ACTIVE PATHWAY IDENTIFICATION AND CLASSIFICATION WITH PROBABILISTIC ENSEMBLES

TIMOTHY HANCOCK HIROSHI MAMITSUKA
timhancock@kuicr.kyoto-u.ac.jp mami@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan

A popular means of modeling metabolic networks is through identifying frequently observed pathways. However the definition of what constitutes an observation of a pathway and how to evaluate the importance of identified pathways remains unclear. In this paper we investigate different methods for defining an observed pathway and evaluate their performance with pathway classification models. We use three methods for defining an observed pathway; a path in gene over-expression, a path in probable gene over-expression and a path of most accurate classification. The performance of each definition is evaluated with three classification models; a probabilistic pathway classifier - HME3M, logistic regression and SVM. The results show that defining pathways using the probability of gene over-expression creates stable and accurate classifiers. Conversely we also show defining pathways of most accurate classification finds a severely biased pathways that are unrepresentative of underlying microarray data structure.

Keywords: Classification; Markov model; Mixture of Experts; Metabolic Pathway.

1. Introduction

A metabolic pathway is a small section of the overall metabolic network that comprises a connected series of chemical reactions which are known to perform a specific function (Figure 1). On the left hand side of Figure 1 we present a simplified diagram of a metabolic pathway of similar structure to those found within on line metabolic network databases such as KEGG [4]. The metabolic compounds of the network are $[C_1, \ldots, C_4]$ are connected by reactions $[R00001, R00002, R00003, R00004]$ which in turn are catalyzed by a set of genes $[G_1, \ldots G_8]$. To define a pathway through the network we have flagged C_1 as the start compound and C_4 as the end. Observations of the genes within Figure 1 come most commonly from microarray experiments which measure the expression levels of the genes within the network. Therefore it is convenient to transform the reaction-compound network shown on the left hand side of Figure 1 into the gene-compound network shown on the right. A pathway through the gene-compound network from C_1 to C_4 can now be defined as a sequence of genes denoting the edges between compounds.

Determining what level of gene expression indicates an active network edge is not

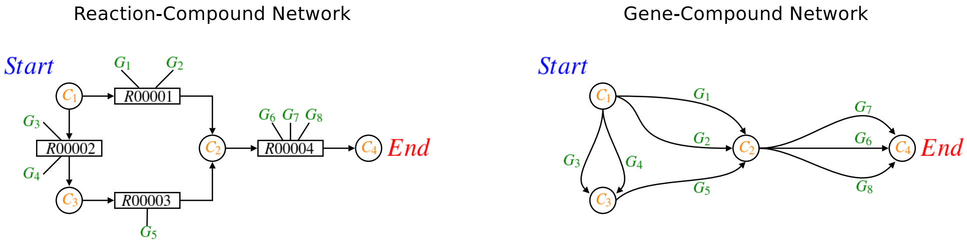

Fig. 1: Diagrammatic representation of a metabolic network.

simple due to the high level of noise inherent within microarrays. By far the most common approach for specifying active genes rely on summary statistics such as fold change and z-scores. However using simple summary statistics to denote an active edge reduces number of observations on the activity of the network down to a single value for each response label. Reducing all microarray observations to a single value could potentially remove important subtle changes in expression structure that can affect which path is taken. Another common approach is to employ a supervised technique to pick genes from a metabolic network that have expression structure useful for classifying a response. However the process of selecting important genes is likely to break network relationships between the genes leading to a disconnected solution which makes any biological pathway interpretation difficult.

In this paper we consider the problem of what is an appropriate definition for an active edge. To ensure that the full structure of the underlying microarray is represented we require a solution that can extract observed pathways from each microarray experiment. We consider three definitions of an active edge:

(1) A high value of scaled gene expression data.
(2) A high probability found within an empirical cumulative distribution function computed for all genes within the network.
(3) A low entropy of the classified label of microarray for each experiment computed by individual gene classifiers.

We perform a comparative analysis across the three gene activity criteria using standard classification models; support vector machines and logistic regression as well as an explicit classifier of metabolic pathways, HME3M [2]. The success of each gene activity criteria is evaluated both in terms of any observed bias in the number pathways extracted and in terms of classification accuracy and stability. In the next section we describe the pathway extraction procedure and each edge activity criteria. We then describe our pathway classifiers and compare the performance of each method on real biological data.

2. Pathway Extraction

In our experiments we extract all valid pathways from each microarray experiment that are observed from the start to the end compounds of a specified network. To do this we treat each microarray experiment, x_i as a single observation of the activity of all genes within a network. For each x_i we also have a response label y_i denoting the experimental conditions. Then using a pre-specified criteria to determine which genes are active within x_i we extract all possible paths from the start node to the end node and label each path with y_i. For example in the example network in Figure 1 if we determine that for a given experiment, x_i, the genes $[G_1, G_4, G_5, G_6]$ are active then 2 possible paths can be extracted and represented as binary strings $[G_1, G_6] = [1, 0, 0, 0, 0, 1, 0, 0]$ and $[G_4, G_5, G_6] = [0, 0, 0, 1, 1, 1, 0, 0]$. Both of these paths are then given the response label y_i. If all eight genes within Figure 1 are found to be active then a total of 12 possible pathways can be extracted. From each microarray experiment we extract all possible paths and their response labels and augment them into our final pathway dataset.

Extracting all possible paths observed over all microarray experiments presents two major issues for an further analysis. Firstly how to determine the criteria which denotes active genes within each microarray experiment, and secondly how to handle the inevitable duplicate pathways that will be extracted. In our experiments we focus on the first issue of assessing gene activity, but also address what effect duplicate paths by first including them within the analysis and then monitoring the performance change observed after their removal. We now describe the three gene activity criteria under consideration.

3. Gene Activity Criteria

3.1. *Scaled Expression*

The z-scaling or normalization of gene expression data is a standard preprocessing step performed before most microarray analysis and is simply,

$$z_i = \frac{x_{ij} - \bar{x}_j}{s_{x_j}} \tag{1}$$

where $\bar{x}_j$ is the mean of all expressions for gene x_j and s_{x_j} is the standard deviation. The effect of normalization is that each gene will have a mean of $\bar{x_j} = 0$ and a standard deviation of $s_{x_j} = 1$. Normalization therefore brings all gene expressions down to a relative scale allowing for easy comparison. However normalization using the z transformation does assume that the distribution of each gene is independent and normally distributed.

3.2. *Empirical CDF*

To overcome the assumptions of a z-score non-parametric procedures such as ranking the gene expressions are often employed. In this paper we rank gene expressions based on their position within the empirical Cumulative Distribution Function (CDF) of all genes. Although an empirical CDF is equivalent to a standard ranking procedure there are two major advantages for its application; firstly that the empirical CDF provides a function capable of looking up new gene expression observations without altering the ranking, and secondly by using an empirical CDF the gene expressions are transformed to probabilities where a probability of 0.5 provides an intuitive of gene activity definition.

To compute the empirical CDF over all genes we first ignore all gene structure and extract a simple vector of individual expressions. Then we compute the probability of each individual expression value by,

$$P(x_{ij}) = \int_{-\infty}^{x_{ij}} P(t).dt = \frac{\text{Number}[x < x_i ij]}{\text{Number}[x]} \tag{2}$$

where Number$[x < x_i]$ indicates the number of individual expressions below x_{ij}. Clearly the final empirical CDF probability of each observation unlike the z-score does not assume any specific distribution.

3.3. *Entropy*

Both the z-score and empirical CDF criteria are unsupervised naive transformations on the gene expression. We also include a model dependent transformation to assess the ability of using a supervised technique to select active genes within a pathway. To do this we perform a logistic regression on each variable individually and extract the posterior probabilities of classification. From the posterior probabilities we compute the entropy of the classification for each observation,

$$\text{Entropy}(x_{ij}) = \sum_{y_l \in y} p(x_{ij}|y = y_l) \log p(x_{ij}|y = y_l) \tag{3}$$

where y_l is a label within the response y and $p(x_{ij}|y = y_l)$ are the posterior probabilities for each observation x_{ij} and each response label y_l. The entropy of the posterior probabilities is an indication of how certain each logistic regression is on its assigned label. The entropy of a random assignment where $p(x_{ij}|y = y_l) = 0.5 \;\; \forall \;\; y_l$ is approximately -0.6931 and as the certainty of the classification increases the entropy tends to 0. Therefore if all labels are classified well the entropy should be small, and a small entropy is an indication that the network structure is well defined for a particular observation. Entropy however does assume a particular model for each gene, and clearly the results will be dependent on the choice of model. In this paper we select logistic regression because of the smooth transition between classes that is provided by the assumed logistic curve.

3.4. *Comparison of Activity Criteria*

In Figure 2 we present a example of the effect of each of our gene activity criteria on 5 simulated genes connected in a single path from the gene 1 to gene 5 (left to

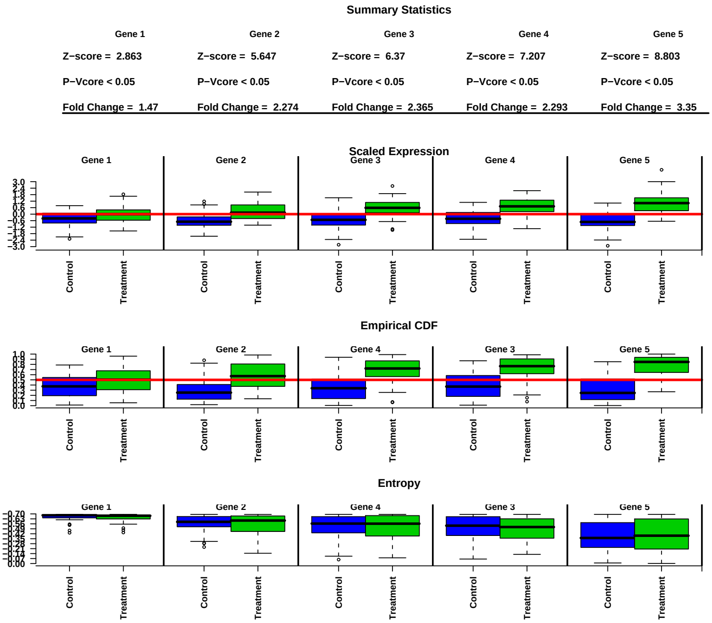

Fig. 2: Boxplots displaying the effect of different gene activity criteria.

right). From left to right we have simulated a subtle increase in gene expression observed in the treatment group. This is observed in the z-scores and corresponding P-values as well as expression fold changes computed for each gene presented at the top of Figure 2. It is clearly seen the z-scores for all genes find the treatment group to be significantly over-expressed at a P-value < 0.05 and therefore we observe the overall path to be active.

However for individual gene activity criteria marked differences in the structure of each transformed gene are observed. For the scaled expression values we observe that the range of expressions within each response label is rather small. The red line across all box plots is drawn at a z-score of 0 where all observations above this line intuitively correspond to active gene expression. However because small expression range we observe that a subtle change in this tolerance value will create a large difference in the number of observations deemed to be active. Conversely for the empirical CDF criteria we observe that the range of expression within each response label is broader but the structure of the overall expression profile is maintained. This implies that if we slightly move the gene activity tolerance a similar gene expression profile will be obtained and therefore we expect pathway extraction using empirical CDF to be more stable than the standard scaled expression approach.

The entropy criteria however presents the opposite interpretation, where an active gene is determined by a low entropy score. It is clear from Figure 2 that there is a large difference in the entropy profiles from the genes showing accurate performances compared to G_1 where inaccurate classifications are observed. The inaccurate classifications of G_1 are in contrast to the summary statistics which show the entire path to be active. This implies that a single inaccurate gene classifier may break the pathway even though the data indicates the pathway to be active. To overcome this problem the entropy tolerance may need to be decreased to unrealistic levels which will result in the number paths extracted to become large and cause serious interpretation problems and performance penalties.

4. Pathway Classification

To compare the performance of each gene activity criteria we analyze the pathways with benchmark classification models penalized logistic regression (PLR) [6] and Support Vector Machines (SVM) with linear, polynomial (degree = 3) and radial basis kernels [1]. Furthermore to analyze the specific pathway differences found by each criteria we also compare the model performances with a specialized pathway classification model HME3M [2].

4.1. *Hierarchical Mixture of Markov Experts (HME3M)*

HME3M is a Hierarchical Mixture of Experts (HME) [3] which uses a Markov mixture model 3M [5] to first cluster the pathways. Then supervision of this cluster analysis is performed by training experts (classification models) on the pathways found at each cluster. This process however is not performed as two discrete steps, but is iteratively optimized with an EM algorithm. The EM optimization passes information from the experts back into the pathway clustering algorithm and vice versa. This flow of information from the experts into the clustering algorithm allows for clusters to be found that will provide optimal classification performances at each expert. Therefore HME3M is an probabilistic ensemble of experts where each expert evaluates the classification accuracy of each pathway cluster.

Combing a HME with the 3M model produces the HME3M likelihood,

$$p(y|x) = \sum_{m=1}^{M} \pi_m p(y|x, \beta_m) p(c_1|\theta_{1m}) \prod_{t=2}^{T} p(c_t, x_t|c_{t-1}; \theta_{tm}) \ . \tag{4}$$

The parameters of (4) are estimated with the EM algorithm by defining the responsibilities variable h_{im} to be the probability that a sequence i belongs to component m, given x, θ_m, β_m and y and using the following E and M steps:
E-Step: Define the responsibilities h_{im}:

$$h_{im} = \frac{\pi_m p(m|x_i, \theta_m) p(y_i|x_i, \beta_m)}{\sum_{m=1}^{M} \pi_m p(m|x_i, \theta_m) p(y_i|x_i, \beta_m)} \tag{5}$$

M-Step: Estimate the Markov mixture and expert model parameters:

(1) Estimate the mixture parameters:

$$\pi_m = \frac{\sum_{i=1}^{N} h_{im}}{\sum_{m=1}^{M} \sum_{i=1}^{N} h_{im}} \quad \text{and} \quad \theta_{tm} = \frac{\sum_{i=1}^{N} \delta(x_{it} = 1) h_{im}}{\sum_{i=1}^{N} h_{im}} \tag{6}$$

(2) Estimate the expert parameters:

We model our experts with a Penalized Logistic Regression (PLR) [6],

$$l(\beta_m | h_{im}) = \underset{\beta_m}{\arg\max} \left\{ \sum_{i=1}^{N} h_{im} \left(y_i \beta_m^T x_i + log(1 + e^{\beta_m^T x_i}) \right) - \frac{\lambda}{2} |\beta_m|^2 \right\} \qquad (7)$$

where X are the paths, y is a binary response variable, h_{im} are HME3M responsibilities for path i in component m, and β_{tm} are the PLR coefficients.

5. Experiments

We perform our experiments on a subsection of the glycolysis pathway for *Arabidopsis thaliana* extracted from KEGG [4] using gene expression information from the AltGenExpress database [7]. The pathway shown in Figure 3 starts at Alpha-D-Glucose and finishes at Pyruvate, contains 41 edges, and for a single observation there are 103680 possible paths. We extract observations from the microarray for two classes; *"rosette leaf"* (n = 21) and *"flower"* (n = 15) and specify "flower" to be target class ($y = 1$) and "rosette leaf" to be the comparison class ($y = 0$).

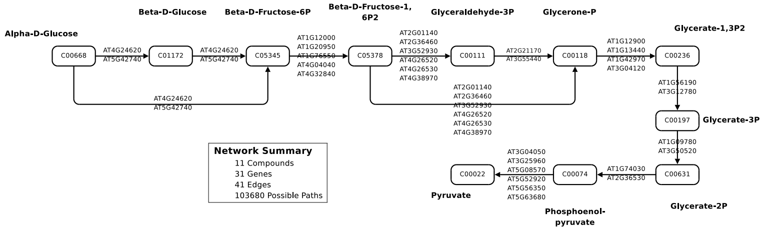

Fig. 3: *Arabidopsis thaliana* glycolysis pathway from Alpha-D-Glucose to Pyruvate.

We select this pathway as in our previous research [2] we used HME3M to show that there is a clear differential pathway expression that can be used to create a stable classifier of flower gylcolysis pathways. However in [2] we only employed a scaled expression tolerance, and did not consider different pathway definitions or duplicate pathways. In this research we focus on the effect of the pathway extraction procedure to see if the same pathways found in [2] can be more efficiently found using different pathway definitions and further assess the effect of removing duplicate pathways from the dataset.

6. Results: Pathway Extraction

We vary the tolerance for each gene activity criteria over a range such that the number of paths extracted from each class for each criteria would be as similar as possible. We vary the scaled expression tolerance between $[-0.14, 0.14]$, the

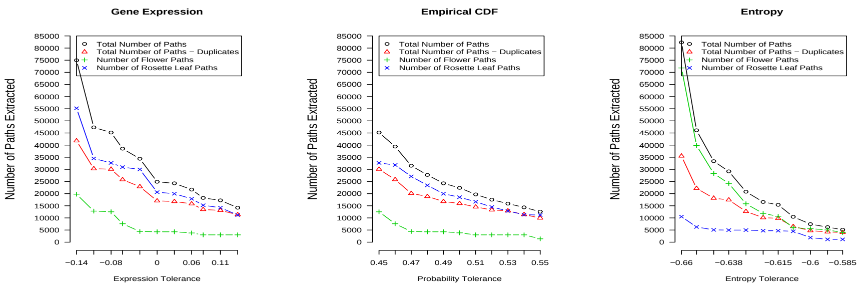

Fig. 4: Number of pathways extracted using different gene activity criteria.

empirical CDF tolerance between $[0.45, 0.55]$ and the entropy tolerance between $[0.54, 0.6]$. The results are presented in Figure 4. In Figure 4 the black line represents the total number of paths extracted and the red line represents the number of paths after duplicate pathways have been removed. The blue line represents the number of paths extracted from the rosette leaf experiments and the green line represents the number of number of paths extracted from the flower experiments.

The pathway extraction profiles agree with the results found in the simulation experiments (Figure 2). We see that the scaled expression criteria at low tolerance values extracts a large number of paths and then small increases in this tolerance rapidly decreases the number of pathways found. In contrast the empirical CDF begins at a smaller number of extracted paths and a constant rate of decrease in the number of pathways extracted is observed as the tolerance level increases.

For the entropy criteria it is clear that the extracted number of paths will dramatically decrease with a small increase in tolerance. The large decrease in pathways observed requires the setting of a small and imprecise entropy tolerance range, $[-0.66, -0.58]$, such that observations from both response labels could be extracted. We also observe that decreasing the entropy is clearly biased to the flower experiments. This bias towards the flower experiments is because they are predicted less accurately by some individual gene logistic regressions. These inaccurate predictions of the flower experiments results means they have a entropy and therefore decreasing the entropy tolerance results in large number of flower pathways to be extracted. This clear bias in the number of extracted pathways towards the flower experiments is likely seriously effect performance and stability of each classifier.

7. Results: Pathway Classification

For the HME3M model over all experiments we set the number of components to be four ($M = 4$) and $\lambda = 1$ to agree with the parameters used in [2]. For all models to cope with the number of pathways extracted inverse 20-fold cross validation is employed and each model is compared based on its test set correct classification rate (CCR). The results are presented in Figure 5.

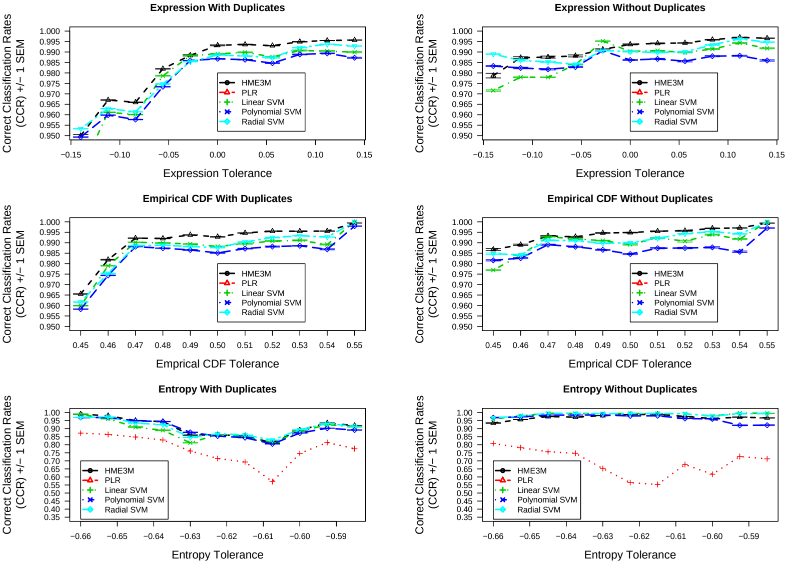

Fig. 5: Performance of each classification model on the pathways extracted with different gene activity criteria and with duplicate pathways removed.

From Figure 5 for the scaled expression models with duplicate pathways the optimal performance is reached after a tolerance of 0 is exceeded. In comparison the empirical CDF models with duplicates pathways reach the same optimal performance earlier in the experiments, after a tolerance of 0.47 is exceeded. For both criteria the once the optimum is reached all models maintain their performance with further increases in tolerance. These observations clearly show that the empirical CDF is extracting a more stable a set of pathways with a structure that is resistant to reasonable changes in tolerance. In contrast the performance profile for the entropy criteria with duplicates pathways is not stable. This is most likely due to the severe bias in the number of pathways extracted for the flower experiments. This severe bias towards the flower observation is artificially inflating the prediction results and removing this bias can be used to explain the decrease in model performances decrease as the entropy tolerance is increased.

Over all models an increase of approximately 2 % in classification performance is observed if duplicate pathways are removed. However removing the duplicate pathways does not increase the optimal performance but only increases model performances at lower tolerance values. The increased stability offered by removing duplicate pathways is dramatically observed for entropy criteria models where unexpected performance decrease is completely reversed. Therefore removing the duplicate pathways is a positive pre-processing step that has the effect of removing noise pathways and reducing the effect of any class bias.

Finally we compare the path structure within each model by correlating the

HME3M posterior probabilities for each model classifying each set of extracted pathways. We use the optimal pathways datasets observed in Figure 5 which are; for a scaled expression tolerance of 0, empirical CDF tolerance of 0.5 and an entropy tolerance of -0.66. We then train dataset specific models on a subset of each dataset individually. We use these models classify all pathways extracted from all datasets. Finally we compare the structure within model by correlating each models predicted posterior probabilities with the posterior probabilities of the dataset specific models. The results are presented in Table 1.

Table 1: Correlation between HME3M modeled pathways for each dataset.

	HME3M Model					
Pathway Dataset	**Expression With Duplicates**	**Expression Without Duplicates**	**Empirical CDF With Duplicates**	**Empirical CDF Without Duplicates**	**Expression With Duplicates**	**Entropy Without Duplicates**
Expression With Duplicates	**1.000**	**0.997**	**0.996**	**0.971**	0.052	0.107
Expression Without Duplicates	**0.997**	**1.000**	**0.996**	**0.971**	0.045	0.085
Empirical CDF With Duplicates	**0.997**	**0.997**	**1.000**	**0.983**	0.037	0.109
Empirical CDF Without Duplicates	**0.989**	**0.988**	**0.987**	**1.000**	0.013	0.069
Entropy With Duplicates	-0.035	0.026	0.028	-0.070	**1.000**	**0.824**
Entropy Without Duplicates	0.046	0.090	0.097	-0.078	**0.798**	**1.000**

The results in Table 1 show that scaled expression models and empirical CDF models produce highly correlated posterior probabilities when the datasets are switched. This clearly shows that models built from pathways extracted from either scaled expression tolerance or an empirical CDF tolerance will have similar structure. In contrast the entropy models clearly identify a different pathway structure which show little or no correlation with those found by other methods. This difference is due to the low entropy tolerance only extracting paths observed within a single class. Therefore the structure found within the entropy tolerance pathways is unlikely to be representative of the underlying microarray data.

8. Conclusions

In this paper we have compared three criteria for defining an observed pathway through a metabolic network. The results clearly show that employing a rank style transformation such as an empirical CDF upon the raw expressions will extract fewer pathways whilst maintaining the structure within the microarray to produce more stable and accurate pathway classifiers. We have also shown that removing duplicate pathways will increase the stability of the performances and reduce bias. Conversely we have also shown that employing a classification model to extract pathways is likely to produce a severe bias within observed pathway set, leading to unstable results that are not representative of the underlying microarray.

9. Acknowledgments

Timothy Hancock was in part supported by JSPS and in part by BIRD and and Hiroshi Mamitsuka was supported in part by BIRD .

References

[1] Dimitdadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A., e1071 - misc functions of the department of statistics, 2002.

[2] Hancock, T., Mamitsuka, H., A markov classification model for metabolic pathways. *Workshop on Algorithms in Bioinformatics (WABI)*, 2009.

[3] Jordan, M., Jacobs, R., Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[4] Kanehisa, M., Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[5] Mamitsuka, H., Okuno, Y., Yamaguchi, A., Mining biologically active patterns in metabolic pathways using microarray expression profiles. *SIGKDD Explorations*, 5(2):113–121, 2003.

[6] Park, M.Y., Hastie, T., Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30-50, 2008.

[7] Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., Lohmann, J.U., A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics*, 37(5):501–506, April 2005.