Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Modeling the Altered Expression Levels of Genes on Signaling Pathways in Tumors As Causal Bayesian Networks

Richard Neapolitan[1], Diyang Xue[2] and Xia Jiang[2]

[1]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.

ABSTRACT: This paper concerns a study indicating that the expression levels of genes in signaling pathways can be modeled using a causal Bayesian network (BN) that is altered in tumorous tissue. These results open up promising areas of future research that can help identify driver genes and therapeutic targets. So, it is most appropriate for the cancer informatics community.

Our central hypothesis is that the expression levels of genes that code for proteins on a signal transduction network (STP) are causally related and that this causal structure is altered when the STP is involved in cancer. To test this hypothesis, we analyzed 5 STPs associated with breast cancer, 7 STPs associated with other cancers, and 10 randomly chosen pathways, using a breast cancer gene expression level dataset containing 529 cases and 61 controls. We identified all the genes related to each of the 22 pathways and developed separate gene expression datasets for each pathway. We obtained significant results indicating that the causal structure of the expression levels of genes coding for proteins on STPs, which are believed to be implicated in both breast cancer and in all cancers, is more altered in the cases relative to the controls than the causal structure of the randomly chosen pathways.

KEYWORDS: signal transduction pathway, bayesian network, gene expression level, breast cancer, causal structure

CITATION: Neapolitan et al. Modeling the Altered Expression Levels of Genes on Signaling Pathways in Tumors As Causal Bayesian Networks. *Cancer Informatics* 2014:13 77–84 doi: 10.4137/CIN.S13578.

RECEIVED: November 5, 2013. **RESUBMITTED:** November 25, 2013. **ACCEPTED FOR PUBLICATION:** November 25, 2013.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: The research reported here was funded in part by grant R00 LM010822 NIH/NLM from the National Library of Medicine.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: xij6@pitt.edu

Introduction

There is evidence that similar cancers have many variations at the molecular level, and each has its own clinical course. This is called the *heterogeneity* of tumors. For example, in the case of human epidermal growth factor receptor 2 (HER2)-amplified breast cancer, the survival of many patients is vastly improved with the drug Herceptin,[1] but less than 50% respond[2] and the drug can be toxic.[3] This phenomenon is not peculiar to HER2-amplified breast cancer. As another example, breast cancer patients with positive estrogen receptor (ER) expression and negative lymph node metastasis (ER+/node–) have better clinical outcomes than other subtypes of patients; however, a sub-population has recurrence. So, we cannot provide optimal treatment for many patients because responses to treatments are often different for patients who have similar clinical features.

A signal transduction pathway (STP) is a network of information flow in the cells that initiates with a signal outside the cell and results in a cellular response. Many aberrant STPs have been associated with various cancers.[4–10] For example, we now know that the ERbB, PI3K–Akt, and Wnt pathways are associated with breast cancer. The signal aberrations associated with a disease often result from one or more mutated genes that code proteins on the pathways. There has been an explosion of new genomic and proteomic datasets providing us with unprecedented and rich resources to reveal the mechanisms of STPs. We have datasets concerning single nucleotide polymorphisms (SNPs), somatic mutations, copy number, methylation levels, and expression levels in both cancerous and non-cancerous tissues.[11–13] We have flow cytometry datasets providing us with simultaneous observations of many signaling molecules in a multitude of individual cells.[14,15]

To develop optimal treatments for cancer patients, it is necessary to address two fundamental issues regarding STPs: (1) the discovery of which STPs are implicated in a cancer or cancer subtype and (2) the prediction of how stimulations and inhibitions will affect the overall activity of the STP.

Using gene expression datasets, a good deal of effort has been devoted to the first issue just mentioned. Initially, techniques such as over-representation analysis[16–18] were employed. Such methods ignore the topology of the network, and hence do not account for key biological information. That is, if a pathway is activated through a single receptor and that protein is not produced, the pathway will be severely impacted. However, a protein that appears downstream may have a limited effect on the pathway. Recently, researchers have developed methods that account for the topology of an STP when analyzing gene expression data to determine whether the STP is implicated in a cancer.[19–21] Signaling pathway impact analysis (SPIA)[19] is a software package (http://bioinformaticsprb.med.wayne.edu/SPIA) for identifying whether a signaling network is relevant in a given condition that accounts for the topology of the network. However, it is not model based, and does not provide a predictive causal model of an STP. PARADIGM[20] creates a model of a single patient rather than the population, and is able to incorporate copy number variations (CNV) and even mutations. Not being population based, it does not provide an overall causal model of the altered STPs in a given cancer.

To address the second issue (the prediction of how stimulations and inhibitions will affect the overall activity of the STP), we need a causal model of the variables related to an STP. A number of studies[14,15,22–24] have shown that STPs can be modeled as causal Bayesian networks (BNs) if each node in the network represents the phosphorylation activity of a protein. A strength of BNs is that they represent probabilistic relationships, and therefore they can manage the noise in biological data. A second strength is that they can model the natural causal relationships in biology.

On the one hand, protein phosphorylation assays are slow, relatively expensive, and can be performed for only a tiny but important fraction of the genome. On the other hand, the gene expression level data are widely available because they are inexpensive and genome wide. As noted previously, methods have already been developed that account for the topology of an STP when analyzing gene expression data to determine whether the STP is implicated in a cancer.[19–21] However, the correlation of gene expression with activity is not well established. Studies show that the protein expression level (abundance) is often not positively correlated with activity[25] and that the gene expression level is often not correlated with protein abundance.[26] Hence, the gene expression levels might be at most loosely correlated with the activity, which means that the causal structure of an STP might not be represented by the relationships among the gene expression levels. More fundamentally, it is an open question as to whether there are causal relationships among the expression levels of genes coding for proteins on an STP.

We investigated this question. Specifically, the central hypothesis to be investigated in this paper is that the expression levels of genes that code for proteins on a given STP are causally related, and that this causal structure is altered when the STP is involved in a particular cancer. If this hypothesis is correct, using the ample gene expression datasets and BN learning algorithms, we can learn the causal network structure of the gene expression levels in an STP that is altered in a given cancer, and then identify driver genes based on the topology of the network.

The Cancer Genome Atlas (TCGA) makes available a breast cancer dataset that contains data on SNPs and the expression levels of 17,814 genes. There are 529 cases and 61 controls for which this information is available. Using these datasets and BN technology, we investigate the causal structure of genes that code for proteins on 5 STPs believed to be associated with breast cancer, 7 STPs believed to be associated with other cancers, and 10 randomly chosen pathways. We obtain significant results indicating that the causal structure of the STPs, which are believed to be implicated in both breast cancer and all cancer, is more altered in the cases relative to the controls than the causal structure of the randomly chosen pathways.

Method

As our method applies BNs to modeling STPs, we first review both of these.

BNs. BNs[27–29] are increasingly being used for uncertain reasoning and machine learning. A BN consists of a directed acyclic graph (DAG) $G = (V, E)$ whose nodeset V contains random variables and whose edges E represent relationships among the random variables, the prior probability distribution of every root variable in the DAG and the conditional probability distribution of every non-root variable given each set of values of its parents. Often the DAG is a causal DAG, which is a DAG containing the edge $X \rightarrow Y$ only if X is a direct cause of Y.[29] The probability distribution of the variables in a BN must satisfy the *Markov condition*, which states that each variable in the network is probabilistically independent of its non-descendents conditional on its parents.

Figure 1 shows a BN representing the causal relationships among variables related to lung disorders. In this BN, h_1, for example, denotes an individual has a smoking history and h_2 denotes the individual does not. Using this BN, we can determine conditional probabilities of interest with a BN inference algorithm.[29]

A BN DAG model consists of a DAG $G = (V, E)$, where V is a set of random variables, and a parameter set θ whose members determine conditional probability distributions for G, but without numerical assignments to the parameters. The task

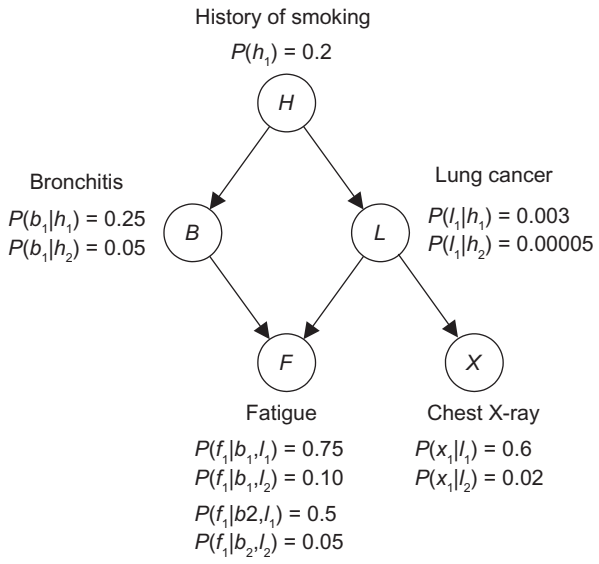
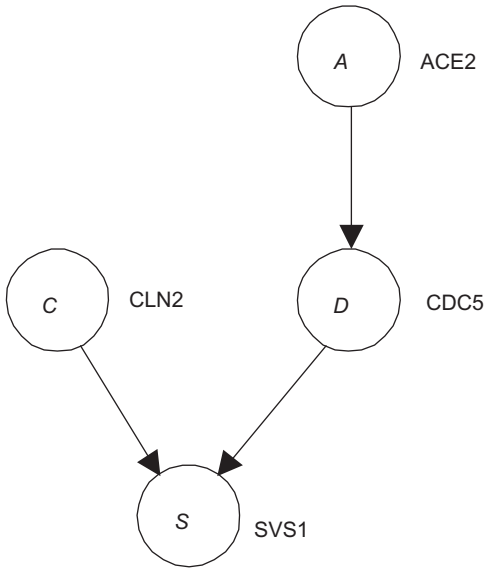


Figure 1. A BN representing a subset of the variables related to lung disorders. There is an edge from node *A* to node *B* if *A* has a direct causal influence on *B*.



$P(s|C = \text{low}, D = \text{low}) = \text{NormalDen}(s; 0.6, 0.1)$

$P(s|C = \text{low}, D = \text{high}) = \text{NormalDen}(s; 1.3, 0.3)$

$P(s|C = \text{high}, D = \text{low}) = \text{NormalDen}(s; 1.1, 0.2)$

$P(s|C = \text{high}, D = \text{high}) = \text{NormalDen}(s; 1.7, 0.4)$

Figure 2. A BN for a small gene regulatory network (based on a figure in Ref 33). Only the conditional probability distribution for node *S* is shown. Each variable is continuously distributed, and defined to be "high" if its value is higher than 1 and "low" if its value is less than 1. The notation $\rho(s|C = \text{low}, D = \text{low}) = \text{NormalDen}(s;0.6,0.1)$ means *S* is normally distributed with mean 0.6 and standard deviation 0.1 if *C* and *D* are both "low."

of learning a BN DAG model from the data is called *model selection*.

In the constraint-based approach,[30] we learn a DAG model from the conditional independencies that the data suggest are present in the generative probability distribution *P*. In the score-based approach, we assign a score to a DAG based on how well the DAG fits the data. The *Bayesian score* is the probability of the Data given the DAG model.[31] A popular variant of this score is the Bayesian Dirichlet equivalent uniform (BDeu) score.[32] If the set of variables in model *G* is $\{X_1, X_2, \ldots, X_n\}$, this score is as follows:

$$P(\text{Data}\,|\,G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha/q_i)}{\Gamma(\alpha/q_i + \sum_{k=1}^{r_i} s_{ijk})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha/(r_i q_i) + s_{ijk})}{\Gamma(\alpha/(r_i q_i))}, \quad (1)$$

where α is a parameter called the prior equivalent sample size, r_i is the number of states of X_i, q_i is the number of different instantiations of the parents of X_i, and s_{ijk} is the number of times in the data that X_i took its kth value when the parents of X_i had their jth instantiation.

Many biological processes have been modeled using BNs including molecular phylogenetics,[33] gene regulatory networks,[34–36] genetic linkage,[37] genetic epistasis,[38–42] and STPs.[14,15,22–24] Figure 2 shows a BN representing a small gene regulatory network.

STPs modeled as BNs. An STP is a network of intercellular information flow initiated when extracellular signaling molecules bind to cell-surface receptors. The signaling molecules become modified, causing a change in their functional capability and affecting a change in the subsequent molecules in the network. This cascading process culminates in a cellular response. Consensus STPs have been developed based on the composite of studies concerning individual STP components. Figure 3 shows part of the consensus STP of human primary naive CD4 T cells, downstream from CD3, CD28, and LFA-1 activation. Kyoto Encyclopedia of Genes and Genomes (KEGG)[43] has a collection of manually drawn pathways representing our knowledge of about 136 pathways. STPs are not thought to be stand-alone networks, but rather they have inter-pathway communication.[44]

If we represent the phosphorylation level of each protein in an STP by a random variable and draw an arc from *X* to *Y* if there is an edge from protein *X* to protein *Y* in the STP, then we are modeling the STP as a BN. For this BN to represent the joint probability distribution of the random variables, the Markov condition must be satisfied. Woolf et al.[22] argue that the steady-state concentrations should satisfy this condition. For example, in Figure 3 the phosphorylation activity of MEK1/2 should be dependent on the phosphorylation activity of PKA because high PKA activity implies high RAF activity, which in turn implies high MEK1/2 activity. However, once we know the activity of RAF, the implication link is broken, which is what the Markov condition entails. Sachs et al.[14] performed a proof of principle study concerning this conjecture, and confirmed this. A number of other papers[15,23,24] successfully modeled STPs as BNs using phosphorylation activity.

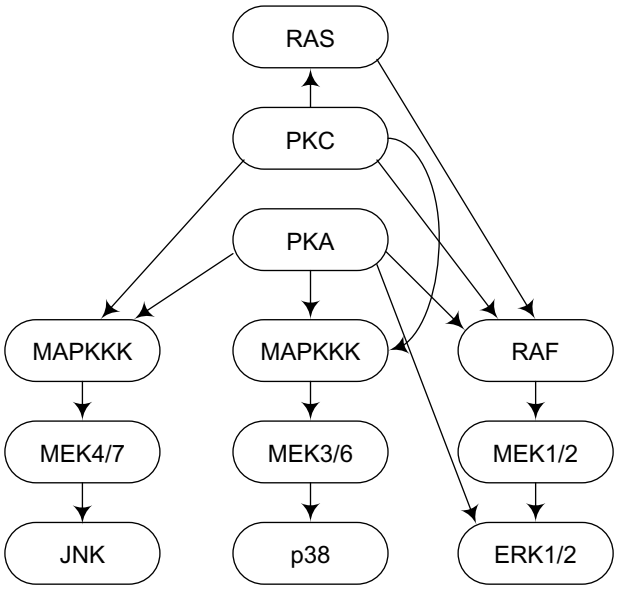


Figure 3. A portion of the consensus STP of human primary naive CD4 T cells, downstream from CD3, CD28, and LFA-1 activation. Arcs are used to illustrate connections between signaling molecules. In some cases, the connections may be indirect and may involve specific phosphorylation sites of the signaling molecules. MAPKKK appears twice because MEK4/7 and MEK3/6 each have a MAPKKK that is its activator. This figure is based on a figure in Ref. 14; see that paper for more details.

As discussed in the Introduction, gene expression level seems to be at most loosely correlated with activity. So, if there are causal relationships among the expression levels of genes coding for proteins on an STP, the BN representing these relationships may not represent the biological flow of an STP. This means it would be difficult to learn STPs from the gene expression levels. However, if our goal is to investigate how variables concerning known STPs are modified in tumors, not to learn the structure of unknown STPs, then the causal structure of the gene expression levels in tumors can provide us with important information. As also mentioned in the Introduction, it is an open question as to whether there are even causal relationships among the expression levels of genes coding for proteins on an STP. This paper investigates this question.

Identifying aberrant STPs using BNs and gene expression level data. In what follows, for simplicity we will say that a gene coding for a protein on an STP is on the STP itself. We assume that we have two sets of data. The first set contains the gene expression levels of all (at least most) genes in a set of cases (tumors) and the second set contains the gene expression levels of all genes in a set of controls. Let STPX be an STP we are investigating, Data_1 be the data concerning the cases for genes on STPX, and Data_2 be the data concerning controls for genes on STPX.

There are two models. Model M_A represents that the same causal structure (BN) is generating both Data_1 and Data_2. In this case, the two datasets can be considered as coming from the same population and are therefore combined. Model M_B represents that two different causal structures (BNs) are generating the data. We compute the log Bayes factor of model M_B relative to model M_A as follows. We first compute

$$P(\text{Data} \mid M_A) = \sum_G P(\text{Data}_1, \text{Data}_2 \mid G) P(G \mid M_A)$$
$$= \frac{1}{m} \sum_G P(\text{Data}_1, \text{Data}_2 \mid G),$$

$$P(\text{Data} \mid M_B) = \left(\sum_G P(\text{Data}_1 \mid G) P(G \mid M_B) \right) \left(\sum_G P(\text{Data}_2 \mid G) P(G \mid M_B) \right)$$
$$= \frac{1}{m^2} \left(\sum_G P(\text{Data}_1 \mid G) \right) \left(\sum_G P(\text{Data}_2 \mid G) \right),$$

where m is the number of possible DAG models containing the variables. In these computations, we are summing over all the DAG models G according to the law of total probability (model averaging) and are assuming that all the DAG models are equiprobable. The likelihoods are computed using the BDeu score (Equation 1). As there is an intractable number of models, we do approximate model averaging using Markov chain Monte Carlo (MCMC) as described in Ref. 29. Next we compute the Bayes factor K as follows:

$$K = \ln \left(\frac{P(\text{Data} \mid M_B)}{P(\text{Data} \mid M_A)} \right). \tag{2}$$

An alternative method would be to approximately learn the most likely DAG model G_1 based on Data_1, the most likely DAG model G_2 based on Data_2, and the most likely DAG model G_3 based on Data_1 and Data_2. Then take the maximum likelihood estimates $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ of the parameters in G_1, G_2, and G_3, and compute the log likelihood ratio L as follows:

$$L = \ln \left(\frac{P(\text{Data}_1 \mid G_1, \hat{\theta}_1) P(\text{Data}_2 \mid G_2, \hat{\theta}_2)}{P(\text{Data}_1, \text{Data}_2 \mid G_3, \hat{\theta}_3)} \right). \tag{3}$$

The larger the value of K or L, the more the data indicate that the causal structure of STPX is altered in the tumorous tissue. The advantage of using the Bayes factor is that it automatically includes a penalty for model complexity. However, it is costly to compute. In our investigations, we approximate the Bayes factor by approximately learning the most likely model and then using the Bayesian information criteria (BIC) to approximate the probability of the data given the model. In the limit, the BIC and the BDeu score (Equation 1) choose the same model.[29]

Evaluation Methodology

It is difficult to assess a pathway analysis model or methodology using real data because the ground truth is not known. In the absence of a gold standard, we can perform our analysis based on the existing biological knowledge. Hence, to investigate whether the causal structure of the expression levels of genes on an STP is altered when the STP is involved in cancer, we compared results obtained using the breast cancer data for 5 STPs implicated in breast cancer, 7 STPs implicated in other cancers, and 10 random pathways. We investigated STPs implicated in other cancers because it is believed that there are commonalities across tumor lineages.[45] The pathways investigated are listed in Table 1. The first column lists the five STPs believed to be implicated in breast cancer. The PI3K pathway is one of the most important pathways in cancer metabolism in general, and has recognized as an important target in breast cancer management for years.[46] Hyperactive Wnt signaling has been shown to contribute to cancer in a wide range of human tissue, and Wnt genes have been identified as oncogenes in mouse mammary tumorigenesis.[47] Over-expression of the *ErbB2* gene occurs in approximately 20% of breast cancers.[48] The Hedgehog[49] and Notch[50] pathways have also been associated with breast cancer, but perhaps less strongly. The second column of Table 1 lists the seven STPs implicated in other cancers, and the last column shows the randomly chosen pathways. We did not investigate the glioma, pancreatic, and colorectal cancer pathways because of their overlap with the PI3K and Wnt pathways. The 10 non-cancerous pathways were obtained by first eliminating the 12 cancer pathways investigated and then randomly choosing 10 pathways from all the pathways in the KEGG database.

Table 1. Pathways investigated.

BREAST CANCER	OTHER CANCER	OTHER PATHWAYS
PI3K	Viral Carcinogenesis	3M Syndrome Ubiquitin Mediated Proteolysis
Wnt	Small Cell Lung Cancer	Salivary Secretion
ERbB	ChronicMyeloid Leukemia	Barth Syndrome Glycerophospholipid Metabolism
Hedgehog	MalignantMelanoma	Dent Disease_Inositol Phosphate Metabolism
Notch	Non-Small Cell Lung Cancer	Alpha-1-Antitrypsin (A1 AT) Deficiency Complement and Coagulation Cascade
	Bladder Cancer	N-Glycan Biosynthesis
	Thyroid Cancer	Viral Myocarditis
		Gallbladder Disease ABC Transporters
		Type II Diabetes Mellitus
		Type I Diabetes Mellitus

The cancer genome atlas (TCGA) makes available a breast cancer dataset that contains data on SNPs and the expression levels of 17,814 genes. There are 529 cases and 61 controls for which this information is available. Using the KEGG database, we identified all the genes related to each of the 22 pathways. We extracted gene expression profiles for the 529 breast cancer patients and 61 controls in the TCGA database. By mapping the gene names of the genes in the gene sets identified using the KEGG pathways and the gene names in the TCGA data, we were able to extract the gene expression profiles for each of the 22 pathways for the 529 patients and 61 controls. All the expression levels were discretized to values *low*, *medium*, and *high* using the equal width discretization technique, which discretizes the data into partitions of K equally sized intervals ($K = 3$ in our application). Using the resultant datasets, we computed the approximate Bayes factor for each of the 22 pathways. We used the BN learning package HUGIN[51] to approximately learn the most probable DAG models, and to calculate the BICs.

All experiments were run using a Dell PowerEdge R515, which has two AMD Opteron™ 4276HE, 2.6 GHz, 8C, Turbo CORE, 8M L2/8M L3, 1600 MHz Max Mem single processors.

Results

Table 2 lists the pathways, along with their Bayes factors, in a decreasing order. It is notable that PI3K, which is "probably one of the most important pathways in cancer metabolism and growth,"[52] scored much higher than all other pathways. The Wnt and ErbB pathways are also near the top of the list. However, the Notch and Hedgehog pathways are not. In general, however, the cancer-related pathways are concentrated at the top of the list. Figure 4 shows the average Bayes factor and standard error for each of the three categories.

Table 3 shows the P-values, obtained using the nonparametric Mann–Whitney test, comparing the 5 breast cancer pathways to the 10 randomly chosen pathways (listed as "other pathways") and comparing all 12 cancer pathways to the 10 randomly chosen pathways. In both cases, the results were significant with the respective P-values of 0.049 and 0.040. These significant values were obtained even though the Hedgehog and Notch pathways, which scored in the bottom half of the list, were included in the cancer sets. However, as mentioned previously, we do not have absolute ground-truth STPs. Perhaps, these pathways are not substantially implicated in breast cancer, which is what our results suggest.

The possibility exists that these significant results were obtained simply because the genes are over or under expressed in cancer-related STPs and the causal structure is not relevant. To test this possibility, we redid the study with all the BNs constrained to having no causal edges. Table 3 shows the resultant P-values which are significant. Figure 5 shows

Table 2. Bayes factors for 22 pathways. There is an "X" if the pathway is implicated in breast cancer or any cancer.

PATHWAY	BREAST CANCER	CANCER	BAYES FACTOR
PI3K	X	X	3343
Viral Carcinogenesis		X	1884
Wnt	X	X	1485
3M Syndrome Ubiquitin Mediated Proteolysis			1262
Small Cell Lung Cancer		X	1034
ERbB	X	X	912
MalignantMelanoma		X	892
Salivary Secretion			873
ChronicMyeloid Leukemia		X	806
Barth Syndrome Glycerophospholipid Metabolism			778
Alpha-1-Antitrypsin (A1 AT) Deficiency Complement and Coagulation Cascade			682
Non-Small Cell Lung Cancer		X	623
Dent Disease_Inositol Phosphate Metabolism			607
Notch	X	X	574
Viral Myocarditis			498
Gallbladder Disease ABC Transporters			443
Hedgehog	X	X	404
Bladder Cancer		X	392
Thyroid Cancer		X	390
Type II Diabetes Mellitus			306
Type I Diabetes Mellitus			143
N-Glycan Biosynthesis			–87

Table 3. *P*-values obtained with causal modeling and without causal modeling.

ALTERNATE HYPOTHESIS	CAUSAL MODELING	NO CAUSAL MODELING
Breast cancer > other	0.049	0.291
All cancer > other	0.040	0.244

the average Bayes factor and standard error for each of the three categories when there are no causal edges. Note that the ranges overlap more than those in Figure 4, which is obtained when causation is modeled. These results support that there is an underlying causal structure among the expression levels of genes on an STP and that this causal structure is altered when an STP is involved in cancer.

All networks learned are fairly complex. As an example, Figure 6 shows the network learned from cases for the ErbB pathway.

Discussion

We analyzed 5 STPs associated with breast cancer, 7 STPs associated with other cancers, and 10 randomly chosen pathways. Based on modeling the relationships among the expression levels of genes on the pathways as causal BNs, we obtained results indicating that the causal structure of the cancer-related STPs is significantly more altered in breast cancer tissue than the randomly chosen pathways. These results support that the expression levels of genes on STPs are causally related and that this causal structure is altered in the tumorous tissue when an STP is involved in cancer.

These results are significant for a number of reasons. First, we can use the methodology to develop a method for investigating whether an STP is involved in cancer, which can be compared to the existing methods.[30,31,53] Second, these results open up a promising area of future research involving

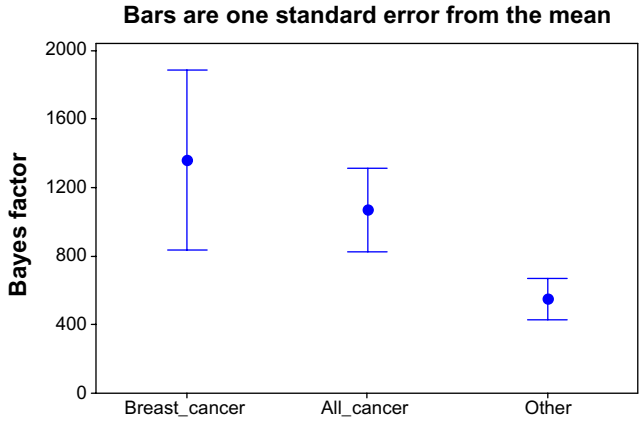


Figure 4. Average Bayes factors and standard error for breast cancer pathways, all cancer pathways, and other pathways, when causation is modeled.

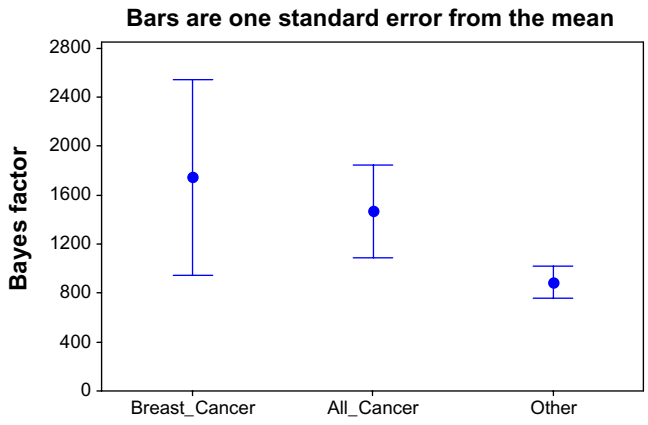


Figure 5. Average Bayes factors and standard error for breast cancer pathways, all cancer pathways, and other pathways, when causation is not modeled.

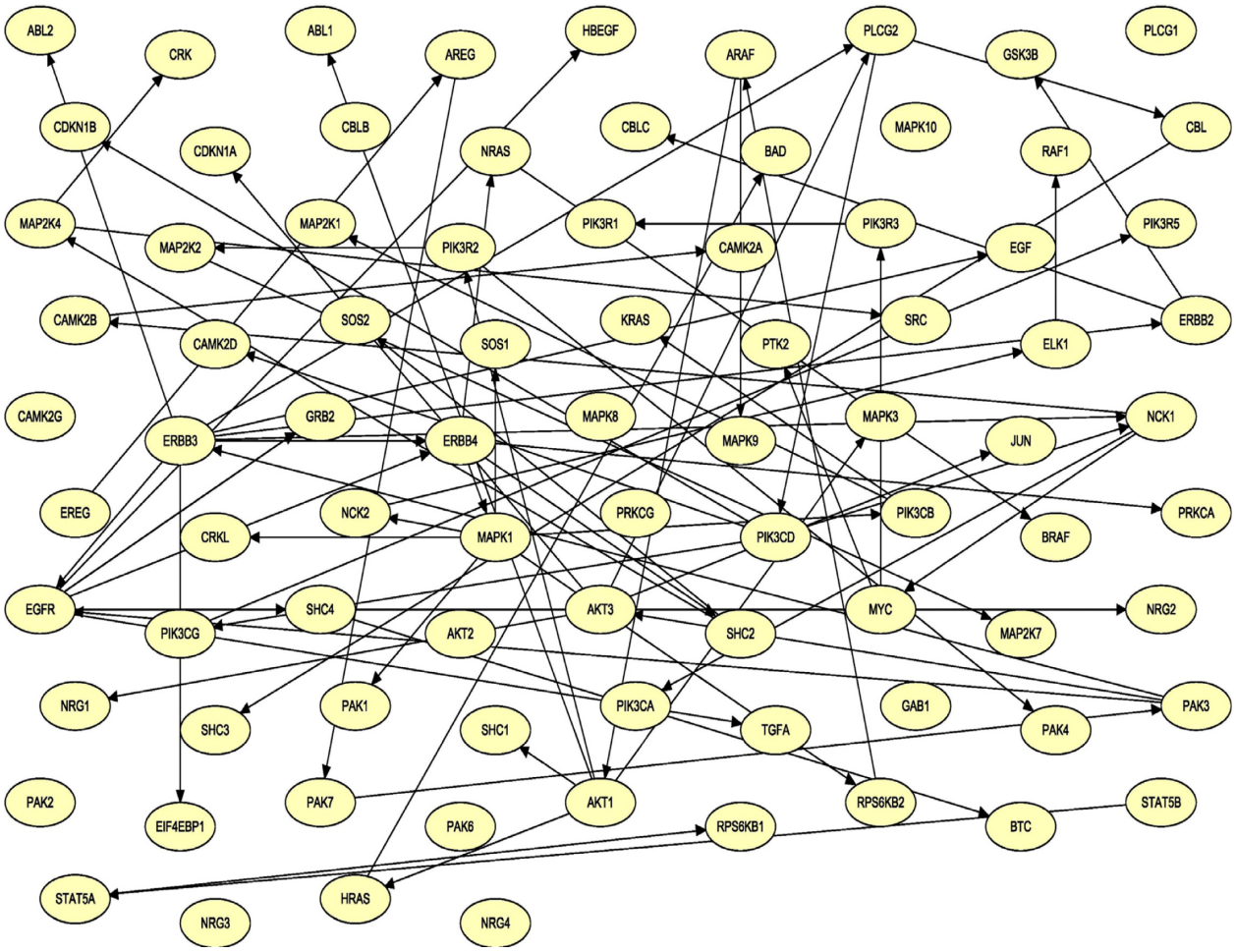


Figure 6. The causal BN learned from breast cancer cases for the ErbB pathway.

the use of BN technology to model the causal relationships among the expression levels of genes on an STP. Using such a network, we can learn possible driver genes, and the effect of genetic variants on these driver genes and therefore on the network. Such investigations would enable us to better identify therapeutic targets in a patient-specific fashion.

In future research, we can implement the Bayes factor calculation (Equation 1), and see if it yields better results than the approximation used in the given studies. Furthermore, we can develop and implement a method that better learns the causal edges among the genes in the STP. Rather than just learning a single highly likely model using a package like HUGIN, we can do approximate model averaging to learn the strength of the edges. Finally, we can develop and test an entire BN that contains both expression levels and genetic causes of expression levels.

Conclusion

We conclude that our study supports that the relationships among the expression levels of genes on an STP can be modeled using a causal BN, and that this network is altered in the tumorous tissue. This result opens up new avenues for identifying driver genes on STPs.

Author Contributions

XJ conceived and designed the experiments. DX processed the data, developed the datasets representing the pathways, and analyzed the data. RN wrote the first draft of the manuscript. XJ contributed to the writing of the manuscript. RN and XJ jointly developed the structure and arguments for the paper. All authors reviewed and approved the final manuscript.

DISCLOSURES AND ETHICS

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

REFERENCES

1. Vogel C, et al. First-line, single-agent herceptin® (trastuzumab) in metastatic breast cancer, a preliminary report. *Eur J Cancer*. 2001;37:25–9.
2. Park JW, et al. Unraveling the biologic and clinical complexities of HER2. *Clin Breast Cancer*. 2008;8:392–401.
3. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature*. 2012;490:61–70.

4. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*. 2012;22(2):398–406.
5. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res*. 2011;22(2):1–12.
6. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol*. 2011;18(3):507–22.
7. Zhao J, Zhang S, Wu LY, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*. 2012;28(22):2940–7.
8. Jebar AH, Hurst CD, Tomlinson DC, Johnston C, Taylor CF, Knowles MA. FGFR3 and Ras gene mutations are mutually exclusive genetic events in urothelial cell carcinoma. *Oncogene*. 2005;24(33):5218–25.
9. Kurose K, Gilley K, Matsumoto S, Watson PH, Zhou XP, Eng C. Frequent somatic mutations in PTEN and TP53 are mutually exclusive in the stroma of breast carcinomas. *Nat Genet*. 2002;32(3):355–7.
10. Xing M, Cohen Y, Mambo E, Tallini G, Udelsman R, Ladenson PW, et al. Early occurrence of RASSF1 A hypermethylation and its mutual exclusion with BRAF mutation in thyroid tumorigenesis. *Cancer Res*. 2004;64(5):1664–8.
11. Ding et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 2008;455:1069–75.
12. Curtis et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
13. The Cancer Genome Atlas. http://cancergenome.nih.gov/.
14. Sachs K, et al. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*. 2005;308:523–9.
15. Sachs K. Bayesian network models of biological signaling pathways. MIT PhD Thesis; 2006. http://hdl.handle.net/1721.1/38865.
16. Drăghici S, et al. Global functional profiling of gene expression. *Genomics*. 2003;81:98–104.
17. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102:15545–50.
18. Tian L, et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA*. 2005;102:13544–9.
19. Tarca A, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25:75–82.
20. Vaske CJ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomic data using PARADIGM. *Bioinformatics*. 2010;26:i237–45.
21. Ng et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*. 2012;21(18):640–6.
22. Woolf P, et al. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics*. 2005;21:741–53.
23. Pe'er D. Bayesian network analysis of signaling networks: a primer. *Sci STKE*. 2005;2005(281):pl4.
24. Sachs K, et al. Bayesian network approach to cell signal pathway modeling. *Sci STKE*. 2002;148:pe38.
25. Tsigankov P, Gherardini PF, Helmer-Citterich M, Späth GF, Zilberstein D. Phosphoproteomic analysis of differentiating Leishmania parasites reveals a unique stage-specific phosphorylation motif. *J Proteome Res*. 2013;12(7):3405–12.
26. Chen G, Gharib TG, Huang CC. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics*. 2002;1(4):304–13.
27. Pearl J. *Probabilistic Reasoning in Intelligent Systems*. Burlington, MA: Morgan Kaufmann; 1988.
28. Neapolitan RE. *Probabilistic Reasoning in Expert Systems*. New York, NY: Wiley; 1989.
29. Neapolitan RE. *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall; 2004.
30. Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd ed. New York: Springer-Verlag; 1993. Boston, MA: MIT Press; 2000.
31. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn*. 1992;9:309–47.
32. Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. *Mach Learn*. 1995;20(3):197–243.
33. Neapolitan RE. Probabilistic reasoning in bioinformatics. Burlington, MA: Morgan Kaufmann; 2009.
34. Segal E, Pe'er D, Regev A, Koller D, Friedman N. Learning module networks. *J Mach Learn Res*. 2005;6:557–88.
35. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3–4):601–20.
36. Friedman N, Koller K. Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks. *Mach Learn*. 2003;50: 95–125.
37. Fishelson M, Geiger D. Optimizing exact genetic linkage computation. *J Comput Biol*. 2004;11:114–21.
38. Jiang X, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. *Genet Epidemiol*. 2010;34(6):575–81.
39. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics*. 2011;12(89):1471–2105.
40. Jiang X, Barmada MM, Cooper GF, Becich MJ. A Bayesian method for evaluating and discovering disease loci associations. *PLoS One*. 2011;6(8):e22075.
41. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S, Cooper GF. A fast algorithm for learning epistatic genomics relationships. In: Proceedings of American Medical Informatics Association (AMIA) Annual Fall Symposium. 2010.
42. Jiang X, Neapolitan RE. Mining strict epistatic interactions from high-dimensional datasets: ameliorating the curse of dimensionality. *PLoS One*. 2012;7(10):e46771.
43. KEGG PATHWAY. http://www.genome.jp/kegg/pathway.html.
44. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*. 2001;2:343–72.
45. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20.
46. Baselga J. Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer. *Oncologist*. 2011;16:12–9.
47. Howe L, Brown A. Wnt signaling and breast cancer. *Cancer Biol Ther*. 2004;3(1): 36–41.
48. http://www.bethyl.com/content/bulletin_ErbB2/.
49. Hui M, Cazet A, Nair R, et al. The Hedgehog signaling pathway in breast development, carcinogenesis and cancer therapy. *Breast Cancer Res*. 2013;15:203.
50. Al-Hussaini H, Subramanyam D, Reedijk M, et al. Notch signaling pathway as a therapeutic target in breast cancer. *Mol Cancer Ther*. 2010;10(1):9–15.
51. http://www.hugin.com/.
52. Baselga J. Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer. *Oncologist*. 2011;16(Suppl 1):12–9.
53. Kjaerulff UB, Madsen AL. *Bayesian Networks and Influence Diagrams*. New York, NY: Springer; 2010.