Modeling anti-*Trypanosoma cruzi* Activity of *N*-Oxide Containing Heterocycles

Mariana Boiani,[†,‡] Hugo Cerecetto,[†] Mercedes González,*[,†] and Johann Gasteiger[‡]

Laboratorio de Química Orgánica, DQO, Facultad de Ciencias-Facultad de Química, Universidad de la
República, Iguá 4225, Montevideo 11400, Uruguay, and Computer-Chemie-Centrum, Universität
Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91042 Erlangen, Germany

Received July 27, 2007

In the present study a systematic approach was used to model the anti-*T. cruzi* activity of a series of *N*-oxide containing heterocycles belonging to four chemical families with a wide structural diversity. The proposed mode of action implies the reduction of the *N*-oxide moiety; however, the biochemical mechanism underlying the anti-*T. cruzi* activity is still unkown. For structural representation two types of descriptors were analyzed: quantum chemical (AM1) global descriptors and properties coded by radial distribution function (RDF). Both types of descriptors point to the relevance of electronic properties. The local-RDF (LRDF) identified an electrophilic center at 4.1−4.9 Å from the oxygen atom of the *N*-oxide moiety, although other properties are required to explain the biological activity. While the mode of action of *N*-oxide containing heterocycles is still unknown, the results obtained here strengthen the importance of the electrophilic character of the molecule and the possible participation of the heterocycle in a reduction process. The ability of these descriptors to distinguish among activity classes was assessed using Kohonen neural networks, and the best clustering descriptors were later used for model building. Different learning algorithms were used for model development, and stratified 10-fold cross-validation was used to evaluate the performance of each classifier. The best results were obtained using *k*-nearest neighbors (*k*-NN) and decision tree (J48) methods combined with global descriptors. Since tree-based methods are easily translated into classification rules, the J48 model is a useful tool in the de novo construction of new *N*-oxide containing heterocycle lead structures.

INTRODUCTION

Chagas' disease is the major endemic disease in South and Central America caused by a trypanosomatid parasite (*Trypanosoma cruzi*).[1] The current treatment relies on two chemotherapeutic agents: Nifurtimox (4-(5-nitrofurfurylindenamino)-3-methylthiomorpholine 1,1-dioxide, Lampit, recently discontinued by Bayer) a nitrofuran derivative and Benznidazole (*N*-benzyl-2-(2-nitro-1*H*-imidazol-1-yl)acetamide, Rochagan, Roche) a nitroimidazole derivative, introduced empirically in the market in 1970.[2] Despite the great effort that has been done into the discovery of unique targets that afford selectivity, the drugs used today have serious side effects. What is more, differences in drug susceptibility among different *T. cruzi* strains lead to varied parasitological cure rates according to geographical area. There is, therefore, an urgent need for the development of new antichagasic drugs.

Our group described for the first time *N*-oxide containing heterocycle derivatives as anti-*T. cruzi* agents,[3−5] representing different skeletal types: benzofuroxans,[6−8] benzimidazole *N*-oxide,[9,10] indazole *N*-oxide,[11] and quinoxaline *N*-oxide.[12] Among these families of compounds a number of promising prototypes for the development of antichagasic drugs have been found and have been subjected to further studies (mammal cytotoxicity, in vivo activity). While the structural features determining the ability of these derivatives to inhibit the growth of *T. cruzi* are not quite clear, a bioreduction process is presumed to be involved. This is supported by the complete loss of biological activity that is verified after reduction of the *N*-oxide moiety[3,6,8] and the correlation between anti-*T. cruzi* activity and the reduction potential observed for some derivatives.[12] However, a more comprehensive picture of the biological activity of *N*-oxide derivatives is still missing. While some of the families have been studied using 2D and 3D QSAR approaches,[13] for others only preliminary SAR have been developed, and so far no attempt to include all the structural families into a single model has been conducted. Besides, a more complete survey of descriptors and model building methods would be desirable.

Drug action is closely related to chemical structure, hence rational drug design requires a knowledge of structure−activity relationships.[14] The application of mathematical models to display biological activity is manifold. On the one hand, it could be used in the design of new compounds in a time and cost-effective manner. On the other hand, it provides information regarding the possible mode of action. A key subject in the generation of QSAR models is choosing a proper description of chemical structure, which can change dramatically the quality of the obtained models. Nowadays, there are several descriptors available,[15] and it is necessary to adopt a protocol to identify information-rich descriptors corresponding to the phenomenon under investigation. Unsupervised methodologies are better suited than supervised ones for this task, and as examples we can mention principal component analysis (PCA) and Kohonen neural networks.[16,17] Since experience shows that no single machine learning

* Corresponding author phone/fax: (5982)5258618-int 216/5250749; e-mail: megonzal@fq.edu.uy.

[†] Universidad de la República.

[‡] Universität Erlangen-Nürnberg.

10.1021/ci7002768 CCC: $40.75 © 2008 American Chemical Society
Published on Web 12/29/2007

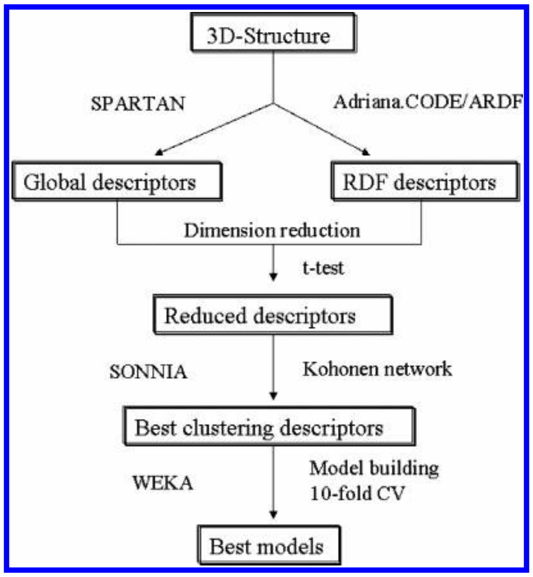

Figure 1. Flow chart showing the strategy used to model anti-*T. cruzi* activity.

Table 1. List of Descriptors Analyzed

symbol	property	dimensionality after reduction
global	μ, SM5, LogP, E_{LUMO}, GAP, HDon, TPSA	7
RDF$\chi\pi$	π-electronegativity[a]	31
RDFqπ	π-charge[a]	32
RDFqσ	σ-charge[a]	13
LRDF$\chi\pi$	π-electronegativity[b]	17
LRDFqπ	π-charge[b]	6
LRDFqσ	σ-charge[b]	32

[a] RDF encoded. [b] Local RDF encoded.

method is appropriate for all possible learning problems, the inclusion of different learning algorithms in the model building process is recommended.

The main goal of the present study is to develop a comprehensive model of anti-*T. cruzi* activity of *N*-oxide containing heterocycles, which will be used in the design of novel structures with desired activities and will help in understanding the mechanism implicated in the biological activity. In order to do this the following methodology was employed (Figure 1). First, two structure representations were generated, and proper descriptors were selected based on the clustering observed using Kohonen neural networks. Second, we studied the structure−activity relationship using PCA and frequency histograms to pursue a common pharmacophoric pattern. Finally, the best clustering descriptors were used in combination with different learning algorithms to develop the models, and stratified 10-fold cross-validation was used to assess performance.

DATA SET

The template structures of the compounds included in this study are shown in Figure 2. A total of 179 compounds from our in-house database belonging to four chemical families with a wide structural diversity were used: benzimidazole *N*-oxides (42 compounds), benzofuroxans (59 compounds), indazole *N*-oxides (12 compounds), and quinoxaline di-*N*-oxides (66 compounds). All the compounds were assayed in vitro against the epimastigote form Tulahuen 2 strain of *Trypanosoma cruzi* using the same protocol.[3] As a first screening procedure the antiproliferative activity of the compounds is determined at a single dose (25 μM), and the biological activity is expressed as percentage of growth inhibition (PGI) (Table 1S, Supporting Information). As an internal criterion compounds presenting a PGI value equal to or greater than 60% ("hits") are considered for further studies. Initially, this same criterion was used to classify the compounds into two activity classes: inactive (PGI < 60%, 117 compounds) and active (PGI ≥ 60%, 62 compounds). Using this categorization 66% of the compounds belong to the inactive class. Second, the inactive class was further subdivided into two classes, considering an intermediate class: inactive (PGI < 20%, 78 compounds), active (59% ≥ PGI ≥ 20%, 39 compounds), and highly active (PGI ≥ 60%, 62 compounds). The latter categorization leads to a more balanced distribution of the compounds into activity classes: highly actives (34%), actives (22%), and inactives (44%).

METHODS

Structure Representation. The first step in the development of QSAR models is the selection of proper descriptors of the chemical structures. In the present study two types of descriptors were analyzed, quantum chemical descriptors and radial distribution function (RDF) descriptors, both relying on 3D-structures (Table 1). From our previous experience

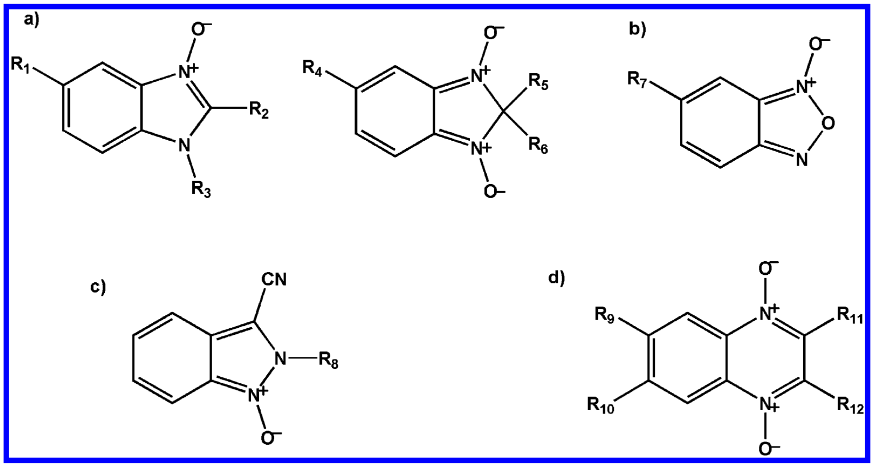

Figure 2. General structures of anti-*T. cruzi* agents from the data set: (a) benzimidazole *N*-oxide derivatives (Bz); (b) benzofuroxan derivatives (Bfx); (c) indazole *N*-oxide derivatives (In); and (d) quinoxaline di-*N*-oxide derivatives (Qx). R_1-R_{12} variable fragments.

quantum-chemical descriptors from AM1 calculations provide a good description of structure−activity relationships of *N*-oxide containing heterocycles, and they were the first descriptors analyzed.[8,10,11] The molecular modeling studies were carried out using the Spartan'04 software package.[18] The Merck molecular force field (MMFF) was used for preliminary structure optimization and conformational search using either systematic search (few degrees of freedom) or Monte Carlo methods (more complicated systems) as implemented in the program package.[19] Due to the rigid nature of the heterocyclic systems only one conformer was observed. After that, a reoptimization step was applied using a semiempirical quantum chemical method (AM1).[20] For this final structure the following properties were calculated: molecular volume (V), molecular area (A), dipolar moment (μ), energy of the molecule in aqueous phase (E_{aq}), solvation energy using Truhlar model SM5.4 (SM5),[21] octanol/water partition coefficient using Ghose-Crippen method (LogP),[22] highest occupied molecular orbital (HOMO) energy (E_{HOMO}), lowest unoccupied molecular orbital (LUMO) energy (E_{LUMO}), HOMO−LUMO energy gap (GAP), molecular electronegativity (χ), and molecular hardness (η). Using the program ADRIANA.code the number of hydrogen bond donors (HDon), the number of hydrogen bond acceptors (HAcc), and the topological polar surface area (TPSA)[23] were determined for all the compounds. These descriptors along with quantum chemical descriptors are termed Global descriptors.

A second type of descriptor used in the study was radial distribution function (RDF) descriptors. While quantum-chemical descriptors provide a global picture of the molecule, RDF descriptors express pharmacophore features as a 3D arrangement of atomic properties. The radial distribution function is a transformation of the 3D molecular structure into a vector of descriptors allowing the consideration of different physicochemical atomic properties. The 3D structures were obtained by the program CORINA as implemented in the program ADRIANA.Code.[24,25] Radial distribution functions (RDF) were used for the codification of atom properties (pi electronegativity ($\chi\pi$), pi charge (qπ), and sigma charge (qσ)) using the program ADRIANA-.Code.[25] The following equation gives the basis of this code:

$$g(r) = \sum_{i=1}^{N-1} \sum_{j>i}^{N} p_i p_j e^{-B(r-r_{ij})^2}$$

Here, p_i and p_j are chosen atomic properties, N is the number of atoms in a molecule, and r_{ij} is the distance between the atoms i and j. B is the temperature factor that determines the accuracy of position of atoms. The parameter B was set to 100 Å^{-2} corresponding to a total resolution of 0.1 Å. r is the running variable and is made discrete to arrive at a fixed-length representation of a molecule by $g(r)$.[26,27]

Furthermore, an RDF-like function was used by selecting a specific atom as focus (the assumed reaction site) and migrating through the 3D molecular structure with this atom in the focus. This representation is called local RDF (LRDF) and is calculated with the program ARDF using the following equation:[28]

$$f(r) = p_c \sum_{i=1,j\neq c}^{N} p_i e^{-B(r-r_{i,c})^2}$$

Similar to the RDF code, the LRDF can be interpreted as a probability distribution to find an atom in a spherical shell around one (the central) atom c. The oxygen of the *N*-oxide moiety was used as the central atom. In the case of di-*N*-oxide derivatives the following criteria was applied: for quinoxaline 1,4-dioxide derivatives the *N*-oxide moiety displaying the most positive nitrogen charge (Mullikens' charge) was chosen; for benzimidazole 1,3-dioxide derivatives, which show no clear difference on nitrogen charge, the *N*-oxide nitrogen 3 was chosen. This atom was selected because it participates in the bioreduction process that is believed to be responsible for the anti-*T. cruzi* activity.[3,4,11,13]

All the RDF and LRDF representations are uniform and invariant under translations and rotations of molecules. The functions are given as vectors calculated with a sampling rate of 0.1 Å. The dimension was set to 118, and $g(r)$ and $f(r)$ were defined in the interval 1.0−12.8 Å.

Preprocessing. The initial number of descriptors was submitted to the following reduction procedure: (1) Descriptors with constant values for all molecules were excluded. (2) Pairwise correlation analysis was then done, and a descriptor was eliminated if the correlation coefficient with another descriptor was equal to or higher than 0.90. (3) Using a statistical *t*-test only descriptors able to discriminate between activity classes were selected (*t*-value >2). All the reduced descriptors were autoscaled using the mean and standard deviation.[29]

Evaluation of the Descriptors. The ability of the descriptors to discriminate between classes was analyzed using Kohonen networks. Kohonen networks are used to obtain two-dimensional maps from higher dimensional data, the general idea is that similar compounds in the chemical space are clustered together in the neuronal space.[29] The software SONNIA (Self-Organizing Neural Network for Information Analysis)[30] was used to generate the Kohonen networks. These networks consist of a *n*-dimensional input layer, where *n* is the number of descriptors used (global, RDF, or LRDF). The training starts with the presentation of a vector (molecule) of input variables to all neurons.[16,31] The neuron that has weights being closest to the input variables is selected as the winning neuron in the learning algorithm. To improve the response to the input in the next epoch, the weights of the so-called winning neuron are adjusted to the input vector as well as the weights of neurons in the neighborhood. The degree of adaptation decreases with increasing distance to the winning neuron. This adaptation is repeated for each vector of an input molecule. After training, the response of the network is calculated for each vector of the data set. Subsequently, the projection of the data set into the two-dimensional space is performed by mapping the activity of each vector into the coordinates of its winning neuron. Different topologies (rectangular, toroidal) as well as different network sizes (X × Y) were analyzed. The initial learning spans were 1/2 X and 1/2 Y, with and initial learning rate of 0.7 and a rate factor of 0.9. The initial weights were randomly initialized, and training was performed until the network stabilized. The resulting output maps were evaluated according to their cluster ability, occupancy, and number of conflicts.

Model Building. The statistical software package WEKA[32] was used for the generation of models. Four methods were tested: *k*-nearest neighbors (*k*-NN),[33] simple logistic regression (Slog),[34−36] decision trees (J48), and multilayer perceptron (MLP).[16] All the methods were used as implemented in WEKA. Initially, all models were built using default parameters (*k* (number of nearest neighbors), *M* (minimum number of instances per leaf), and *a* (number of neurons in the hidden layer)), and later a search for optimum parameters was conducted ($1 < k < 20$, $2 < M < 20$, $2 < a < 12$). All the compounds were used to build the models. Models were generated for each data set without variable selection or in combination with the variable-selection filter implemented in WEKA (attribute evaluator combined with either the BestFirst or the ExahustiveSearch method). Despite decision tree (J48) and simple logistic regression (Slog) implement automatic variable selection, the inclusion of variable selection prior to model building was also studied. In general, its inclusion improves the statistics except for multilayer perceptron (MLP) models, which could imply that this method is suffering from overfitting. Then, combinations of the data sets using variable selection were analyzed. To assess the performance of the models obtained, the percentage of correctly classified molecules (%Correct class.) from 10-fold cross-validation was used, averaged over 10 entire repeats of the cross-validation using different random seeds. Using the information from the confusion matrix, averaged over 10 repeats, the Matthews correlation coefficient (C_{Matthews}) was calculated using the following formula[37,38]

$$C_{\mathrm{Matthews}} = \frac{(\mathrm{PN}) - (\mathrm{N^fP^f})}{\sqrt{(\mathrm{N} + \mathrm{N^f})(\mathrm{N} + \mathrm{P^f})(\mathrm{P} + \mathrm{N^f})(\mathrm{P} + \mathrm{P^f})}}$$

where P represents true positive, N true negative, Pf false positive, and $\mathrm{N^f}$ false negative.

This coefficient returns a value between −1 and 1. The higher the value of the C_{Matthews}, the more reliable is the classification result. It is difficult to find a single number to describe performance properly, but C_{Matthews} has been proposed as a more reliable descriptor than %Correct class.[39] The latter may favor classifiers giving more false positives when the negative class is much larger than the positive one. From the confusion matrix the following per class statistic were derived: (1) recall (sensitivity), which is obtained by dividing the number of correctly classified compounds of a given class by the total number of compounds in that class; (2) precision, which is obtained by dividing the number of correctly classified compounds of a given class by the total number of compounds which were predicted to belong to this class; and (3) F-measure, which is calculated as 2 *recall* × *precision/(recall* + *precision*) and combines recall and precision into a single efficiency measure (it is the harmonic mean of the precision and recall).

RESULTS AND DISCUSSION

Analysis of Descriptors. Overall, a better clustering is observed when a 2-class categorization is used instead of three classes. Three descriptors show a clustering with approximately 50 conflicts (2 classes) and occupancy of more than 75%: π-electronegativity and σ-charge encoded by global RDF and global descriptors (Table 2). The latter presents a lower number of conflicts, but it should be borne in mind that more than one property is included. Most of the descriptors were able to discriminate among structural families, even when no clustering of activity classes was observed. Only σ-charge shows a similar clustering of activity classes and family. This difference could be due to the distances included. While electronegativity spans from 1.2 to 10 Å, only distances above 6.1 Å were relevant for σ-charge pointing to atoms outside the heterocyclic skeleton.

Table 2. Best Clustering Descriptors Using Kohonen Networks[a]

	conflicts[b]			
	3 classes	2 classes	family[c]	occupancy[d] (%)
global	53	29	2	79
RDF$\chi\pi$	70	52	9	78
RDFqσ	73	49	57	81

[a] A 17 × 7 network with toroidal topology was used. [b] Number of compounds in conflictive neurons (neurons holding compounds from different classes). [c] Coded by chemical family (indazole, benzimidazole, benzofuroxan, quinoxaline). [d] Percentage of neurons occupied.

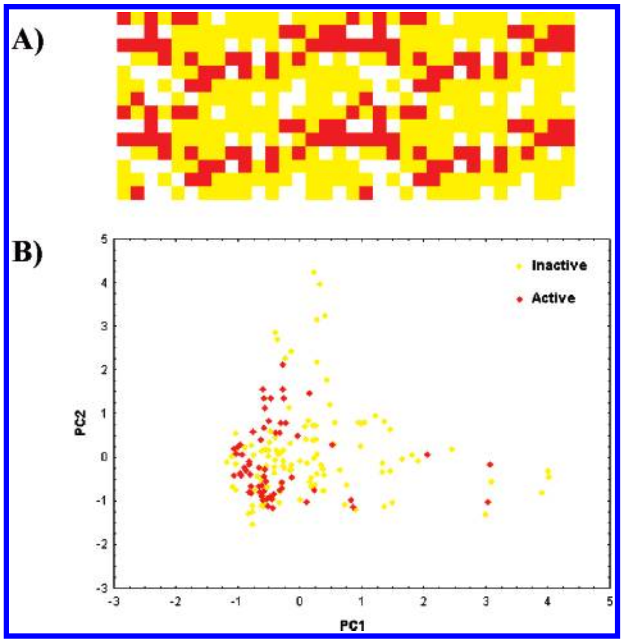

Figure 3. (A) Map of most frequent output coded by activity class (active = red, inactive = yellow). To illustrate the clustering of the different classes, four toroidal maps were arranged like tiles to indicate the closed nature of a toroidal surface. (B) Plot of the compounds in the plane generated by the first two components coded by activity class (active = red, inactive = yellow).

Another unsupervised method used to analyze the descriptors was principal components analysis (PCA).[40] While all the variance observed in the descriptor is placed in the Kohonen map, in PCA this variance is split into components. In Figure 3 the plots of the first two components (Figure 3B) and the Kohonen map (Figure 3A) for global descriptors are compared. Clearly, the grouping of compounds belonging to the same activity class is better visualized in the Kohonen map, which implies that more than two components are required for a correct description of the biological activity. The first two components, that explain 57% of the variance, are related to the polar character of the molecule (SM5, TPSA, HDon, μ). The LUMO energy is found in the third component with an important contribution (18%), and LogP and GAP are represented in the fourth and fifth components,

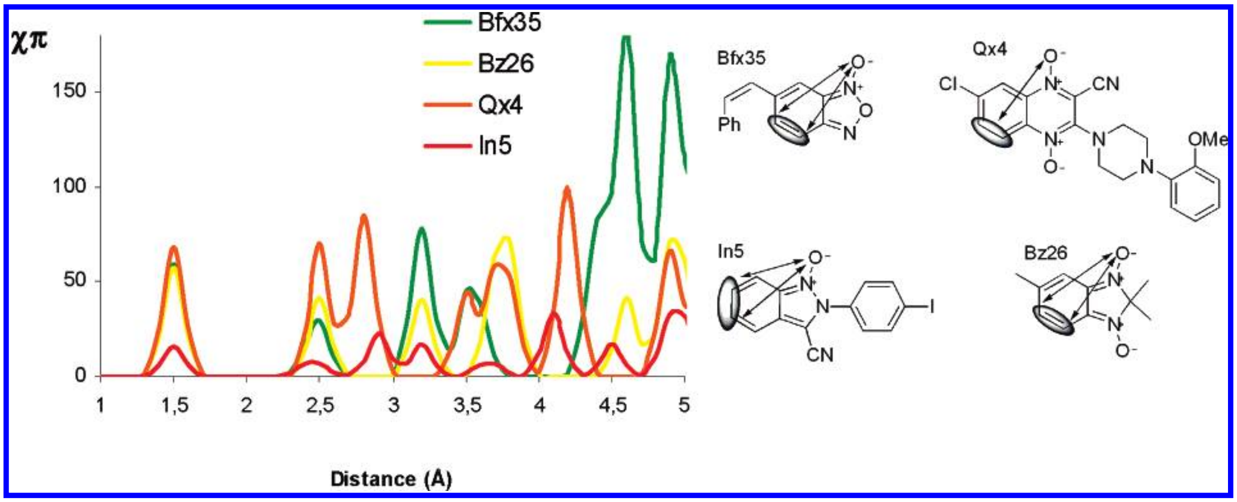

Figure 4. Property weighted ($\chi\pi$) LRDF codes of four *N*-oxide derivatives belonging to the active class. Only distances including heterocyclic atoms are shown. The structure of the four compounds is displayed marking the electrophilic center at 4.1−4.9 Å from the oxygen atom of the *N*-oxide moiety.

respectively. By inspection of frequency histograms for both activity classes the following conclusion could be drawn. Active compounds are less polar (lower dipolar moment, lower SM5, lower TPSA, no HDon moieties, and higher logP) and softer electrophiles (lower GAP, lower LUMO energy) than inactive compounds. This high electrophilic character could be related to the participation of these derivatives in a bioreduction process or in a reaction with a biological nucleophile.

In search of a common pharmacophoric pattern, π-electronegativity ($\chi\pi$) and σ-charge (qσ) were further analyzed using LRDF within each family. In contrast to the rest of the families, when the values of $\chi\pi$ and qσ for the atoms of the benzofuroxan skeleton are analyzed (*t*-test), no difference between activity classes is observed. A similar trend is observed for global descriptors related to electronic properties (E_{LUMO}, GAP, and μ). Although this result is in agreement with the CoMFA models developed for this system,[13] and could explain the poor clustering observed for this family, it questions the importance of the heterocycle for anti-*T. cruzi* activity. However, comparison of $\chi\pi$ of active compounds for the four chemical families shows a common electrophilic center at 4.1−4.9 Å from the oxygen atom of the *N*-oxide moiety (Figure 4). This points out an important pharmacophoric feature for anti-*T. cruzi* activity in accordance with the reactivity observed for these compounds. *N*-Oxide containing heterocycles are prone to nucleophilic substitution due to the deactivating effect of the *N*-oxide group.[41] The failure of electronic properties ($\chi\pi$, qσ) to discriminate between activity classes could be ascribed to the little variance capture by them. In general, *N*-oxide derivatives share a similar electronic structure, and difference among activity classes required the inclusion of other properties.

Model Building. The results obtained for 2-class categorization are shown in Table 3. The performance of each classifier was studied using the percentage of correct classification (% Correct class.) and the Matthews coefficient (C_{Matthews}) from stratified 10-fold cross-validation. The latter combines assessment of both recall and precision into a single number and is usually regarded as a more balanced measured than percentage of correct classification.[42] In general, global descriptors outperformed RDF descriptors ($\chi\pi$ + qσ), in agreement with the clustering observed in the Kohonen networks, validating again the ability of these descriptors to distinguish among classes. Regression methods (Slog, MLP) give better results when both types of descriptors are combined, while the reverse is observed for *k*-NN and J48. Regarding learning methods, locally approached algorithms such as *k*-NN and J48 performed better than globally approached ones (Slog, MLP), which could imply that activity classes are not homogeneous. Considering correct classification and C_{Matthews}, the best models were obtained using nearest neighbors and decision tree methods combined with global descriptors and simple logistic using both types of descriptors (Table 4). From these three models *k*-NN and J48 display better statistics than Slog in terms of recall and precision. Besides, the decision tree model could be considered a valuable tool in the design of new derivatives due to its simplicity, only four properties are required (E_{LUMO}, SM5, LogP, HDon) (Figure 5).

Table 3. Percentage of Correct Classification Using 10-Fold CV (Mean ± Standard Deviation)

	% Correct class. (C_{Matthews})		
method	Global	RDF[a]	Global + RDF
Slog	74 ± 2 (0.42)	72 ± 1 (0.34)	**80 ± 2 (0.57)**
MLP	76 ± 1 (0.49)	73 ± 3 (0.41)	79 ± 2 (0.55)
k-NN[b]	**81 ± 1 (0.60)**	71 ± 2 (0.33)	76 ± 2 (0.48)
J48[c]	**79 ± 10 (0.59)**	78 ± 2 (0.51)	75 ± 2 (0.47)

[a] RDF encoded $\chi\pi$ (dimension 31) plus RDF encoded qσ (dimension 13). [b] $k = 3$. [c] $M = 2$.

Table 4. Best Models Obtained

method	descriptors	recall[a]	precision[a]	F-measure
Slog	RDF$\chi\pi$ (1.7, 2.6, 3.1, 3.2, 4.4, 5.2, 6.2, 6.3),[b] RDFqσ (7.1, 9.3),[b] E_{LUMO}, SM5, LogP, GAP, HDon	0.69	0.75	0.72
k-NN	E_{LUMO}, SM5, LogP, GAP, HDon	0.74	0.74	0.74
J48	E_{LUMO}, SM5, LogP, HDon	0.82	0.82	0.82

[a] Averaged from 10-fold CV. [b] Distance in Å.

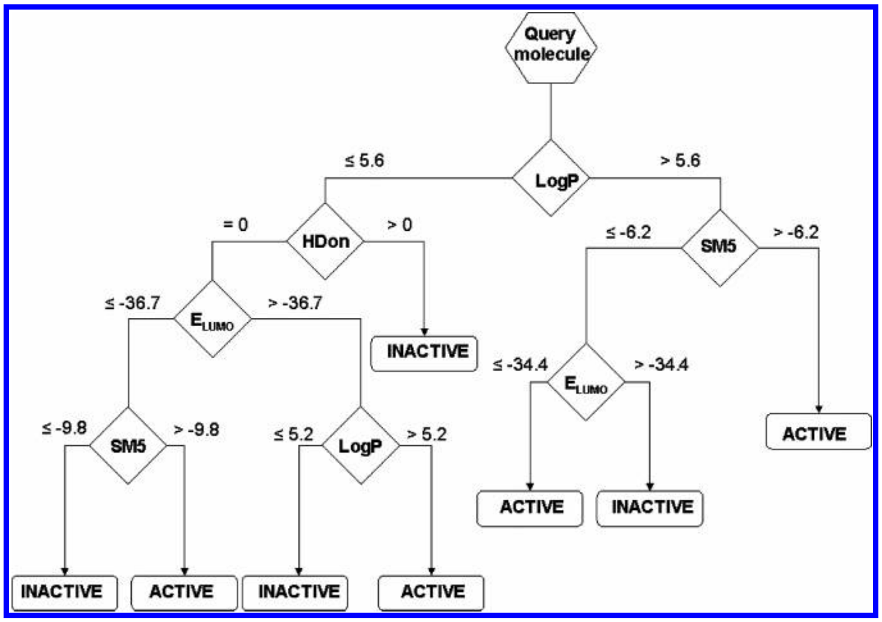

Figure 5. Decision tree model generated using global descriptors. E_{LUMO} and SM5 are expressed in kcal/mol.

Regarding classification within structural families using the best models obtained, benzofuroxan derivatives show a higher percentage of misclassification than the other families. While the percentage of misclassification for indazole, quinoxaline, and benzimidazole derivatives is around 10%, almost 20% of benzofuroxans derivatives are misclassified. Besides, the misclassification pattern differs significantly for this family when both models (*k*-NN, J48) are compared, whereas almost the same molecules are wrongly predicted for the other three families. Among indazole, quinoxaline, and benzimidazole derivatives wrongly classified compounds belong mainly to the "intermediate class" (PGI close to 60%) or represent unique structures, but no clear trend is observed for benzofuroxans.

Overall, 3-class categorization leads to qualitatively similar models but of poorer performance than 2-class categorization. While a two-class description is certainly less informative, it is a more reliable way of describing these data. Since any value of PGI used as a class boundary would be artificial, the criterion used was operational, our internal definition of a hit compound (PGI ≥ 60%). One of the main goals of the present study was to obtain a model that could be used in the design of new structures focusing the synthesis in the most promising ones. If a value of PGI of 40% is chosen, then a more balanced distribution of compounds among classes (60% inactive, 40% active) is obtained and the performance of the models is slightly improved (data not shown), but the number of compounds to be tested is also increased. Besides, a more rigorous definition of the active class helps in establishing the structural requirements for the biological activity.

Finally, the properties E_{LUMO}, SM5, LogP, and HDon were included in all the models using global descriptors and were the only ones present in the tree model, pointing to their relevance for the anti-*T. cruzi* activity. In the model that include RDF descriptors the distances selected for π-electronegativity emphasize the importance of the heterocyclic atoms (distances < 6 Å), while σ-charge distribution is relevant on the substituent. While the mode of action of *N*-oxide derivatives is still unknown, the results obtained here strengthen the importance of the electrophilic character of the molecule and the possible participation of the heterocycle in a reduction process. On the other hand, the development of a single QSAR model including the four structural families supports similar requirements for biological activity.

CONCLUSIONS

Although QSAR models can be used to address the mode of action of drugs, it is quite difficult to overcome problems associated with heterogeneity of enzymes and receptors. Herein, two types of descriptors were used to study the anti-*T. cruzi* activity: radial distribution function (RDF) and global descriptors. The main difference between the properties described using RDF or quantum chemistry methods is that the former gives a pharmacophoric structure in terms of 3D distribution. Both types of descriptors point to the relevance of electronic properties, and LRDF identified an electrophilic center at 4.1−4.9 Å from the oxygen atom of the *N*-oxide moiety. However, other properties are required to explain the biological activity. The results obtained here support the implication of a bioreduction process in the anti-*T. cruzi* activity of *N*-oxide containing heterocycles, though different enzymes could be involved.

Finally, we have succeeded in modeling the four structural families into a single model. According to the statistics observed, the best models were developed using *k*-NN and J48 both displaying a percentage of correct classification near 80%. The decision tree model using global descriptors displays the best statistics. Besides, tree-based methods are easily translated into classification rules which turn this model into a useful tool in the de novo construction of novel *N*-oxide lead structures.

ACKNOWLEDGMENT

M.B. is grateful to DAAD (Germany) for a Research Fellowship at the Computer-Chemie-Centrum and to PEDECIBA (Uruguay) for a Ph.D. fellowship.

Supporting Information Available: Structure and PGI values for all the compounds included in the study (Table 1S).

This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) World Health Organization. *Thirteenth Programme Report. UNDP/World Bank/World Health Organization programme for research and training in tropical diseases*; World Health Organization: Geneva, Switzerland, 1997.

(2) Urbina, J. Chemotherapy of Chagas Disease. *Curr. Pharm. Des.* **2002**, *8*, 287−295.

(3) Cerecetto, H.; Di Maio, R.; González, M.; Risso, M.; Saenz, P.; Seoane, G.; Denicola, A.; Peluffo, G.; Quijano, C.; Olea-Azar, C. 1,2,5-Oxadiazole *N*-oxide Derivatives and related compounds as Potential Antitrypanosomal Drugs. Structure-Activity Relationships. *J. Med. Chem.* **1999**, *42*, 1941−1950.

(4) Aguirre, G.; Cerecetto, H.; Di Maio, R.; González, M.; Porcal, W.; Seoane, G.; Denicola, A.; Ortega, M. A.; Aldana, I.; Moge-Vega, A. Benzo[1,2-*c*]1,2,5-oxadiazole *N*-oxide Derivatives as Potential Antitrypanosomal Drugs. Structure-Activity Relationships. Part II. *Arch. Pharm. (Weinheim, Ger.)* **2002**, *335*, 15−21.

(5) Olea-Azar, C.; Rigol, C.; Mendizábal, F.; Brinones, R.; Cerecetto, H.; Di Maio, R.; González, M.; Porcal, W.; Risso, M. Electrochemical and Microsomal Production of Free Radicals from 1,2,5-oxadiazole *N*-oxide as Potential Antiprotozoal Drugs. *Spectrochim. Acta, Part A* **2003**, *59*, 69−74.

(6) Aguirre, G.; Boiani, L.; Cerecetto, H.; Di Maio, R.; González, M.; Porcal, W.; Thomson, L.; Tórtora, V.; Denicola, A.; Möller, M. Benzo[1,2-c]1,2,5-oxadiazole *N*-oxide Derivatives as Potential Antitrypanosomal Drugs. Part III. Substituents-Clustering Methodology in the search of new active compounds. *Bioorg. Med. Chem.* **2005**, *13*, 6324−6335.

(7) Olea-Azar, C.; Rigol, C.; Mendizábal, F.; Cerecetto, H.; Di Maio, R.; González, M.; Porcal, W.; Morello, A.; Repetto, Y.; Maya, J. D. Novel Benzo[1,2-*c*]1,2,5-oxadizole *N*-oxide derivatives as Antichagasic Agents: Chemical and Biological Studies. *Lett. Drugs Des. Dev.* **2005**, *2*, 294−301.

(8) Porcal, W.; Hernandez, P.; Aguirre, G.; Boiani, L.; Boiani, M.; Merlino, A.; Ferreira, A.; Di Maio, R.; Castro, A.; González, M.; Cerecetto, H. Second generation of 5-Ethenylbenzofuroxan derivatives as Inhibitors of *Trypanosoma cruzi* growth: Synthesis, biological evaluation and strucutre-activity relationships. *Bioorg. Med. Chem.* **2007**, *15*, 2768−2781.

(9) Aguirre, G.; Boiani, M.; Cerecetto, H.; Gerpe, A.; González, M.; Fernández Sainz, Y.; Denicola, A.; Ochoa de Ocáriz, C.; Nogal, J.; Montero, D.; Escário, J. Novel Antiprotozoal Products: Imidazole and Benzimidazole N-Oxide Derivatives and Related Compounds. *Arch. Pharm. (Weinheim, Ger.)* **2004**, *337*, 259−270.

(10) Boiani, M.; Boiani, L.; Denicola, A.; Torres, S.; Serna, E.; Vera de Bilbao, N.; Sanabria, L.; Yaluff, G.; Nakayama, H.; Rojas de Arias, A.; Vega, C.; Rolan, M.; Gomez-Barrios, A.; Cerecetto, H.; González, M. 2*H*-Benzimidazole 1,3-dioxide derivatives: A new family of water-soluble anti-trypanosomatid agents. *J. Med. Chem.* **2006**, *49*, 3215−3220.

(11) Gerpe, A.; Aguirre, G.; Boiani, L.; Cerecetto, H.; González, M.; Olea-Azar, C.; Rigol, C.; Maya, J. D.; Morello, A.; Piro, O.; Arán, V. J.; Azqueta, A.; Lopez de Cerain, A.; Moge-Vega, A.; Rojas de Arias, A.; Yaluff, G. Indazole *N*-oxide derivatives as antiprotozoal agents: Synthesis, biological evaluation and mechanism of action studies. *Bioorg. Med. Chem.* **2006**, *14*, 3467−3480.

(12) Aguirre, G.; Cerecetto, H.; Di Maio, R.; González, M.; Montoya Alfaro, M. E.; Jaso, A.; Zarranz, B.; Ortega, M. A.; Aldana, I.; Moge-Vega, A. Quinoxaline *N*,*N*′-dioxide derivatives and related compounds as growth inhibitors of *Trypanosoma cruzi*. Structure-activity relationships. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3835−3839.

(13) Aguirre, G.; Boiani, L.; Boiani, M.; Cerecetto, H.; Di Maio, R.; González, M.; Porcal, W.; Denicola, A.; Piro, O.; Castellano, E.; Sant'Anna, C. M. R.; Barreiro, E. J. New potent 5-substituted Benzofuroxans as Inhibitors of *Trypanosoma cruzi* growth. Quantitative Structure-Activity relationship studies. *Bioorg. Med. Chem.* **2005**, *13*, 6336−6346.

(14) Silverman, R. B. *The Organic Chemistry of drug Design and drug Action*; Academic Press: San Diego, CA, 1992.

(15) Gasteiger, J. A Hierarchy of Structure Representations. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; pp 1034−1061.

(16) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(17) Wagner, S.; Hofmann, A.; Siedle, B.; Terfloth, L.; Merfort, I.; Gasteiger, J. Development of a Structural Model for NF-κB Inhibition of Sesquiterpene Lactones Using Self-Organizing Neural Networks. *J. Med. Chem.* **2006**, *49*, 2241−2252.

(18) *Spartan'04 Windows, 1.0.1*; Wavefunction, Inc.: Irvine, CA, 2003.

(19) *Spartan'04 Windows: Tutorial and User's Guide*; Wavefunction: Irvine, CA, 2003.

(20) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(21) Storer, J. W.; Giesen, D. J.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G.; Liotard, D. A. Solvation Modeling in Aqueous and Nonaqueous Solvents: New Techniques and a Re-examination of the Claisen Rearrangement. In *Structure and Reactivity in Aqueous Solution, ACS Symposium Series*; Cramer, C. J., Truhlar, D. G., Eds.; American Chemical Society: Washington, DC, 1994; Vol. 568.

(22) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for 3D structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(23) Clark, D. E. Rapid calculation of polar molecular surface area and it application to the prediction of transport phenomena.1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.

(24) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(25) *ADRIANA.Code*, *1.0*; Molecular Networks GmbH: Erlangen, Germany, 2006. http://www.molecular-networks.com.

(26) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.

(27) Hemmer, M. C.; Gasteiger, J. Prediction of Three-dimensional molecular structures using information from infrared spectra. *Anal. Chim. Acta* **2000**, *420*, 145−154.

(28) *Computer program ARDF*; Computer-Chemie-Centrum, Universität Erlangen-Nuremberg: Erlangen, Germany 2005.

(29) Mazzatorta, P.; Vracko, M.; Jezierska, A.; Benfenati, E. Modeling toxicity by using supervised Kohonen neural networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 485−492.

(30) *SONNIA software 4.1*; Molecular Networks GmbH: Erlangen, Germany 2005. http://www.molecular-networks.com.

(31) Kohonen, T. *Self-Organizing maps*, 3rd ed.; Springer: New York, 2001.

(32) Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.

(33) Witten, I. H.; Eibe, F. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann: San Francisco, CA, 2000.

(34) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning*; Springer: New York, 2001.

(35) Harrel, F. E. J. *Regression Modeling Strategies- With applications to Linear Models, Logistic Regression and Survival Analysis*; Springer-Verlag: New York, 2001.

(36) Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; John Wiley & Sons, Inc.: 2000.

(37) Frimurer, T.; Bywater, R. Improving the odds in discriminating "Drug-like" from "Non Drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.

(38) Schneider, G.; Wrede, P. Artificial neural networks for computed-based molecular design. *Prog. Bioph. Mol. Biol.* **1998**, *70*, 175−222.

(39) Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. O. Chemoinformatic-Based Classification of Prohibited Substances Employed for Doping in Sport. *J. Chem. Inf. Model.* **2006**, *46*, 2369−2380.

(40) Jolliffe, T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

(41) Albini, A.; Pietra, S. *Heterocyclic N-oxides*; CRC Press: Boca Raton, FL, 1991.

(42) Baldi, P.; Brunak, S. *Bioinformatics: the machine learning approach*, 2nd ed.; Cambridge, MA, 2001.

CI7002768