**A Novel Method for Bayesian Network Analysis of the**

**Influence of Genetic Factors on Diseases**

**Neena Parikh1, Amin Zollanvari, PhD2, Gil Alterovitz, PhD1, 2, 3**

**1Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA; 2Children’s Hospital Informatics Program, Boston, MA; 3Division of Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Boston, MA**

**ABSTRACT**

***Motivation:*** *Several studies have been done that integrate several genome-wide experiments pertaining to a single disease. What this project aims to achieve is an automated method for constructing accurate Bayesian networks based on publicly available findings pertaining to any disease.*

***Results:*** *An automated pipeline was successfully constructed. Integrative models were made based on gene expression data obtained from GEO experiments relating to four different diseases using Bayesian statistical methods. Many of these models demonstrated a high level of accuracy and predictive ability. The approach described in this paper can be applied to any complex disorder and can include any number and type of genome-wide experiments.*

**INTRODUCTION**

In recent years, several genome-wide technologies have been developed. As the amount of publicly available genomic data increases, the necessity for varying bioinformatics methodologies to conduct relevant analysis also increases. Much has been written about the obstacles presented by translating discoveries made via genomic data to medicine1, 2, 3, 4. In order to circumvent these obstacles, the discipline of translational informatics has emerged – a discipline that focuses on the development of analytic and interpretive methods to evaluate the increasing amounts of biological data into diagnostics and therapeutics for the clinical environment5. Many researchers have used an integrative approach to isolate genes that are associated with certain diseases. For example, such research has been widely done with respect to type I diabetes6 and obesity7. These studies, among others, have demonstrated the efficacy of utilizing data from several genetic and microarray studies.

**Gene Expression Omnibus**

The Gene Expression Omnibus (GEO) is a publicly accessible repository available through the World Wide Web8. This project was initiated in response to an increasing demand for a public database of high-throughput gene expression data. The GEO offers a flexible platform for submission and retrieval of heterogeneous data sets from high-throughput gene expression and genomic hybridization experiments. The GEO separates primary (user-submitted) data into three main categories: *platforms*, *samples*, and *series*. Basically, a platform is a list of probes that define which molecules may be detected; a platform may reference many samples that have been submitted by multiple users. A sample describes the conditions under which it was handled, the manipulations it underwent, and the abundance measurement of each element derived form it; each sample references a single platform that is used to generate its molecular data. A series record organizes samples into the meaningful sets of data that make up an experiment, providing a focal point and description of the whole study.

Selected primary records undergo an upper-level of rendering into *DataSet* and *Profile* records. A Dataset record represents a curated collection of biologically and statistically comparable GEO samples. DataSets are created by GEO staff from reassembled series records. Profiles are then derived from DataSets, consisting of the expression measurements for an individual gene across all samples in a DataSet. In this study, all gene expression data was obtained from DataSet records.9

**Bayesian Networks and Multi-nets**

In a Bayesian approach to statistical analysis, the data analysis process begins with an already established probability distribution, referred to as the *prior distribution*. This process consists of using previously obtained sample data to update the prior distribution into a *posterior distribution*. The basic tool for this process is the Bayes’ theorem.10

When dealing with complex problems, graphical models can help break down the complex systems into simpler parts, allowing for the analysis of a single variable. A Bayesian network is a directed, acyclic graphical structure. The two main components of a Bayesian network are called *nodes* and *edges*; nodes are representative of objects that have states (for example, a phenotype), and edges correspond to the conditional probability distribution of the end node (known as a *child*) on the source node (known as a *parent*). Every node in the network, given the states of all of its parents, is statistically independent of all other non-descendant nodes. A simple Bayesian network is depicted in Figure 1.

Using a Bayesian network as a tool for statistical analysis produces several advantages. For instance, such a network features a lesser element of complexity in that it represents only a subset of dependencies rather than the complete joint probability distribution. Additionally, a Bayesian network has the capability to integrate a variety of factors into a single model due to the use of conditional dependencies. Additionally, a Bayesian network results in a much stronger predictive value than does other comparable analysis methods.10

In the research conducted prior to the writing of this paper, a certain type of Bayesian model known as the Bayesian multinet was constructed. The multinet was introduced in a companion paper11. Rather than being a simple Bayesian network as described above and portrayed in Figure 1, a Bayesian multinet is a set of distinct yet related Bayesian networks12. In this model, individuals are separated into unique classes; each network is created from the relevant information relating to a single class of multiple individuals. The aim of using a multinet structure as opposed to a simple Bayesian model is to create a more accurate and precise model of the dependencies between certain genes for different groups of subject.

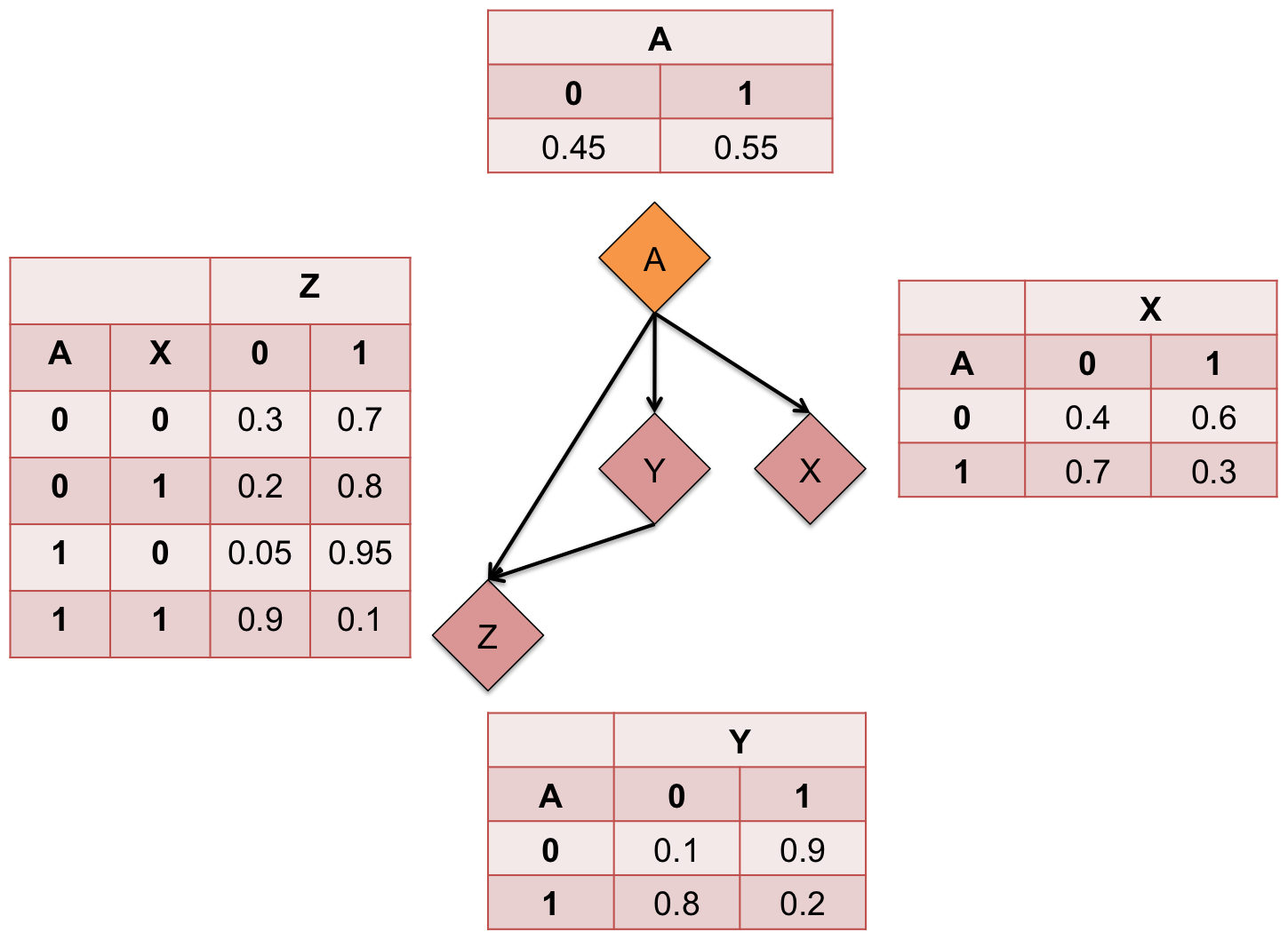
**Goals of Research**

Several studies have been done that merge gene expression data from multiple experiments with respect to a single disease. Though some of these studies have indeed used an automated approach to such research, they have focused on single diseases or disorders.

This study aims to construct an automated framework that allows for the integration of genome-wide expression data (using the GEO) in regard to any disease or disorder, while also creating a predictive model for disease-related phenotypes that illustrates the relationship between various genetic factors and pathways in order to aid future treatment and study of these diseases.

**Diseases**

For this paper, the following disorders were studied: Huntington’s Disease, Obesity, Leukemia, and Lymphoma. These four diseases were selected based on the number and content of their respective related experiments in the GEO.

****

**Figure 1:** Example of a Bayesian network. In this example, A is the founder; X, Y, and Z are the children of A; A and Y are the parents of Z.

**METHODS**

**Downloading experiments from the GEO and disease mapping**

In order to create a mapping of all GEO DataSet files to diseases, a process similar a study conducted at the Stanford University School of Medicine5 was used. Several methods were presented in this paper; the GENOTEXT system13, 14 was initially explored as a possibility for creating the DataSet-disease mappings. However, it was eventually decided that a related but simpler method would be used for the purposes of this study.

All available GEO DataSet files were downloaded and read in order to obtain relevant information. Each experiment was mapped to disease(s) using their corresponding PubMed identification numbers and Medical Subject Heading terms.

**Merging experiments and obtaining expression data**

After generating this list of diseases and their associated GEO DataSets, a “pipeline” program in the R language was used to merge the GEO experiments and obtain all relevant gene expression data. The user first selects the disease to consider (i.e., Lymphoma), and using the disease-DataSet mapping list, looks up the relevant DataSet file names. This list of DataSet file names is fed into the program in R, which automatically generates two text files that can be used to create the multinets.

Firstly, of the DataSet experiments fed into the program, one needs to determine which of the experiments are “interesting” – for instance, in order to be relevant to the purpose of this study, an experiment must contain data comparing control versus non-control subjects in a manner that relates to the phenotype associated with the disease. After obtaining all “interesting” experiments, these experiments must be merged in some way. As gene expression values are unit-less and relative, each experiment’s data must be normalized with respect to all other experiments; this was done by selecting one experiment as the “reference” experiment and adjusting all other experiment’s values relative to the reference experiment. Subsequently, all samples from all experiments were merged according to control versus non-control samples.

The top differentially expressed genes were found for each experiment individually via a number of steps. Firstly, a linear model was fit to each gene by using the design matrix of the microarray experiment, with rows corresponding to arrays and columns to coefficients to be estimated. Using this linear model, moderated t-statistics were computed by empirical Bayes shrinkage of the standard error toward a common value. This method was used to rank genes in order of evidence for differential expression. Finally, a table of the top-ranked genes from the linear model fit was extracted.16 In this experiment, several trials were conducted; each one included a different number of top-ranked genes that were extracted.

The genes that were studied were the ones that intersected across all experiments.

**Constructing multinets and receiver operating characteristic (ROC) curves**

The multinet classifiers for the separate classes of data were then generated. The Bayesian multinet classifier was validated using 10-fold cross-validation17. In essence, the samples are divided into ten approximately equally sized subsets. In repeated simulations, nine of these subsets are used for training and the tenth is used for testing. The area under the ROC curve (AUROC) value of the multinet’s cross-validation procedure indicates the model’s accuracy.

The Receiver Operating Characteristic (ROC) is a curve that relates the true positive to the false positive rate for different thresholds. The AUROC is a statistical measure of robustness; the closer the AUROC is to 1, the closer the true positive ration is to 1, and the more accurate the predictive model. AUROC is a more comprehensive statistical measure of robustness than just predictive accuracy18.

**RESULTS**

Several trials were done with each disease, varying the number of top differentially expressed genes that were examined in each experiment. For each trial, the AUROC values were observed as a measure of accuracy. The genes that were determined to be the most prominent were also investigated in more detail. As a basis for comparison, several trials were also conducted with single experiments rather than a network of several experiments used to create the predictive model. The results from each experiment are shown in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Disease** | **# Experiments** | **% Top Genes** | | **# Common Genes** | **AUROC** |
| Huntington's Disease | 4 | | 15% | 16 | 0.75 |
| Obesity | 12 | | 35% | 21 | 0.783 |
|  | 10 | | 35% | 41 | 0.7926 |
| Leukemia | 12 | | 25% | 13 | 0.7736 |
| Lymphoma | 6 | | 15% | 27 | 0.8669 |

**Table 1:** Summary of results from each disease.

**DISCUSSION**

For each experiment, it was found that often times the most prominent genes were not present among the very top proportion of differentially expressed genes, but rather more toward the middle of the spectrum. For instance, none of the 21 intersecting genes among the top 35% of differentially expressed genes across all twelve GEO experiments relating to Obesity were present in the gold standard Obesity Gene Map list19; however, when observing the 180 intersecting genes among the top 18000 genes from each experiment, there were six genes also present on the gold standard list. Such a finding indicates that certain genes may have been overlooked in the past when studying these diseases; further research may focus on the effect and prominence of these seemingly “less” differentially expressed genes.

With respect to the genetic factors related to Huntington’s Disease, the automated pipeline found there to be several genes that have already been studied in the context of the disease (*RALA, CBX5, CALM3, GLG1, GLUL, MAPK8IP1, IMMT, MAP3K8, CDKN1A*)20-26; however, certain genes were also discovered that have not been researched in regard to Huntington’s Disease. It has been shown that some of these genes (*SCT, LMNB1, IVD*)27-30 have some relation to certain neurodegenerative diseases or are involved in pathways that are relevant to the development of Huntington’s Disease. Although the study for which this paper was written did not look into such genes into great detail, these findings demonstrate that the methods explored in this project can be used for discovery of novel disease-related genes.

The model created to represent genetic factors relating to leukemia also resulted in similar findings – several genes in this model have already been studied in varying degrees with respect to leukemia (*WT1, PDE4DIP, NCAM1, AKAP13, SLC35E1, HFE, JUN*)31-40; however, a few novel genes were also presented in the model (*IVD, SYNJ2, TTLL3*). Again, such findings demonstrate the power of this project’s methods in discovery of novel disease-specific genes.

**Contributions**

New methods that this project explored:

1. Automated (rather than manual) mapping of GEO experiments to specific diseases
2. A structure for automated construction of Bayesian networks with respect to analysis of the influence of genetic factors on specific diseases

New findings that this project discovered:

1. The accuracy of predictive models is greater when constructed via an integrative approach rather than a single-experiment approach.
2. Certain genes that may not be as greatly differentially expressed may still hold as much prominence as those genes at the top of the list.
3. The framework presented in this paper can be used for discovery of novel disease-related genes.

**ACKNOWLEDGEMENTS**

The work was supported by the Boston Children’s Hospital and the Harvard – MIT Division of Health Sciences and Technology.

**REFERENCES**

1. Snyderman R. The clinical researcher – an ‘emerging’ species. Jama, 2004, 291(7): 882-883.

2. Lenfant C. Shattuck lecture – clinical research to clinical practice – lost in translation? New England Journal of Medicine, 2003, 349(9): 868-874.

3. Rees J. Complex disease and the new clinical sciences. Science, 2002, 296(5568): 698-700.

4. Schwartz K, Vilquin JT. Building the new translational highway: toward new partnerships between academia and the private sector. Nature Medicine, 2003, 9(5): 493-495.

5. Butte AJ, Chen R. Finding Disease-Related Genomic Experiments Within an International Repository: First Steps in Translational Bioinformatics. AMIA Symposium Proceedings, 2006: 106-110.

6. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, et al. Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. Genome Research, 2002, 12, 232-243.

7. English SB, Butte AJ. Evaluation and Integration of 49 Genome-wide Experiments and the Prediction of Previously Unknown Obesity-related Genes. Bioinformatics Advance Access, Oct 2007.

8. Wheeler DL, Church DM, Edgar R, Federhen S, Hemlberg W, Madden TL, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res., 2004, 1(32, Database issue): D35-40.

9. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 2002, 30 (1): 207-210.

10. Alterovitz G, Liu J, Afkhami, Ramoni MF. Bayesian methods for proteomics. Proteomics, 2007, 7.

11. Zollanvari A, Wu A, Alterovitz G. Multinet Bayesian Networks for Integrative Genomic Discovery: Application to Genetic Epistatic Interactions in HIV. Unpublished (2011).

12. Zollanvari A, Merlob A, Huynh K, Dreyfuss JM, Ramoni RB, McGeachie MM, et al. Multi- level Interactions Underlie Complex Behavior: Prognosis of Nicotine Dependence in a Holistic Bayesian Framework. Unpublished (2011).

13. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. National Biotechnology, 2006, 24(1): 55-62.

14. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 2004, 1(32, Database issue): D267-270.

15. Davis S. Using the GEOquery package. Genetics Branch, National Cancer Institute, National Institutes of Health, 2010. Accessed Jul 2011, <http://www.bioconductor.org/>.

16. Smyth GK. Lineral models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 2004, 3(1): Article 3.

17. Braga-Neto U, Dougherty E. Exact performance of error estimators for discrete classifiers. Pattern Recognition, 2005, 38: 1799–1814.

18. Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. Critical Care, 2004, 8 (6): 508–512.

19. Rankinen R, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, et al. The Human Obesity Gene Map: The 2005 Update. Obesity, 2006, 14: 529-644.

20. Strand AD, Aragaki AK, Shaw D, Bird T, Holton J, Turner C, et al. Gene expression in Huntington’s disease skeletal muscle: a potential biomarker. Human Molecular Genetics, 2005, 14(13): 1863-1878.

21. Gene Expression Atlas – Summary for CBX5. European Bioinformatics Institute, 2011. <http://www.ebi.ac.uk>

22. Rakhilin SV, Olson PA, Nishi A, Starkova NN, Fienberg AA, Nairn AC, et al. A Network of Control Mediated by Regulator of Calcium/Calmodulin-Dependent Signaling. Science, 2004, 306(5696): 698-701.

23. Willmroth F, Beaudet AL. Structure of the murine E-selectin ligand 1 (ESL-1) gene and assignment to Chromosome 8. Mammalian Genome, 1999, 10: 1085-1088.

24. Jablonski MM, Freeman NE, Orr WE, Templeton JP, Lu L, Williams RW et al. Genetic Pathways Regulating Glutamate Levels in Retinal Müller Cells. Neurochem Res., 2011, 36(4): 594, 603.

25. Behrens PF, Franz P, Woodman B, Lindenberg KS, Landwehrmeyer GB. Impaired glutamate transport and glutamate-glutamine cycling: downstream effects of the Huntington mutation. Brain, 2002, 125(8): 1908-1922.

26. Boxall R, Porteous DJ, Thomson PA. DISC1 and Huntington’s Disease – Overlapping Pathways of Vulnerability to Neurological Disorder? PLoS One, 2011, 6(1): e16263.

27. Whitmore TE, Holloway JL, Lofton-Day CE, Maurer MF, Chen L, Quinton TJ, et al. Human secretin (SCT): gene structure, chromosome location, and distribution of mRNA. Cytogenet Cell Genet., 2000, 90(1-2): 47-52.

28. Schuster J, Sundblom J, Thuresson AC, Hassin-Baer S, Klopstock T, Dichgans M, et al. Genomic duplications mediate overexpression of lamin B1 in adult-onset autosomal dominant leukodystrophy (ADLD) with autonomic symptoms. Neurogenetics, 2011, 12(1): 65-72.

29. Hendrickson SL, Lautenberger JA, Chinn LW, Malasky M, Sezgin E, Kingsley LA, et al. Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression. PLoS One, 2010, 5(9): e12862.

30. Yoshida T, Kato K, Yokoi K, Oguri M, Watanabe S, Metoki N, et al. Association of genetic variants with hemorrhagic stroke in Japanese individuals. International Journal of Molecular Medicine, 2010, 25(4): 649-656.

31. Spassov BV, Stoimenoc AS, Balatzenko GN, Genova ML, Peichev DB, Konstantinov SM. Wilms’ tumor protein and FLT3-internal tandem duplication expression in patients with de novo acute myeloid leukemia. Hematology, 2011, 16(1): 37-42.

32. Ishikawa Y, Kiyoi H, Naoe T. Prevalence and clinical characteristics of N-terminally truncated WT1 expression in acute myeloid leukemia. Leukemia Res., 2011, 35(5): 685-688.

33. Huret JL. Leukemia Section: Short Communication. Atlas of Genetics and Cytogenetics in Oncology and Hematology, 2006.

34. Wattanawaraporn R, Singhsilarak T, Nuchprayoon I, Mutirangura A. Hypermethylation of TTC12 gene in acute lymphoblastic leukemia. Leukemia, 2007, 21: 2370-2373.

35. Raponi M, Harousseau JL, Lancet JE, Löwenberg B, Stone R, Zhang Y, et al. Identification of Molecular Predictors of Response in a Study of Tipifarnib Treatment in Relapsed and Refractory Acute Myelogenous Leukemia. Clinical Cancer Research, 2007, 13: 2254-2260.

36. Paulsson K, Heidenblad M, Strömbeck B, Staaf J, Jönsson G, Borg A, et al. High-resolution genome-wide array-based comparative genome hybridization reveals cryptic chromosome changes in AML and MDS cases with trisomy 8 as the sole cytogenetic aberration. Leukemia, 2006, 20: 840-846.

37. Dorak MT, Burnett AK, Worwood M. HFE gene mutations in susceptibility to childhood leukemia: HuGE review. Genetic Medicine, 2005, 7(3): 159-168.

38. Viola A, Pagano L, Laudati D, D’Elia R, D’Amico MR, Ammirabile M, et al. HFE gene mutations in patients with acute leukemia. Leuk. Lymphoma, 2006, 47(11): 2331-2334.

39. Veneri D, Franchini M, Krampera M, de Matteis G, Solero P, Pizzolo G. Analysis of HFE and TFR2 gene mutations in patients with acute leukemia. Leukemia Res., 2005, 29(6): 661-664.

40. Kollmann K, Heller G, Ott RG, Scheicher R, Zebedin-Brandl E, Schneckenleithner C, et al. c-JUN promotes BCR-ABL-induced lymphoid leukemia by inhibiting methylation of the 5’ region of Cdk6. Blood, 2011, 117(15): 4065-4075.