BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 12 2011, pages 1667–1674
doi:10.1093/bioinformatics/btr269

Systems biology Advance Access publication May 5, 2011

Pathway analysis of high-throughput biological data within a Bayesian network framework

Senol Isci[1], Cengizhan Ozturk[1], Jon Jones[2] and Hasan H. Otu[3,4,*]

[1]Bogazici University, Institute of Biomedical Engineering, 34342, Istanbul, Turkey, [2]Department of Urology, Johannes Gutenberg University, 55131 Mainz, Germany, [3]Department of Bioengineering, Istanbul Bilgi University, 34060, Istanbul, Turkey and [4]Department of Medicine, BIDMC Genomics Center, Harvard Medical School, Boston, MA 02115, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Most current approaches to high-throughput biological data (HTBD) analysis either perform individual gene/protein analysis or, gene/protein set enrichment analysis for a list of biologically relevant molecules. Bayesian Networks (BNs) capture linear and non-linear interactions, handle stochastic events accounting for noise, and focus on local interactions, which can be related to causal inference. Here, we describe for the first time an algorithm that models biological pathways as BNs and identifies pathways that best explain given HTBD by scoring fitness of each network.

Results: Proposed method takes into account the connectivity and relatedness between nodes of the pathway through factoring pathway topology in its model. Our simulations using synthetic data demonstrated robustness of our approach. We tested proposed method, Bayesian Pathway Analysis (BPA), on human microarray data regarding renal cell carcinoma (RCC) and compared our results with gene set enrichment analysis. BPA was able to find broader and more specific pathways related to RCC.

Availability: Accompanying BPA software (BPAS) package is freely available for academic use at http://bumil.boun.edu.tr/bpa.

Contact: hotu@bidmc.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 7, 2010; revised on April 12, 2011; accepted on April 13, 2011

1 INTRODUCTION

High-throughput biological data (HTBD) are generated in a variety of ways, including through deep sequencing and microarrays. These data can provide a snapshot of regulatory processes in the cell by inferring gene networks from experimental data. In particular, Bayesian network (BN) models have gained popularity for the task of learning biological pathways from microarray gene expression data (Friedman *et al.*, 2000; Imoto *et al.*, 2002). In gene network modeling studies using BNs, nodes generally represent expression level of a gene and edges represent relationship between genes. BN models capture both linear and non-linear interactions between sets of random variables and handle stochastic events in a probabilistic framework accounting for noise. This results in the emphasis of only the strong relations in the observed data. BNs are, therefore, viable candidates for modeling gene regulation systems, where stochastic effects and large amounts of noise are expected. Furthermore, BNs are able to focus on local interactions, where each node is directly affected by a relatively small number of nodes and interactions defined by a BN can be related to causal inference (Friedman *et al.*, 2000). These properties are similarly observed in biological networks justifying the use of BNs in exploring pathways in the setting of gene interaction networks, using HTBD.

Arguably the most popular HTBD type is microarrays, where an identification of differentially expressed genes between two groups of samples initially relied on individual gene analysis (IGA). An alternative approach, called pathway analysis, functional enrichment analysis or gene set analysis (GSA) (Nam and Kim, 2008), which focuses on directly determining predefined gene sets or classes that are significantly regulated, has received a great deal of attention. GSA methods score groups of genes and can identify genes that exhibit subtle changes at an individual level, but show concordant enrichment within a set (Subramanian *et al.*, 2005). Here we propose a new method of pathway analysis, which uses a graph theoretic approach and BN theory to evaluate whether a putative pathway successfully describes the underlying HTBD. Previously described GSEA (Subramanian *et al.*, 2005) or Gene Ontology (Hosack *et al.*, 2003) based methods do not take into account the connectiveness of analyzed gene lists. There have been methods proposed to take into account the GO graph topology (Alexa *et al.*, 2006), overlap between GO categories in the GO hierarchy (Lu *et al.*, 2008) or modeling interactions between GO categories (Bauer *et al.*, 2010) in assessing the significance of enrichment of a GO term based on experimental data. However, none of these methods takes into account the network or structure defining the relation between the genes in each category.

The proposed method [Bayesian Pathway Analysis (BPA)] considers the topology via which genes interact with each other when analyzing a group of genes. Using pathway information from global databases, we model each biological pathway as a BN after merging repeating entries and, if necessary, solving for cyclicity, while preserving the dependencies entailed by the original pathway. We consider the resulting BN, which is a graphical representation of gene interactions rendered by the given pathway, with non-informal, uniform belief priors. We quantify the degree to which observed experimental data fit this BN using Bayesian Dirichlet equivalent (BDe) score calculation, where the BN is updated with input data

*To whom correspondence should be addressed.

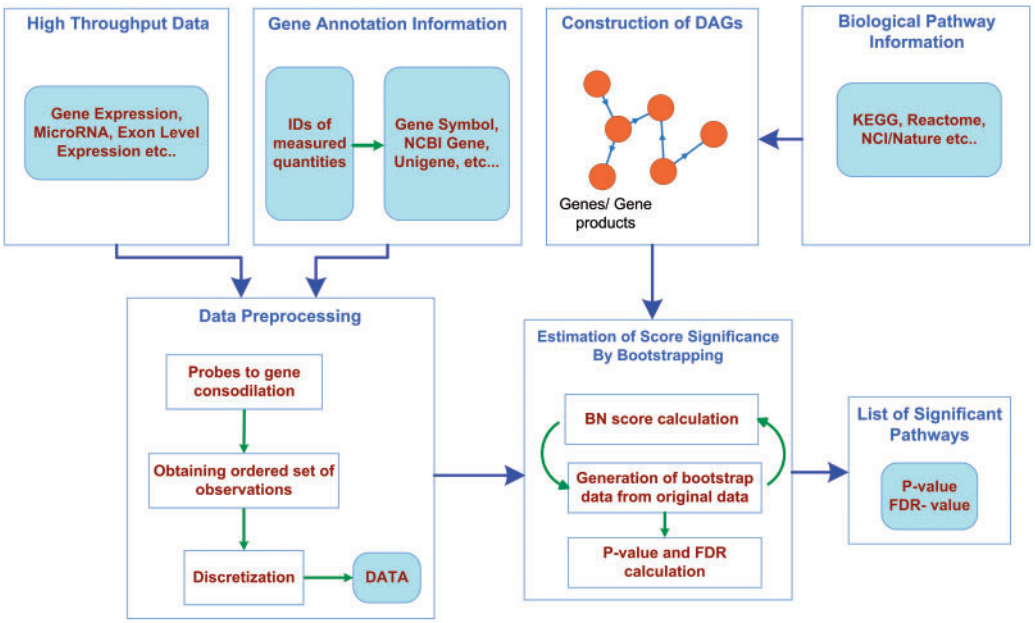

Fig. 1. DAGs are created from pathways in external databases, where, if necessary, multiple occurrences of same nodes are merged and cyclicity is resolved preserving dependencies entailed by the original pathway. Nodes in DAGs and microarray probes are mapped using gene annotation information. Multiple probes mapped to the same node are replaced by a robust average value followed by discretization of all signal values. A score metric, which is a measure of probability of data given the fixed network structure, is calculated. Significance of this score (p-value) is assessed by randomization of input data via bootstrapping. FDR values are calculated to account for multiple hypothesis testing.

during score calculation. We assess statistical significance for the score of each pathway by testing it against datasets generated by applying randomization via bootstrapping. Results are evaluated in forms of nominal p-values and false discovery rate (FDR) values correcting for multiple hypotheses testing. Overall workflow used in BPA is depicted in Figure 1.

Renal cell carcinoma (RCC) is the sixth leading cause of cancer deaths in the United States and has no established biomarker for early detection or follow-up (Jones *et al.*, 2008; Mills *et al.*, 2009). Most common histological subtypes of RCC are clear-cell RCC (cRCC) and papillary RCC (pRCC) generally related to deficiencies in von Hippel–Lindau, rapamycin complex 1 kinase (mTOR) and fumarate hydratase (Brugarolas, 2007). Treatment for metastatic RCC includes rather unspecific application of cytokine therapies (e.g. interferon-α and interleukin-2) and more targeted use of receptor tyrosine kinase inhibitors (Sorafenib and Sunitinib), mTOR inhibitors (Everolimus and Temsirolimus) and monoclonal anti-vascular endothelial growth factor antibody therapy with Bevacizumab (Brugarolas, 2007; Mills *et al.*, 2009). These therapies offer acceptable response rates (30–40%), but are not beneficial in overall survival. Therefore, a more detailed analysis of molecular pathways underlying the RCC is essential. We have previously analyzed transcriptional profiling of various RCC subtypes and in a separate study obtained a predictive proteomic signature that distinguishes between interleukin-2 therapy responders and non-responders (Jones *et al.*, 2005; Jones *et al.*, 2008). Others have also used gene expression and proteomic approaches to gain insight into the molecular pathways governing RCC (Furge *et al.*, 2007; Gumz *et al.*, 2007; Koeman *et al.*, 2008; Kort *et al.*, 2008; Lenburg *et al.*, 2003; Perroud *et al.*, 2006; Wang *et al.*, 2009; Yang *et al.*, 2005). In addition to cRCC and pRCC, we used BPA to analyze the rare RCC subtype chromophobe RCC (chRCC) and other renal malignancies, such as transitional cell cancers of the renal pelvis (TCC) and Wilms' tumors (WT) or benign renal tumors such as oncocytomas (OC).

2 METHODS

2.1 Pathway information retrieval

Biochemical network data of pathways were retrieved from KEGG (Kanehisa *et al.*, 2008), NCI/Nature Pathway Interaction Database (Schaefer *et al.*, 2009), Reactome (Vastrik *et al.*, 2007) and HumanCyc (Romero *et al.*, 2005) representing molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes and diseases.

2.2 Construction of directed acyclic graphs

A BN is a compact graphical representation of the joint probability distribution over a set of random variables in the form of a directed acyclic graph (DAG), where nodes represent random variables. The DAG encodes assertions of conditional independence, which are generally represented as a set of conditional probability tables (CPT). When modeling pathways as BNs, we first merge repeating entries (for example, Smad2/3 is present at several locations in the KEGG TGFβ pathway, see Supplementary Figures S1 and S2), as a single node in the DAG while conserving edge relations. Cyclic paths are eliminated using Spirtes' method (Spirtes, 1995). In this procedure, graph representation of structural equation models (SEM) is converted to collapsed acyclic graphs such that d-separations in the collapsed graph entails the same independency relations defined by the model (Pearl, 2000). Cyclegroups (set of all cycles sharing at least one node) are found using Tarjan's algorithm (Tarjan, 1972). For a given BN, all d-separations

are conditional independencies and every conditional independency implied by the BN is identified by *d*-separations (Neapolitan, 2004). Therefore, our way of solving cyclicity preserves distributional features explained by the pathway after it has been converted to a DAG.

2.3 Microarray data preprocessing and discretization

BPA assumes normalized data as an input. First, IDs used in the array platform, corresponding to a given node in the pathway representation, are pooled and one representative signal value per node is calculated using the one-step Tukey's biweight algorithm (Hoaglin *et al.*, 2000). Currently, BPA addresses experimental designs consisting of two groups of samples (e.g. cancer versus normal) though generalization of this framework to multiple groups is straightforward. For a given BN (converted from a pathway), we obtain observed fold changes (FCs) for genes in this BN (pathway) by pairwise comparisons of samples in each group. This approach provides a distribution of FC values and a reasonable dataset size used to score each BN. Let $\boldsymbol{SG_1}$ and $\boldsymbol{SG_2}$ represent two groups of samples in the dataset with $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ samples in each group, respectively. Let $\boldsymbol{g_{i1j}}$ and $\boldsymbol{g_{i2k}}$ be the expression values of the $\boldsymbol{i}$-th node in the pathway in $\boldsymbol{j}$-th and $\boldsymbol{k}$-th samples in the sample groups $\boldsymbol{SG_1}$ and $\boldsymbol{SG_2}$, respectively, where $\mathbf{1}\leq \boldsymbol{j}\leq \boldsymbol{C_1}$ and $\mathbf{1}\leq \boldsymbol{k}\leq \boldsymbol{C_2}$. Let $\boldsymbol{X_i}$, $\mathbf{1}\leq \boldsymbol{i}\leq \boldsymbol{N}$, represent the random variable for the $\boldsymbol{i}$-th node in a BN with $\boldsymbol{N}$ nodes. An ordered set of observations, $\boldsymbol{O}$, for the dataset is obtained by pairwise comparison of all samples in sample groups $\boldsymbol{SG_1}$ and $\boldsymbol{SG_2}$. The $\boldsymbol{l}$-th element of $\boldsymbol{O}$, $\boldsymbol{o_l}$, corresponds to comparison of $\boldsymbol{j}$-th and $\boldsymbol{k}$-th samples in the sample groups $\boldsymbol{SG_1}$ and $\boldsymbol{SG_2}$ such that $\mathbf{l}=\boldsymbol{j}-\mathbf{1}\times\boldsymbol{C_2}+\boldsymbol{k}$, where $\mathbf{1}\leq \boldsymbol{j}\boldsymbol{C_1}$ and $\mathbf{1}\leq \boldsymbol{k}\leq \boldsymbol{C_2}$. Thus, the cardinality of $\boldsymbol{O}$ is $\boldsymbol{C_1}\times\boldsymbol{C_2}$. Each $\boldsymbol{o_l}$ is a vector with dimension equaling the number of nodes in the pathway, $\boldsymbol{N}$, such that the $\boldsymbol{i}$-th element of $\boldsymbol{o_l}$, $\boldsymbol{o_{li}}$, equals $\boldsymbol{g_{i2k}}/\boldsymbol{g_{i1j}}$, where $\boldsymbol{j}$ and $\boldsymbol{k}$ are related to $\boldsymbol{l}$ as described above. The data matrix $\boldsymbol{D}$, with elements $\boldsymbol{d_{li}}$ is obtained from $\boldsymbol{O}$ such that $\boldsymbol{d_{li}}$ equals 1 if $\boldsymbol{o_{li}}<0.5$ or $\boldsymbol{o_{li}}>2$ (i.e. a gene is dysregulated) and 2 otherwise. To this end, we have converted each pathway into a BN, where nodes of the BN represent nodes in the pathway and node random variables are identified by discretized FC values. Nodes of BNs are assumed to follow Dirichlet distribution and are initialized using the equivalent sample size method for prior beliefs (Neapolitan, 2004). The matrix $\boldsymbol{D}$, which consists of $\boldsymbol{N}$ columns and $\boldsymbol{C_1}\times\boldsymbol{C_2}$ rows, is sequentially evaluated row by row, where each row is used to update the Dirichlet distribution parameters used at the nodes of BN. Upon conclusion of evaluation of $\boldsymbol{D}$, the score for the BN is calculated as described below.

2.4 Bayesian score metric

Following discretization, the nodes in the BN model represent discrete random variables with a multinomial distribution. Dirichlet distribution is chosen as the conjugate prior to the multinomial distribution. For a given BN model, the probability of observing data is (Neapolitan, 2004):

$$P(\text{Data}|\text{Model})=\prod_{i=1}^{N}\prod_{j=1}^{q_i}\frac{\Gamma(N_{ij})}{\Gamma(N_{ij}+M_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(a_{ijk}+s_{ijk})}{\Gamma(a_{ijk})}$$

where $\boldsymbol{N}$ is the number of nodes, $\boldsymbol{q_i}$ is the number of different states of node's parents and $\boldsymbol{r_i}$ is the set of values a node can take on. $\boldsymbol{N_{ij}}$ is the sum of corresponding Dirichlet distribution hyper-parameters $\boldsymbol{a_{ijk}}$. $\boldsymbol{M_{ij}}$ is the number of times that the parents of node $\boldsymbol{i}$ take on configuration $\boldsymbol{j}$ in the dataset. Of these $\boldsymbol{M_{ij}}$ cases, $\boldsymbol{s_{ijk}}$ is the total number of times in the sample that node $\boldsymbol{i}$ is observed to have value $\boldsymbol{k}$ when its parents take on configuration $\boldsymbol{j}$. The equation above is used as a score metric and named Bayesian scoring criterion (BSC) (Heckerman *et al.*, 1995). Hyper-parameters, $\boldsymbol{a_{ijk}}$, can be determined using the equivalent sample size method (i.e. sum of the initial Dirichlet parameters, used at each node, have the same total), in which case the score is called the BDe (Neapolitan, 2004). Details of the scoring scheme are found in the Supplementary Material.

2.5 Estimation of score significance by randomization via bootstrapping

At this point, we have BNs converted from pathways, a data matrix $\boldsymbol{D}$ for each BN representing observed, discretized FC values for genes (nodes) represented in the BN and BDe score calculated for the BN using the observed $\boldsymbol{D}$ matrix. We assess statistical significance of the BDe score, $\boldsymbol{S_n}$, calculated for n-th BN by using randomization via bootstrapping. We use a data-generating process and estimate a distribution of the score $\boldsymbol{S}$ given the null hypothesis that scores are the result of pure chance. For a one-tailed test with a rejection region in the upper tail, the bootstrap p-value for $\boldsymbol{S_n}$, $P(S_n)$, is estimated by the proportion of randomized samples that yield a score $> \boldsymbol{S_n}$. If we have $\boldsymbol{B}$ randomized datasets, then

$$P(S_n)=\frac{1}{B}\sum_{k=1}^{B}I(S_k>S_n)$$

where $\boldsymbol{I}$ is the indicator function yielding $\mathbf{1}$ if the Bayesian score is better than the original network score and $\mathbf{0}$ otherwise and $\boldsymbol{S_k}$ is the score of the BN using $\boldsymbol{k}$-th randomized dataset. As $\boldsymbol{B}$ goes to infinity, the estimated p-value will tend to the ideal p-value and the error in estimation will be kept minimal (Davison and Hinkley, 1997; Efron and Tibshirani, 1993).

The process for generating the randomized samples is as follows: suppose dataset $\boldsymbol{D}$ is composed of $\boldsymbol{M}$ cases for a total of $\boldsymbol{N}$ genes and can be considered as an $\boldsymbol{M}\times\boldsymbol{N}$ matrix where $\boldsymbol{l}$-th row $\boldsymbol{d_l}=[\boldsymbol{d_{l1}}, \boldsymbol{d_{l2}}, \ldots, \boldsymbol{d_{lN}}]$, $\mathbf{1}\leq\mathbf{l}\leq\boldsymbol{M}$ and $\boldsymbol{d_{li}}$ is the value of the $\boldsymbol{i}$-th node (gene) in the $\boldsymbol{l}$-th instance of input data. For each node $\boldsymbol{X_i}$, we sample with replacement $\boldsymbol{M}$ instances from the $\boldsymbol{i}$-th column, $[\boldsymbol{d_{1i}}, \boldsymbol{d_{2i}}, \ldots, \boldsymbol{d_{Mi}}]^T$, of the original data matrix $\boldsymbol{D}$ and obtain the newly formed column of the bootstrapped data matrix $\boldsymbol{D_k}$. The BDe score for this new data matrix is calculated and the whole process is repeated $\boldsymbol{B}$ times. The approach we adopt here has previously been described and applied to phylogeny reconstruction using molecular sequences (Brown, 1994; Davison and Hinkley, 1997). Bootstrap alone, which is generally used to establish confidence, would not be fitting to assess significance in the current setting. Therefore, we provide randomization via bootstrapping, which provides an approximation of the null distribution. When scoring a BN, the rows of $\boldsymbol{D}$, which hold information reflecting the dependency relation between nodes of BN, are considered sequentially in order to update the parameters of each node on the BN. We randomize rows of $\boldsymbol{D}$ by changing the structure of columns of $\boldsymbol{D}$ via sampling with replacing *each column of* $\boldsymbol{D}$ *separately*. Querying each pathway database that holds few hundreds of networks generates a multiple hypothesis testing problem (utilized KEGG database contributes over 200 pathways). We address this issue by calculating FDR, using Benjamini–Hochberg procedure, applied on p-values calculated for each pathway (Benjamini and Hochberg, 1995).

The overall complexity of the BPA is $O(N^2+E^2+BC^2G)$, where N is the number of nodes, E is the number of edges, G is the number of genes, C is the number of samples and B is the number of bootstrap datasets. Modeling of pathways as DAGs is quadratic in N and E for Spirtes' algorithm and linear in N and E for Tarjan's algorithm. Data discretization is quadratic in C and linear in G. BDE score calculation is $O(BN^2G)$. Modeling pathways as DAGs is done off line, which does not lead to the computational time of a single analysis.

3 RESULTS

3.1 Identification of data fitting to network

We tested our method to assign significance to BDe scores on eight synthetic binary BNs of different sizes and the well-known Alarm and Asia BNs. The Alarm BN is designed to identify anesthesia problems with 37 nodes of 2, 3 or 4 states (Beinlich *et al.*, 1989) and the eight-node binary Asia BN is designed to calculate the probability of a patient having tuberculosis (Lauritzen and Spiegelhalter, 1988). All 10 networks

Table 1. Scores and *p*-values of scores for synthetic, Alarm and Asia BNs

BN name	No. of nodes	Data following CPT		Data inconsistent with CPT	
		Score	*p*-value	Score	*p*-value
Alarm	37	−9955	$<5\times10^{-4}$	−22600	0.56
Asia	8	−2221	$<5\times10^{-4}$	−2926	0.54
BN1	19	−9344	$<5\times10^{-4}$	−10213	0.62
BN2	8	−3569	$<5\times10^{-4}$	−3874	0.54
BN3	21	−10844	$<5\times10^{-4}$	−12763	0.55
BN4	36	−20074	$<5\times10^{-4}$	−21746	0.59
BN5	18	−9607	$<5\times10-^{4}$	−10245	0.50
BN6	29	−15859	$<5\times10^{-4}$	−17122	0.64
BN7	19	−9804	$<5\times10^{-4}$	−10996	0.65
BN8	53	−29937	$<5\times10^{-4}$	−32262	0.67

along with their CPTs can be found in the Supplementary Material. Synthetic BNs have (average ± SD) 25.38 ± 13.87 nodes, 24.38 ± 13.87 links and 1.93 ± 0.08 average degree reflecting a spectrum of typical biological networks (see Supplementary Material). For each BN, we used two datasets to calculate BDe scores and their significance: one that follows the underlying CPT and one that does not. For a given BN, we randomly fixed Dirichlet hyper-parameters for each node and generated data that follow this CPT using the Bayes Net Toolbox (BNT) for Matlab (http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html). CPTs calculated from this data are considered ideal as they are based on data following the fixed CPT for a given BN. For inconsistent CPTs, we chose Dirichlet hyper-parameters to be equal for each node and obtained data using BNT. Therefore, the underlying CPTs calculated from this randomly generated data are considered to be the non-ideal CPT as they are not based on data reflecting dependency structure implied by the given BN. We used datasets with a size of 1000, the bootstrap test count was chosen to be 2000 and an equivalent sample size of 1 for Dirichlet hyper-parameters was used during BDe score calculation. The results summarized in Table 1 show that when the data following underlying CPTs are used, *p*-values indicate strong significance and are very close to zero ($P < 5\times10^{-4}$). Conversely, when data generated using CPTs not consistent with independencies entailed by the BNs are used, *p*-values are severely deteriorated. Therefore, datasets produced from inconsistent CPTs are quickly detected by BPA.

3.2 Sample size

In real life microarray experiments, the number of samples rarely exceeds 100 due to technical and financial limitations rendering limited number of observations to assess change in expression of a given gene. In order to see the effect of this limitation on BPA, we tested BNs in Table 1 with varying sizes of datasets (using CPTs that follow the underlying BN structure: dataset size 20–200, and CPTs that are non-ideal for the underlying BN structure: dataset size 20–300) and calculated the significance of corresponding BDe scores. For each dataset size, 50 runs have been performed. The average *p*-value of the runs, each obtained using 1000 bootstrapped samples, and associated standard errors are shown in Figure 2A and B. In case of ideal CPTs, *p*-values start to get lower after a small increase in the sample size with highest attainable significance ($P < 10^{-3}$) at sizes >140. This is a dataset size that can be generated by the proposed method in an experimental setting, where one has 12 samples in each of the two groups [BPA would generate $12\times12=144$ observations for each BN (see Section 2.3)]. In case of non-ideal CPTs, *p*-values remain high regardless of the dataset size (see also Table 1). These results suggest that BPA can successfully be used with datasets commonly seen in real experimental settings.

3.3 Change in pathway structure

Biological pathways may be incomplete as some of the nodes and/or edges for a given cascade of events may not have been identified yet. In order to test for the effect of missing edges and nodes on the significance of BDe scores calculated by BPA, we systematically removed all possible $\boldsymbol{k}$ edge combinations, $\mathbf{1 \leq k \leq 5}$, for BNs listed in Table 1. For each $\boldsymbol{k}$, we calculated the average *p*-value obtained by removing different combinations using a dataset size of 140 (following the underlying CPT) with 1000 bootstraps. We repeated the same procedure by removing all combinations of nodes. Results are shown in Figure 2C and D. In both cases, all BNs maintain significant results despite removal of up to five edges and/or nodes, except for BN2 (eight nodes and seven edges), BN5 (18 nodes and 17 edges) and Asia BN (eight nodes and eight edges), which have smallest number of nodes and edges among the 10 synthetic networks. Note that the effect of node removal is more severe as when a node is removed so are all the edges connected to it. Robustness to node/edge removal is possibly due to the factorized scoring metric and the BNs ability to focus on local interactions, where each node is directly affected by a relatively small number of parent nodes and interactions. The synthetic networks tested follow average node/edge distribution in typical biological pathways and the removal of up to five nodes/edges is likely to be very high compared to the pathway error instances seen in real biological pathways, which makes application of BPA for pathway analysis possible.

3.4 Application to synthetic datasets

We first looked at whether our method of solving cyclicity could create large cliques and inversely affect the BPA's overall performance. We generated 20 synthetic directed graphs containing cyclic paths following an SEM (Bollen, 1989), using TETRAD IV (http://www.phil.cmu.edu/projects/tetrad/tetrad4.html) (Scheines, 1996). The corresponding acyclic collapsed versions of synthetic cyclic graphs show ~2.7–2.8 times the increase in number of nodes, number of edges, maximum degree, average degree and density of the networks, on average. The BPA was run on a 1000 size dataset with 1000 bootstraps and resulted in significant *p*-values (lowest attainable) in all cases when we generated data that follow SEMs. These results suggest that our method of handling cyclicity may generate large cliques; however, the BPA is not adversely affected by the proposed method of generating DAGs from biological pathways (for details, see Supplementary Material).

We then compared performance of the BPA with GSEA v2 (Subramanian *et al.*, 2007) and a model-based approach, GlobalTest (Goeman *et al.*, 2004) using Bioconductor v2.7, GlobalTest v5.4.0 package on synthetic datasets approximating real microarray data. We generated synthetic transcriptional regulatory networks and produced simulated gene expression data with noise using SynTReN v1.12 (Van den Bulcke *et al.*, 2006). We created 60 synthetic

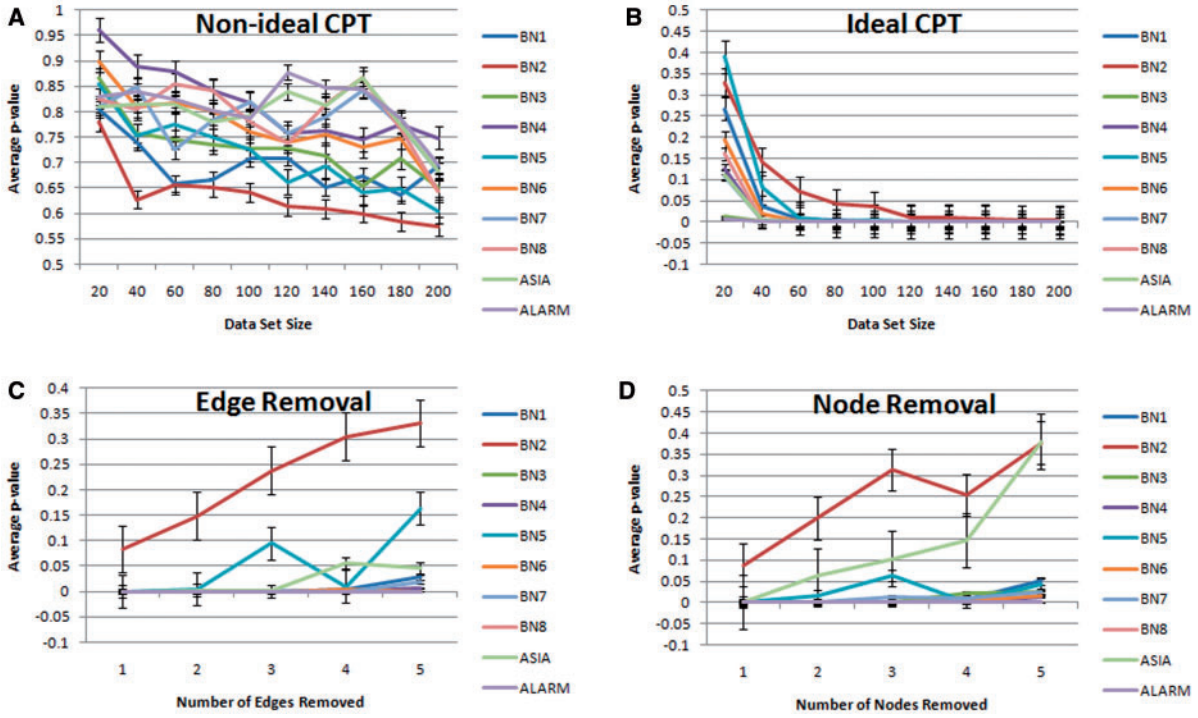

Fig. 2. BPA performance for BNs listed in Table 1: (**A**) data follow underlying CPTs; (**B**) data do not follow underlying CPTs; (**C**) progressive removal of edges in BNs; and (**D**) progressive removal of nodes in BNs. In each case, average *p*-values of 50 runs have been calculated. Dataset sizes are 20–200 in (A), 20–300 in (B) and 140 in (C) and (D) to depict better resolution, plateau and real-life settings, respectively.

Table 2. Average ± SD of fraction of pathways accurately called active or inactive by BPA, GSEA and GlobalTest (GT)

BPA 2 level		BPA 3 level		GSEA	GT
FC 2	FC 3	FC 2	FC 3		
0.825 ± 0.047	0.838 ± 0.042	0.825 ± 0.047	0.838 ± 0.042	0.583 ± 0.001	0.400 ± 0.024

networks (58/60 have cycles) with sizes ranging from 2 to 200. Details of the networks' parameters are included in the Supplementary Material. We randomly selected 25 out of 60 pathways to be active and SynTReN generated corresponding expression datasets for 20 test and 20 normal samples with 2249 synthetic genes adding a 4% noise level. For all three methods, we used 1000 bootstraps and chose a nominal *p*-value and FDR cutoff values of 0.05 and 0.25, respectively. We assessed accuracy (if a network—or corresponding gene set—is correctly called active/inactive) of the three algorithms for 10 simulated datasets and provide the results in Table 2.

BPA was tested for FC cutoff values (CO) of 2 and 3, and discretization levels of 2 (i.e. if a gene's FC is above CO or below 1/CO, i.e. a gene is dysregulated, we insert a 1 in the observation data matrix D, otherwise we insert a 2) and 3 (i.e. we insert a 1, 2 or 3 in the observation data matrix D, if a gene's FC is above CO, below 1/CO, between CO and 1/CO, inclusive, respectively). These results suggest that BPA outperforms both GSEA and GlobalTest and there is no significant change in using two or three levels of discretization in BPA with a slight improvement in performance when a FC cutoff of 3 is used. We used two-level discretization with an FC cutoff of 2 when applying BPA to real datasets. A two-level discretization seems more natural as we do not keep activator/repressor information when modeling biological pathways as DAGs and an FC cutoff of 2 allows BPA to capture subtle changes.

3.5 Application to real RCC dataset

We applied BPA on real RCC datasets in order to identify the underlying molecular mechanisms of the disease (Table 3). In each experiment, every cancer subtype was individually compared to the normal samples generating 16 datasets in total. BNs corresponding to biological pathways are scored using BDe with an equivalent sample size of 1 for Dirichlet hyper-parameters and a selected subset of those that remain significant after 1000 bootstraps are shown in Table 4. We list CPU times for BPA in the Supplementary Material, which is 30 ± 10 min (average ± SD) for the 16 analyzed datasets, where the time given is for the complete analysis of a single dataset. Running times range from 17 to 57 min. Analysis was performed on an Intel Core 2 Duo CPU E6550 2.33 GHz processor with Windows XP 32bit OS. We also analyzed the 16 datasets using GSEA and GlobalTest. In all three methods, the *p*-value and FDR cutoff values were chosen to be 0.05 and 0.25, respectively. In order to avoid the problem of pathway alignment that would arise if multiple pathway sources were used, we limited our analysis to the KEGG pathway database for this exemplary case. In case of GSEA and GlobalTest,

we used MSigDB v2.5 (group CP under C2). The complete list of significant pathways for each method and a comparative analysis are included in the Supplementary Material.

Most of the pathways deemed significant by BPA agree with those found in the literature, using genomic and proteomic approaches. For example, arginine and proline metabolism, citrate cycle [tricarboxylic acid (TCA) cycle], purine metabolism, fatty acid metabolism, pyruvate metabolism, glycolysis/gluconeogenesis, valine, leucine and isoleucine degradation pathways have been shown to be important in RCC analyzed using a proteomic approach (Perroud *et al.*, 2006). On the other hand, significant pathways found in different subtypes show notable agreement among datasets analyzed. Using BPA, we found 25 pathways significant in at least half of the datasets; this number was only nine for GSEA (see Supplementary Material). In BPA, on an average, 10.6 datasets were found significant by each of the 25 pathways (for a total dataset occurrence of 265), while GSEA's average was 9.3 significant datasets per pathway (for a total dataset occurrence of 84). When we considered pathways deemed significant for at least one dataset, we found 129 pathways discovered by BPA yielding 571 dataset occurrences in total and 121 pathways discovered by GSEA resulting in 390 dataset occurrences. BPA was able to find ∼63% of pathways discovered by GSEA. Overall, these results indicate that BPA found a greater pathway base related and specific to RCC as compared to GSEA. We believe this enhancement in performance is due to the ability of BPA to take into account connectedness of genes that make up a pathway; whereas in the GSEA analysis, such genes are only considered as a list and no topological information is incorporated into the analysis. GlobalTest results indicated 199 significant pathways (out of 206) for 2974 dataset occurrences yielding ∼14.95 average datasets (out of 16) deemed significant

Table 3. Datasets used in BPA analysis of malignancies in kidney

Dataset name	Number and types of samples	GEO no.
Lenburg *et al.*	17 (8 N, 9 cRCC)	GSE 781
Jones *et al.*	92 (23 N, 32 cRCC, 11 pRCC, 6 chRCC, 12 OC, 8 TCC)	GSE 15641
Furge *et al.* and Yang *et al.*	47 (12 N, 35 pRCC)	GSE 7023 and GSE 2748
Gumz *et al.*	20 (10 N, 10 cRCC)	GSE 6344
Kort *et al.*	79 (12 N, 10 cRCC, 17 pRCC, 6 chRCC, 7 OC, 27 WT)	GSE 11024
Koeman *et al.*	32 (12 N, 10 chRCC, 10 OC)	GSE 8271
Wang *et al.*	22 (12 N, 10 cRCC)	GSE 14762

Table 4. Selected significantly regulated pathways (p-value < 0.05, FDR < 0.25)

Pathway/dataset	Lenburg	Jones	Yang	Gumz	Kort	Koeman	Wang
Alanine and aspartate metabolism	c^{a}	$c^{a,b}$ p^{b} ch^{b} O^{b} T^{a}	p^{b}	c^{b}	W^{b}		$c^{a,b}$
Arachidonic acid metabolism	c^{a}	c^{b} p^{b} $ch^{a,b}$ O^{b} T^{b}	p^{b}		O^{b} $W^{a,b}$	O^{b}	c^{b}
Arginine and proline metabolism	$c^{a,b}$	c^{b} $p^{a,b}$ ch^{b} O^{b} $T^{a,b}$	p^{b}	c^{b}	p^{b} W^{a}		$c^{a,b}$
Cell cycle		$c^{a,b}$ p^{b} ch^{b} O^{b} $T^{a,b}$	p^{b}		p^{b} $W^{a,b}$	O^{b}	$c^{a,b}$
Citrate cycle (TCA cycle)	$c^{a,b}$	$c^{a,b}$ ch^{b} O^{b} T^{b}		c^{b}	O^{b} W^{b}	O^{b}	c^{b}
Drug metabolism—cytochrome P450		c^{b} p^{b} ch^{b} O^{b} T^{b}			W^{b}		
ECMreceptor interaction		c^{b} p^{b} ch^{b} T^{b}	p^{b}		p^{b} W^{b}		c^{b}
Fatty acid metabolism	$c^{a,b}$	$c^{a,b}$ $p^{a,b}$ ch^{b} O^{b} $T^{a,b}$	$p^{a,b}$	$c^{a,b}$	$c^{b}p^{a,b}O^{b}$ W^{b}	ch^{b} O^{b}	$c^{a,b}$
Focal adhesion		c^{b} p^{b} ch^{b} O^{b} T^{b}	p^{b}		p^{b} W^{b}	O^{b}	c^{b}
Galactose metabolism		c^{b} ch^{b}			W^{b}		c^{b}
Glutamate metabolism	c^{a}	$c^{a,b}$ $p^{a,b}ch^{b}$ O^{b} T^{b}	p^{b}		p^{b} W^{b}	ch^{b}	c^{b}
Glycolysis/gluconeogenesis	$c^{a,b}$	$c^{a,b}$ p^{b} ch^{b} O^{b} T^{b}	p^{b}	$c^{a,b}$	c^{b} $p^{b}O^{b}W^{a,b}$	ch^{b} O^{b}	c^{b}
Glycosphingolipid biosynthesis		c^{b} ch^{b} O^{b} T^{b}	p^{b}		W^{b}	O^{b}	c^{b}
Inositol phosphate metabolism		$c^{a,b}$ $p^{a,b}ch^{b}$ O^{b} T^{b}					c^{a}
Insulin signaling pathway		c^{b} $p^{a,b}$ ch^{b} O^{b} T^{b}			W^{b}		c^{b}
MAPK signaling pathway		c^{b} $p^{b,a}$ ch^{b} T^{b}			W^{b}		c^{b}
Metabolism of xenobiotics by cyt. P450	c^{a}	c^{a} p^{b} $ch^{a,b}$ O^{a}		c^{a}	c^{b} O^{a} $W^{a,b}$	O^{a}	$c^{a,b}$
Natural killer cell mediated cytotoxicity		$c^{a,b}$ p^{b} ch^{b} T^{b}	p^{b}		p^{b} W^{b}		$c^{a,b}$
Nicotinate and nicotinamide metabolism	$c^{a,b}$	c^{b} p^{b} $ch^{a,b}O^{b}$ T^{b}		c^{b}	$W^{a,b}$	ch^{b}	c^{b}
Nitrogen metabolism	c^{a}	$c^{a,b}$ $p^{a,b}$ ch^{b} O^{b} $T^{a,b}$	p^{b}	c^{a}	p^{b} $W^{a,b}$		$c^{a,b}$
One carbon pool by folate	$c^{a,b}$	c^{b} p^{b} ch^{b} $O^{a,b}$ T^{b}	p^{b}	c^{b}	p^{b} W^{b}		c^{b}
p53 Signaling pathway		c^{a} ch^{b}	p^{a}		W^{a}		c^{a}
Pentose phosphate pathway		c^{b} p^{b} ch^{b} O^{b} T^{b}	p^{b}		p^{b} W^{b}		c^{b}
Propanoate metabolism	c^{a}	$c^{a,b}$ p^{a} ch^{b} T^{a}		c^{a}	$W^{a,b}$		c^{a}
Purine metabolism	c^{b}	c^{b} p^{b} ch^{b} O^{b} T^{b}	p^{b}	c^{b}	$c^{b}p^{b}ch^{b}O^{b}W^{b}$	ch^{b} O^{b}	c^{b}
Pyruvate metabolism	$c^{a,b}$	$c^{a,b}$ ch^{b} O^{b} $T^{a,b}$	p^{b}		ch^{b} O^{b} $W^{a,b}$	ch^{b} O^{b}	c^{b}
Pyrimidine metabolism		c^{b} p^{b} ch^{b} O^{b} T^{b}	p^{b}		p^{b} $W^{a,b}$		c^{b}
Retinol metabolism	c^{b}	c^{b} p^{b} ch^{b} O^{b} T^{b}	p^{b}		p^{b} $W^{a,b}$	ch^{b} O^{b}	c^{b}
Urea cycle metabolism of amino groups	$c^{a,b}$	$c^{a,b}$ $p^{a}ch^{a,b}$ O^{a} $T^{a,b}$	p^{a}		p^{a} ch^{a} W^{b}	O^{a}	c^{a}
Valine leucine and isoleucine degr.	$c^{a,b}$	$c^{a,b}$ $p^{a,b}$ ch^{b} O^{b} $T^{a,b}$	p^{b}	$c^{a,b}$	$c^{b}p^{b}ch^{b}O^{b}W^{a,b}$	ch^{b} O^{b}	$c^{a,b}$

Boldface pathways are shown to be important in RCC using an experimental proteomic approach. [a]GSEA; [b]BPA. c: cRCC; p: pRCC; ch: chRCC; O: OC; T: TCC; and W: WT.

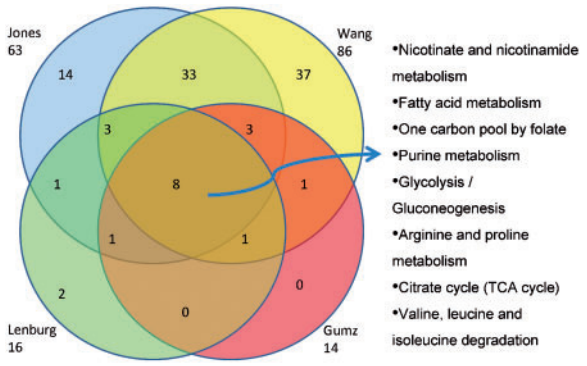

Fig. 3. Venn diagram depicting pathways shared by BPA analysis of Jones, Lenburg, Gumz and Wang cRCC datasets. Eight pathways at the intersection of all four analyses are indicated.

for each pathway. We include complete results of GlobalTest in Supplementary Material as pathway selection with this method showed little specificity (∼97% of tested pathways found significant for ∼90% of the datasets). Similar behavior for GlobalTest have been observed previously by other studies potentially due to distributional assumptions (that regression coefficients for the genes come from the same normal distribution) and errors in empirical covariance estimates, made by this approach, leading to high false positive rates, especially in cases with small sample sizes (Gatti *et al.*, 2010; Liu *et al.*, 2007). Furthermore, GlobalTest loses significant power when a given gene list contains correlated genes, which holds true for genes in a given pathway (Mansmann and Meister, 2005).

Out of 12 pathways shown in Table 4, which were shown to be related to RCC using an experimental proteomic approach, BPA found 117 dataset occurrences for which these pathways were significant (∼61% of possible $12 \times 16 = 192$ dataset occurrences), while GSEA could only identify 50 (∼26%) dataset occurrences.

BPA was also able to yield high consensus among pathways found significant for a given RCC subtype. In Figure 3, we show the overlap of pathways found significant for cRCC subtype in four different datasets. More than 70% of pathways found in four datasets are shared by at least two datasets, while eight pathways were common to all. Among these eight pathways, six of them (except for nicotine and folate pathways) have been shown to be activated in RCC based on a proteomic approach (Perroud *et al.*, 2006).

Results summarized in Table 4 put forth molecular mechanisms that are not only subtype specific, but also commonly seen in different RCC tumors. The complexity/relevance of some of these pathways can be exemplified by the activation of the insulin signaling pathway through the activation of the insulin growth factor receptor-1 that activates the PI3K/Akt signaling pathway. PI3Ks catalyze the conversion of phosphatidylinositol bisphosphate (PIP) 2 to PIP 3 (inositol phosphate metabolism). PIP 3 acts as a second messenger to activate Akt. Akt mediates the activation of mTOR that is responsible for its effects on cell growth. In addition to activating the PI3K/Akt/mTOR pathway, IGFR-1 also activates the Ras/MAPK/Raf/MEK/ERK mitogenic signaling pathway (MAPK signaling pathway). Subsequently, this leads to an activation of the cell cycle transition through stimulation of cyclin D1 (Cell cycle) and to increased cell proliferation. These prominent cellular pathways have been the objective for most of the targeted therapies now used in metastatic RCC. mTOR inhibitors act on PI3K/Akt/mTOR pathway leading to an inhibition of protein synthesis and cell cycle arrest, while some receptor tyrosine kinase inhibitors (Sorafenib) also affect Raf, blocking the MAPK mitogenic signaling pathway. In addition, both approaches inhibit angiogenesis that has a strong impact in RCC tumorigenesis and progression. Interestingly, further studies are underway analyzing possible composite or sequential therapies blocking these pathways for the identification of the optimal therapeutic approach. However, different targets within other pathways, described here, may lead to additional successful results and should be explored further. According to our analysis, one of these novel pathways could be the glyoxylate and dicarboxylate metabolism, which has been associated with lung cancer, but not with RCC (Creighton *et al.*, 2003). Indeed, metabolism-related pathways have been shown to play a role in RCC progression and would therefore be reasonable targets for further in-depth analysis and *in vitro* testing.

4 DISCUSSION

We described a method that models biological pathways as BNs and determines the fitness of given microarray data using BDe score. The proposed method overcomes representation, mapping, data discretization and cyclicity problems that arise in modeling pathways as BNs. We have chosen multinomial BNs with Dirichlet priors because: (i) their posterior can be efficiently calculated in closed form; (ii) they capture non-linear interactions; and (iii) they render a plain model requiring less parameter adjustment. Moreover, algorithms scoring mutinomial BNs have low time complexity. Alternative models such as linear Gaussian models, Gaussian process networks or regression models are usually preferred in the task of structure learning. Linear Gaussian models and regression models in BN setting can only detect linear dependencies between the child and parent variables. In Gaussian process BN models, Gaussian process priors are used as parametric families to model non-linear relations. However, the problem then becomes one of selecting the best fitting covariance function and the number of its hyperparameters in Gaussian process modeling, which requires various approximations and assumptions that may not be suitable in HTBD settings (Friedman and Nachman, 2000).

RCC represents a spectrum of genetically diverse epithelial tumors with a common derivation from the renal tubular epithelium and a variable clinical course. Approximately 30% of cases present with metastatic disease at initial diagnosis and 30% of initially organ confined cases develop metastases during later follow up. Since there are no reliable biomarkers available, patient management remains problematic despite improving understanding of the underlying molecular mechanisms. In particular, treatment of advanced RCC still poses a great challenge, as the RCC is resistant to chemo- and radiation therapy and cytokine-based therapies offer only low clinical response rates with considerable toxicity. The advent of targeted therapy has brought exciting therapeutic options with promising clinical results, although the clinical benefit with respect to overall survival is only marginal. However, high-throughput technologies that analyze the entire genome and proteome promise to elucidate the heterogeneity of this disease and eventually enable a patient-tailored, individualized treatment. In contrast to the analysis of single genes, gene pathways enable us to see the context of complex interactions and to understand the biologic relevance

of their expression. The plethora of pathways presented in our manuscript mirror complex biologic processes in kidney tumors and is often closely intertwined.

Overall, we believe the proposed approach provides a unique perspective that merges BN theory and HTBD analysis. Most BN models employed on HTBD use time series experimental designs in order to increase the size of the observed data. We have overcome this bottleneck and provided a tool that can be used with most common experimental settings interpreting the results within the context of known biological pathways. Moreover, existing BN approaches on HTBD generally focus on building networks from input data, which makes these approaches applicable on a few dozens of genes due to the complexity of structure learning algorithms. Given the fact that high-throughput platforms generate data for tens of thousands of genes, the proposed approach makes use of relevant experimental information and is applied to the complete dataset within the context of known biological pathways. Our simulations on synthetic and real datasets show that BPA is able to successfully find molecular mechanisms that best describe underlying HTBD.

ACKNOWLEDGEMENT

We thank Dr Khalid Sayood for critical review of the manuscript.

Funding: Grant from The Dubai Harvard Foundation for Medical Research (to H.H.O., in part).

Conflict of Interest: none declared.

REFERENCES

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Bauer,S. *et al.* (2010) GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.*, **38**, 3523–3532.

Beinlich,I.A. *et al.* (1989) The ALARM monitoring system: a case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, Springer, Berlin, pp. 247–256.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B*, **57**, 289–300.

Bolen,K. (1989) *Structural Equations with Latent Variables*. John Wiley & Sons, New York, pp. 80–117.

Brown,J.K. (1994) Bootstrap hypothesis tests for evolutionary trees and other dendrograms. *Proc. Natl Acad. Sci. USA*, **91**, 12293–12297.

Brugarolas,J. (2007) Renal-cell carcinoma–molecular pathways and therapies. *N. Engl. J. Med.*, **356**, 185–187.

Creighton,C. *et al.* (2003) Gene expression patterns define pathways correlated with loss of differentiation in lung adenocarcinomas. *FEBS Lett.*, **540**, 167–170.

Davison,A.C. and Hinkley,D.V. (1997) *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, UK.

Efron,B. and Tibshirani,R. (1993) *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability 57*. Chapman & Hall, New York.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Friedman,N. and Nachman,I. (2000) Gaussian process networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI-00)*. Morgan Kaufmann, Stanford, CA, pp. 211–219.

Furge,K.A. *et al.* (2007) Detection of DNA copy number changes and oncogenic signaling abnormalities from gene expression data reveals MYC activation in high-grade papillary renal cell carcinoma. *Cancer Res.*, **67**, 3171–3176.

Gatti,D.M. *et al.* (2010) Heading down the wrong pathway: on the influence of correlation within gene sets, *BMC Genomics*, **11**, 574.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Gumz,M.L. *et al.* (2007) Secreted frizzled-related protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma. *Clin. Cancer Res.*, **13**, 4740–4749.

Heckerman,D. *et al.* (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.*, **20**, 197–243.

Hoaglin,D.C. *et al.* (2000) *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York, pp. 339–400.

Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Imoto,S. *et al.* (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.

Jones,J. *et al.* (2005) Gene signatures of progression and metastasis in renal cell cancer. *Clin. Cancer Res.*, **11**, 5730–5739.

Jones,J. *et al.* (2008) Proteomic identification of interleukin-2 therapy response in metastatic renal cell cancer. *J. Urol.*, **179**, 730–736.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Koeman,J.M. *et al.* (2008) Somatic pairing of chromosome 19 in renal oncocytoma is associated with deregulated EGLN2-mediated [corrected] oxygen-sensing response. *PLoS Genet.*, **4**, e1000176.

Kort,E.J. *et al.* (2008) The E2F3-Oncomir-1 axis is activated in Wilms' tumor. *Cancer Res.*, **68**, 4034–4038.

Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application on expert systems. *J. R. Stat. Soc.*, **50**, 157–224.

Lenburg,M.E. *et al.* (2003) Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data. *BMC Cancer*, **3**, 31.

Liu,Q. *et al.* (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

Lu,Y. *et al.* (2008) A probabilistic generative model for GO enrichment analysis. *Nucleic Acids Res.*, **36**, e109.

Mansmann,U. and Meister,R. (2005) Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.*, **44**, 449–453.

Mills,E.J. *et al.* (2009) Metastatic renal cell cancer treatments: an indirect comparison meta-analysis. *BMC Cancer*, **9**, 34.

Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.

Neapolitan,R.E. (2004) *Learning Bayesian Networks*. Prentice Hall, NJ, USA.

Pearl,J. (2000) *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, UK.

Perroud,B. *et al.* (2006) Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Mol. Cancer*, **5**, 64.

Romero,P. *et al.* (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.

Schaefer,C.F. *et al.* (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.

Scheines,R. (1998) The TETRAD project: constraint based aids to causal model specification. *Multivariate Behavioral Res.*, **33**, 65–117.

Spirtes,P. (1995) Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*. Morgan Kaufmann, Montreal, Quebec, Canada, pp. 491–549.

Subramanian,A. *et al.* (2007) GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.

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.

Tarjan,R. (1972) Depth-first search and linear graph algorithms. *SIAM J. Comput.*, **1**, 146–160.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

Vastrik,I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.

Wang,Y. *et al.* (2009) Regulation of endocytosis via the oxygen-sensing pathway, *Nat. Med.*, **15**, 319–324.

Yang,X.J. *et al.* (2005) A molecular classification of papillary renal cell carcinoma. *Cancer Res.*, **65**, 5628–5637.