Novel Application of Junction Trees to the Interpretation of Epigenetic Differences among Lung Cancer Subtypes

Arturo Lopez Pineda, MS[1], and Vanathi Gopalakrishnan, PhD[1]
[1]The PRoBE Lab, Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA

Abstract

In this era of precision medicine, understanding the epigenetic differences in lung cancer subtypes could lead to personalized therapies by possibly reversing these alterations. Traditional methods for analyzing microarray data rely on the use of known pathways. We propose a novel workflow, called Junction trees to Knowledge (J2K) framework, for creating interpretable graphical representations that can be derived directly from *in silico* analysis of microarray data. Our workflow has three steps, preprocessing (discretization and feature selection), construction of a Bayesian network and, its subsequent transformation into a Junction tree. We used data from the Cancer Genome Atlas to perform preliminary analyses of this J2K framework. We found relevant cliques of methylated sites that are junctions of the network along with potential methylation biomarkers in the lung cancer pathogenesis.

Introduction and Background

Lung cancer is the leading cause of human cancer death in the United States, with estimated yearly casualties of over 160,000 [1]. Among all lung cancers the most frequent subtypes are adenocarcinoma (ADC) and squamous cell carcinoma (SCC), accounting for 38.5% and 20% of all cases respectively [2]. Identifying molecular differences between these two subtypes is important to enable clinicians to select patients who will likely benefit from a given drug regimen, and also in selecting those patients who will avoid toxicity from the treatment [3]. It has been suggested that the ADC and SCC develop through distinct pathogenetic pathways, resulting in epigenetic alterations [4].

DNA methylation is an epigenetic alteration that creates molecular changes to the environment of the DNA. This alteration occurs when a methyl group is attached to a specific location of the DNA, typically in sites where the sequence cytosine-phosphate-guanine (CpG) is abundant. This epigenetic alteration has the effect of silencing gene transcription, potentially removing important functions in the protein pathways. Recent studies have found that DNA methylation can have an effect on the progression [5], and recurrence [6], of the cancer into a more aggressive form.

Distinct DNA methylation signatures between ADC and SCC have been found in studies targeting candidate genes in lung cancer [7]. However, there is still a need for understanding the mechanisms of DNA methylation for future epigenetic therapies, such as reversal of DNA methylation. This therapy has shown promising results using a technique called active demethylation that promotes DNA repair [8].

Traditional methods for analysis of DNA methylation microarray technology often investigate open research questions such as: How do differentially methylated sites interact among one another? Is there a network topology that might help discover differences between lung cancer subtypes? Balbin et al. [9] recently proposed a framework for the reconstruction of gene network topology in lung cancer combining multiple 'omic' data and then identifying functional networks associated with them. This framework relies on the identification of differentially expressed pathways from the 'omic' data. To achieve this task they use SPIA [10], an algorithm that combines the evidence of enrichment analysis (log-fold change differential expression) from the 'omic' data and the perturbation it has on the known pathways in KEGG [11]. Martini et al. [12] proposed an algorithm for the reconstruction of relevant network topologies. They start with known pathways found in KEGG which they transform into Junction trees[13] for human interpretation. Pradhan et al. [14] used microarray data to select the differentially expressed genes and then used the known interactions in the BioGrid [15] platform to reconstruct the network topology.

This paper describes a novel workflow, called Junction trees to Knowledge (J2K), for performing *in silico* analysis of DNA methylation datasets. It addresses the post-classification problem of characterizing biomarkers that are responsible for disease classification at the sub-clinical level. It makes use of Bayesian networks (BNs) [16], which traditionally have been used in other domains to perform probabilistic inference; and Junction trees (JTs) [13], which have been vastly applied to propagate belief over a network and compute exact posterior probabilities [17].

While there are many computational algorithms that can assist in the creation of JTs, their application to modeling 'omic' data is relatively new. To our knowledge, the use of junction tree representation to simplify the BN has not been explored adequately with biomedical data. While the literature supports their use for efficient identification of proteins from tandem mass spectra [18], it is unclear whether a representation that is created for purely computational efficiency can also provide biologically relevant results. We believe that our research will help us understand this aspect better.

Materials and Methods

Our workflow, as shown in Figure 1, first discretizes the features in the data using MDLPC [19], and selects those that best distinguish the target class via feature selection with the ReliefF [20] algorithm. Then it builds a BN using EBMC [21], and finally it transforms the directed network into a JT [13]. The remaining parts of this section describe these algorithms, including the in-house developed JT creation algorithms.

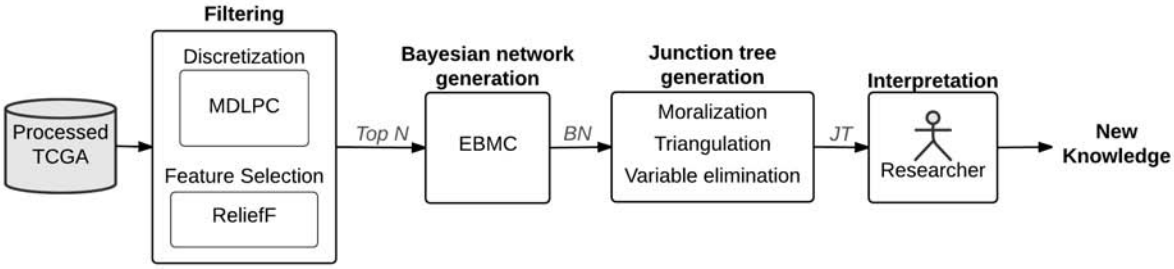

Figure 1. Empirical workflow of TCGA data to directed graph (BN) to undirected graph (JT) to Knowledge (J2K)

Dataset

The Cancer Genome Atlas (TCGA) is a public repository of genomic data supervised by the National Cancer Institute (NCI) that aims to characterize human cancers. We extracted DNA methylation intensity profiles for 197 tumor samples (65 ADC and 132 SCC) from the TCGA data portal for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC, [22]). The microarray platform used for analysis was the Illumina® Infinium HumanMethylation 27k (27,578 variables representing methylation sites).

Data Preprocessing

Discretization. The methylation intensity for each methylation site in the microarray data is a continuous value. We partitioned this value into intervals using the Fayyad and Irani's Minimum Description Length Principle Cut (MDLPC) [19]. This algorithm selects a cut point that minimizes the joint entropy of the resulting subintervals. It then continues to partition recursively until no cut point can be selected. The resulting set of intervals constitutes the values for that methylation site. Features refer to variable-value pairs.

Feature Selection. A subset of the most informative features from the microarray data were selected using the ReliefF feature selection algorithm [20]. This is a multivariate filter that sequentially evaluates every instance to estimate how well a feature can distinguish the target class given the instances in the same neighborhood. In the end, the top scoring features are retrieved. The ordering from this selection is also used in subsequent elements of our workflow.

Bayesian Networks

A Bayesian network [16] is a probabilistic graphical model that explains a given set of discrete data. It comprises a set of nodes (methylation sites) that represent random variables and a set of arcs among the nodes that represent probabilistic dependence. The posterior probability of an event (disease) D occurring, given that an event (symptom) S is observed is calculated using the well-known Bayes' formula [23].

The *Efficient Bayesian Multivariate Classification (EBMC)* [21] is a recent method for learning a BN from data. It uses a greedy search to find a constrained BN that best predicts a target node, similar to the Bayesian Rule Learning (BRL) algorithm [24]. It initially starts with an empty model and then it identifies a set of nodes that are parents of the target and predicts it well. EBMC then transforms the temporary network structure into a statistically equivalent one (Augmented Naïve Bayes or ANB [25]) where the parents of the target become children of the target with arcs

among them. It then iterates the whole process until no set of parents can be added to the target node to improve the prequential score [26]. EBMC has been shown to perform well at binary outcome prediction using high-dimensional discrete datasets [27]. This method differs from other BN-learning algorithms in at least two ways: (a) no ordering of features is required and (b) the representation of the learned BN is an augmented Naïve Bayes structure.

Junction Trees

A Junction tree [13] is a tree-structured undirected graph, whose nodes correspond to cliques of features (or methylated sites in our case), and whose links connect pairs of cliques that have features in common. A clique is a subset of nodes in an undirected graph where any two nodes are connected by an edge. In order to create a JT, three steps are needed:

1. Moralization. Starting from a directed graph, such as a BN, the directionality of the edge is removed by connecting or 'marrying' the set of nodes that share common children but do not have direct edges between them. This yields an undirected moral graph.
2. Triangulation. In the undirected moral graph, all cycles containing four or more nodes must be triangulated. This process involves iteratively adding extra edges to eliminate such cycles of four or more nodes (chord-less cycles). There are an exponential number of triangulation possibilities, depending on the number of nodes involved in the cycle, which have been solved by using a predefined ordering of nodes to be triangulated. Triangulation is an NP-hard problem. We order the nodes for this triangulation process based on the ReliefF scoring to make the algorithm efficient. The triangulated graph containing cliques is used for node elimination, which creates a JT.
3. Node Elimination. A new tree-structured undirected graph (empty JT) is first constructed by following the node elimination algorithm. Then, a node is selected for elimination, and its containing clique is added to the JT. Next, the node and its incident edges are eliminated from the triangulated graph and the process is repeated until no other nodes are available. Node elimination is also an NP-hard problem. Hence, a predefined ordering of nodes has to be used. The JT must satisfy a property called the 'junction property' (*running intersection property*), meaning that if a feature is contained in two cliques, then it must also be contained in every clique on the path that connects them. The order in which the nodes were eliminated is based on the ReliefF scoring.

Results and Discussion

The MDLPC [13] algorithm transforms all continuous methylation profiles into discrete binned data containing at least two bins for each methylation site. Since MDLPC is a supervised discretization method, it also has the side effect of discarding those features with only one bin. For the datasets that we used, the algorithm selected 7,908 features out of 27,579.

The ReliefF algorithm [20] ranks all features according to their impact in differentiating between the classes. To demonstrate proof of concept, we selected the top ranked 30 methylated sites to create BNs. The area under the receiver operating characteristic curve (AUC) when evaluating these BNs over a stratified 10-fold cross-validation is 0.988. This level of classification performance is reasonable for this problem, because the two subtypes of lung cancer are known to be fairly distinct.

The structure of the BN created using the EBMC algorithms applied to the data is shown in Figure 2. As expected, there are multiple connections between the target node (class node) and the rest of the selected nodes. There are two fairly distinct clusters of 10 interconnected methylated sites (see Figure 2). The network topology produced by EBMC is an augmented naïve Bayes (ANB) structure. The corresponding gene IDs to where those methylation sites are located is used after this step (i.e. methylation site cg18515587 is located in gene SELENBP1).

As a conceptual method, we proposed the use of JTs as way of clarifying the structure of BNs. In Figure 3, we show the corresponding JT representation of the BNs seen in Figure 2. In the EBMC-derived JT representation we can start to think of new hypothesis with a greater biological relevance. For example, looking at the central clique (central circle) it is easy to see that there are key molecules that are worthwhile examining carefully, because a perturbation in this clique would have an impact on the entire structure. The central clique has the following genes: SELENBP1, AKR1B10, LPP, SH3BP1, BCDIN3, KRTHB4, MAPK8, TCF2. We used the suite NextBio® to test the association of this clique to different tissues and diseases in the known literature and curated studies. This suite finds that the central clique is associated with the epithelial cells of nasal turbinates, and the epithelial cells of bronchial large airways, and that it also correlates with esophageal cancer cell line OE21. The clique (and specially SELENBP1) is associated to the Selenium binding protein which is considered to be a tumor suppressor and a prognostic marker [28].

The BN shown in this manuscript is a Bayesian network classifier (BNC) created using the EBMC learning algorithm. The resulting structure of this BN improves the classification of the target node but does not to capture all the probabilistic relationships between features. We plan to use other BN learning algorithms to further explore this novel research area.

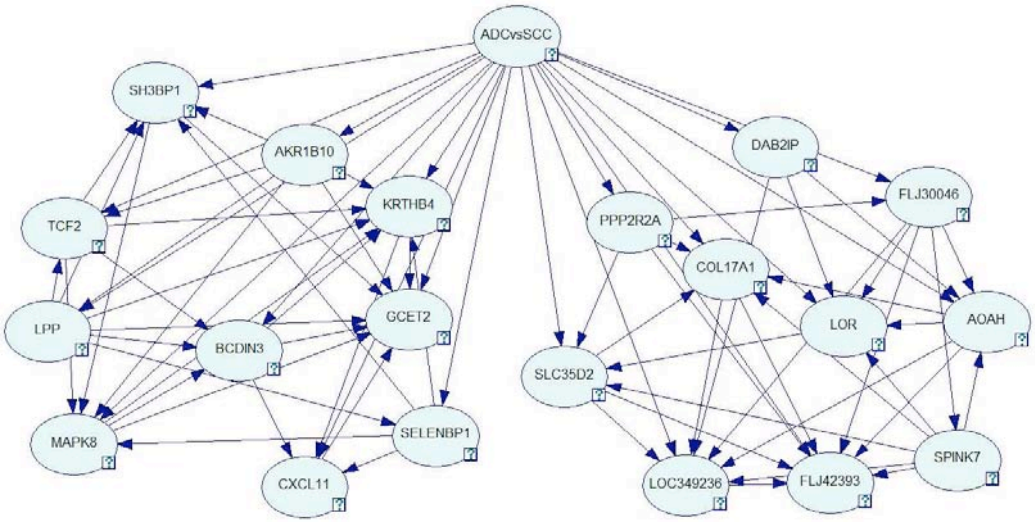

Figure 2. EBMC-generated BN model for the classification task ADC_{tumor} vs SCC_{tumor}.

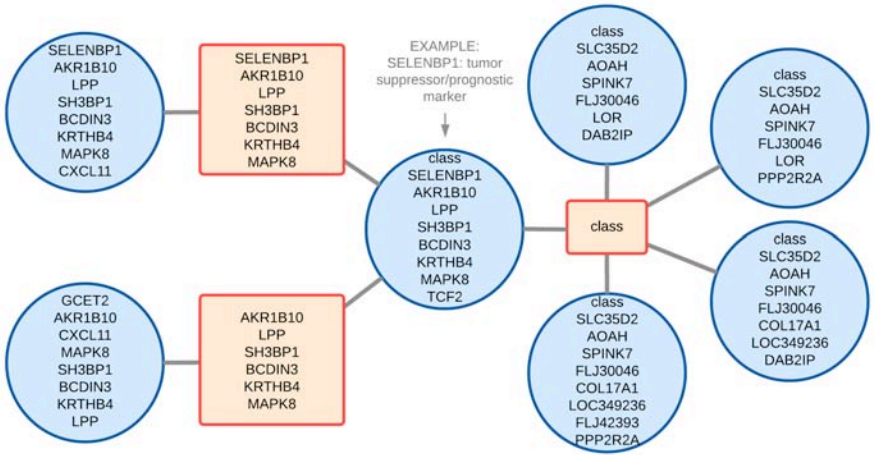

Figure 3. EBMC-derived JT, where the squares represent junctions while the circles represent cliques. An example is provided to show the importance of the JT to identify central cliques with important genes.

Conclusion

In this research, we have tested a novel concept called J2K framework to transform biological data from directed to undirected graphs via the application of JT generation algorithms. We applied a series of algorithms for transforming epigenomic data of lung cancer into a graphical representation that is interpretable for human researchers. Our study can easily be generalized into other types of epigenomic and genomic data, and we plan on testing it with other biomedical datasets. Particularly, we have found cliques of methylated sites that are of interest for the differentiation of lung cancer subtypes.

Acknowledgments: We thank Dr. Gregory F. Cooper for providing the EBMC java code.

Grant support: The research reported in this publication was supported by the following grants from the National Institutes of Health: National Cancer Institute Award Number P50CA90440, National Library of Medicine Award Number R01LM010950, and National Institute of General Medical Sciences Award Number R01GM100387. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1 Cancer statistics, 2012. 2012;**62**:10–29. doi:10.3322/caac.20138

2 Cruz Dela CS, Tanoue LT, Matthay RA. Lung Cancer: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine* 2011;**32**:605–44. doi:10.1016/j.ccm.2011.09.001

3 Langer CJ, Besse B, Gualberto A, *et al.* The evolving role of histology in the management of advanced non-small-cell lung cancer. *J Clin Oncol* 2010;**28**:5311–20. doi:10.1200/JCO.2010.28.8126

4 Lockwood WW, Wilson IM, Coe BP, *et al.* Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One* 2012;**7**:e37775–5. doi:10.1371/journal.pone.0037775

5 Towle R, Truong D, Hogg K, *et al.* Global analysis of DNA methylation changes during progression of oral cancer. *Oral Oncol* 2013;**49**:1033–42. doi:10.1016/j.oraloncology.2013.08.005

6 Sato T, Arai E, Kohno T, *et al.* DNA methylation profiles at precancerous stages associated with recurrence of lung adenocarcinoma. *PLoS One* 2013;**8**:e59444–4. doi:10.1371/journal.pone.0059444

7 Rauch TA, Wang Z, Wu X, *et al.* DNA methylation biomarkers for lung cancer. *Tumor Biol* 2012;**33**:287–96. doi:10.1007/s13277-011-0282-2

8 Barreto G, Schäfer A, Marhold J, *et al.* Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature* 2007;**445**:671–5. doi:10.1038/nature05515

9 Balbin OA, Prensner JR, Sahu A, *et al.* Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat Commun* 2013;**4**:2617–7. doi:10.1038/ncomms3617

10 Tarca AL, Draghici S, Khatri P, *et al.* A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**:75–82. doi:10.1093/bioinformatics/btn577

11 Ogata H, Goto S, Sato K, *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;**27**:29–34. doi:10.1093/nar/27.1.29

12 Martini P, Sales G, Massa MS, *et al.* Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res* 2013;**41**:e19–9. doi:10.1093/nar/gks866

13 Lauritzen SL, Spiegelhalter DJ. *Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems*. 1988. doi:10.2307/2345762

14 Pradhan MP, Desai A, Palakal MJ. Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst Biol* 2013;**7**:141–1. doi:10.1186/1752-0509-7-141

15 Stark C, Breitkreutz B-J, Reguly T, *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9. doi:10.1093/nar/gkj109

16 Neapolitan RE. *Probabilistic Reasoning in Expert Systems*. 2012.

17 Serang O. The probabilistic convolution tree: efficient exact Bayesian inference for faster LC-MS/MS protein inference. *PLoS One* 2014;**9**:e91507–7. doi:10.1371/journal.pone.0091507

18 Serang O, Noble WS. Faster Mass Spectrometry-Based Protein Inference: Junction Trees Are More Efficient than Sampling and Marginalization by Enumeration. *IEEE/ACM Trans Comput Biol and Bioinf*;**9**:809–17. doi:10.1109/TCBB.2012.26

19 Fayyad U, Irani K. BEACON eSpace at Jet Propulsion Laboratory: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. 1993.

20 Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence* 1997;**7**:39–55. doi:10.1023/A:1008280620621

21 Cooper GF, Hennings-Yeomans P, Visweswaran S, *et al.* An efficient bayesian method for predicting clinical outcomes from genome-wide data. *AMIA Annu Symp Proc* 2010;**2010**:127–31.

22 The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;**489**:519–25. doi:10.1038/nature11404

23 Daly R, Shen Q, Aitken S. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review* 2011;**26**:99–157. doi:10.1017/S0269888910000251

24 Gopalakrishnan V, Lustgarten JL, Visweswaran S, *et al.* Bayesian rule learning for biomedical data mining. *Bioinformatics* 2010;**26**:668–75. doi:10.1093/bioinformatics/btq005

25 Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Mach Learn* 1997;**29**:131–63. doi:10.1023/A:1007465528199

26 Kontkanen P, Myllymäki P, Silander T, *et al. On supervised selection of Bayesian networks*. Morgan Kaufmann Publishers Inc. 1999.

27 Jiang X, Cai B, Xue D, *et al.* A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. *J Am Med Inform Assoc* 2014;**21**:e312–9. doi:10.1136/amiajnl-2013-002358

28 Yang W, Diamond AM. Selenium-binding protein 1 as a tumor suppressor and a prognostic indicator of clinical outcome. *Biomark Res* 2013;**1**:15–5. doi:10.1186/2050-7771-1-15