APPLICATIONS OF BAYESIAN NETWORK MODELS IN STUDYING ACUTE MYELOID LEUKEMIA (AML)

by

Rupesh Agrahari

A thesis submitted to the Graduate College of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Computer Science
May 2016

Committee Members:

Habil Zare, Chair

Byron J. Gao

Nihal Dharmasiri

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

I would like to dedicate this thesis to my family. A special gratitude to my parents, Suman Devi and Madan Prasad Gupta for their unconditional love and support. My late grandmother has always been a source of optimism and morality for me, and she will always guide me through her teachings.

I dedicate this work to my brothers Ritesh and Rahul and my sisters Rashmi and Ritika for all their love, support, and encouragement I needed for my success. I would like to thank all my family and friends for being a part of my life.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Habil Zare, for his patience, support, and encouragement throughout my graduate studies. He always suggested what was good for me and my career. He has been an adviser who listens to problems and gives solutions, no matter how small or big they were. I learned a lot by being a part of his *Oncinfo* lab, and I am sure that these experiences will help me shape my career. Habil collaborated with Dr. Aly Karsan to get the BCCA dataset, which was used as a validation data in this study. He wrote various parts of the code including the differential expression analysis of BCCA data, computation of eigengenes, projection of BCCA data on MILE data, and plotting the BNs. He also wrote the scripts used to run the complete project pipeline locally and on super computer clusters.

I would like to thank Dr. Amir Foroushani for his continuous help and guidance. His insightful comments and constructive criticisms were extremely crucial to my work and they helped me improve my approach for doing research. He gathered the MILE data from GEO repository that was used as an input to this study. Amir wrote various parts of the code including the code for finding of hub genes, plotting miller scores, processing data to be used for BN learning, learning BN using processed data, and finding the consensus network based on top third best scoring individual networks.

The input given by Habil and Amir were of crucial importance to this thesis. They guided and helped me to resolve the technical issues I faced during my work.

My thanks also go to the members of my thesis committee, Dr. Byron Gao and Dr. Nihal Dharmasiri for their feedback to my abstract and thesis document. Their feedback were immensely helpful and played important role in

my thesis writing. Nihal's feedback helped me improve my thesis in a way that a non-technical reader could understand it. Nihal's comments also helped me incorporate biological details related to my research work. Byron's comments focused the *machine learning* aspect of my work and that helped me improve my thesis by incorporating comparison between my study and other similar research works.

I would like to acknowledge Dr. Aly Karsan and Rod Docking for providing BCCA data for this study. We used BCCA data to validate our computational model. Dr. Karsan provided useful feedback at different stages of this work which helped us improve our strategy for research. Rod mapped the large RNA-Sequence database to a subset dataset (BCCA) for our study.

I acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this study. URL: http://www.tacc.utexas.edu

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

My thesis aims at designing a computational model to analyze gene expression data to improve cancer diagnosis, specifically Acute Myeloid Leukemia (AML), which is a type of aggressive blood cancer. As part of a team of researchers in the *Oncinfo Lab*, I used Bayesian networks (BN) to model gene expression data. A BN is a probabilistic graphical model where a set of random variables represent nodes of a Directed Acyclic Graph (DAG). The edges of the DAG model the conditional dependencies between the random variables. We used established clustering methods to cluster data and group similar genes together. Specifically, we applied Weighted Gene Co-Expression Network Analysis (WGCNA) as a clustering mechanism to cluster our gene expression data. For each cluster of genes, we used principal component analysis (PCA) to compute a single value, called an *eigengene*. Eigengenes were represented by nodes in the BN and dependency among those eigengenes were modeled by the edges of the BN. The rational for using a BN in this framework is that it can model gene expressions and dependencies, enabling us to use probability theory to make scientific predictions. The application of our BN model is to identify AML patients from another type of hematological malignancy. I performed the classification of patients using a cross-validation technique and tested the performance on an independent dataset. Moreover, I trained my model on a training dataset with 366 samples and evaluated the performance on a test dataset with 74 samples. The accuracy of predictions on train and test datasets were 93.5% and 84%, respectively. Further improvements to the methodology are required to improve its accuracy and make it appropriate for clinical use.

I. INTRODUCTION

Acute Myeloid Leukemia (AML) is a cancer of the myeloid line of blood cells in which bone marrow produces abnormal white blood cells, red blood cells, or platelets. AML is the most common acute leukemia affecting adults, and its incidence increases with age. It is a rare and aggressive type of blood cancer, accounting for about 1.2% of total cancer deaths in the USA (Jemal et al., 2002)

Myelodysplastic Syndromes (MDS) is a disease that affects bone marrow and blood. MDS is characterized by ineffective hematopoiesis, the ineffective production of blood cells and platelets in the bone marrow (Albitar et al., 2002). MDS is relatively mild and easily managed, but it can grow more severe over time and even turn into AML. MDS has a high risk of developing into AML, either gradually or rapidly (Shi et al., 2004). "Approximately 30% of patients with MDS will progress and develop into AML" (Wang et al., 2011). MDS can be argued to be preleukemia (Shi et al., 2004) or pre-AML but studies show that MDS is a discrete entity, that is different from AML, and thus cannot be simply said to be preleukemia (Albitar et al., 2002). This behavior makes it important to analyze and compare the two diseases to gain a better biological insight.

This study is inspired by, and builds upon the co-expression network analysis and Bayesian Network (BN). To perform co-expression network analysis, we used Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder & Horvath, 2008). WGCNA is a technique that can be used to perform various operations including data reduction, feature selection, clustering, and data exploration (Horvath, 2011). We used WGCNA, specifically, as a clustering mechanism that groups similar genes together into same groups (clusters) based on their coexpression values. The results of coexpression analysis are the gene modules that contain genes with similar expression.

We summarize the biological information of each gene module in one eigengene using Principal Component Analysis (PCA) (Jolliffe, 2002). PCA is a

statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Baker et al., 1993). We used these eigengenes to train a Bayesian network (BN) that could be used as a predictive model. BNs represent a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG)(Christofides & Theo-ry, 1975; Jensen, 1996).

A BN consists of a Directed Acyclic Graph (DAG) and either a set of conditional probability tables for discrete data or probability density functions for continuous data. The structure of a DAG is defined by two sets: the set of nodes (vertices) and the set of directed edges. The nodes represent random variables and are drawn as circles labeled by the variable names. The edges represent direct dependency among the variables and are drawn by arrows between nodes (Ben-Gal, 2007). In a DAG, if there is a directed edge coming from node 'X' to another node 'Y' then X is called a "parent node" of Y and Y becomes a "child node" of X. In this study, each node is an observed variable modeling the expression value of an eigengene. Figure I.1 shows the schematic diagram explaining the methodology for computing a BN model using the MILE study data.

It is important to determine how well our predictive model predicts the type of disease of a given patient. Cross validation (CV) is a model validation technique used to assess how the results of a statistical analysis will generalize to an independent data set (Refaeilzadeh et al., 2009; Kohavi et al., 1995). I used 5-fold Cross Validation (CV) that performs five rounds of CV to validate the model. After getting the individual predictions performed by the 5-fold CV, I took majority vote on the results to determine consensus predictive capability of the model. The consensus predictive model was used to perform prediction on train (MILE) and test (BCCA) datasets. We performed all our statistical analysis using R programming Language.

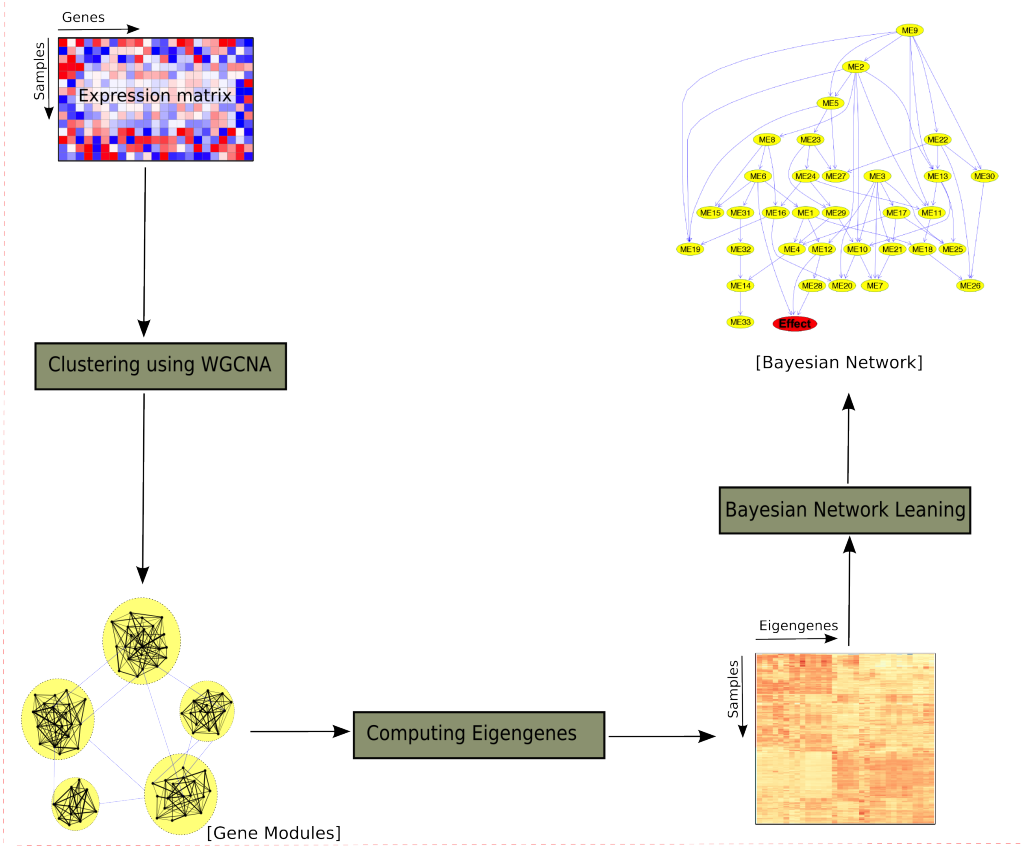


Figure I.1: **Schematic view of the methodology.** The input is the gene expression profile (matrix), gathered and processed from MILE study. We Applied WGCNA for clustering, to find the coexpression network built according to the correlation between gene pairs. We summarized the biological information of each module of genes into an eigengene vector. A BN is fitted to the eigengenes to delineate the relationships between modules.

II. METHODOLOGY

We downloaded expression profiles from Gene Expression Omnibus (GEO) repository (series accession number GSE15061)(Mills et al., 2009). The dataset is part of the expression MIcroarray analysis for diagnosis of LEukemia (MILE) series and consists of 202 AML, 164 MDS, and 69 non-leukemia samples. We selected only the AML and MDS samples for computing the predictive model. Apart from the MILE study dataset, we also have access to RNA sequencing (RNA-Seq) data from peripheral blood cells or bone marrow blast from British Columbia Cancer Agency (BCCA) (Ozsolak & Milos, 2011). The BCCA dataset contains 133 AML and 22 MDS samples for 51,019 transcripts of genes.

We have clinical data for both MILE and BCCA datasets. Clinical data contains information about the actual disease type of each of the samples. For BCCA samples, we also have information such as age, gender, and disease sub-type. In this study, clinical information was used as the gold standard to validate our BN model.

Pre-processing of Data

Data preprocessing is a data mining concept that involves eliminating unwanted information or noise from the input data, and transforming the data into a format that is more relevant and informative. We used R script to retrieve Differentially Expressed (DE) data from GEO repository by eliminating irrelevant samples and transforming it using logarithm base 2 to improve its interpretability. To refine the large dataset downloaded from GEO repository, we asked for the top 18,250 differentially expressed probes of genes corresponding to relatively large sample size of 202 AML and 164 MDS cases.

The downloaded dataset consisted of probes that mapped to one or more genes, which means that there was a many-to-many relationship among the

probes and genes. We filtered the 18,250 DE probes so that each of the probes were mapped to a single Entrez gene. We ended up with 13,294 probes mapping to 9,178 Entrez genes. In this processed dataset, each probe maps to exactly one gene, but for some genes, there can be several such probes. Althogh we reduced the size of data, its behavior and pattern was kept intact, which allowed us to study and perform our analysis easily and in a more manageable fashion.

We processed the BCCA dataset by keeping only the samples that had same disease type as MILE dataset samples. The processed BCCA dataset contains 52 AML and 22 MDS samples.

Clustering

Clustering is one of the most widely used unsupervised learning techniques for data mining (Jain et al., 1999). We used established clustering technique, Weighted Gene Co-expression Network Analysis (WGCNA), to identify gene modules in the data based on co-expression analysis. In this project, we applied WGCNA only on the AML samples of the MILE dataset. Genes with relatively higher correlation with other genes for the 202 AML samples were clustered together into one cluster (Langfelder & Horvath, 2008). A gene module is defined as a set of co-expressed genes to which the same set of transcription factors binds (Bar-Joseph et al., 2003). Here, gene modules are the clusters that contain a set of co-expressed genes. The co-expression among the genes is calculated using Pearson correlation (Benesty et al., 2009). Pearson correlation is a measure of linear correlation between two variables X and Y, giving a value between +1 and -1 inclusive, where +1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. It is widely used as a measure of the degree of linear dependence between two variables.

Finding interesting modules and genes

After clustering MILE data using WGCNA, we ended up with 33 gene modules (Figure III.2), which we wanted to use for learning our BN. Before moving forward, we also wanted to find the modules that were most interesting to us, which meant that we had to determine the modules and genes that show high dependency and/or are shown to be related with AML/MDS in various studies. This process involved two major steps.

Finding hub genes

A hub gene is a node/gene that has the highest intra-modular connectivity within a module. Since hub nodes have been found to play an important role in many networks, highly connected hub genes are expected to play an important role in biology as well (Langfelder et al., 2013). In this process, we used Pearson correlation (Benesty et al., 2009) to calculate intra-modular connectivity within the gene modules (i.e. the connectivity of nodes to other nodes within the same module).

Calculating miller scores

A way to identify modules that are interesting to our study is to find out how frequently genes within the module were reported by other studies. For this, we used data from Miller & Stamatoyannopoulos (2010) study that systematically surveyed 25 published reports of gene expression profiling in AML (Miller & Stamatoyannopoulos, 2010). We used this survey to score the modules based on their known association with AML. Figure II.1 shows the enrichment of modules in genes associated with AML that is reported in the Miller & Stamatoyannopoulos (2010) study. For instance, in the figure, the red bar reports the number and percentage of genes in each module that were reported to be related to AML in at least two studies.

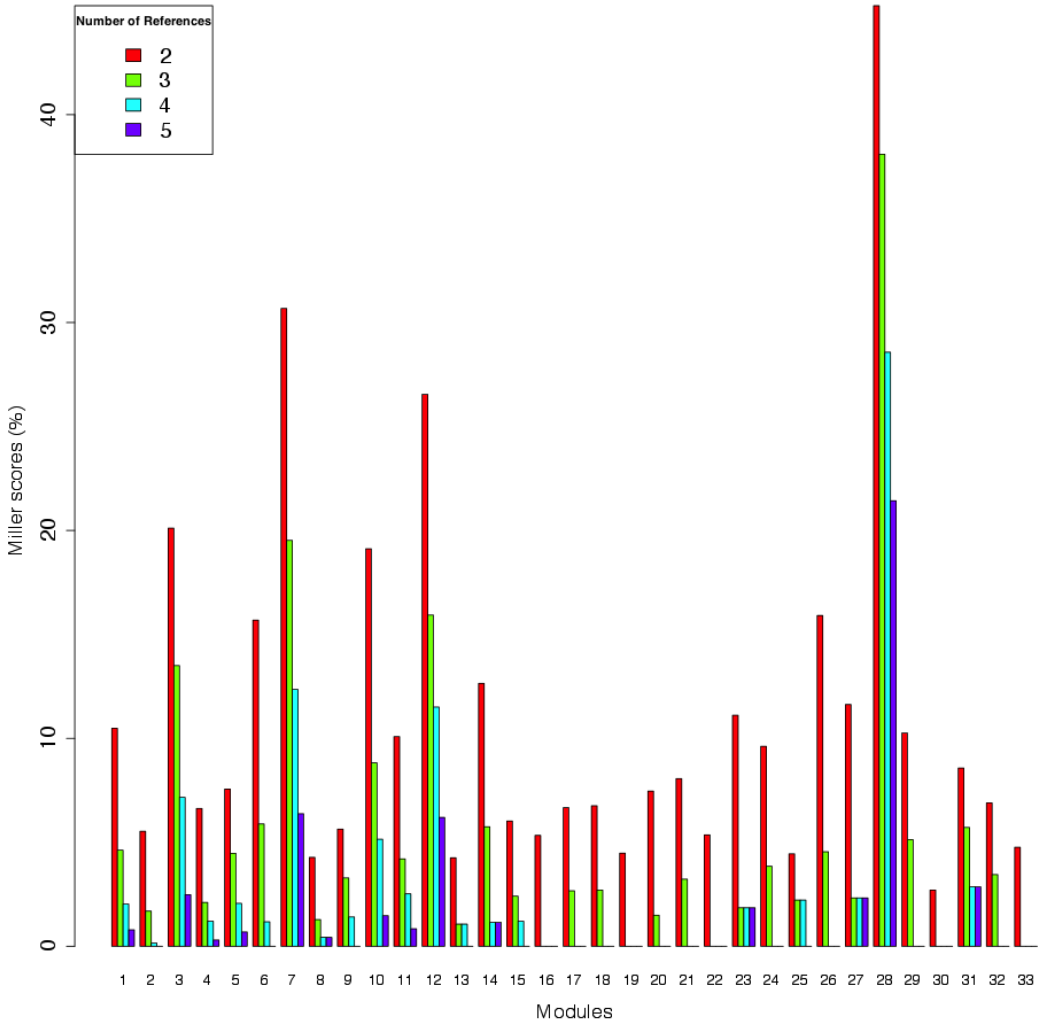


Figure II.1: **Miller scores.** The 33 modules identified using MILE dataset are sorted on the x-axis based on their sizes. The y-axis shows the percentage of genes in each module that were reported to be related to AML in at least 2 (red), 3 (green), 4 (blue), and 5 (purple) studies according to Miller et al. survey (Miller & Stamatoyannopoulos, 2010).

Computing module eigengenes

Our strategy for computing a computational network is to use the modules of genes as nodes of BN. Because the nodes of BN can only be represented by a single random variable, we cannot use the whole module of genes as a node. To overcome this, we summarized the biological information of each module in one eigengene using Principal Component Analysis (PCA)(Jolliffe, 2002).

An eigengene of a module is a weighted average of expressions of all genes in that module. The weights are adjusted so that the loss in the biological information is minimized (Jolliffe, 2002; Oldham et al., 2006). Computation of eigengenes transformed expression data from "sample $\times$ genes" space to "sample $\times$ eigengenes" space (Alter et al., 2000), which can be used for further analysis. Figure II.2 shows a graphical representation of the steps we followed to compute module eigengenes from our gene expression dataset.

Data processing for BN learning

Our data was continuous and needed to be processed before we could use it for BN learning. First, our data needed to be discretized. Discrete data allows us to model complex non-linear interactions between genes without resorting to computationally prohibitive calculations over continuous distributions (Yu et al., 2002). We used 3-interval discretization, because it has been found to be optimal for BN learning (Yu et al., 2002).

After processing the existing expression data, we needed a marker node in our network that can show the correlation between nodes with AML. To map the correlation between nodes and the disease, we introduced a marker node "Effect" whose expression value is "1" for AML disease samples and "0" for MDS disease samples. We used clinical data to determine the samples belonging to either AML or MDS disease.

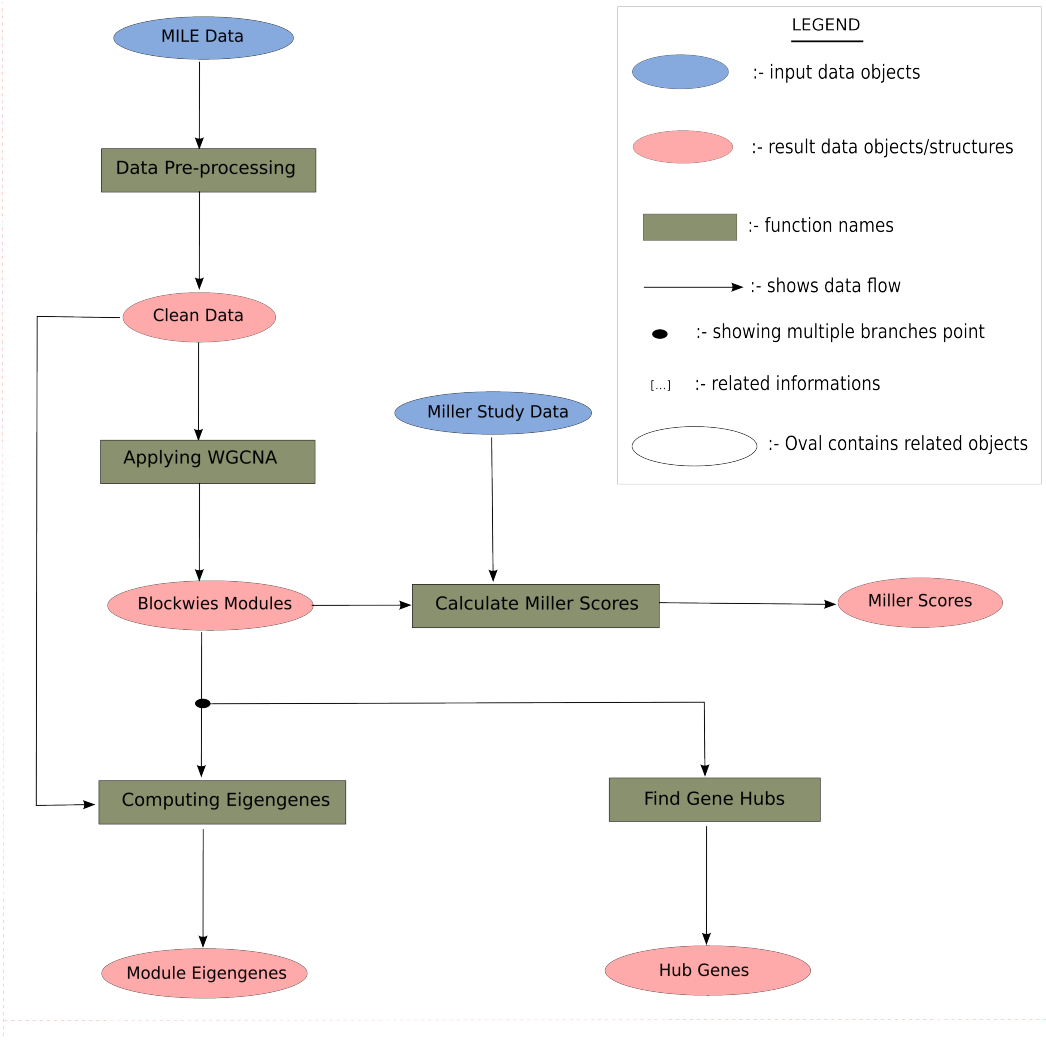


Figure II.2: **Graphical presentation of the steps for data processing.** MILE data is downloaded and cleaned through noise elimination and transformation. WGCNA is used as a clustering technique to compute gene modules. Final outcome is the module eigengene vector which will be used for BN learning.

To find out the nodes that are correlated with AML, we used Markov blanket concept of BN. A Markov blanket for a node is a set of nodes composed of the parents of the node, children of the node and the parents of children of that node (Pearl, 2014). Markov blanket of a node is the only knowledge needed to predict the behavior of that node (Pearl, 2014). We blacklisted the children of Effect node by eliminating all the edges coming from it. This way, the Markov blanket for Effect node consists only its parent nodes. Our assumption was that the parent nodes of the marker node could be the modules enriched with genes that are associated with AML.

Learning the bayesian network

We used the discretized processed data to learn the BN. A study by Yu et al. (2002) suggests that a greedy search method with random restarts, employing Bayesian Dirichlet equivalent (BDe) scoring metric (Heckerman et al., 1995), and being given data discretized with 3-interval discretization is best BN inference algorithm for recovering the simulated genetic pathways. We learned our BN from 500 networks with BDe scoring and called the resulting networks as candidate networks. The BDe scores were used as a measure of the goodness of the network. We took an average of the one third candidate networks with the best score and created a consensus network that maps the overall behavior of the individual networks.

Figure II.3 shows the steps we followed to compute the consensus network.

Cross validation

A single round of cross validation involves partitioning the data into complementary subsets, performing the analysis on one subset (training set), and validating the analysis on the other subset (validation set or testing set) (Refaeilzadeh et al., 2009). In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance

Figure II.3: **Graphical presentation of the steps for BN learning.** Input data is received in the form of eigengene values. After processing individual BNs are learned for 500 random restarts and corresponding scores are calculated. Finally, consensus network is computed based on the top third individual networks with best BDe scores.

of being validated against (Refaeilzadeh et al., 2009). To reduce variability, multiple rounds of cross-validation are performed using different random partitions and the validation results are averaged over the rounds. In this study, we used five rounds of CV, called the 5-fold CV, to validate our model. In each round of 5-fold CV, 1/5th of the dataset were considered as "testing/validation" set and 4/5th of the dataset were considered as "training" set.

Average voting and performance measurement

In 5-fold CV, we computed BN model on the "training" sets and used them to predict disease type on the "validation"/"testing" sets. We ended up with five different models for classification. We used confusion matrix, also known as error matrix, to find other statistical performance measurements such as accuracy, sensitivity, and precision (Stehman, 1997; Fawcett, 2006). Accuracy is the proportion or percentage of correcly predicted labels over all predictions. Sensitivity, alson known as recall, measures the proportion of positive samples that are correctly identified as such. Precision, also known as positive predictive value, measures the proportion of actual positive samples in the population being tested (Bishop, 2007; James et al., 2013). Positive samples are the samples that we want to identify in our study. We considered "AML" samples as positives and "MDS" samples as negatives. Figure II.4 shows the steps we followed to perform the CV and measure the performances.

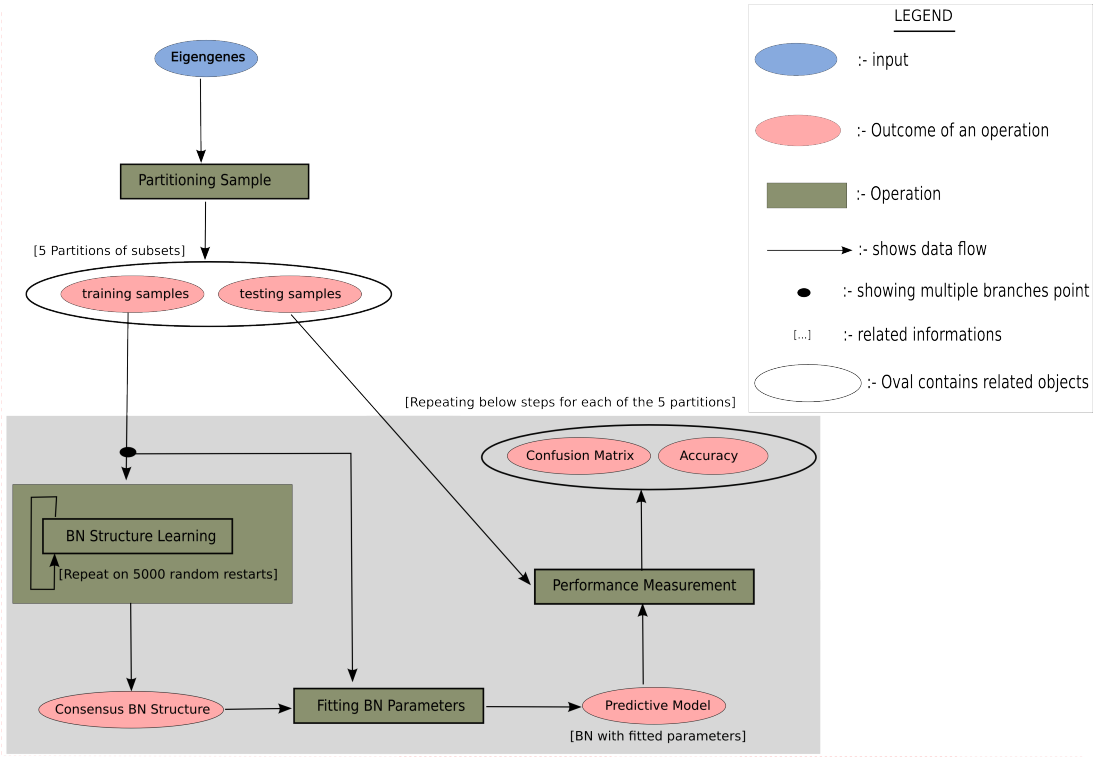


Figure II.4: **Steps for 5-Fold CV and performance measurement**. We partitioned the eigengene matrix into five sets of training and testing subsamples. For each partition, we use training samples to learn the BN and testing samples to measure the performance. Final outcome is the confusion matrix along with performance measures.

III. RESULTS

Data processing and clustering

We used MILE dataset to compute the predictive model. The MILE dataset consists of 366 samples. We applied WGCNA on AML samples and identified 33 gene modules as clusters of genes that have high correlation in the 202 AML cases. The cluster dendrogram (Figure III.1) shows the module assignments of the 33 gene modules where each color represents exactly one module. WGCNA could not confidently assign 4,125 genes to any module because they hardly correlate with any other gene. We grouped those 4,125 genes in "module 0" and considered them as outliers for this study. The genes in module 0 are represented by "gray" color in figure III.1. Table III.1 lists the module assignments and their sizes. We plotted the distribution of modules and sizes (Figure III.2) ignoring the outliers.

Computing eigengenes and learning BN structure

We computed eigengene values for each of the modules. The resulting eigengene matrix contains 366×33 elements for the 33 modules obtained from WGCNA analysis. The rows of the eigengene matrix corresponds to the data samples and columns correspond to modules. We plotted heatmap for the expression of all eigengenes in MILE dataset (Figure III.3). In the heatmap, eigengenes show significantly different patterns in the samples (rows) for the two disease types. We hypothesized that the eigengenes are important biological signatures that can predict the disease type solely based on gene expression. To validate this hypothesis, we computed a predictive model that can use eigengenes expression to predict the disease type.

We processed data by applying 3-interval discretization and then adding "Effect" marker node to it. We used the processed data to learn the BN

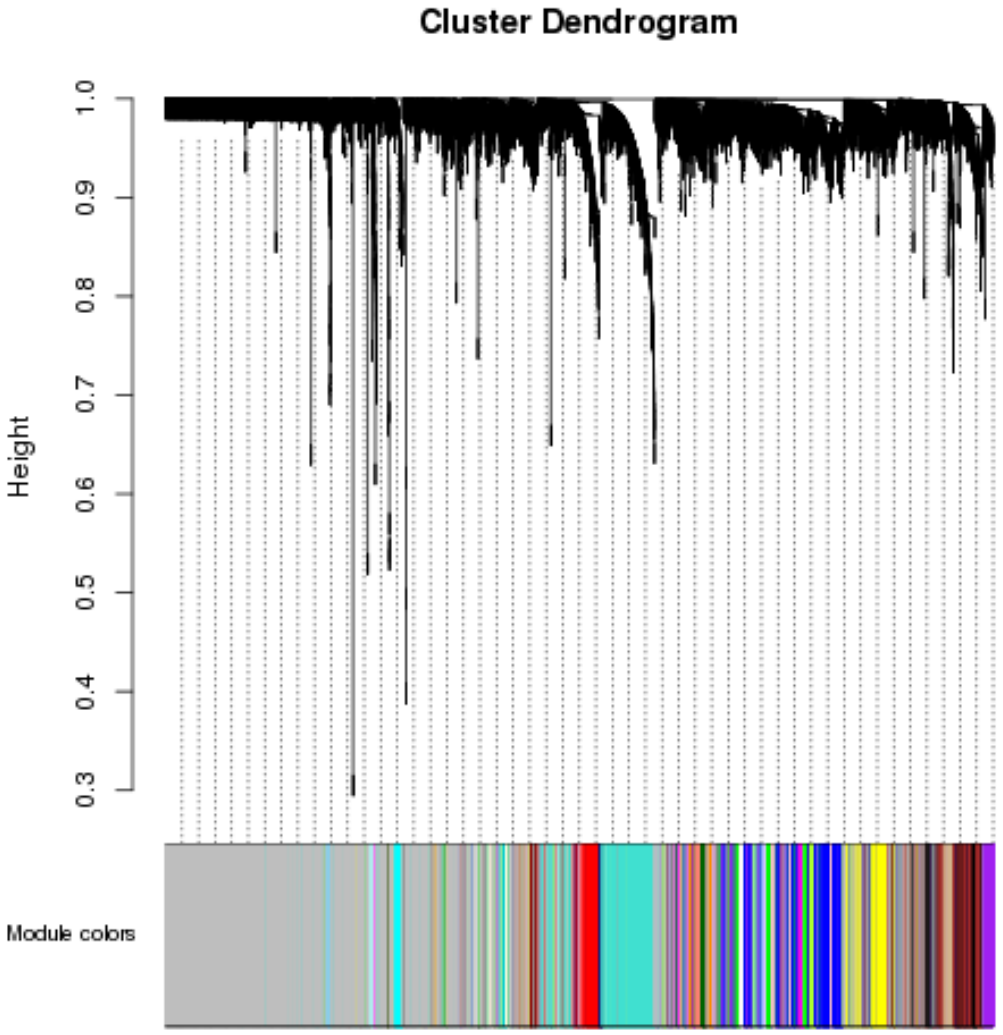


Figure III.1: **Cluster dendrogram for AML.** WGCNA assigned genes into 33 different modules. Each module assignment is represented by a color. The module color "gray" represents the genes that did not correlate to any other gene properly.

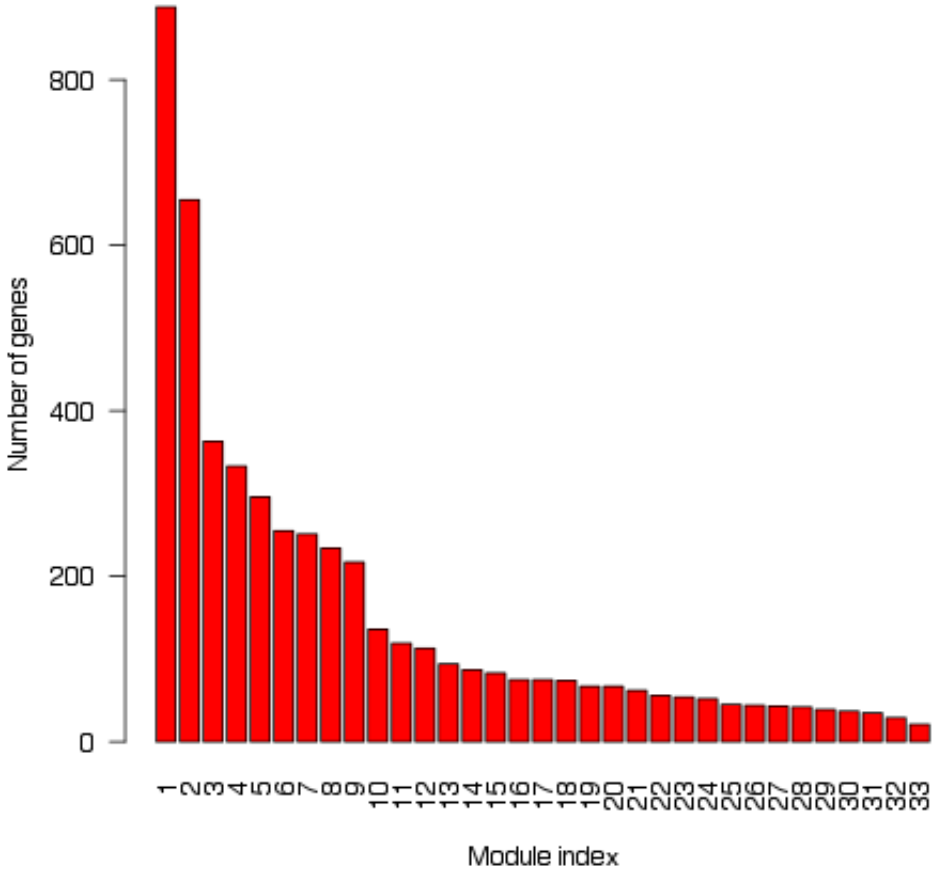


Figure III.2: **The distribution of module sizes.** The 33 modules are sorted on the x-axis based on their size (i.e. the number of genes they contain).

Table III.1: Modules and sizes
Table lists the modules and the number of genes they contain (size of the module). "Module 0" contains 4,125 outlier genes that were ignored in this study.

Module	**Size**	**Module**	**Size**	**Module**	**Size**
0	4125	1	888	2	655
3	363	4	333	5	296
6	255	7	251	8	234
9	217	10	136	11	119
12	113	13	94	14	87
15	83	16	75	17	75
18	74	19	67	20	67
21	62	22	56	23	54
24	52	25	45	26	44
27	43	28	42	29	39
30	37	31	35	32	29
33	21				

structure. As suggested by the Yu et al. (2002) study, we learned multiple BN structures to get networks using random restarts.

To decide the number of networks needed for this study, we experimented with various number of networks including 10, 50, 100, 200, and 500 networks. The BNs learned from lower number of networks (below 500) did not converge well and had ample room for improvement (Figure III.4). BN learning from 500 networks converged well with very little room for improvement (Figure III.5). Because the BN structure with 500 random networks converged well, repeating the experiment for a higher number of networks may not result into a different predictive model. Figures III.6 and III.7 shows the consensus BN structures learned from 500 and 5,000 random networks, respectively. The two models have similar structure including 'module 6' and 'module 12' being the parent nodes of 'Effect' node in both the networks which suggests that these two modules could contain genes that are related with AML. We chose 500 networks for our experiment.

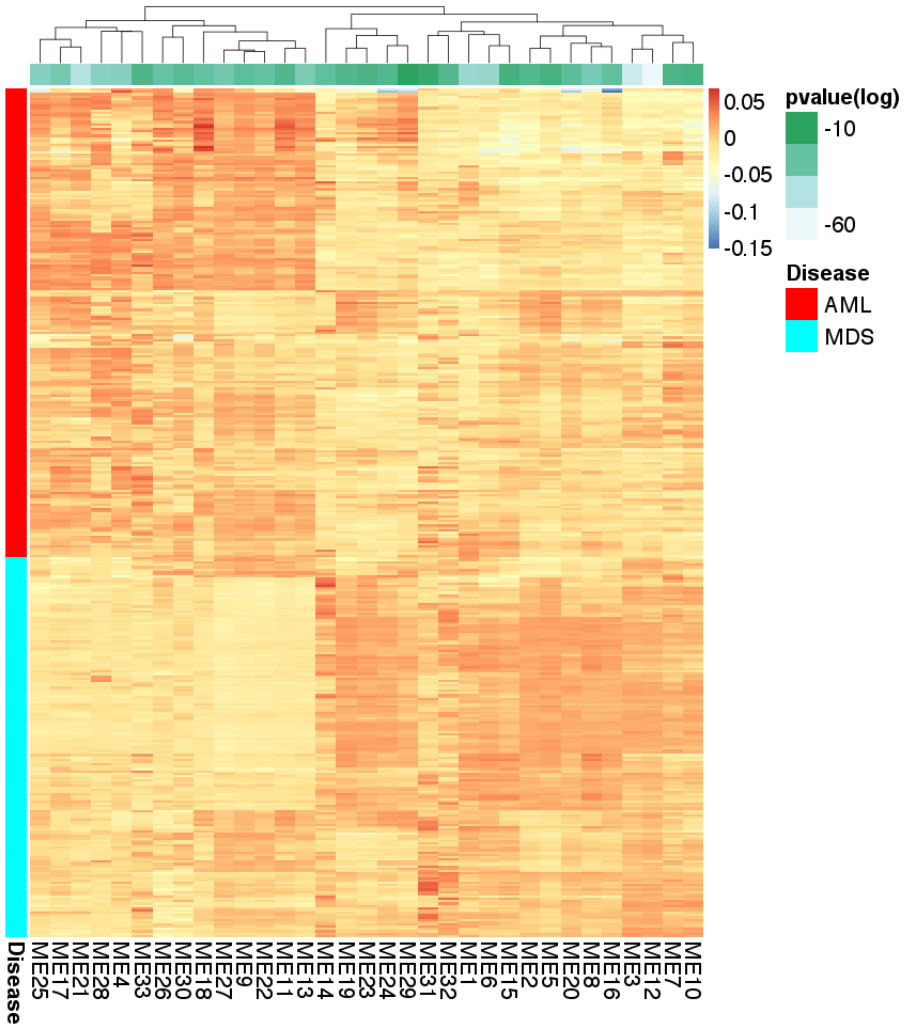


Figure III.3: **Expression of eigengenes in MILE dataset.** Eigengenes show different pattern in the samples (rows) for the two disease. Modules (columns) are clustered together based on the similarity of expression in MILE dataset.

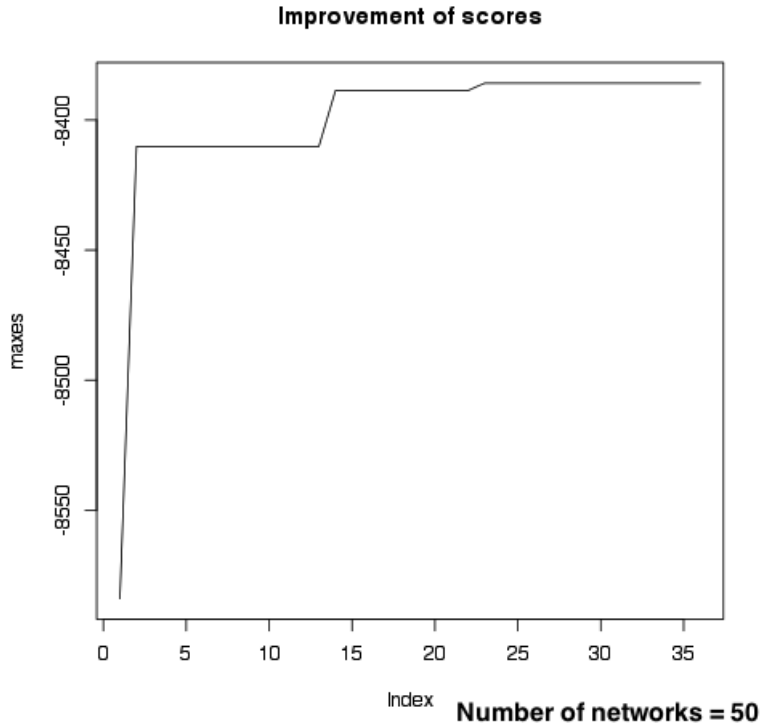


Figure III.4: **Plot for improvement of score from 50 random networks.** The plot shows the improvement of score for learning BN on MILE data from 50 random networks. There is ample room for improvement.

Fitting parameters and 5-fold cross validation

BNs can handle uncertainty using the theory of probability. To make a predictive BN model, we needed to fit the conditional dependency tables for the nodes of BN. We used *bn.fit()* function of *bnlearn* package (Nagarajan et al., 2013) to fit BN parameters. The resulting network with fitted parameters is called predictive model and we can use it for predicting the disease type of patients.

To validate our strategy, we used 5-fold CV. We partitioned the processed input data into five random partitions keeping 1/5th of data as "validation" set and 4/5th of the data as "training" set. Iteratively, we learned BN structure and fitted the parameters on "training" samples of each of the partitions. As a result, we computed five different predictive models and used them to predict the disease types on their respective "validation" samples. The (a) sections of Appendix A show five BN structures computed while performing 5-fold CV.

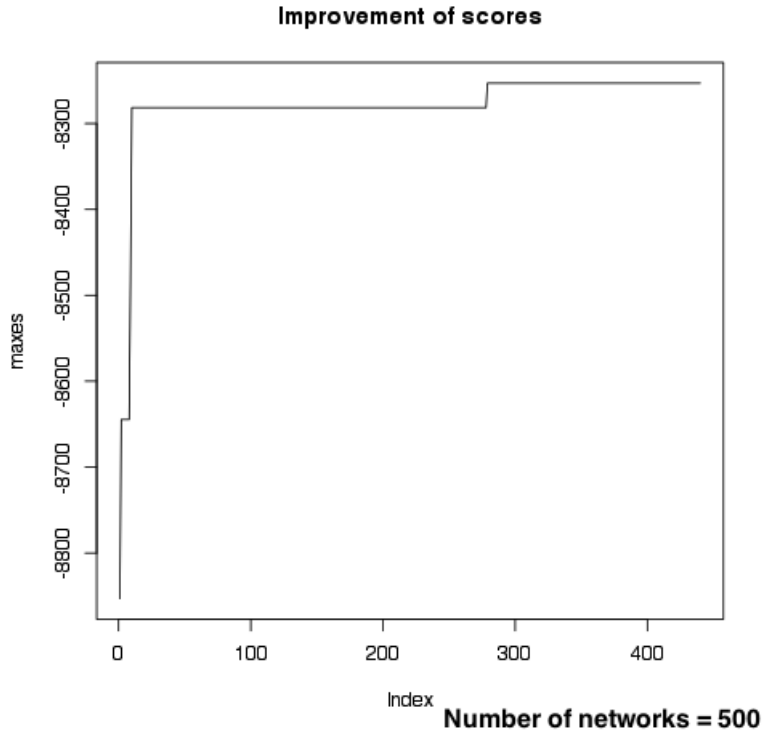


Figure III.5: **Plot for improvement of score from 500 random networks.** The plot shows the improvement of score for learning BN on MILE data from 500 random networks. The model converged well and thus chose 500 networks for learning our predictive BN model.

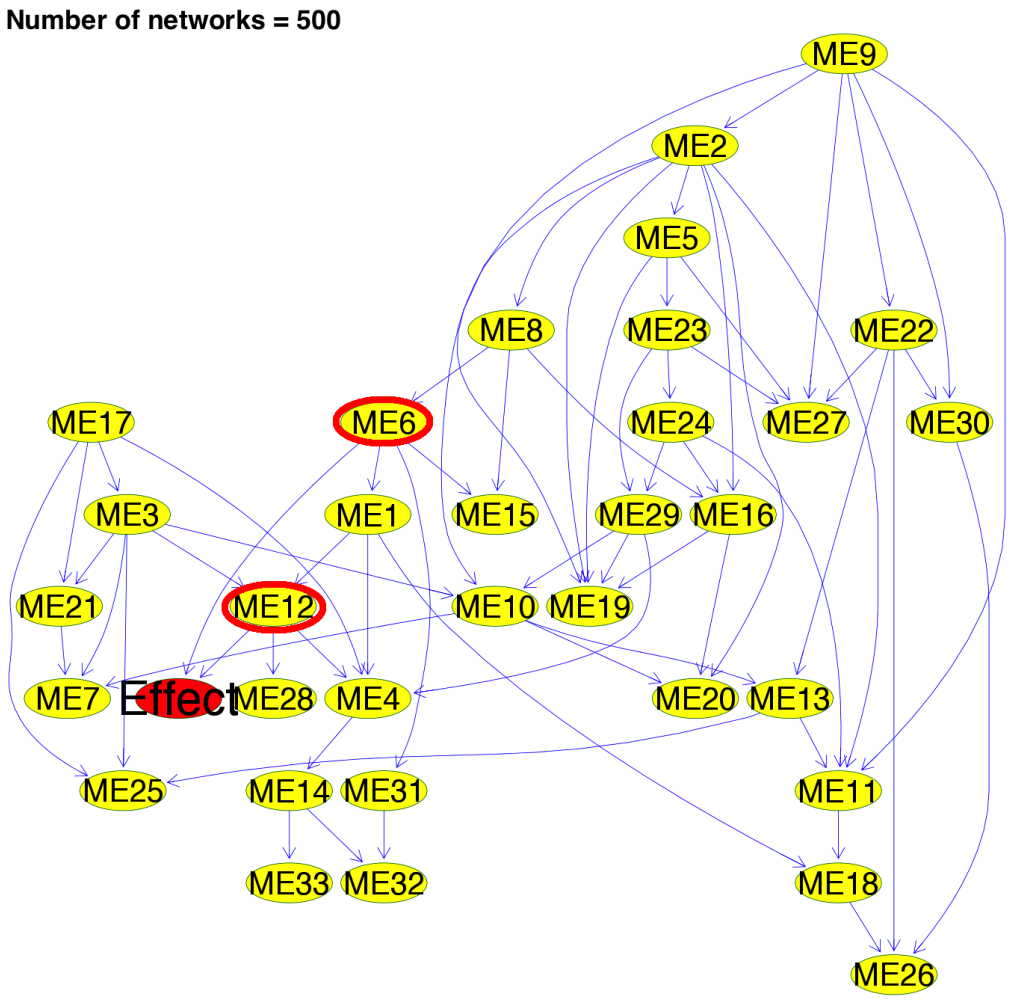


Figure III.6: **Consensus BN structure for 500 random networks.** This consensus BN was computed on MILE data for 500 networks computed using random restarts. Modules are represented as nodes and the edges denote the dependency between connected nodes. Effect is the marker node for AML disease.

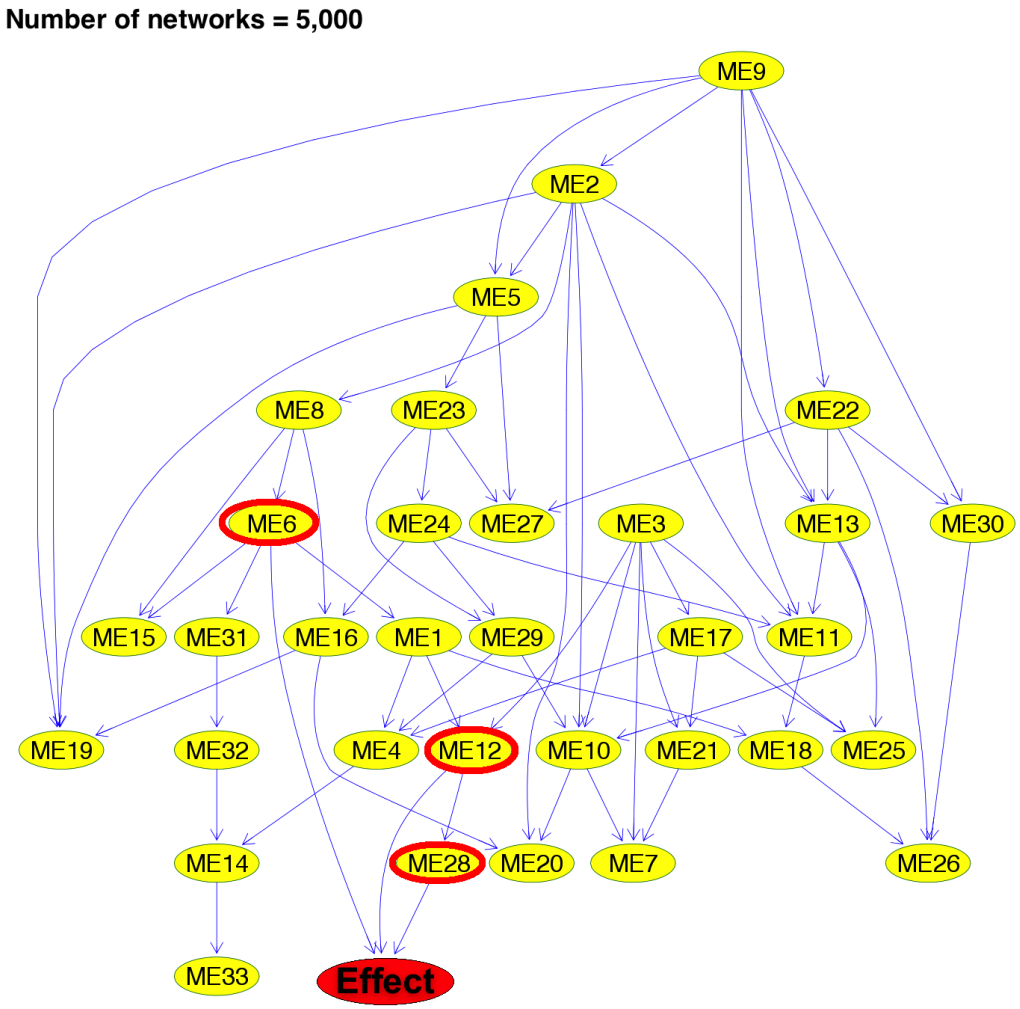


Figure III.7: **Consensus BN structure for 5,000 random networks.** This consensus BN was computed on MILE data for 5,000 networks computed using random restarts. Modules are represented as nodes and the edges denote the dependency between connected nodes. Effect is the marker node for AML disease.

Measuring performance and average voting

As a result of the validation performed using five predictive models, we got five confusion matrices along with accuracy, sensitivity, precision (positive predictive value), and other statistical measures. The (b) sections of Appendix A show individual confusion matrices and the performance measures for all the partitions computed while performing 5-fold CV. The range of accuracy on the validation set of 5-fold CV was 78.5%-94.6% with an average accuracy of 88%, while the range of accuracy on the training set of the 5-fold CV was 88%-97% with an average accuracy of 93.2% (Table III.2).

I performed majority voting of the five individual predictions to get consensus prediction for 366 samples of MILE data. After taking the majority vote, we found out that 24 samples were misclassified. The confusion matrix and statistical measurements for the majority voting, summarized over the results of five predictive models computed using 5-fold CV (Figure III.8), shows that the accuracy of the consensus prediction is 93.4%. The recall (sensitivity) and precision were 90.1% and 97.85%, respectively. Table III.2 lists the performance of individual predictive models, and their mean on both training and validation data sets. There was not a considerable difference between the mean accuracies for training and validation sets; however, we found that the consensus accuracy (93.4%) was slightly better than the mean accuracies on training (93.2%) and validation (88%) sets of MILE dataset. Also, the consensus accuracy was better than the mean accuracy of the models (92.2%) on the complete MILE dataset (Table III.3). This shows that taking an average vote of the predictions could result in a better predictive model. Appendix B contains a table of all the individual model predictions, and majority vote of individual predictions on MILE dataset.

```
Confusion Matrix and Statistics for the majority vote results on MILE data

            Actual
Prediction AML MDS
       AML 182   4
       MDS  20 160

               Accuracy : 0.9344              95% CI : (0.904, 0.9575)
    No Information Rate : 0.5519              P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8686           Mcnemar's Test P-Value : 0.0022

             Sensitivity : 0.9010                      Specificity : 0.9756
          Pos Pred Value : 0.9785                  Neg Pred Value : 0.8889
              Prevalence : 0.5519                  Detection Rate : 0.4973
   Detection Prevalence : 0.5082                Balanced Accuracy : 0.9383

       'Positive' Class : AML
```

Figure III.8: **Performance measurements for the average vote.** The confusion matrix shows that 20 MDS samples were misclassified as AML and 4 AML samples were misclassified as MDS. Consensus accuracy is 93.4% along with 97.85% recall (sensitivity) and 90.1% precision (positive predictive value).

Prediction using the model on BCCA data

One of the most intriguing facts about BNs is that they can handle uncertainty using probability theory. MILE data is a microarray data which was used to train the predictive model in this study. We have access to 155 samples of RNA Sequence (RNA-Seq) data (BCCA dataset). The technologies used for calculating the differential expressions in MILE and BCCA dataset are different. We validated our model on BCCA data to see how well it predicts the patients. BCCA dataset contains 155 samples of data with various disease types, including 74 samples of gene expressions with the same disease type as MILE dataset. For this experiment, we kept only the samples with disease type same as MILE dataset and thus, we ended up with 54 AML samples and 22 MDS samples of BCCA data.

We used the five predictive models to predict the disease type on 74 samples of BCCA dataset and computed five different results. After taking a majority vote on the individual predictions, we found out that our consensus

model misclassified 12 samples. The confusion matrix and statistical measures calculated for the prediction results on BCCA data (Figure III.9) shows that our model predicted the patients with 83.8% accuracy, 84.5% precision, and 94.3% recall. Table III.4 lists the performance of individual predictive models, mean of the performances, and the performance of average vote on BCCA dataset. The mean accuracy of the predictive models was 77% while the consensus accuracy was 83.8%. Consensus network allows us to make our model safe from noise or accidental predictions and testing on both MILE and BCCA data shows that the performance of the consensus predictive model is better than the mean of five individual predictive models. Appendix C contains a table of all the individual predictions, and majority vote of individual predictions on BCCA dataset.

We computed five predictive models as a result of 5-fold CV and tested BCCA dataset on each of those models. The predictions by consensus is better than the individual models predictions (Table III.4). Figures in Appendix D show the performances of each of the individual predictive models on BCCA dataset.

Table III.2: Statistical measures on training and test subsets of MILE dataset. Table lists the statistical performance measurements computed as result of 5-fold CV predictive models on the training (4/5) samples and testing (1/5) samples of MILE dataset.

	Training Set			**Testing Set**		
Model	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
Model 1	96.9	95.7	98.7	78.4	71.8	84.8
Model 2	92.8	89.6	97.3	89.2	84.2	94.1
Model 3	91.5	88.8	87.4	86.5	80.1	94.5
Model 4	88	83.1	94.3	94.6	95.2	95.2
Model 5	96.6	96.9	96.9	91.5	95.1	90.1
Mean	**93.2**	**90.8**	**94.9**	**88**	**85.2**	**91.8**

Comparing prediction results

Our consensus predictive BN model performed prediction on the MILE dataset with an accuracy of 93.5%. The model misclassified 20 AML samples and

Table III.3: Statistical measures on overall MILE dataset.
Table lists the statistical performance measurements by all the predictive models on 366 samples of MILE dataset. The performance of our study is better than the predictive performance of Mills et al. (2009) study on MILE dataset (Table III.5).

Model	**Accuracy**	**Precision**	**Recall**
Model 1	93.2	96.3	91.1
Model 2	92.1	96.8	88.6
Model 3	90.4	95.1	87.1
Model 4	89.4	94.5	85.6
Model 5	95.6	95.6	96.5
Mean	**92.2**	**95.7**	**89.8**
Consensus	**93.5**	**97.9**	**90.1**
Mills et al. (2009) study	73.8	93.1	69.6

Table III.4: Statistical measures on BCCA dataset.
Table lists the statistical performance measurements by all the predictive models on 74 samples of BCCA dataset.

Model	**Accuracy**	**Precision**	**Recall**
Model 1	81.1	85.2	88.5
Model 2	68.9	83.7	69.2
Model 3	81.1	82.8	92.3
Model 4	79.7	80.3	94.2
Model 5	74.3	79	86.5
Mean	**77**	**82.2**	**86.2**
Consensus	**83.8**	**84.5**	**94.2**

```
Confusion Matrix and Statistics for the majority vote results on BCCA data

          Actual
Prediction AML MDS
       AML  49   9
       MDS   3  13

               Accuracy : 0.8378              95% CI : (0.7339, 0.9133)
    No Information Rate : 0.7027              P-Value [Acc > NIR] : 0.005718

                  Kappa : 0.5787           Mcnemar's Test P-Value : 0.148915

            Sensitivity : 0.9423                      Specificity : 0.5909
         Pos Pred Value : 0.8448                   Neg Pred Value : 0.8125
             Prevalence : 0.7027                   Detection Rate : 0.6622
   Detection Prevalence : 0.7838                Balanced Accuracy : 0.7666

       'Positive' Class : AML
```

Figure III.9: **Performance measurements for the average vote on BCCA data.** The confusion matrix shows that all the 3 MDS cases were misclassified to be AML and 9 AML cases were misclassified to be MDS. Consensus accuracy is 83.8% along with 84.5% recall (sensitivity) and 94.2% precision (positive predictive value).

4 MDS samples, making a total of 24 misclassified samples. The accuracy of the consensus model on only AML samples was 90.1% and on only MDS samples was 97.6%.

A study by Mills et al. (2009) was conducted on MILE dataset and it performed classification with an approximate accuracy of 93% on AML samples and 50% on MDS samples. Table III.5 shows the confusion matrix for the actual and predicted samples by Mills et al. (2009) study. The accuracy, precision, and recall for that study was 73.8%, 93.1% and 69.6%, respectively (Table III.3). The study uses a diagnostic classification (DC) model, developed for the MILE study, to distinguish leukemia from MDS and from nonleukemic conditions (Mills et al., 2009). DC model was based on a margin tree graph that was generated following a method previously established in the use of high-dimensional classification of cancer microarray data (Haferlach et al., 2010; Tibshirani & Hastie, 2007). Margin tree classifiers such as DC model choose the number of classes required, seek the line that partitions the classes into groups, calculate the maximum

margin[1] for the classes, and then use approaches such as greedy (Cormen et al., 2001), single linkage, or complete linkage (Manning et al., 2008, p. 350) to group the samples into chosen classes (Tibshirani & Hastie, 2007).

Table III.5: Confusion matrix for Mills et al. (2009) study.
Table shows the confusion matrix for Mills et al. (2009) study. "AML" samples are considered positive.

		Actual	
		AML	**MDS**
Predicted	AML	188	82
	MDS	14	82

Mills et al. (2009) study predicted AML and MDS samples of MILE data with an accuracy of 93% and 50%, respectively, while our study predicted AML and MDS samples of MILE data with an accuracy of 90.1% and 97.6%. Our predictive model has slightly less accuracy on AML samples of MILE data but the performance of our model on MDS samples of MILE data is exceptional. Also, the overall performance of our model is better than Mills et al. (2009) study (Table III.3).

Apart from MILE data predictions, we also used the predictive BN models to predict disease type on BCCA dataset and computed a consensus model based on the individual results. On BCCA dataset the consensus accuracy was 83.8% along with 94.3% accuracy on only AML and 60% accuracy on only MDS samples. The accuracy on MDS samples of BCCA dataset are not as good as the accuracy on MILE dataset MDS samples but still its better than the MDS sample accuracy of Mills et al. (2009) study. The accuracy on AML samples of BCCA dataset (94.3%) is better than the accuracy on AML samples of MILE dataset (90.1%) and accuracy on AML samples of Mills et al. (2009) study (93%).

The DC model or the margin tree graph used for classifying the diseases in Mills et al. (2009) study compare the performance of the "margin tree" to the closely related "all-pairs" (one versus one) support vector machine, and nearest

[1]The margin is the minimum distance of the data points to the decision line.

centroids on a number of cancer microarray data sets. They found that the margin tree has accuracy that is competitive with other methods and offers additional interpretability in its putative grouping of the classes (Tibshirani & Hastie, 2007). Because our model performs better predictions, we can conclude that our model is better than "margin tree graph", "support vector machine", and "nearest centroids" in predicting AML/MDS disease type of samples.

Analyzing predictive model structures

We computed five predictive models as a result of 5-fold CV. "Effect" node in the predictive model shows the correlation of the modules with AML disease. Table III.6 lists the parent nodes of Effect for each of the models. The most common modules that are parent of Effect node are modules 1, 4, 12, 28, and 30 (Table III.7).

As per the BN structure learned from 500 networks (Figure III.6), modules 6 and 12 are the modules that could be enriched with the genes that correlate with AML. Module 6 contains genes that are related to "Cell cycle" pathway and Module 12 contains genes that are related to "Extra cellular Matrix" pathway. A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell (Schuster et al., 1999). Further experiments could be conducted to study the genes that could be related to AML.

Table III.6: Parent nodes of Effect node.
Table lists the parent nodes of "Effect" node, as predicted by five predictive models as a result of 5-fold CV. "ME" stands for "Module Eigengenes" and it denotes the "gene modules".

Model	**Parent nodes of Effect**
Model 1	ME1, ME4, ME12, ME28, ME30
Model 2	ME1, ME3, ME4, ME21
Model 3	ME1, ME4, ME12
Model 4	ME4, ME12
Model 5	ME6, ME12, ME14, ME28, ME30

Table III.7: Frequency of parent nodes of Effect node.
The table shows the frequency of the parent nodes of "Effect" node in table III.6. Module 1, 4, and 12 were modules that were chosen as parent of Effect node by majority of the predictive models.

Module	**4**	**12**	**1**	**28**	**30**	**3**	**6**	**14**	**21**
Frequency	4	4	3	2	2	1	1	1	1

IV. DISCUSSION

Biological processes in a cell often require coordination between multiple genes. We used gene network analysis to model the interaction between genes (Langfelder et al., 2013). Using network analysis, we can examine the differences in gene expression profiles of samples affected by AML or MDS diseases (Figure III.3). Our study shows how a Bayesian network learned on module eigengenes can be used as a predictive model to predict the disease type of patients solely based on their gene expression data.

We built a predictive model based on MILE data and validated it on BCCA data. On the training dataset, our model was able to predict the disease type of MILE samples with 93.5% accuracy. However, when used on the test dataset, it predicted disease type of BCCA samples with an accuracy of 83.8%. The experiments show that, training a BN on microarray data (MILE) and testing it on RNA-Seq data (BCCA) gives comparatively less accurate predictions, however, the predictions are better than other studies (Mills et al. (2009)).

Our BN modeled MILE and BCCA dataset that were generated using different technologies, therefore, we believe that BNs are capable of modeling heterogeneous data. Further studies could use this capability of BNs in analyzing data acquired using different sources or technologies. A suggestion for future experiments would be to reverse the process and see how well a predictive model based on BCCA data can predict samples of MILE data. It is also interesting to test the results of using a smaller dataset for training and a larger dataset for validation.

This study has the potential to scale to experiments that study large network of genes by analyzing co-expressions or BNs to identify causal relationships between genes. The results of such experiments could be useful in pinpointing the cause and origin of diseases and can potentially aim to find out

novel treatment plans.

Novelty

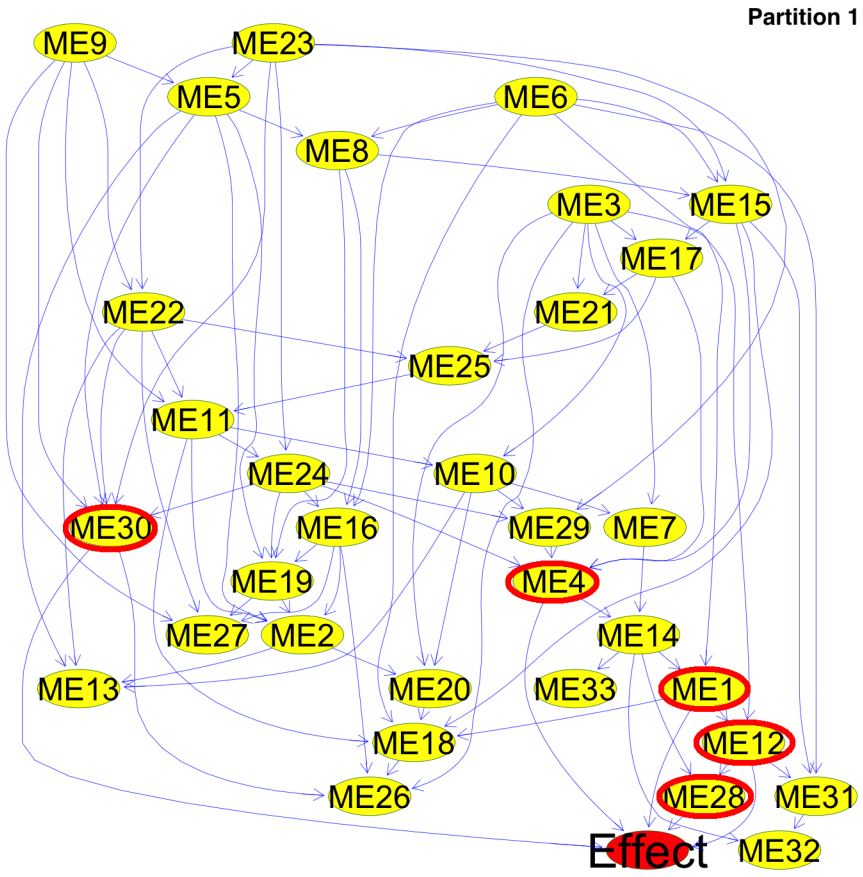
An eigengene summarizes the biological information in a module with a single value. In this study, they were the features to compute our model. Eigengenes can map thousands of genes with small number of features which allows us to study large number of genes in a single model. In contrast, the studies that tried to use BNs to study interaction between genes, had a restriction on the number of genes (Yu et al., 2002). Those studies could use only few genes as features, but our model can use thousands of genes as features (Zhang et al., 2013; Friedman & Koller, 2003).

Applications

This study also finds out the module of genes that could be related to the AML. We found out that the genes within modules 1, 3, 4, 6, 12, 14, 21, 28, and 30 (Table III.6, III.7, and figure III.6) could be related to AML. This information narrows down the search space for the studies that try to find out the genes that cause of the disease.

APPENDIX SECTION

APPENDIX A

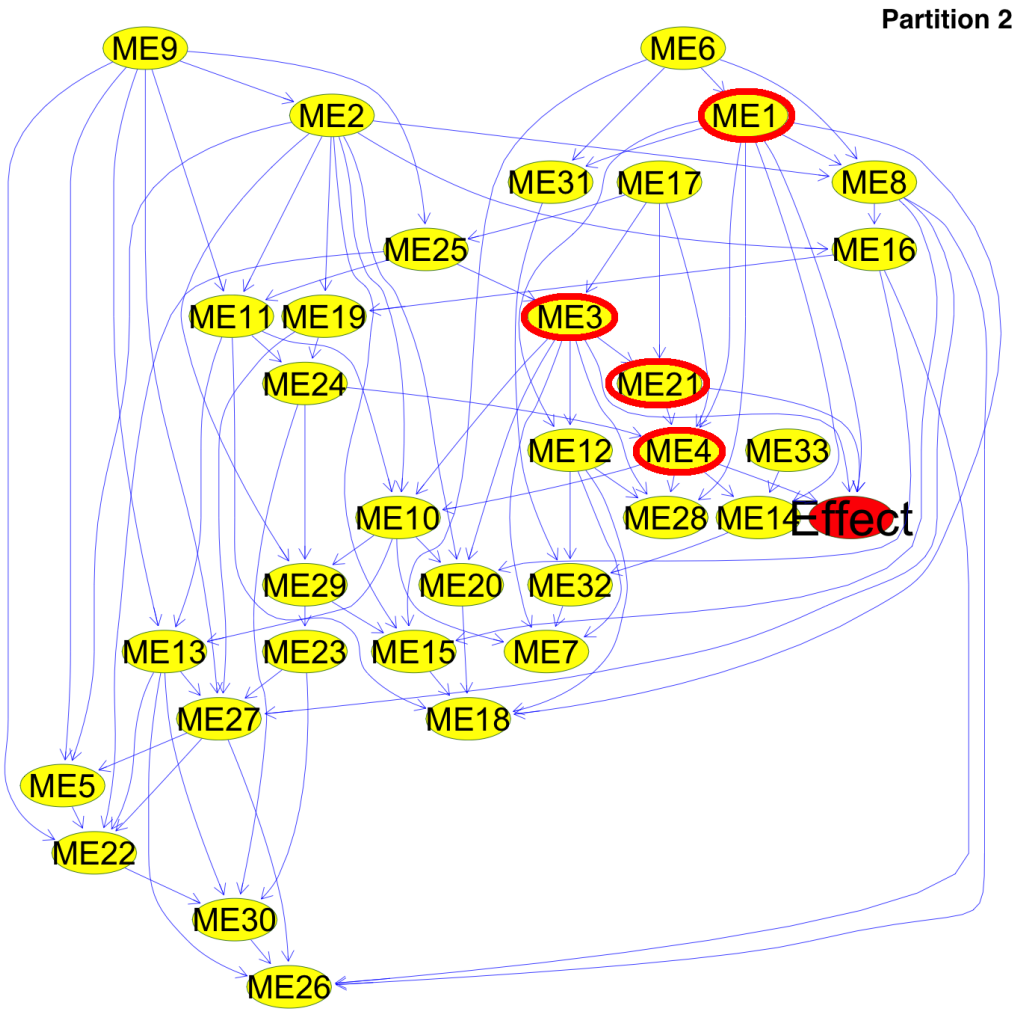


(a) BN structure.

```
Partition 1: Confusion matrix and statistics

           Actual
Prediction AML  MDS
       AML 184    7
       MDS  18  157

               Accuracy : 0.9317                          95% CI : (0.9008, 0.9553)
    No Information Rate : 0.5519              P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8628           Mcnemar's Test P-Value : 0.0455

            Sensitivity : 0.9109                      Specificity : 0.9573
         Pos Pred Value : 0.9634                  Neg Pred Value : 0.8971
             Prevalence : 0.5519                  Detection Rate : 0.5027
   Detection Prevalence : 0.5219               Balanced Accuracy : 0.9341

       'Positive' Class : AML
```

(b) Statistical measures.

BN structure and statistical measures for partition 1 of 5-fold CV.

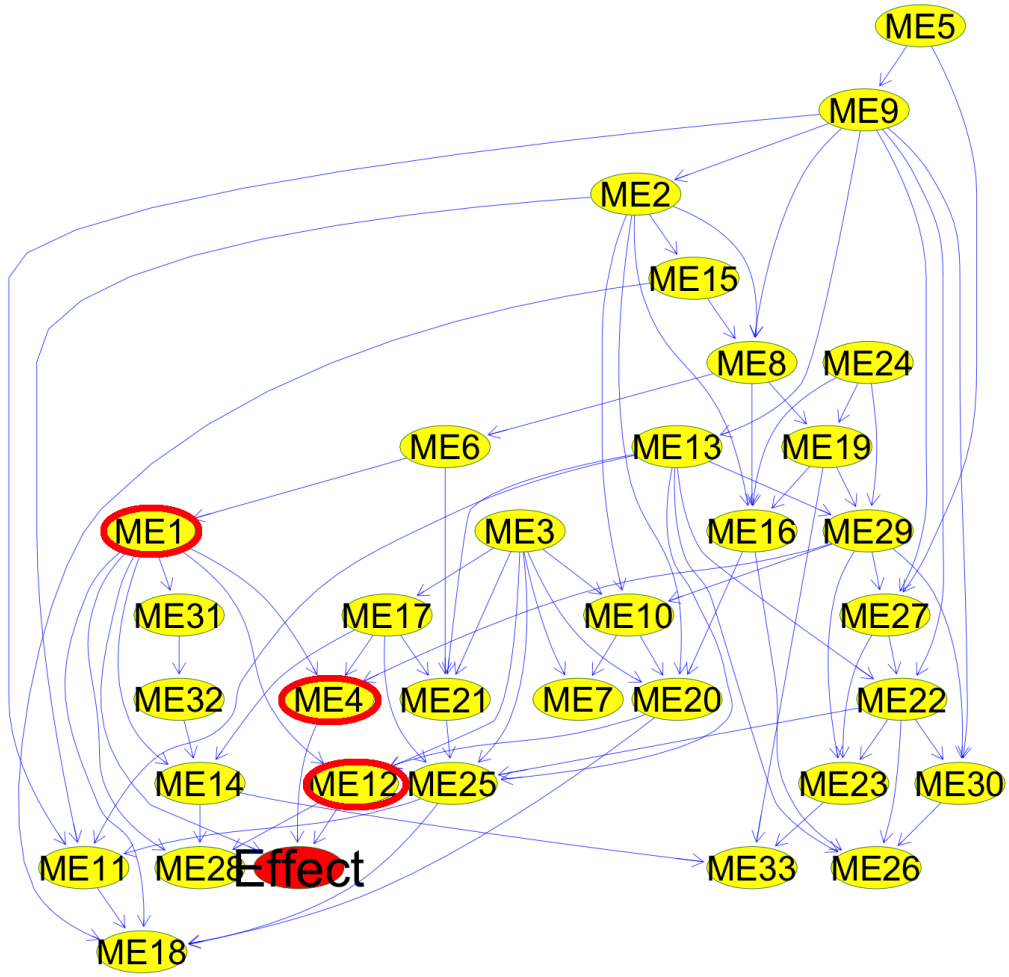


(a) BN structure.

```
Partition 2: Confusion matrix and statistics

          Actual
Prediction AML MDS
       AML 179   6
       MDS  23 158

               Accuracy : 0.9208                          95% CI : (0.8882, 0.9463)
    No Information Rate : 0.5519              P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8414           Mcnemar's Test P-Value : 0.002967

            Sensitivity : 0.8861                      Specificity : 0.9634
         Pos Pred Value : 0.9676                   Neg Pred Value : 0.8729
             Prevalence : 0.5519                   Detection Rate : 0.4891
   Detection Prevalence : 0.5055                 Balanced Accuracy : 0.9248

       'Positive' Class : AML
```

(b) Statistical measures.

BN structure and statistical measures for partition 2 of 5-fold CV.

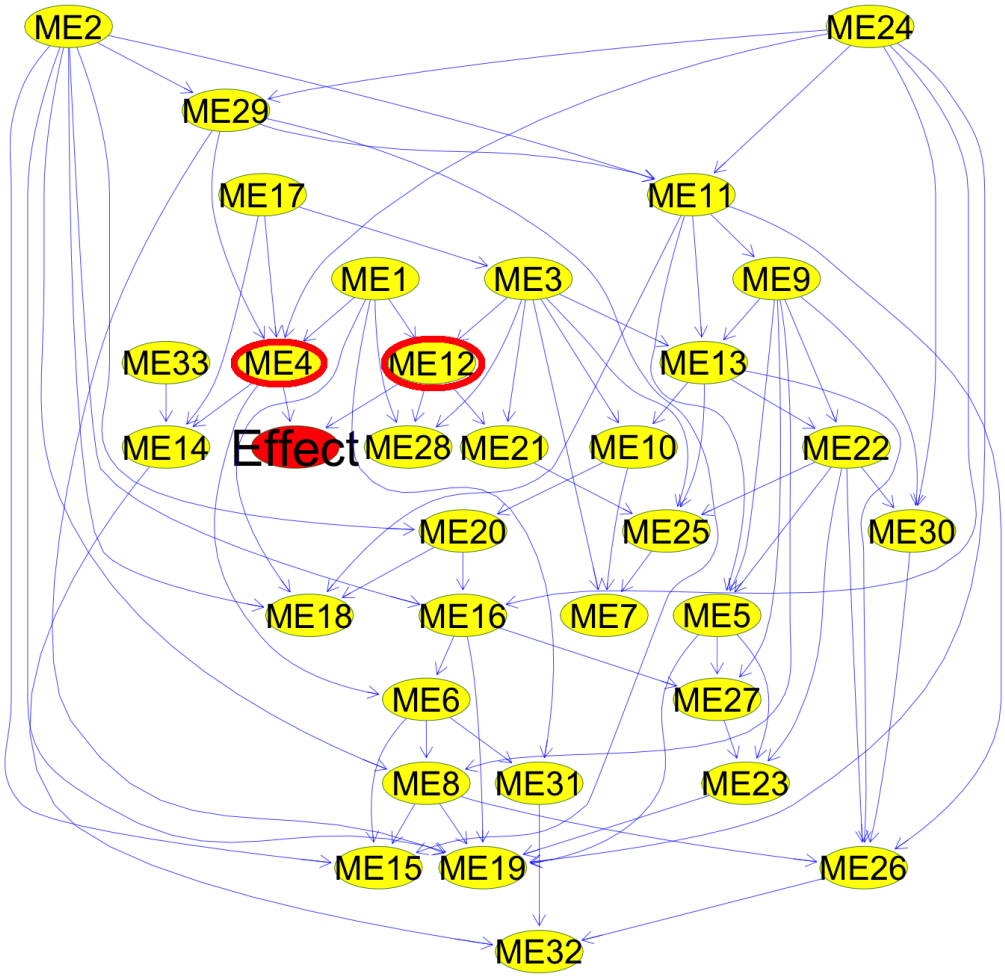


(a) BN structure.

```
Partition 3: Confusion matrix and statistics

            Actual
Prediction AML MDS
       AML 176   9
       MDS  26 155

               Accuracy : 0.9044                          95% CI : (0.8695, 0.9325)
    No Information Rate : 0.5519              P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8085           Mcnemar's Test P-Value : 0.006841

            Sensitivity : 0.8713                      Specificity : 0.9451
         Pos Pred Value : 0.9514                   Neg Pred Value : 0.8564
             Prevalence : 0.5519                   Detection Rate : 0.4809
   Detection Prevalence : 0.5055                 Balanced Accuracy : 0.9082

       'Positive' Class : AML
```

(b) Statistical measures.

BN structure and statistical measures for partition 3 of 5-fold CV.

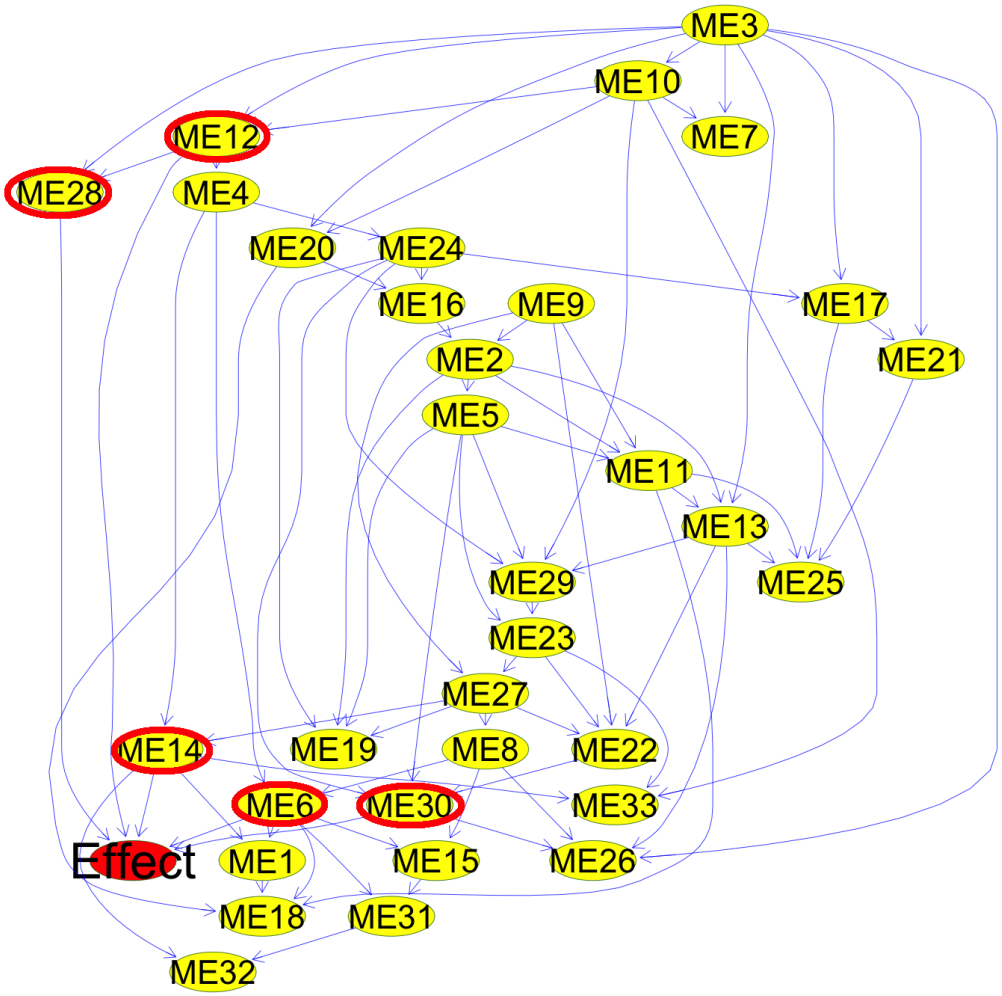


(a) BN structure.

```
Partition 4: Confusion matrix and statistics

          Actual
Prediction AML MDS
       AML 173  10
       MDS  29 154

               Accuracy : 0.8934                          95% CI : (0.8572, 0.9231)
    No Information Rate : 0.5519              P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7869           Mcnemar's Test P-Value : 0.003948

            Sensitivity : 0.8564                      Specificity : 0.9390
         Pos Pred Value : 0.9454                   Neg Pred Value : 0.8415
             Prevalence : 0.5519                   Detection Rate : 0.4727
   Detection Prevalence : 0.5000                 Balanced Accuracy : 0.8977

       'Positive' Class : AML
```

(b) Statistical measures.

BN structure and statistical measures for partition 4 of 5-fold CV.



(a) BN structure.

```
Partition 5: Confusion matrix and statistics

           Actual
Prediction AML MDS
       AML 195   9
       MDS   7 155

               Accuracy : 0.9563                          95% CI : (0.93, 0.9748)
    No Information Rate : 0.5519            P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9115         Mcnemar's Test P-Value : 0.8026

            Sensitivity : 0.9653                    Specificity : 0.9451
         Pos Pred Value : 0.9559                Neg Pred Value : 0.9568
             Prevalence : 0.5519                Detection Rate : 0.5328
   Detection Prevalence : 0.5574              Balanced Accuracy : 0.9552

       'Positive' Class : AML
```

(b) Statistical measures.

BN structure and statistical measures for partition 5 of 5-fold CV.

Average vote results on MILE dataset

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376265	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376379	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376187	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376274	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376060	AML	AML	MDS	AML	AML	AML	AML	TRUE
GSM376311	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376138	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376290	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376189	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376071	AML	AML	MDS	AML	AML	AML	AML	TRUE
GSM376070	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376141	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376364	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376409	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376142	AML	AML	AML	MDS	AML	AML	AML	TRUE
GSM376114	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376321	MDS	AML	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376348	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376215	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376305	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376157	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376163	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376394	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376241	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376410	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376386	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376292	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376094	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376411	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376051	AML	MDS	AML	MDS	MDS	MDS	MDS	FALSE
GSM376283	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376174	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376186	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376120	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376273	MDS	MDS	AML	AML	AML	MDS	AML	FALSE
GSM376498	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376067	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376092	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376072	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376408	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376315	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376123	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376254	AML	AML	AML	MDS	MDS	AML	AML	TRUE
GSM376126	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376340	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376151	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376357	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376414	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376423	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376168	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376256	AML	AML	MDS	AML	AML	AML	AML	TRUE
GSM376136	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376372	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376260	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376329	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376058	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376183	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376235	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376418	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376095	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376359	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376402	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376360	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376185	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376266	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376358	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376213	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376374	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376375	MDS	MDS	MDS	MDS	AML	MDS	MDS	TRUE
GSM376193	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376316	MDS	MDS	AML	MDS	MDS	MDS	MDS	TRUE
GSM376342	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376052	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376399	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376267	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376272	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376393	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376086	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376066	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376326	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376257	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376074	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376279	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376137	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376361	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376276	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376236	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376188	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376178	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376330	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376325	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376209	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376155	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376242	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376110	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376378	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376199	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376366	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376149	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376291	MDS	MDS	MDS	AML	MDS	MDS	MDS	TRUE
GSM376207	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376109	AML	AML	AML	MDS	AML	AML	AML	TRUE
GSM376054	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376262	AML	AML	AML	AML	MDS	AML	AML	TRUE
GSM376062	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376134	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376415	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376407	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376308	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376204	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376196	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376080	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376239	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376313	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376338	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376173	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376144	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376404	MDS	MDS	MDS	AML	AML	MDS	MDS	TRUE
GSM376400	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376268	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376057	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376050	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376076	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376210	AML	AML	AML	AML	MDS	MDS	AML	TRUE
GSM376113	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376049	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376063	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376169	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376293	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376282	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376425	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376323	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376322	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376307	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376365	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376250	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376264	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376089	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376426	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376331	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376143	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376085	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376055	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376205	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376312	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376097	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376339	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376059	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376107	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376401	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376093	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376380	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376077	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376277	MDS	MDS	MDS	AML	AML	MDS	MDS	TRUE
GSM376195	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376111	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376167	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376309	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376353	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376296	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376192	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376354	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376385	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376162	AML	AML	AML	AML	MDS	AML	AML	TRUE
GSM376384	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376299	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376153	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376346	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376146	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376119	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376125	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376159	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376334	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376218	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376135	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376271	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376310	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376075	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376497	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376203	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376412	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376172	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376381	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376324	MDS	MDS	AML	MDS	MDS	MDS	MDS	TRUE
GSM376166	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376201	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376270	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376251	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376336	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376087	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376295	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376333	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376180	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376341	MDS	MDS	MDS	MDS	AML	MDS	MDS	TRUE
GSM376191	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376300	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376212	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376261	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376128	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376140	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376287	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376171	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376100	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376388	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376053	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376238	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376317	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376289	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376373	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376301	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376245	AML	AML	AML	MDS	MDS	AML	AML	TRUE
GSM376129	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376098	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376344	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376165	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376337	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376068	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376347	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376499	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376170	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376208	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376106	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376345	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376416	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376421	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376150	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376198	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376078	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376395	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376184	AML	AML	MDS	AML	AML	AML	AML	TRUE
GSM376088	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376145	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376352	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376117	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376389	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376104	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376083	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376211	AML	AML	AML	AML	MDS	MDS	AML	TRUE
GSM376156	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376244	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376318	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376139	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376417	MDS	AML	AML	AML	AML	AML	AML	FALSE
GSM376064	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376383	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376275	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376099	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376133	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376124	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376367	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376420	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376090	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376406	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376382	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376319	MDS	MDS	MDS	AML	AML	MDS	MDS	TRUE
GSM376403	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376278	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376084	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376368	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376154	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376101	AML	AML	AML	MDS	MDS	AML	AML	TRUE
GSM376255	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376132	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376246	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376148	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376306	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376065	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376269	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376387	MDS	MDS	MDS	AML	AML	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376200	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376096	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376391	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376237	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376243	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376351	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376320	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376392	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376130	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376073	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376103	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376355	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376161	AML	AML	AML	MDS	MDS	AML	AML	TRUE
GSM376335	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376056	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376121	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376115	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376363	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376362	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376102	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376147	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376127	AML	AML	AML	AML	AML	MDS	AML	TRUE
GSM376253	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376152	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376286	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
GSM376122	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376176	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376179	AML	MDS	AML	AML	AML	AML	AML	TRUE
GSM376328	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376079	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376350	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376247	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376248	AML	MDS	AML	AML	AML	MDS	AML	TRUE
GSM376298	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376281	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376302	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376164	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376332	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376427	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376197	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376182	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376252	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376240	AML	MDS	AML	AML	AML	AML	AML	TRUE
GSM376349	MDS	AML	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376419	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376327	MDS	MDS	AML	MDS	MDS	MDS	MDS	TRUE
GSM376061	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376116	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376158	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376397	MDS	AML	AML	AML	AML	MDS	AML	FALSE
GSM376202	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376376	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376216	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376112	AML	MDS	AML	AML	AML	MDS	AML	TRUE
GSM376314	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376285	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376294	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376263	AML	AML	AML	AML	AML	MDS	AML	TRUE
GSM376396	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376206	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376091	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376249	AML	MDS	AML	AML	MDS	AML	AML	TRUE
GSM376303	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376190	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376082	AML	MDS	AML	AML	AML	AML	AML	TRUE
GSM376105	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376398	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376370	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376118	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376405	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376413	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376284	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376280	MDS	AML	MDS	AML	AML	AML	AML	FALSE
GSM376081	AML	MDS	MDS	MDS	MDS	AML	MDS	FALSE
GSM376217	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376214	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376422	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376369	MDS	AML	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376424	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376343	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376194	AML	AML	MDS	MDS	MDS	AML	MDS	FALSE
GSM376288	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376131	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376177	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376297	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376175	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376160	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376108	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376371	MDS	AML	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376377	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
GSM376258	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376390	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376259	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376069	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376356	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
GSM376181	AML	AML	AML	AML	AML	AML	AML	TRUE
GSM376304	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

APPENDIX C

Average vote results on BCCA dataset

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
A08838	AML	AML	AML	AML	AML	AML	AML	TRUE
A08843	AML	AML	AML	AML	AML	AML	AML	TRUE
A08852	AML	AML	AML	AML	AML	AML	AML	TRUE
A08853	AML	AML	AML	AML	AML	MDS	AML	TRUE
A08855	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08856	AML	AML	AML	AML	AML	AML	AML	TRUE
A08858	AML	AML	AML	AML	AML	AML	AML	TRUE
A08860	AML	AML	AML	AML	AML	AML	AML	TRUE
A08861	AML	AML	AML	AML	AML	MDS	AML	TRUE
A08862	AML	AML	AML	AML	AML	AML	AML	TRUE
A08863	AML	MDS	MDS	AML	AML	AML	AML	TRUE
A08864	AML	AML	AML	AML	AML	AML	AML	TRUE
A08865	AML	AML	AML	AML	AML	AML	AML	TRUE
A08866	AML	AML	AML	AML	AML	AML	AML	TRUE
A08867	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08868	AML	AML	AML	AML	AML	MDS	AML	TRUE
A08869	AML	AML	AML	AML	AML	AML	AML	TRUE
A08870	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08871	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08873	AML	AML	AML	AML	AML	AML	AML	TRUE
A08874	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08876	AML	AML	AML	AML	AML	AML	AML	TRUE
A08877	AML	AML	AML	AML	AML	AML	AML	TRUE
A08878	AML	AML	AML	AML	AML	AML	AML	TRUE
A08879	AML	AML	AML	AML	AML	AML	AML	TRUE
A08880	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08881	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08883	AML	AML	AML	AML	AML	AML	AML	TRUE
A08884	AML	AML	AML	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
A08885	AML	AML	AML	AML	AML	AML	AML	TRUE
A08886	AML	AML	AML	AML	AML	AML	AML	TRUE
A08887	AML	AML	AML	AML	AML	AML	AML	TRUE
A08888	AML	AML	AML	AML	AML	AML	AML	TRUE
A08890	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08891	AML	AML	AML	AML	AML	AML	AML	TRUE
A08892	AML	AML	MDS	AML	AML	AML	AML	TRUE
A08893	AML	AML	AML	AML	AML	AML	AML	TRUE
A08894	AML	AML	AML	AML	AML	AML	AML	TRUE
A08895	AML	AML	AML	AML	AML	AML	AML	TRUE
A08896	AML	AML	MDS	MDS	AML	AML	AML	TRUE
A08897	AML	MDS	MDS	AML	AML	AML	AML	TRUE
A08898	AML	AML	AML	AML	AML	AML	AML	TRUE
A08899	AML	AML	AML	AML	AML	AML	AML	TRUE
A08900	AML	AML	AML	AML	AML	MDS	AML	TRUE
A08901	AML	AML	AML	AML	AML	AML	AML	TRUE
A08902	AML	MDS	AML	AML	AML	AML	AML	TRUE
A08912	AML	AML	AML	AML	AML	AML	AML	TRUE
A15343	MDS	AML	MDS	AML	AML	AML	AML	FALSE
A15344	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15346	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15348	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15353	MDS	MDS	MDS	MDS	AML	AML	MDS	TRUE
A15302	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15308	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15311	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15317	MDS	MDS	MDS	MDS	AML	AML	MDS	TRUE
A15321	MDS	AML	MDS	AML	AML	AML	AML	FALSE
A15328	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15336	AML	AML	MDS	AML	AML	AML	AML	TRUE

Continued on next page

Table – continued from previous page

Sample	Actual	Part 1	Part 2	Part 3	Part 4	Part 5	Vote	Accuracy
A15337	AML	MDS	MDS	MDS	MDS	MDS	MDS	FALSE
A15338	AML	MDS	MDS	MDS	MDS	MDS	MDS	FALSE
A15340	AML	AML	MDS	AML	AML	AML	AML	TRUE
A15341	AML	MDS	AML	MDS	MDS	MDS	MDS	FALSE
A15362	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15365	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15367	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15372	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15376	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
A15378	MDS	MDS	MDS	MDS	MDS	AML	MDS	TRUE
A15381	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE
A15384	MDS	AML	AML	AML	AML	AML	AML	FALSE
A15388	MDS	MDS	MDS	AML	AML	MDS	MDS	TRUE
A15389	MDS	MDS	AML	AML	AML	MDS	AML	FALSE
A15390	MDS	MDS	MDS	MDS	MDS	MDS	MDS	TRUE

```
Confusion matrix and statistics on BCCA data using BN for partition 1:

          Actual
Prediction AML MDS
       AML  46   8
       MDS   6  14

               Accuracy : 0.8108             95% CI : (0.703, 0.8925)
    No Information Rate : 0.7027             P-Value [Acc > NIR] : 0.02469

                  Kappa : 0.535           Mcnemar's Test P-Value : 0.78927

             Sensitivity : 0.8846                    Specificity : 0.6364
          Pos Pred Value : 0.8519                 Neg Pred Value : 0.7000
              Prevalence : 0.7027                 Detection Rate : 0.6216
   Detection Prevalence : 0.7297               Balanced Accuracy : 0.7605

       'Positive' Class : AML
```

Performance measures on BCCA data for first partition of 5-fold CV.

```
Confusion matrix and statistics on BCCA data using BN for partition 2:

          Actual
Prediction AML MDS
       AML  36   7
       MDS  16  15

               Accuracy : 0.6892             95% CI : (0.571, 0.7917)
    No Information Rate : 0.7027             P-Value [Acc > NIR] : 0.65390

                  Kappa : 0.3346          Mcnemar's Test P-Value : 0.09529

             Sensitivity : 0.6923                    Specificity : 0.6818
          Pos Pred Value : 0.8372                 Neg Pred Value : 0.4839
              Prevalence : 0.7027                 Detection Rate : 0.4865
   Detection Prevalence : 0.5811               Balanced Accuracy : 0.6871

       'Positive' Class : AML
```

Performance measures on BCCA data for second partition of 5-fold CV.

```
Confusion matrix and statistics on BCCA data using BN for partition 3:

          Actual
Prediction AML MDS
       AML  48  10
       MDS   4  12

               Accuracy : 0.8108              95% CI : (0.703, 0.8925)
    No Information Rate : 0.7027              P-Value [Acc > NIR] : 0.02469

                  Kappa : 0.5085          Mcnemar's Test P-Value : 0.18145

            Sensitivity : 0.9231                    Specificity : 0.5455
         Pos Pred Value : 0.8276                 Neg Pred Value : 0.7500
             Prevalence : 0.7027                 Detection Rate : 0.6486
   Detection Prevalence : 0.7838               Balanced Accuracy : 0.7343

       'Positive' Class : AML
```

Performance measures on BCCA data for third partition of 5-fold CV.

```
Confusion matrix and statistics on BCCA data using BN for partition 4:

          Actual
Prediction AML MDS
       AML  49  12
       MDS   3  10

               Accuracy : 0.7973              95% CI : (0.6878, 0.8819)
    No Information Rate : 0.7027              P-Value [Acc > NIR] : 0.04551

                  Kappa : 0.45            Mcnemar's Test P-Value : 0.03887

            Sensitivity : 0.9423                    Specificity : 0.4545
         Pos Pred Value : 0.8033                 Neg Pred Value : 0.7692
             Prevalence : 0.7027                 Detection Rate : 0.6622
   Detection Prevalence : 0.8243               Balanced Accuracy : 0.6984

       'Positive' Class : AML
```

Performance measures on BCCA data for fourth partition of 5-fold CV.

```
Confusion matrix and statistics on BCCA data using BN for partition 5:

          Actual
Prediction AML MDS
       AML  45  12
       MDS   7  10

               Accuracy : 0.7432              95% CI : (0.6284, 0.8378)
    No Information Rate : 0.7027              P-Value [Acc > NIR] : 0.2660

                  Kappa : 0.3424           Mcnemar's Test P-Value : 0.3588

            Sensitivity : 0.8654                      Specificity : 0.4545
         Pos Pred Value : 0.7895                   Neg Pred Value : 0.5882
             Prevalence : 0.7027                   Detection Rate : 0.6081
   Detection Prevalence : 0.7703                 Balanced Accuracy : 0.6600

       'Positive' Class : AML
```

Performance measures on BCCA data for fifth partition of 5-fold CV.

REFERENCES

Albitar, M., Manshouri, T., Shen, Y., Liu, D., Beran, M., Kantarjian, H. M., ... others (2002). Myelodysplastic syndrome is not merely "preleukemia". *Blood*, *100*(3), 791–798.

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, *97*(18), 10101–10106.

Baker, P., Keck, C., Mott, F., & Quinlan, S. (1993). Nlsy child handbook, revised edition: A guide to the 1986-1990 national longitudinal survey of youth. *Columbus, Ohio: Center for Human Resource Research. Ohio State University*.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., ... others (2003). Computational discovery of gene modules and regulatory networks. *Nature biotechnology*, *21*(11), 1337–1342.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer.

Ben-Gal, I. (2007). Bayesian networks. *Encyclopedia of statistics in quality and reliability*.

Bishop, C. (2007). *Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn.* Springer, New York.

Christofides, N., & Theo-ry, G. (1975). *An algorithmic approach*. New York: Academic Press Inc.

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2001). Introduction to algorithmsm, chap. 16. *Greedy Algorithms. MIT Press, Cambridge*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Friedman, N., & Koller, D. (2003). Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, *50*(1-2), 95–125.

Haferlach, T., Kohlmann, A., Wieczorek, L., Basso, G., Te Kronnie, G., Béné, M.-C., ... others (2010). Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. *Journal of Clinical Oncology*, *28*(15), 2529–2537.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, *20*(3), 197–243.

Horvath, S. (2011). *Weighted network analysis: applications in genomics and systems biology*. Springer Science & Business Media.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jemal, A., Thomas, A., Murray, T., & Thun, M. (2002). Cancer statistics, 2002. *CA: a cancer journal for clinicians*, *52*(1), 23–47.

Jensen, F. V. (1996). *An introduction to bayesian networks* (Vol. 210). UCL press London.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Langfelder, P., & Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 1.

Langfelder, P., Mischel, P. S., & Horvath, S. (2013). When is hub gene selection better than standard meta-analysis? *PloS One*, *8*(4), e61505.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval* (Vol. 1) (No. 1). Cambridge university press Cambridge.

Miller, B. G., & Stamatoyannopoulos, J. A. (2010). Integrative meta-analysis of differential gene expression in acute myeloid leukemia. *PLoS One*, *5*(3), e9466.

Mills, K. I., Kohlmann, A., Williams, P. M., Wieczorek, L., Liu, W.-m., Li, R., ... others (2009). Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of aml transformation of myelodysplastic syndrome. *Blood*, *114*(5), 1063–1072.

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). Bayesian networks in r. *Springer*, *122*, 125–127.

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences*, *103*(47), 17973–17978.

Ozsolak, F., & Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, *12*(2), 87–98.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532–538). Springer.

Schuster, S., Dandekar, T., & Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology*, *17*(2), 53–60.

Shi, J., Shao, Z.-H., Liu, H., Bai, J., Cao, Y.-R., He, G.-S., ... others (2004). Transformation of myelodysplastic syndromes into acute myeloid leukemias. *Chinese Medical Journal*, *117*(7), 963–967.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, *62*(1), 77–89.

Tibshirani, R., & Hastie, T. (2007). Margin trees for high-dimensional classification. *The Journal of Machine Learning Research*, *8*, 637–652.

Wang, L., Gao, C., & Chen, B. (2011). [research progress on mechanism of mds transformation into aml]. *Zhongguo shi yan xue ye xue za zhi/Zhongguo bing li sheng li xue hui= Journal of experimental hematology/Chinese Association of Pathophysiology*, *19*(1), 254–259.

Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., & Jarvis, E. D. (2002). Using bayesian network inference algorithms to recover molecular genetic regulatory networks. In *International conference on systems biology* (Vol. 2002).

Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., ... others (2013). Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, *153*(3), 707–720.