Prediction and Discrimination of Pharmacological Activity by Using Artificial Neural Networks

María José Castro[1], Wladimiro Díaz[2], and Juan Lucas Domínguez[1]

[1] Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46071 València, Spain
mcastro@dsic.upv.es, jldoru@hotmail.com

[2] Departament d'Informàtica
Universitat de València
Dr. Moliner, 50, 46100 Burjassot (València), Spain
Wladimiro.Diaz@uv.es

Abstract. The design of new medical drugs is a very complex process in which combinatorial chemistry techniques are used. For this reason, it is very useful to have tools to predict and to discriminate the pharmacological activity of a given molecular compound so that the laboratory experiments can be directed to those molecule groups in which there is a high probability of finding new compounds with the desired properties. A suitable set of topological indices that describe the molecular structure is used in this work. Two discrimination problems and two prediction problems are studied, using multilayer perceptrons to discriminate/predict. A large amount of different configurations are tested, yielding very good performances.

1 Introduction

The design of new medical drugs possessing desired chemical properties is a challenging problem in the pharmaceutical industry. The traditional approach for formulating new compounds requires the designer to test a very large number of molecular compounds, to select them in a blind way, and to look for the desired pharmacological property. Therefore, it is very useful to have tools to predict and to discriminate the pharmacological activity of a given molecular compound so that the laboratory experiments can be directed to those molecular groups in which there is a high probability of finding new compounds with the desired properties.

The tools that have been developed for this purpose are based on finding the relationship between a molecule's chemical structure and its properties. Given that the properties of a molecule come from its structure, the way the molecular structure is represented has special relevance. In this work, the molecular structure is described by a reduced set of 62 topological indices. This paper describes

a neural network based approach for solving the problem of activity prediction and discrimination based on the structural representation of the molecule.

Two discrimination problems and two prediction problems are studied, using multilayer perceptrons to discriminate/predict. A large amount of different configurations are tested, yielding to very good performances.

2 The Molecular Representation

The chosen set of molecular descriptors should adequately capture the phenomena underlying the properties of the compound. It is also important for these descriptors to be obtained without a lot of computational effort since they have to be computed for every molecule whose property needs to be predicted or discriminated.

The molecular topology is an alternative to the methods based on the "exact" description of the electronic attributes of a molecule calculated by mechanical-quantum methods. These molecular descriptors, which are based on graph theory, allow us to describe a molecule as a set of quantized numerical indices and it requires a lower calculation effort than other methods. They consider molecular structure as planar graphs where atoms are represented by vertices and chemical bonds are represented by edges. The topological indices have information about the number and kind of bonds that exist between the atoms as well as other structural attributes (size, branching factor, cycles, etc.) [1–3]. Searching for the set of indices which best adjust to this problem is a very complex task.

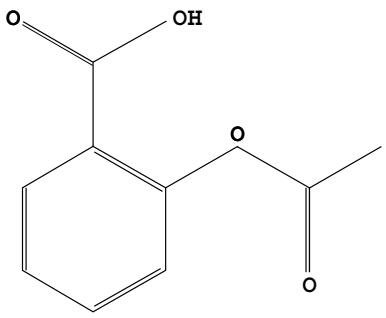
In this work, a set of 62 indices has been selected [4–6]. Fourteen of these indices are related to the molecular attributes of the compound; for example, the total number of atoms of a certain element (carbon, nitrogen, oxygen, sulphur, fluorine, chlorine, ...), the total number of bonds of a certain type (simple, double or triple), the number of atoms with a specific vertex degree, distance between the bonds, etc...

The remaining forty-eight topological indices include different topological information, such as the number of double bonds at distance 1 or 2, and the minimum distance between pairs of atoms, which are counted as the number of bonds between atoms. These indices are classified into six groups which are associated to the most frequent elements that constitute the molecules with pharmacological activity: nitrogen, oxygen, sulphur, fluorine, chlorine, bromine, and a general group in which the distances between pairs of atoms are considered without identifying the type of atom.

As an example, the set of topological indices of a chemical compound so well-known as the acetylsalicylic acid (*aspirin*) is shown in Figure 1.

3 Activity Discrimination and Prediction Problems

The case studies are of interest in the field of medicine. Two discrimination problems and two prediction problems were studied using the topological descriptors of the molecules explained above.



$\{\ 9, 0, 4, 0, 0, 0, 0, 8, 5, 0, 4, 5, 4, 0,$
$4, 2, 0, 0, 0, 0, 0, 0, 0, 0, 5, 10, 8, 11, 7, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,$
$0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 13, 17, 16, 15, 11, 6, 0, 0, 0, 0, 0, 0\ \}$

Fig. 1. ***Top.*** Molecular structure of the acetylsalicylic acid (*aspirin*). The hydrogen-suppressed graph is shown, in which every unlabeled vertex represents a carbon atom and every double edge represents a double bond. ***Bottom.*** The set of topological indices: 9 carbon atoms, 4 oxygen atoms, 8 simple bonds, 5 double bonds, 4/5/4 atoms with a vertex degree equal to one, two and three, respectively, 4 double bonds at distance one, 2 double bonds at distance two, 5/10/8/11/7/7 atoms with a distance of one/two/three/four/five/six from the oxygen atoms, 13/17/16/15/11/6 atoms with a distance of one/two/three/ four/five/six between them. (Null values are skipped.)

3.1 Activity Discrimination Problems

The properties studied were analgesic and antidiabetic discrimination. The objective was to train a classifier and evaluate it.

- *Analgesic discrimination problem.* The purpose of this experiment was to determine whether a molecule has analgesic activity or not. A dataset of 985 samples with potential pharmacological activity was used.
- *Antidiabetic discrimination problem.* In this case, we wanted to determine whether a molecule presents antidiabetic activity. A dataset of 343 samples was used.

3.2 Activity Prediction Problems

The properties considered were antibacterial activity and solubility. The objective of the case studies was to implement a predictor and evaluate its performance.

- *Antibacterial activity prediction problem.* We wanted to predict the minimum inhibitory concentration of antibacterial activity. A dataset of 111 samples was used.

Table 1. Datasets for the experimentation. For the activity discrimination problems, the active and inactive molecule percentages for each dataset are indicated in parenthesis.

		Number of samples		
	Problem	*Active*	*Inactive*	*Total*
Activity	Analgesic	172 (17.5%)	813 (82.5%)	985
discrimination	Antidiabetic	180 (52.5%)	163 (47.5%)	343
Activity	Antibacterial			111
prediction	Solubility			92

– *Solubility prediction problem.* In this case, we were interested in predicting the solubility capability of the molecules. A dataset of 92 samples was used.

The datasets for the four problems are shown in Table 1. In order to train the models, we split each dataset into training (75%) and test (25%) sets. The partitions were performed randomly (taking into account that the percentages of active and inactive molecules were homogeneous for the datasets of the activity discrimination problems).

4 Artificial Neural Networks for Structure-Activity Relationship Modeling

Classification of complex data has been addressed by various statistical and machine learning techniques. Although these methodologies have been successfully applied in a variety of domains, there are some classification tasks, particularly in medicine or chemistry, which require a more powerful, yet flexible and robust technique to cope with extra demands concerning limited datasets and complexity of interpretation. In this context, the use of artificial neural networks becomes an excellent alternative.

We used multilayer perceptrons (MLPs) for structure-activity discrimination and prediction. The number of input units was fixed by the number of topological descriptors of the molecules (62 topological indices). The input data of each dataset was discretized by dividing by the maximum value of all the indices.

There was only one output unit corresponding to the property being discriminated or predicted. The data for the activity discrimination problems were labeled with 1, -1, or 0: a value of 1 indicates that the molecule has pharmacological activity, a value of -1 indicates that the molecule is inactive, and a value of 0 indicates undetermined activity. Therefore, we use the hyperbolic tangent function, defined in the interval $[-1, 1]$, as the activation function.

The concentration and solubility levels for the activity prediction problems were discretized between 0 and 1, so we used the sigmoidal activation function.

The training of the MLPs was carried out using the neural net software package "SNNS: Sttutgart Neural Network Simulator" [7]. In order to successfully use

Table 2. MLP topologies and learning algorithms studied.

Topology:	One hidden layer:	2, 4, 8, 16, 32, 64
	Two hidden layers:	2-2, 4-4, 8-8, 16-16, 32-32, 64-64
Training algorithm:	Backpropagation without momentum term	
	Learning rate:	0.1 0.2 0.4 0.7 0.9 1.5 2.0
Training algorithm:	Backpropagation with momentum term	
	Learning rate:	0.1 0.2 0.4 0.7 0.9
	Momentum term:	0.1 0.2 0.4 0.7 0.9
Training algorithm:	Quickprop	
	Learning rate:	0.1 0.2 0.3
	Quick rate:	1.75 2 2.25

neural networks, a number of considerations has to be taken into account, such as the network topology, the training algorithm, and the selection of the algorithm's parameters [7–9]. Tests were conducted using different network topologies: a hidden layer with 2, 4, 8, 16, 32 and 64 units or two hidden layers with an equal number of hidden units (2, 4, 8, 16, 32 or 64). Several learning algorithms were also studied: the incremental version of the backpropagation algorithm (with and without momentum term) and the quickprop algorithm. Different combinations of learning rate (LR) and momentum term (MT) as well as different values of the maximum growth parameter (MG) for the quickprop algorithm were proved (see Table 2). In every case, a validation criterion (25% of the training data was randomly selected for validation) was used to stop the learning process and to select the best configuration.

5 Experimental Results

5.1 Activity Discrimination Problems Experiments

The output values of the MLPs are between -1 and 1 (due to the hyperbolic tangent activation function). In the learning stage, -1 is assigned to the molecule that does not have pharmacological activity (analgesic or antidiabetic) and 1 to the molecule that do have it. After training the MLP models for the activity discrimination problems, the classification criterion was the following: if the molecule is *inactive* and

- the output achieved with the MLP is in the interval $[-1, -0.5]$, it is counted as correct;
- if the output is in the interval $]-0.5, 0[$ the result is counted as undetermined;
- finally, if the output is in the interval $[0, 1]$, it is an error.

When testing an *active* molecule the classification criterion was similar:

- it is considered to be correctly classified when the output value of the MLP is between 1 and 0.5;

Table 3. MLP performance (in %) for the activity discrimination problems.

Discrimination	*Configuration*	*Success*		
problems	*MLP topology, algorithm and parameters*	*Active*	*Inactive*	*Total*
Analgesic	{62−16−1}			
	Backpropagation (LR=0.1)	61.90%	91.67%	86.59%
Antidiabetic	{62−4−4−1}			
	Backpropagation (LR=0.2, MT=0.1)	93.33%	95.12%	94.19%

- if the output is in the interval $]0.5, 0[$, it is counted as undetermined;
- if the output is between 0 and -1, it is considered an error.

In the experimentation with potential analgesic activity, the best performance on the validation data was achieved using an MLP of one hidden layer of 16 units, trained with the standard backpropagation algorithm with a learning rate equal to 0.1. We then tested this trained MLP with the test data (not yet used so far), obtaining an overall success percentage of 86.59%, with no sample classified as undetermined. If we analyze these results considering the group (active or inactive), we get a success percentage of 61.90% in the active group and a success percentage of 91.67% in the inactive group.

For the antidiabetic activity discrimination problem, we reached the best performance on the validation data with an MLP of two hidden layers of 4 units each, trained with the backpropagation algorithm (LR=0.2 and MT=0.1), achieving a percentage of classification equal to 94.19% on the test data. If we analyze the results considering the active and inactive groups we get a success percentage of 93.33% and 95.12%, respectively. The performance of the discrimination experiments are shown in Table 3.

5.2 Activity Prediction Problems Experiments

Structure-activity prediction was achieved with high accuracy. For the antibacterial prediction problem, of all the networks tested, the most suitable one (on the validation data) turned out to be an MLP of one hidden layer of 64 units, trained with the standard backpropagation algorithm, using a learning rate equal to 0.1. This network was capable of predicting the minimum inhibitory concentration of antibacterial activity with a mean square error lower than 0.05 on unseen data, the test dataset (see Table 4).

The best performance on the validation data for the solubility prediction problem was achieved using an MLP of one hidden layer of 32 units, trained with the standard backpropagation algorithm with a learning rate equal to 0.1. This MLP could predict the solubility capacity of a molecule with a mean square error of 0.02 on test data.

Table 4. MLP mean square error (MSE) for the prediction problems.

Prediction problems	*Configuration* *MLP topology, algorithm and parameters*	*MSE*
Antibacterial	{62−64−1}	
	Backpropagation (LR=0.1)	0.05
Solubility	{62−32−1}	
	Backpropagation (LR=0.1)	0.02

6 Conclusions and Future Work

In this work, the viability of the use of artificial neural networks for structure-activity discrimination and prediction have been shown based on the structural representation of the molecules. Two discrimination problems and two prediction problems were studied, using multilayer perceptrons to discriminate and predict different properties of the molecular compounds.

The experiments performed with the analgesic group allow to determine whether a given molecule is active or inactive with a classification percentage of 86.59%. Better results were obtained with the antidiabetic group, with a success classification rate of 94.19%.

On the other hand, structure-activity prediction was achieved with high accuracy: antibacterial activity can be predicted with a mean square error of 0.05; the solubility capacity of a molecule can be predicted with a 0.02 mean square error.

Before ending we would like to remark that this work is only the first step towards an automatic methodology for designing new medical drugs. Thus, the following step will be the inverse problem of constructing a molecular structure given a set of desired properties [10].

Acknowledgements

We are grateful to Dr. Mr. Facundo Pérez and Dra. Ms. María Teresa Salabert, from the Chemistry and Physics Department of the Pharmacy Faculty of the Universitat de València, for their help in supplying the datasets and specially for their supervision in getting the topological indices and the elaboration of the samples used in the experimentation. The authors also wish to thank Cristina Adobes Martín for the software developed to calculate the topological indices.

Financial support for this work was provided by TIC2000-1153 of Spanish government.

References

1. A. T. Balaban, editor. *Chemical Applications of Graph Theory.* Academic Press, 1976.

2. N. Trinajstic. *Chemical Graph Theory*. CRC Press, Boca Raton, FL, 1976.
3. L. B. Kier and L. H. Hall. *Molecular Connectivity in Structure-Activity Analysis*. John Willey and Sons, New York, 1986.
4. Cristina Adobes. Diseño e implementación de herramientas para la predicción de propiedades moleculares. Master's thesis, Facultat de Informàtica, Universitat de València, 2000. (In Spanish.).
5. Wladimiro Díaz, María José Castro, Cristina Adobes, Facundo Pérez, and María Teresa Salabert. Discriminación de la actividad farmacológica utilizando técnicas conexionistas. In *Actas de la IX Conferencia de la Asociación Española para la Inteligencia Artificial*, volume I, pages 233–241, Gijón (Spain), November 2001.
6. Juan L. Domínguez, María José Castro, and Wladimiro Díaz. Discriminación y predicción de propiedades de fármacos mediante redes neuronales. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 2002.
7. A. Zell et al. *SNNS: Stuttgart Neural Network Simulator. User Manual, Version 4.2*. Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Germany, 1998.
8. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *PDP: Computational models of cognition and perception, I*, pages 319–362. MIT Press, 1986.
9. C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.
10. Venkat Venkatasubramanian, King Chan, and James M. Caruthers. Computer-Aided Molecular Design Using Neural Networks and Genetic Algorithms. *Computers and Chemical Engineering*, 18(9):833–844, 1994.