10

Prediction of Novel Pathway Elements and Interactions Using Bayesian Networks

Andrew P. Hodges, Peter Woolf and Yongqun He ξ
University of Michigan, Ann Arbor, MI,
USA

1. Introduction

Signalling and regulatory pathways that guide gene expression have only been partially defined for most organisms. Given the increasing number of microarray measurements, it may be possible to reconstruct such pathways and uncover missing connections directly from experimental data. One major question in the area of microarray-based pathway analysis is the prediction of new elements to a particular pathway. Such prediction is possible by independently testing the effects of added genes or variables on the overall scores of the corresponding expanded networks. A general network expansion framework to predict new components of a pathway was suggested in 2001 (Tanay and Shamir, 2001). Many machine learning approaches for identifying hidden or unknown factors have appeared in the literature recently (Gat-Viks and Shamir, 2007; Hashimoto, et al., 2004; Herrgard, et al., 2003; Ihmels, et al., 2002; Needham, et al., 2009; Parikh, et al., 2010; Pena, et al., 2005; Tanay and Shamir, 2001; Yu and Li, 2005).

Compared to existing pathway expansion methods based on correlation, Boolean, or other strategies (Hashimoto, et al., 2004; Herrgard, et al., 2003; Ihmels, et al., 2002; Tanay and Shamir, 2001), Bayesian network-based expansion methods provide distinct advantages. A Bayesian network (BN) is a representation of a joint probability distribution over a set of random variables (Friedman, et al., 2000). Bayesian networks are able to identify causal or apparently causal relationships (Friedman, et al., 2000), and can be used to predict both linear and nonlinear functions. Furthermore, BN analysis is robust to error and noise and easily interpretable by humans. Bayesian network-based expansion has been used for gene expression data analysis (Gat-Viks and Shamir, 2007; Pena, et al.). We have recently developed an algorithm termed "BN+1" which implements Bayesian network expansion to predict new factors and interactions that participate in a specific pathway (Hodges, et al., 2010; Hodges, et al., 2010). This algorithm has been tested using *E. coli* microarray data (Hodges, et al., 2010) and verified with a synthetic network (Hodges, et al., 2010).

This Book Chapter aims to first provide a detailed review on different computational methods for pathway element prediction, introduce how a BN analysis is typically performed, and then describe how this BN+1 algorithm works. We will also introduce our MARIMBA software program (http://marimba.hegroup.org) which can implement the BN+1 algorithm along with many other useful features. So far, the success of BN+1 in new pathway element prediction has been demonstrated in prokaryotic *E. coli* system. This paper will introduce our new study of applying BN+1 to predict new pathway elements for

eukaryotic B-cell receptor (BCR) pathway using high throughput microarray data from perturbed B-cells obtained from the Alliance for Cellular Signalling (AfCS) (Zhu, et al., 2004). Finally, we will present current challenges and possible future directions in this field.

2. Overview of different computational methods for prediction of new pathway elements

In this section, we describe several existing methods for pathway expansion. By pathway expansion, we mean the expansion of a known set of variables with some biological role or function to include novel interacting or downstream variables. This definition is highly flexible and can be used for a variety of biological and biomedical situations.

2.1 Correlation methods and pathway expansion

Some of the most prevalent approaches used towards analyzing high-throughput datasets are correlation-based methods. Correlation methods attempt to identify the degree of similarity or dissimilarity between two or more variables (*e.g.*, the expression profiles of two genes) using simple computational distance metrics, such as Manhattan and Pearson metrics (Herrero, et al., 2001). An underlying assumption is that cellular processes often require the participation of multiple gene products which are expected to show correlated expression patterns as well as physical interactions (Meier and Gehring, 2008).

To predict new pathway elements using correlation methods, one or more genes (or other biological entities) are usually selected initially as a target of interest for comparison. A correlation is then determined between each other gene's (or entity's) expression pattern and that of the gene of interest. Those correlations appearing above some established threshold or ranking are then represented as either edges in a network or as a dendrogram in an expression-based heatmap diagram. For example, Herrgard et al defined subset of variables with specific modular behaviors and network structure using correlations and linear multiple regression (Herrgard, et al., 2003). These modules are then expanded to identify other neighboring variables with likely interactions or influences with the module-based sub-networks. Tanay et al (2001) introduced a fitness function-based approach for expanding sets of variables in literature models (Tanay and Shamir, 2001).

One advantage of these correlation-based methods is the ability to compute all pair-wise correlations for genes or features on a gene expression microarray or other high-throughput datasets. However, the correlation networks themselves do not imply any directionality for the interactions, such as which gene activates or represses a correlated gene, or whether those genes are instead co-regulated by another biological entity. The types and sometimes directionality of interactions must be determined using one or more analysis procedures, such as gene enrichment, promoter analysis, and context-dependent (or condition-dependent) analysis (Meier and Gehring, 2008). The correlation-based methods are often sensitive to the underlying distance metrics and assumptions, and are easily misinterpreted when the wrong metrics are employed. In addition, nonlinear (*e.g.* biphasic) interactions cannot usually be detected using correlation-based methods.

2.2 Clustering-based identification of new pathway elements

Various clustering method can be used to group genes based on expression values and identify potential new genes to specific pathways. Unsupervised and supervised clustering

methods have been developed (Raychaudhuri, et al., 2001). Unsupervised clusting methods, such as hierarchical clustering (Eisen, et al., 1998), self-organizing maps (Tamayo, et al., 1999), and model-based clustering (*e.g.*, CRCView (Xiang, et al., 2007)), arrange genes and samples in groups/clusters based solely on the similarities in gene expression. Supervised methods, including EASE (Hosack, et al., 2003) and gene set enrichment analysis (GSEA) (Subramanian, et al., 2005), use sample classifiers and gene expression to identify hypothesis-driven correlations. The Gene Ontology program (GO) is frequently used for gene enrichment analysis by many software programs, for example, DAVID (Huang da, et al., 2009) and GOStat (Beissbarth and Speed, 2004). One major disadvantage of such clustering-based methods on identifying new pathway elements is that detailed gene-gene interactions and directionalities cannot be predicted.

2.3 Boolean network-based pathway expansion

In Boolean network modelling, originally introduced by Kauffman (Kauffman, 1969) (Kauffman, 1969) (Kauffman, 1969), gene expression is quantized to only to two levels: ON and OFF. The gene expression level (state) of each gene is functionally related to the expression states of some other genes using logical rules. Probabilistic Boolean Networks (PBN) share the appealing rule-based properties of Boolean networks, but are robust in the face of uncertainty (Shmulevich, et al., 2002). Hashimoto et al. proposed a method to grow genetic regulatory networks from seed genes based on PBN analysis (Hashimoto, et al., 2004). In their study, Boolean functions were implemented towards globally expanding a set of seed genes from known literature-extracted interactions for vascular endothelial growth factor pathway genes using melanoma and glioma data (Hashimoto, et al., 2004). The output of this algorithm depends on the PBN-based objective function. The disadvantage of this approach is that the two-level representation in Boolean network often oversimplifies the complex biological systems.

2.4 Mutual information-based method

Mutual information-based methods have been used for modelling, refining, and expanding biological pathways. In probability theory and information theory, the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. Recent reports by Luo et al. (Luo, et al., 2008; Luo and Woolf, 2010; Watkinson, et al., 2009) and others have shown the utility and improved modelling of using three-way and higher mutual information influences for a given variable. However, the assembly of these multi-parent interactions into larger global networks is yet a challenging issue.

2.5 Bayesian network pathway refinement and expansion

Bayesian networks have recently been widely used for biological pathway reconstruction and expansion. Since this is the major topic of this book chapter, we will introduce it in more details in the next sections.

3. Bayesian network (BN) analysis

In this section, we introduce Bayesian networks and their uses in biomedical research. Most specifically, models generated for understanding biological pathways and relevant gene regulatory networks are discussed.

3.1 Introduction to Bayesian networks

One exciting development in bioinformatics research was the advent and application of Bayesian networks (BN) in biological research. Basically, BNs are graphical representations of statistical interdependencies amongst sets of nodes. BNs model interactions amongst sets of variables (*e.g.* genes, proteins) as probabilistic dependencies or influences. Judea Pearl introduced the notion of Bayesian networks in 1985 (Pearl, 1985; Pearl, 1988) to emphasize three aspects: (i) Often subjective nature of the input data information; (ii) Reliance on Bayes's conditioning as the basis for information updating; and (iii) Distinction between causal and evidential modes of reasoning. Bayesian networks were later implemented by Heckerman et al, Friedman et al, and various other research labs towards biological research (Cooper and Herskovits, 1992; Friedman, et al., 2000; Heckerman, 1995).

Specifically, a BN for a set of variables $X = \{X_1, X_2, ...,X_n\}$ consists of (1) a network structure S that encodes a set of conditional independence assertions about variables in X, and (2) a set P of conditional probability distributions associated with each variable (Heckerman, 2008). Together, these components denote the joint probability distribution for X. The BN structure S is a directed acyclic graph, meaning that the network is hierarchical and has both top-level and terminal nodes and no directed paths which eventually return to them. We use Pa_i to denote the parents of node Xi in S as well as the variables corresponding to those parents. Given structure S, the joint probability distribution for X is given by

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{pa}_i) \tag{1}$$

Different methods have been developed to learn BN structures and will be introduced in detail next.

3.2 Learning Bayesian networks (BNs)

The problem of learning a Bayesian network can be stated as follows: given a training dataset of independent instances, find a network that best matches the dataset. The common approach to this problem is to introduce a statistically sound scoring function that evaluates each network with respect to the training dataset and to search for the optimal network based on this score.

To dissect the processes of learning BNs, we summarize five major steps as follows:

1. Data selection and pre-processing
2. Prior definition (including variables and edges)
3. Selection of network searching strategy (e.g., simulated annealing, greedy)
4. BN execution with a specific scoring method
5. Results output and analysis

These steps will be introduced in detail here for gene expression data analysis:

3.2.1 Data selection and preprocessing

BN is a powerful tool for analyzing high throughput data, e.g., DNA microarray data. Pre-processing is usually required to normalize raw data and possibly filter out those genes that do not show significant changes over all conditions.

3.2.2 Prior definition (including variables and edges)

After selecting appropriate data and variable sets for investigation, settings for the BN simulation must be chosen. Initially, assumptions must be made as to whether structural priors (e.g. the requirement of certain interactions to appear in a model) should be included or not in the BN analysis. It is not necessary to assume any structural priors for the initial set of variables. However, structural priors can be implemented, especially in cases where the biological interactions to be represented are well-established and also fully represented in the underlying biological data used for modelling.

3.2.3 Set up network searching strategy

Once the prior is specified, the BN learning becomes finding a structure that maximizes the BN score according to a BN scoring function. This problem is proven to be NP-complete (Chickering, 1996). Thus heuristic search is needed. The decomposition of the score is crucial for the optimization problem. For example, a local search procedure that changes one edge at a time can efficiently evaluate the gains of a specified score made by adding, removing, or reversing an edge. An example of such a procedure is a greedy random search algorithm with random restarts. Although this procedure does not necessarily achieve a global maximum, it reaches a local maximum and does perform well in practice (Friedman, et al., 2000). Another commonly used method is simulated annealing search algorithm with a temperature schedule that allows for "reannealing" as the temperature is lowered (Heckerman, 1995). Other BN searching strategies include stochastic hill-climbing and genetic algorithm (Friedman, et al., 2000).

3.2.4 Bayesian network scoring approaches

The key part of BN learning is to determine a scoring metric that compares networks and identifies the most likely or 'best supported' networks. Bayesian network scoring is based upon conditional probabilities. One commonly used scoring method is the BDe score (Cooper and Herskovits, 1992; Heckerman, 1995), which is a posterior probability defined as:

$$P(M \mid D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{\left(N_{ij} + r_i - 1\right)!} \prod_{k=1}^{r_i} N_{ijk}\,!\,, \tag{2}$$

where n is the number of variables, q_i is the number of parent configurations for given variable i, r_i is the arity of variable i, N_{ij} is the number of observations with selected parent configuration q_i, N_{ijk} is the number of observations of child in state k with parent configuration q_i (Cooper and Herskovits, 1992). The calculation of this score is implemented in many software programs such as BANJO (Smith, et al., 2006).

Another BN scoring method is the Bayesian Information Criterion (BIC), which was specifically designed to compensate for overfitting (Schwarz, 1978).

3.2.5 Bayesian network analysis software

Many BN analysis software programs are available. Dr. Kevin Murphy provides an excellent summary of existing software packages for Bayesian network modelling (http://www.cs.ubc.ca/~murphyk/Bayes/bnsoft.html). Table 1 lists selected BN software programs from Dr. Murphy's website and other resources.

Name	Source	API	GUI	Undir	Exec	Free	Inference	Exp	Reference
Banjo	Java	Y	N	D	W,U, M	Y	N	N	(Bose, et al., 2006)
BayesiaLab	N	N	Y	C,G	W,U, M	N	Jtree, Gibbs	N	(Conrady and Jouffe, 2011)
BNT	Matlab, C	Y	N	D,U	W,U, M	Y	Several options	N	(Murphy, 2001)
BNJ	Java	N	Y	D	C	Y	Jtree, IS	N	http://bnj.sourceforge.net
Causal Explorer	Matlab, C/C++	Y	N	D	W,U, M	Y	N	N	(Aliferis, et al., 2003)
Deal	R	Y	Y	D	I	Y	N	N	(Bøttcher and Dethlefsen, 2003)
Genie	C++	Y	Y	D	C	Y	Jtree	N	(Druzdzel, 1999)
Java Bayes	Java	Y	Y	D	C	Y	Jtree, Varelim	N	http://www.cs.cmu.edu/~javabayes/Home/
LibB	N	Y	N	D	W,L	Y	N	N	http://www.cs.huji.ac.il/labs/compbio/LibB/
MARIMBA	N	N	Y	D	I	Y	N	Y	(Hodges, et al., 2010; Hodges, et al., 2010)
miniTUBA	N	N	Y	D	I	Y	N	N	(Xiang, et al., 2007)
openBUGS	Y	Y	Y	D	W,U, M	Y	Gibbs	N	(Lunn, et al., 2000; McCarthy, 2007)
OpenPNL	C++	Y	Y	D	W,L	Y	Jtree, Gibbs	N	http://sourceforge.net/projects/openpnl/
PEBL	Python	Y	Y	D	W,U, M	Y	N	N	(Shah and Woolf, 2009)
WinMine	N	N	Y	D,U	W	Y	N	N	(Chickering, 2002)

Notes: The catergories listed include: Source, source code; API, application program interface for programmatic access; GUI, graphical user interface; Undir, ability to handle undirected graphes; Exec, the type of execution, including W:Windows, U:Unix, L:Linux, M:Mac, I:OS-independent, or C:any with compiler; Free, the availability of the software as either free (e.g. academic) or commercial; Inference, inferencing ability; Exp, ability for network expansion; and Ref, references.

Table 1. Selected software programs for BN analysis.

3.2.6 BN result output and analysis

To visualize BN results, different methods can be performed. For example, BANJO uses DOT type of BN result output (Reference: http://www.graphviz.org/Documentation/dotguide.pdf). MARIBMA uses DOT and can also export networks as .sif format for use in Cytoscape (http://www.cytoscape.org). Since different BNs are available, it is crucial for a user to select 'best-scoring' networks and/or generate consensus networks. Often methods are also needed to build weighted networks based on computational analysis or from literature and other database queries.

4. Bayesian network expansion methods

Bayesian network (BN) expansion is an approach that is built upon the BN method and aims to identify new pathway elements that participate in a specified network. In this section, we will introduce basic BN expansion methods and then focus on describing our internally developed BN+1 algorithm and its implementation.

4.1 General BN expansion

Compared to the otehr network expansion methods described above, Bayesian network-based expansion methods provide distinct advantages, such as prediction of both linear and nonlinear functions, robustness in noise data analysis, and identification of causal or appearly causal influences representing interactions among genes. In general, Bayesian network expansion can be defined as the addition of new variables to an existing network, followed by rescoring and ranking of those variables.

BN-based expansion has been used for gene expression data analysis (Gat-Viks and Shamir, 2007; Pena, et al.). For example, Pena et al. reported an algorithm AlgorithmGPC that also grows BN models from seed genes (Pena, et al.). This approach starts with one single gene and builds networks around this gene through expansion and pruning with a set number of genes. Gat-Viks et al also generated a Bayesian network-based refinement and expansion method (Gat-Viks and Shamir, 2007). A main limitation of this approach is that it requires high quality of prior knowledge on the signaling pathways. The topology of the biological pathways may not be consistent with networks learned from transcriptional gene expression data obtained via DNA microarray studies. Therefore, a fixed topology as initial seed network may not be appropriate for robust network expansion simulaions.

Other BN expansion methods have also been published (Needham, et al., 2009; Parikh, et al., 2010). These approaches differ from each other but all showed different levels of success in identifying new pathway elements. In the following two sections, we will introduce our BN+1 algorithm (Hodges, et al., 2010; Hodges, et al., 2010), and how it is implemented in the MARIMBA software.

4.2 The BN+1 algorithm

In our recent study, we developed an algorithm termed "BN+1" which implements Bayesian network expansion to predict new factors and interactions that participate in a specific pathway (Hodges, et al., 2010; Hodges, et al., 2010). Broadly, the BN+1 algorithm iteratively tests to see if any single variable added to a given pathway will significantly improve the likelihood of the overall network. This approach is based on the observation that those variables which are hidden and regulate or are regulated by a network are more likely ranked with high posterior probability scores. Using a compendium of microarray gene expression data obtained from *Escherichia coli*, the BN+1 algorithm predicted many novel factors that influence the *E. coli* reactive oxygen species (ROS) pathway. Some of the predicted new ROS and biofilm regulators (*e.g.*, *uspE* and its interaction with *gadX*) were further experimentally verified (Hodges, et al., 2010). In another study, a synthetic network was also designed to further evaluate this algorithm. Based on the synthetic data analysis, the BN+1 method is able to identify both linear and nonlinear relationships and correctly identify variables near the starting network (Hodges, et al., 2010).

The BN+1 algorithm is specified in Figure 1. A few notes are provided here in our BN+1 implementation:

1. The selection of seed (or core) genes is an important step. The seed genes can be selected from an existing pathway database, from literature survey, or from internal experimental results. Since it is computationally expensive to calculate BNs using a large number of variables, it is often necessary to filter out some genes from an initial list using different criteria, for example, filtering out those genes that do not have significant changes among all microarray chips.
2. While we use a top network structure generated from initial core gene simulation as a prior, we prefer not to fix the core network structure for subsequent network expansion. This preference makes our approach differ from a commonly used method of fixing the prior structure. Our argument is that the prior structure is often determined by many layers of studies, including DNA, RNA and protein data analyses. When only RNA transcriptomic data are used, such prior structure may not hold. The fixture of a prior structure would result in obtaining suboptimal networks that do not match the datasets used for BN simulation.

BN+1 Algorithm
Input: *N* variables (e.g., genes) from a dataset (e.g., microarray dataset) with *L* observations each.

Data Preprocessing (Optional)
Filter out *m* variables (e.g., via coefficient of variation (c.v.) <= 1.0)
Number of possible variables for analysis: *N*= *N*-*m*.

BN Core Network Searching
Select *K* variables from the set of *N* variables (e.g. from a pathway database).
Construct matrix data file *D* with *K***L* observations using *K* variables and *L* observations.
Select settings for BN simulation, including data discretization (e.g. q3 quantization), searcher strategy (e.g. simulated annealing), and structural priors.
Execute BN simulation (e.g. using BANJO).
Save top BN network topology *C*

Iterative Core Expansion
Assign the core topology *C* as unfixed structural prior for BN searching
For each variable *a* in the set {*N*-*K*}, do:
    Generate new data file *D** by concatenating *L* observations for *a* to data file *D*
    Select settings for BN simulation.
    Execute BN simulation.
    Save top network and its posterior probability for *a*.
Rank each variable according to posterior probability.

Output: Rank-ordered BN+1 results.

Fig. 1. BN+1 algorithm.

4.3 Implementation of BN+1 using MARIMBA

MARIMBA is implemented using a three-tiered architecture built on two Dell Poweredge 2580 servers which run the Redhat Linux operating system. Users submit analysis requests and database queries through the web. These queries are then processed using PHP, Perl, Python, JavaScript, and SQL (middle-tier, application server based on Apache) against a MySQL (version 5.0) relational database (back-end, database server). The result of each query is then presented to the user in the web browser.

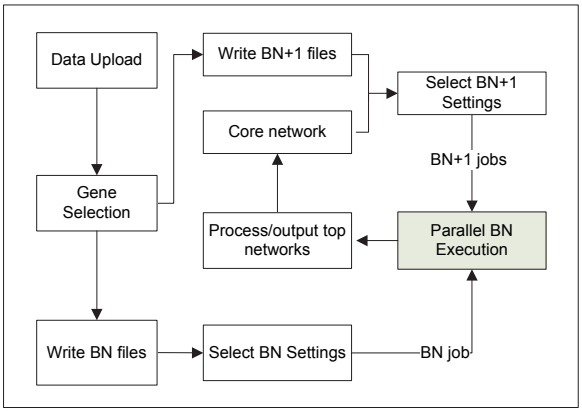


Fig. 2. Workflow of MARIMBA implementation of BN+1 algorithm.

The main MARIMBA system architecture and pipeline for analysis of project data is described in Figure 2 and contains the following steps:

1. *Data Upload:* A user can upload gene expression data using specified format.
2. *Gene Selection:* A gene list as core gene set for BN modeling will be specified by the user.
3. *Write up BN and BN+1 files:* The basic MARIMBA-formatted files are then generated dynamically by MARIMBA for use in Bayesian modeling.
4. *Selection of BN parameters:* BN simulation settings were selected after completing the data and gene selection processes, respectively. A static BN simulation was created to analyze the microarray data. Many settings can be selected by the user. For example, a user can select simulated annealing or greedy method as the network searcher method. We prefer simulated annealing due to its improved performance over greedy searches when no prior knowledge of underlying structure is available (Hartemink, et al., 2002). In our simulated annealing analysis, a relatively low cooling factor is often implemented to allow less restrictive searching of the sample space and potentially identify as many equivalence classes for the top-scoring network as possible. Currently, up to 1,000 networks can be stored in MARIMBA for each run.
5. *Execution of BN and BN+1modeling:* BN files are submitted via the online interface in MARIMBA. Each dataset is transferred to the server prior to simulation by a parallel computer cluster. Each agent runs a unique BANJO simulation. The core BN network is employed as a fixed topology/prior knowledge network in the BN analysis. A BN+1 simulation can also be implemented as defined by a user.
6. *BN Result display and interpretation:* MARIMBA displays top-scoring networks of BN and BN+1 simulations to the user. The images of top-scoring networks are converted directly from their original dot files and are displayed as jpeg images on-the-fly. In addition, MARIBMA is able to calculate conserved edges over a selected number of stored networks. To calculate the conserved edges, MARIMBA determines core BN models by model averaging and equivalence class searching. Here, model averaging is defined as inclusion of an edge between two genes if that edge appeared in more than X percent of the top-scoring networks with identical score. Furthermore, The BN+1 visualization environment in MARIMBA displays a plot for posterior probabilities of saved networks, thus enabling comparison of networks for relevance and likelihood.

Compared to the standalone BANJO system (Bose, et al., 2006), MARIMBA is web-based and allows seamless integration of user project management, analysis construction, BN submission and execuation in a parallel computing environment, and analysis and visualization of results. The user-friendly GUI environment simplifies dataset selection, probeset/gene inclusion, observational file processing, and settings selection for BN execution. Such features are necessitated for efficient querying by biologists who wish to use such BN tools to analyze their data. In addition, the BN+1 algorithm execution and project management is a unique feature in MARIBMA and does not exist in BANJO or any other software program.

5. Use case study: Application of BN+1 to BCR pathway modelling

5.1 Introduction

As an example of the challenge of merging a pathway model and gene expression data, this study focuses on the B-cell receptor pathway (BCR) as described by KEGG (Kanehisa and Goto, 2000; Kanehisa, et al., 2010). The BCR pathway is an integral component of the adaptive immune response mechanism by which B cells respond to foreign antigens (Lucas, et al., 2004). The BCR pathway involves in the activation of specific protein kinase C (PKC) isoforms that induces ultimate activation of the NF-κB transcription factor. Multiple protein species accumulate at the cell membrane in a signalosome complex and are linked to the B cell receptor. Signal propagation from the BCR via kinase-mediated phosphorylation cascades to downstream effectors such as Nfkb, NFAT (nuclear factor of activated T cells), and AP1 is either enhanced or reduced via signalosome interactions with co-stimulatory or co-inhibitory complexes, respectively. BCR signaling guides many important functions such as anergy, B cell ontogeny, and immune response, and is linked to the several imporant pathways: MAPK, coagulation/complement cascades, and actin cytoskeleton (Kanehisa and Goto, 2000; Kanehisa, et al., 2010). NF-κB plays a crucial role in the antigen-induced B lymphocyte proliferation, cytokine production, and B cell survival (Lucas, et al., 2004). While the KEGG pathway database includes a manually curated BCR pathway, this pathway is still considered incomplete (Lucas, et al., 2004).

5.2 Microarray data processing and BN analysis methods

We used gene expression data from perturbed B-cells obtained from the Alliance for Cellular Signaling (AfCS) (Lee, et al., 2006; Zhu, et al., 2004). This dataset is especially attractive because the same tissues were treated with combinations of ligands that perturb different B cell pathways. The AfCS study gathered 424 microarray chips measuring gene expression in B cells from *M. musculus* splenic extracts that are exposed to 33 different ligands (Lee, et al., 2006; Papin and Palsson, 2004; Zhu, et al., 2004). Briefly, B cells purified from splenic preparations from 6- to 8-wk-old male C57BL/6 mice were treated in triplicates or quadruplicates with medium alone, or one of 33 different ligands for 0.5, 1, 2, and 4 h (AfCS protocol PP00000016). RNA was extracted following standard AfCS protocol PP00000009. An Agilent cDNA microarray chip that contains 15,494 cDNA probes printed on 15,832 spots was used. It represents 10,615 unique MGI gene matches (Lee, et al., 2006). Each Agilent array was hybridized with Cy5-labeled cDNA prepared from splenic B cell RNA and Cy3-labeled cDNA prepared from RNA of total splenocytes used as an internal reference (AfCS protocol PP00000019). Hence, each Agilent microarray chip provides one unique observation of relative expression level per selected probe. The arrays were scanned

using Agilent Scanner G2505A, and images were processed using the Agilent G2566AA Feature Extraction software version A.6.1.1. The microarray raw data were downloaded from the AfCS repository at ftp://ftp.afcs.org/pub/datacenter/microarray/.

Microarray data were discretized for each variable in the Bayesian networks using quantile normalization with three bins. Though triplicate or quadruplicate microarray experiments were available in most cases per unique treatment and time of drug administration, we assume that each experiment provides an independent source of information. In this analysis, we did not use all BCR pathway genes. We sought to answer here whether expansion of a sub-network from the BCR pathway would preferentially recover other BCR pathway genes. This assumption is advantageous in that the number of variables allows significantly faster simulation searches for the BN and BN+1 simulations. Particularly, those genes most specifically involved in Nfkb-mediated transcriptional regulation were chosen from the KEGG BCR pathway.

A set of 10,000 top-scoring BNs was generated using the eight variables (the core) and 424 observations. Among the eight variables, two variables are Nfkbie probe sets, and two are Ikbkb probe sets. In many cases, one gene has multiple probe sets. We chose to separate them as different variables in our BN analysis since often these probe sets have different values with low correlation (Fig. 3). This BN analysis was accomplished by running 100 independent simulations and saving the top 100 simulations for each of those runs.

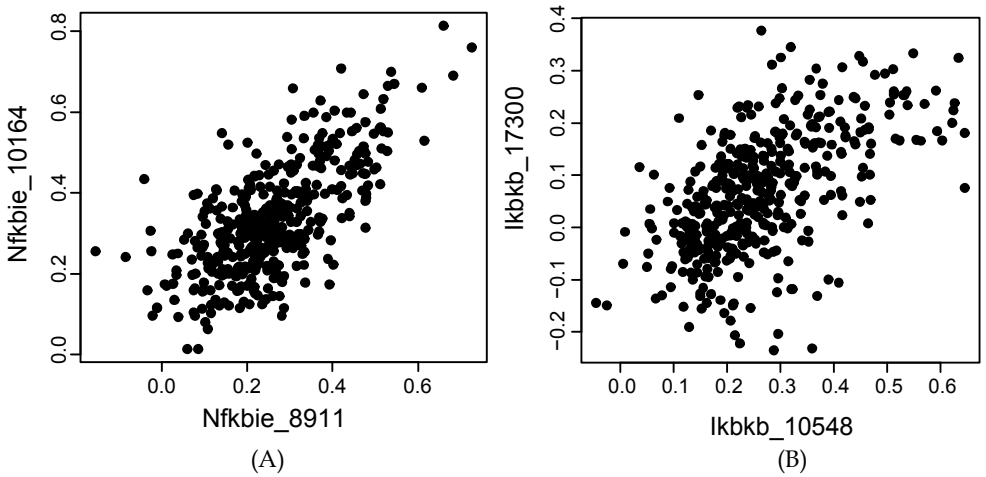


Fig. 3. Scatter plots for Nfkbie and Ikbkb probes from AfCS study. Agilent probe identifiers are listed next to each respective gene. This figure indicates that the probe sets Nfkbie_10164 and Nfkbie_8911 correlate relatively well with a Pearson correlation coefficient of 0.69 (A). However, the correlation between Ikbkb_17300 and Ikbkb_10548 is low (Pearson correlation coefficient: 0.58) (B).

5.3 Results

5.3.1 Defining the core network

Fig. 4 depicts the shared set of interactions appearing in all of the top networks sharing the same best score.

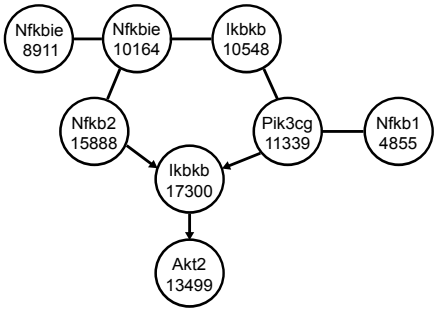


Fig. 4. Consensus of top scoring Bayesian networks for eight probes representing BCR receptor signaling pathway genes. Gene symbols and corresponding Agilent probe identifiers are represented in nodes in the network. Directed edges represent those influences appearing in the same direction in all top-scoring Bayesian networks, while undirected edges appear at least once in the opposite direction though appearing cumulatively with 100% frequency in all of the top networks.

Compared with the KEGG BCR pathway, the consensus network found in our BN analysis (Fig. 2) has a 75% overlap with known interactions (3 out of 4 were correctly predicted), with only one interaction missing (Fig. 5).

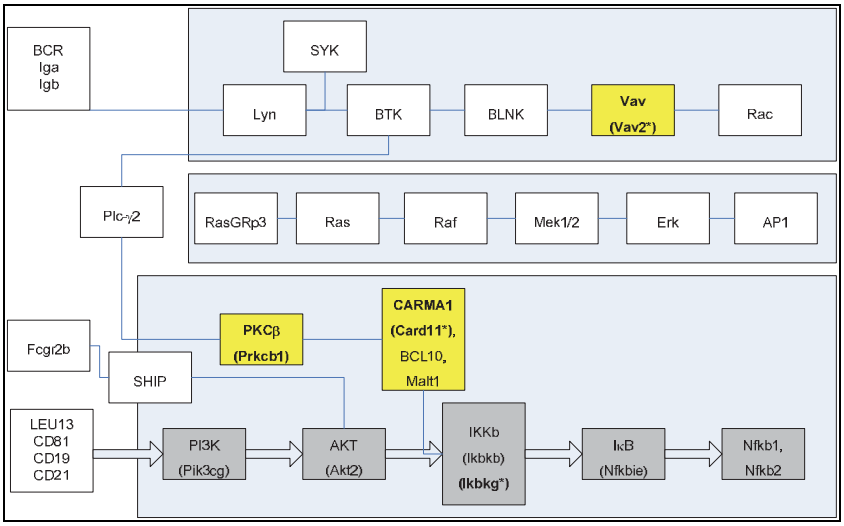


Fig. 5. Schematic representation of the BN+1 analysis results in the content of KEGG BCR pathway. The three blue boxes represent three major sub-networks within the BCR pathway with distinct regulatory and functional roles. The BN core network was defined using members from the third sub-network (dark grey boxes) which reflect major components of Nfkb signalling. Bolded gene names are those genes which were not included in the core network, yet were recovered during BN+1 analysis in the top 100 results. Note that not all members of the listed Nfkb signalling pathway were included in the core network (e.g. Ikbkg), and in some cases were not available on the microarray platform.

5.3.2 Defining BN+1 genes

One of the top-scoring networks used to generate the consensus shown in Fig. 4 was used as a core network for subsequent BN+1 expansion. BN+1 searching was executed for 14,353 individual probes with 50 million networks searched per probe. If only those genes in close neighbourhood in the KEGG BCR pathway are considered, out of 19 selected genes, nine genes were found to be connected to the core network in our analysis. Furthermore, four of these nine genes are in close proximity (within top 10% of top-scoring BN+1 genes with at least one connection to the core network) with these core genes in the KEGG protein signalling pathway: Card11, Prkcb1, Ikbkg, and Vav2. These results suggest that the neighbourhood of transcriptional regulation around the core network as well as distance between the elements in the protein signalling pathway are related to each other.

Analysis of the top BN+1 variables recovered during simulation revealed several interesting results. First, the top set of BN+1 variables is listed in Table 1.

Rank	Agi_ID	GeneID	Symbol	BN1_score	Neighbors
1	11062	77619	Prelid2	-3402.0	Nfkb2
2	9502	20744	Strbp	-3517.0	Nfkbie
3	14138	20823	Ssb	-3545.2	Nfkb2
4	6276	12530	Cdc25a	-3569.2	Nfkb2
5	11361	108829	Jmjd1c	-3586.8	Ikbkb(both), Pik3cg
6	14614	75964	Trappc8	-3587.8	Ikbkb, Pik3cg
7	15876	108786	Cxcl13*	-3593.1	Nfkb2
8	10759	73132	Slc25a16	-3594.8	Ikbkb, Pik3cg
9	5275	67887	Tmem66	-3596.0	Nfkb1, Pik3cg
10	9036	109339	2700018L05Rik	-3599.1	Pik3cg

Notes: Identifier information for each ranked gene is provided, including Agilent probe ID (Agi_ID), Entrez gene ID (GENEID), and gene symbol. Probe variables from the core network which directly connect to the BN+1 variables in the top-scoring networks are listed in the "Neighbors" column.

Table 2. Top ten predicted BN+1 genes.

Many interesting findings were observed from this analysis. Many genes, for example, the Sjorgen syndrome antigen B gene (Ssb) (Brenet, et al., 2009), have been shown to be associated with the Nf-kB and BCR pathways. Ssb plays an important role in polysome translation (Brenet 2009), and is an early DNA-damage responder in apoptotic cells and those treated with cytotoxic chemicals (Al-Ejeh, et al., 2007). Interestingly, we identified Jmjd1c, a member of the jumonji family proteins, as a top predicted gene in our BN+1 simulation. Jmjd1c is conserved in several mammalian species and has documented roles in metal ion binding, oxidoreductase activity, and transcriptional regulation (Katoh, 2007). The murine Jmjd1c mRNA is expressed in multiple tissues, including hematopoietic and undifferentiated ES stem cells, fertilized egg, pancreatic islet, etc (Katoh, 2007). Jmjd1c has a promoter region orthologous to humans with binding sites for the AP-1 transcription factor, which is considered a member of the BCR signalling pathway and is included in the KEGG

representation as AP1 (downstream of the Raf/MEK sub-network in Figure 5 though not in our core network). Fig. 6 illustrates the strongly-correlated relationships uncovered between the Jmjd1c genes and connected core network members. As another example, the Cxcl13 is a chemokine ligand in B cells with a C-X-C motif. It has already been established that Cxcl13 induction requires activation of canonical and non-canonical Nf-kB pathways (Suto, et al., 2009), which confirms the prediction of this gene in our network. These data strongly support the predictions generated by our analysis.

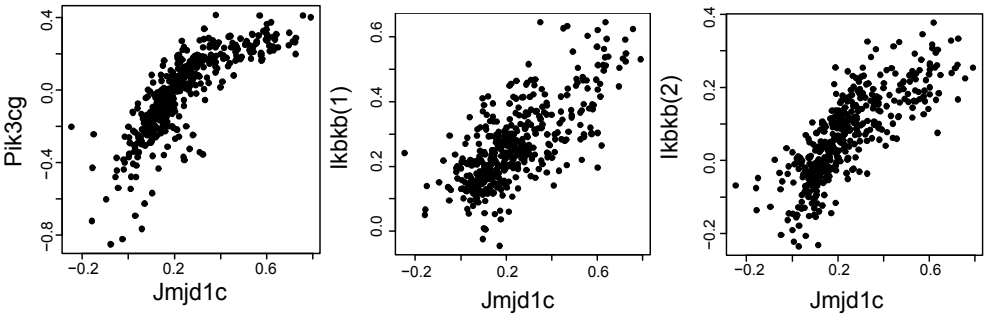


Fig. 6. Scatter plot of expression values for core genes Pik3cg and Ikbkb (both probes) versus BN+1 gene Jmjd1c. A non-linear association between Pik3cg and Jmjd1c is observed (A). A roughly linear relation is observed between Jmjd1c and Ikbkb(1) (Pearson correlation coefficient: 0.71) (B) and between Jmjd1c and Ikbkb(2) (Pearson correlation coefficient: 0.79) (C).

One property of interest, as shown in the table, is that the core genes which recruit the top BN+1 genes are not always the same. From this analysis and previous studies, we have observed that BN+1 variables which show high correlations to at least one core network variable often appear as top BN+1 results. However, in some cases, the BN+1 variable may connect to multiple variables in the core network, and yet show moderate to low correlations with each of them. It is observed that many BN+1 variables have multiple core network variables as parent nodes in the predicted top network. Multi-parent relationships are less common, though statistically more meaningful due to the nature of the implemented conditional probability tables in BDe scoring.
Different methods, such as clustering and GO gene enrichment, can be used to further analyze BN+1 genes.

5.3.3 Clustering analysis of core genes and BN+1 genes

A clustering method provides a way to group BN+1 genes based on gene expression values. A heapmap clustering analysis was performed using 8 probe sets in the core network and 10 probe sets from the BN+1 analysis (Fig. 7). As shown in this heatmap, all NF-κB genes (core genes in our BN simulation) are clustered together, indicating their close association. Our analysis also found that Jmjd1c is closely associated with these NF-κB genes. This further strengthens our BN+1 prediction of the important role of this gene in the NF-κB pathway in B cell signalling.

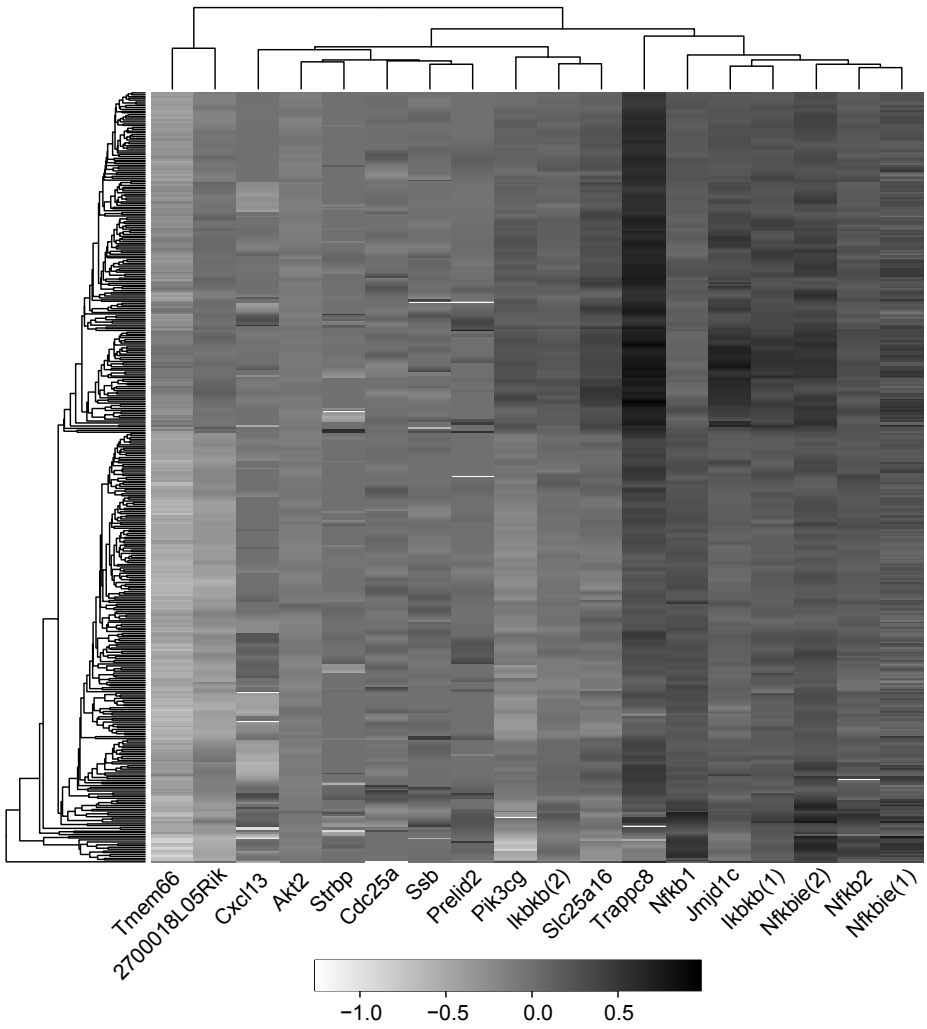


Fig. 7. Heatmap of expression data for top BN+1 and core variables. Parentheses indicate specific probe identities.

5.3.4 GO enrichment of predicted BN+1 genes

Our previous studies indicate that the top few hundred BN+1 genes (i.e. those genes predicted by the BN+1 algorithm) often interact with the seed gene network and biologically active relevant to the pathway of interest (Hodges, et al., 2010; Hodges, et al., 2010). A GO gene enrichment analysis was performed using 250 top BN+1 genes (Table 3). Given the nature of the Nfkb-selected core network and their roles in nuclear localization and transcriptional initiation, it was not surprising that many of the recovered genes show some nuclear compartmentalization. Interestingly, many apoptotic and death-related genes were enriched (Table 3).

Term	*Count*	*P-Value*	*Benjamini P-value*
Biological Process			
GO:0009987~cellular process	106	8.29E-06	0.00981
GO:0070227~lymphocyte apoptosis	4	1.44E-04	0.0823
GO:0008219~cell death	15	1.73E-04	0.0663
GO:0016265~death	15	2.20E-04	0.0634
GO:0048569~post-embryonic organ development	4	2.36E-04	0.0546
GO:0006915~apoptosis	14	2.65E-04	0.0512
GO:0012501~programmed cell death	14	3.12E-04	0.0517
Cellular Compartment			
GO:0005622~intracellular	125	2.93E-08	5.68E-06
GO:0044424~intracellular part	119	4.32E-07	4.19E-05
GO:0043229~intracellular organelle	105	2.76E-06	1.78E-04
GO:0043226~organelle	105	2.84E-06	1.38E-04
GO:0043231~intracellular membrane-bounded organelle	93	4.08E-05	0.00158
GO:0043227~membrane-bounded organelle	93	4.24E-05	0.00137
GO:0005634~nucleus	58	0.001749	0.0474

Notes: Entrez gene identifiers were input for the top 250 BN+1 results into the DAVID tool for GO analysis. The 250 results mapped to 188 unique *Mus musculus* and seven unknown species genes, revealing that some of the top genes were represented by multiple Agilent probes in the top results. Benjamini-derived p-values of 0.01 were used as cutoffs here.

Table 3. GO enrichment results for top 100 predicted variables in the BN+1 analysis.

6. Conclusion

In this paper, different bioinformatics methods for network expansion and detection of new pathway elements are surveyed. Bayesian network-based expansion methods are specifically introduced. Particularly, we outline our BN+1 Bayesian network method that can be used to iteratively compare BDe scores and rank those genes that are likely critical to a specific pathway or network. BN+1 has been successfully demonstrated in *E. coli* system and synthetic data simulation. In this paper, we first demonstrate its use in BCR pathway, a eukaryotic signalling pathway. Our study shows that BN+1 can also be used to predict pathway elements and gene interactions in important eukaryotic pathways. Therefore, the BN+1 algorithm appears to be a generic BN expansion system that can be used to study other prokaryotic and eukaryotic pathways.

Many future directions are envisioned. For example, we can extend the BN+1 algorithm to BN+2, BN+3, or BN+n algorithm by iteratively adding more than one variable to the seed gene network. The principle used in the development of the BN+1 algorithm can also be used for dynamic BN analysis. We are currently in the process of developing a DBN+1 algorithm and using it for temporal data analysis.

7. Acknowledgment

This research was supported in part by NIH Grant U54-DA-021519, NIH Training Grant (5 T32 GM070449-04), 2008 Rackham Spring/Summer Research Grant at the University of Michigan, and the University of Michigan Bioinformatics Program.

8. References

Al-Ejeh, F., *et al.* (2007) In vivo targeting of dead tumor cells in a murine tumor model using a monoclonal antibody specific for the La autoantigen, *Clin Cancer Res*, 13, 5519s-5527s.

Aliferis, C.F., *et al.* (2003) Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery. Citeseer, pp. 371–376.

Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, 20, 1464-1465.

Bose, R., *et al.* (2006) Phosphoproteomic analysis of Her2/neu signaling and inhibition, *Proc Natl Acad Sci U S A*, 103, 9773-9778.

Bøttcher, S.G. and Dethlefsen, C. (2003) *deal: A package for learning Bayesian networks*. Department of Mathematical Sciences, Aalborg University.

Brenet, F., *et al.* (2009) Akt phosphorylation of La regulates specific mRNA translation in glial progenitors, *Oncogene*, 28, 128-139.

Chickering, D.M. (1996) Learning Bayesian networks is NP-complete, *Learning from data: Artificial intelligence and statistics v*, 112, 121–130.

Chickering, D.M. (2002) The winmine toolkit, *Microsoft Research MSR-TR-2002-103*.

Conrady, S. and Jouffe, L. (2011) Breast Cancer Diagnostics with Bayesian Networks.

Cooper, G.F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9, 309-347.

Druzdzel, M.J. (1999) SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: a development environment for graphical decision-theoretic models. JOHN WILEY & SONS LTD, pp. 902-903.

Eisen, M.B., *et al.* (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, 95, 14863-14868.

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

Gat-Viks, I. and Shamir, R. (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast, *Genome Res*, 17, 358-367.

Hartemink, A.J., *et al.* (2002) Combining location and expression data for principled discovery of genetic regulatory network models, *Pac Symp Biocomput*, 437-449.

Hashimoto, R.F., *et al.* (2004) Growing genetic regulatory networks from seed genes, *Bioinformatics*, 20, 1241-1247.

Heckerman, D. (2008) A tutorial on learning with Bayesian networks, *Innovations in Bayesian Networks*, 33-82.

Heckerman, D., Geiger, D. (1995) Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning*, 20, 197-243.

Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, 17, 126-136.

Herrgard, M.J., Covert, M.W. and Palsson, B.O. (2003) Reconciling gene expression data with known genome-scale regulatory network structures, *Genome Res*, 13, 2423-2434.

Hodges, A., Woolf, P. and He, Y. (2010) BN+ 1 Bayesian network expansion for identifying molecular pathway elements, *Communicative & Integrative Biology*, 3, 59-64.

Hodges, A.P., *et al.* (2010) Bayesian network expansion identifies new ROS and biofilm regulators, *PLoS One*, 5, e9513.

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

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc*, 4, 44-57.

Ihmels, J., *et al.* (2002) Revealing modular organization in the yeast transcriptional network, *Nat Genet*, 31, 370-377.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res*, 28, 27-30.

Kanehisa, M., *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res*, 38, D355-360.

Katoh, M. (2007) Comparative integromics on JMJD1C gene encoding histone demethylase: conserved POU5F1 binding site elucidating mechanism of JMJD1C expression in undifferentiated ES cells and diffuse-type gastric cancer, *Int J Oncol*, 31, 219-223.

Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets, *J Theor Biol*, 22, 437-467.

Lee, J.A., *et al.* (2006) Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation, *BMC Bioinformatics*, 7, 237.

Lucas, P.C., McAllister-Lucas, L.M. and Nunez, G. (2004) NF-kappaB signaling in lymphocytes: a new cast of characters, *J Cell Sci*, 117, 31-39.

Lunn, D.J., *et al.* (2000) WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility, *Statistics and Computing*, 10, 325-337.

Luo, W., Hankenson, K.D. and Woolf, P.J. (2008) Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information, *BMC Bioinformatics*, 9, 467.

Luo, W. and Woolf, P.J. (2010) Reconstructing transcriptional regulatory networks using three-way mutual information and Bayesian networks, *Methods Mol Biol*, 674, 401-418.

McCarthy, M.A. (2007) *Bayesian methods for ecology*. Cambridge Univ Pr.

Meier, S. and Gehring, C. (2008) A guide to the integrated application of on-line data mining tools for the inference of gene functions at the systems level, *Biotechnol J*, 3, 1375-1387.

Murphy, K. (2001) The bayes net toolbox for matlab, *Computing science and statistics*, 33, 1024-1034.

Needham, C.J., *et al.* (2009) From gene expression to gene regulatory networks in Arabidopsis thaliana, *BMC Syst Biol*, 3, 85.

Papin, J.A. and Palsson, B.O. (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis, *Biophys J*, 87, 37-46.

Parikh, A., *et al.* (2010) New components of the Dictyostelium PKA pathway revealed by Bayesian analysis of expression data, *BMC Bioinformatics*, 11, 163.

Pearl, J. (1985) *Bayesian networks: A model of self-activated memory for evidential reasoning*. Computer Science Department, University of California.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pena, J.M., Bjorkegren, J. and Tegner, J. (2005) Growing Bayesian network models of gene networks from seed genes, *Bioinformatics*, 21 Suppl 2, ii224-229.

Raychaudhuri, S., *et al.* (2001) Basic microarray analysis: grouping and feature reduction, *Trends Biotechnol*, 19, 189-193.

Schwarz, G. (1978) Estimating the dimension of a model, *The annals of statistics*, 461-464.

Shah, A. and Woolf, P. (2009) Python environment for Bayesian learning: inferring the structure of Bayesian Networks from knowledge and data, *The Journal of Machine Learning Research*, 10, 159-162.

Shmulevich, I., *et al.* (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, 18, 261-274.

Smith, V.A., *et al.* (2006) Computational inference of neural information flow networks, *PLoS Comput Biol*, 2, e161.

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

Suto, H., *et al.* (2009) CXCL13 production by an established lymph node stromal cell line via lymphotoxin-beta receptor engagement involves the cooperation of multiple signaling pathways, *Int Immunol*, 21, 467-476.

Tamayo, P., *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc Natl Acad Sci U S A*, 96, 2907-2912.

Tanay, A. and Shamir, R. (2001) Computational expansion of genetic networks, *Bioinformatics*, 17 Suppl 1, S270-278.

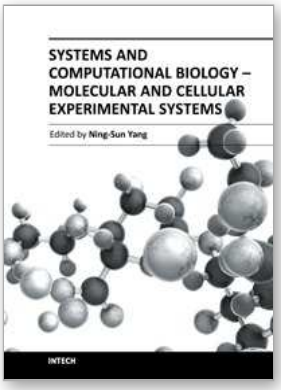
Watkinson, J., *et al.* (2009) Inference of Regulatory Gene Interactions from Expression Data Using Three Way Mutual Information, *Annals of the New York Academy of Sciences*, 1158, 302-313.

Xiang, Z., *et al.* (2007) miniTUBA: medical inference by network integration of temporal data using Bayesian analysis, *Bioinformatics*, 23, 2423-2432.

Xiang, Z., Qin, Z. and He, Y. (2007) CRCView: A web server for analyzing and visualizing microarray gene expression data using model-based clustering, *Bioinformatics*.

Yu, T. and Li, K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis, *Bioinformatics*, 21, 4033-4038.

Zhu, X., *et al*. (2004) Analysis of the major patterns of B cell gene expression changes in response to short-term stimulation with 33 single ligands, *J Immunol*, 173, 7141-7149.



Systems and Computational Biology - Molecular and Cellular Experimental Systems

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-280-7
Hard cover, 332 pages
Publisher InTech
Published online 15, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book presents a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Andrew P. Hodges, Peter Woolf and Yongqun He ξ (2011). Prediction of Novel Pathway Elements and Interactions Using Bayesian Networks, Systems and Computational Biology - Molecular and Cellular Experimental Systems, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-280-7, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-molecular-and-cellular-experimental-systems/prediction-of-novel-pathway-elements-and-interactions-using-bayesian-networks

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821