

CORONARY ARTERY DISEASE PREDICTION USING DNA MICROARRAYS, NEURAL NETWORKS AND OTHER STATISTICAL ANALYSIS TOOLS

W.C. CHIN¹ AND C.K. THAM²

Department of Electrical and Computer Engineering, National University of Singapore

Abstract: *This paper aims to illustrate a novel approach of complex disease prediction, exemplified by a coronary artery disease (CAD) study that we have developed. This multidisciplinary approach straddles fields of microarray technology and genetics, neural networks (NN), data mining and machine learning, as well as traditional statistical analysis techniques, namely principal components analysis (PCA) and factor analysis (FA). A description of the biological background of the study is given, followed by a detailed description of how the problem has been modeled for analyses by neural networks and FA. A committee learning approach for NN have been used to improve generalisation rates. It has been shown that our NN approach was able to yield promising prediction results despite using only the most fundamental network structures. More interestingly, through the statistical analysis process, genes of similar biological functions have been clustered. In addition, a gene marker involved in breaking down lipids has been found to be the most correlated to CAD.*

Keywords: DNA microarray, committee learning, neural networks, data mining, machine learning, principal components analysis, factor analysis.

1. Introduction

Coronary artery disease (CAD), or ischaemic heart disease is not a single disease but is a combination of many individual diseases as listed under the 9th Revision of the International Classification of Diseases (1975). It includes acute myocardial infarction and angina pectoris among others. It is a complex multifactorial process that involves lipid deposition on arteries of the heart, macrophages, blood pressure, rheology of blood flow, smooth muscle proliferation, thrombogenesis, platelet aggregation, insulin resistance and other factors. Every year, millions of deaths worldwide are attributed to CAD and more than half of them are found in developed countries.

With the growing affluence of Singapore, a small island state located at the southern tip of the Malaysian peninsula, the disease patterns and health needs of Singaporeans have changed widely. Chronic degenerative diseases such as cancer and cardiovascular disease have emerged as the major causes of death and hence, finding cost-effective methods to control CAD is one of the challenges for public health in Singapore today.

The risk factors for CAD had been documented and among the more established ones are: family history (genetic factors), plasma lipid, lipoprotein, plasma lipoprotein (a), diet, gender, elevated blood pressure, physical inactivity etc [1].

¹ Corresponding author: Chin Wei Chuen, DSO National Laboratories, 20 Science Park Drive Singapore 118230. Tel: 65-67727148. Email: cweichue@dso.org.sg

² Corresponding author: Dr CK Tham, National University of Singapore, Department of Electrical and Computer Engineering, 10 Kent Ridge Crescent Singapore 119260. Tel: 65-68747959. Email: eletck@nus.edu.sg

2. Background

Microarray technology [2][3] is rapidly advancing and its applications to mutation detection diagnostics, gene discovery, gene expression and mapping have already been well demonstrated. In this study, we applied a DNA microarray-based genotyping method to obtain genetic information that is subsequently combined with other risk factors to predict and individual's risk of a complex disease. A chip-based minisequencing method [4] was used for genotyping. Alleles were discriminated by fluorescence-labeled dideoxynucleotides that are complementary to the polymorphic sites in amplified target DNA. The dideoxynucleotides were extended from primers that were pre-spotted on microarray glass slides. The intensities of fluorescent signals were analysed to yield the genotype calls and constitute an individual's genetic profile. This system is similar to the laboratory information management systems and databases described by Bassett [5].

Neural networks and other statistical pattern recognition techniques were employed for analysis of the database. The overall workflow implementation is illustrated in Figure 1 below.

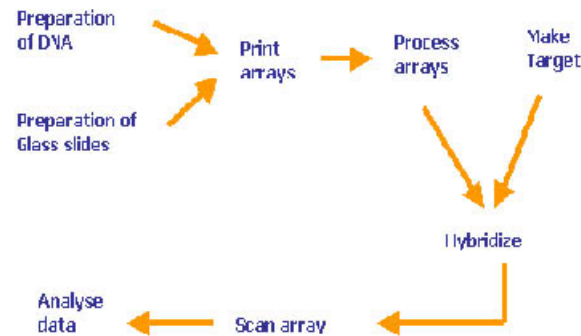


Figure 1. Experimental procedures (workflow) illustrating how DNA microarray preparation and eventually leads to data analysis

3. The HEART Dataset

Prior to the workflow implementation, a prototype dataset was needed for the development of the neural network algorithms. *HEART* contains the medical profiles of 704 human subjects. Of these, 56% were healthy at the time of recruitment and 44% were angiographically diagnosed to have coronary artery disease (CAD), i.e. with a minimum 50% stenosis in at least one of the major coronary arteries.

HEART consists of the CAD status of the patient, “0” for healthy and “1” for diseased and 29 other risk factors or ‘attributes’ of CAD. Among these were 19 CAD candidate gene markers and 10 non-genetic factors. Each genetic attribute denotes the *genotype* of a patient with respect to the gene concerned while each of the non-genetic or (environmental) attribute measures different aspects of the patient's phenotype or medical condition.

3.1. Adaptation of HEART

The *HEART* dataset contains 23 categorical and 6 continuous attributes. Of the 23 categorical attributes, 19 were genetic and 4 were non-genetic. The non-genetic attributes consist of information such as subjects' race, sex, smoking habits and family history etc. The term “non-genetic attributes” is loosely defined here as any attribute that was not obtained by genotyping, since race, sex and family histories were all determined by genes. The continuous attributes consist of information such as plasma cholesterol levels, body-mass index (BMI) and age. The sections below describe how the *HEART* dataset were treated prior to network training.

3.1.1. Binarisation of the dataset

As the *HEART* attributes are made up of a comprehensive combination of the genetic and physical traits of each individual, an adaptation method is needed to convert the mixture of categorical and continuous inputs into a suitable format for the prediction algorithms. The approach adopted here is by means of converting of all input attributes into a *binary stream* that serves as a single input vector for each subject. This process will be termed *binarisation*. The binarisation process involves converting all input factors into a categorical form such that a digit '1' would represent the "class" for the attribute in question.

Continuous attributes were binarised by splitting each of the individual factors into several histograms, or "bins" that reflect their relative positions in that group. Taking the "age" variable for example: the age range for *HEART* subjects is from 18 to 81 years. Dividing this range of 63 years into 6 different bins, with each bin having a range of ten years (except those at the ends of the distribution), a hypothetical subject that is 35 years of age will fall into the 30-39 year old bin and hence be classified as "0 1 0 0 0 0" for his "age" input, where the '1' in the stream indicates the '30-39' bin. It follows that the other bins are indicated '0'. For multi-categorical data, conversion was straightforward such that an attribute that has n possible classes ($1, 2, 3, \dots, n$) would be represented by n bits of numbers. The salient feature for this approach is each variable can only have one "bin" activated (=1) at any time and the number bins can be further increased to accommodate any input level of granularity desired.

Similarly, a categorical attribute (such as gene "x") having 6 classes (0,1,2,3,4,5) would be transformed into a continuous stream of 6 binary numbers, such that if a subject belongs to class "4", he/she would have "0 0 0 0 1 0" as his/her "x" attribute.

3.1.2. Modified cholesterol attributes and analysing data from various aspects

As the lipid profile levels (cholesterol attributes) of all diseased subjects had apriori (to the admission of the study) been drastically modified by lipid-lowering drugs, their cholesterol attributes cannot be used for prediction. Hence, *HEART* will be studied from three different aspects, namely "*FULL*" (consisting of all attributes less cholesterol), "*GENES*" (consisting of genetic attributes only) and "*ENV*" (consisting of environmental attributes less cholesterol). Essentially, after adaptation of the dataset, the *FULL*, *GENES* and the *ENV* aspects mapped into 93-bit, 65-bit and 28-bit binary input vectors respectively.

4. The "paradigm-shift" and the emergence of neural networks in complex diseases research

Due to the complex interplay of biochemical, environmental and genetic factors, the etiology of CAD is relatively ill understood hitherto. In recent decades, however, Strohman [6] had pointed out the need for a "paradigm-shift" in the study of complex diseases, (including CAD), because "*traditional tools and approaches in the fields of genetics and epidemiology are incompatible with what is now known as the biology of CAD*". This gives rise to the need to develop new analytical tools appropriate to the new paradigm. Specifically, the new tool should be able to achieve two new objectives [7], over its pre-processors:

- a. Allow for more than one model for different subsets of the data.
- b. Allow for complex interactions between attributes and the class.

In addition, the new tools (applies to the old tools as well) should be able to allow for attributes that are not independent and to minimise the validation and generalisation errors in each of the models generated. Other considerations are to minimise the number of models generated to fully explain the data and to allow for symmetry in classification errors.

In this context, neural networks (NNs) arise as a natural tool for the task as they are relatively easy to develop compared to traditional algorithms like logistic regression (LR) [8], which entail formal statistical training. Multiple models for any subset of the data can be readily analysed using NNs and complex interrelationships between variables are allowed since NNs do not assume dependence or independence of input variables. Incidentally, any data set that can be analysed using traditional techniques like LR can also be used to develop a NN model. And because predictor variables in a NN undergo a non-linear transformation at each computational and output node, implicit detection of complex non-linear relationships between independent and dependent variables can be readily accomplished [9].

5. The committee learning approach to CAD prediction using neural networks

The common experimental approach for the two NN algorithms investigated in this project is by means of committee learning [10] or also commonly known as “ensemble of classifiers”. This approach of prediction is gaining popularity in the machine-learning field, indicated by Dietterich [11]. Incidentally, a committee of several NN models is created and each NN is sequentially trained on individual subsets of the training data. Each subset of the training data is a randomised extraction from all the available samples. For all simulations in this project, the ratio of training to validation samples has been set to 2:1. The structure of the network models are all fixed and predetermined.

For prediction, individual predictions of each model were summed up and then ensemble averaged to produce the final prediction. Experiments have shown that such methods are superior (in terms of generalisation rates) to training individual network models that does not utilise the full set of available training data.

6. The Multi-Layer Perceptron (MLP)

The MLP [10] is one of the most popular and most important class of NNs that has been applied successfully to solve many difficult problems. In view of its popularity, robustness and relative ease of development, it is the algorithm of choice for this context. In addition, the validation and training results that MLP is able to achieve can readily be used as a benchmark index for other algorithms to compare against.

6.1. Design of the multi-layer perceptron

Design of the MLP network structure usually involve extensive preliminary cross validation [10] experiments to set its network free parameters like the number of hidden nodes and the learning rate. Other computationally intensive algorithms that make use of Bayesian techniques to infer network structure and weights have been investigated [12] [13]. In this project, a simple method for model selection [14] had been used for the sake of economy of effort. (Essentially, it is using extensive cross-validation empirical trials to determine the network structure that works on the benchmark dataset; and then subsequently reusing the network for other datasets). The following subsections describe the design process for the optimum MLP structure that would be replicated in all subsequent experiments.

6.1.1. Determining the number of hidden nodes

The problem of determining the number of hidden nodes is commonly solved by cross validation, which usually involves exhaustive measuring the average validation error of different network models and selecting the one with the minimum validation error.

Reflecting practical approaches to the problem, a quick method has been devised to select the optimum number of hidden nodes. The approach is to compare the validation set performances of holdout experiments with 2, 5, 10 and 15 hidden nodes using the entire training data. The MLP configuration that yields the minimum test set error is hence selected

for subsequent experiments. This configuration will be replicated for all other datasets to avoid having to run separate model selection experiments for every subsequent dataset.

6.1.2. Determining the learning rate

The learning rate of MLPs is another tricky variable to set. “Backprop approximates the trajectory in weight space computed by the method of steepest descent” [10]. Therefore, the smaller we make the learning rate parameter, η , the smaller but smoother is the change in synaptic weights from one iteration to the next. This comes with a trade-off of the need for a larger number of training epochs. However, too large a learning rate, realized by large a η value, would cause the weights to change abruptly, causing oscillations and instability of the network.

Theoretically, for determining the ‘optimal’ values of the learning rate parameter η , any of the following three rules [10] may be used:

- a. The η parameter that on average yields convergence to a local minimum in the error surface of the network with the least number of epochs.
- b. The η parameter that, on average or in the worst-case, convergence to the global minimum in the error surface with the least number of epochs.
- c. The η parameter that on average yields convergence to the network configuration that has the best generalisation over the entire input space, with the least number of epochs.

Again, reflecting practical approach to the problem, a learning rate, η was determined empirically in preliminary holdout experiments involving η values of 0.001, 0.005, 0.025, 0.075 and 0.375. Similarly, the MLP configuration that yields the minimum test set error is hence selected for all subsequent experiments.

6.1.3. The final MLP configuration

From the above model selection experiments, the final configuration of the MLP utilises 10 hidden nodes and 1 output node. The optimal learning rate was found to be 0.005 and the initial starting weight parameters were set to range from -0.02 to 0.02, sampled from a random, uniform distribution. A momentum constant of 0.9 is used for all simulations as a means to accelerate the convergence of the backpropagation algorithm [10].

6.1.4. Training method for MLP

For all following simulations using MLPs, the batch-mode of the backpropagation algorithm using gradient-descent [10] is followed. The cost function to be minimized is the mean sum-of-squares error function [15].

For complexity control, a hybrid method inspired by Carl Rasmussen [12] has been adopted. It combines ideas from cross validation and early stopping [10] to obtain the best of both worlds. It has been reported that this method is competitive with other methods, e.g. MLPs trained with Markov Chain Monte Carlo (MCMC) [12]. Each cross validation ‘fold’ was set to run for 100 epochs, but the weight parameters for the epoch with the minimum validation error were saved. Lastly, ensemble-learning ideas sets in, whereby 50 ‘folds’ of cross validation experiments were run so as to generate 50 network models for prediction. The combined performance of 50 networks is then ensemble averaged to produce the final prediction.

7. The Hierarchical Mixture of Experts (HME)

A more recent algorithm, the HME, is compared to MLP in this project. HME is reputed to have superior generalisation capabilities and has a particular advantage over MLPs due to of the elimination of the *temporal-crosstalk* phenomenon (that is, the process of “learning” a

new sample may actually undo what had been learnt in the previous sample) that has always plagued MLPs. Also, by means of its *divide-and-conquer* [10] strategy, subset of the input space can be better and more accurately modeled than MLPs. Furthermore, its close resemblance to decision trees (CART) [16], and its belief network formulation propels it to prominence in this complex disease risk estimation and prediction application.

For a detailed treatment of the mathematical foundations of HME and the closely related EM algorithm for optimisation, the description in the authoritative paper by Jordan and Jacobs [17] was followed. In particular, the notation introduced by Waterhouse [13] is directly relevant in this context.

7.1. Design of the hierarchical mixtures of experts

In HME model selection, the use of automatic learning methods had been investigated Waterhouse [13] and Rasmussen [12]. Such methods are actually “approximates” to the number of gating and expert networks required for each given dataset. It also avoids having to rely on the crude approach of running preliminary simulations to determine the network structure each time there is a new dataset. Two types of HME architectures were implemented: namely, the fixed-architecture HME (which basically has the number of experts and gates fixed in advance and remains fixed throughout training) and the growing-architecture HME (in which the experts and gates are grown and pruned during training). In this project, for the purpose of comparing the utilities of HME with MLP, only the fixed-architecture HME is investigated in this paper. In the following subsections, the two fixed-architecture HME structures (the single-layer HME and the multi-layer HME) and related training methods are covered.

7.1.1. The design of the single-layer HME

For the single-layer HME, the depth of the network is fixed at 1. Only one gating network is required. The only calculation that needs to be done is to determine the number of experts to use. The number of experts, I is decided by balancing the number of weight parameters and the number of points in the training set; i.e.

$$I = \frac{N}{(d+1)(k+1)}$$

It is interesting to note from the formula that more resources would be allocated to expert networks.

7.1.2. The design of the multi-layer HME

For the multi-layer HME, the number of branches is fixed at 2 to reflect practical approaches to the problem. The only free parameter to determine for the binary tree is the tree depth, M :

$$M = \frac{1}{\log 2} (\log(N + 2(d+1)) - \log(d+1) - \log(k+2))$$

where, similar to above, N is the number of training samples, D is the input dimension and K is the output dimension.

For the multi-layer HME, proportionally more resources would be allocated to gating networks instead.

7.1.3. Training methods for HME

For ease of interpretation, linear experts and gates have been used for all models. The probability distribution of the target generative process is the exponential of the negative of the cross-entropy between the targets and the outputs of the expert. The optimisation process is a combination of the EM and conjugate gradient algorithms [15].

Complexity control takes the shape of early stopping via cross validation in committee learning. However, a number of additional search rules [13] have been implemented in the hope of less crude rules for training. For example, searches for additional minima after finding the first minimum are pursued such that the final iteration is at least 1.5 times the last minimum occurred. The number of iterations each for committee member has been set to range between 30 to 200 epochs. The number of committee members is dynamically determined to range from 3 to 50. Committee members are not further spawned if the last member took more than $1.875 \times N$ seconds to execute, where N is the number of points in the training set. (*In actual simulations however, all experiments proceeded to the minimum number of epochs (30) and have spawned the maximum number (50) of committee members.*)

8. Simulation Results on HEART

This section presents all the simulation results obtained. A comparison is made based on the validation and generalisation results obtained via MLP, single-layer HME and the multi-layer HME. The *GENES* dataset has 308 validation points and 103 test points while all other aspects have 211 validation points and 71 test points. The validation and generalisation results presented are the ensemble-averaged statistics for 50 committee machines.

<i>FULL dataset</i>			
	<i>MLP</i>	<i>Single-Layer HME</i>	<i>Multi-Layer HME</i>
<i>Rec (V)</i>	174.4	177.5	177.7
<i>Recognition rate (V)</i>	82.65%	84.12%	84.22%
<i>Rec (G)</i>	58	57	57
<i>Recognition rate (G)</i>	81.69%	80.28%	80.28%

Table 1. Classification results for the *FULL* dataset

<i>GENES dataset</i>			
	<i>MLP</i>	<i>Single-Layer HME</i>	<i>Multi-Layer HME</i>
<i>Rec (V)</i>	183.1	183.9	183.1
<i>Recognition rate (V)</i>	59.45%	59.71%	59.45%
<i>Rec (G)</i>	77	74	77
<i>Recognition rate (G)</i>	74.76%	71.84%	74.76%

Table 2. Classification results for the *GENES* dataset

<i>ENV dataset</i>			
	<i>MLP</i>	<i>Single-Layer HME</i>	<i>Multi-Layer HME</i>
<i>Rec (V)</i>	176.6	180.6	180.3
<i>Recognition rate (V)</i>	83.70%	85.59%	85.45%
<i>Rec (G)</i>	62	59	59
<i>Recognition rate (G)</i>	87.32%	83.10%	83.10%

Table 3. Classification results for the *ENV* dataset

where “(V)” is the validation set results, “(G)” is the generalisation set results and “Rec” is the number of correctly ‘recognised’ samples

9. Statistical analyses using PCA and FA

Arising from the poor *GENES* validation rates, a detailed analysis of the inter-relationships between *HEART* candidate gene markers was desired. Two statistical algorithms, namely: Principal Components Analysis (PCA) [10] [18] and Factor Analysis (FA) [12] [13] have hence been brought to light in the hope of using them to identify the *latent constructs* of the dataset. A latent construct is some linear combination of the elemental attributes of a dataset that can be an entirely different type of object or attribute altogether. Latent constructs may manifest as one or more of the *HEART* candidate genes and knowledge of these constructs has important consequences clinically as they can be determined early in an individual’s life so that interceptive procedures can be employed, e.g. aggressively modifying the diet, to prevent the development of the disease.

In this aspect, PCA / FA was chosen to perform *dimensionality reduction* [10] so that a reduced subset of the attributes could be used as the NN inputs. The data reduction module of SPSS has been invoked for this purpose. The principal components method of factor extraction, which is essentially a principal components analysis procedure, was chosen as a default option for FA. Hence, both PCA and FA are performed in a single simulation on SPSS. The simulation is split into two different sets. The first simulation proceeds with a FA without the CAD status of *HEART* and the second proceeds with CAD included.

Variable	Communality	* Factor	Eigenvalue	Pct of Var	Cum Pct
Gene 1	.36567	* 1	3.32372	17.5	17.5
Gene 2	.45275	* 2	2.45231	12.9	30.4
Gene 3	.67619	* 3	2.19873	11.6	42.0
Gene 4	.51481	* 4	1.49997	7.9	49.9
Gene 5	.65822	* 5	1.41973	7.5	57.3
Gene 6	.61907	* 6	1.11695	5.9	63.2
Gene 7	.50459	* 7	1.05530	5.6	68.8
Gene 8	.23395	*			
Gene 9	.46675	*			
Gene 10	.72889	*			
Gene 11	.74516	*			
Gene 12	.97535	*			
Gene 13	.97369	*			
Gene 14	.99090	*			
Gene 15	.98621	*			
Gene 16	.97478	*			
Gene 17	.86242	*			
Gene 18	.45558	*			
Gene 19	.88172	*			

Figure 2. Principal Components Analysis Results excluding CAD

Results of the first simulation showed that 7 factors, latent constructs, or simply principal components, accounted for about 69% of the variance in the data distribution. Each of the 7 principal components is a unique combination of the 19 genetic markers of *HEART*. These 7 factors are next subjected to FA. The resulting rotated factor matrix showing the correlations of each gene to one another are shown in Figure 3. (The shaded numbers are the relative

contributions of each gene towards the 7 newly formed factors that are deemed statistically significant, i.e. >0.4.)

Rotated Factor Matrix							
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Gene1	-0.00546	0.12312	0.01982	0.20025	0.05378	0.55311	-0.03421
Gene2	0.02824	0.04486	0.10133	0.06613	-0.09516	-0.50144	-0.4181
Gene3	0.01372	0.05379	0.01588	-0.03189	-0.05596	-0.06252	0.81535
Gene4	-0.03947	0.70616	-0.07996	-0.00962	0.06003	0.06648	-0.00888
Gene5	-0.005	0.7939	-0.07056	0.01111	-0.00509	0.14662	0.03586
Gene6	-0.00277	0.76292	0.09401	-0.03653	-0.06708	-0.14525	0.03528
Gene7	0.01521	0.69395	0.0893	0.0063	-0.07475	-0.08665	-0.04097
Gene8	0.06907	-0.12886	0.10805	-0.06185	-0.02644	0.44314	-0.001
Gene9	0.01517	-0.08215	0.16608	0.09189	0.03901	-0.51939	0.39046
Gene10	0.04481	-0.0954	-0.07109	0.07794	0.83447	0.08891	0.04894
Gene11	0.01562	0.02088	-0.00513	-0.03877	0.8584	-0.04231	-0.06567
Gene12	0.15736	0.02468	0.07079	-0.97159	-0.01829	-0.01454	0.02103
Gene13	-0.1738	-0.01346	-0.07956	0.96755	0.02138	0.01117	-0.01557
Gene14	0.98683	-0.01297	-0.04442	-0.11625	0.02584	0.02649	0.00675
Gene15	0.98434	-0.01644	-0.05036	-0.11412	0.02946	0.02294	0.00835
Gene16	-0.97825	0.0086	0.0437	0.12123	-0.02004	-0.02704	0.00017
Gene17	-0.08786	0.04589	0.91168	-0.07251	-0.04249	-0.09716	-0.07016
Gene18	0.02654	-0.04266	0.58826	0.0177	0.03962	0.24825	0.20858
Gene19	-0.07241	0.03117	0.92023	-0.09785	-0.08962	-0.07233	-0.07645

Figure 3. Rotated Factor Matrix depicting the Factor Analysis Results excluding CAD

Interestingly, the clustered genes (shaded numbers) agree exactly with their corresponding biological functions. Factor 1 is contributed mainly by genes 14, 15 and 16, which are part of a family of genes responsible forming blood clot whereby Factor 2 is contributed mainly by genes 4, 5, 6 and 7, which are a family of proteins closely associated with cholesterol. Factor 3 is contributed mainly by genes that are precursors to blood formation while Factor 4 is contributed mainly by the 2 more polymorphic sites on a gene subunit of the blood clotting gene family. Factor 5 is mainly contributed by 2 blood anti-coagulation gene markers that are closely related while Factor 7 is contributed mainly by another protein that is associated with cholesterol.

Lastly, Factor 6 is contributed mainly by genes 1,2 and 8, 9, the former two being responsible for a regulation of blood pressure and the latter two being related to cholesterol and to breaking down lipids. It has not been part of our aim to specifically cluster the genes in this way but the results has helped to reveal the statistical relevance of each of the 19 genetic attributes.

In the second simulation, the correlation of CAD with the genetic markers was examined. Figure 4 shows the rotated factor matrix with CAD included.

Rotated Factor Matrix								
	Factor1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
CAD	0.06784	-0.02426	0.0392	0.06473	0.03043	0.67658	-0.04213	0.02841
Gene1	-0.00851	0.11698	0.03716	0.20672	0.04908	0.33215	0.39756	-0.01056
Gene2	0.02498	0.0149	0.11728	0.05047	-0.08106	0.02789	-0.71041	-0.21052
Gene3	0.0161	0.02935	0.0269	-0.03364	-0.04307	-0.10111	-0.02277	0.89226
Gene4	-0.04288	0.6823	-0.05661	-0.01281	0.06359	0.2367	-0.09447	0.10184
Gene5	-0.00719	0.78435	-0.05829	0.0124	-0.00465	0.18056	0.06351	0.07602
Gene6	-0.00307	0.77886	0.07211	-0.03715	-0.06461	-0.17822	-0.04397	-0.01673
Gene7	0.01563	0.71403	0.06802	0.00838	-0.07534	-0.14087	0.0193	-0.117
Gene8	0.07133	-0.08509	0.08045	-0.04712	-0.04222	-0.0623	0.63172	-0.22514
Gene9	0.02334	-0.03554	0.11546	0.09558	0.02898	-0.64865	-0.01787	0.10675
Gene10	0.04391	-0.09884	-0.07135	0.07779	0.83654	0.01995	0.07273	0.05902
Gene11	0.01639	0.02146	-0.00505	-0.03791	0.85633	-0.00106	-0.01102	-0.08971
Gene12	0.15811	0.02468	0.07185	-0.97127	-0.01849	-0.00769	0.01114	0.01855
Gene13	-0.17456	-0.01277	-0.08162	0.96721	0.02166	-0.00096	-0.0091	-0.01587
Gene14	0.98665	-0.01193	-0.04428	-0.11563	0.02566	0.02197	0.0196	0.00675
Gene15	0.98413	-0.01512	-0.05075	-0.1136	0.02935	0.01656	0.01839	0.00745
Gene16	-0.97787	0.00797	0.04352	0.12091	-0.0202	-0.02398	-0.01432	-0.00411
Gene17	-0.08668	0.0515	0.90571	-0.07315	-0.04202	-0.1135	-0.0741	-0.08025
Gene18	0.02594	-0.06205	0.60941	0.01905	0.04307	0.16327	0.12345	0.29131
Gene19	-0.07138	0.03841	0.9133	-0.09815	-0.08982	-0.11228	-0.04716	-0.09289

Figure 4. Rotated Factor Matrix depicting the Factor Analysis Results including CAD

As evident from Figure 4, with CAD included, a total of 8 latent constructs (or factors) were discovered. Factor 6 was found to be highly correlated to gene 9, a gene (or enzyme) responsible for breaking down lipids. And the score for gene 9 (-0.65) showed that there was a negative correlation.

The gene has since been used to cross-validate with a MLP network but results have not been better.

10. Discussion

A critical discussion and review of the experiments are presented in this section.

10.1. MLP and HME simulation results

10.1.1. Results on FULL

This aspect utilises the maximum number of attributes possible. Respectable performances have been achieved. Average validation and generalisation rates of the three methods are 83.7% and 80.75% respectively. A point to note is that this has been achieved with the simplest of MLP and HME architectures employed. (Other means to improve the generalisation rate e.g. pruning and weight regularisation has yet to be employed.)

10.1.2. Results on GENES

The results of the *GENES* experiment illustrate three important points. Firstly, the *GENES* dataset is not exhaustive enough for the network to generalise well. Secondly, the environmental factors in this context seem to play the primary role for the network to generalise. Thirdly, the ensemble of classifiers approach serves to improve generalisation rate more for weak classifiers than for strong classifiers. Average validation and generalisation rates for this dataset are 59.54% and 73.79% respectively.

The *GENES* training dataset (923 data points) consists of 529 CAD positives (1's), which takes up about 57.3% of the dataset. A validation recognition rate of 59.54% offers only a slight advantage (3.9% improvement) over randomly guessing all as CAD positive. This could be due to the lack of sufficient training data for the network to “learn” in the supervised training phase and hence the network simply “overfits” the biased dataset’s representation of the gene pool of the population.

The poor validation set performance of *GENES* as compared to the benchmark *FULL* dataset, hints at the possibility that the environmental factors are the primary factors correlating to CAD. The addition of the candidate gene markers actually brings down the entire NN performance. This will be analysed in the next dataset.

The estimated generalisation rate obtained by cross validation, (which is actually equivalent to the average validation rate of 59.54%) is markedly underestimated, when the actual average generalisation rate of 50 classifiers (73.79%) scored a 23.9% improvement over the former. This is a contrast to the notion that generalisation performance is usually not as high as validation set performances due to the “unseen” examples.

Dietterich [11] offered an explanation of this phenomenon, citing: “*An ensemble can be more accurate than its component classifiers only if the individual classifiers disagree with one another*”. An intuitive example is to consider that there is an ensemble of 3 classifiers: $\{h_1, h_2 \text{ and } h_3\}$ and a new case x . when the 3 classifiers are identical (i.e. all being strong classifiers), then when $h_1(x)$ is wrong, $h_2(x)$ and $h_3(x)$ are also wrong. However, if the errors made by the classifiers are uncorrelated, then when $h_1(x)$ is wrong, $h_2(x)$ and $h_3(x)$ may not necessarily be wrong. Hence, a majority vote would correctly classify x . This explains why

the ensemble of 50 weak classifiers for gene-only can achieve a dramatic improvement in generalisation performance.

10.1.3. Results on ENV

The effect of environment factors on CAD has been studied. Average validation and generalisation rates are 84.91% and 84.50% respectively marking 1.5% and 4.6% higher validation and generalisation rates than the *FULL* aspect. This goes to suggest that the presence of genes would negatively impact the recognition rates. However, this does not dismiss the utility of genes in this context. As mentioned earlier, CAD etiology is polygenic and multifactorial and is an area where all research strategies hitherto have been largely inadequate. Trying to use only environmental factors to explain the data would be a huge fallacy, as complex gene-environment interactions are known to have effects on a person's risk of CAD. This particular high recognition rates for environment-only dataset only serves to explain that in complex disease studies, the contribution of each genetic marker is often subtle and could easily be masked by non-genetic factors. We are currently addressing this issue by applying a different strategy for handling the genetic information in our prediction algorithm.

10.2. PCA and FA results

The results of PCA/FA in this project have been largely illuminating. PCA/FA revealed the inter-relationships between the various genetic attributes. Such "relationships" can be attributed to their *linkage disequilibrium*, which is the association of two linked alleles more frequently than would be expected by chance [19]. The pattern of linkage relationships among polymorphisms in a gene is an important factor in the choice of DNA markers for association studies of complex diseases. Generally, loci located within a few kilobases of one another are expected to show a high degree of linkage disequilibrium.

The clustering of gene 9 with CAD may suggest a relatively greater contribution of this gene to CAD risk prediction compared to the other lipids. However, further verification is required to establish the role of gene 9 in CAD risk prediction and its interactions with other genes and environmental factors

It is needful to highlight the pitfall of using a case-control study design such as this for disease prediction. CAD is a chronic disease that has a late onset. As such, subjects who were recruited as healthy controls at the time of study may, in time, become a CAD case. This has led to inaccuracies in our classification of study subjects' CAD status.

11. Conclusion

This paper has examined and compared the performances of three neural network methods of in the predicting coronary artery disease. Attributes were a combination of genetic and phenotypic factors. Results of the simulation runs indicated that the techniques used are promising, as superior recognition rates have been achieved using the most fundamental network architectures. The committee learning neural network approach on the *GENES* dataset has been largely successful, scoring a 23.9% improvement over a single classifier approach. Furthermore, the *HEART* dataset was further analysed using two statistical tools, namely principal components analysis (PCA) and factor analysis (FA). As a result of which, these techniques were able to cluster together genes of the same biological 'family' and hence reduced the number of input attributes for disease risk prediction. A lipid metabolism gene was also identified as an important gene for further studies of its predictive role in CAD risk assessment.

References

- [1] Heng, C.K., "Candidate genes for Coronary Artery Disease", PhD Thesis, *National University of Singapore, Department of Paediatrics*, 1996.
- [2] Brown, P.O., Botstein, D., "Exploring the new world of the genome with DNA microarrays", *Nature Genetics Supplement*, Volume 21,33-37,1999.
- [3] Hacia, G., "Resequencing and mutational analysis using oligonucleotide microarrays", *Nature Genetics Supplement*, Volume 21,42-47,1999.
- [4] Syvänen AC., "From Gels to Chips:"Minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms", *Hum Mutat* 13:1-10, 1999.
- [5] Bassett et al, Eisen, M.B., Boguski et al, "Gene expression informatics-it's all in your mine", *Nature Genetics Supplement*, 21:51-55,1999.
- [6] Strohman, R., "Ancient genomes, wise bodies, unhealthy people: limits of a genetic paradigm in biology and medicine", *Perspectives in Biology and Medicine*, 37(1):112-145, 1993.
- [7] Congdon, C.B., "A comparison of genetic algorithms and other machine learning systems on a complex classification task from common disease research", PhD Thesis, *University of Michigan, Computer Science and Engineering Department*, 1995.
- [8] Hosmer, D.W. and Lemeshow, S. Jr., "Applied Logistic Regression", *John Wiley and Sons*, 1989
- [9] Tu, J., "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", *J Clin Epidemiol*, 49(11):1225-1231, 1996.
- [10] Haykin, S., "Neural Networks – A Comprehensive Foundation", *Prentice Hall*, 351-357,213-218, 193-194,169-171,171-173,215-217,161-175,373,392-442,401,1990.
- [11] Dietterich, et al, "Machine-Learning Research Four Current Directions", *American Association for Artificial Intelligence, AI Magazine*, 1997.
- [12] Rasmussen et al, "Evaluation of Gaussian processes and other methods for non-linear regression", PhD thesis, *University of Toronto*, 1996.
- [13] Waterhouse, S.R., "Classification and Regression using mixtures of experts", PhD Thesis, *University of Cambridge*, 1996.
- [14] Chin, W.C., "Risk estimation of heart disease using DNA microarrays and neural networks ", FYP Thesis, *National University of Singapore*, 27-29, 2000.
- [15] Bishop, C.M., "Neural networks for Pattern Recognition", *Clarendon Press, Oxford*, 89-97,274-282, 1995.
- [16] Breiman, L., Friedman, J. H., Olshen, R.A., & Stone, C.J., "Classification and Regression Trees", *Belmont, CA: Wadsworth International Group*, 1984.
- [17] Jordan, M.I., & Jacobs, R.A., "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, 6(2), 181-214, 1994.
- [18] Sharma, S., "Applied Multivariate Techniques", *Wiley*, 58-89,90-143, 1996.
- [19] Emery, A.E.H., Mueller, R.F., "Elements of medical genetics", *Churchill living stone*, 367, 369, 1988.