*Proceedings of 2012 4th International Conference on Machine Learning and Computing*
***IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore***

# ANN Based Application of Pharmacogenetics to Personalized Cancer Treatment

Bekir Karlik[1+] and Emre Oztoprak[2]

[1] Department of Computer Engineering of Mevlana University Konya, Turkey
[2] Medicine Faculty of Mevlana University Konya, Turkey

**Abstract.** Artificial Neural Networks (ANN) may probably be the single most successful technology in the last two decades which has been widely used in a large variety of applications in biomedical areas. ANN is used in pharmaceutical (pharmacokinetic and pharmacogenetic) areas to model complex interactions and predict the nonlinear relationship between causal factors and response variables. The aim of this study is indicate a novel approach on application of pharmacogenetics to personalized cancer treatment using data of TPMT polymorphisms and ANN.

**Keywords:** artificial neural networks, personalized medicine, pharmacogenetics, TPMT, drug, cancer.

## 1. Introduction

Pharmacogenomics aims to understand pharmacological response with respect to genetic variation. Essential to the delivery of better health care is the use of pharmacogenomic knowledge to answer questions about therapeutic, pharmacological or genetic aspects [1]. Pharmacogenomics is the application of genomic technologies to drug discovery and development, as well as for the elucidation of the mechanisms of drug action on cells and organisms. DNA microarrays measure genome-wide gene expression patterns and are an important tool for pharmacogenomic applications, such as the identification of molecular targets for drugs, toxicological studies and molecular diagnostics. Genome-wide investigations generate vast amounts of data and there is a need for computational methods to manage and analyze this information [2].

As pharmacogenetic researchers gather more detailed and complex data on gene polymorphisms that effect drug metabolizing enzymes, drug target receptors and drug transporters, they will need access to advanced statistical tools to mine that data. These tools include approaches from classical biostatistics, such as logistic regression or linear discriminant analysis, and supervised learning methods from computer science, such as support vector machines and ANN [3-4]. Warnecke-Eberz U. et al. (2010) used TaqMan low-density arrays and analysis by ANN predicts response to neoadjuvant chemoradiation in esophageal cancer. Neoadjuvant radiochemotherapy of locally advanced esophageal cancer is only effective for patients with major histopathological response. A total of 17 genes were selected to predict histopathologic tumor responses to chemoradiation (cisplatin, 5-fluorouracil, 36 Gy). For gene-expression analysis quantitative TaqMan low-density arrays were applied. Expression levels in pretreatment biopsies of 41 patients (cT2-4, Nx, M0) were compared with the degree of histopathologic regression in resected specimens applying univariate, multivariate and ANN analysis. Multivariate analysis of the marker combination provided response prediction with 75.0% sensitivity, 81.0% specificity and 78.1% accuracy. ANN analysis was the best predictive model for major histopathologic response (80% sensitivity, 90.5% specificity and 85.4% accuracy), representing a clinically practical system. Low-density-array RT-PCR analyzed by ANN predicts histopathologic response to neoadjuvant chemoradiation in their patient collective, and could be used to further individualize treatment strategies in esophageal cancer [5]. Lin E. Et al. (2006) demonstrated that a

+ Corresponding author. Tel.: +90332 (4444243); fax: +90332(2411111).
*E-mail address*: bkarlik@mevlana.edu.tr

trained ANN model is a promising method for providing the inference from factors such as single nucleotide polymorphisms (SNPs), viral genotype, viral load, age and gender to the responsiveness of interferon [6]. Chao-Cheng Lin et al. (2008) aimed to train and validate ANN, using clinical and pharmacogenetic data, to predict clozapine response in schizophrenic patients. Five pharmacogenetic variables and five clinical variables were collated from 93 schizophrenic patients taking clozapine, including 26 responders. ANN analysis was carried out by training the network with data from 75% of cases and subsequently testing with data from 25% of unseen cases to determine the optimal ANN architecture. Then the leave-one-out method was used to examine the generalization of the models. The optimal ANN architecture was found to be a standard feed-forward, fully-connected, back-propagation multilayer perceptron. The overall accuracy rate of ANN was 83.3%, which is higher than that of logistic regression (LR) (70.8%). By using the area under the receiver operating characteristics curve as a measure of performance, the ANN outperformed the LR. The gene polymorphisms should play an important role in the prediction [7]. Serretti and Smeraldi (2004) tested a neural network strategy for a combined analysis of two gene polymorphisms. A Multi-Layer Perceptron model showed the best performance and was therefore selected over the other networks. One hundred and twenty one depressed inpatients treated with fluvoxamine in the context of previously reported pharmacogenetic studies were included. The polymorphism in the transcriptional control region upstream of the 5HTT coding sequence (SERTPR) and in the Tryptophan Hydroxylase (TPH) gene was analyzed simultaneously. A multi-layer perceptron network composed by 1 hidden layer with 7 nodes was chosen. 77.5 % of responders and 51.2% of non-responders were correctly classified. Finally, they performed a comparison with traditional techniques. A discriminant function analysis correctly classified 34.1 % of responders and 68.1 % of non-responders. Overall, their findings suggest that ANN may be a valid technique for the analysis of gene polymorphisms in pharmacogenetic studies. The complex interactions modeled through ANN may be eventually applied at the clinical level for the individualized therapy [8]. Cosgun E. et al. (2011) have applied three machine learning approaches: Random Forest Regression (RFR), Boosted Regression Tree (BRT) and Support Vector Regression (SVR) to the prediction of warfarin maintenance dose in a cohort of African Americans. They have developed a multi-step approach that selects SNPs (Single Nucleotide Polymorphism), builds prediction models with different subsets of selected SNPs along with known associated genetic and environmental variables and tests the discovered models in a cross-validation framework. Preliminary results indicate that their modelling approach gives much higher accuracy than previous models for warfarin dose prediction. A model size of 200 SNPs (in addition to the known genetic and environmental variables) gives the best accuracy. The R(2) between the predicted and actual square root of warfarin dose in this model was on average 66.4% for RFR, 57.8% for SVR and 56.9% for BRT. Thus RFR had the best accuracy, but all three techniques achieved better performance than the current published R(2) of 43% in a sample of mixed ethnicity, and 27% in an African American sample [9].

E. Himes et al. (2009) sought to relate candidate gene SNP data with bronchodilator response and measure the predictive accuracy of a model constructed with genetic variants. Bayesian networks, multivariate models that are able to account for simultaneous associations and interactions among variables, were used to create a predictive model of bronchodilator response using candidate gene SNP data from 308 Childhood Asthma Management Program Caucasian subjects. The model found that 15 SNPs in 15 genes predict bronchodilator response with fair accuracy. Bayesian networks are an attractive approach to analyze large-scale pharmacogenetic SNP data because of their ability to automatically learn complex models that can be used for the prediction and discovery of novel biological hypotheses [10]. Larder B. et al. (2008) describe that the development and application of ANN models as alternative tools for the interpretation of HIV genotypic drug resistance data. A large amount of clinical and virological data, from around 30,000 patients treated with antiretroviral drugs, has been collected by the HIV Resistance Response Database Initiative (RDI, www.hivrdi.org) in a centralized database. Treatment change episodes (TCEs) have been extracted from these data and used along with HIV drug resistance mutations as the basic input variables to train ANN models [11]. Sabbagh and Darlu (2006) investigated the ability of several pattern recognition methods to identify the most informative markers in the CYP2D6 gene for the prediction of CYP2D6 metabolizer status. Four data-mining tools were explored: decision trees, random forests, artificial neural networks, and the multifactor dimensionality reduction (MDR) method. Marker selection was performed separately in eight population samples of different ethnic origin to evaluate to what extent the most informative markers differ across ethnic groups. Their results show that the number of polymorphisms required predicting CYP2D6 metabolic phenotype with a high accuracy can be dramatically reduced owing to the strong haplotype block structure observed at CYP2D6. MDR and neural networks provided nearly identical results and performed the best. Data-mining methods, such as MDR and neural networks, appear as promising tools to improve the

efficiency of genotyping tests in pharmacogenetics with the ultimate goal of pre-screening patients for individual therapy selection with minimum genotyping effort [6].

Pharmacogenomics is a new field which uses genetic information to estimate drug treatment response. The person's drug-therapeutic or toxic molecular genetic basis of response to context clarification on the new drugs and genes engaged in the discovery of the target points to a branch of sciences. Currently, there are only a few pharmacogenetic diagnostic tests available, and clinical guidelines for pharmacogenetically tailored therapy are lacking [12]. In clinical pharmacology, detailed data about the complex molecular mechanisms of the interactions between drug(s) and organism become available. Most notably, the target genes of many drugs are being discovered and the differential genes expression induced by drugs can be investigated by microarray techniques. However, genetic variation can account for as much as 95 percent of variability in drug disposition and effects [13]. One of the best-developed examples of pharmacogenetics applied to clinical practice is the enzyme thiopurine methyltransferase (TPMT). TPMT is responsible for the degradation of azathioprine and mercaptopurine, which are commonly used to treat acute leukemia, inflammatory bowel disease, rheumatoid arthritis, and transplant immune suppression. Drug metabolizing enzymes participate in the neutralizing of xenobiotic and biotransformation of drugs. Polymorphisms in the drug-metabolizing enzyme coding genes alter the activity of these enzymes for some substrates. Thiopurine S-methyltransferase (TPMT) is a cytosolic enzyme that catalyzes the S-methylation of aromatic and heterocyclic sulfhydryl compounds like 6-mercaptopurine (6MP), which is used to treat patients with acute lymphoblastic leukemia (ALL). Polymorphisms in the genes encoding cytochrome p450 (CYP) and thiopurine S-methyl transferase (TPMT) enzymes catalyze the metabolic reactions of several drugs. These polymorphisms might be responsible for adverse drug reactions. TPMT activity is related to the outcome and/or toxicity of therapy. Patients with inherited very low levels of TPMT activity are at increased risk for thiopurine-induced toxicity, when treated with standard doses of these drugs [14].

ANN despites its recent rapid growth in the implementation in various fields of applications, have their potential in clinical pharmacology largely unexplored. This study presents an ANN approach for personalized cancer treatment using pharmacogenomic data which provides to describe data of TPMT polymorphisms. For this purpose, a supervised learning named error back-propagation training algorithm which consists of presenting input and output data to the multi-layered perceptron architecture of neural network. This training is considered complete when the neural network reaches a user defined performance level. This level signifies that the network has achieved the desired statistical accuracy as it produces the required outputs for a given sequence of inputs. The codes of the pharmagenucleotides in the genes are used letters of alphabet such as A, C, G, T which are not numerical values. To prevent this handicap the ASCII codes of these letters can be used as input values of NN.

## 2. Material and Methods

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. ANN is an adaptable system that can learn input-output relationships through repeated presentation of data and is capable of generalizing to new, previously unseen data. It can be trained by submitting several sets of input data with their associated output, and during the training, the network learns to associate particular sets of inputs with particular outputs by adapting its free parameters. Neural networks are sometimes called machine learning algorithms, because changing of its connection weights (training) causes the network to learn the solution to the problem. The strength of connection between the neurons is stored as a weight-value for the specific connection. The system learns new knowledge by adjusting these connection weights. The learning ability of a neural network is determined by its architecture and by the algorithmic method chosen for training. There are two types of learning as supervised and unsupervised. In supervised learning, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. In unsupervised learning, all the observations are assumed to be caused by latent variables, that is, the observations are assumed to be at the end of the causal chain. The universal approximation theorem for neural networks states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated arbitrarily closely by a multi-layer perceptron with just one hidden layer. This result holds only for restricted classes of activation functions. Some of the most commonly used activation functions are to solve non-linear problems such as Sigmoid and Hyperbolic

Tangent Function [15]. Multi-layer networks use a variety of learning techniques, the most popular being *back-propagation*. The Back Propagation NN works in two modes, a supervised training mode and a production mode. The training can be summarized as follows [16]:

1. Start by initializing the input weights for all neurons to some random numbers between 0 and 1, then:
2. Apply input to the network.
3. Calculate the output.
4. Compare the resulting output with the desired output for the given input. This is called the *error*.
5. Modify the weights and threshold q for all neurons using the *error*.

The challenge is to find a good algorithm for updating the weights and thresholds in each iteration (step 4) to minimize the *error*. The structure of multi-layer perceptron (MLP) is like 31:31:3, which means 31neurons of input layer, 31 neurons of hidden layer, and 3 neurons of output layer. In this study, the back-propagation learning algorithm is used since it is the most popular supervised learning algorithm. Neural Network System primarily consists of 2 steps:

## 2.1. Training of Neural Network

TPMT data is used to train Neural Network. Training data was collected from 4 healthy people (control) and 8 leukemia patients. Genomic DNA was extracted from peripheral blood by using proteinase K/salting out method. Primer design and restriction enzyme analysis were performed according to previous studies The samples without any TPMT*2, *3A, *3B and *3C mutations were genotyped as TPMT wild type allele (TPMT*1), the samples with one deficient allele (TPMT*1/*2, *1/*3C, *1/*3B, *1/*3A) were genotyped as heterozygous and the samples with two deficient alleles (TPMT*2/*3C, *2/*3B, *3C/*3B, *2/*3A etc.) were genotyped as homozygous. The samples that carried both the G460A and A719G mutations were named TPMT *3A [17-18]. TPMT data was collected by Fatih University, Department of Biology, in Istanbul. Each class of output determines Cancer (continue to drugs), Cancer (discontinue to drugs), and No cancer (no drug) as shown in Fig.1. This software named Prosthesis has been developed by B. Karlik using Delphi.

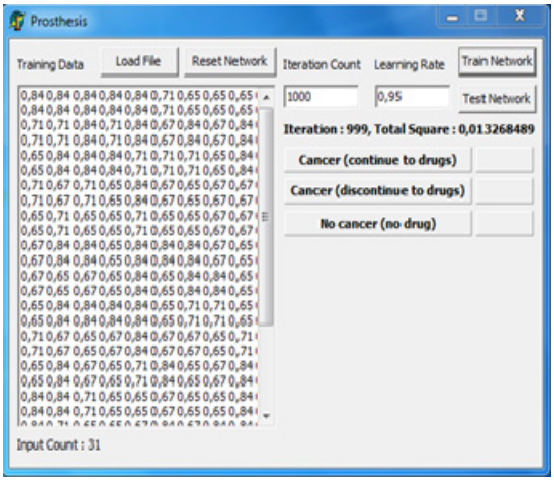


Fig. 1: User interface of program.

Where; Train: Train the network with given parameters. Load File: it is to load for both input and desired output data. Test: Test the network for new inputs. Reset: Reset network for the initial weights. Iteration: number of iteration, Learning rate is to find optimum learning rate (here we found as 0.95). Error describes total Mean Square Error (MSE).

## 2.2. Testing of Neural Network

Testing data is recorded from 8 persons. Each class of output determines normal healthy person and person with leukemia cancer which consists of 31 normalized data. To run the program, first of all, training_data.txt must be loaded by using Load File button. Then by clicking the Train Network button, Train operation is done. After that, again by using the Load File button, test_data.txt is loaded and test operation is done by using the Test Network button.

# 3. Results and Discussion

One of the main factors preventing a more efficient use of new pharmacological treatments for chronic diseases such as hypertension, cancer, Alzheimer disease or obesity is represented by the difficulty of

predicting “a priori” the chance of response of the single patient to a specific drug. This study presents a new application of ANN methods to detect drug therapy of leukemia cancer. Proposed study supports use of personalized drug therapy in clinical practices. This tool can be used for treatment of variety diseases with similar characteristics. This method may provide tools for clinical association studies and help find genetic TPMT (or SNPs) involved in responses to therapeutic drugs or adverse drug reactions. Moreover, this preliminary study can be improved and applied to solve similar treatment problems. In future work, we can compare these two methods with the other classifiers methods such as fuzzy classifier, RBF, LVQ, etc.

# 4. References


[1] M. Dumontier and N. Villanueva-Rosales. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*. March 2009, **10**(2): 153-163.

[2] M. Ringnér, et al. Analyzing array data using supervised methods. *Pharmacogenomics*. 2002, **3**(3): 403-415.

[3] W. Shannon , R. Culverhouse, J. Duncan. Analyzing microarray data using cluster analysis. *Pharmacogenomics*. 2003, **4**(1): 41-52.

[4] Audrey Sabbagh and Pierre Darlu. Data-Mining Methods as Useful Tools for Predicting Individual Drug Response: Application to CYP2D6 Data. *Hum Hered* . 2006, **62**: 119–134.

[5] U. Warnecke-Eberz, R. Metzger, E. Bollschweiler, S.E. Baldus, R. P. Mueller, H. P. Dienes, A. H. Hoelscher, P. M. Schneider. TaqMan low-density arrays and analysis by artificial neuronal networks predict response to neoadjuvant chemo radiation in esophageal cancer. *Pharmacogenomics*. 2010, **11**(1): 55-64.

[6] E. Lin, Y. Hwang, S. C. Wang, Z.J. Gu, E. Y. Chen. An artificial neural network approach to the drug efficacy of interferon treatments. *Pharmacogenomics*. 2006, **7**(7): 1017-24.

[7] C.C. Lin, Y.C. Wang, J. Y. Chen, Y.J. Liou, Y. M. Bai, I. C. Lai, T.T. Chen, H. W. Chiu, Y. C. Li. Artificial neural network prediction of clozapine response with combined pharmacogenetic and clinical data. *Computer Methods and Programs in Biomedicine*. 2008, **91**: 91–99.

[8] A. Serretti and E. Smeraldi. Neural network analysis in pharmacogenetics of mood disorders. *BMC Medical Genetics*. 2004, **5**(27) doi:10.1186/1471-2350-5-27.

[9] E. Cosgun, N. A. Limdi, C. W. Duarte. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics*. May 2011, **15**:27(10), pp. 1384-1489.

[10] B. E. Himes, A. C. Wu, Q. L. Duan, et al. Predicting response to short-acting bronchodilator medication using Bayesian networks. *Pharmacogenomics*. 2009, **10**(9), 1393–1412.

[11] B. Larder, D. Wang, A. Revell. Application of artificial neural networks for decision support in medicine. *Methods Mol Biol*. 2008, **458**: 23-36.

[12] D. E. Lanfear, H. L. McLead. Pharmacogenetics using DNA to optimize drug therapy. *American Family Physician*. 2007, **76**(8): 1179-1182.

[13] A. G. Floares. Using computational intelligence to develop intelligent clinical decision support systems. *Proc. of 6th Int. Con. on Comp. Intel. Methods for Bioinfor. & Biostatistics*. 2010, ISBN:3-642-14570-1 978-3-642-14570.

[14] S. Zhou. Clinical Pharmacogenomics of Thiopurine S-methyltransferase. *Current Clin. Pharma.* 2006, **1**: 119-128.

[15] B. Karlik, A. V. Olgac. Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *Inter. Journal of Artificial Intelligence and Expert Systems*. 2011, **1**(4): 111-122.

[16] M. A. Hussain. C. R. Che Hassan, K. S. Loh, K. W. Mah. Application of Artificial Intelligence Techniques in Process Fault Diagnosis. *Journal of Engineering Science and Technology*. 2003, **2**(3): 260-270.

[17] M. M. Ameyaw, E. S. Collie-Duguid, R. H. Powrie, et al. Thiopurine methyltransferase alleles in British and Ghanaian populations. *Hum Mol Genet*. 1999, **8**: 367-370.

[18] M. A. Sayitoglu, I. Yildiz., O. Hatirnaz, U. Ozbek. Common Cytochrome p4503A (CYP3A4 and CYP3A5) and Thiopurine S-Methyl Transferase (TPMT) Polymorphisms in Turkish Population. *Turk J Med Sci*., 2006, **36**: 11-15.