Научно-исследовательский институт
клинической и экспериментальной лимфологии
филиал ФИЦ ИЦиГ СО РАН

Integrated bioinformatics analysis of the gene structure of polygenic diseases to develop effective methods for their early prognosis

BGRS / SBioMed-2022

Prokofiev V., Shevchenko A., Konenkov V.

Research Institute of Clinical and Experimental Lymphology – Branch of the Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia of Cytology and Genetics, SB RAS, Novosibirsk, Russia

INTRODUCTION

It is impossible to imagine modern medical genetics of polygenic human diseases without the widespread use of intelligent analysis and machine learning methods, in particular, such as simulation of stochastic processes, MDR analysis, bioinformatic analysis of complex biological networks based on graph theory, which allow replacing the original, largely stochastic, object with its image - a mathematical model with further study of this model using computer-implemented computational logic algorithms or heuristic [1, 2, 3]. In connection with the above, it is very relevant to conduct comparative clinical and genetic studies with the analysis of combined genotypes in combination with various variants of mathematical modeling, the results of which can become the basis for the development of fundamentally new ways to predict the development of socially significant human diseases, their early diagnosis and prevention.

1. Moore J.H., Folkert W.A., Scott M.W. Bioinformatics challenges for genome-wide association studies. *Bioinformatics Review*. 2010;26:445–455.

2. Lvovs D., Favorova O.O., Favorov A.V. A polygenic approach to the study of polygenic diseases. *Acta Naturae*. 2012;4:59-71.

3. Deo R.C. Machine Learning in Medicine. *Circulation*. 2015;132(17):1920–1930.

PATIENTS AND METHODS

The study of the features of the genetic structure of cytokines, angiogenesis and vascular remodeling factors in healthy individuals and patients with various forms of diseases - myocardial infarction, breast cancer, rheumatoid arthritis, female infertility of unknown origin and type 2 diabetes mellitus was carried out. 526 healthy and 1074 with various pathology people were examined. The study was conducted using the SNP methods, various variants of data mining and machine learning (simulation modeling, MDR method, bioinformatic analysis of gene networks) [4, 5].

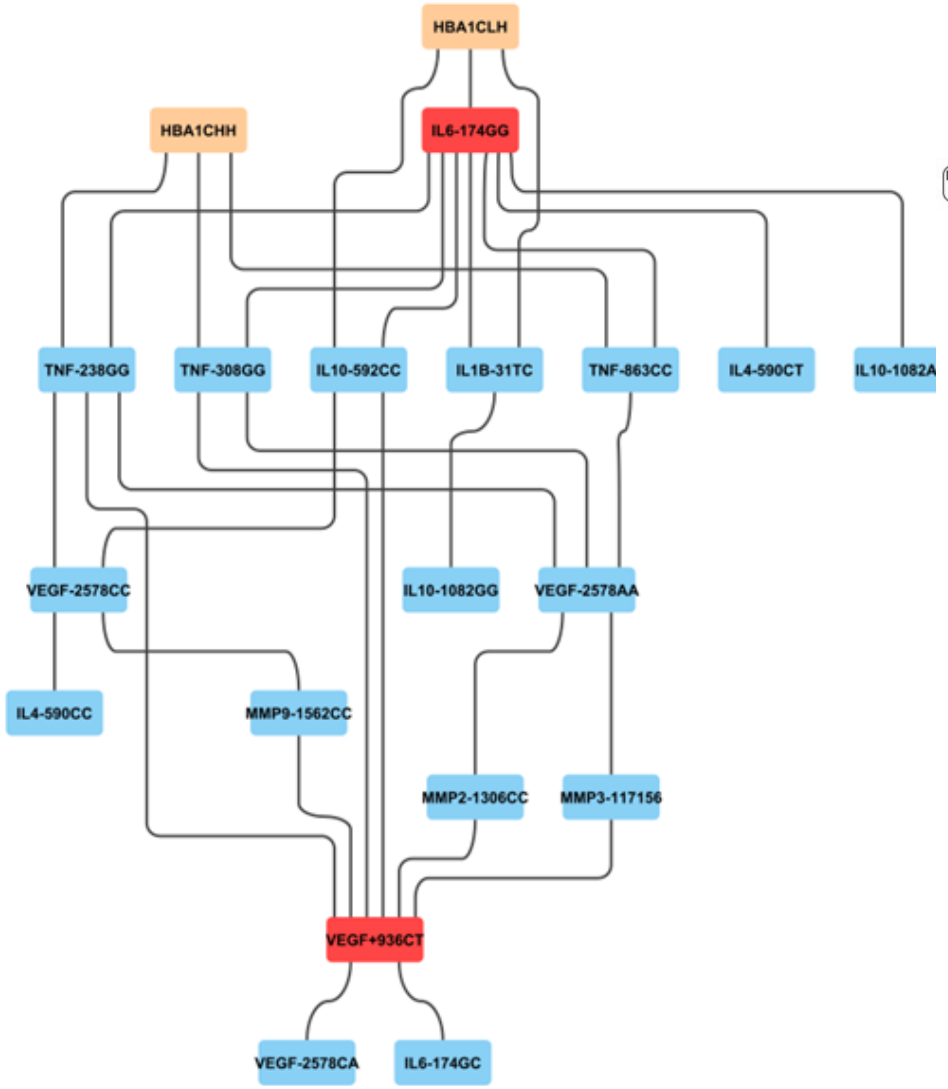
Statistical processing of the results was carried out based on a methodological, including a comprehensive computer analysis of gene chains of various dimensions with an analysis of their structural and functional organization in terms of their possible pathogenetic role in the development of the particular polygenic disease [6]. This approach allowed us to identify highly informative gene ensembles associated with the development of each of the studied diseases.

4. Lance W.H, Marylyn D.R. and Moore J.H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics.* 2003;19(3):376–382.

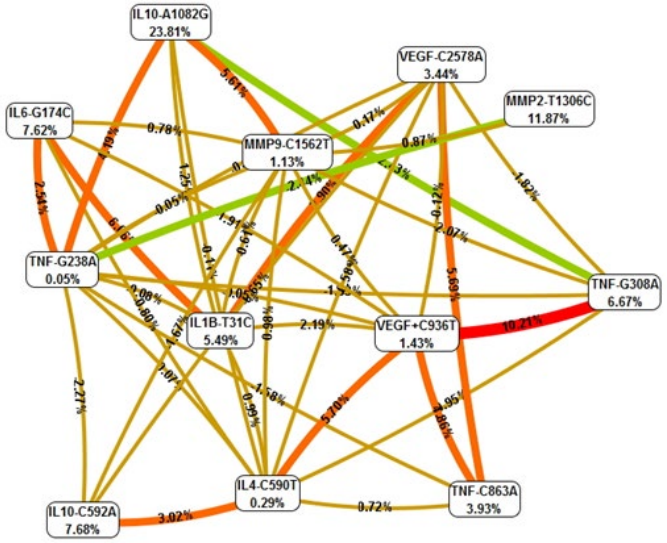
5. Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B. and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504.

6. Prokofiev V. Konenkov V., Koroleva E., Shevchenko A., Dergacheva T. and Novikov A. The structure of the cytokine gene network in uterine fibroids. *2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB).* 2020:261-264. doi: 10.1109/CSGB51356.2020.9214588

A FRAGMENT OF THE GENE NETWORK OF CYTOKINES IN DIABETIC RETINOPATHY



An entropy graph reflecting intergenic interactions in the resulting model of primary infertility of unclear genesis



RESULTS

With the help of these genetic complexes and combined methods of mathematical modeling, bioinformatic matrices were created, on the basis of which, for each of the multifactorial diseases, prognostic algorithms with an optimal ratio of sensitivity and specificity indicators were developed, ensuring maximum accuracy of the method (up to 94%) with a minimum set of prognostic predictors. It is shown that for the construction of prognostic models of polygenic diseases with the greatest predictive potential, it can be effective to use simulation modeling based on a computational algorithm embedded in the modified procedure of sequential recognition A. Wald. To improve the quality of predictive models, it is necessary to use a combined mathematical modeling method that combines simulation modeling, MDR analysis and graphical visualization of gene networks using the Cytoscape platform.

OPPERATIONAL CHARACTERISTICS OF PREDICTIVE MODELS

Diseases	Sensitivity (%)	Specificity (%)	Accuracy (%)
Myocardial infarction	78.5	87.6	81.1
Breast cancer	85.3	79.7	82.6
Rheumatoid arthritis	89.6	88.5	88.8
Diabetic retinopathy	87.8	100.0	94.2
Primary infertility	93.8	80.0	86.1

CONCLUTION

The results of the conducted research can serve as a scientific basis for the development of effective technologies for the prediction, diagnosis and treatment of the polygenic diseases.