were common to both. Starting from 1282 genes that were selected as differentially expressed in run-1, we searched for their corresponding normalized expression levels in run-2. Only 706 of them had any valid measurements in run-2 based on ANOVA and so only these were used for one set of experiments. Similarly, starting from the 978 differentially expressed genes of run-2, we found 932 of them having any valid normalized expression values in run-1 based on ANOVA.

Two sets of models were trained using GO tags from the Biological Process (BP) and Molecular Function (MF) GO hierarchies respectively and one without any tags (No GO). For data with 706 genes, this resulted in 480 and 363 total GO tags from the BP and MF hierarchies respectively. The total number of GO tags for the other data with 932 genes were 569 and 400 from the BP and MF hierarchies, respectively. Note that we included all the ancestors of the GO annotations for the genes by tracing the GO hierarchies starting from each annotation all the way to the root of the corresponding hierarchy.

We considered stage prediction corresponding to our hypothesis of the two stages of blood vessel growth (angiogenesis and maturation). During the first stage (*angiogenesis*), relevant microvessel segments relax their normal vessel structure leading to the expansion of the microvasculature via the addition of new vessel segments. Subsequently, the newly formed vessels differentiate into the varied elements of a normal vasculature including arterioles, venules and capillaries finally leading to a mature vascular network (*maturation*) through vessel adaptation. We hypothesized that the first two time points correspond to angiogenesis and the next four to maturation. This is based on the experimental results described in Section 5.2. For the stageless models of Sections 2.5 and 2.2, stage probabilities were computed from marginal time probabilities as described in Section 3.2. In each case predicted stage was the one with the highest probability.

4.2. Cho yeast cell cycle data

The yeast cell cycle data has been used to identify and study the periodic fluctuations of mRNA levels of cell cycle regulated genes. Synchronized cells derived from cultures of *Saccharomyces cerevisiae* and arrested at time 0 were allowed to go through mitotic cell division over a period of 160 min. The method of synchronization was based on the temperature sensitive *cdc28-13* allele. Expression levels of all the genes in the yeast genome were sampled at 10 min intervals resulting in a total of 17 measurements including time 0. The cells underwent almost two complete cell divisions within this period and hence the data represents measurements over two cell cycles. The transition from the first to second cycle is approximately at about 90 min from the start. Each cell cycle is divided into 4 phases: G1, S, G2, and M based on bud size, cellular position of nucleus and standardization to previously known cell cycle regulated genes [15]. The S phase corresponds to genome duplication, M phase to nuclear division, which are separated from each other by the two gap phases G1 and G2.

The raw yeast cell cycle data was preprocessed by log transformation followed by geometric normalization. Measurements from both the cell cycles were used and prediction accuracies were computed by training models on data from each cell cycle separately. Stage prediction accuracies were computed in a similar fashion as the Hoying data. Table 1 lists the time samples corresponding to the various phases of cell cycle for both the cycles.

Previous work using this dataset identified 421 [15] genes as cell cycle regulated. We used only the measurements for these genes in

Table 1
Assignment of time points to various stages of the yeast cell cycle in the Cho dataset. The data has two complete cell cycles with the transition from the first to second cycle occurring at 90 min from start. The assignment of time samples from the two cycles are listed in separate columns.

Phase	Time points (cycle-1)	Time points (cycle-2)
G1	0, 10, 20	90, 100
S	30, 40	110, 120
G2	50, 60	130, 140
M	70, 80	150, 160

Table 2
Assignment of time points to the S and Non-S phases of the human cell cycle in the Whitfield dataset. This dataset has 3 complete cell cycles with samples recorded every hour over a period of 46 h, with a different number (14, 14 and 17) and assignment of samples to the two phases of each cycle. Each column shows the phase assignment for the samples, denoted by the measurement hour, within a cycle.

Phase	Time points (cycle-1)	Time points (cycle-2)	Time points (cycle-3)
S	0–3	14–18	28–32
Non-S	4–13	19–27	33–44

our experiments. This resulted in a total of 1275 and 491 tags including the ancestors in the BP and MF GO hierarchies respectively. Models were trained with (BP or MF) and without GO tags followed by stage prediction.

4.3. Whitfield human cell cycle data

Similar to the yeast, genes regulated periodically during the human cell cycle have been identified using the HeLa cancer cell line [16]. Three different synchronization methods were used to arrest cells in the S phase (double thymidine block) and M phase (thymidine-nocodazole block and mitotic shake-off) of the cell division cycle. Data from the third double thymidine block study was used for the experiments here. After release from the double thymidine block, when the cells entered S phase, gene expressions were monitored every hour up to 46 h during which the cells went through three successive cell division cycles. The cell divisions occurred after approximately 13 h, 27 h and 44 h relative to the 0 time point (release from the block). This yielded 14, 14 and 17 measurements for the first, second, and third cycles, respectively. cDNA arrays probing nearly 30,000 human genes were used to measure gene expression levels.

In the human cell cycle data the time points corresponding to the S phase are clearly delineated from that of the other phases but it is not the case with the other remaining phases. There seems to be an overlap between the time points corresponding to the M, G1 and G2 phases. So we split the time points over one cell cycle into two phases: S and Non-S, where the Non-S phase is the superset of the remaining phases. This leads to the correspondence between samples of the three cell cycles and the two phases shown in Table 2.

We use a subset of the 1134 genes identified as cell cycle regulated for our experiments. This is the set of 1099 genes for which the publicly available data table has non-empty rows (http://genome-www.stanford.edu/Human-CellCycle/HeLa/). The total number of GO tags for the included genes is 567 and 222 in the BP and MF hierarchies respectively. Models were trained using measurements from one full cell cycle out of the three available cycles and used to predict stage labels on the samples of the other two cycles. This led to a total of three training/test pairs with two cycles in the test set for each pair.

5. Results

Using the protocol for each dataset described above we computed the mean of prediction accuracies over 10 different EM runs for each of the models. Each of the EM runs were initialized with different randomly chosen cluster responsibilities for genes and run for 40 iterations, which we observed to be sufficient for convergence to the nearest optimum. We then averaged the resulting means over held-out sets to compute the overall mean prediction accuracy. The standard error of the overall mean was computed by dividing the standard deviation of the means over EM runs by the square root of the number of such means.

5.1. Optimum number of clusters

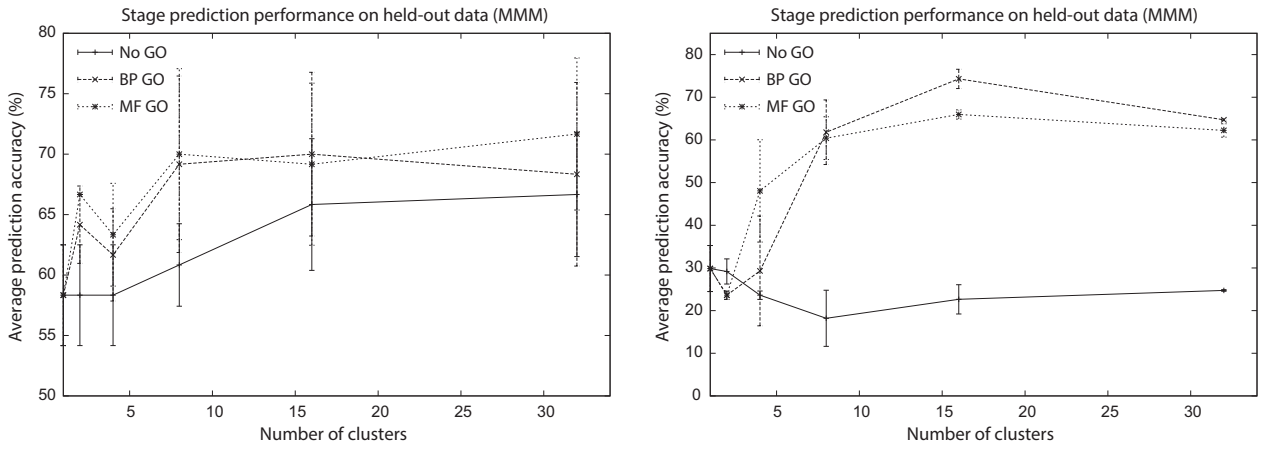
We keep the number of groupings almost the same in the experiments to make a fair comparison across the models with and without using GO tags. In order to choose a good number of clusters, we study the held-out prediction behavior as a function of the number of clusters in the model. We perform this study for the simplest multimodal mixture model (MMM) by varying the number of clusters from 1 to 32 exponentially, i.e. doubling each time. The stage prediction results on the three time course datasets for held-out samples are plotted in Fig. 3. Based on the plots, the optimum number of clusters is 16. However, we choose the number of clusters to be 15, a number divisible by 3, which supported model topologies that we have omitted for the sake of brevity. Alternatively, it is possible to use other model selection criteria such as AIC [27] or BIC [28] for choosing the number of clusters in our models.

5.2. Estimation of biological stages

As described in Section 3.3, it is possible to estimate stage boundaries in a time course dataset given that the number of stages is known. Assuming a moderately small total number of time points in the dataset, a brute force search for all possible groupings of time points into stages is performed to determine the grouping that leads to the best training likelihood. We first performed this search for the datasets where the ground truth was known. Specifically, we tried all possible groupings for the Cho and Whitfield datasets where the total number of time points within different cycles was 8, 9 and 14, 14, 17, respectively. Our best estimate of the stage boundaries was off by at most one time point of the ground truth boundaries in the five experiments in these two datasets.

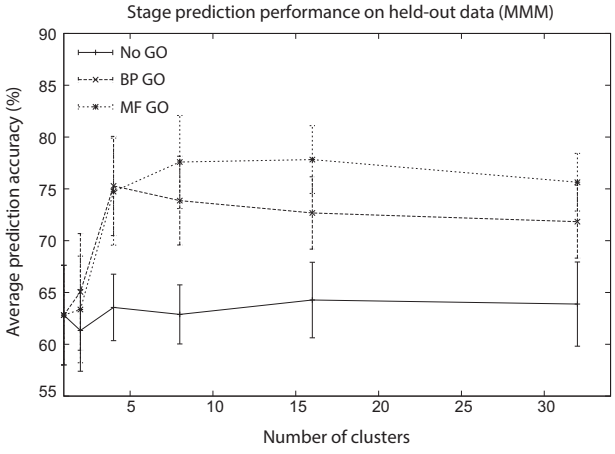
The angiogenesis data was recorded to study the process of blood vessel growth in a microvascular construct. Blood vessel growth can be hypothesized to involve roughly two stages (angiogenesis and maturation) as described in Section 4.1. In our data, we do not know beforehand as to where the switch from one stage to the other occurs within the recorded time series. Thus we estimated the switch points based on training data likelihood as well as stage prediction accuracy on test data using the SS-MMM model. Following the brute force search approach, similar to the other datasets, we computed the log likelihood of the training data for each possible stage boundary estimate (day 3, 7, 14, 21 and 28) and averaged them across the EM initializations and the two experimental runs (Section 4.1). The results in Table 3 indicate that day 7 is the best stage switch estimate among all the possibilities.

Further we used each of the above models for stage prediction on the dataset (reciprocally) not used for training. Here we are validating the de novo boundary using the assumption that the two data sets have the same stages in lieu of ground truth. Specifically, the “ground truth” switch point used for testing was the same one used for pooling in the training data. We plot the average prediction accuracies as a function of the switch point for both training and held out data



(a) Hoying

(b) Cho



(c) Whitfield

Fig. 3. Averaged stage prediction accuracies over held-out samples of the various datasets as a function of the number of clusters used in the MMM model. A prediction strategy based on random guess would have a stage prediction accuracy of 50%, 25%, and 50% for the Hoying, Cho, and Whitfield datasets, respectively.

Table 3
Estimation of the best switch point for Hoying data set using training data likelihood. The table lists the average log likelihood (LL) of training samples offset from the maximum for all possible switch points, equivalent to log of the likelihood ratio w.r.t. the maximum likelihood, with (BP and MF) and without (No GO) using GO tags. The best estimate in all cases is day 7.

Switch point	LL (No GO)	LL (BP)	LL (MF)
3	−182	−181	−173
7	0	0	0
14	−17.7	−11.6	−8.38
21	−25.1	−13.6	−7.63
28	−137	−133	−130

in Fig. 4. The switch point (day 7) leading to the most accurate prediction on held out samples is a second good estimate of the true switch, which agrees with the first estimate. This second process is similar to cross validation and provides an estimate of the most likely switch point in the absence of such information a priori, we use the day 7 switch point for all experiments on the angiogenesis data. For the other datasets we use the stage switch points that came with the datasets as provided in Sections 4.2 and 4.3.

5.3. Model parameters

The parameter counts for each of our models are listed in Table 4.

Table 4
Parameter settings of the various models. Each cluster is modeled by a Gaussian distribution in each of the models. The notation *#a/#b* indicates that there are as many *a*'s as given by the column entries for every *b* of the particular model.

	MMM	Pooled-MMM		SS-MMM		TS-MMM
	#Clusters	*#Pools*	*#Clusters*	*#Pools*	$\frac{\#Clusters}{\#Pools}$	$\frac{\#Clusters}{\#Time-points}$
Hoying data	15	2	15	2	15	15
Cho data	15	4	15	4	15	15
Whitfield data	15	2	15	2	15	15

The above choices lead to grouping of the genes' measurements (expression levels and GO tags) into almost the same number (15) of groups. This is regardless of whether grouping is done considering expression profiles across the entire time course (MMM and Pooled-MMM), individual time points (TS-MMM) or pooling expression levels across a subset of time points (SS-MMM and Pooled-MMM). Keeping the number of base clusters as similar as possible in all the groupings enables a fair comparison across models that are quite different in how they group the data.

5.4. Cluster GO term posteriors

All our models learn a posterior over GO terms for each of the clusters. These posteriors could be useful for further interpretation

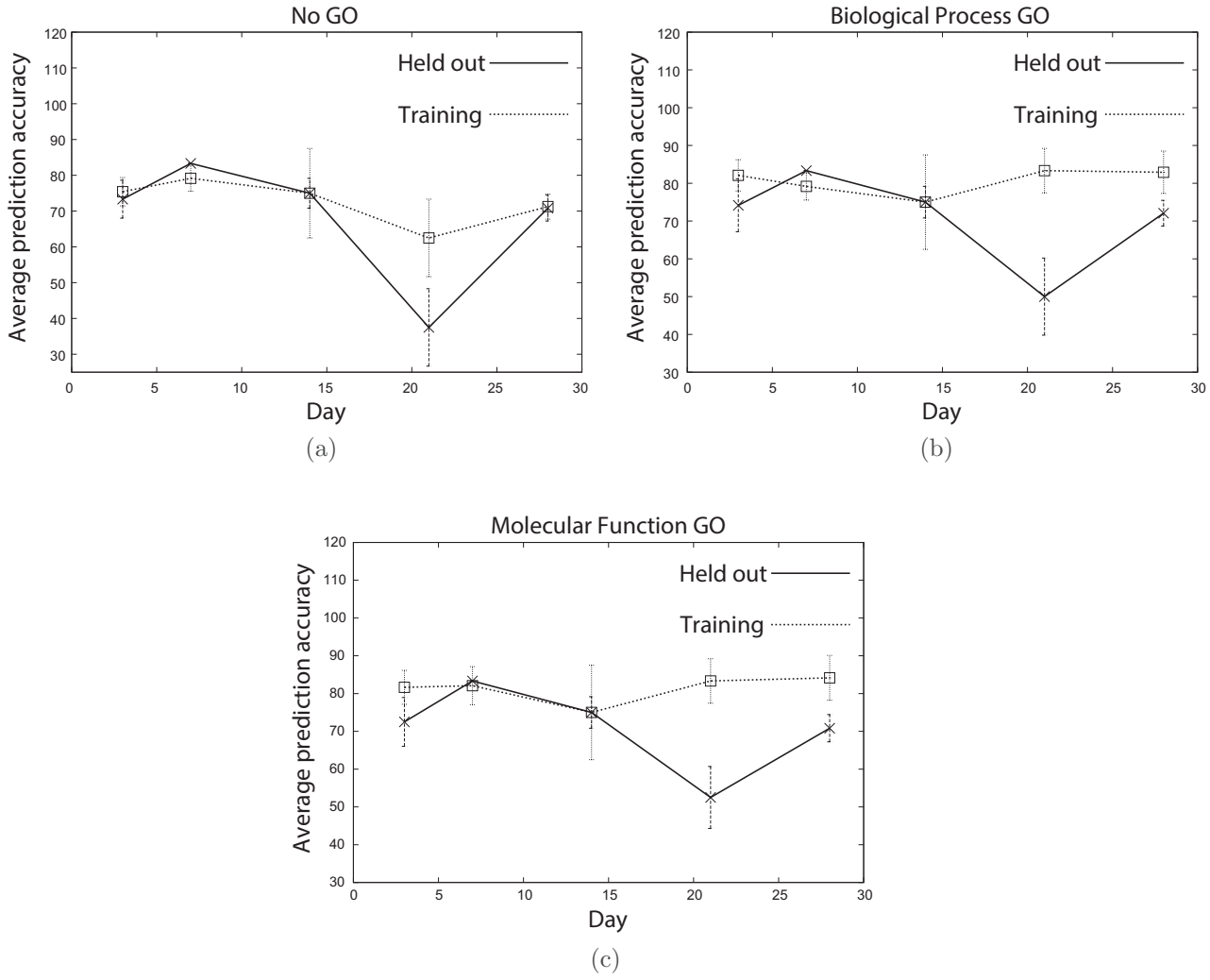


Fig. 4. Estimation of the switch point between the angiogenesis and maturation stages for the Hoying dataset. The plots show the stage prediction accuracy as a function of different assumed switch points (days 3, 7, 14, 21 and 28). The three plots correspond to models without using GO tags (top-left) and using tags from the Biological Process (top-right) or Molecular Function (bottom) GO hierarchies. Based on average stage prediction accuracy over held out data, day 7 is the optimal estimate of the time of stage switch.

of the clusters. For example, the most probable GO terms within a cluster could be indicative of the predominant Biological Process or Molecular Function that the genes within that cluster are involved. Since our main focus in this paper is to evaluate the performance of multimodal clustering on phenotype prediction we have not extensively examined the GO term posteriors for biological interpretation. Fig. 5 shows the first 20 most likely GO terms of the Cho yeast cell cycle data from the Biological Process ontology in a few clusters of the MMM model. The GO terms were post-processed using the REViGO tool [29] to eliminate any redundancies in the terms.

5.5. Comparison between models

The stage prediction results on held-out samples of the three time-course datasets using each of the models are shown in Fig. 6. The results suggest that the proposed models learn about the underlying distribution, as stage prediction performance on held out data is almost always above chance (excepting the MMM and SS-MMM models on the yeast cell cycle data, without using GO tags). Further, the models that use GO tags almost always perform better or at least equally well within error relative to their corresponding models that do not make use of GO tags. This improvement in prediction performance is most evident for the yeast cell cycle data. Regardless of the model, the average prediction accuracy is close to chance prediction (25%) when GO tags are not used. Using GO tags from the BP tree, this improves to around 70.3%, 38.54%, 68.40%, and 76.67% for the MMM, TS-MMM, SS-MMM and Pooled-MMM models respectively. A similar improvement to around 65.4%, 45.49%, 84.44%, and 73.54% is observed using the MF GO tags. In the case of human cell cycle data, the most significant improvement from about 49.1% (No GO) to about 76.51% (BP) and 73.57% (MF) is observed using the Pooled-MMM model. Using GO tags in the other models results in a similar but a slightly less significant improvement. The prediction performance on the angiogenesis data using models that use GO tags is almost as good or slightly better as the ones not using them. Note that this data has only very few samples (6) and so an anomalous result is quite likely. The prediction results offer clear evidence that GO tags provide independent and useful information that can augment measured expression levels of genes.

There seems to be a merit to using models that pool expression levels within a stage, on the task of learning about biological stages. On almost all the datasets, the pooled sample models seem to help the task of stage prediction. This is because appropriately constrained models generalize better than their unconstrained versions. With the knowledge of biological stages, the extra constraints of sharing Gaussian parameters across time samples of a stage lead to simpler models in the case of pooled sample models (SS-MMM and Pooled-MMM). On the other hand their corresponding

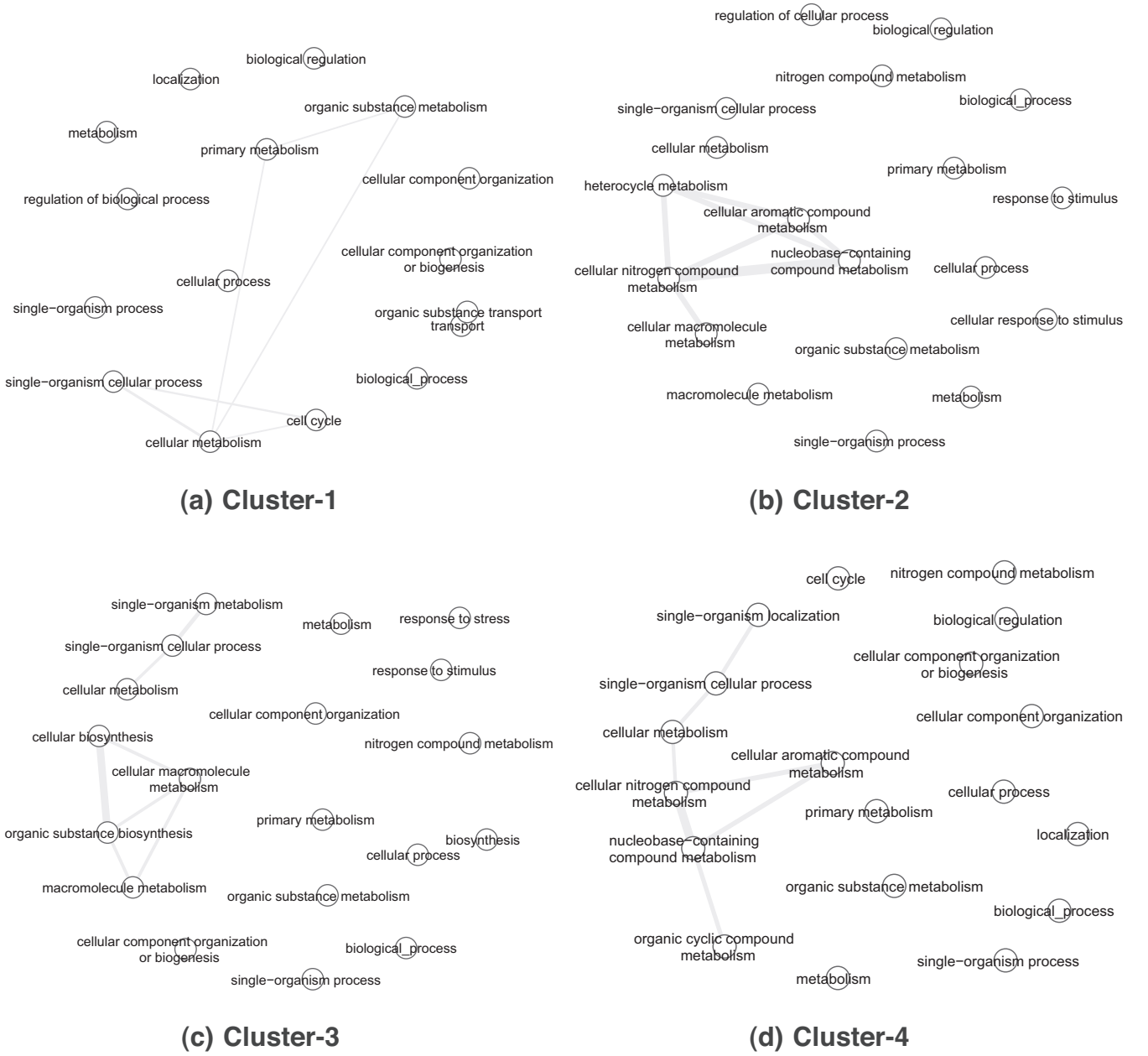


Fig. 5. Most likely GO terms in a few clusters of the MMM model on the Cho yeast cell cycle data. The 20 most likely terms were chosen based on their posterior likelihoods in the clusters. The terms were post-processed to eliminate any redundancies in the GO terms using the REVIGO [29] tool. The same tool was used to plot the visualization of the terms and their relationships.

special cases (TS-MMM and MMM) with fewer constraints tend to overfit leading to relatively poor generalization. This is similar to regularization in model fitting or introducing appropriate priors in Bayesian inference. Note that some of the other constraints, e.g., all the genes within a group making phase transitions together, might agree more with some datasets than others. Models without such constraints are easy to construct but they have a tendency to fit noise more than the underlying biological phenomena.

5.6. Comparison to a discriminative model: multiclass SVM

We experimented with a discriminative model for stage prediction on our datasets, specifically a Multiclass Support Vector Machine (SVM) with a linear kernel. We used the standard formulation of multi-class classification with a different weight vector for each class within the structured output SVM (SVM-struct) framework [30]. The prediction results are plotted in Fig. 6e. Note that GO tags are not useful in discriminative modeling as they remain the same across all the gene expression samples and do not help differentiate between samples. Therefore we did not use GO tags for these experiments. The experimental protocol is the same as for the previous experiments for each dataset. The results are competitive with some of the generative models proposed in this paper that use GO tags. This suggests that it is possible to obtain a high prediction accuracy without using GO tags using a discriminative model such as a multiclass SVM. However, generative models can be used to make a number of inferences other than stage prediction, as exemplified by the determination of stage boundaries developed here. These models can also be used to simulate new data points by sampling. Discriminative methods [31–33] lack this ability and learn a strong model with fewer parameters only to discriminate between classes.

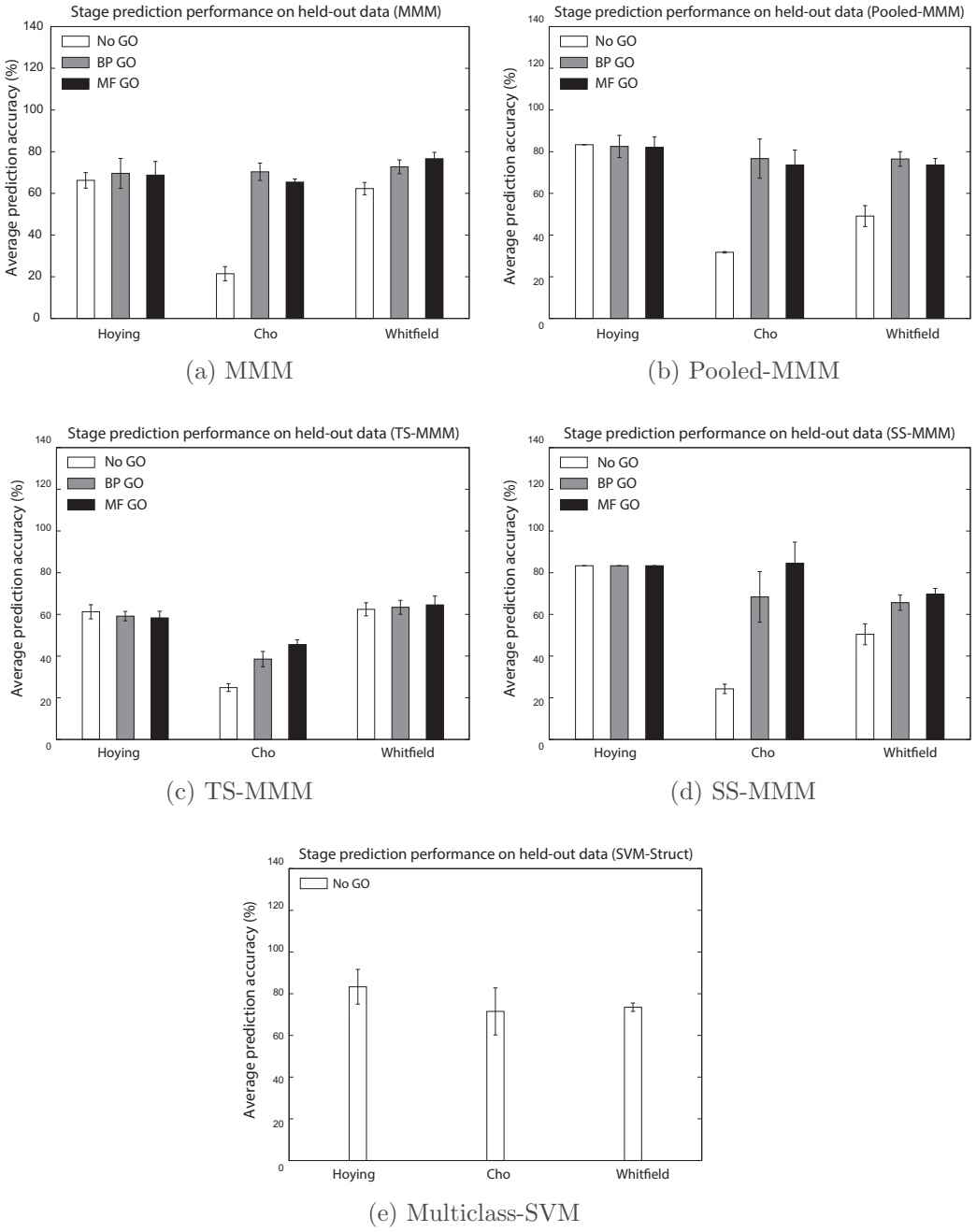


Fig. 6. (a)–(d) Averaged stage prediction accuracies over held out samples of the various datasets using the different models. The pooled sample models (SS-MMM and Pooled-MMM) use a Bayesian stage prediction scheme (see Section 3.1), whereas the TS-MMM and MMM models use a time-based stage prediction scheme (see Section 3.2). A prediction strategy based on random guess would have a prediction accuracy of 50%, 25%, and 50% on the Hoying, Cho, and Whitfield datasets, respectively. (e) Average prediction accuracy on the held out samples of the Hoying, Cho and Whitfield datasets using a multiclass SVM. GO tags are not used for these experiments as they remain the same across all the samples.

6. Discussion

In this work, we introduce a multimodal probabilistic generative framework for modeling a number of microarray datasets. Multimodal models enable us to incorporate an independent source of information in the form of Gene Ontology (GO) tags for analyzing microarray gene expression data. These are terms from an evolving controlled vocabulary to describe genes and gene products. The most common approach to use these terms is to examine gene clusters or candidate gene groups by doing an enrichment analysis. We deviate from this approach and use these terms in clustering genes in a continuous-discrete space of gene expression levels and GO terms. Although we use Gaussians in this work, different sources of noise can be easily modeled with appropriate parametric probability distributions in the generative framework. A number of inferences including phenotype or biological stage inference can be done employing a Bayesian methodology. The proposed models do not assume anything about how the gene expression data were measured. Expression data measured using other technologies is easily handled within our modeling framework.

We propose four generative multimodal models for time course gene expression datasets to cluster genes differently based on their behavior over the entire time course, biological stages or individual time points. We use these models to infer biological stage boundaries assuming the number of stages, and stage prediction on novel data discussed in detail below. To determine stage boundaries, we choose

the ones leading to the largest likelihood. To compute the likelihood we choose to use the SS-MMM model. Using this paradigm, we are able to estimate the true stage boundaries to within a single time point relative to ground truth in Cho and Whitfield datasets. We use the same procedure to estimate the stage boundary in the Hoying dataset where no such prior information is available. We also validate that the boundary we find is also the best boundary for stage prediction on the assumption that the two runs of the experiment have the same stages.

Our results based on biological stage prediction for various datasets suggest that GO tags provide useful information to obtain better gene clusters for the task of prediction. Significant improvements in prediction accuracies using GO tags are observed on the yeast and human cell cycle data, whose genes and annotations have been long studied. The improvements are less significant with the Hoying angiogenesis data possibly due to various factors such as the selection of genes, their annotations or the dataset size. Models that pool samples across time points (Pooled-MMM and SS-MMM) are regularized versions of corresponding models (MMM and TS-MMM) that do not assume pooling and lead to better generalization in terms of prediction accuracy.

The held-out prediction behavior on all the datasets improves with increasing number of clusters, up to about 16 clusters, when GO tags are used. This is not necessarily true with models not using GO tags. This implies that when more patterns (clusters) of gene behavior are allowed, GO tags can help delineate these patterns in a way useful for phenotype prediction. From Fig. 3, it can be seen that by increasing the number of clusters from 1 to 16 the average held-out prediction accuracy improves from around 30% to 70% when GO tags are used with the yeast cell cycle data. A similar improvement from around 63% to around 75% is observed with the human cell cycle data. Note that with models not using GO tags in the two cases the average prediction accuracy remains almost the same over the range of number of clusters. It ranges between 15% and 30% in the case of yeast cell cycle data and between 60% and 65% in the case of human cell cycle data.

The above improvement due to using GO tags is more significant in the two cell cycle datasets than the angiogenesis dataset. This is perhaps due to two possible reasons. One is that the number of samples in the Hoying angiogenesis data is small (6 per run, see Section 4) leading to estimates of standard errors that are relatively large. The second is the quality of GO tags for the genes chosen for the experiment. For the two cell cycle datasets, the chosen genes have been very well studied by other research groups and are known to be involved in the process of cell cycle regulation. Yeast and human genes have been thoroughly studied and annotated. The genes used for the Hoying data are chosen based on differential expression. More work is needed to ascertain that the selected genes are all involved in the blood vessel growth process and have reliable GO annotations.

7. Software

A Matlab package implementing the proposed methods will be made available online at the URL: http://www.ivilab.org/software.html.

Acknowledgments

This material is based upon work supported in part by the Arizona Biomedical Research Commission (ABRC).

Appendix A. Expectation maximization (EM)

We describe the EM algorithm update equations for training each of our models described in this work. In our implementation, the EM algorithm is initialized with random assignments to cluster responsibilities (expected values of hidden variables) of the E-step. Most of the equations follow the notation in Section 2 and other equations in the paper, along with a few new terms introduced below as needed.

A1. MMM

The parameters of the MMM model include the cluster priors $P(c)$, means ($\boldsymbol{\mu}_c(t)$), variances ($\boldsymbol{\sigma}_c^2(t)$) of the clusters c at each of the time points t in the training data, and the cluster conditional Gene Ontology term distributions $P(o|c)$ for each term o. Let $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c^2$ represent the corresponding mean and variance vectors obtained by concatenating these parameters across the time points t. The E-step and the M-step update equations are given as follows.

A1.1. E-step

$$\gamma_c^g = P(c|\boldsymbol{e}^g, O^g) \tag{A.1}$$

$$\propto P(c)\, p(\boldsymbol{e}^g, O^g|c) \tag{A.2}$$

where γ_c^g is the expected value of the hidden variable (cluster responsibility) corresponding to cluster c for gene g. The form of $p(\boldsymbol{e}^g, O^g|c)$ is given by Eq. (1).

A1.2. M-step

$$P(c) = \frac{\sum_g \gamma_c^g}{\sum_c \sum_g \gamma_c^g} \tag{A.3}$$

$$\boldsymbol{\mu}_c = \frac{\sum_g \gamma_c^g \boldsymbol{e}^g}{\sum_g \gamma_c^g} \tag{A.4}$$

$$\boldsymbol{\sigma}_c^2 = \frac{\sum_g \gamma_c^g (\boldsymbol{e}^g - \boldsymbol{\mu}_c)^2}{\sum_g \gamma_c^g} \tag{A.5}$$

$$P(o|c) = \frac{\sum_g \delta^g(o)\gamma_c^g}{\sum_o \sum_g \delta^g(o)\gamma_c^g} \tag{A.6}$$

where $\delta^g(o)$ is an indicator function which is 1 if gene g is annotated with the term o, otherwise 0.

A2. Pooled-MMM

The EM update equations for this model are almost the same as the MMM model, except that shared parameter values for all pooled time points within a stage are used in computing the likelihood in the E-step (Eq. (2)). Similarly, expression values for all pooled time points within a stage are averaged to obtain the corresponding shared parameter estimates. The M-step updates for only the shared parameters: stage s specific mean ($\boldsymbol{\mu}_c(s)$) and variances ($\boldsymbol{\sigma}_c(s)$) are given below.

A2.1. M-step

$$\boldsymbol{\mu}_c(s) = \frac{\sum_g \gamma_c^g \left(\frac{1}{|s|}\sum_{t\in s} \boldsymbol{e}_t^g\right)}{\sum_g \gamma_c^g} \tag{A.7}$$

$$\boldsymbol{\sigma}_c^2(s) = \frac{\sum_g \gamma_c^g \left(\frac{1}{|s|}\sum_{t\in s} \left(\boldsymbol{e}_t^g - \boldsymbol{\mu}_c(s)\right)^2\right)}{\sum_g \gamma_c^g} \tag{A.8}$$

where $|s|$ is the number of samples within stage s .

A3. SS-MMM

The parameters of the SS-MMM model include the univariate stage and cluster specific means ($\boldsymbol{\mu}_{c_s}$), variances ($\boldsymbol{\sigma}^2_{c_s}$), priors ($P(c_s)$), and ontology term conditional distributions ($P(o|c_s)$) for every cluster c_s of a stage s.

A3.1. E-step

Each gene admits two responsibility variables, one for sampling the expression from a stage specific cluster $\gamma^g_{c_s}(\boldsymbol{e}^g_s)$ and the other for sampling a GO tag o from a stage specific cluster $\gamma^g_{c_s}(o)$

$$\gamma^g_{c_s}(\boldsymbol{e}^g_s) = P(c_s|\boldsymbol{e}^g_s) \tag{A.9}$$

$$\propto P(c_s)p(\boldsymbol{e}^g_s|c_s) \tag{A.10}$$

$$\gamma^g_{c_s}(o) = P(c_s|o, \boldsymbol{e}^g_s) \tag{A.11}$$

$$\propto P(c_s|\boldsymbol{e}^g_s)P(o|c_s) \tag{A.12}$$

The likelihoods in the above equations are computed from the current parameter estimates in the EM iterations.

A3.2. M-step

$$P(c_s) = \frac{\sum_g \left(\gamma^g_{c_s}(\boldsymbol{e}^g_s) + \sum_o \frac{\delta^g(o)\gamma^g_{c_s}(o)}{M}\right)}{\sum_{c_s}\sum_g \left(\gamma^g_{c_s}(\boldsymbol{e}^g_s) + \frac{\delta^g(o)\gamma^g_{c_s}(o)}{M}\right)} \tag{A.13}$$

$$\boldsymbol{\mu}_{c_s} = \frac{\sum_g \gamma^g_{c_s}(\boldsymbol{e}^g_s)\left(\frac{1}{|s|}\sum_{t\in s}\boldsymbol{e}^g_t\right)}{\sum_g \gamma^g_{c_s}(\boldsymbol{e}^g_s)} \tag{A.14}$$

$$\boldsymbol{\sigma}^2_{c_s} = \frac{\sum_g \gamma^g_{c_s}(\boldsymbol{e}^g_s)\left(\frac{1}{|s|}\sum_{t\in s}(\boldsymbol{e}^g_t - \boldsymbol{\mu}_{c_s})^2\right)}{\sum_g \gamma^g_{c_s}(\boldsymbol{e}^g_s)} \tag{A.15}$$

$$P(o|c_s) = \frac{\sum_g \delta^g(o)\frac{\gamma^g_{c_s}(o)}{M}}{\sum_o\sum_g \delta^g(o)\frac{\gamma^g_{c_s}(o)}{M}} \tag{A.16}$$

In the above equations, M is the total number of stages, $|s|$ is the total number of samples in a stage s, and $\delta^g(o)$ is the indicator function for denoting whether the gene g is annotated with the term o.

A3.3. TS-MMM

The EM update equations are the same as for the SS-MMM model except that stage s is replaced by time point t and number of stages M is replaced by the total number of time points N.

References

[1] M. Ashburner, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[2] S.Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Use and misuse of the gene ontology annotations, Nat. Rev. Genet. 9 (2008) 509–515.
[3] P. Khatri, S. Drăghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics 21 (2005) 3587–3595.
[4] S. Grossman, S. Bauer, P.N. Robinson, M. Vingron, Improved detection of over-representation of gene-ontology annotations with parent child analysis, Bioinformatics 23 (2007) 3024–3031.
[5] A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating go graph structure, Bioinformatics 22 (2006) 1600–1607.
[6] Y. Lu, R. Rosenfeld, I. Simon, G.J. Nau, Z. Bar-Joseph, A probabilistic generative model for go enrichment analysis, Nucleic Acids Res. 36 (17) (2008) e109.
[7] D. Wu, G.K. Smyth, Camera: a competitive gene set test accounting for inter-gene correlation, Nucleic Acids Res. 40 (17) (2012) e133.
[8] D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J.E. Visvader, G.K. Smyth, Roast: rotation gene set tests for complex microarray experiments, Bioinformatics 26 (2010) 2176–2182.
[9] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M. Jordan, Matching words and pictures, J. Mach. Learn. Res. 3 (2003) 1107–1135.
[10] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, in: Proceedings of International Conference on Computer Vision, vol. II, pp. II:408–415.
[11] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of pictures, in: Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems, vol. 16, pp. 553–560.
[12] P. Carbonetto, N. de Freitas, Why can't José read?: the problem of learning semantic associations in a robot environment, in: Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-linguistic Data, pp. 54–61.
[13] M. Masseroli, D. Chicco, P. Pinoli, Probabilistic latent semantic analysis for prediction of gene ontology annotations, in: Proceedings of the IEEE International Joint Conference on Neural Networks, IEEE, 2012, pp. 1–8.
[14] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.
[15] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 2 (1998) 65–73.
[16] M.L. Whitfield, A.J. Sherlock, G. Saldanha, C.A. Murray, J.I. ane Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M Hurt, P.O. Brown, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, Mol. Biol. Cell 13 (2002) 1977–2000.
[17] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C. Lee, J.M. Trent, L.M. Staudt, J.H. James, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P.O. Brown, The transcriptional program in the response of human fibroblasts to serum, Science 283 (1999) 83–87.
[18] Y. Cheng, G.M. Church, Biclustering of expression data, in: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 2000, pp. 93–103.
[19] P. Kirk, J.E. Griffin, R.S. Savage, Z. Ghahramani, D.L. Wild, Bayesian correlated clustering to integrate multiple datasets, Bioinformatics 28 (2012) 3290–3297.
[20] R.S. Savage, Z. Ghahramani, J.E. Griffin, B.J.D. la Cruz, D.L. Wild, Discovering transcriptional modules by Bayesian data integration, Bioinformatics 26 (2010) 158–167.
[21] S. Rogers, M. Girolami, W. Kolch, K.M. Waters, T. Liu, B. Thrall, H.S. Wiley, Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models, Bioinformatics 24 (2008) 2894–2900.
[22] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. 39 (1977) 1–38.
[23] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[24] J. Bilmes, A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical Report, ICSI (U. C. Berkeley), 1998.
[25] S.S. Nunes, K.A. Greer, C.M. Stiening, H. Chen, K.R. Kidd, M.A. Schwartz, C.J. Sullivan, H. Rekapally, J.B. Hoying, Implanted microvessels progress through distinct neovascularization phenotypes, Microvasc. Res. 79 (2010) 10–20.
[26] K.A. Greer, M.R. McReynolds, H.L. Brooks, J.B. Hoying, Carma: a platform for analyzing microarray datasets that incorporate replicate measures, BMC Bioinform. 7 (2006) 149.
[27] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (1974) 716–723.
[28] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[29] F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, Revigo summarizes and visualizes long lists of gene ontology terms, PLoS One 6 (2011) e21800.
[30] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, J. Mach. Learn. Res. 6 (2005) 1453–1484.
[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2009.
[32] D. Amaratunga, J. Cabrera, Y.-S. Lee, Enriched random forests, Bioinformatics 24 (2008) 2010–2014.
[33] D. Amaratunga, J. Cabrera, Y. Cherckas, Y.-S. Lee, Ensemble Classifiers, in: Volume 8 of Collections, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 235–246.