Molecules **2010**, *15*, 4875-4889; doi:10.3390/molecules15074875

OPEN ACCESS

molecules

ISSN 1420-3049

www.mdpi.com/journal/molecules

Article

Machine Learning Techniques for Single Nucleotide Polymorphism—Disease Classification Models in Schizophrenia

Vanessa Aguiar-Pulido, José A. Seoane, Juan R. Rabuñal, Julián Dorado, Alejandro Pazos and Cristian R. Munteanu *

Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, S/N, 15071 A Coruña, Spain; E-Mails: vanesa.aguiar@udc.es (V.A.-P.); jseoane@udc.es (J.A.S.); juanra@udc.es (J.R.R.); julian@udc.es (J.D.); apazos@udc.es (A.P.)

* Author to whom correspondence should be addressed; E-Mail: muntisa@gmail.com; Tel.: +34 981 167 000 Ext. 1302; Fax: +34 981 167 160.

Received: 4 June 2010; in revised form: 8 July 2010 / Accepted: 9 July 2010 / Published: 12 July 2010

Abstract: Single nucleotide polymorphisms (SNPs) can be used as inputs in disease computational studies such as pattern searching and classification models. Schizophrenia is an example of a complex disease with an important social impact. The multiple causes of this disease create the need of new genetic or proteomic patterns that can diagnose patients using biological information. This work presents a computational study of disease machine learning classification models using only single nucleotide polymorphisms at the HTR2A and DRD3 genes from Galician (Northwest Spain) schizophrenic patients. These classification models establish for the first time, to the best knowledge of the authors, a relationship between the sequence of the nucleic acid molecule and schizophrenia (Quantitative Genotype – Disease Relationships) that can automatically recognize schizophrenia DNA sequences and correctly classify between 78.3–93.8% of schizophrenia subjects when using datasets which include simulated negative subjects and a linear artificial neural network.

Keywords: DNA molecule; SNP; schizophrenia; artificial neural networks; evolutionary computation

---

1. Introduction

Disease computational studies use diverse types of data, such as the structure and physical/chemical properties of a protein and DNA/RNA molecules, blood proteome mass spectra, DNA microarray results, disease biomarkers and concentration of the metabolites in physiological liquids. Schizophrenia, which is a common disease, can be defined as a heterogeneous syndrome characterized by perturbations in language, perception, thinking, social relationships and will. There is not a set of symptoms which uniquely characterize the disease, and even though researchers have been looking for a unique cause of schizophrenia for years with no success, most of them have concluded that schizophrenia would be the consequence of several cumulative effects of certain risk factors (genetic and environmental) [1]. Several studies of families, twins and foster-children confirmed and have allowed quantification of the contribution of genetics to schizophrenia [2]. After this, molecular genetics techniques started to be used to identify the genes that caused the disease [3]. These genes are not the genes of schizophrenia themselves, but rather they may transmit a set of characteristics which would increase the risk of developing the disease.

One of the most studied genes in relation to schizophrenia susceptibility is DRD3. As well as HTR2A, it is considered to be an important target for several antipsychotic drugs [4,5]. 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 [6], as the DRD3 messenger RNA is predominantly expressed in the limbic system, a region thought to be dysfunctional in this disease [7,8].

Association studies involving these functional candidate genes have systematically focused on a limited set of Single Nucleotide Polymorphisms (SNPs), generally based on previously reported small contributions of these markers of risk of susceptibility to schizophrenia. More specifically, SNP T102C (rs6313) at HTR2A and SNP Ser9Gly (rs6280) at DRD3 have been extensively analyzed in several schizophrenia case-control studies [9]. A SNP [10] is a single nucleotide site where two (of four) different nucleotides occur in a high percentage (*i.e.*, at least 1 %) of the population.

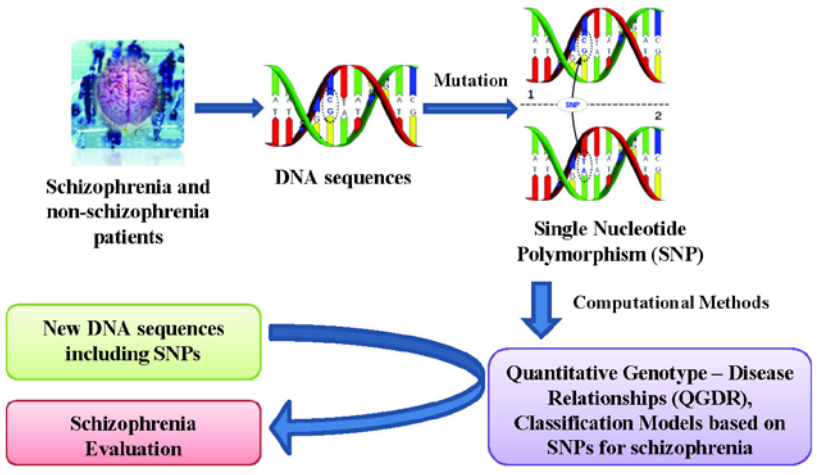
There are several studies on SNPs, such as that one in [11], where a method is presented for haplotype partitioning based on pairwise analysis of SNPs. A block-based approach for mapping a single locus trait was applied to blocks of different methods in a case-control study. Results show that any block-based association test is considerably more efficient than the conventional single site association trait and, in particular, the method presented performed best accuracy, even when a low marker density was available. Another study on SNPs is that one presented in [12]. In this paper, the use of two feature importance ranking measures (the modified t-test and F-statistics) is proposed to rank a large amount of SNPs and then the greedy manner together with a classifier are used in order to determine a desirable feature subset, which leads to the highest classification accuracy with the minimum size. Results show that both ranking methods are efficient at determining the important SNPs and they both find nearly the same amount of them. However, the first measure tends to be better in terms of classification accuracy. Compared to other methods, the results obtained in this paper are better.

There exist several genetic data simulation packages. Among those, we encounter coalescent-based methods [13], which have been used for population based simulation in genetic studies, such as GENOME [14]. This method was developed to overcome previous limitations. HAP-SAMPLE [15], 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 also exist forward-time population simulations, such as easyPOP [16], FPG [17], FREGENE [18], simuPOP [19] and genomeSIMLA [20]. 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 scale genetic and also phenotypic variation data is presented in [21]. 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.

Models based on Machine Learning have been extensively used to analyze complex diseases, such as diabetes [22], hepatitis [23], rheumatoid arthritis [24], *etc*. However, not many studies have been carried out on variation analysis in schizophrenia using Machine Learning algorithms [25]. Statistical models were the most used for this type of complex disease.

Quantitative Structure - Activity Relationships (QSARs) are widely used for predicting protein properties [26] and Quantitative Protein (or Proteome)-Disease Relationships (QPDRs) [27-33] for disease prediction. Recent works using complex networks of proteins or mass spectra of the human serum proteome have contributed to create theoretical models for cancer diagnosis and screening for cancer-related molecules in the case of colorectal [34,35], breast [34,36] and prostate [37-39] cancers. In a similar way, a Quantitative Genotype - Disease Relationship (QGDR) can be established in order to automatically evaluate schizophrenia DNA sequences using SNP data. Methods such as artificial neural networks [40], support vector machines [41], evolutionary computation [42,43] and other Machine Learning techniques [44] have been used in order to find the best classification models. This work presents a study of schizophrenia QGDR classification using only single nucleotide polymorphisms from Galician patients [9]. Thus, this information of the DNA molecule will be used as the input for several machine learning techniques that search for the best classification model capable of evaluating new schizophrenia DNA sequences (see Figure 1).

Figure 1. Flow chart of the QGDR model classification between the DNA structure (SNPs) and schizophrenia.



2. Results and Discussion

Two hundred and fifty two (252) QGDR classification models have been obtained using SNPs at two schizophrenia-related genes (each of them or both), twelve machine learning techniques and seven datasets, starting from the original data and using extra simulated negative (control) subjects (see Table 1). In terms of classification the subjects are organized in two groups: *Schizo* and *non-Schizo*. These models describe relationships between the DNA information (SNPs) and schizophrenia.

Table 1. The classification models obtained for the evaluated schizophrenia patients using the SNP information at DRD3 and HTR2A.

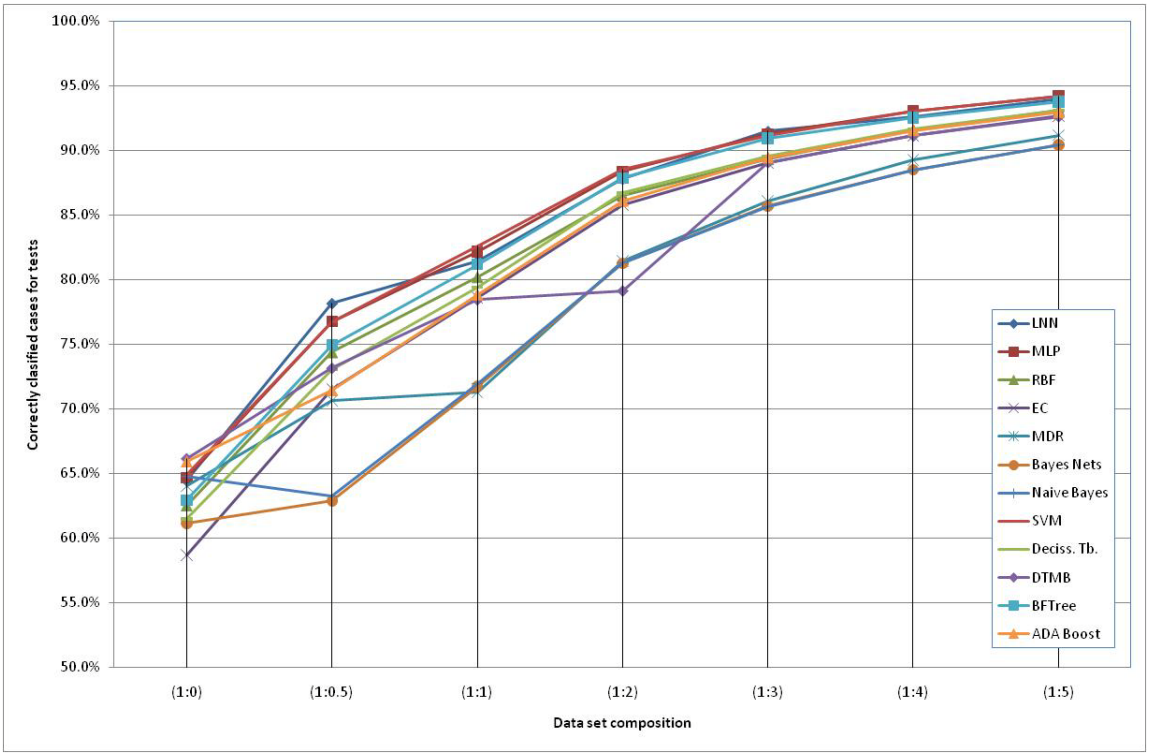
Data set	Gene	LNN	MLP	RBF	EC	MDR	Bayes Nets	Naïve Bayes	SVM	Decis. Tb.	DTNB	BFTree	AdaBoost
SNP (1:0)	**DRD3**	62.9%	59.5%	58.9%	56.6%	60.0%	62.5%	61.6%	64.8%	62.2%	59.5%	61.3%	63.4%
	HTR2A	62.4%	62.9%	63.7%	57.5%	64.0%	61.9%	66.6%	65.2%	61.0%	62.3%	62.8%	63.5%
	Both	64.5%	64.7%	62.5%	58.7%	64.0%	61.2%	64.8%	64.9%	61.5%	66.2%	62.9%	65.9%
SNP (1:0.5)	**DRD3**	74.6%	72.9%	71.5%	71.0%	60.5%	71.3%	71.0%	75.4%	73.5%	70.4%	73.7%	71.3%
	HTR2A	75.9%	75.5%	73.6%	71.7%	74.2%	62.2%	62.9%	77.4%	73.2%	70.9%	74.5%	71.4%
	Both	78.2%	76.8%	74.4%	71.5%	70.7%	62.9%	63.3%	76.8%	73.1%	73.2%	75.0%	71.4%
SNP (1:1)	**DRD3**	80.5%	79.5%	78.5%	78.2%	69.8%	77.9%	76.2%	81.4%	79.6%	77.1%	79.4%	78.6%
	HTR2A	80.7%	81.7%	80.2%	78.5%	71.0%	71.9%	72.3%	83.0%	79.8%	76.8%	81.2%	78.8%
	Both	81.4%	82.2%	80.2%	78.6%	71.3%	71.7%	72.0%	82.6%	79.4%	78.5%	81.2%	78.8%
SNP (1:2)	**DRD3**	87.0%	86.1%	85.8%	85.4%	79.4%	84.8%	83.2%	87.7%	86.6%	80.4%	86.1%	85.2%
	HTR2A	88.0%	88.1%	86.3%	85.9%	81.4%	81.3%	81.6%	88.8%	86.5%	76.2%	87.6%	86.1%
	Both	87.8%	88.4%	86.5%	85.8%	81.4%	81.3%	81.3%	88.5%	86.7%	79.2%	87.9%	86.1%
SNP (1:3)	**DRD3**	89.9%	89.5%	88.9%	88.4%	84.8%	89.4%	86.9%	90.6%	89.5%	87.6%	89.5%	88.7%
	HTR2A	90.4%	90.7%	89.3%	89.1%	85.9%	85.7%	85.9%	91.4%	89.7%	86.5%	90.3%	89.4%
	Both	91.5%	91.3%	89.3%	89.1%	86.1%	85.7%	85.6%	91.2%	89.5%	89.1%	90.9%	89.4%
SNP (1:4)	**DRD3**	91.9%	91.7%	91.3%	90.9%	87.4%	91.5%	89.2%	92.5%	91.6%	90.3%	91.5%	90.7%
	HTR2A	92.6%	92.7%	91.8%	91.2%	88.5%	88.6%	88.6%	93.2%	91.7%	88.5%	92.4%	91.5%
	Both	92.6%	93.0%	91.6%	91.2%	89.3%	88.5%	88.5%	93.0%	91.6%	91.1%	92.5%	91.5%
SNP (1:5)	**DRD3**	93.9%	93.1%	93.0%	92.1%	88.4%	92.9%	90.8%	93.6%	93.1%	91.8%	92.9%	92.2%
	HTR2A	93.2%	93.9%	92.9%	92.6%	91.2%	90.5%	90.5%	94.3%	93.1%	90.0%	93.5%	92.9%
	Both	93.9%	94.2%	93.1%	92.6%	91.2%	90.4%	90.4%	94.2%	93.1%	92.6%	93.8%	92.9%

Notes: LNN = Linear Neural Networks, MLP = Multilayer Perceptron; RBF = Radial Base Functions; EC = Evolutionary Computation; MDR = Multifactor Dimensionality Reduction; Bayes Nets = Bayesian Networks; SVM = Support Machine Vectors; Decis. Tb. = Decision Tables; DTNB = Decision Table Naïve Bayes Hybrid Classifier; BFTree = Best-First decision Tree classifier; AdaBoost = Adaptative Boosting.

The models generated using the original dataset correctly classify only 66.6% of the schizophrenic subjects when using the HTR2A gene and the Naïve Bayes method. This low accuracy can be due to the reduced number of subjects available and an increased number of "3" values of the SNPs (unknown data). Therefore, we included additional simulated subjects obtained with the HAP-SAMPLE software [15] in the negative group (*non-Schizo*), maintaining the capacity to evaluate positive subjects (cases) for the models. Thus, seven datasets have been created, labeled as SNP (1:*n*),

where 1:*n* (*n* = 0, 0.5, 1, 2, 3, 4, 5) is the proportion between the real subjects (positive and negative) and the simulated negative subjects (see details in the Experimental and Theoretical Section). The graphical representation of the evolution of the best classification depending on the additional number of simulated negative subjects is shown in Figure 2. It can observed that the classification percentages do not increase significantly after adding five parts of simulated negative subjects. Among the best models, we propose the following two QGDR models which correspond to simple linear artificial neural networks (LNN).

Figure 2. Correctly classified subjects depending on the simulated negative data for both genes; the dataset labels represent the proportion between real subjects (positive and negative = case and control) and simulated negative subjects.



The first model (Model 1) includes only a minimum number of simulated negative subjects, SNP (1:0.5): 260 real positive subjects, 354 real negative subjects and 307 simulated negative subjects for schizophrenia, a total of 921 subjects. It is based on 40 SNPs (at DRD3: rs7631540, rs6808291, rs1486012, rs9824856, rs2134655, rs963468, rs3773678, rs167771, rs226082, rs1486009, rs6280, rs7638876, rs9825563, rs1354348; at HTR2A: rs3889066, rs7329640, rs10507544, rs7333412, rs3125, rs6314, rs6308, rs1058576, rs1923884, rs2296972, rs9316233, rs659734, rs1928042, rs2770296, rs582385, rs1928040, rs731779, rs985934, rs9534505, rs6304, rs6305, rs2070036, rs6313, rs1328685, rs731244, rs10507547) and the model used was a LNN with 40 inputs and 152 neurons, which correctly classifies 78.2% of the subjects of the test group. The area under the receiver operating characteristic curve (AUC-ROC) for the cross-validation group (0.8405) shows that the model is not random (see Figure 3).

The second model (Model 2) includes a maximum number of simulated negative subjects, SNP (1:5): 260 real positive subjects, 354 real negative subjects and 3070 simulated negative subjects for

schizophrenia, a total of 3,684 subjects. The model is based only on two SNPs (rs7329640 and rs985934) at HTR2A: a LNN with two inputs and eight neurons, which correctly classifies 93.2% of the subjects of the test group. The AUC-ROC for the cross-validation group (0.9439) demonstrates the goodness of the model (see Figure 4).

Figure 3. Area under the receiver operating characteristic curve (AUC-ROC) for LNN 40:152-1:1 (Model 1).

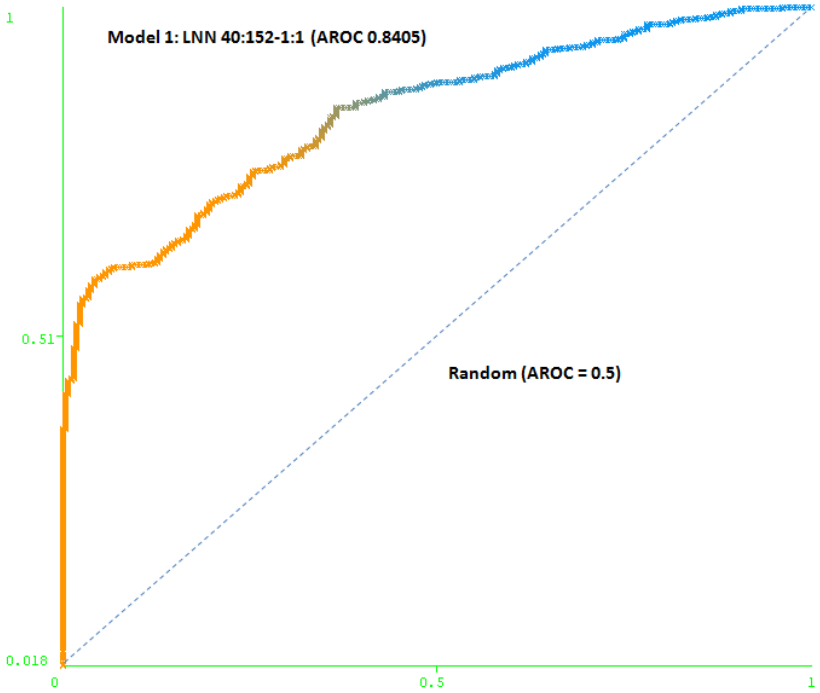


Figure 4. Area under the receiver operating characteristic curve (AUC-ROC) for LNN 2:8-1:1 (Model 2).

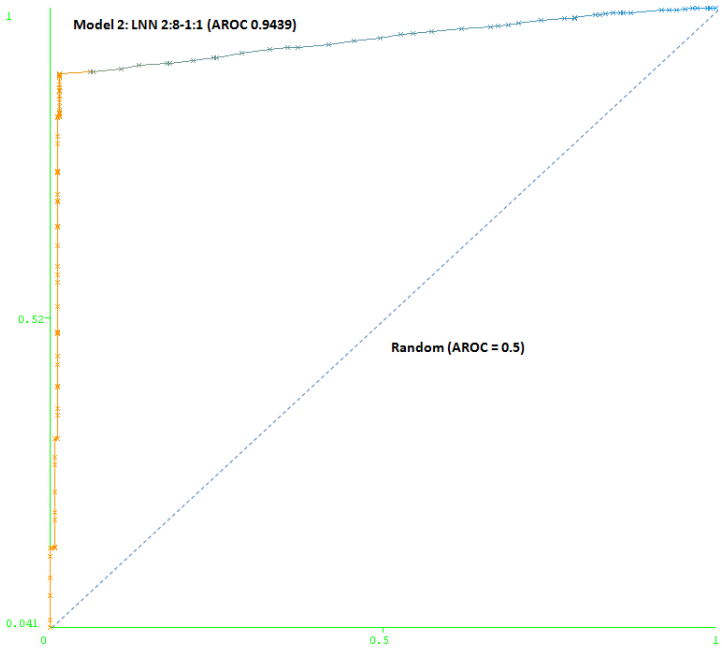


Table 1 shows that the classification accuracy percentages are in the range of 56.6–66.6% for SNP(1:0), 60.5–78.2% for SNP(1:0.5), 69.8–83.0% for SNP(1:1), 76.2–88.8% for SNP(1:2), 84.8–91.5% for SNP(1:3), 87.4–93.2% SNP(1:4) and 88.4–94.3% for SNP(1:5). In general, we can observe that the genotype information from the HTR2A gene is classifying more accurately than when considering the SNPs at DRD3 and using the Support Machine Vectors (SVM) technique [45]. There are two exceptions to this performance, with small differences, in the schizophrenia classification for SNP (1:0.5) and for SNP (1:3), where the maximum accuracy percentages correspond to LNN using information from both genes. Despite the fact that an MLP is more complex than an LNN, the first one obtains almost the same classification scores as the LNN. Finally, Evolutionary Computation (EC) [46] obtains better classification scores when the second gene or both genes together are considered, as a higher number of SNPs is taken into account and, thus, there is more information.

3. Experimental and Theoretical Section

3.1. Subjects and Genotyping

The case-control subjects consisted of 260 unrelated patients (65% males) being treated by the Galician Mental Health Service for schizophrenia and 354 unrelated blood negative donors (45% males) recruited from the Galician Blood Transfusion Centre (staff at the University of Santiago de Compostela and patients attending the University of Santiago de Compostela Hospital Complex). The study protocol was approved by the Bioethics Committee of the University of Santiago de Compostela (for details see [9]). In order to extract genomic DNA from white blood cells of peripheral venous blood from control and case subjects a standard protocol has been used. SNP genotyping was performed using the MassARRAY SNP genotyping system (Sequenom Inc., San Diego, CA, USA) [47]. Re-genotyping of random samples, which represented a total of 600 successfully replicated genotypes, revealed an accuracy rate of >99.9%.

3.2. Datasets

Seven datasets have been used containing the following 48 SNPs at the DRD3 and HTR2A genes associated with schizophrenia from the Galician population [9]: rs4682148, rs7631540, rs6808291, rs1486012, rs9824856, rs2134655, rs963468, rs3773678, rs167771, rs226082, rs10934256, rs1486009, rs6280, rs7638876, rs9825563, rs1354348, rs9283560 (17 SNPs at DRD3) and rs3889066, rs7329640, rs10507544, rs7333412, rs3125, rs6314, rs6308, rs1058576, rs6561333, rs1923884, rs2296972, rs9316233, rs659734, rs1928042, rs2770296, rs9316235, rs582385, rs1928040, rs731779, rs985934, rs9534505, rs6304, rs6305, rs2070036, rs2070037, rs6313, rs1328685, rs731244, rs1360020, rs10507546, rs10507547 (31 SNPs at HTR2A). SNPs can take different values: 0 if homozygous (both copies of a given gene have the same allele) for 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 unknown.

Additional negative subjects have been generated using the simulation tool named HAP-SAMPLE [15]. HAP-SAMPLE is a web application for simulating SNP genotypes for case-control and affected-child trio studies by re-sampling from Phase I/II HapMap SNP data. Providing a list of SNPs to be "genotyped," along with a disease model file that describes causal SNPs and their effect sizes, the

application returns two sets of simulated genotypes from case and control subjects. We discarded the case subjects. Thus, a file was created with a different number of control subjects, which were added to case subjects from real clinical data. This data was modified in order to introduce genotyping errors taking into account the error frequencies of the real data.

In addition to the original dataset that contains 260 positive subjects and 354 negative subjects SNP (1:0), we obtained six datasets by including 307, 614, 1,228, 1,842, 2,456 and 3,070 simulated negative subjects. The datasets were named: SNP (1:0.5), SNP (1:1), SNP (1:2), SNP (1:3), SNP (1:4) and SNP (1:5), where the label represents the proportion between the real subjects (positive and negative) and the simulated negative subjects.

3.3. QGDR models

The classification models have been obtained with the following methods: Linear Neural Networks, Multilayer Perceptron, Radial Base Functions, Bayesian Networks, Naïve Bayes, Support Machine Vectors, Decision Tables, Decision Table Naïve Bayes Hybrid Classifier, Best-First decision Tree classifier, Adaptative Boosting (all of them from Weka 3.6.2 [48]), Evolutionary Computation and Multifactor Dimensionality Reduction.

Artificial Neural Networks (ANN) have been extensively used for classification problems. More specifically, the simple Perceptron [49], also known as Linear Neural Network (LNN), has been utilized. This technique uses a linear network model, with no hidden layers, to perform classification. The Multilayer Perceptron (MLP) [50] has also been utilized. Other types of networks considered were Radial Base Functions (RBF) [51]. In this type of network, the neurons of the hidden layer perform a calculation function instead of the activation function of the MLP. The general scheme for an ANN with only one hidden layer is presented in Figure 5.

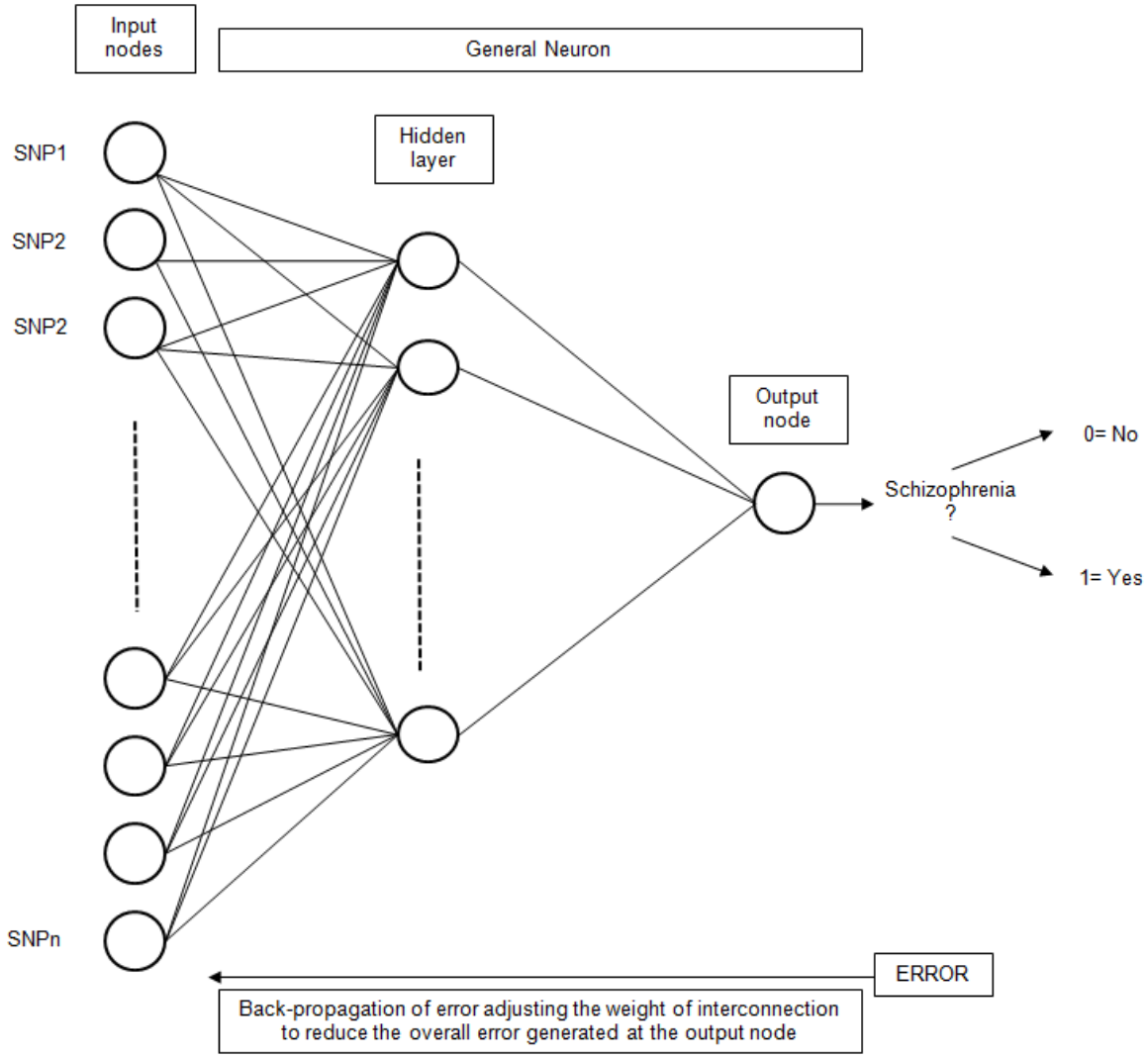
As well as the MLP, Support Machine Vectors (SVM) are nonlinear classifiers. SVM induce linear separators or hyperplanes in the space of characteristics. This type of classifier has proved to be very useful when dealing with high dimensionality problems [45].

Bayesian methods have also been applied to this problem. These 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 [52] (which assumes that the attributes are independent), and Bayesian Networks [53].

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 [54] have been used, 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 Naïve Bayes Hybrid Classifier (DTNB) DTNB [55]. A similar model was used to infer decision trees, following a hybrid approach between the decision trees and the Naïve Bayes classifier, called Best-First decision Tree classifier (BFTree) [56].

Finally, we tried a boosting meta-algorithm. This algorithm consists in combining multiple classification models that complement each other. The Adaptative Boosting (AdaBoost) [57] method builds the models iteratively, weighting the instances differently in each iteration. The new models classify the instances that the previous models do not classify correctly.

Figure 5. The general structure of an ANN for schizophrenia classification based on SNP inputs.



Multifactor Dimensionality Reduction (MDR) [58,59] is a data mining approach designed to detect and characterize nonlinear 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 nonparametric 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 [60-62] such as schizophrenia [63-65], although it can also be applied to other domains.

The technique of Evolutionary Computation (EC) [46] used in this paper is based on genetic algorithms (GAs) [66]. A GA is a search method based on Charles Darwin's Theory of Evolution [67]. Algorithms based on GAs make a population evolve through random actions similar to those existing in biological evolution (mutations and genetic recombination, as well as selections with a certain

criteria called fitness). The fitness is used to decide which individuals are selected, *i.e.*, the more suitable individuals are the higher likelihood they will reproduce. More specifically, the method considered here follows the Iterative Rule Learning (IRL) approach [68,69]. Thus, the result of this method is a set of rules which are used to classify the input data. Like MDR, this method tries to find relationships between attributes or variables and a binary outcome. It has mostly been applied to biomedical data; however, it is still in development.

For each classification, the data has been split into two groups: *Schizo* (positive/case subjects) and *non-Schizo* (negative/control subjects). The SNPs have categorical values of "0" if homozygous for the first allele, "1" if heterozygous, "2" if homozygous for the second allele "3" for unknown genotypes. The 10-fold cross-validation method [70-72] has been used to verify the accuracy of the models. The efficiency of the models that evaluate if a patient has schizophrenia is mainly due by the number of correct classifications when using the test set. In addition, these models have been constructed using the SNPs at only one of the two genes or at both of them. Therefore, the classification results have been obtained using 12 machine learning techniques and seven datasets that include different percentages of simulated negative subjects, that is, 252 classification models to be tested.

4. Conclusions

This work presents a disease computational study of schizophrenia based on DNA molecule information provided by SNPs and proposes for the first time, to the best knowledge of the authors, two classification models for schizophrenia evaluation. 252 classification models have been obtained using SNPs at two schizophrenia-related genes (each of them or both), twelve machine learning techniques and seven datasets. The best relationships between the DNA molecule sequence and schizophrenia evaluated 78.3–93.8% of the DNA sequence from schizophrenia patients, for datasets with extra simulated negative subjects. In future work, QGDR models will be extended to other types of complex diseases, such as colorectal cancer and cardiovascular diseases, and the best models will be implemented online for free access.

Acknowledgements

The work of Vanessa Aguiar-Pulido is supported by a grant from the General Directorate of Quality and Management of Galicia's University System of the Xunta. Munteanu C. R. and José A. Seoane acknowledge the funding support for a research position by "Isidro Parga Pondal" Program and an "Isabel Barreto" grant from Xunta de Galicia (Spain), respectively. This work is supported by the "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, by the "Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies", Ibero-NBIC Network (209RT0366) funded by CYTED (Ciencia y Tecnología para el Desarrollo) and by the COMBIOMED Network, the grant (Ref. PIO52048 and RD07/0067/0005), funded by the Carlos III Health Institute.

References

1. Chinchilla Moreno, A. *Las Esquizofrenias. Sus Hechos Y Valores Clínicos Y Terapéuticos*; Elsevier Masson: Barcelona, Spain, 2007.
2. Sham, P. Genetic epidemiology. *Br. Med. Bull.* **1996**, *52*, 408-433.
3. Sáiz, J.; Fañanás, L. Introducción: Genética y Psiquiatría. *Monogr. Psiquiatr.* **1998**, *10*, 1-3.
4. 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. *J. Pharmacol. Exp. Ther.* **1989**, *251*, 238-246.
5. Sokoloff, P.; Levesque, D.; Martres, M.P.; Lannfelt, L.; Diaz, G.; Pilon, C.; Schwartz, J.C. The dopamine D3 receptor as a key target for antipsychotics. *Clin. Neuropharmacol.* **1992**, *15*, 456A-457A.
6. Utsunomiya, K.; Shinkai, T.; De Luca, V.; Hwang, R.; Sakata, S.; Fukunaka, Y.; Chen, H.I.; Ohmori, O.; Nakamura, J. Genetic association between the dopamine D3 gene polymorphism (Ser9Gly) and schizophrenia in Japanese populations: evidence from a case-control study and meta-analysis. *Neurosci. Lett.* **2008**, *444*, 161-165.
7. Suzuki, M.; Hurd, Y.L.; Sokoloff, P.; Schwartz, J.C.; Sedvall, G. D3 dopamine receptor mRNA is widely expressed in the human brain. *Brain Res.* **1998**, *779*, 58-74.
8. Talkowski, M.E.; Mansour, H.; Chowdari, K.V.; Wood, J.; Butler, A.; Varma, P.G.; Prasad, S.; Semwal, P.; Bhatia, T.; Deshpande, S.; Devlin, B.; Thelma, B.K.; Nimgaonkar, V.L. Novel, replicated associations between dopamine D3 receptor gene polymorphisms and schizophrenia in two independent samples. *Biol. Psychiat.* **2006**, *60*, 570-577.
9. Dominguez, E.; Loza, M.I.; Padin, F.; Gesteira, A.; Paz, E.; Paramo, M.; Brenlla, J.; Pumar, 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. *Schizophr. Res.* **2007**, *90*, 123-129.
10. den Dunnen, J.T.; Antonarakis, S.E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **2000**, *15*, 7-12.
11. Katanforoush, A.; Sadeghi, M.; Pezeshk, H.; Elahi, E. Global haplotype partitioning for maximal associated SNP pairs. *BMC Bioinformatics* **2009**, *10*, 269.
12. Zhou, N.; Wang, L. Effective selection of informative SNPs and classification on the HapMap genotype data. *BMC Bioinformatics* **2007**, *8*, 484.
13. Kingman, J. F. Origins of the coalescent. 1974-1982. *Genetics* **2000**, *156*, 1461-1463.
14. Liang, L.; Zollner, S.; Abecasis, G.R. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics* **2007**, *23*, 1565-1567.
15. 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*, 2581-2588.
16. Balloux, F. EASYPOP (version 1.7): a computer program for population genetics simulations. *J. Hered.* **2001**, *92*, 301-302.
17. Hey, J. *FPG: A Computer Program for Forward Population Genetic Simulation*, 2004. http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#FPG/ (accessed on 5 May 2010).

18. 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*, 1725-1731.
19. Peng, B.; Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **2005**, *21*, 3686-3687.
20. Edwards, T.L.; Bush, W.S.; Turner, S.D.; Dudek, S.M.; Torstenson, E.S.; Schmidt, M.; Martin, E.; Ritchie, M.D. Generating Linkage Disequilibrium Patterns in Data Simulations using genomeSIMLA. *Lect. Notes Comput. Sci.* **2008**, *4973*, 24-35.
21. Li, J.; Chen, Y. Generating samples for association studies based on HapMap data. *BMC Bioinformatics* **2008**, *9*, 44.
22. 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 Genet.* **2010**, *11*, 26.
23. Saangyong, U; Dong-Hoi, K.; Young-Woong, K.; Sungwon, C; Jaeyoun, C.; Jin, K. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. *Expert Systems* **2009**, *26*, 60-69.
24. 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 identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. *Genes Immun.* **2010**, *11*, 199-208.
25. Nicodemus, K.K.; Callicott, J.H.; Higier, R.G.; 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 DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Hum. Genet.* **2010**, *127*, 441-452.
26. Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, The Netherlands, 1999.
27. Barabasi, A.L.; Bonabeau, E. Scale-free networks. *Sci. Am.* **2003**, *288*, 60-69.
28. 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. *Mol. Divers.* **2004**, *8*, 401-412.
29. Barabasi, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101-113.
30. Barabasi, A.L. Sociology. Network theory-the emergence of the creative enterprise. *Science* **2005**, *308*, 639-641.
31. González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. *Curr. Top. Med. Chem.* **2007**, *7*, 1025-1039.
32. Ferino, G.; Gonzalez-Diaz, 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. *Biochem. Biophys. Res. Commun.* **2008**, *372*, 320-325.
33. Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F.M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750-778.

34. Munteanu, C.R.; Magalhaes, A.L.; Uriarte, E.; Gonzalez-Diaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* **2009**, *257*, 303-311.
35. Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J. Theor. Biol.* **2009**, *261*, 449-458.
36. Vilar, S.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *J. Comput. Chem.* **2008**, *29*, 2613-2622.
37. González-Díaz, H.; Ferino, G.; Prado-Prado, F.J.; Vilar, S.; Uriarte Villares, E.; Pazos, A.; Munteanu, C.R. Protein Graphs in Cancer Prediction. In *An Omics Perspective on Cancer Research*; Cho, W.C.S., Ed.; Springer Netherlands: Amsterdam, The Netherlands, 2010; doi:10.1007/978-90-481-2675-0_7.
38. González-Díaz, H.; Ferino, G.; Podda, G.; Uriarte, E. Discriminating Prostate Cancer Patients from control group with connectivity indices. *ECSOC* **2008**, *12*, G1:1-G1:10.
39. Ferino, G.; Delogu, G.; Podda, G.; Uriarte, E.; González-Díaz, H. Quantitative Proteome-Disease Relationships (QPDRs) in Clinical Chemistry: Prediction of Prostate Cancer with Spectral Moments of PSA/MS Star Networks. In *Clinical Chemistry Research*; Mitchem, B.H., Sharnham, C.L. Eds.; Nova Science Publishers: New York, NY, USA, 2009.
40. Diederich, J. *Artificial Neural Networks: Concept Learning*; IEEE Press: Piscataway, NJ, USA, 1990; p. 141.
41. Byvatov, E.; Schneider, G. Support vector machine applications in bioinformatics. *Appl. Bioinformatics* **2003**, *2*, 67-77.
42. Eberbach, E. Toward a theory of evolutionary computation. *Biosystems* **2005**, *82*, 1-19.
43. Rowland, J.J. Model selection methodology in supervised learning with evolutionary computation. *Biosystems* **2003**, *72*, 187-196.
44. Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Pearson Addition Wesley: Boston, ML, USA, 2006.
45. Vapnik, V. *Statistical Learning Theory*; John Weily and Sons: New York, NY, USA, 1998.
46. Aguiar Pulido, V.; Seoane Fernández, J.A.; Freire, A.; Munteanu, C.R. Data Mining in Complex Diseases Using Evolutionary Computation. *Lect. Notes Comput. Sci.* **2009**, *5517*, 917-924.
47. Costas, J.; Torres, M.; Cristobo, I.; Phillips, C.; Carracedo, A. Relative efficiency of the linkage disequilibrium mapping approach in detecting candidate genes for schizophrenia in different European populations. *Genomics* **2005**, *86*, 280-286.
48. Waikato, T.U.O. Weka Machine Learning Project. http://www.cs.waikato.ac.nz/ml/weka/ (accessed on 5 May 2010).
49. Rosenblatt, F. *Principles of Neurodynamics; Perceptrons and The Theory of Brain Mechanisms*; Spartan Books: Washington, DC, USA, 1962.
50. Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, NY, USA, 1995.
51. Buhmann, M.D. *Radial Basis Functions: Theory and Implementations*; Cambridge University Press: Cambridge, UK, 2003.

52. John, G.H.; Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, August 18-20, 1995; Morgan Kaufmann: San Mateo, CA, USA, 1995; pp. 338-345.
53. Bouckaert, R.R. *Bayesian Networks in Weka*; Technical report, Computer Science Department. University of Waikato: Tauranga, New Zealand, 2004.
54. Kohavi, R. The Power of Decision Tables. In *Proceedings of 8th European Conference on Machine Learning*, Heraclion, Greece, April 25-27, 1995; Levrac, N., Wrobel, S., Eds.; Springer-Verlag Publisher: London, UK, 1995; pp.174-189.
55. Mark Hall, E.F. Combining Naive Bayes and Decision Tables. In *Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS)*, Coconut Grove, Florida, May 15–17, 2008; AAAI Press: Menlo Park, CA, USA, 2008.
56. Shi, H. Best-first Decision Tree Learning. MSc Thesis, University of Waikato, Hamilton, New Zealand, 2007.
57. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, Desenzano sul Garda, Italy, June 28 to July 1, 1996; Saitta, L., Ed., Morgan Kaufmann: San Francisco, CA, 1996; pp. 148-156,
58. 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. *J. Theor. Biol.* **2006**, *241*, 252-261.
59. Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **2009**, *10*, 392-404.
60. Greene, C.S.; Sinnott-Armstrong, N.A.; Himmelstein, D.S.; Park, P.J.; Moore, J.H.; Harris, B.T. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics* **2010**, *26*, 694-695.
61. 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.D.; Edwards, T.L.; Van Steen, K. FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. *PLoS One* **2010**, *5*, e10304.
62. 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. *Hum. Hered.* **2010**, *69*, 60-70.
63. 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*, 49-54.
64. 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.; Perez, L.; Leon, M.; de Fonseca, F. R.; Phillips, C.; Torres, M. Association of schizophrenia with DTNBP1 but not with DAO, DAOA, NRG1 and RGS4 nor their genetic interaction. *J. Psychiatr. Res.* **2008**, *42*, 278-288.

65. Yasuno, K.; Ando, S.; Misumi, S.; Makino, S.; Kulski, J.K.; Muratake, T.; Kaneko, N.; Amagane, H.; Someya, T.; Inoko, H.; Suga, H.; Kanemoto, K.; Tamiya, G. Synergistic association of mitochondrial uncoupling protein (UCP) genes with schizophrenia. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **2007**, *144B*, 250-253.
66. Holland, J.H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, USA, 1975.
67. Darwin, C. *On the Origin of Species by Means of Natural Selection*; John Murray: London, UK, 1859.
68. Venturini, G. SIA: A supervised inductive algorithm with genetic search for learning attributes based concepts. In *Proceedings of the 6th European Conference on Machine Learning*, Vienna, Austria, April 5-7, 1993; Brazdil, P., Ed.; Springer Verlag: Vienna, Austria, 1993; pp. 280-296.
69. González, A.; Herrera, F. Multi-stage genetic fuzzy systems based on the iterative rule learning approach. *Mathware Soft Comput.* **1997**, *4*, 233-249.
70. McLachlan, G.J.; Do, K.-A.; Ambroise, C. *Analyzing Microarray Gene Expression Data*. Wiley-Interscience: Hoboken, NJ, USA, 2004.
71. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, August 20-25, 1995; Morgan Kaufmann Publisher: San Francisco, CA, USA, 1995; Volume 2, pp. 1137-1143.
72. Picard, R.; Cook, D. Cross-Validation of Regression Models. *J. Amer. Statist. Assn.* **1984**, *79*, 575–583.