Unravelling Gene Interactions to Find the Cause of Artherosclerosis, a Multigenic Disease, Using an Artificial Neural Network

W Dassen, W Spiering, P de Leeuw, P Smits, WA Dijk, H Spruijt, E Gommer[1], C Bonnemayer, PA Doevendans[1]

Maastricht University, Dept of Cardiology, Maastricht, The Netherlands, [1] Interuniversitairy Cardiology Institute of The Netherlands, Utrecht, The Netherlands

Abstract

To understand the etiology of multigenic disease like atherosclerosis a polymerase chain reaction based gene array containing 65 single nucleotide polymorphisms (SNP) was analyzed. To asses the possibilities of pattern recognition techniques in detecting unfavorable genetic combinations two approaches were analyzed. A selection of these 65 single nucleotide polymorphisms formed the input to both binary logistic regression models and to self-learning artificial neural networks. Repeated analysis showed that both methods performed equal. Further research to improve the differentiating power of both methods should focus first on decreasing the number of otherwise indeterminable polymorphisms.

1. Introduction

The advance in molecular genetics helps us to move from monogenetic disease to the unraveling of more complex diseases including multigenic disease states as atherosclerosis.(1;2) Whereas in monogenetic disease very often one base change in the coding region of a gene is sufficient to cause a disease, in multigenic disease the effects of one base change are modest, and therefore a composition of several unfavorable changes could explain the disease etiology. Subtle genetic variation involving one or several bases occur approximately every 1000 base pairs These changes can be positioned throughout the gene or in between genes. Therefore only a minority will effect the coding sequence (exons). Even if the coding sequence is involved the amino-acid order may be unchanged due to redundancy in the genetic code. Some changes in introns or the regulatory part of the gene could have an effect on transcription of the gene and lead to an increase or reduction of protein product.(3) Little is known on the influences of base changes in the extragenic regions. If a specific DNA change is found in the population in more then 1% of the individuals, the variation is indicated by the term polymorphism. When only one base is involved these differences between individuals are indicated by the term single nucleotide polymorphisms (SNP).

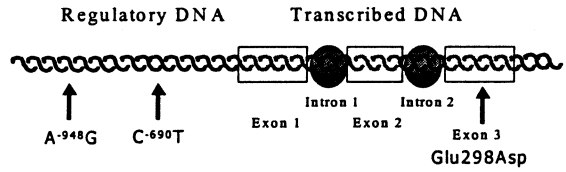


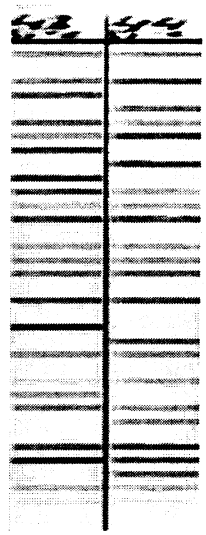
Figure 1. Structure of the endothelial nitric oxide gene. Arrows indicate the position of three different polymorphisms

Depending on the localization of intragenic SNPs they are indicated by a minus sign when the position is preceding the transcription initiation site (regulatory DNA). For instance the A-948G polymorphism in the eNOS gene (figure 1) indicates an adenine to guanine polymorphism at position 948 bases before the first exon.(4). If a change occurs in an exon it could result in an amino acid change (eg Glu298Asp). Here, the polymorphism leads to replacement of amino acid 298 glutamine by arginine.(5-7) In order to upscale SNP analysis and to unravel multigenic disease different approaches can be followed. SNPs can be analyzed in separate biochemical reactions, using the polymerase chain reaction (PCR). For the detection of one SNP two primers have to be used. By mixing primers several SNPs can be detected in one reaction. In the array used here, 4 membranes have been preloaded with DNA templates to which PCR product can hybridize. The hybridization product can be recognized by antibody based chemical reaction.

0276-6547/01 $17.00 © 2001 IEEE

Computers in Cardiology 2001;28:373-376.

Figure 2. In the left panel de genes spotted on the filter are indicated, with in the middle the polymorphism tested. On the picture (right panel) a filter is shown where for each variation one or two bands can be detected, depending on the genotype.



gene	variant
cystathionine β-synthase	ile278thr
methylenetetrahydrofolate reductase	C677T
endothelial nitric oxide synthase	A(-948)G
	C(-690)T
	glu298asp
angiotensin converting enzyme	Ins/del
angiotensin II receptor type 1	A1166C
angiotensinogen	met235thr
atrial natriuretic peptide	G664A
	T2238C
α-adducin	gly460trp
epithelial Na channel α subunit	trp493arg
	ala663thr
G protein β3 subunit	C825T
β2 adrenergic receptor	arg16gly
	gln27glu
stromelysin	5A/6A

As man has a double set of chromosomes (paternal and maternal) there are two variants of each gene (allele). For each position two variants are possible leading to three possibilities for each SNP. For instance an individual can be homozygous for the A at position –948, or for the G at this position. Alternatively, heterozygousity can exist at this particular locus (A and G allele, figure 2). (8) A more sophisticated approach is based on microarrays which can be used for high density SNP mapping. Currently more then 2 million SNPs have been identified (9)

To evaluate endothelial function and atherosclerosis we tested a polymerase chain reaction based gene array containing 65 single nucleotide polymorphisms (SNP) in disease related genes as documented in previous studies. Representing a classical pattern recognition problem, the analysis of combinations of unfavorable genes was evaluated using both classical statistical techniques and self-learning neural networks.

2. Methods

The array was performed on DNA samples of 89 patients with diabetes mellitus and impaired endothelial function and 47 healthy controls. Endothelial function was assessed by measuring changes in forearm blood flow after pharmacological interventions. Every SNP has three possible outcomes resulting in an almost infinite number of unique combinations. To evaluate whether self-learning techniques could be applied in this type of pattern recognition, an artificial neural network was conceived, trained and tested with an independent test set.

The experiment comprised a number of subsequent steps. Using univariate analysis all 65 variables were evaluated to asses if they could contribute in differentiating healthy and diabetes mellitus cases. Only those 10 variables with a $p<0.1$ were accepted in the study. With these variables it was tested whether a binary logistic regression model could be constructed that would classify all 136 cases correctly.

Subsequently, the 136 cases were randomized 10 times. After every randomization 86 cases were selected as a training set and the remaining 50 cases as an independent test set, resulting in 10 training sets and 10 corresponding test sets. Each training set was used to generate a neural network, that could be evaluated using the complementary test set. Subsequently all neural networks were analyzed further separately. Their training sets were used to build 10 logistic regression models, and evaluated with the corresponding test sets. Finally the overall performance of these logistic models including all cases from both training and test set was quantified.

2.1. The neural network

The input of this network was formed by those 10 out of these 65 variables that had reached a significance level of less than 10% using univariant analysis after exclusion of missing values, representing 38 different options. These 38 different categories were formed by 10 variables each representing three possibilities plus 8 out of these 10 variables showing at least one missing value. From the combined 136 patients 86 were randomly selected as training and 50 as test set. To generalize the results this randomization was repeated 10 times resulting in 10 different networks. The artificial neural network was composed of 38 input neurons, 38 hidden and one output neuron. The neural networks were constructed using Brainmaker Professional. All standard settings were selected to make these 10 resulting networks comparable.

2.2. The binary logistic regression model

All logistic regression models in this study were calculated using the SPSS 9 software package. All variables were entered forward stepwise conditional in the model. For each variable in the equation, the coefficient (B), its standard error, the estimated odds ratio [exp(B)], the confidence interval for exp(B) and the log-likelihood if removed from the model, was calculated. Finally for each of the 136 cases both the observed and the predicted group was stored together with the predicted probability.

The distribution of the 10 selected SNP in the sick and the healthy group is given in table 1. The two columns indicate the number of cases in the sick and the healthy group. The digits at the right site indicate in which of the the logistic regression models 1 – 10 SNP was entered.

SNP included	sick	healthy	included in :
Apo(a)C93T			**0**
CC	53	10	
CT	17	0	
TT	0	3	
missing	19	34	
ApoCIII T3206G			**1 to 10**
GG	4	8	
TG	43	12	
TT	42	17	
missing	0	10	
PON1 Gln192Arg			**0**
AA	5	4	
GA	27	28	
GG	51	15	
missing	6	0	
LDLR Ncol+-			**1,3,6,7,8**
++	37	13	
+-	29	24	
--	16	2	
missing	7	8	
TNFβ Thr26Asn			**0**
AA	10	3	
TA	53	22	
TT	21	20	
Missing	5	2	
ACE ins/del			**0**
DD	19	8	
ID	37	26	
II	22	5	
missing	11	8	
ADRB2 Gln27Glu			**2,3**
GlnGln	22	19	
GlnGlu	52	24	
GluGlu	15	3	
missing	0	1	
TNFα G(-376)A			**4,5**
AA	3	0	
GA	21	3	
GG	63	43	
Missing 2	1		
TNFα G(-308)A			**3**
AA	6	1	
GA	39	14	
GG	44	32	
TNFβ Thr26Asn			**4,5,10**
AA	7	1	
TA	59	25	
TT	23	21	

Table 1. The distribution of those single nucleotide polymorphisms analyzed in the study.

3. Results

The performance of the 10 created training and test sets are given in table 2. Training of the network took a median of 55 runs (range 35-82). The evaluation of the trained network resulted in a mean correct classification of 73.8 ± 4.2% (range 68-82). The logistic regression models resulted in a mean correct classification of 71.4 ± 5.1% . This difference is not statistically different.

Rand. #	Runs NN	Test NN %	Train LR %	Test LR %	SNP included
1	35	82	74.4	75.5	A,B
2	82	72	75.6	77.6	A,C
3	64	74	87.2	69.4	A,B,C,E
4	40	78	82.6	69.4	A,D,F
5	55	76	80.2	76.0	A,D,F
6	55	70	80.2	68.2	A,B
7	59	68	79.1	65.3	A,B
8	58	70	80.2	64.0	A,B
9	55	74	74.4	78.0	A
10	35	74	79.1	70.2	A,F

Table 2. The results of the 10 neural networks and of the corresponding binary regression models. The columns represent: number of randomization, number of training runs required for 100% training, % correct predicted by NN test set, % correctly trained by Logistic regression model, % correctly predicted test cases by logistic regression model. The right column indicated which SNP were entered in the logistic regression model: A = ApoCIII T3206G, B = LDLR Ncol+-, C = ADRB2 Gln27Glu, D = TNFα G(-376)A, E = TNFα G(-308)A F = TNFβ Thr26Asn

3.1. Overall results

To evaluate the behavior of single cases, the classification 1 (healthy) or 2 (diseased) was added up in these 10 logistic regression models independent of the fact whether they were used as a training or as a test case. Table 3. illustrates the distribution of the classification for both the diseased and the healthy group. A sum of 10 represents 10 times healthy, 11 represents one wrong classification , etc. A total of 20 indicates 10 times the classification disease. In 16 healthy and in none of the disease cases the SNPs were misinterpreted consequently.

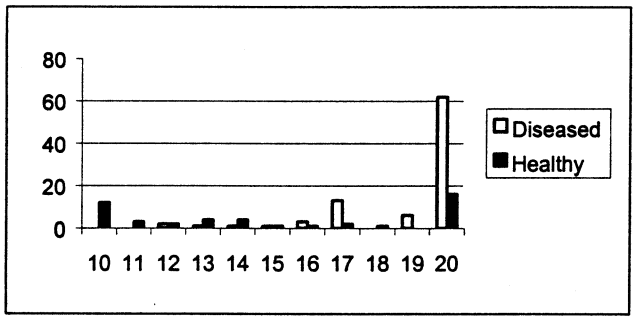


Figure 3. The distribution of all classifications.

4. Discussion

These results demonstrate that pattern recognition techniques like self-learning artificial neural network might have a place in determining whether certain combinations of polymorphisms are hidden in large-scale arrays. However, a significant number of polymorphisms could not be interpreted.

Additional improvement could be obtained by refining the quality of the array and by quantifying the certainty of the outcome of each polymorphism determined using such an array. In some instances the staining results in bands that do not reach the threshold for band recognition. Interestingly, when we study a group of diabetes patients the most predominant gene combinations found are in the area of the lipoproteins. This is not surprising as diabetes is a metabolic disease involving both glucose handling, but also other metabolic pathways including lipids.

References

[1] Doevendans PA, van Empel V, Spiering W, van der Zee R. Clinical Perspectives of Molecular Cardiology. Ned Tijds. Klin Chem 26,48-51. 2001.

[2] Collins FS. Shattuck lecture--medical and societal consequences of the Human Genome Project. N Engl J Med 1999; 341:28-37.

[3] Doevendans PA, van Gilst WH, Laarse A, van Bilsen M. Molecular cardiology part I: Gene transcription in the cardiovascular system. Cardiologie 1997; 4:221-229.

[4] Hingorani AD. Polymorphisms in endothelial nitric oxide synthase and atherogenesis: John French Lecture 2000. Atherosclerosis 2001; 154:521-527.

[5] Jachymova M, Horky K, Bultas J, Kozich V, Jindra A, Peleska J, Martasek P. Association of the Glu298Asp polymorphism in the endothelial nitric oxide synthase gene with essential hypertension resistant to conventional therapy. Biochem Biophys Res Commun 2001; 284:426-430.

[6] Guzik TJ, Black E, West NE, McDonald D, Ratnatunga C, Pillai R, Channon KM. Relationship between the G894T polymorphism (Glu298Asp variant) in endothelial nitric oxide synthase and nitric oxide-mediated endothelial function in human atherosclerosis. Am J Med Genet 2001; 100:130-137.

[7] Hibi K, Ishigami T, Tamura K, Mizushima S, Nyui N, Fujita T, Ochiai H, Kosuge M, Watanabe Y, Yoshii Y, Kihara M, Kimura K, Ishii M, Umemura S. Endothelial nitric oxide synthase gene polymorphism and acute myocardial infarction. Hypertension 1998; 32:521-526.

[8] Cardiovascular Genetics: for clinicians. Doevendans PA, Wilde AA, editors 2001. Dordrecht, Kluwer.

[9] Doevendans PA, Mummery C. Pluripotent Stem Cells: Biology and Applications. Neth Heart J 2001; 9:103-107.

Address for correspondence:

Willem RM Dassen,PhD, FESC
Dept of Cardiology,
Maastricht University
POBox 616 , 6200MD Maastricht ,
The Netherlands
w.dassen@cardio.azm.nl