Current Topics in Medicinal Chemistry. 2013; 13(5): 675-684

Applied computational techniques on schizophrenia using genetic mutations

Vanessa Aguiar-Pulido, Marcos Gestal, Carlos Fernandez-Lozano, Daniel Rivero, Cristian R. Munteanu

lnformation and Communications Technologies Department. Computer Science Faculty, University of A Coruña. Campus de Elviña s/n, 15071 Spain

Abstract:
Schizophrenia is a complex disease, with both genetic and environmental influence. Machine learning techniques can be used to associate different genetic variations at different genes with a (schizophrenic or non-schizophrenic) phenotype. Several machine learning techniques were applied to schizophrenia data to obtain the results presented in this study. Considering these data, Quantitative Genotype – Disease Relationships (QDGRs) can be used for disease prediction. One of the best machine learning-based models obtained after this exhaustive comparative study was implemented online; this model is an artificial neural network (ANN). Thus, the tool offers the possibility to introduce Single Nucleotide Polymorphism (SNP) sequences in order to classify a patient with schizophrenia. Besides this comparative study, a method for variable selection, based on ANNs and evolutionary computation (EC), is also presented. This method uses half the number of variables as the original ANN and the variables obtained are among those found in other publications. In the future, QDGR models based on nucleic acid information could be expanded to other diseases.

Keywords: Bioinformatics, Data mining, Machine learning, Neural networks, Schizophrenia, SNP, Support vector machines.

1. INTRODUCTION

The study of diseases with computational models uses different molecular information such as structure and physical/chemical properties of a protein and DNA/RNA molecules, blood proteome mass spectra, DNA microarrays, disease biomarkers and concentration of the metabolites in physiological liquids. Schizophrenia is a common mental disorder defined as a heterogeneous syndrome characterized by perturbations in language, perception, thinking, social relationships and will as a consequence of several cumulative effects of certain (genetic and environmental) risk factors [1] or epigenetics [2]. Due to the impact of this disease molecular genetics techniques have been used to identify the genes related with this disorder.

The computational methods are focused on finding the relationships between schizophrenia and molecular information. Quantitative Structure - Activity Relationships (QSARs) are widely used for predicting protein properties [3] and Quantitative Protein (or Proteome) - Disease Relationships (QPDRs) [4-10] for disease prediction. The numerical data used for these classifications consisted in topological indices or molecular descriptors obtained with the Graph/Complex Network theory [11-14]. Several QSAR/QPDR models based on protein structure and proteome mass spectra have been obtained for cancer [15-18], especially for breast and colorectal cancer [19, 20] and prostate cancer [21]. Additional applications have been published for protein interactions in parasites [22-24].

In a similar way, a QGDR can be established in order to automatically evaluate schizophrenia DNA sequences using Single Nucleotide Polymorphisms (SNP) data A SNP [25] is a single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population, that is, at 1east in 1%. Thus, SNPs can be used as inputs in disease computational studies such as pattern searching or classification models. Models based on machine learning have been extensively used to analyse complex diseases, such as diabetes [26], hepatitis [27] and rheumatoid arthritis [28]. However. not many studies have been carried out on variation analysis in schizophrenia using Machine Learning algorithms [29, 30]. Statistical models were the most used for this type of complex disease.

Ban *et al.* [26] analyzed the importance of gene-gene interactions in Type 2 diabetes mellitus (T2D) susceptibility by investigating 408 SNPs in 87 genes involved in major T2D-related pathways in 462 T2D patients and 456 healthy controls from the Korean cohort studies. They used the support vector machine (SVM) method to differentiate between cases and controls using SNP information in a 10-fold cross-validation test and they achieved a 65.3% prediction rate with a combination of .14 SNPs in 12 genes by using the radial basis function (RBF)-kernel SVM. As the high-throughput technology tor genome-wide SNPs improves, it is likely that a much higher prediction rate with biologically more interesting combination of SNPs can be acquired by using· this method. Thus, SVM-based feature selection method in this study found novel association between combinations of SNPs and T2D in a Korean population.

Uhmn *et al.* [27] used several machine learning techniques to predict the susceptibility to chronic hepatitis from SNP data integrated with several feature selection algorithms to identify a set of SNPs relevant to the disease. They applied a backtracking technique to a couple of feature selection algorithms, forward selection and backward elimination, and showed that it was beneficial to find the best solutions by experiment. The experimental results show that the decision rule was able to distinguish between chronic and normal hepatitis with a maximum accuracy of 73.20%, whereas the accuracy of the support vector machine was 67.53% and that of the decision tree was 72.68%. It was also shown that the decision tree and decision rule are potential tools to predict susceptibility to chronic hepatitis from SNP data.

Briggs *et al.* [28] studied the genetic interactions (epistasis) with a statistical approach, by combining several analytical methods. Thus, they used a multi-stage analysis that incorporated supervised machine learning and methods of association testing, to investigate epistatic interactions with a well-established genetic factor (PTPN22 1858T) in a complex autoimmune disease such as rheumatoid arthritis (RA). The analysis consisted of four principal stages: Stage I (data reduction) - identifying candidate chromosomal regions in 292 affected sibling pairs, by predicting PTPN22 concordance using multipoint identity-by-descent probabilities and a supervised machine learning algorithm (Random Forests); Stage II (extension analysis) - testing detailed genetic data within candidate chromosomal regions for epistasis with PTPN22 1858T in 677 cases and 750 controls using logistic regression; Stage III (replication analysis) - confrrmation of epistatic interactions in 947 cases and 1.756 controls; Stage IV (combined analysis) - a pooled analysis including all 1624 RA cases and 2,506 control subjects for final estimates of effect size. A total of seven replicating epistatic interactions were identified. The results demonstrate that the SNP variants within CDH13. MY03A, CEP72 and near WFDCl showed significant evidence for interaction with PTPN22, affecting susceptibility to RA.

One of the most studied genes related to schizophrenia susceptibility is DRD3. Same as HTR2A, it is considered to be an important target for several antipsychotic drugs [31]. HTR2A encodes one of the receptors for serotonin and DRD3 encodes one subtype of the five dopamine receptors, both neurotransmitters. More specifically, Dopamine 3 receptors (DRD3) are concentrated in limbic regions of the brain, which are associated with cognitive, emotional and endocrine functions. Thus, it may be particularly relevant to schizophrenia, as the DRD3 messenger RNA is predominantly expressed in the limbic system, a region thought to be dysfunctional in this disease [32]. Association studies involving these functional candidate genes have systematically focused on a limited set of SNPs, generally based on previously reported small contributions of these markers of risk of susceptibility to schizophrenia. More specifically, SNP TI 02C (rs6313) at HTR2A and SNP Ser9Gly (rs6280) at DRD3 have been extensively analyzed in several schizophrenia case-control studies [33].

The serotonin transporter gene (SLC6A4) and its promoter (5-HTTLPR) polymorphism have been the focus of a large number of association studies of behavioral traits and psychiatric disorders such as schizophrenia. However, large-scale genotyping of the polymorphism has been very difficult. Lu *et al.* [30] reported the development and validation of a 5-HTTLPR genotype prediction model. The single nucleotide polymorphisms (SNPs) from the 2,000 kb region surrounding 5-HTTLPR were used to construct a prediction model through a newly developed machine learning method, multicategory vertex discriminant analysis with 2.147 individuals from the Northern Finnish Birth Cohort genotyped with the Illumina 370K SNP array and manually genotyped for 5-HTTLPR polymorphism. The prediction model was applied to SNP genotypes in a Dutch/German schizophrenia case-control sample of 3,318 individuals to test the association of the polymorphism with schizophrenia. The prediction model of eight SNPs achieved a 92.4% accuracy rate and a 0.98±0.01 area under the receiving operating characteristic. Thus, evidence tor an association of these SNPs with schizophrenia was observed (P=0.05, odds ratio=1.105). This prediction model provides an effective substitute of manually genotyped 5-HITLPR alleles, providing a new approach for large scale association studies of this polymorphism.

The current review will present details about the comparative study of machine learning disease classification models using only SNPs at the HTR2A and DRD3 genes in Galician (Northwestern Spain) ·schizophrenic patients. Methods such as ANNs [34], SVMs [35-37], EC [38-40] and other machine learning techniques [41) have been used to find the best classification models.

Once this comparison was finished. the machine learning-based method which obtains the best results in Ref. [42] was implemented online as SNP-Schizo (http://bioaims.udc.es/SNPSchizo.php) in the Bio-AIMS server. This too] also includes an approach for variable selection, based on ANNs and evolutionary computation (EC).

2. MATERIALS AND METHODS

Fig. 1 summarizes the workflow followed by this approach. Firstly, data from patients is genotyped in order to obtain SNP sequences. After that, computational methods are applied to this data in order to obtain QGDR classification models. Finally, the models obtained are evaluated using new data. Thus, this procedure allows establishing relationships between SNP sequences and the predisposition to the disease.

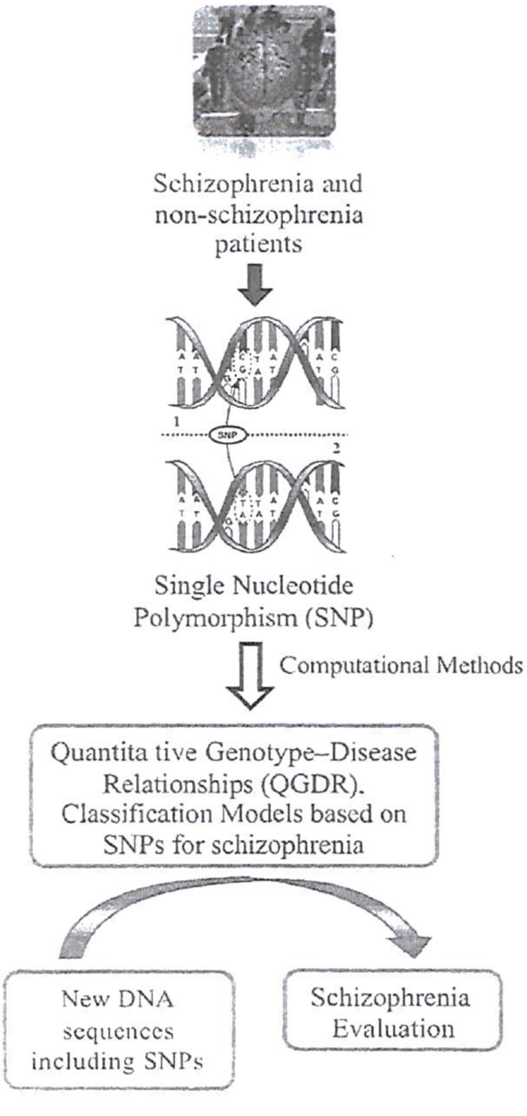

Schizophrenia and non-schizophrenia patients
1
SNP
2
Single Nucleotide Polymorphism (SNP)
Computational Methods
Quantita tive Genotype–Disease Relationships (QGDR). Classification Models based on SNPs for schizophrenia
New DNA sequences including SNPs
Schizophrenia Evaluation

Fig. (1). QGDR Modcl classification.

2.1. Schizophrenia Data

For the comparative study, schizophrenia data collected from Galician patients [33] were used. These data contained 48 SNPs at the DRD3 and HTR2A genes, which are associated to schizophrenia. These SNPs were encoded taking different values:

- 0 if homozygous (both copies of a given gene have the same allele) tor the first allele (one of a number of alternative forms of the same gene occupying a given position on a chromosome),
- 1 if heterozygous (the patient has two different alleles of a given gene).
- 2 if homozygous for the second allele or
- 3 if unknown .

The original dataset contained 260 positive subjects (genetically predisposed to schizophrenia) and 354 negative subjects (not predisposed), a total of 614 patients.

To perform further tests, six other datasets were obtained from the original one. This was carried out by adding negative subjects generated with the HAP-SAMPLE [43] simulation tool. These data were modified to include genotyping errors (represented as value 3) taking into account the error frequencies of the real data, but choosing randomly which positions were modified. Thus, these datasets included 307, 614, 1.228, 1.842, 2,456 and 3.070 simulated negative subjects. Datasets were named following the pattern 1:N. where this label represents the proportion between the real subjects (positive and negative) and the simulated negative subjects.

However, there are several genetic data simulation packages. Among those, we consider the coalescent-based methods [44], which have been used for population-based simulation in genetic studies, such as GENOME [45]. This method was developed to overcome previous limitations. HAPSAMPLE [43], which is the simulator used in this paper, uses the existing Phase I/II HapMap data to resample existing phased chromosomes to simulate datasets. There are also forward-time population simulations. such as easyPOP [46], FPG [47], FREGENE [48], simuPOP [49] and genomeSlMLA [50]. The last method can simulate realistic patterns of LD in both family-based and case-control datasets and, unlike other similar packages, has proved to be an effective platform for simulating large scale genetic data. Another program capable of generating large sea le genetic as well as phenotypic variation data is presented in ref. [51]. This program generates genotypes/phenotypes by perturbing real data, with the aim of creating a large number of replicates that share similar properties with real data. Nevertheless, since HAP-SAMPLE is an association simulator for candidate regions and was specifically designed for simulating SNP genotypes for case-control studies, it was the most appropriate simulator given the nature of the original data used in this study.

2.2. Machine Learning Methods

ANNs [52, 53] have been extensively used for classification problems. More specifically, a multilayer Linear Neural Network (LNN) has been used. This technique uses a linear network model, as the activation functions is linear, and always has an equivalent single layer counterpart [54]. The Multilayer Perceptron (MLP) [52] has also been used. Other types of networks considered were the Radial Base Functions (RBF). In this type of network. the neurons of the hidden layer perform a calculation function instead of an activation function of the MLP.

Same as the MLP, SVMs belongs to non-linear classifiers. SVMs induce·linear separators or hyperplanes in the space of characteristics. This type of classifier has proved to be very useful when dealing with high dimensionality problems [55].

Bayesian methods are based on Bayes' theory of probability. Not only they allow performing classification, but they also allow finding relationships among attributes. Several of these methods have been used, such as Naive Bayes [56] (which assumes that the attributes are independent and Bayesian Networks [57].

The following techniques allow obtaining classification models based on "IF-THEN-ELSE" rules or on hierarchical structures such as trees. More specifically, rule inference models·from Decision Tables [58] are obtained by building a decision table majority classifier. This type of method evaluates feature subsets using best-first search and uses the nearest-neighbor method to determine the class for each instance that is not covered by the decision table or by the Decision Table Naive Bayes Hybrid Classifier (DTNB) [59]. A similar model was considered to infer decision trees, following a hybrid approach between the decision trees and the Naïve Bayes classifier, called Best-First decision Tree classifier (BFTree) [60].

A boosting meta-algorithm was also included in this study. This algorithm consists of combining multiple classification models that complement each other. The Adaptive Boosting (AdaBoost) [61] method builds the models iteratively, weighing the instances differently in each iteration. The new models classify the instances that the previous models did not classify correctly.

Multifactor Dimensionality Reduction (MDR) [62, 63] is a data mining approach designed to detect and characterize non-linear interactions among discrete attributes or variables that influence a binary outcome (for example, case-control status). It is a constructive induction algorithm which reduces the. Original n-dimensional model to a one-dimensional model, repeating this procedure for each possible n-factor combination and selecting the combination that maximizes the case-control ratio of the high-risk group. This method is considered to be a non-parametric alternative to traditional statistical methods. The MDR software combines attribute selection, attribute construction and classification with cross-validation. This method has mostly been used to detect gene-gene interactions or epistasis in genetic studies of common human diseases [64-66] such as schizophrenia [67-69], although it can also be applied to other domains.

2.3. Improving Machine Learning Methods by Means of Variable Selection

Once the comparative study is presented. a novel approach based on a previous variable selection will be discussed. This new approach uses Genetics Algorithms (GA) and ANNs in a first stage to establish which are the most relevant variables within the data. In the second stage, the classification stage, ANNs and SVMs will be used.

ANN and GA .for Variable Selection

GAs [70-72] represent a search method based on Charles Darwin's theory of Evolution [73]. This algorithm makes a population evolve through random actions similar to those existing in biological evolution such as mutations and genetic recombination, as well as selections with a certain criterion called fitness. The fitness is used to decide which individuals are selected, i.e., the most suitable individuals are those with the higher likelihood they will reproduce. Thus, the result of this method is a set of rules which are used to classify the input data. Thus, this method tries to find relationships between attributes or variables and a binary outcome [74-77].

ANN-GA approach [38. 39] uses “pruned” search, which starts by considering all the variables and gradually discards groups of them. The remaining set of variables is used to classify the samples, and the results are used to determine how relevant the discarded variables were for the classification. This process can be continued as long as the classification results are equal, or at least similar, to those obtained using the overall set of variables. Therefore, the GA determines how many and which variables will be considered for the classification. An ANN was included within the GA to evaluate the fitness values of the individuals. The use of an “inner” ANN to evaluate fitness avoids definition and optimization of more formal equations and, remarkably, yields generality to the approaches presented herein. As the goal is to determine which solutions, out of those provided by the GA, represent good starting points to get acceptable classification models, it is not required to fully train such ANN; instead, extending the training up to the point where the ANN starts converging is enough.

In other words, the pruned search consists of a stepwise approach by which the GA steadily reduces the number of variables characterizing the samples, until an optimal subsect is obtained. Each individual in the genetic population is initially described by d genes, each representing an original variable (using a binary encoding, each gene can be either 0 or 1).

Fitness will guide the pruning process (a black-box approach) to get individuals that, besides classifying as accurately as possible, use less variables. Eq. (1) defines how fitness can be described according to two parameters: the number of variables used to classify the samples and the quality of their classifications, calculated using the Mean Square Error (MSE) of the inner-ANN. Eq. (2) (employed here) shows that fitness will favor those individuals with less active genes (the denominator being the total number of variables).

fitness(individual;) = f(classitication$_i$) + f(Selected variables) (1)

fitness(individual$_i$) = MSE(ANN$_i$) + #1's genotype individual$_i$ / #total variables (2)

A good characteristic of Eqs. (1) and (2) is that they can be tailored. For instance, fitness may consider the cardinality (i.e . number of variables that have been select for classification) or the percentage of variables (regarding the overall initial set of variables) being used.

Similar approaches were applied in the diagnosis of dermatological diseases [78], prediction of outcome [79] or heart problems [80] among others. In other fields this kind or approaches are also widely used [81-83]

ANN and SVM.for Classijication

After the variables have been selected, a classification algorithm has been applied in order to build the classification model. In the variables selection phase, a simple ANN model was built tor the fitness score, but in the classification phase, this ANN has been replaced by a more complex model with more complex training. Several models have been tested, mainly SVMs and different ANN models. These ANN and SVM models have been developed using the Weka software [84], specifically, the MI.P and Sequential Minimal Optimization (SMO) algorithm implementations.

3. RESULTS

3.1. Comparative Study

252 QDGR classification models were obtained after applying machine learning techniques to the data described previously. Seven datasets were used. The Weka software package [84] was used to perform the comparative study. This work presents the results achieved with the best responses provided by the most representative algorithms included in this software. In addition to LNN, the following techniques were applied to the datasets: MLP, RBF, EC, MDR. Naive Bayes, Bayes Networks, SVM, Decision Tables, DTNB, BFTree and AdaBoost.

After carrying out this comparative study, the neural network model was implemented online. This approach consists of a type of ANN, hereafter referred to as LNN, that has a linear activation function in all neurons. More specifically, it is a multilayer neural network, with 40 neurons in the first layer, 152 in the second layer and 1 neuron as output. The number of input neurons was selected according to the results obtained from several feature selection methods (Best First [85], Linear Forward Selection [86], FCBF Search [87], Genetic Search [88], Scatter Search [89] and Random Search [90]). After several runs, it was proved that taking as input only the 40 neurons selected by the previous selection methods, the method achieved good results.

A graphical representation of the evolution of the different methods is shown in (Fig. 2). As said before, 1:N represents the proportion between the real subjects and the simulated negative subjects. Thus, the first dataset does not include any simulated subject and the last dataset includes 5 simulated subjects per real one.

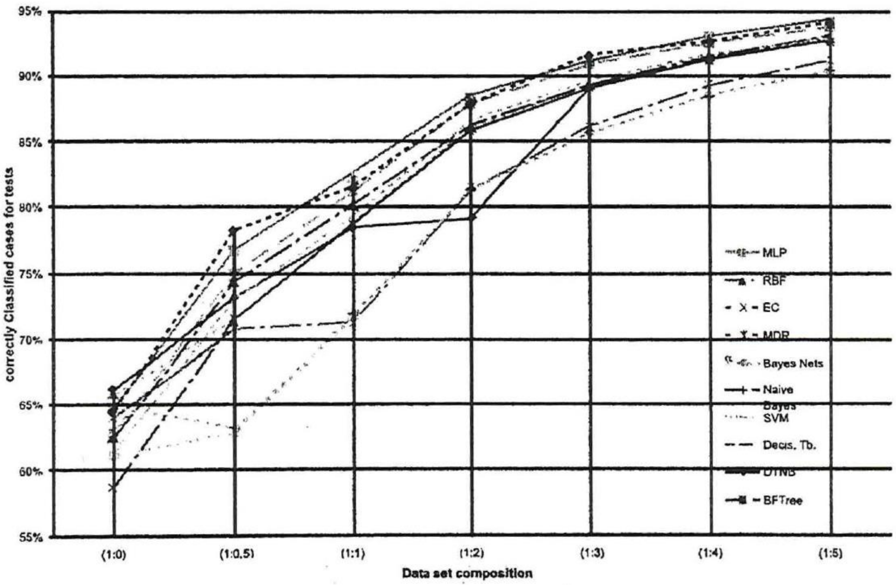


Fig. (2). Classification results of the different methods.

For each method, the percentage of correctly classified subjects is shown for each dataset. It can be observed that the classification percentages do not increase significantly after adding five parts of simulated subjects. Thus, we will focus on the results obtained for the dataset which contains the lowest number of simulated subjects, that is, the 1:0.5 dataset.

Classification accuracy percentages range from 56.6 to 66.6% for 1:0, which is the original dataset. For the datasets which included simulated subjects, these percentages range from 60.5 to 78.2% for 1:0.5, 69.8 to 83.0% for 1:1, 76.2 to 88.8% for 1:2, 84.8 to 91.5% for 1:3. 87.4 to 93.2% for 1:4 and 88.4 to 94.3% for 1:5.

Among the best models, the LNN described above is proposed. This QGDR model includes only a minimum of simulated subjects (1:0.5). Thus, this dataset is made up of 921 subjects: 260 real positive subjects, 354 real negative subjects and 307 simulated negative subjects for schizophrenia As mentioned previously, this neural network is based on 40 SNPs which are taken as an input and it has a hidden layer of 152 neurons. This technique obtained 78.2% in test accuracy When the 1:0.5 dataset was used as input.

The LNN achieves good results for all the datasets, as it is simpler and less computationally expensive than other methods. In (Fig. 3), the area under the receiver operating curve (AUC-ROC) for the cross-validation group (0.8405) demonstrates that this model is not a random one. In addition, for this model, the threshold is 0.8.

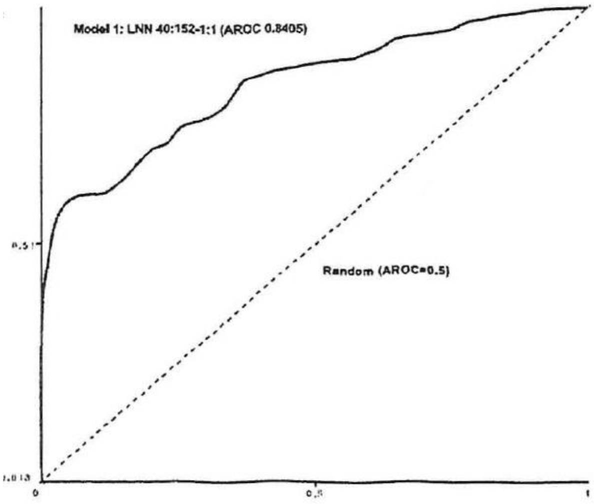


Fig. (3). Area under the receiver operating characteristic curve (AUC-ROC).

3.2. Results After Variable Selection

It is important to point out that ii1 order to characte1ize a complex disease, 40 SNPs are too many. Therefore, another approach based on genetic algorithms, artificial neural networks and SVMs, which was described previously, was applied to the 1:0.5 dataset. The best SVM classification model was a Weka SMO implementation with a complexity of 5. building logistics models [91], using a polynomial kernel [84]. Thus, the model is based on a LNN 40:1 52-1-l, data set 1:0.5 and it has a test accuracy of 78.2%.

This method obtains similar results to the LNN in terms of classification scores and AUC-ROC values (Table 1**)** using less than half of the variables: only 17 variables were considered, instead of the 40 variables required by the previous LNN.

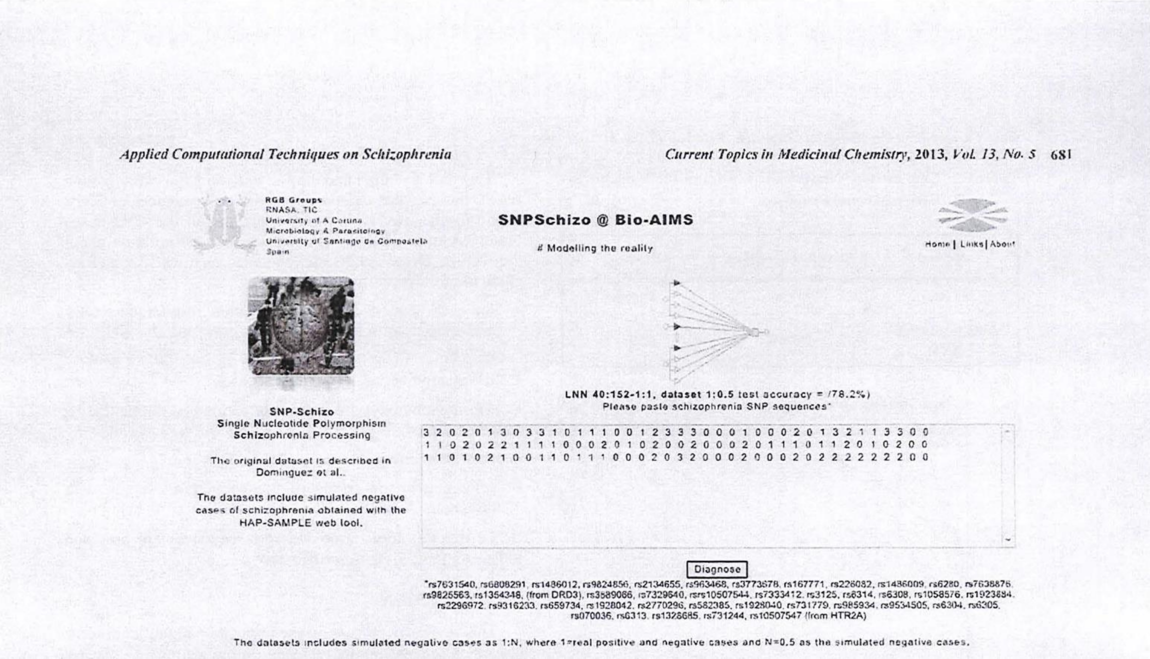
Table 1. Comparison between the LNN and the Variable Selection Method Proposed.

Method	Classification Scores	AUC-ROC
LNN	78.20%	0.8405
Variable selection - SVM	76.98%	0.824

3.3. Single Nucleotide Polymorphism Schizophrenia Processing (SNP-Schizo)

Bio-AIMS (Biomedical Artificial Intelligence Model Server) [16, 23. 92, 93] is a portal that offers theoretical models based on Artificial intelligence, Computational Biology and Bioinformatics to study Complex Systems in OMICS (Genomics, Transcriptomics, Metabolomics, Reactomics) that are relevant for Cancer, Neurosciences, Cardiovascular diseases, Parasitology, Microbiology and other Biomedical research in general (http://bio-aims.tic.udc.es/). It is the result of the collaboration between several scientific institutions. This portal includes two parts: TargetPred (Target Prediction) and DiscasePred (Disease Prediction). The DiseastPred part includes biomedicine applications for predicting human diseases from different data sources, such as genotypcs. Future tools will be implemented based on the published models using EEG recordings and blood proteome mass spectra for epilepsy and colorectal cancer.

SNPSchizo (Single Nucleotide Polymorphism Schizophrenia Processing) [94] is the result of an online implementation (http://miaja .tic.udc.es/Bio-AIMS/SNPSchizo.php) of the previously described machine learning method which takes as input SNPs from two different genes related to schizophrenia and performs a classification [42] (see Fig. 4). The interface of this tool was implemented using PHP, XHTML and Python, and the method was implemented using Java and Weka's [84] APls. The tool is running on Apache HTTP- Server.



Applied Computational Techniques on Schizophrenia *Current Topics in Medicinal Chemistry*, **2013**, *Vol. 13, No. 5* 681

RGB Groups
RNASA. TIC
University of A Coruna
Microbiology & Parasitology
University of Santiago de Compostela
Spain

SNPSchizo @ Bio-AIMS

Modelling the reality

SNP-Schizo
Single Nucleotide Polymorphism
Schizophrenia Processing

The original dataset is described in Dominguez et al..

The datasets include simulated negative cases of schizophrenia obtained with the HAP-SAMPLE web tool.

LNN 40:152-1:1, dataset 1:0.5 test accuracy = /78.2%)
Please paste schizophrenia SNP sequences*

```
3 2 0 2 0 1 3 0 3 3 1 0 1 1 1 0 0 1 2 3 3 3 0 0 0 1 0 0 0 2 0 1 3 2 1 1 3 3 0 0
1 1 0 2 0 2 2 1 1 1 1 0 0 0 2 0 1 0 2 0 0 2 0 0 0 2 0 1 1 1 0 1 1 2 0 1 0 2 0 0
1 1 0 1 0 2 1 0 0 1 1 0 1 1 1 0 0 0 2 0 3 2 0 0 0 2 0 0 0 2 0 2 2 2 2 2 2 2 0 0
```

*rs7631540, rs6808291, rs1486012, rs9824856, rs2134655, rs963468, rs3773678, rs167771, rs226082, rs1486009, rs6280, rs7638876, rs9825563, rs1354348, (from DRD3), rs3589066, rs7329640, rsrs10507544, rs7333412, rs3125, rs6314, rs6308, rs1058576, rs1923884, rs2296972, rs9316233, rs659734, rs1928042, rs2770296, rs582385, rs1928040, rs731779, rs985934, rs9534505, rs6304, rs6305, rs070036, rs6313, rs1328685, rs731244, rs10507547 (from HTR2A)

The datasets includes simulated negative cases as 1:N, where 1=real positive and negative cases and N=0.5 as the simulated negative cases.

Fig. (4). SNP-Schizo web tool.

This tool is simple and easy to use. To get a classification result, the user has to introduce a list of sequences of SNPs in the format used by the tool and click on the "Diagnose” button. A new window will pop up with information about the results. These results can be saved as a text file and include the following information:

- For each sequence: the classification result (genetically predisposed to schizophrenia or not) and the SNP sequence.
- Information about the original dataset.
- Information about the method implemented online and its test accuracy.
- Input SNP order.
- Reference to the article of the comparative study with these data.

To test this tool, three example sequences are provided following the coding described above.

4. CONCLUSION

This review is presenting the applied computational techniques on schizophrenia, focusing on the genotype – disease relationships based on information of the nucleic acids such as the genetic mutations: Several machine learning methods have been described including a method for variable selection based on an ANN and a GA. To test the different methods, real clinical data of an association study on Galician patients (Spain) who suffered from schizophrenia using the DRD3 and HTR2A genes have been used, as well as simulated data which were generated with specialized software.

In complex diseases such as schizophrenia, the factors involved in increasing the risk of developing a disease do not correspond to one or two genes. There is a combination of values from different sets of SNPs, as well as a great influence due to environmental factors, which increase the risk of developing this complex disease.

The classification results obtained with the original data are not good with any of the presented methods. When the number of control subjects in the training sets is increased using simulated data, the developed method improves its classification accuracy, obtaining better results than with those methods which provide objective infi1rmation about SNPs, obtaining a model based on rules or on trees. One of the models that obtain the best classification scores was implemented online as a free web tool named SNP-Schizo. The model implemented was the result of applying a LNN to the dataset that contained the lowest number of simulated subjects.

It is also interesting to observe which variables (or SNPs) are taken into account by tbc different methods which perform variable selection. Costas *et al.* used a sliding window approach and confirmed the existence of a common protective haplotype, which included the SNPs rs963468, rs2134655, rs 1486012 and rs7631540 at DRD3, against schizophrenia [95]. The ANN and GA for variable selection approach presented in this paper is capable of finding three of the four previous SNPs (rs2134655, rs1486012 and rs7631540) if 17 variables are considered. However, if 21 variables are considered, this approach finds all the SNPs included in the publication, as well as rs9824856, which is located at the same region. The same results in terms of classification scores and AUC-ROC values are obtained considering either 17 or 21 variables.

This review demonstrated the power of machine leaning in obtaining genotype - disease classifications using molecular structure information such as the genetic mutations and it proposes the application on other diseases. The results can be used to determine the future gene targets for new drugs or genetic treatments.

CONFLICT OF lNTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGE.MENTS

Vanessa Aguiar-Pulido and Cristian R. Munteanu acknowledge the funding support for a research position by the "Plan 12C" and “Isidro Parga Pondal” programs from Xunta de Galicia (Spain) and the European Social Fund (ESF), respectively, being the former one also co-funded by FEDER. This work is supported by the following projects: “Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies", Ibero-NBIC Network (209RT-0366l funded by CYTED (Spain). "Development of new image analysis techniques in 2D Gel for biomedical research” (ref.10SIN 105004PR) funded by Xunta de Galicia. RD07/0067/0005 funded by the Carlos III Health institute, and “Galician Network for Colorectal Cancer Research” (REGICC, Ref. 2009/58) from the General Directorate of Scientific and Technologic Promotion of the Galician University System of Xunta de Galicia.

DISCLOSURE

Part of information included in this article has been previously published in Molecules, "Machine Leaning Techniques for Single Nucleotide Polymorphism-Disease Classification Models in Schizophrenia" and in Lecture Notes in Computer Science. "'SNP-Schizo: A Web Tool for Schizophrenia SNP Sequence Classification".

REFERENCES

[1] Picchioni, M.M.; Murray, R.M. Schizophrenia. *BMJ (Clinical researched, 2*007, *335,* (7610), 91-95.

[2] Svrakic, D.M.; Zonunski, C.F.; Svrakic, N.M.; Zwir, I.; Cloninger, C.R. Risk architecture of schizophrenia: the role of epigenetics. *Current opinion in psychiatry,* 2013. *[Epub ahead of print]*.

[3] Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and QSPR.* Gordon and Breach: The Netherlands, 1999.

[4] Barabasi, A.L.; Bonabeau, E. Scale-free networks. *Scientific American,* 2003, *288,* (5), 60-69.

[5] Balaban, A.T.; Basak, S.C.; Beteringhe, A.; Mills, D.; Supuran, C.T. QSAR study using topological indices for inhibition of carbonic anhydrase II by sulfanilamides and Schiff bases. *Molecular Diversity.* 2004, *S,* (4), 401-412.

[6] Barabasi, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nature Reviews: Genetics,* 2004, *5,* (2), 101-113.

[7] Barabasi, A.L. Sociology. Network theory-the emergence of the creative enterprise. *Science.* 2005, *308,* (5722), 639-641.

[8] González-Díaz, H.; Villar, S.; Santana, L.; Uriarte, E.; Medicinal Chemistry and Bioinformatics - Current Trends in Drug's Discovery with Networks Topological Indices. *Current Topics in Medical Chemistry,* 2007, 7, (10), 1025-1039.

[9] Ferino, G.; González-Díaz, H.; Delogu, G.; Podda, G.; Uriarte, E. Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. *Biochemical and biophysical research communications,* 2008, 372, (2), 320-325.

[10] González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E. Proteomics, networks and connectivity índices. *Proteomics.* 2008, 8, (4), 750-778.

[11] Randic, M. Novel shape descriptors for molecular graphs. *J Chem Inf Comput Sci,* 2001, 41, (3), 607·613.

[12] Balaban, A.T., Chemical graphs. XXXIV Five new topological indices for the branching of tree-like graphs. *Theor Chim Acta,* 1979, 53, 355-375.

[13] Balaban, A.T.; Balaban, T.S., New vertex invariants and topological indices of chemical graphs based on infonnation on distances. *J Math Chem,* 1991, 8, 383-397.

[14] González-Díaz, H.; Pérez-Montoto, L.G.; Duardo-Sánchez, A.; Paniagua, E.; Vázquez-Prieto, S.; Vilas, R.; Dea-Ayuela, M.A.; Bolas-Fernández, F.; Munteanu, C.R.; Dorado, J.; Costas, J.; Ubeira, F.M. Generalized lattice graphs for 2D-visualization of biological information. *Journal of theoretical biology,* 2009, 261, (1). 136-147.

[15] González-Díaz, H.; Ferino, G.; Prado-Prado, F.J.; Vilar, S.; Uriarte, E.; Pazos, A.; Munteanu, C.R In *An Omics Perspective on Cancer Research.* Cho, W.C.S., Ed.; Springer, 2010.

[16] Munteanu, C.R.; Vázquez, J.M.; Dorado, J.; Sierra, A.P.; Sánchez González, A.; Prado-Prado, F.J.; González-Díaz, H., Complex network spectral moments for ATCUN motif DNA cleavage: first predictive study on proteins of human pathogen parasites. *Journal of proteome research,* 2009, 8, (11), 5219-5228.

[17] Vázquez-Naya, J.M.; Martínez-Romero, M.; Porto-Pazos, A.B.; Novoa, F.; Valladares-Ayerbes, M.; Pereira, J.; Munteanu. C.R; Dorado. J., Ontologies of drug discovery and design for neurology, cardiology and oncology. *Curr Phann Des.* 2010, 16, (24). 2724-2736.

[18] Martínez-Romero, M.; Vázquez-Naya, J.M.; Rabunal. J.R; Pita Fernández. S.; Macenlle, R.; Castro-Alvariño, J.; López-Roses, L.; Ulla, J.L.; Martínez-Calvo, A.V.; Vázquez, S.; Pereira, J.; Porto Pazos, A.B.; Dorado, J.; Pazos, A.; Munteanu, C.R. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Current drug metabolism,* 2010, 11, (4), 347-368.

[19] Munteanu, C.R.; Magalhaes, A.L.; Uriarte. E.; González-Díaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *Journal of theoretical biology,* 2009, 257, (2), 303-311.

[20] Martínez-Romero, M.; Vázquez-Naya, J.M.; Rabuñal, J.R.; Pita Fernández, S.; Macenlle, R.; Castro-Alvariño, J.; López-Roses, L.; Ulla, J.L.; Martínez-Calvo, A.V.; Vázquez, S.; .Pereira, J.; Porto Pazos, A.B.; Dorado, J.; .Pazos, A.; Munteanu, C.R. Artificial lntelligence Techniques for Colorectal Cancer Drug Metabolism: Ontologies and Complex Networks. *Current drug metabolism,* 2010, 11, (4), 347-368.

[21] Cruz-Monteagudo, M.; Munteanu, C.R.; Borges, F.; Cordeiro, M.N.; Uriarte, E.; González-Díaz. H. Quantitative Proteome Property Relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. *Bioorganic & medicinal chemistry,* 2008, 16, (22), 9684-9693.

[22] González-Díaz, H.; Muino, L.; Anadon, A.M.; Romaris, F.; Prado Prado, F.J.; Munteanu. C.R.; Dorado. J.; Sierra. A.P.; Mezo. M.; González-Warleta. M.; Garate, T.; Ubeira, F.M. MISS-Prot: web server for self/non-self discrimination of protein residue networks in parasites: theory and experiments in Fasciola peptides and Anisakis allergens. *Mol Biosyst,* 2011, 7, (6), 1938-1955.

[23] Rodríguez-Soca, Y.; Munteanu, C.R.; Dorado, J.; Pazos, A; Prado Prado, F.J.; González-Díaz, H. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *Journal of proteome research,* 2010, 9, (2), 1182-1190.

[24] Rodríguez-Soca, Y.; Munteanu, C.R.; Dorado, J.; Rabuñal. J.; Pazos, A.; González-Díaz, H. Plasmod-PPI: a web-server predicting complex biopolymer targets in Plasmodium with entropy measures of protein-protein interactions. *Polymer,* 2010, 51, (1), 264-273.

[25] den Dunnen, J.T.; Antonarakis, S.E. Mutation nomenclature extensions und suggestions to describe complex mutations: a discussion. *Human mutation,* 2000, 15. (1), 7-12.

[26] Ban, H.J.; Heo, J.Y.; Oh, K.S.; Park, K.J. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *BMC genetics,* 2010, 11, 26.

[27] Saangyong Uhnm, D.-H.K.; Young-Woong Ko; Sungwon Cho; Jaeyoun Cheong and Jin Kim. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. *Expert Systems,* 2009, 26,(1), 10.

[28] Briggs, F.B.; Ramsay, P.P.; Madden, E.; Norris, J.M.; Holers, V.M.; Mikuls, T.R.; Sokka. T.; Seldin, M.F.; Gregersen, P.K.; Criswell, L.A.; Barcellos. L.F. Supervised machine learning and logistic regression identities novel epistatic risk factors with PTPN22 for rheumatoid arthritis. *Genes and immunity,* 2010, 11, (3), 199-208.

[29] Nicodemus, K.K.; Callicott, J.H.; Higier, RG.; Luna, A. ; Nixon, D.C.; Lipska, B.K.; Vakkalanka, R.; Giegling, I.; Rujescu, D.; Clair, D.S.; Muglia, P.; Shugart, Y.Y.; Weinberger, D.R. Evidence of statistical epistasis between DJSCI, CIT and NDELI impacting risk for schizophrenia: biological validation with functional neuroimaging. *Human genetics.*

[30] Lu, A.T.; Bakker, S.; Janson, E.; Cichon, S.; Cantor, R.M.; Ophort, R.A. Prediction of serotonin transporter promoter polymorphism genotypes from single nucleotide polymorphism arrays using machine learning methods. *Psychiatric genetics,* 2012, 22, (4), 182-188.

[31] Meltzer, H.Y.; Matsubara, S.; Lee, J.C. Classification of typical and atypical antipsychotic drugs on the basis of dopamine D-1. D-2 and serotonin2 pKi values. *The Journal of pharmacology and experimental therapeutics,* 1989, 251, (1), 238-246.

[32] Suzuki, M.; Hurd, Y.L.; Sokoloft, P.; Schwartz, J.C.; Sedvall, G. D3 dopamine receptor mRNA is widely expressed in the human brain. *Brain research.* 1998, 779, (1-2), 58-74.

[33] Domínguez, .E.; Loza, M.L; Padin, F.; Gesteira, A.; Paz, E.; Paramo, M.; Brenlla, J.; Plunar, E.; Iglesias, F.; Cibeira, A.: Castro, M.; Caruncho, H.; Carracedo, A.; Costas, J. Extensive linkage disequilibrium mapping at HTR2A and DRD3 for schizophrenia susceptibility genes in the Galician population. *Schizophrenia research,* 2007, 90, (1-3), 123-129.

[34] Diederich, J. *Artificial neural networks: concept learning.* IEEE Press Piscataway, NJ. USA. 1990.

[35] Byvatov, E.; Schneider, G. Support vector machine applications in bioinformatics. *Applied Bioinformatics,* 2003, 2, (2). 67-77.

[36] Byvatov, E.; Schneider, G., Support vector machine applications in bioinformatics. *Appl Bioinformatics,* 2003, 2, (2), 67-77.

[37] Cristianini, N.; Shawe-Taylor, J. A*n introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge University Press, New York, LISA.1999.

[38] Gomez-Carracedo, M.P.; Gestal, M.; Dorado, J.; Andrade, J.M. Chemically driven variable selection by focused multimodal genetic algorithms in mid-IR spectra. *Analytical and Bioanalytical Chemistry,* 2007, 389, (7-8), 2331-2342.

[39] Gestal, M.; Gómez-Carracedo, M.P.; Andrade, J.M.; Dorado, J.; Fernández, E.; Prada. D.; Pazos, A. Selection of variables by genetic algorithms to classify apple beverages by artificial neural networks. *Applied Artificial Intelligence,* 2005, 19, (2), 181-198.

[40] Aguiar Pulido, V.; Seoane Fernández, J.A.; Freire, A.; Munteanu, C.R. Data Mining in Complex Diseases Using. Evolutionary Computation. *Lecture Notes in Computer Science,* 2009, 5517, (Part I), 917-924.

[41] Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining.* Pearson Addition Wesley: Boston. Maryland, 2006.

[42] Aguiar-Pulido, V.; Seoane, J.A.; Rabuñal, J.R.; Dorado, J.; Pazos, A.; Munteanu, C.R., Machine learning techniques for single nucleotide polymorphism - disease classification models in schizophrenia. *Molecules (Basel, Switzerland).* 2010, 15, (7). 4875-4889.

[43] Wright, F.A.; Huang, H.; Guan, X.; Gamiel, K.; .Jeffries, C.; Barry, W.T.; de Villena, F.P.; Sullivan, P.F. ; Wilhelmsen, K.C.; Zou, F. Simulating association studies: a data-based resampling. method for candidate regions or whole genome scans. *Bioinformatics,* 2007, 23, (19), 2581-2588.

[44] Kingman, J.F. Origins of the coalescent. 1974-1982. *Genetics.* 2000, 156, (4), 1461-1463.

[45] Liang, L.; Zollner, S.; Abecasis, G.R. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics,* 2007, 23, (11), 1565-1567.

[46] Balloux, F. EASYPOP (version 1.7): a computer program for population genetics simulations. *The Journal of heredity,* 2001, *92,* (3), 301-302.

[47] Hey, J. A computer program for forward population genetic simulation. 2004.

[48] Hoggart, C.J.; Chadeau-Hyam, M.; Clark, T.G.; Lampariello. R.; Whittaker, J.C.; De Iorio, M.; Balding, D.J. Sequence-level population simulations over large genomic regions. *Genetics.* 2007, 177, (3), 1725-1731.

[49] Peng, B.; Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics.* 2005, 21, (18), 3686-3687.

[50] Edwards, T.L.; Bush, W.S.; Turner, S.D.; Dudek, S.M.; Torstenson, E.S.; Schmidt, M.; Martin, E.; Ritchie, M.O. Generating Linkage Disequilibrium Patterns in Data Simulations using genomeSIMLA. *LNCS: Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics,* 2008, 4973, (2008), 24-35.

[51] Li, J.; Chen, Y. Generating samples for association studies based on HapMap data. *BMC bioinformatics.* 2008, *9,* 44.

[52] Bishop, C. *Neural Networks for pattern recognition.* Oxford University Press New York. 1995.

[53] Haykirt, S.; Network, N. A comprehensive foundation. *Neural Networks.* 2004, 3.

[54] Zhang, W. *Computational Ecology: Artificial Neural Networks and their Applications.* World Scientitic Publishing Company Incorporated. 2010.

[55] Vapnik, V. *Statistical Learning Theory.* John Weily and Sons: New York, 1998.

[56] John, G.H.; Langley, P. In *Estimating Continuous Distributions in Bayesian Classifiers.* Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, San Mateo; Morgan Kaufmann,199S; pp 338-345.

[57] Bouckaert, R.R. *Bayesian networks in Weka.* Computer Science Department University of Waikato: 2004.

[58] .Kohavi, R. In *The Power of Decision Tables.* Proceedings of 8th European Conference on Machine Learning. Heraclion, Greece: Springer-Verlag, 1995: pp 174-189.

[59] Hall, M.; Frank, E. In *21st Florida Artificial Intelligence Society Conference.* AAAI Press: Florida, USA, 2008, pp 318-319.

[60] Shi, H. *Best-first decision tree learning.* Hamilton: New Zeland, 2007.

[61] Yoav Freund. R.E.S. In *Experiments with a new boosting algorithm.* Thirteenth International Conference on Machine Learning. San Francisco, 1996.

[62] Moore, J.H.; Gilbert, J.C.; Tsai, C.T.; Chiang, F.T.; Holden, T.; Barney, N.; White, B.C. A flexible computational framework for detecting, characterizing. and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology,* 2006, 241, (2), 252-261.

[63] Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nature reviews: genetics.* 2009, 10, (6), 392-404.

[64] Greene, C.S.; Sinnott-Armstrong, N.A.; Himmelstein, D.S.; Park, P.J.; Moore, .I.H.; Harris, B.T. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics,* 2010, 2*6*,(5), 694-695.

[65] Cattaert, T.; Urrea, V.; Naj, A.C.; De Lobel, L.; De Wit, V.; Fu, M.; Mahachie John, J.M.; Shen, H.; Calle, M.L.; Ritchie, M.O.; Edwards, T.L.; Van Steen, K. FAM-MDR: a flexible family-bused multifactor dimensionality reduction technique to detect epistasis using related individuals. *PloS one*, 5, (4), e10304.

[66] He, H.; Oetting, W.S.; Brott, M.J.; Basu, S. Pair-wise multifactor dimensionality reduction method to detect gene-gene interactions in a case-control study. *Human heredity,* 2009, 69, (1). 60-70.

[67] Kang, S.G.; Lee, H. J.; Choi, J. E.; Park, Y.M.; Park, J.H.; Han, C.; Kim, Y.K.; Kim, S.H.; Lee, M.S.; Joe, S. H.; Jung, I.K.; Kim, L. Association Study between Antipsychotics- Induced Restless Legs Syndrome and Polymorphisms of Dopamine D1, D2, D3, and D4 Receptor Genes in Schizophrenia. *Neuropsychobiology,* 2008, 57, (1-2). 49-54.

[68] Vilella, E.; Costas, J.; Sanjuan, J.; Guitart, M.; De Diego, Y.; Carracedo, A.; Martorell, L.; Valero, J.; Labad, A.; De Frutos, R.; Najera, C.; Molto, M.D.; Toirac, I.; Guillamat, R.; Brunet, A.; Valles, V.; Pérez, L.; León, M.; de Fonseca, F.R.; Phillips, C.; Torres, M. Association of schizophrenia with DTNBP1 but not with DAO, DAOA, NRGI and RGS4 nor their genetic interaction. *Journal of psychiatric research*, 2008, 42, (4), 278-288.

[69] Yasuno, K.; Ando, S.; Misumi, S.; Makino, S.; Kulski, J.K.; Mumtake, T.; Kaneko, N.; Amagane, H.; Someya, T.; Inoko, H.; Suga, H.; Kanemoto, K.; Tamiya, G. Synergislic association of mitochondrial uncoupling protein (UCP) genes with schizophrenia. *American Journal of Medical Genetics. Neuropsychiatric Genetics,* 2007, 144B, (2), 250-253 .

[70] Holland, J. Adaptation in natural and artificial systems: an introductory analysis with applications to biology. control, and artificial intelligence. University of Michigan Press, 1975.

[71] Gestal, M.; Rivero, D.; Rabuñal, J.R.; Dorado, J.; Pazos, A. *Introducción a los Algoritmos Genéricos y Programación Genética.* Servicio de Publicaciones Universidade da Coruña, 2010.

[72] Hernández, J.A.; Dorado, J.; Gestal, M.; Porto, A.B. Avances en Algoritmos Evolutivos. *INTELIGENCIA ART1FICIAL Y COMPUTACIÓN AVANZADA,* 35.

[73] Darwin, C. On the Origin of Species by Means of Natural Selection. John Murray. 1859.

[74] Guo, L.; Rivero, D.; Dorado, J.; Munteanu, C.R.; Pazos, A. Automatic feature extraction using genetic programming: An application to epileptic EEG classification. *Expert Systems with Applications.* 2011, 38, (8), 10425-10436.

[75] Rivero, D.; Dorado, J.; Rabuñal, J.R.; Pazos, A. Modifying genetic programming for artificial neural network development for data mining. *Soft Computing-A Fusion of Foundations. Methodologies and Applications,* 2009, 13, (3), 291-305.

[76] Rivero, D.; Dorado, J.; Rabuñal, J.R.; Pazos, A.; Pereira. J. Artificial neural network development by means of genetic programming with graph codification. *Transactions on Engineering, Computing and Technology.* 2006, 16. 209-214.

[77] Cantú-Paz, E. In *Genetic and Evolutionary Computation Conference.* Springer-Verlag: Seattle, Washington, USA, 2004, pp 959-970.

[78] Fidelis, M.V.; Lopes, H.; Freitas, A. In *Evolutionary Computation. 2000. Proceedings of the 2000 Congress on* IEEE, 2000; Vol. 1, pp 805-810.

[79] Dybowski, R.; Gant, V.; Weller, P.; Chang, R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *The Lancet,* 1996, 347, (9009), 1146-1150.

[80] Anbarasi, M.; Anupriya, E.; Iyengar, N. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology,* 2010, 2, (10). 5370-5376.

[81] Gestal, M.; Andrade, J.M. In *Encyclopedia of Artificial lntelligence;* Information Science Reference: Hershey, EEUU, 2008. pp 581 -588.

[82] Seoane, J.; Aguiar, V.; Gestal, M.; Dorado, J.; Pazos, A. In *Association analysis in complex diseases using evolutionary computation,* 6th Intelligent System for Molecular Biology, Toronto, Canada, 2008; p 22.

[83] Gestal, M.; Cancela, A.; Andrade, J.; Gómez-Carracedo, M. In *Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications.* Sugaraman, V., Ed.; Information Science Reference: Hershey, New York, USA, 2007, pp 274-292.

[84] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, LA. The WEU Data Mining Software: An Update. *SIGKDD Explorations,* 2009, 11, (1).

[85] Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach (2nd Ed.).* Prentice Hall: Upper Saddle River, New Jersey: Prentice Hall, 2003.

[86] Gutlein, M.; Frank, E.; Hall, M.; Karwath, A. In *In Proceedings of Symposium on Computational lntelligence and Data Mining,* IEEE Computer Society: Nashville, TN, 2009, pp 332-339.

[87] Yu, L.; Liu, H. In Proceedings of the Twentieth International Conference on Machine Learning. 2003, pp 856-863.

[88] Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co.: Boston, MA, 1989.

[89] García López, F.; García Torres, M.; Melian Batista, B.; Moreno Pérez, J .A.; Moreno-Vega, J.M. Solving feature subset selection problem by a Parallel Scatter Search. *European Journal of Operational Research,* 2006, 169. (2), 477-489.

[90] Lin, H.; Setiono, R. In *13th International Conference on Machine Learning:* Bari, Italy, l996, pp 319-327.

[91] Landwehr, N.; Hall, M.; Frank, E. Logistic model trees. *Machine Learning,* 2005, 59, (1), 161-205. .

[92] González-Díaz, H.; Prado-Prado, F.J.; García- Mera, X.; Alonso, N.; Abeijón, P.; Caamano, O.; Yanez, M.; Munteanu, C.R ; Pazos Sierra, A.; Dea-Ayuela, M.A.; Gómez-Muñoz, M.T.; Garijo, M,M.; Sansano, J.; Ubeira, F.M. MIND-BEST: web server for drugs & target Discovery: design, synthesis, and assay of MAO-B inhibitors and theoretic-experimental study of G3PD protein from Trichomona gallinae. *Journal of proteome research*, 2010.

[93] Concu, R.; Dea-Ayuela, M.A.; Pérez-Montoto, L.G.; Prado-Prado, F.J.; Uriarte, E.; Bolas-Fernández, F.; Podda, G.; Pazos, A.; Munteanu, C.R.; Ubeira, F.M; González-Díaz, H. 3D entropy and moments prediction of enzyme clases and experimental-theoretic study of peptide fingertips in Leishmania parasites. *Biochimica et biophysica acta,* 2009, 1794, (12), 1784-1794.

[94] Aguiar-Pulido, V.; Seoane, J.; Munteanu, C.; Pazos, A. SNP-Schizo: A Web Tool for Schizophrenia SNP Sequence Classification. *LNCS: Advances in Computational Intelligence*, 2011, 6692, 252-259.

[95] Costas, J.; Carrera, N.; Domínguez, E.; Vilella, E.; Martorell, L.; Valero, J.; Gutiérrez-Zotez, A.; Labad, A.; Carracedo, A. A common haplotype of DRD3 affected by recent positive selection is associated with protection from schizophrenia. *Human genetics,* 2009, 124, (6), 607-613.