A Discrete Bayesian Network Framework for Discrimination of Gene Expression Profiles

Nikolay Balov
Department of Biostatistics and Computational Biology
University of Rochester Medical Center
Email: nikolay_balov@urmc.rochester.edu

***Abstract*—Using gene expression profiles for predicting phenotypic differences that result from cell specializations or diseases poses an important statistical problem. Graphical statistical models such as Bayesian networks may improve the prediction accuracy by identifying alternations in gene regulations due to the experimental conditions.**

We consider a discrete Bayesian network model that represents pairs of experimental classes by networks that share a common graph structure but have distinct probability tables. We apply a score-based network estimation procedure that maximizes the KL-divergence between class probabilities. The proposed method performs an implicit model selection and does not involve additional complexity penalization parameters. Classification of gene profiles is performed by comparing the likelihood of the estimated class networks.

We evaluate the performance of the new model against support vector machine, penalized linear regression and linear Gaussian networks. The classifiers are compared by prediction accuracy across 9 independent data sets from breast and lung cancer studies. The proposed method demonstrates a strong performance against the competitors.

***Keywords*-Gene expression, classification, Bayesian networks, discrete models**

I. Introduction

In this article we develop a Bayesian network approach with a novel supervised learning procedure to classification problems found in gene expression data. In the usual setting we have observations on N genes measured on n microarrays under two different experimental conditions. It is of interest to develop an algorithm that discriminates between these two conditions and is able to assign gene profiles to their corresponding classes.

Many classical approaches such as the linear discriminant analysis are ill suited for large N small n settings. Other models, such as LASSO [Tibshirani(1996)] and support vector machines (SVM) [Cortes *et al.*(1995)], either ignore the possible gene interactions or are unable to explicitly reveal them. On the other hand, the most essential feature of Bayesian network (BN) models is their ability to identify related genes and economically quantify their association [Spirtes *et al.*(2000b)], [Friedman *et al.*(2000)], [Imoto *et al.*(2003)]. It has already been demonstrated that BNs can compete with some state-of-the-art models in terms of discrimination and classification power [Ibrahim *et al.*(2002)] and [Helman *et al.*(2004)], but their widespread application is nevertheless not evident.

A major issue in applying BNs to analysis of gene expression data, which typically is of small sample size, is choosing the complexity of the underlying graph structure - too simple models lack representation power, too complex ones may overfit the data. One approach that addresses this model selection problem employs the Bayesian paradigm and performs maximum posterior estimation [Heckerman *et al.*(1995)]. Unfortunately, this solution requires either costly Monte Carlo implementation [Ibrahim *et al.*(2002)], or some heuristic approximation procedures [Helman *et al.*(2004)] and its effectiveness ultimately depends on the choice of prior. Alternatively, the so called constrained-based learning algorithms such as the PC algorithm [Spirtes *et al.*(2000a)] rely on specifying tuning parameters (the α-level for the conditional independence tests in PC) in order to choose the 'right' complexity of the model. Similarly, the score-based learning methods such as the penalized likelihood estimation [Buntine(1996)] depend on the choice of penalization parameters. In the theory of statistical learning it is an accepted practice for the model parameters to be tuned using some cross-validation (CV) procedure. For the purpose of structure learning however, this approach can be computationally prohibitive or can fail in very small sample settings. Moreover, the theoretical validity of CV is questionable [Braga-Neto(2007)]. We propose a discrete BN framework with a scoring learning algorithm that addresses the model selection issue by explicitly including the class information in the optimization function. It can be applied to any ordered gene set and, as we demonstrate below, is both feasible to carry out and effective in classification.

Discrete (or Categorical) Bayesian networks (CBNs) represent associations between categorical random variables through directed acyclic graphs (Section II-A). Multinomial representation of the conditional probabilities allows for detecting non-linear relationships that are inaccessible by, for example, linear Gaussian Bayesian networks (LBNs). Although the application of CBNs to continuous gene expression data is accompanied by loss of information due to discretization, in many cases (for examples see Figure 2 below), CBNs provide more sensitive representation of

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

gene associations than LBNs. The possibility of improving classification accuracy by mitigating the gene expression noise through discretization has already been noted [Shmulevich and Zhang (2002)]. In the context of microarray data, another advantage of discretization is its robustness to the so-called lab effect inherent in many multi-laboratory studies.

We start with the Maximum Likelihood (ML) principle for CBN estimation (Section II-A). Instead of optimizing a log-likelihood function however, we utilize the available class information and consider a function based on the KL divergence between the conditional probability estimated from the whole sample and the one corresponding to the first class only (see Eq. (3)). The proposed graph structure estimator maximizes thus defined function properly scaled. Then we provide a classification algorithm (Section II-C) that represents the experimental conditions by two distinct networks sharing the already estimated graph structure but having class-specific probabilities. The edges in the graph represent gene regulations that differ significantly between the classes. The classification of new observations is performed by comparing the likelihoods of the fitted networks.

In Section III, the proposed method is tested on 9 data sets - 6 breast cancer and 3 lung cancer studies - grouped in pairs by phenotypic and class criteria. The performance of 4 algorithms - the proposed one, SVM, LASSO and a LBN-based classifier using the PC algorithm - are compared on a collection of KEGG pathways as well as on sets of differentially expressed genes. The proposed classifier demonstrates superior performance in terms of several prediction measures across the considered data sets. We conclude with a discussion on some of the limitations and possible extensions of our method.

II. Methods

CBN is a probability model based on a directed acyclic graph (DAG) G with random categorical node-variables X_i, $i = 1, .., N$. Associated with each X_i is a set of nodes pa_i called parent set, such that $X_j \in pa_i$ whenever there is a directed edge connecting X_j and X_i in G. The number of nodes in pa_i is denoted by $|pa_i|$. We shall write $k \in X_i$ to index the states of X_i, that is, the possible categorical values for the node, and $j \in pa_i$ to index the combination of parent states of X_i. The second component of a CBN is a table of conditional probabilities $P(X_i|pa_i)$. We assume that for each i, $X_i|pa_i$ follows a multinomial distribution with parameter $P_i = \{P_{i,kj}\}_{k,j}$, where $P_{i,kj} \equiv P(X_i = k|pa_i = j)$. Hence $\sum_{k\in X_i} P_{i,kj} = 1$, for all $j \in Pa_i$, while $P_{i,j} \equiv \sum_{k\in X_i} P_{i,kj} = P(Pa_i = j)$. With $[X_i]$ and $[pa_i]$ we shall denote the number of states of X_i and pa_i, respectively. Clearly, $[pa_i] = \prod_{X_a\in pa_i}[X_a]$. The complexity of a CBN $(G, P = \{P_i\})$ is defined as $df(G) \equiv \sum_{i=1}^{N}[pa_i]([X_i] - 1)$ and equals the number of parameters needed for specifying the probability table P (*df* for degree of freedom).

For any DAG, there is a causal order of its nodes such that the parents of each node appear always earlier in this order. For simplicity, in the exposition below we shall assume that there is a causal order of the nodes coinciding with their index order and write $X_1 \prec X_2 \prec ... \prec X_N$.

A. Estimating Discrete Bayesian Networks from Labeled Data

We are interested in developing a CBN-based model suitable for two-class discrimination problems. The procedure we propose utilizes the class information to find a CBN that achieves maximum class separation in terms of an appropriate measure. The graph structure of the proposed estimator is then used for defining class-specific models.

Let $\{x^s\}_{s=1}^n$ be an n-sample of independent observations on $\{X_i\}_{i=1}^N$ and let each observation x^s have a label $c^s \in \{0, 1\}$ that assigns it to one of two classes; c^s are assumed observations on a class random variable C. We denote the labeled sample with $D_n = \{(x^s, c^s)\}_{s=1}^n$.

Let us first ignore the class-labels. The log-likelihood of a CBN (G, P) with respect to $\{x^s\}_{s=1}^n$ is

$$l(G, P|D_n) = \sum_{i=1}^{N} \sum_{j\in pa_i} \sum_{k\in X_i} n_{i,kj} \log P_{i,kj}, \qquad (1)$$

where $n_{i,kj} \equiv \sum_{s=1}^n 1_{\{x_i^s=k, pa_i^s=j\}}$ and $n_{i,j} \equiv \sum_k n_{i,kj}$. Using the ML principle, for a fixed DAG G, the parameter P can be easily estimated by maximizing $l(G, P|D_n)$ as a function of P

$$\max_P l(G, P|D_n) = \sum_{i=1}^{N} \sum_{j\in pa_i} n_{i,j} \sum_{k\in X_i} \hat{P}_{i,kj} \log \hat{P}_{i,kj}, \quad (2)$$

where $\hat{P}_{i,kj} \equiv n_{i,kj}/n_{i,j}$. The essential problem is thus to estimate the graph structure G. It is easy to verify that by adding new edges to G, the likelihood (2) cannot decrease. Hence, the ML estimation can easily result in over-fitting the data. A standard solution of this problem is, instead of maximizing the log-likelihood, to consider a scoring criterion of the form $l(G|D_n) - \lambda_n df(G)$ [Buntine(1996)], where λ_n is a penalization parameter indexed by the sample size n. Common choices include BIC, $\lambda_n = 0.5log(n)/n$, and AIC, $\lambda_n = 1/n$. Unfortunately, there is no consensus on what should be an optimal penalization and usually data-driven approaches such as CV are used for choosing λ_n. Below, we address this issue by proposing a log-likelihood-based scoring function that can be optimized to perform estimation and model selection without involving additional penalization parameters.

Similarly to $P_{i,kj}$, let us define the conditional probabilities pertaining to the first experimental class, $Q_{i,kj} = P(X_i = k|pa_i = j, C = 0)$ and let $\hat{Q}_{i,kj}$ be the corresponding point estimators as in (2). That is, $\hat{Q}_{i,kj} \equiv n^0_{i,kj}/n^0_{i,j}$,

where $n^0_{i,kj} \equiv \sum_{s=1}^{n} 1_{\{x_i^s=k, pa_i^s=j, c^s=0\}}$. We then consider the function

$$\mathcal{S}(G|D_n) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j\in pa_i}\sum_{k\in X_i} n_{i,kj}\log(\hat{P}_{i,kj}/\hat{Q}_{i,kj}) \quad (3)$$

and for a given class $\mathcal{G}$ of DAGs with nodes $\{X_i\}_{i=1}^N$, we propose to estimate G as

$$\hat{G} = arg\max_{G\in\mathcal{G}} \frac{\mathcal{S}(G|D_n)}{df(G)}. \quad (4)$$

We call $G \mapsto S(G|D_n)/df(G)$ scoring function of G with respect to the labeled sample D_n. We shall tentatively refer to $\hat{G}$ as BNKL estimator. Figure 1 shows a scoring function example for some real data.

Equivalently, $S(G|D_n)$ can be expressed as

$$\mathcal{S}(G|D_n) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j\in pa_i} n_{i,j} d_{KL}(\hat{P}_{i,j}||\hat{Q}_{i,j}),$$

where d_{KL} denotes the Kullback-Leibler (KL) divergence between the multinomial distributions $\hat{P}_{i,j} = \{P_{i,kj}\}_k$ and $\hat{Q}_{i,j} = \{Q_{i,kj}\}_k$. In other words, the optimization problem (4) aims at finding a network that achieves maximum separation between the classes in terms of KL-divergence. From the properties of the latter, we always have $\mathcal{S}(G|D_n) \geq 0$. Note that (3) reduces to (1), up to an additional constant, if $Q_{i,j}$ are uniform distributions, $Q_{i,j} = \{1/[X_i]\}_i$. Thus, in absence of labeled data, the latter case can be seen as default assumption for a reference class distribution. In the light of this observation we can think of $S(G|D_n)$ as a natural extension of the log-likelihood $l(G|D_n)$ to two-class problems.

The factor $df(G)$ in the denominator of (4) is needed for performing proper model selection. Let κ be the odds ratio for the first class, $\kappa = P(C=0)/P(C=1)$. Using results from the large sample statistical theory it can be shown that, under the null hypothesis $H_0 : P_{i,j} = Q_{i,j}$, for all i, j, $2n\kappa\mathcal{S}(G|D_n)$ is asymptotically $\mathcal{X}^2$ distributed with $df(G)$ degree of freedom (a formal proof is out of the scope of this paper). This fact explains the particular form of (4) - if H_0 is true, for sufficiently large fixed sample size, $E(\mathcal{S}(G|D_n))/df(G) \approx const$ and the G-scores will be properly normalized for comparison.

Next, we consider some computational aspects of solving (4). The class of all possible DAGs is super-exponential to the number of nodes N, which makes the problem prohibitive unless some strong constraints are imposed. In the current implementation of our method, we assume that the causality order of the nodes is known, and so is the maximum parent size M (these are the same structure constraints used by [Cooper and Herskovitz(1992)] in their K2 algorithm which however implements a completely different methodology). Formally, we consider collections of DAGs

$$\mathcal{G} = \{G|X_1 \prec X_2 \prec ... \prec X_N, |pa_i| \leq M, \forall i\}. \quad (5)$$

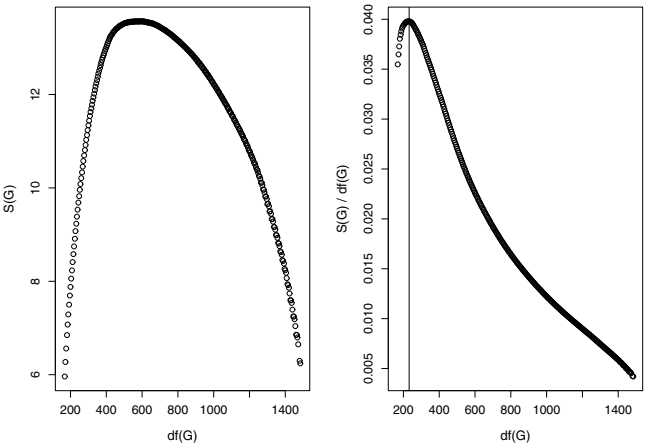

Figure 1. The S-function of Eq. 3 (left panel) and the scoring function of Eq. 4 (right panel) for the LUNG1 data set and the Small Cell Lung Cancer pathway with 84 genes. The complexity of the fitted estimator $\hat{G}$ is indicated by a vertical line.

In the real data tests below we use $M = 2$. For classes $\mathcal{G}$ as in (5), the optimal DAG $\hat{G}$ with respect to some data can be found by an efficient exhaustive search. We anticipate the conditions in (5) to be further relaxed, but this is a subject of an ongoing work we are not ready to report here.

B. Discrimination of Gene Expression Profiles

We return to the main goal of this investigation - developing a CBN-based classifier for two-class problems. We have shown, Eq. (5), how we can choose a graph structure to separate a labeled sample. In the view of our approach, it is natural to assume that the class representing networks share a common DAG but have distinct probability tables.

More specifically, we assume that: (i) we are given a (discrete) sample D of observations on X_i, $i = 1, .., N$; (ii) the class sub-samples $D_0 = D \cap \{c = 0\}$ and $D_1 = D \cap \{c = 1\}$ come from two CBNs, (G, P_0) and (G, P_1), with DAG G and probability tables P_0 and P_1; (iii) G is such that $X_1 \prec X_2 \prec ... \prec X_N$ and its maximum parent size is M. Then we propose the following classification algorithm.

Training:

1) Fit a DAG $\hat{G}$ to D according to (4).
2) Define two CBNs $(\hat{G}, \hat{P}_0)$ and $(\hat{G}, \hat{P}_1)$ by estimating the class-specific conditional probability tables $\hat{P}_0(\hat{G}|D_0)$ and $\hat{P}_1(\hat{G}|D_1)$ as in (2).

Let x be an observation on the node-genes X_i's.

Classification:

1) Calculate the log-likelihoods $l_0 = l(\hat{G}, \hat{P}_0|x)$ and $l_1 = l(\hat{G}, \hat{P}_1|x)$.
2) Assign x to the class with greater log-likelihood l.

In the case of gene expression data, the above algorithm is preceded by a discretization step - any sample $\{y^s\}_{s=1}^n$ of observations on a continuous gene expression vector $(Y_i)_{i=1}^N$ is transformed into categorical sample $\{x^s\}_{s=1}^n$. It

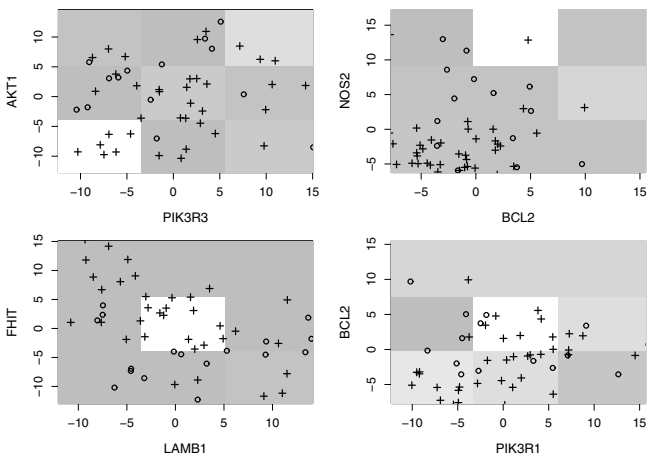

Figure 2. Example of gene expression discretization and 3-nomial representation of gene interactions. Used are 4 pairs of genes from the Small Cell Lung Cancer pathway and LUNG1 data set. Different point shapes for the classes are used. Overlaid on the cross plots are the discretization regions shaded according to the KL-values $p_{ij}\log(p_{ij}/q_{ij})$, $i,j=1,2,3$ (regions with higher values are shown lighter).

is an accepted practice for gene expression levels to be discretized into 3 categories - 'under-expressed', 'baseline' and 'over-expressed'. In our experiments we employ a 3-category uniform discretization as follows. After excluding 5% of the most extreme values (to guard against outliers), the range of Y is divided into equal intervals and an observation y is assigned a categorical value x according to the interval into which y falls. In $training \rightarrow test$ prediction problems, the discretization parameters (cut-off points) are determined strictly from the training sample and are used to discretize the test sample. Our choice of uniform discretization is largely due to its simplicity and good performance in practice.

Figure 2 shows some examples of gene expression discretization and modeling gene interactions with 3-nomial distributions. For instance, the directed edge $(AKT1 \rightarrow PIK3R3)$ have probabilities $p_{i,j} = P(X_{PIK3R3} = i|X_{AKT1} = j)$ which are $p_{i1} = (0.40, 0.47, 0.13)$, $p_{i2} = (0.33, 0.50, 0.17)$ and $p_{i3} = (0.29, 0.50, 0.21)$. The probabilities corresponding to the first class only are $q_{i1} = (0.00, 0.67, 0.33)$, $q_{i2} = (0.72, 0.14, 0.14)$ and $q_{i3} = (0.17, 0.83, 0.00)$. The KL-divergence between p and q is large enough to indicate a significant class difference in the gene interaction (p-val$\approx$0). On the other hand, no significant class difference in the linear interaction between the genes is detected (p-val=0.18) - a manifestation of discrete representation higher sensitivity compared to the linear one.

The BNKL algorithm is implemented using the *catnet* package for **R**. To illustrate the algorithm complexity, we present average training times of a BNKL classifier with 3 categories per gene and M=2, for sample size $n = 250$ and different network sizes N.

N	10	25	50	100	200	250	500
time(sec)	0.02	0.06	0.2	1.3	8	14	161

The results are in agreement with the theoretical complexity of the algorithm $O(n * N^{M+1})$.

C. Benchmark Classifiers

We compare the proposed algorithm to 3 state-of-the-art classifiers. The first one is SVM with radial kernel as implemented in the *e1071* package for **R**. The kernel parameter γ is tuned via CV on the training data for optimal performance. The benchmark performance of SVM is well established [Statnikov *et al.*(2005)]. The second classifier, LASSO, performs a penalized linear regression on the binary class variable. The algorithm is available in the *lars* package for **R**. The third reference classifier employs a LBN model as follows: (1) a DAG $\hat{G}$ is fitted to the combined sample $D_0 \cup D_1$ using the PC algorithm [Kalisch and Bhlmann (2007)] with Gaussian test for conditional independence at $\alpha = 0.05$ (see *pcalg* package for **R**); (2) for each i, two distinct sets of $(Y_i|pa_i)$ regression parameters are estimated for each class separately; (3) a test sample is assigned to the class which model has greater likelihood. Note that, SVM, LASSO and LBN+PC (which we refer to as PC), in contrast to BNKL, are applied directly to continuous observations on $(Y_i)_{i=1}^N$.

III. Results

In this section we test the performance of the BNKL classifier against SVM, LASSO and PC by prediction accuracy across independent data sets. We consider 9 data sets arranged in groups by phenotypic and class compatibility:
(1) two breast cancer sets, GSE1456 (n=159) and GSE3494 (n=233), which we refer to as BRES1 and BRES2, with subjects classified by their survival status.
(2) four breast cancer sets, GSE2990 (n=183), GSE7390 (n=198), GSE20711 (n=90) and GSE2034 (n=286), which we refer to as BRER1, BRER2, BRER3 and BRER4, with subjects classified by their estrogen receptor (ER) status.
(3) three lung cancer data sets, GSE10245 (n=58), GSE18842 (n=46), GSE31799 (n=49), which we refer to as LUNG1, LUNG2 and LUNG3, of tumor samples from non-small cell lung cancer classified into adenocarcinoma (AC) and squamous cell carcinoma (SCC) subtypes. Prior to the analysis the raw microarray sets are RMA normalized and the probe expression levels across sample records are standardized.

The classifiers are applied on two types of ordered gene subsets: differentially expressed (DE) genes identified by two-sample t-test and a collection of curated pathways. In the former case, the genes are ordered by decreasing p-values. The considered collection of KEGG pathways is available at the GSEA website (*c2.cp.kegg.v3.0.symbols.gmt*). It contains 186 gene sets of

Table I
PREDICTION ACCURACY USING DIFFERENTIALLY EXPRESSED GENES (DES) SELECTED BY *t-test* ON THE TRAINING DATA. D_{10}: TOP 10 DES. D_{100}: TOP 100 DES. ALSO GIVEN ARE THE OVERALL RANKS. ROW BEST RESULTS ARE IN BOLD.

GSE data sets		classifiers			
training→test	genes	BNKL	SVM	LASSO	PC
BRES1→BRES2	D_{10}	**59.75**	50.17	51.31	50.00
	D_{100}	**58.21**	52.58	56.29	50.36
BRES2→BRES1	D_{10}	**60.68**	53.31	51.66	51.25
	D_{100}	**66.52**	60.35	51.24	53.31
BRER1→BRER2	D_{10}	**63.14**	58.663	52.34	50.00
	D_{100}	**86.96**	63.69	70.79	50.00
BRER2→BRER1	D_{10}	80.85	79.46	80.59	**82.86**
	D_{100}	79.38	**82.27**	81.27	75.59
BRER3→BRER4	D_{10}	77.72	72.93	**80.11**	64.01
	D_{100}	79.70	74.85	**81.68**	69.58
BRER4→BRER3	D_{10}	**79.46**	59.38	60.42	50.00
	D_{100}	**71.43**	69.35	58.04	50.00
LUNG1→LUNG2	D_{10}	**87.50**	81.25	76.56	54.69
	D_{100}	**84.38**	81.25	76.56	50.00
LUNG2→LUNG1	D_{10}	65.42	69.72	**87.22**	50.00
	D_{100}	**90.42**	67.22	78.75	50.00
LUNG1→LUNG3	D_{10}	89.05	88.27	**90.76**	75.00
	D_{100}	**91.55**	88.27	85.00	50.00
LUNG3→LUNG1	D_{10}	86.67	89.17	**90.42**	77.78
	D_{100}	90.69	89.17	80.42	**93.19**
LUNG2→LUNG3	D_{10}	65.00	64.74	**66.64**	50.00
	D_{100}	**90.78**	65.52	85.43	50.00
LUNG3→LUNG2	D_{10}	**85.94**	84.37	**85.94**	75.00
	D_{100}	90.63	92.19	87.50	**93.75**
Total Rank		**39**	60	53	86

variable size, from 10 to 389. We apply the BNKL model using the orders of the genes in the pathways.

To evaluate the performance of the algorithms, we have implemented across data set prediction for pairs of compatible data sets. The performance is measured as balanced accuracy in percents according to the formula ACC=50(TP/P + TN/N), where P and N are the number of test observations in the classes, while TP and TN are the number of correctly assigned observations to the first and second class, respectively. The 'random guess' procedure thus has accuracy of 50 on average and so does any algorithm that assigns all observations to one class. Another alternative measure we utilize is the area under the Precision (TP/(TP+FP)) - Recall (TP/(TP+FN)) curve or AUPR, again in percents from 0 to 100.

A. Classification with Differentially Expressed Genes

It is standard practice in microarray studies to identify groups of DE genes employing a two-sample test and use them as biomarkers for discriminating between the involved experimental conditions. It is then an important question of whether models that employ between-gene interactions can improve the performance upon the former essentially univariate approach.

In Table I we present DE-based prediction results for 12 $(training, test)$-pairs. For each training set, the t-test p-values for all genes are calculated and ordered decreasingly. Then the top 10 (D_{10}) and top 100 (D_{100}) DE genes are used for classification. The performance numbers achieved by LASSO, a multi-linear regression, are indicative for what is the best possible prediction accuracy if the biomarker approach is followed. For each prediction case (table row) the classifiers are ranked according to their ACC and then the total ranks are calculated. As evident, BNKL most often achieves best accuracy and has the best total rank of 39, followed by LASSO with total rank of 53. We are aware that this difference in performance may be due to (i) discrete multinomial vs. continuous gene expression representation and/or (ii) multi-variate (employing between-gene associations) vs. uni-variate (no gene associations) modeling. Further tests are needed to resolve this confounding. In any case, the results clearly indicate the merits of incorporating the BNKL model in a biomarker framework.

We observe that the performance does not necessarily increase by going from D_{10} to D_{100}, because, we reason, the more genes are considered the more difficult the model selection problem becomes. Note that these top genes are selected out of some tens of thousands and direct interactions between them seem unlikely. Consequently, simpler networks, as those selected by BNKL, are expected to perform better. A closer look reveals that, with a few exceptions, the PC algorithm chooses too complex, overfitting networks (numbers not shown), which explains its relatively lower performance.

B. Classification Using Pathways

The notion of pathways has been developed to relate gene systems to a broad range of biological functions. Recent methodologies such as Gene Set Enrichment Analysis (GSEA) [Subramanian *et al.*(2005)], have focused on identifying pathways that discriminate different experimental conditions. In the context of CBNs, we utilize pathways as priors to facilitate inference and lessen the computational complexity. First, the limited number of genes in the pathways makes the BNKL learning relatively fast (see the performance numbers above). Second, based on prior evidence, the genes in the pathways are related and it is thus reasonable to search for class differences due to their interactions. The latter assumption however is not essential for our algorithm. Although, when no interactions are detected, BNKL is equivalent to a naive discrete classifier.

We apply the competing classifiers on each of the considered KEGG pathways. For a data set pair (A, B), we run and report the accuracy of across data set predictions $A \rightarrow B$ and $B \rightarrow A$ (in $training \rightarrow test$ notation). In this way, for each pathway we obtain one binary prediction vector for the test sample (recall, we use 0 and 1 as class labels). The pathway-average prediction vector, with components between 0 and 1 giving the class confidence for each test observation, is what we use to calculate an AURP value measuring the prediction performance over all pathways. Table II reports ACC and

Table II
KEGG PATHWAY-AVERAGE PERFORMANCE MEASURED BY PREDICTION ACCURACY ACC AND AUPR (IN PERCENTS), AND THE CORRESPONDING TOTAL RANKS.

GSE data sets	classifiers			
training → test	BNKL	SVM	LASSO	PC
ACC				
BRES1→BRES2	**61.57**	50.00	50.00	53.94
BRES2→BRES1	**57.85**	50.00	50.00	51.68
BRER1→BRER2	**71.98**	57.81	53.13	63.72
BRER2→BRER1	**79.59**	79.33	76.39	50.34
BRER3→BRER4	**87.29**	51.67	83.59	50.00
BRER4→BRER3	71.58	58.18	50.00	**78.72**
LUNG1→LUNG2	**79.69**	73.44	73.44	53.13
LUNG2→LUNG1	**87.64**	68.75	70.97	56.25
LUNG1→LUNG3	**85.78**	85.00	80.78	55.00
LUNG3→LUNG1	90.42	90.42	**93.19**	50.00
LUNG2→LUNG3	**90.00**	65.51	75.09	53.45
LUNG3→LUNG2	79.69	**82.81**	76.56	50.00
Total Rank	**15**	30	32	37
AUPR				
BRES1→BRES2	**42.18**	38.89	36.03	41.23
BRES2→BRES1	**41.53**	38.76	32.79	41.18
BRER1→BRER2	**94.26**	93.54	90.63	93.81
BRER2→BRER1	94.88	95.42	95.22	**96.01**
BRER3→BRER4	92.81	93.99	**94.23**	93.26
BRER4→BRER3	69.29	63.49	46.89	**73.77**
LUNG1→LUNG2	**89.93**	85.66	89.92	87.88
LUNG2→LUNG1	**97.20**	96.14	95.29	94.94
LUNG1→LUNG3	**94.24**	92.09	90.51	89.63
LUNG3→LUNG1	**97.81**	97.77	97.53	96.77
LUNG2→LUNG3	92.54	**95.42**	95.16	93.51
LUNG3→LUNG2	**96.43**	92.86	89.28	85.45
Total Rank	**22**	29	36	33

Table III
CLASSIFIER COMPARISON BASED ON INDIVIDUAL PATHWAY PREDICTION ACCURACY. SHOWN ARE THE MEDIAN DIFFERENCES BETWEEN BNKL AND SVM, LASSO AND PC, AND THE P-VALUES (IN PARENTHESES) OF MANN-WHITNEY TEST FOR ZERO DIFFERENCE (THOSE LESS THAN 0.001 ARE SET TO 0).

GSE data sets	BNKL-SVM	BNKL-LASSO	BNKL-PC
BRES1→BRES2	4.41 (0)	4.62 (0)	2.71 (0)
BRES2→BRES1	3.69 (0)	5.53 (0)	2.27 (0)
BRER1→BRER2	6.76 (0)	8.61 (0)	4.41 (0)
BRER2→BRER1	-2.75 (0)	-1.92 (0)	13.08 (0)
BRER3→BRER4	11.41 (0)	-1.19 (0.70)	13.96 (0)
BRER4→BRER3	3.79 (0)	10.27 (0)	6.03 (0)
LUNG1→LUNG2	3.13 (0)	6.25 (0)	12.50 (0)
LUNG2→LUNG1	8.19 (0)	5.42 (0)	12.57 (0)
LUNG1→LUNG3	-0.95 (0.011)	5.99 (0)	11.08 (0)
LUNG3→LUNG1	-1.52 (0.001)	4.93 (0)	13.54 (0)
LUNG2→LUNG3	5.86 (0)	2.41 (0)	10.68 (0)
LUNG3→LUNG2	-1.56 (0)	2.90 (0.002)	16.29 (0)

AUPR, as well as the total rank for each of the competing algorithms. The BNKL classifier achieves the highest overall scores with respect to both measures.

Reported in Table III is another performance comparison based on the achieved ACC on each individual pathway. There, the pathway accuracies of the benchmark classifiers are subtracted from that of BNKL and Mann-Whitney test is performed on the resulting 3 sets of 186 differences. A significant positive difference indicates better performance of BNKL, while a significant negative one favors the competing classifier. According to the reported results, only SVM is a real competitor of BNKL, and still, BNKL is better in 8 out of the 12 considered cases.

Table IV
PREDICTION ACCURACY FOR (LUNG1, LUNG2) ALONG WITH SOME OF THE TOP PERFORMING PATHWAYS.

	BNKL	SVM	LASSO	PC
Mean ACC	74.67	69.19	69.16	62.61
Top Pathways				
VASOPRESSIN REGULATED WATER R	**86.13**	74.38	68.28	72.81
SPHINGOLIPID METABOLISM	**85.77**	70.62	71.72	58.12
ADHERENS JUNCTION	**85.68**	70.17	67.45	59.24
HEDGEHOG SIGNALING PATHWAY	**85.05**	70.95	73.78	72.52
GAP JUNCTION	**84.76**	72.50	75.66	56.09
PROSTATE CANCER	**84.60**	73.28	72.61	59.84
CHRONIC MYELOID LEUKEMIA	**84.58**	67.99	72.16	64.84
VEGF SIGNALING PATHWAY	**84.53**	68.77	67.49	59.90
UBIQUITIN MEDIATED PROTEOLY	**83.96**	75.47	73.16	57.97
TGF BETA SIGNALING PATHWAY	**83.78**	69.86	69.17	70.02
ARGININE AND PROLINE METABOL	**83.70**	74.55	72.55	70.52
GLYCOSPHINGOLIPID BIOSYNTHES	**83.30**	78.59	70.31	64.97
...				
SMALL CELL LUNG CANCER	**80.35**	71.11	65.15	62.03

Table IV shows the pathway-based ACC performance averaging LUNG1→LUNG2 and LUNG2→LUNG1 prediction cases. The mean classification accuracy over all pathways is reported as well as the top pathways with best performance. Among them is the Small Cell Lung Cancer pathway. In Figure 3 we plot the DAG representing this pathway as estimated by BNKL - a simple network with 16 edges and complexity 232 (see also the related Figure 1) - telling us that BNKL favors parsimonious models (for comparison, the PC algorithm reconstructs a much more complex network with 82 edges, not shown).

IV. CONCLUSION

We have proposed a novel discrete Bayesian network classifier for analysis of gene expression data. The experimental conditions of interest are modeled by networks sharing a common graph structure but having distinct probability tables representing the difference in the implicated gene regulations. The application of multinomial conditional probabilities allows for representing complex, non-linear gene associations. The learning algorithm implements a scoring criteria based on the KL-divergence between class probabilities. The graph structure of the network is chosen such as to maximize the class differences. By conducting across independent data set prediction, we have demonstrated that the performance of the proposed approach is comparable, if not superior, to some state-of-the-art classification algorithms.

In its current implementation, the BNKL classifier can be applied on any ordered set of genes of reasonable size

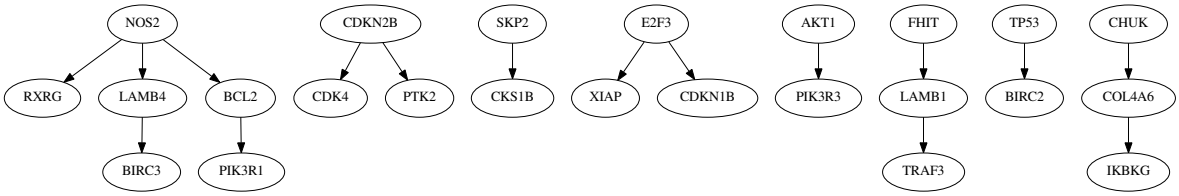

Figure 3. The Small Cell Lung Cancer pathway for the LUNG1 data set as estimated by BNKL. The total number of genes in the pathway is 84 but here only the connected ones are shown. Note that BNKL does not necessarily detect all gene associations present in the data but only those that exhibit significant class differences.

($N < 1000$) such as gene pathways or sets of differentially expressed genes. Unfortunately, the application of BNKL to a complete probe set with some 20,000 genes is computationally prohibitive. However, we believe that the restrictions in (5) can be further relaxed with only slight increase of the complexity of the algorithm. First, the fixed gene order condition can be dropped in favor of some more optimal order such as the one of decreasing KL-divergence between the marginal class distributions of the nodes. Also, a whole genome analysis would be possible, although at some loss of full optimality, if the gene-node parents are chosen from sets of potential parents with manageable sizes. These potential parents can be selected according to some degree of association with the child genes.

Finally, we would like to report on further validation and testing on the proposed here framework by including more diverse data sets and benchmark classifiers. Such investigation would not be possible without the available and constantly expanding public repositories of gene expression data.

Funding: This work is supported by NIH grant K99LM009477 from the National Library of Medicine. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

REFERENCES

[Braga-Neto(2007)] Braga-Neto, U (2007). Fads and fallacies in the name of small-sample microarray classification. *IEEE Sig. Proc. Mag., 24, 91-99.*

[Buntine(1996)] Buntine, W (1996). A guide to the literature on learning graphical models. *IEEE Transactions on knowledge and data engineering, Vol.8, No.2.*

[Cortes *et al.*(1995)] Cortes, C, Vapnik, V (1995). Support-vector networks, *Machine Learning 20, 273-297.*

[Cooper and Herskovitz(1992)] Cooper, G and Herskovitz, E (1992). A Bayesian method for the induction of probabilistic networks from data, *Machine Learning 9, 330-347.*

[Friedman *et al.*(2000)] Friedman, N, Linial, M, Nachman, I and Peer., D (2000). Using Bayesian networks to analyze expression data, *J. Comput. Biol., 7:601620.*

[Heckerman *et al.*(1995)] Heckerman, D, Geiger, D and Chickering, D (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 10, 197-243.*

[Helman *et al.*(2004)] Helman, P, Veroff, R, Atlas, S and Willman, C (2004). A Bayesian Network Classification Methodology for Gene Expression Data, *J. Comput. Biol., 11(4):581-615.*

[Ibrahim *et al.*(2002)] Ibrahim, J, Chen, M, Gray, R (2002). Bayesian Models for Gene Expression With DNA Microarray Data, *J. Amer. Stat. Ass., 97(457):88-99.*

[Imoto *et al.*(2003)] Imoto S, Higuchi T, Goto H, Tashiro K, Kuhara S, (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, IEEE Comp. Sys. Bioinformatics (CSB03) 2: 104113.

[Kalisch and Bhlmann (2007)] Kalisch, M, Bhlmann, P (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm, *Machine Learning Research 8, 613-636.*

[Shmulevich and Zhang (2002)] Shmulevich, I and Zhang, W (2002). Binary analysis and optimization-based normalization of gene expression data, *Bioinformatics, 18(4):555-565.*

[Spirtes *et al.*(2000a)] Spirtes, P, Glymour, C, and Scheines, R (2000). Causation, Prediction, and Search, *The MIT Press, 2nd edition.*

[Spirtes *et al.*(2000b)] Spirtes, P, Glymour, C, Scheines, R, Kauffman, S, Aimale, V, Wimberly, F (2000). Constructing Bayesian network models of gene expression networks from microarray data, *Proc. Atl. Symp. on Comp. Biology , 1-5.*

[Statnikov *et al.*(2005)] Statnikov, A *et al.* (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, *Bioinformatics, 21(5):631-643.*

[Subramanian *et al.*(2005)] Subramanian, A *et al.*(2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. USA, 102, 15545-15550.*

[Tibshirani(1996)] Tibshirani, R (1996). Regression shrinkage and selection via the LASSO, *Jour. Royal Stat. Soc., Series B 58, 267-288.*