A Hybrid Dynamic Bayesian Network Approach for Modelling Temporal Associations of Gene Expressions for Hypertension Diagnosis

Arinze Akutekwe and Huseyin Seker

Abstract— **Computational and machine learning techniques have been applied in identifying biomarkers and constructing predictive models for diagnosis of hypertension. Strategies such as improved classification rules based on decision trees have been proposed. Other techniques such as Fuzzy Expert Systems (FES) and Neuro-Fuzzy Systems (NFS) have recently been applied. However, these methods lack the ability to detect temporal relationships among biomarker genes that will aid better understanding of the mechanism of hypertension disease. In this paper we apply a proposed two-stage bio-network construction approach that combines the power and computational efficiency of classification methods with the well-established predictive ability of Dynamic Bayesian Network. We demonstrate our method using the analysis of male young-onset hypertension microarray dataset. Four key genes were identified by the Least Angle Shrinkage and Selection Operator (LASSO) and three Support Vector Machine Recursive Feature Elimination (SVM-RFE) methods. Results show that cell regulation FOXQ1 may inhibit the expression of focusyltransferase-6 (FUT6) and that ABCG1 ATP-binding cassette sub-family G may also play inhibitory role against NR2E3 nuclear receptor sub-family 2 and CGB2 Chromatin Gonadotrophin.**

I. Introduction

Nearly one out of three adults in the United States have hypertension and an estimated $47.5 Billion is spent on the disease each year [1]. Hypertension or high blood pressure is defined as systolic or diastolic blood pressure greater than or equal to 140mmHg and 90mmHg, respectively. Called a silent killer because it has no symptoms, hypertension is one of the major risk factors for developing heart disease. With risk factors such as smoking, alcohol and obesity, an estimated 32% of men and 29% of women in England have hypertension or have been treated of it [2]. In line with technological advances in the post-genome era, complex and high dimensional proteomic and gene expression profiles have been extracted and computational approaches to analyze and discover potential key biomarkers and relationships among genes is therefore necessary to understand the disease mechanism.

Various computational approaches have been proposed and applied in recent research on hypertension such as an improved C4.5 classification algorithm based on maximal information for identification of key hypertension biomarkers [3], Fuzzy Expert Systems (FES) and Neuro-Fuzzy Systems (NFS) for the diagnosis of hypertension [4] with a conclusion that NFS was more appropriate than FES. These methods are only able to analyze classification performance and perform "black-box" knowledge representation (NFS) but are not able to model temporal relationship among hypertension genes and key potential biomarkers for hypertension diagnosis.

This study therefore aims to investigate temporal association of high quality hypertension genes and discover potential time dependent biomarkers across two time points using a proposed two-stage computational approach [5]. At the first stage feature selection and classification are carried out using five different methods. High quality features based on best classification performance are selected. At the second stage, Dynamic Bayesian Network is used to model temporal associations across two time points of significant arcs. This promising hybrid method is discussed in the following sections in detail.

II. Methods and Materials

A. Feature Selection and Classification

In high-dimensional settings, feature selection helps to select the most important features in order to reduce their number, avoid over-fitting and at the same time, retain best class discriminatory information as much as possible.

There are many methods for feature selection widely practiced which can be grouped into three main techniques namely filter, wrapper and embedded techniques [6]. In this study, five wrapper feature selection techniques will be used as the wrapper-based methods provide information about interaction between the features selected, which generally select high-quality feature subsets.

i. Support Vector Machine Recursive Feature Elimination (SVM-RFE)

The SVM-RFE was used for selecting predictors relevant to cancer classification problem [7]. The method generates the ranking of features based on backward feature elimination by training a SVM. In this study, we considered the three popular kernels of the SVM: linear (SVM-RFE-L), Non-Linear (SVM-RFE-NL) and Gaussian radial basis function kernels of the SVM-RFE algorithm (SVM-RFE-RBF). For classification problems, SVM involves the minimization of the error function

The authors are with the Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK (email: aakutekwe@dmu.ac.uk, hseker@dmu.ac.uk). Corresponding author: hseker@dmu.ac.uk

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

$$\frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i \quad (1)$$

$$\text{s.t} \quad y_i(w^T\phi(x_i)+b) \geq 1-\xi_i, \xi_i \geq 0, i=1,....,N$$

where C is the capacity constant, w is the vector of coefficients, *b* is a constant and ξ_i represents parameters for handling non-separable data inputs.

ii. Random Forest Recursive Feature Elimination (RF-RFE)

The random forest method generates many trees using the idea of bagging in tandem with random feature selection and computes the ensemble of trees using

$$f(x) = \sum_{m=1}^{M}\frac{1}{M}f_m(x) \quad (2)$$

where *M* represents different trees to be trained on different subsets of data chosen at random and f_m is the *m*'th tree. The feature to split in each node in the tree is selected as the best among a set of randomly selected features. After generating a large number of trees, the most popular class in the trees is selected [8]. Recursive elimination is carried out by successively eliminating the least important variable based on decreased classification accuracy.

iii. Least Angle Shrinkage and Selection Operator (LASSO)

The lasso is an efficient regularization method to improve prediction performance and prevent over-fitting. It performs shrinkage and continuous subset selection for linear and logistic regression via an L_1-norm regularization penalty. The objective function is to minimize the sum of squared errors with a bound on the sum of absolute values of the coefficients [9]. The lasso estimate is defined by the maximum likelihood

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 \quad (3)$$

$$\text{subject to} \sum_{j=1}^{p}|\beta_j| \leq s$$

where $\sum_{j=1}^{p}|\beta_j|$ is the L_1 lasso penalty and *s* is the standardized tuning parameter that determines the amount of shrinkage.

B. Dynamic Bayesian Network (DBN)

Bayesian Networks (BN) are directed acyclic graphs (DAG) having nodes that represent random variables [10]. Each node x_i has a conditional probability distribution $p(x_i|parents(x_i))$ with its parent nodes and the joint probability distribution of all nodes given by:

$$p(x_1,....,x_n) = \prod_{i=1}^{n}p(x_i \mid parents(x_i)) \quad (4)$$

Dynamic Bayesian Networks (DBNs) are Bayesian Networks that aid modeling of associations arising from temporal dynamics in time between features of interest which otherwise cannot be performed using static Bayesian Network. It was proven in [11] that in modeling a DBN, arcs defining dependence relationships among variables of successive time points can be represented when a vector auto-regressive process (VAR) model of order 1 (VAR(1)) is assumed. Hence, a DBN can be expressed as

$$X_{(t)} = \alpha + \phi X_{(t-1)} + \omega_{(t)} \quad (5)$$

where ϕ is a k×k transition matrix that expresses the dependence of $X_{(t)}$ on $X_{(t-1)}$. $\omega_{(t)}$ is the vector white noise process assumed to be multivariate normal with mean zero and covariance matrix $E\{\omega_{(t)},\omega_{(t)}'\} = \sum_{\omega}$. The stochastic transition model for the process is of first-order Markovian. For all time points $t > 0$, the random variables $X(t) = (X_1(t),...,X_i(t),...,X_k(t))$ observed at time *t* are conditionally independent given the random variables $X(t-1)$ at previous time *t-1*. This implies that at any time point, simultaneously observed variables are conditionally independent given their immediate past. Hence a VAR(1) process whose covariance matrix $\sum_{\omega}$ is diagonal can be represented as a DBN where the arcs of the DBN model are identified by the non-zero elements of the matrix ϕ.

DBNs are used to represent directed graphical stochastic models of dynamical systems that are problem specific. Variants of the Hidden Markov Models (HMMs), which are tools that can represent probability distributions over sequence of observations, can be considered to be DBNs [12]. HMMs, which can be seen as special cases of DBNs, are ubiquitous for modeling time series data where they are used to encode structures that are implied and not fully expressed in a DBN [13]. This allows modelling of temporal feedback loops that are common in biological pathways, where parent genes inhibit or slow down the expression and chemical reaction of child genes [11, 14].

Different shrinkage algorithms for learning and inference of DBN models for biological pathways using regularized estimators are studied in [14]. In this paper, we applied the G1DBN algorithm [11, 14] for DBN modelling in order to determine significant temporal relationships among features that yield high accuracies selected by the feature selection methods.

III. Results and Discussion

The hypertension pre-processed dataset was obtained from [15]. It contained gene expression profiles of male young-onset hypertension with age range of 20-50 years. The microarray chip contained 39,200 polynucleotide data, of which 22,184 were mapped to the human genome. There were 77 cases of male young-onset (age 37.6 ± 7.2) hypertensive and 82 male normotensive controls (age 36.9 ± 6.6). These make up a total of 159 observations and 22,184 predictor genes used in this study. To be able to determine relevant temporal interaction among the hypertension genes, feature selection was first carried out using five different methods in order to select best features with the highest class discriminatory information. For all the selection methods, a 10-fold cross validation was performed and their performances were averaged. We used popular performance

criteria which are sensitivity, specificity, accuracy, standard deviation of accuracy (Std.Acc), false positive rate (Type 1 Error), Balanced Classification Rate (BCR), F1-score and Matthew's Correlation Coefficient (MCC), as shown in Table I, to rank the performance of the methods and selected features. Mathematical expressions of these measurements can be found in [12].

There were a total of 101 features selected by the LASSO which had non-zero coefficients. The selection was achieved using the value of the standardization parameter *s* obtained from 10-fold cross validation. For the SVM-RFE methods, cost of constraint violation C was varied between 1 and 10. Finally, C=1 that yielded best accuracy was therefore chosen. For SVM-RFE NL and SVM-RFE RBF, the best gamma value of 0.0000451 corresponding to the inverse of dataset dimension was chosen. For SVM-RFE Non-Linear, moderate polynomial degree of 3 was used and the best-selected features were ranked based on the highest accuracy of classifier. The SVM-RFE Linear, SVM-RFE Non-Linear, and SVM-RFE Radial selected 137, 45 and 49 best features respectively. For the RF-RFE, the experiment was run for 200, 500 and 1000 iterations each for 1000, 2000 and 5000 randomly initialized forest trees. 320 top features selected by backward elimination on random forests had overall best performance. Table 1 shows the results of the best model with best performance criteria.

For modelling the temporal associations of selected features using DBN, we considered the features selected by the LASSO and SVM-RFE-Linear methods as they yielded higher accuracies of 99% and disregarded the feature selected by the RF-RFE, SVM-RFE Non-Linear and SVM-RFE Radial methods which had lower accuracies.

The Dynamic Bayesian Network for the genes selected by the LASSO and SVM-RFE-Linear was modelled using the G1DBN algorithm with each observation taken as a time point. The G1DBN algorithm performs DBN modelling in two main steps. The first step infers a first order dependence score matrix (S1) which contains the score of each edge of the DBN. S1 and edge selection threshold alpha1, obtained in the first step are used with edge selection threshold alpha2 in the second step, to infer the score of each edge of a DBN describing full order dependencies between successive variables. The smallest score refers to the most significant edge. In order to obtain optimized DBN, threshold values of alpha1 and alpha2 for the two steps are found to be 0.5 and 0.05, respectively, which were used to prune the edges of the DBN model.

We discovered 351 and 460 directed arcs describing full order conditional dependencies among the selected features of LASSO and SVM-RFE Linear respectively. The score matrices of the discovered arcs ranged from 0.000 to 0.05. Fig 1 shows the DBN model of 13 genes selected by the LASSO with most significant inferred temporal arcs across time points having scores of 0.000. Fig 2 shows the DBN model of 15 genes selected by SVM-RFE Linear with most significant inferred temporal arcs having scores of 0.000. The current meanings of the represented genes were verified from Gene Expression Omnibus [17] and genes without gene symbols were represented by their UniGene IDs (starting

TABLE I. SUMMARY OF BEST PERFORMANCE CRITERIA

Performance Criteria	Feature Selection Wrapper Methods				
	RF-RFE	***LASSO***	***SVM-RFE-L***	***SVM-RFE-NL***	***SVM-RFE-RBF***
	Classifiers				
	Decision Tree	***L1 Logistic Regression***	***SVM-L***	***SVM-NL***	***SVM-RBF***
no. of genes	320	**101**	137	45	49
Sensitivity	0.6292	**0.9857**	0.9889	0.9260	0.9975
Specificity	0.5854	**1.0000**	0.9850	0.8470	0.9514
Accuracy	0.6067	**0.9900**	0.9900	0.8867	0.9600
Std.Acc	0.1235	**0.0211**	0.0281	0.0892	0.0562
Type 1 Err.	0.4146	**0.0139**	0.0145	0.1530	0.0486
BCR	0.6073	**0.9923**	0.9894	0.8865	0.9595
F1-Score	0.5965	**0.9929**	0.9850	0.8949	0.9531
MCC	0.2164	**0.9873**	0.9739	0.7800	0.9197

BCR: The Balanced Classification Rate (Harmonic mean of the Precision and Recall).
Type 1 Err: The False Positive Rate (FPR). Std.Acc: The standard deviation of mean accuracy.

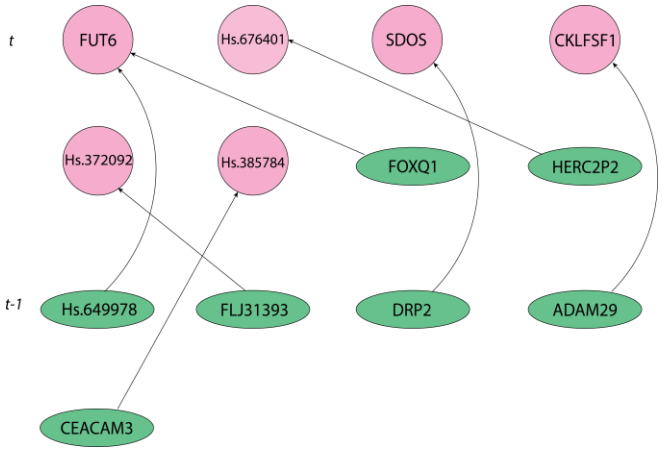

Figure 1. Dynamic Bayesian Network Model of Hypertension genes showing temporal relationships of key features. It is worth noting that the green ellipses are features of the predicted parents at time t-1 which inhibit the children shown in the pink circles at time t.

with Hs.) in the diagrams. From Fig 1 results, the human transcribed locus gene (Hs.649978) and cell regulation FOXQ1 may inhibit the expression of focusyltransferase-6 (FUT6). The ABCG1 ATP-binding cassette sub-family G may also play inhibitory role against NR2E3 nuclear receptor sub-family 2 and CGB2 Chromatin Gonadotrophin as shown in Fig 2.

We extracted 22 more significant gene subsets, which were picked up in common by the LASSO and the SVM-RFE Linear selection methods and inferred their DBN model. Fig 3 shows the temporal associations of 11 genes with six most significant edges of lowest scores. The result show that CRABP2 cellular retinoic acid binding protein 2 may be highly associated in time with DKK3 dickkopf WNT signaling pathway inhibitor 3.

There were a total of four key genes selected in common by the more robust LASSO and the three SVM-RFE methods and are shown in Table II. These genes that are consistent

TABLE II. KEY GENES COMMONLY SELECTED BY LASSO AND SVM-RFE METHODS

UniGene ID	Symbol	Name
Hs.666652	null	Human transcribed locus
Hs.73839	RNASE3	ribonuclease; RNase A family; 3 (eosinophil cationic protein)
Hs.497626	PLXNA2	Human protein-coding gene PLXNA2
Hs.656129	null	CDNA FLJ36210 fis, clone THYMU2000155

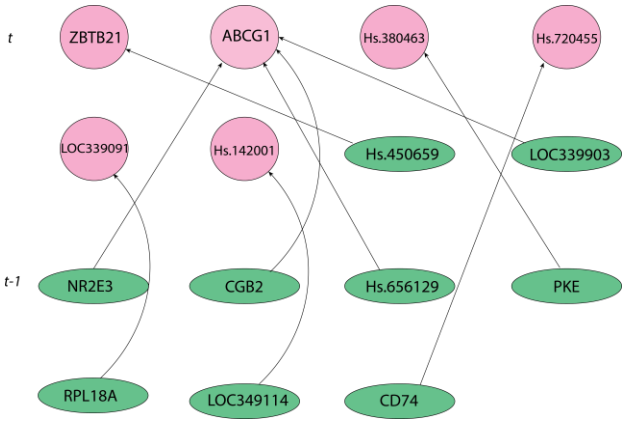

Figure 2. DBN Model of Hypertesion genes showing temporal relationship among top genes selected by SVM-RFE Linear method.

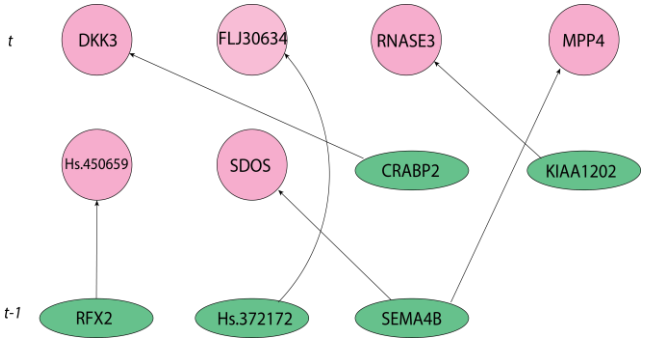

Figure 3. DBN of Hypertension genes showing temporal relationship among key features selected by both the LASSO and SVM-RFE Linear methods (the green spheres are features of predicted parents at time t-1 which inhibit features in red circles (children) at time t.

across the LASSO and SVM-RFE methods of high accuracies might be highly associated with the hypertension disease.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, a hybrid Dynamic Bayesian Network modeling and inference made up of five feature selection methods and Dynamic Bayesian Network was successfully applied to the discovery of possible key biomarkers for the diagnosis of hypertension.

The results obtained from comprehensive analyses showed that the features selected by the LASSO and SVM-RFE Linear methods yielded the best predictive performance. DBN defining a VAR process of order 1 was used to perform inference on temporal relationship of the selected subsets from time *t* to *t-1*. Significant temporal relationships between genes FOXQ1 and FUT6 and between ABCG, NR2E3 and CGB2 were also discovered. We also discovered that DKK3 dickkopf WNT signaling pathway inhibitor 3 may be highly associated with CRABP2 cellular retinoic acid binding protein 2 in hypertension diagnosis.

Further examination will be carried out in the databases to find out the relationship between the modeled genes and the disease in gene expression databases. Other DBN algorithms that do not depend on regularized estimators such as Gibbs Sampling and Loopy Belief Propagation could also be experimented with and results compared.

REFERENCES

[1] Centre for Disease Control and Prevention (CDCP) [Online] Available from: http://www.cdc.gov/bloodpressure/facts.htm [Accessed Apr. 2014].

[2] British Heart Foundation (BHF) [Online] Available from: http://www.bhf.org.uk/research/heart-statistics/risk-factors/blood-pressure.aspx [Accessed Apr. 2014].

[3] Z. Wei, Z. Xuan, C. Junjie, "Study on classification rules of hypertension based on decision tree," *Software Engineering and Service Science (ICSESS), 4th IEEE International Conference on*, pp.93-96, May 2013.

[4] S. Das, P.K. Ghosh and S. Kar, "Hypertension diagnosis: A comparative study using fuzzy expert system and neuro fuzzy system," *IEEE International Conference on FUZZY SYSTEMS*, pp.1-7, July 2013.

[5] A. Akutekwe and H. Seker. "Two-Stage Computational Bio-Network Discovery Approach for Metabolites: Ovarian Cancer as a Case Study." (to appear in) *Proc. of IEEE-BHI 2014*, Valencia, Spain, 1-4 June 2014.

[6] Y. Saeys, I. Inza and P. Larrañaga. "A review of feature selection techniques in bioinformatics" *Bioinformatics,* Vol.23(19): pp.2507-7, 2007.

[7] I. Guyon, J. Weston, S. Barnhill and V. Vapnik "Gene selection for cancer classification using support vector machines". *Machine learning*, Vol.46(1-3), pp. 389-422. Jan. 2002

[8] L. Breiman. "Random Forests" *Machine Learning*. Vol. 45(1), pp. 5-32, Jan. 2001.

[9] T. Hastie, R. Tibshirani and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2011.

[10] S. Russell and P. Norvig, *Artificial intelligence : a modern approach* Upper Saddle River, N.J.: Prentice Hall, 2010.

[11] S. Lebre. "Stochastic process analysis for Genomics and Dynamic Bayesian Networks inference." PhD dissertation, Université d'Evry-Val d'Essonne, 2007.

[12] K. Murphy. "*Machine learning: a probabilistic perspective*". Cambridge MA: MIT Press, 2012.

[13] D. Koller and N. Friedman*." Probabilistic graphical models: principles and techniques"*. Cambridge MA: MIT press, 2009.

[14] R. Nagarajan, M. Scutari and S. Lèbre, "*Bayesian Networks in R with Applications in Systems Biology"*. New York: Springer, 2013.

[15] K.S, Lynn; L. Li-Lan; L. Yen-Ju; W. Chiuen-Huei; S. Shu-Hui; L. Ju-Hwa; L. Wayne; H. Wen-Lian and P. Wen-Harn, "A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data." *Bioinformatics* Vol.25(8): pp.981-988, Feb. 2009.

[16] National Center for Biotechnology Information, U.S. National Library of Medicine, Gene Expression Omnibus [Online] Available from: https://www.ncbi.nlm.nih.gov/gene [Accessed Apr. 2014]