MSC 62C10, 92C50 **DOI: 10.14529/jcem150402**

GENOME-WIDE ANALYSIS OF GENETIC ASSOCIATIONS FOR PREDICTION OF POLYGENIC HYPERCHOLESTEROLEMIA WITH BAYESIAN NETWORKS

A.V. Sulimov[1,2], sulimovv@mail.ru,
A.N. Meshkov[3], meshkov@lipidclinic.ru,
I.A. Savkin[1,2], is@dimonta.com,
E.V. Katkova[1,2], katkova@dimonta.com,
D.C. Kutov[1,2], dk@dimonta.com,
Z.B. Hasanova[4], zukhra@yandex.ru,
N.V. Konovalova[4], miirka@mail.ru,
V.V. Kukharchuk[4], v_kukharch@mail.ru,
V.B. Sulimov[1,2], vs@dimonta.com.
[1]Research Computer Center of Lomonosov Moscow State University, Moscow, Russian Federation.
[2]Dimonta Ltd., Moscow, Russian Federation.
[3]National Research Center for Preventive Medicine of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation.
[4]Russian Cardiology Research and Production Complex of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation.

The genome-wide analysis of genetic associations with lipid metabolism indicators was carried out using the technology of Bayesian networks (BN). It was performed to diagnose polygenic hypercholesterolemia on the basis of genetic data of the Russian population of patients. The data of 1,200 patients was analyzed. 196725 SNPs as well as clinical data, lipid profile indicators – different types of cholesterol – were obtained for each of them. The genome-wide association analysis (GWAS) and the statistical method of Pearson's chi-squared test were used for the initial selection of the most significant parameters. Two of the patient states related to a lipid metabolism were studied. These states are the level of LDL-C (low density lipoprotein) and the level of HDL-C (high density lipoprotein). The Bayesian networks having the simplest topology – naive – were used to predict the level of lipoprotein. The construction of ROC-curves and the calculation of the area under these curves (AUC) were used to assess a quality (reliability) of the prediction. AUC value increased from 0,5 for the initial BN to 0,9 after selecting of significant parameters using the GWAS method or the Pearson one. A further increase in AUC to 0,99 and decrease in the number of prognostic parameters to 150 was performed using Bayesian network optimization with respect to the number of parameters-nodes. Here the optimized function was value of AUC. The ambiguity of obtaining prognostic parameters at various ways of initial reducing the number of network nodes using the methods of GWAS and Pirson is shown. Low values of AUC were obtained for an independent control group of patients, despite very good results on the quality of the predictions, which were obtained on the training set. Further application of the proposed methodology is possible after the substantial reduction of the number of SNPs on the base of the analysis of the respective molecular mechanisms.

Keywords: GWAS; LDL-C; HDL-C; SNP; bayesian networks.

Introduction

Cardiovascular diseases (CVD) of the atherosclerotic origin are the leading cause of death and disability of the population in many countries worldwide [1]. Dyslipidemia associated with elevated level of LDL-C (low density lipoprotein cholesterol) and reduced level of HDL-C (high density lipoprotein cholesterol) is the main factor in the development of diseases associated with atherosclerosis [1–3]. Increased level of LDL-C could be due to secondary causes – factors of nutrition, taking drugs, a number of diseases – as well as to genetic factors [4–6]. Hereditary hypercholesterolemia may be monogenic or polygenic. The first one is in the case of a familial hypercholesterolemia [7]. Nowadays there are more than 90 gene loci associated with lipid metabolism, which is identified by genome-wide association studies, (GWAS) [8]. At the same time, the currently available data confirm the existence of racial and population differences in the magnitude of the risk of single nucleotide polymorphisms (SNP), which are associated with the disease [5, 6]. GWAS researches on indicators of lipid metabolism have not been conducted in the Russian population. In addition, there are no algorithms for the prediction of the development of polygenic dyslipidemia taking into account the patient's genetic data. Therefore, the aim of this paper is to conduct a genome-wide analysis of genetic associations with lipid metabolism indicators using Bayesian network technology. That is, the construction of these networks, their training, optimization and evaluation of the quality of predictions by using the trained Bayesian network for diagnosis of polygenic hypercholesterolemia based on genetic data.

1. Materials and Methods

The investigation of SNP was performed within the framework of realization of the research project "The approbation and adoption of new algorithms for the prevention, diagnosis and treatment of atherosclerosis in practice of outpatient clinics of Moscow West Administrative District". The main aspects of design and research methods are published previously [7, 8]. 1,200 patients with different values of the cardiovascular risk, calculated on the scale of SCORE (Systematic Coronary Risk Evaluation), were included in the study [9]. The subgroup 1 – patients with low to moderate risk of cardiovascular events ($< 5\%$), the subgroup 2 – patients without clinic of coronary heart disease (CHD), but with high and very high risk of cardiovascular events ($\geq 5\%$), the subgroup 3 – patients with ischemic heart disease (angina, myocardial infarction, revascularization). The patients were divided into three equal subgroups according to their number. The analysis of risk factors includes: age, sex, lipid profile indicators – cholesterol (TC), triglycerides (TG), low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C).

1.1. Isolation of DNA and Identification of SNP

DNA was extracted from 300 μl of a frozen blood with EDTA using Qiagen DNA bloodminikit according to the instructions. The quality of the isolated DNA was checked by the method of electrophoresis on 0,8 % agarose gel. DNA concentration was determined using the instrument Nanodrop ND-1000. 200 ng of genomic DNA was used for genotyping on the microarrays Cardio-Metabochip (Illumina) by protocol of data of the microarrays Infinium HD Ultra [10]. Genomic DNA was subjected to genome-isothermal amplification

with the following crushing to smaller fragments. These fragments were hybridized to a microarray, and one fluorescent nucleotide, which determines the value of polymorphism, was completed. Then the signal was amplified using fluorescently labeled antibodies. Scanning of the microarrays was performed on the instrument BeadArrayReader (Illumina) using the program BeadScan. Analysis of the genotypes was carried out in the program GenomeStudioGenotypingModule (Illumina).

1.2. Bioinformatic Analysis of Data

There were examined 1200 DNA samples from different patients, each sample of 196725 SNP. Average quality of genotyping (callrate) was 93,7%. The program PLINKv1.07 was used for the control of quality of the samples and for the genome-wide analysis of the associations (GWAS) [11].

Tests on the quality of genotyping of each sample, on the Hardy-Weinberg equilibrium, on the quality of SNP genotyping in a group, on the frequency of the minor allele (MAF)$\geq$ 5% were carried out during the control of quality of the samples and SNP. Total, 101 188 SNPs and 1182 samples left after all control procedures. Average quality of genotyping (callrate) is 99,8%. Then allied samples were deleted. To this end IBDs (IdentityByDescent) between all possible pairs of samples were checked. The pairs of samples with PI_HAT>0,8 (Proportion IBD) were considered allied and one random sample of the pair was removed. As a result 1121 samples have remained after this test.

The method GWAS, which is one of the fastest growing trends in the modern medicine, was used to investigate the associations between genotypes and phenotypes [11, 12]. It is based on an analysis of the causal association of SNPs and phenotypic characteristics. Here single nucleotide polymorphisms play the role of variables values. For the first time, this algorithm was successfully applied in 2005 to identify the genetic factors, which send for the macular degeneration [13]. The additional hypotheses (the dominant model, recessive one and others) are often used during the work with GWAS. The using of additional models allows to affect on the frequency of cases. The value of p, which is a numerical characteristic of the causal association of SNPs and phenotypes, is calculated as a result of GWAS performance:

$$p = \int_0^{\chi^2} \frac{t^{(f-2)/2} e^{-t/2}}{2^{f/2}\Gamma(f/2)} dt \qquad (1)$$

Single nucleotide polymorphisms with $p < 0,0001$ were selected in this work.

1.3. Technology of Bayesian Networks

To predict the development of polygenic dyslipidemia we used the probabilistic model, which is popular in medicine and is based on Bayesian networks (BN) [14–18]. This model is a successful combination of the probability theory and the graph theory. It allows you to combine different types of data – both clinical and huge databases of SNPs. The BN apparatus is represented by a directed acyclic graph, such that each its node is associated with a single variable and a table of the conditional probabilities [19]. The variables corresponding to the BN nodes may be clinical, molecular-biological and/or genetic parameters of patients, their age, the methods, which are used in a treatment,

the outcomes of the disease and other information. BN nodes are connected by arrows. The direction of each arrow corresponds to the assumed cause relations. Each node of BN has a finite set of mutually exclusive states. It depends on assumed discrete values of the variable, which is represented by this node. A node is called a parent, if the arrows go out of it, and a child, if the arrows go in it. The probability of a state of the node depends on the current status of parent nodes. The tables of conditional probabilities contain the probabilities of observations of the node states, which are obtained for various states of the parent nodes. To set these tables one can use the expert opinions or the results of the BN training procedure. BN inference procedure allows to determine the probabilities of the nodes states depending on the experimental information about the values of other nodes. Consider the prediction of a patient condition in the case of BN using. One of the network nodes corresponds to the variable that determines the patient condition and it is called the target variable or the endpoint or the outcome, and other nodes are relevant to the factors affecting on the state of CT. The choice of BN topology significantly affects the values of the conditional probabilities. In this work, we used a naive topology of Bayesian networks (Fig. 1), that is, all nodes have one common parent (root node "A").

Corresponding to the target variable.

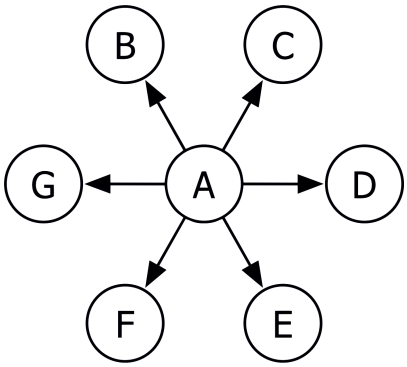

Fig. 1. Bayesian network having naive topology. The root node "A" corresponds to the projected condition of the patient (CT)

In the case of personalized medicine, target variable is used as the root node "A". All other nodes are called leaves and they correspond to various parameters of patients. Naive topology has several advantages. Firstly, a simple expression for the total probability is obtained in this model. Secondly, the smallest number of tables of conditional probabilities is achieved. Thirdly, variables do not influence on each other. It reduces the minimal number of data, which is needed for the BN training. It is particularly important for work with a small number of patients.

To assess the predictive ability of the BN and to compare different BNs by this characteristic, we used the method of ROC-curves [20, 21]. The value of the area under the ROC-curve (AUC) was used for a numerical measure of the quality of BN predictions: the closer AUC value is to 1, the better quality of the prediction of the trained network becomes. The most accurate method "except by one" was used to construct the ROC-curve. Each step of the construction of the ROC-curve is as follows. One patient is taken out of the database, training of the BN is performed on the remaining patients, and then the outcome for the eliminated patient is predicted by the trained BN, that is

using parameters of this patient we predict his outcome determining the corresponding conditional probability. This probability is added to the ROC-table together with the real evidence that the root variable has. This procedure is repeated for each patient and as a result the whole ROC-table containing the actual outcome (real evidence) and the conditional probability of this outcome for each patient is filled. After that the table was sorted by the conditional probability. Using it, we built the ROC-curve and calculated the value of AUC.

1.4. Data Preparation

The input data consist of genetic and clinical information of 1121 patients. The genetic information was contained in the output files of Genotyping module of the GenomeStudio program [22]. Clinical information was stored as Excel spreadsheets and contained the clinical data and the various risk factors of patients. We investigated two risk factors associated with the lipid metabolism: the level of LDL-C and HDL-C. Each risk factor in the existing database is corresponded to a continuous clinical parameter. To use Bayesian networks it is needed to convert continuous parameters to digital ones. The values of the first target, LDL-C, were determined in the following manner: 0 – well, if the level of LDL-C $\leq$ 4,9 $mmol/l$ and 1 – sick, if the level of LDL-C $>$ 4,9 $mmol/l$. The values of the second target, HDL-C, were determined in the following manner: 0 – well, if the level of HDL-C $>$ 1,2 mmol/l 1 – sick, if the level of HDL-C $\leq$ 1,2 $mmol/l$.

Each genetic GenomeStudio program file contained the information about the values of 196725 SNP parameters for one of the patients (see Fig. 2). Each SNP contained the difference by size in one nucleotide (A, C, G or T) in the DNA sequence in the homologous segments of homologous chromosomes (see the third and the fourth columns in Fig. 2). SNP set in all genetic files was the same, but sometimes there were omissions. SNP was used as a Bayesian network parameters (see the first column in Fig. 2). For further work it is needed to glue the genetic and clinical data into a single database. Since none of the existing tabular editors are not able to work with such a large number of parameters, it is needed to collect at once the database in CSV format. Thereto the GeNA program was developed for this purpose. The input of the program was a list of genetic files, a file with sampling of used genetic parameters, a file with the clinical data and the target variable. The output of GeNA was the db01 database and the corresponding BN with naive topology and the root node corresponding to the target variable.

Db01 database contained the values of these genetic parameters and values of all clinical parameters. To check the results of selection of network nodes, we divided the

SNP Name	Sample ID	Allele1 - Top	Allele2 - Top	Sample Name	Sample Group			Sample Index		SNP Index		SNP Aux	GC Score
chr1:109457160	5605703060_R03C01	C	C	147_4_44	27	1	0	0.8609	1	109457160	0.8316	1.0000	[T/G]	BOT TOP
chr1:109457233	5605703060_R03C01	C	C	147_4_44	27	2	0	0.7725	1	109457233	0.8139	1.0000	[T/G]	BOT BOT
chr1:109457614	5605703060_R03C01	-	-	147_4_44	27	3	0	0.0000	1	109457614	0.0000	0.0000	[T/C]	BOT TOP
chr1:109457618	5605703060_R03C01	A	A	147_4_44	27	4	0	0.5787	1	109457618	0.6683	0.5704	[T/C]	BOT TOP
chr1:109457943	5605703060_R03C01	A	A	147_4_44	27	5	0	0.8236	1	109457943	0.8053	1.0000	[T/C]	BOT BOT
chr1:109458224	5605703060_R03C01	G	G	147_4_44	27	6	0	0.8915	1	109458224	0.9012	0.9964	[A/G]	TOP BOT
chr1:109458469	5605703060_R03C01	A	A	147_4_44	27	7	0	0.9345	1	109458469	0.8981	1.0000	[A/G]	TOP BOT
chr1:109458772	5605703060_R03C01	C	C	147_4_44	27	8	0	0.8906	1	109458772	0.8554	1.0000	[T/G]	BOT TOP
chr1:109459424	5605703060_R03C01	A	A	147_4_44	27	9	0	0.9016	1	109459424	0.9105	0.9486	[A/G]	TOP TOP
chr1:109460430	5605703060_R03C01	G	G	147_4_44	27	10	0	0.7836	1	109460430	0.8210	1.0000	[T/C]	BOT BOT
chr1:109461426	5605703060_R03C01	A	A	147_4_44	27	11	0	0.9368	1	109461426	0.9008	1.0000	[A/G]	TOP BOT
chr1:109464943	5605703060_R03C01	G	G	147_4_44	27	12	0	0.8685	1	109464943	0.8373	1.0000	[A/G]	TOP TOP
chr1:109465107	5605703060_R03C01	T	T	147_4_44	27	13	0	0.9206	1	109465107	0.9298	1.0000	[T/A]	BOT BOT

Fig. 2. File with the genetic information of the patient

resulting database randomly into two parts in a ratio of 90% for the selection database db03 and 10% for the control database db04. The partitioning was made using the standard function of the spreadsheet editor *RAND()*. The binary target parameters correspond to the root nodes of BN. Leaf nodes corresponded to clinical parameters (sex and age), as well as to all the genetic parameters, containing from 2 to 3 values.

1.5. Prediction of BN on the Basis of SNP Data

The SNP database contains almost two hundred thousand genetic parameters. We used the ANN program in the study of the quality of prediction of the patients states on the base of various BNs using the SNP database and obtained the conditional probabilities of the target, which are close to zero for one outcome, and close to unity for another outcome. Analysis of the inference procedure showed that such tables of conditional probabilities are obtained by multiplication of the probabilities of rare values (SNP option value, which is extremely rare for the studied patients). An example of a table of conditional probabilities for the SNP parameter containing rare value is presented in Table 1.

The multiplication of a large number of rare values in the inference procedure led to a significant change of the final probability value. To solve this problem one can exclude the values of nodes with rare values (0,0001) from the inference procedure. For this purpose the program ANN-RARE was developed. This program replaced the probability of rare values (0,2; 0,0001) to the unit ones (1; 1) during the inference procedure. It was equivalent to disuse this evidence, that ceased to influence on the final probability. Application of the new program ANN-RARE allowed to obtain conditional probabilities for the patient outcome in the range from 0 to 1.

Assessment of predictions quality for the targets LDL-C and HDL-C on the naive BNs containing all the nodes of existing SNP parameters (almost two hundred thousand) gave the value of AUC equal to 0,5. This result shows the randomness of predictions. To improve predictions, it was decided to exclude from the BN all the nodes unrelated to a given target. To assess the causal association of nodes we used methods based on criterion χ^2. The results of the research obtained by the GWAS method (see section 1.2) were used, as well as the statistical method of Pearson's chi-squared criterion (briefly – Pearson method, see section 1.6).

Table 1

Table of conditional probabilities containing the rare value of "CC" for the outcome "1"

Node value	AA	AC	CC
target = 0	0,3	0,5	0,2
target = 1	0,4999	0,5	0,0001

1.6. Selection of SNP Parameters by Pearson Method

Pearson's chi-squared criterion is widely used to test the hypothesis that the distribution $X\{x_1, \ldots, x_N\}$ corresponds to some theoretical distribution law [23]. To use this criterion for an analysis of the database SNPs we implemented the Pirson program . This program considers the value of each parameter separately and the value of the target. Theoretical distribution was obtained on the assumption that the parameter does not depend on the given target, that is the condition of independence:

$$P(Value_i, target_j) = P(Value_i) * P(target_j), \tag{2}$$

where $Value_i$ – i-th value of the parameter, and $target_j$ – j-th value of the target, is hold. Let us consider the work of the program Pirson for the example of the parameter containing three values (AA, AC, and CC) and the target has two values (0 and 1) (Table 2).

Table 2

Example of partition of values in groups (N_{ji} – the number of patients with a given value of the parameter $Value_i$ and a given value of $target_j$)

Node value	$Value_1 =$ **AA**	$Value_2 =$ **AC**	$Value_3 =$ **CC**
$target_0 = 0$	N_{01}	N_{02}	N_{03}
$target_1 = 1$	N_{11}	N_{12}	N_{13}

Consider the calculation of the theoretical distribution for example of the first cell.

1. To count the number of all patients, that is, the sum over all indices $N = \sum\limits_{j,i} N_{ji}$.

2. To calculate the values $N_{Value_1} = \sum\limits_{j=1,2} N_{j1}$ and $N_0 = \sum\limits_{i=1,2,3} N_{0i}$, where N_{Value_1} – the number of patients with the value of the parameter $Value_1$, N_0 – the number of patients with $target = 0$.

3. To calculate $N_{01}^{theor} = \frac{N_{Value_1}}{N}\frac{N_0}{N}N$, where N_{01}^{theor} – the theoretical expected value of the number of patients whose $Value_i = Value_1$ and $target = 0$. It is calculate on the assumption of independence of variables: $P(Value_1, target_0) = P(Value_1) * P(target_0)$.

4. To calculate the deviation of theoretical one from the observed one $\chi_{01}^2 = \frac{(N_{01} - N_{01}^{theor})^2}{N_{01}^{theor}}$.

5. To calculate the sum by all cells $\chi^2 = \sum\limits_{j,i} \chi_{ji}^2$.

If $\chi^2 < \chi_{crit}^2(p, f)$ [24], then the variables are independent. Here the number of freedom degrees $f = (m-1)(n-1)$, where m – the number of states that the parameter can take, n – the number of states that the target can take ($m = 3$, $n = 2$ in case, which we consider).

The value of $1 - p$ is called the level of significance. This value is the probability of a deviation of the hypothesis about the independence of distributions, if this hypothesis is true. Since the value of χ^2 is proportional to N, then to compare the causal association of variables to each other one should normalize $\chi^2_{norm} = \frac{\chi^2}{N} N_{max}$, where N_{max} – max N for all variables. It allows to work with the missing data. This algorithm allowed to choose about 7000 relevant variables. It led to a significant increase of the quality of BN prediction.

1.7. BN Optimization by the Number of Parameters

To increase the predictive ability of BN we applied the novel method, which was already tested by us on other problems. It is the optimization of naive BN with respect to the number of nodes with the target function – value AUC [17, 18, 25]. The optimization algorithm was also changed slightly. Namely, we excluded from the inference procedure the nodes with rare values (0,0001). Naive network topology is particularly useful for using the optimization algorithm. It so, because such topology is invariant with respect to exclusion of nodes from BN except the root one – network topology is also naive, when you delete any number of leaf nodes from BN. Network optimization algorithm is based on the analysis of the possible networks, which are composed from various combinations of the network nodes. Each such combination necessary includes the root node of the BN under study. This root node corresponds to a given target. The algorithm is iterative: At every iteration we calculate AUC of the entire network, then calculate AUC of every network that can be obtained by excluding one variable (one node) from the entire network (except for the root variable). From the obtained networks we choose the one having the maximal value of AUC. If this value is larger than one of the entire network, we begin the new iteration of the algorithm with this newly obtained network. We stop the algorithm execution as soon there is no possibility to improve the AUC value by excluding a variable from the network. To speed up the work of the algorithm it is possible to remove several nodes from the original network at the end of the iteration. It means to remove nodes such that the removing of each of them separately led to an increase of the values of AUC of the original network. This optimization allows to select a set of variables that are critical for the predictive ability of the network and, accordingly, to eliminate variables, which affect weakly or negatively on it. As a result, the predictive ability of the BN significantly increases. [17, 18, 25].

2. Results

2.1. Genome-Wide Association Analysis

The genome-wide association analysis among 1121 patients was performed using the PLINK program. Also, using this program the Bonferroni procedure for multiple tests of significance was performed and values of p<0,05 were considered as significant. The result of research was identification of one SNP (rs7412) in the APOE gene significantly associated with the LDL-C level: p=6,654e-09, $p(Bonferroni)$=0,0006733. We have not identified other SNP, which are significantly associated (with correction for multiple tests of significance) to the levels of LDL-C and HDL-C.

2.2. Evaluation of Quality of BN Prediction

The selection of parameters by GWAS and Pearson statistical methods on full database db01 allowed to reduce the number of BN nodes to several thousand. Next step of the BN optimization with respect to the number of nodes was performed using the AUC value as a target function. The quality of the prediction of BN thus obtained is shown in Fig. 3 a (LDL-C) and Fig. 3b (HDL-C). These diagrams reflect a change of AUC value depending on BN containing different number of nodes. The initial selection of the parameters by the GWAS method on a database containing two hundred thousands of SNP parameters was performed by the program PLINK and allowed to improve the quality of BN prediction up to AUC = 0,8, and to reduce the number of nodes-leaves up to 6 thousand. The same selection of parameters by the Pearson method was performed by the program Pirson and allowed to improve the quality of BN prediction up to AUC = 0,97, and to reduce the number of nodes-leaves up to 3 thousand. Further optimization of the network with respect to the number of nodes (see section 1.7) increased the quality of predictions up to AUC = 0,99 and reduced the set of critical parameters up to 200.

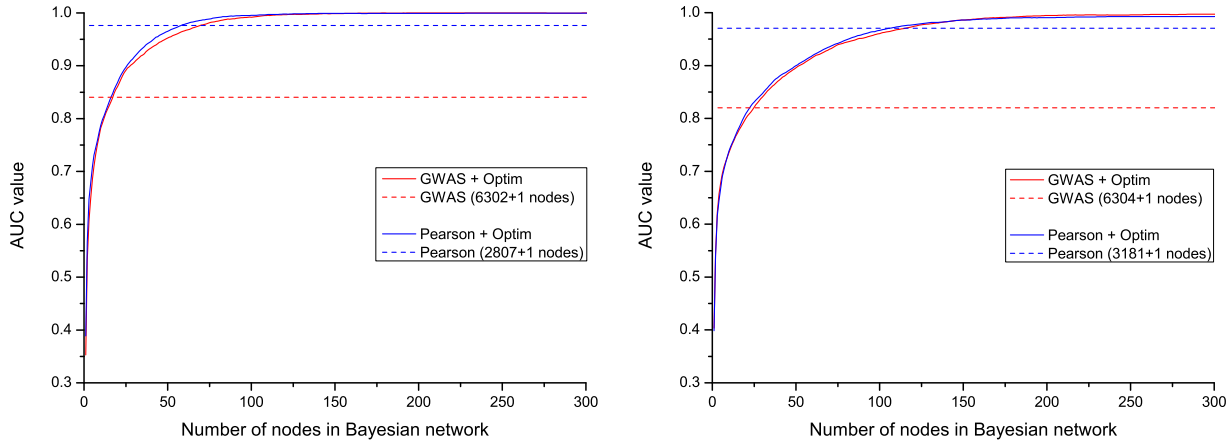

Fig. 3. Change of AUC for optimized BN for LDL-C: a) and HDL-C; b) in the database db01, where GWAS and Pearson indicate methods of initial selection of parameters and Optim points to further optimization of networks with respect to the number of nodes

2.3. Prediction of the Control Group

The control group of patients was selected during the step of preparation of data (the control database db04). Prediction for the control group was performed with BN obtained by selection of parameters using the statistical Pearson method. Two BN were investigated for each target. One BN was obtained using the Pirson program on the whole database db01 and other one was obtained also using the Pirson program but on the selection database db03. To predict the states of the target for the control group, BN were trained by ANN-RARE program on the same databases which were used to select the parameters. ROC-curves for respective BN are presented in Fig. 4a and Fig. 4b. Prediction for the db04 control group showed that in spite of the same quality of prediction, BN which were

obtained by selection of parameters by Pearson method on the whole database (db01) give much better results than BN, which were obtained by the selection on the selection database (db03). This result means that the selection of the parameters by the statistical methods, which are based on the criterion χ^2, strongly connects selected SNP parameters and the database which was used in the selection procedure. Also it means that the further use of the BN data to predict new patients is not possible.

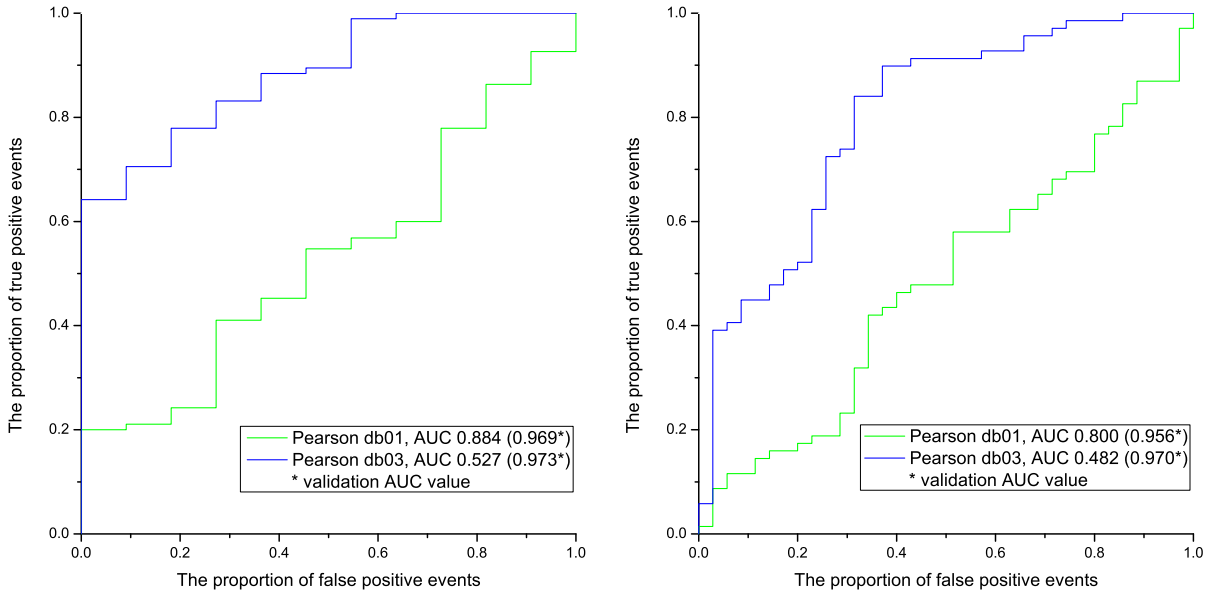

Fig. 4. ROC-curves of LDL-C: a) and HDL-C; b) prediction for the control group db04 obtained with bayesian networks trained on the database db03. Here "Pearson db01" and "Pearson db03" specify the bayesian networks created by selection of the parameters by Pearson method on the databases db01 and db03, respectively. Values of AUC correspond to the prediction quality for the control group db04, values of AUC in brackets correspond to the validation quality for the database db03

Conclusions

The research of database of 1121 Russian patients containing both clinical and genetic parameters – single nucleotide polymorphisms (SNP) – was performed using Bayesian network technology.

The genome-wide analysis of genetic associations with lipid metabolism indicators for diagnosis of polygenic hypercholesterolemia was performed. As a result one single nucleotide polymorphism (SNP rs7412) in the APOE gene which was significantly associated with the level of low density lipoprotein cholesterol ($p = 6,654e^{-9}$, $p(Bonferroni) = 0,0006733$) was discovered. No other SNP significantly associated (adjusted for multiple tests of significance) with the level of LDL-C and HDL-C were detected.

The selection of BN parameters using statistical GWAS and Pearson methods, as well as optimization of the BN with respect to the number of nodes, allows to improve the quality of the prediction of the level of lipoprotein cholesterol having low density and high

density up to AUC = 0,99 significantly, and to reduce the number of prognostic parameters up to 200. Despite the high quality of BN prediction on the database of selection, the obtained networks do not provide a good prediction for the control group, which is not part of the database of selection. Perhaps this is due to a very large number of genetic parameters in comparison with the number of patients and with comparatively high level of noise in the values of SNP parameters. Further application of the methodology, which is proposed in this article, is possible, if the number of SNP is substantially reduced using analysis of the molecular mechanisms, which connect genetic characteristics and lipid metabolism indicators.

The research is supported by the Russian Science Foundation grant (project № 14-50-00029).

References

1. Roger V.L., Go A.S., Loyd-Jones D.M., et al. Heart Disease and Stroke Statistics – 2012 Update: A Report from the American Heart Association. *Circulation*, 2012, vol. 125, issue 1, pp. e2–e220. DOI: 10.1161/CIR.0b013e31823ac046.
2. Kannel W.B., Dawber T.R., Kagan A., Revotskie N., Stokes J. 3rd. Factors of Risk in the Development of Coronary Heart Disease – Six Year Follow-up Experience. The Framingham Study. *Ann Intern Med*, 1961, vol. 55, issue 1, pp. 33–50. DOI: 10.7326/0003–4819–55–1–33.
3. Keys A. Coronary Heart Disease in Seven Countries: I. The Study Program and Objectives. *Circulation*, 1970, vol. 41, no. 4S, pp. I-1–I-8. DOI: 10.1161/01.CIR.41.4S1.I–1.
4. Gianfagna F., Cugino D., Santimone I., Iacoviello L. From Candidate Gene to Genome-Wide Association Studies in Cardiovascular Disease. *Thromb Res*, 2012, vol. 129, issue 3, pp. 320–324. DOI: 10.1016/j.thromres.2011.11.014.
5. Teo Y.Y., Sim X. Patterns of Linkage Disequilibrium in Different Populations: Implications and Opportunities for Lipid-Associated Loci Identified from Genome-Wide Association Studies. *Curr Opin Lipidol*, 2010, vol. 21, issue 2, pp. 104–115. DOI: 10.1097/MOL.0b013e3283369e5b.
6. Zhou W., Christiani D.C. East Meets West: Ethnic Differences in Epidemiology and Clinical Behaviors of Lung Cancer Between East Asians and Caucasians. *Chinese Journal of Cancer*, 2011, vol. 30, issue 5, pp. 287–292. DOI: 10.5732/cjc.011.10106.
7. Urazalina S.Zh., Titov V.N., Vlasik T.N., Balahonova T.V., Karpov Ju.A., Kuharchuk V.V., Bojcov S.A. [The Relationship of Concentration Lipoprotein-Associated Phospholipase A2 Secretory Markers of Subclinical Atherosclerotic Lesions the Arterial Wall in Patients with Low and Medium Risk on a Scale SCORE]. *Terapevticheskii arhiv*, 2011, vol. 83, no. 9, pp. 29–35. [Уразалина, С.Ж. Взаимосвязь концентрации ассоциированной с липопротеинами секреторной фосфолипазы А2 с маркерами субклинического атеросклеротического поражения артериальной стенки у пациентов с низким и средним риском по шкале SCORE / С.Ж. Уразалина, В.Н. Титов, Т.Н. Власик, Т.В. Балахонова, Ю.А. Карпов, В.В. Кухарчук, С.А. Бойцов // Терапевтический архив. – 2011. – Т. 83, № 9. – С. 29–35.]

8. Urazalina S.Zh., Rogoza A.N., Balahonova T.V. [The Value of the Markers Preclinical Lesions Wall of the Carotid Artery to Determine the Magnitude of Cardiovascular Risk on a Scale Recommendations EOAG / EOC (2003, 2007)]. *Cardiovascular Therapy and Prevention*, 2011, vol. 10, no. 7, pp. 74–80. [Уразалина, С.Ж. Значение маркеров доклинического поражения стенки сонной артерии для определения величины сердечно-сосудистого риска по шкале Рекомендаций ЕОАГ/ЕОК (2003, 2007) / С.Ж. Уразалина, А.Н. Рогоза, Т.В. Балахонова // Кардиоваскулярная терапия и профилактика. – 2011. – Т. 10, № 7. – С. 74–80.]

9. Jdanov D.A., Deev A.D., Jasilionis D., Shalnova S.A., Shkolnikova M.A., Shkolnikov V.M. Recalibration of the SCORE Risk Chart for the Russian Population. *European Journal of Epidemiology*, 2014, vol. 29, issue 9, pp. 621–628. DOI: 10.1007/s10654–014–9947–7.

10. Voight B.F., Kang H.M., Ding J., Palmer C.D., Sidore C., et al. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genet*, 2012, vol. 8, issue 8, pp. e1002793. DOI: 10.1371/journal.pgen.1002793.

11. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. PLINK: a Toolset for Whole-Genome Association and Population-Based Linkage Analysis. *American Journal of Human Genetics*, 2007, vol. 81, issue 3, pp. 559–575. Available at: http://pngu.mgh.harvard.edu/purcell/plink/ (accessed 9 October 2015).

12. Jian W., Sanjay S. A Powerful Hybrid Approach to Select Top Single-Nucleotide Polymorphisms for Genome-Wide Association Study. *BMC Genetics*, 2011, vol. 12, no. 3, pp. 3–9. DOI: 10.1186/1471–2156–12–3.

13. Klein R.J., Zeiss C., Chew E.Y., Tsai J.-Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, NY)*, 2005, vol. 308, issue 5720, pp. 385–389. DOI: 10.1126/science.1109557.

14. Lucas P.J., van der Gaag L.C., Abu-Hanna A. Bayesian Networks in Biomedicine and Health-Care. *Artif Intell Med*, 2004, vol. 30, issue 3, pp. 201–214. DOI: 10.1016/j.artmed.2003.11.001.

15. Gevaert O., Smet F.D., Timmerman D., Moreau Y., Moor B.D. Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks. *Bioinformatics*, 2006, vol. 22, issue 14, pp. e184–e190. DOI: 10.1093/bioinformatics/btl230

16. Sulimov A.V., Vtyurina D.N., Romanov A.N., Maslennikov E.D., Sulimov V.B., Kurochkin I.N. et al. Expert Systems of Personalized Medicine: the Use of Bayesian Networks to Predict the State of Patients. *Post-Genomic Research and Technology.* Moscow, Publishing House of Moscow University, 2011, pp. 641–702.

17. Gens G.P., Sulimov A.V., Moiseeva N.I., Ovsij O.G., Vel'sher L.Z., Rybalkina E.Ju., Selezneva I.I., Savkin I.A. [Search Approaches to Forecasting Outcomes Breast Cancer Using Bayesian Networks]. *Onkologiia. Zhurnal imeni P.A. Gerzena*, 2014, no. 9, pp. 37–46. [Генс, Г.П. Поиск подходов к прогнозированию исходов рака молочной

железы с помощью байесовских сетей / Г.П. Генс, А.В. Сулимов, Н.И. Моисеева, О.Г. Овсий, Л.З. Вельшер, Е.Ю. Рыбалкина, И.И. Селезнева, И.А. Савкин // Онкология. Журнал им. П.А. Герцена. – 2014. – № 9. – С. 37–46.]

18. Maslennikov E.D., Sulimov A.V., Savkin I.A., Evdokimova M.A., Zateyshchikov D.A., Nosikov V.V., Sulimov V.B. An Intuitive Risk Factors Search Algorithm: Usage of the Bayesian Network Technique in Personalized Medicine. *Journal of Applied Statistics*, 2014, vol. 42, issue 1, pp. 71–87. DOI: 10.1080/02664763.2014.934664.
19. Jensen F.V., Nielsen T.D. *Bayesian Networks and Decision Graphs*, New York, Springer Verlag, 2007.
20. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. *HP Labs Tech Report HPL-2003-4*, 2004.
21. Obuchowski N.A. ROC Analysis. *American Journal of Roentgenology*, 2005, vol. 184, issue 2, pp. 364–372. DOI: 10.2214/ajr.184.2.01840364.
22. Available at: http://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html (accessed 9 October 2015).
23. Van der Waerden B.L. *Mathematische statistik*. Berlin, Gottingen, Heidelberg, Springer-Verlag, 1960.
24. Jay L. Devore, Kenneth N. Berk Modern Mathematical Statistics with Applications. *Springer Texts in Statistics*, 2012, p. 200. DOI: 10.1007/978-1-4614-0391-3.
25. Gens G.P., Sulimov V.B., Moiseeva N.I., Sulimov A.V., Ovsij O.G. [A Method of Predicting Outcomes of Breast Cancer]. *Patent Russian Federation № 2563437*, Date of Application 26.06.2014, Date of Issue 20.09.2015. [Генс, Г.П. Способ прогнозирования исходов рака молочной железы / Г.П. Генс, В.Б. Сулимов, Н.И. Моисеева, А.В. Сулимов, О.Г. Овсий // Патент Российской Федерации № 2563437, заявлено 26.06.2014, опубл. 20.09.2015.]

Aleksey V. Sulimov, lead programmer, laboratory of Computational Systems and Applied Programming Technologies, Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation, sulimovv@mail.ru.

Aleksey N. Meshkov, candidate of medicine, head of the laboratory of Molecular Genetics, National Research Center for Preventive Medicine of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation, meshkov@lipidclinic.ru.

Igor A. Savkin, programmer, laboratory of Computational Systems and Applied Programming Technologies, Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation, is@dimonta.com.

Ekaterina V. Katkova, candidate of physical and mathematical sciences, researcher, laboratory of Computational Systems and Applied Programming Technologies, Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation, katkova@dimonta.com.

Danil K. Kutov, programmer, laboratory of Computational Systems and Applied Programming Technologies, Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation, dk@dimonta.com.

Zuhra B. Hasanova, junior researcher, Russian Cardiology Research and Production Complex of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation, zukhra@yandex.ru.

Nina V. Konovalova, junior researcher, Russian Cardiology Research and Production Complex of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation, miirka@mail.ru.

Valery V. Kukharchuk, corresponding member of Russian Academy of Sciences, doctor of medicine, professor, Russian Cardiology Research and Production Complex of the Ministry of Healthcare of the Russian Federation, Moscow, Russian Federation, v_kukharch@mail.ru.

Vladimir B. Sulimov, doctor of physical and mathematical sciences, head of the laboratory of Computational Systems and Applied Programming Technologies, Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation, vs@dimonta.com.

Received October 9, 2015

УДК 519.226.3, 616.153.922 **DOI: 10.14529/jcem150402**

ПОЛНОГЕНОМНЫЙ АНАЛИЗ ГЕНЕТИЧЕСКИХ АССОЦИАЦИЙ ДЛЯ ПРЕДСКАЗАНИЯ ПОЛИГЕННОЙ ГИПЕРХОЛЕСТЕРИНЕМИИ С ИСПОЛЬЗОВАНИЕМ БАЙЕСОВСКИХ СЕТЕЙ

А.В. Сулимов, А.Н. Мешков, И.А. Савкин, Е.В. Каткова, Д.К. Кутов, З.Б. Хасанова, Н.В. Коновалова, В.В. Кухарчук, В.Б. Сулимов

Проведен полногеномный анализ генетических ассоциаций с показателями липидного обмена с применением технологии байесовских сетей для постановки диагноза полигенной гиперхолестеринемии на основе генетических данных российской популяции пациентов. Были проанализированы данные 1200 пациентов, для каждого из которых кроме клинической информации, показателей липидного профиля – различных видов холестерина, были получены 196725 однонуклеотидных полиморфизмов (SNP). Для первоначального отбора наиболее значимых параметров использовался полногеномный анализ ассоциаций (GWAS) и статистический метод критерия согласия Пирсона.

Были исследованы два состояния пациента связанные с липидным обменом: уровень ХС-ЛПНП (липопротеины низкой плотности) и ХС-ЛПВП (липопротеины высокой плотности). Для предсказания уровня липопротеинов использовались байесовские сети простейшей топологии – наивной, а для оценки качества (надежности) предсказания применялось построение ROC-кривых и вычисление площади под этими кривыми (AUC). После отбора значимых параметров с помощью методов GWAS или Пирсон величина AUC повышалась от 0,5 для начальной сети до 0,9. Дальнейшее повышение AUC до 0,99 и уменьшение числа прогностических параметров до 150 проводилось с помощью оптимизации байесовской сети по числу узлов-параметров, где целевой функцией была величина AUC. Показана неоднозначность получения прогностических параметров при различных способах первоначального уменьшения числа узлов сети с помощью метода GWAS и Pirson. Несмотря на очень хорошие результаты по качеству предсказания, полученные на обучающей выборке, для независимой контрольной группы пациентов были получены не высокие значения AUC. Дальнейшее применение предложенной в настоящей статье методологии возможно при существенном уменьшении числа SNP на основе анализа молекулярных механизмов.

Ключевые слова: GWAS; LDL-C; HDL-C; SNP; байесовская сеть.

Сулимов Алексей Владимирович, ведущий программист, лаборатория вычислительных систем и прикладных технологий программирования, Научно-исследовательский вычислительный центр, Московский государственный университет имени М.В. Ломоносова (г. Москва, Российская Федерация), sulimovv@mail.ru.

Мешков Алексей Николаевич, кандидат медицинских наук, заведующий, лаборатория молекулярной генетики, ФГБУ "Государственный научно-исследовательский центр профилактичепской медицины" Министерства здравоохранения Российской Федерации (г. Москва, Российская Федерация), meshkov@lipidclinic.ru.

Савкин Игорь Алексеевич, программист, лаборатория вычислительных систем и прикладных технологий программирования, Научно-исследовательский вычислительный центр, Московский государственный университет имени М.В. Ломоносова (г. Москва, Российская Федерация), is@dimonta.com.

Каткова Екатерина Владимировна, кандидат физико-математических наук, научный сотрудник, лаборатория вычислительных систем и прикладных технологий программирования, Научно-исследовательский вычислительный центр, Московский государственный университет имени М.В. Ломоносова (г. Москва, Российская Федерация), katkova@dimonta.com.

Кутов Данил Константинович, программист, лаборатория вычислительных систем и прикладных технологий программирования, Научно-исследовательский вычислительный центр, Московский государственный университет имени М.В. Ломоносова (г. Москва, Российская Федерация), dk@dimonta.com.

Хасанова Зухра Биляловна, младший научный сотрудник, ФГБУ "Российский кардиологический научно-производственный комплекс" Министерства здравоохранения Российской Федерации (г. Москва, Российская Федерация), zukhra@yandex.ru.

Коновалова Нина Валерьевна, младший научный сотрудник, ФГБУ "Российский кардиологический научно-производственный комплекс" Министерства здравоохранения Российской Федерации (г. Москва, Российская Федерация), miirka@mail.ru.

Кухарчук Валерий Владимирович, член-корреспондент РАН, доктор медицинских наук, профессор, ФГБУ "Российский кардиологический научно-производственный комплекс" Министерства здравоохранения Российской Федерации (г. Москва, Российская Федерация), v_kukharch@mail.ru.

Сулимов Владимир Борисович, доктор физико-математических наук, заведующий, лаборатория вычислительных систем и прикладных технологий программирования, Научно-исследовательский вычислительный центр, Московский государственный университет имени М.В. Ломоносова (г. Москва, Российская Федерация), vs@dimonta.com.

Поступила в редакцию 9 октября 2015.