J Mol Model (2008) 14:975–985
DOI 10.1007/s00894-008-0332-x

ORIGINAL PAPER

A neural networks study of quinone compounds with trypanocidal activity

Fábio Alberto de Molfetta ·
Wagner Fernando Delfino Angelotti ·
Roseli Aparecida Francelin Romero ·
Carlos Alberto Montanari ·
Albérico Borges Ferreira da Silva

Received: 9 January 2008 / Accepted: 4 June 2008 / Published online: 16 July 2008

Abstract This work investigates neural network models for predicting the trypanocidal activity of 28 quinone compounds. Artificial neural networks (ANN), such as multilayer perceptrons (MLP) and Kohonen models, were employed with the aim of modeling the nonlinear relationship between quantum and molecular descriptors and trypanocidal activity. The calculated descriptors and the principal components were used as input to train neural network models to verify the behavior of the nets. The best model for both network models (MLP and Kohonen) was obtained with four descriptors as input. The descriptors were T_5 (torsion angle), QTS1 (sum of absolute values of the atomic charges), VOLS2 (volume of the substituent at region B) and HOMO−1 (energy of the molecular orbital below HOMO). These descriptors provide information on the kind of interaction that occurs between the compounds and the biological receptor. Both neural network models used here can predict the trypanocidal activity of the quinone compounds with good agreement, with low errors in the testing set and a high correctness rate. Thanks to the nonlinear model obtained from the neural network models, we can conclude that electronic and structural properties are important factors in the interaction between quinone compounds that exhibit trypanocidal activity and their biological receptors. The final ANN models should be useful in the design of novel trypanocidal quinones having improved potency.

Keywords Quinone · Trypanocidal activity · Neural network · Multilayer perceptrons · Kohonen models

Introduction

American trypanosomiasis—or Chagas disease—is a parasitic disease endemic to Latin America, where it is a major cause of heart disease with 18–20 million people infected and over a 100 million at risk. It is caused by infection with the flagellate protozoan *Trypanosoma cruzi*, which is transmitted to humans by triatomine vectors (kissing bugs) or through contact with infected blood [1].

Trypanosoma cruzi, a haemoflagelete protozoan (family Trypanosomatidae, order Kinetoplastida), is the etiological agent of Chagas disease and its life cycle involves obligatory passage through vertebrate (mammals, including human) and invertebrate (hematophagus triatomine insects) hosts. Transmission of the infective trypomastigote form occurs mainly by vector insect bite (80–90%), blood transfusion (5–20%) and congenital routes (0.5–8.0%). The chronic disease is characterized by cardiac, digestive or neurological disturbances [2]. Furthermore, the development of an effective vaccine has been hampered by the complex biology and high adaptability through antigenic

F. A. de Molfetta · C. A. Montanari · A. B. F. da Silva (✉)
Departamento de Química e Física Molecular,
Instituto de Química de São Carlos,
Universidade de São Paulo, CP 780,
13560-970 São Carlos, SP, Brasil
e-mail: alberico@iqsc.usp.br

W. F. D. Angelotti
Departamento de Físico-Química, Instituto de Química,
Universidade Estadual de Campinas, CP 6154,
13081-970 Campinas, SP, Brasil

R. A. F. Romero
Departamento de Ciências da Computação e Estatística,
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, CP 668,
13560-970 São Carlos, SP, Brasil

variation of the protozoan causative agent. Thus, the main line of defense against parasitic infections has been through chemotherapy [3]. Problems associated with trypanocidal drugs include their limited efficacy, human toxicity, high cost, and the emergence of drug-resistant trypanosome strains [4]. In view of this, the development of new classes of readily accessible compounds with trypanocidal activity and improved pharmacological properties is imperative.

Quinones, particularly 1,4-naphthoquinones (1,4-NQs), are widespread among the secondary metabolites of plants and microorganisms. They can also be prepared synthetically and are widely produced by the chemical industry as organic dyes. Interest in 1,4-NQ is not restricted to the chemistry of dyes; a wide spectrum of biological activities is described for them, including antitumor, wound healing, anti-inflammatory, antiparasitic and cytotoxic activities, among others. These biological activities have justified the large number of studies found in the literature aimed at the synthesis and evaluation of either natural quinones or their analogues as potential pharmacological agents [5].

Artificial neural networks (ANNs) are widely used as pattern recognition methods to learn relationships in a way similar to that used by the human brain. ANNs have become an important modeling technique in numerous areas of chemistry and pharmacy. Because they have a great capacity for adaptability, they are recommended as a powerful tool for pattern classification and for building predictive models [6, 7].

In this work, neural network models for predicting the trypanocidal activity of 28 quinone compounds are investigated. With the aim of modeling the nonlinear relationship between quantum and molecular descriptors and trypanocidal activity, we employed ANNs such as multilayer perceptrons (MLP) and Kohonen models. In the present work, two types of variables were used as input to train the neural network models: calculated descriptors and principal components (PCs). In this case, the most important PCs were then taken as network inputs instead of the original data. Both models were built in order to verify the behavior of the nets.

Methods

Artificial neural networks provide a powerful technique for modeling nonlinear relationships [8]. The ANN technique was applied in order to discover the possible existence of non-linear relationships between activity and molecular descriptors that are ignored in linear approaches.

Neural networks are, therefore, commonly applied in pharmaceutical research to analyze the complex relationships that exist among the structure of molecules and their physicochemical or biological properties, with the goal of identifying which structural features are of pharmacological importance [9]. For this purpose, we have used molecular descriptors. These descriptors allow structural information to be used as the input required for training the neural networks.

The quinone compounds examined here have been reported in the literature as powerful and selective trypanocidal agents [10]. The central structure, numbering and chemical structure of the 28 quinone compounds studied in this work are presented in Fig. 1.

The geometry optimizations of the quinone compounds were performed with the initial structures (see Fig. 1) by using the DFT/B3LYP functional [11]. The choice of the DFT method was made because recent studies have demonstrated that the DFT/B3LYP method leads to excellent results for the analysis of geometries and energies [12, 13]. After obtaining the minimum energy conformation for each compound, molecular properties (variables or descriptors) of the 28 quinone compounds were calculated using the DFT/B3LYP functional with the 6-31G* basis set, as implemented in the GAUSSIAN 98 computational package [14]. This basis set is the standard basis set for calculations involving up to medium size systems. These descriptors are: total energy, energy of the highest occupied molecular orbital (HOMO), energy of the molecular orbital below HOMO (HOMO−1), energy of the lowest unoccupied molecular orbital (LUMO), differences of some molecular orbital energies, bond orders over all bonds that comprise the basic skeleton of the quinone compounds studied, molecular hardness, molecular softness, dipole moment, molecular volume, torsion angle, atomic charges Qi, i=1, 2,...,18 (see Fig. 1), and the sum of the atomic charges of the substituents at regions A and B (see Fig. 1); atomic charges were calculated with the CHELPG [14] (Charges from Electrostatic Potentials using a Grid based method) option in the GAUSSIAN 98 program [15]. The octanol/water partition coefficients were calculated using the program XlogP [16], and the topological descriptors were evaluated from the Dragon 3.0 molecular package [17], totaling 209 calculated descriptors for each molecule.

The whole data set was autoscaled along all the variables, i.e., normalized and centered on the mean, so that they could be compared to each other on the same scale.

After calculation of the atomic and molecular descriptors, Fisher's weights of these descriptors were obtained and the more significant descriptors were selected, i.e., those that had greatest Fisher weights were considered to have a high ability in the discrimination (separation) between active and inactive compounds.

Fisher's weights allow evaluation of how useful a variable is to discriminate the samples between groups. This tool uses the variance and the difference between averages for each variable for a group of representative samples (training set) to calculate a score related to the ability of the variable to indicate differences between groups [18].

Results and discussion

Multi-layer perceptrons

Multi-layer perceptrons (MLP) are feedforwarded multilayer networks that provide flexible frameworks for non-linear function estimation. An MLP consists of formal neurons or nodes, and the connections (weights) among them. In a MLP architecture, the neurons are arranged in layers (an input layer, one or more hidden layers, and an output layer), and the connections are unidirectional from input to output. Adjacent layers are fully connected but no connections exist among neurons within the same layer. This architecture computes a numerical output value, *f*(x), for a given numerical input vector x, which is the row of the X matrix corresponding to a given object (molecule, species, etc...). A formal neuron sums up incoming signals multiplied by the connection weights, subtracts a threshold value (or bias θ),

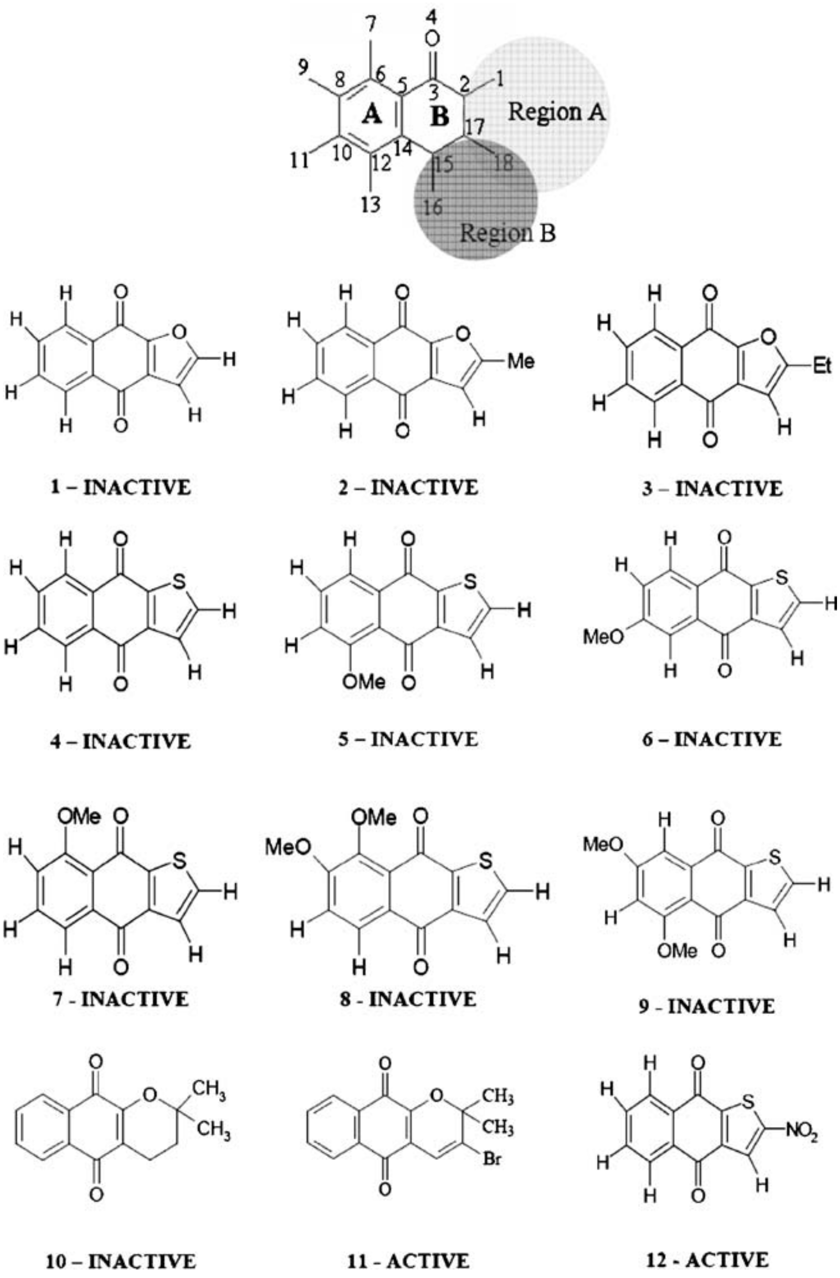
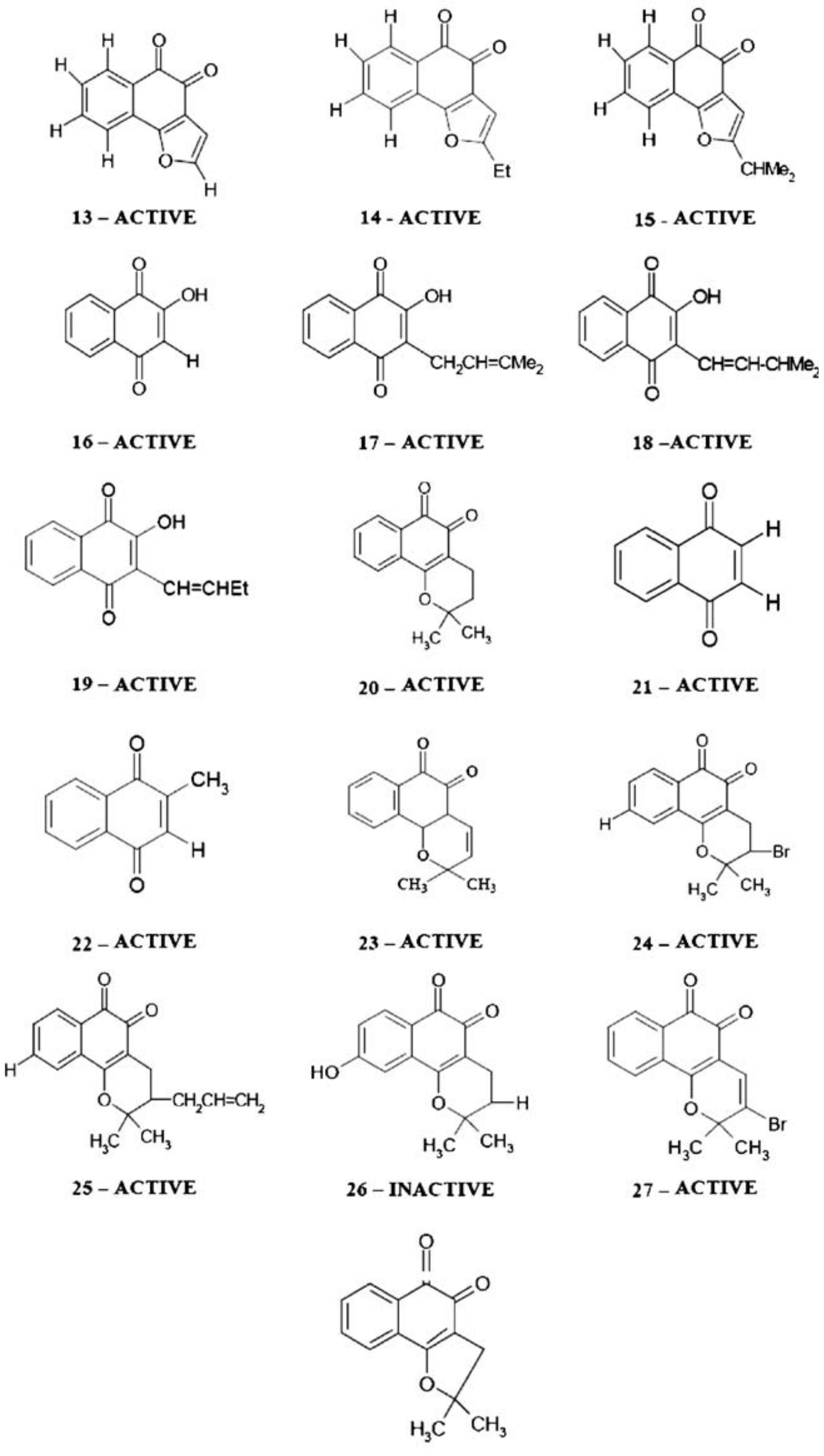


Fig. 1 The central chemical structure, numbering used and chemical structure of the 28 quinone compounds studied

Fig. 1 (continued)



13 – ACTIVE

14 - ACTIVE

15 - ACTIVE

16 – ACTIVE

17 – ACTIVE

18 –ACTIVE

19 – ACTIVE

20 – ACTIVE

21 – ACTIVE

22 – ACTIVE

23 – ACTIVE

24 – ACTIVE

25 – ACTIVE

26 – INACTIVE

27 – ACTIVE

28 – ACTIVE

and calculates an output signal by using the so-called transfer function. Neurons can have different transfer functions. Input neurons simply distribute the descriptor data to the hidden layer neurons without any further computation. Hidden layer neurons typically have a sigmoidal transfer function:

$$\mathrm{sf}(\mathrm{input}) = \frac{1}{1 + e^{-input}} \tag{1}$$

that limits the neuron's output signal to values between 0 and 1. The output layer neurons usually have sigmoidal or linear transfer functions, depending on the application. The whole network represents a non-linear relationship that can be written for each output as:

$$\widehat{y}= \mathrm{f(x)} = \sum_h sf\left[\left(\sum_i x_i w_{ih} - \theta_i\right)\right] w_h - \theta_h \tag{2}$$

where w_{ih} is the connection weight between the input node i with the hidden node h and w_{h} are the connection weights between each hidden node h with the final output considered, y. The values of θ_{i} and θ_{h} are the biases corresponding to the input and hidden layers. The difference between $\hat{y}$ and y (target) is the target error, which is subsequently back-propagated to modify the weights in order to attain the best fit [19].

In order to reduce the number of descriptors to be considered, principal component analysis (PCA) aims at data reduction through linear combinations of the original variables. PCA seeks to group these correlated variables, generating a new set of variables called principal components (PCs). These PCs are built as linear combinations of the original variables and have the important property of being completely uncorrelated. The first new axis, PC1, is chosen in the direction that maximizes the variance; the second axis must be chosen orthogonal to the first and in the direction to describe as much variance left as possible and so on.

The initial data matrix, represented by X, is decomposed into two matrices, T and P, where

$$\mathrm{X} = \mathrm{TP}^{\mathrm{T}} \tag{3}$$

In Eq. 3, T is known as the scores matrix and represents the position of the samples in the new coordinate system. The matrix P in Eq. 3 is known as the loadings matrix and describes how the new axes, i.e., the PCs, are built from the original variables. The samples are mapped through scores and variables by the loadings in the new low dimensional vector space defined by the PCs [20].

The data set was randomly split to form training and test data sets. The training and test data sets comprised 20 and 8 compounds, respectively. The test set was used to monitor the overall performances of the trained network. Once the best topology of the network was obtained and the convergence criterion was reached, a leave-one-out cross-validation procedure was also employed to validate the performances of the resulting network results.

The prediction results of the different models studied are presented in terms of root-mean-square (RMS) prediction error. The RMS error is defined as

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n} \left(y - y_p\right)^2}{n}} \tag{4}$$

where y is the true value, y_{p} the predicted value and n is the number of samples.

The numbers of hidden neurons were first optimized by trying several network architectures, varying the number of neurons in the hidden layer from 1 to 20. The architecture was obtained by averaging triplicate RMS values of the training set for each neuron; the model with minimum RMS for the training set was selected.

In a first step, PCA analysis with 209 descriptors was carried out for the 28 compounds shown in Fig. 1. The first seven PCs explained 83.8% of the data variance, and was enough to describe the data set. The seven PCs for each compound were used as input to the neural networks, reducing the size of this input from 209 to 7.

The network had one input layer, one hidden layer and one output layer. The output layer is the class of active and inactive compounds. The hidden layer, with minimum RMS using seven PCs, was 13 neurons. Accordingly, the final MLP architecture was 7-13-2.

This model showed high values for RMS and with one correctness rate of the net around 25% for the test set, which indicates the low predictive power of the model. The neural network of this data set revealed that there are a large number of descriptors with many intercorrelations and redundancies.

We next constructed another neural network model using the best 12 descriptors, which were selected according to the values of Fisher's weight. These descriptors were then used as input to the neural networks. Furthermore, the model with minimum RMS for the training set was selected with 20 neurons. Accordingly, the final MLP architecture for this net was 12-20-2.

The descriptors were T1, T2, T3 and T_5 (torsion angles), BO (bond order), QTS1 (sum of absolute values of the atomic charges), VOLS2 (volume of the substituent at region B) and HOMO−1 (energy of the molecular orbital below HOMO), BIC (Balaban index), E2v (weighted by atomic van der Waals volumes), E1e (weighted by atomic Sanderson electronegativities) and E2p (weighted by atomic polarizabilities).

The BIC is calculated using a topological matrix of distances in the vertex- and edge-weighted graph reflecting

Table 1 Models for 12 descriptors, selected by Fisher's weight with rightness rate, and training (RMS1) and test (RMS2) errors

	Compounds of test set	Rightness rate (%)	RMS1	RMS2
Model 1	9, 10, 22, 23, 24, 25, 27, 28	87.5	0.0154	0.125
Model 2	1, 10, 22, 23, 24, 25, 27, 28	87.5	0.0184	0.125
Model 3	9, 2, 22, 23, 24, 25, 27, 28	100	0.0132	0
Model 4	9, 10, 3, 23, 24, 25, 27, 28	87.5	0.0171	0.125
Model 5	9, 10, 22, 4, 24, 25, 27, 28	87.5	0.0186	0.125
Model 6	5, 9, 10, 22, 23, 25, 26, 27	75	0.0027	0.178
Model 7	6, 9, 10, 22, 23, 25, 27, 28	87.5	0.0185	0.125
Model 8	1, 2, 7, 22, 23, 24, 25, 26	87.5	0.0023	0.125

all types of atoms and chemical bonds in a given molecule [21]. The other three topological indices (E2v, E1e and E2p) are WHIN (weight holistic invariant molecular) descriptors, and these descriptors contain information about the whole 3D molecular structure in terms of size, shape, symmetry and atom distribution. These indices are calculated from the *x*, *y*, and *z*-coordinates of a 3D structure of the molecule, usually from a spatial conformation of minimum energy, within different weighting schemes in a straightforward manner and represent a very general approach to describe molecules in a unitary conceptual framework [22].

In this method, we constructed eight models randomly separated into training and test sets (see Table 1) with 70% of the compounds for training and 30% for testing sets, where the model kept a constant ratio between active and inactive compounds.

In Table 1 (see model 6) we can see that compounds 10 and 26 were classified incorrectly as active compounds. This model presents both compounds, and has a 75% rightness rate. From Table 1, the rightness rate of the net stayed around 87.5% for the test set, which indicates an improvement in the predictive power of the model regarding the first model.

In order to verify the behavior of the net, 12 descriptors selected by Fisher's weights were subjected to PCA analysis. In this case, five PCs for each compound were used as input to the neural networks, where these components explained 90.4% of the total data variance.

The architecture chosen and used for comparison was that which produced the minimum error. For this model, the architecture was 5-10-2. Therefore, in this case, using the PCs as input to the neural network and the same eight models with the compounds used before (see Table 2), the results were similar to those for the 12 descriptors. The rightness rate of the net stayed around 75% for the test set and, depending on the model, the compounds (10 and 26) were classified incorrectly as active compounds and one compound was classified as inactive (25).

The results obtained by PCs, which contain most of the variability in the data set, were thus capable of describing in a similar way the results obtained previously, albeit in a much lower dimensional space.

In a previous work [20], four descriptors—T_5, QTS1, VOLS1 and HOMO−1 (see Table 3)—were important for the separation between active and inactive compounds, and these were used as input to the neural networks. The process of choosing the net architecture was as described above. For this model, the architecture was 4-10-2.

With the same compounds, and using the leave-one-out crossvalidation procedure, the rightness rate of the net increased to 87.5% for the test set and only one compound (26) was classified as active incorrectly (Table 4).

The two classes of quinone compounds were labeled, that is, active compounds were labeled by code (1 0) and inactive compounds as (0 1). From Table 5 we can see that compound 26 was labeled as (0.9198 0.0104), but it should

Table 2 Models of the data set, with five principal components (PCs) with rightness rate, and training (RMS1) and test (RMS2) errors

	Compounds of test set	Rightness rate (%)	RMS1	RMS2
Model 1	9, 10, 22, 23, 24, 25, 27, 28	75	0.0164	0.160
Model 2	1, 10, 22, 23, 24, 25, 27, 28	75	0.0151	0.179
Model 3	9, 2, 22, 23, 24, 25, 27, 28	87.5	0.0156	0.125
Model 4	9, 10, 3, 23, 24, 25, 27, 28	75	0.0168	0.179
Model 5	9, 10, 22, 4, 24, 25, 27, 28	75	0.0172	0.160
Model 6	5, 9, 10, 22, 23, 25, 26, 27	75	0.0032	0.177
Model 7	6, 9, 10, 22, 23, 25, 27, 28	75	0.0172	0.177
Model 8	1, 2, 7, 22, 23, 24, 25, 26	87.5	0.0027	0.125

Table 3 Four molecular descriptors calculated for the quinone compounds studied and used as input to neural network with indication of activity. *T_5* Torsion angle, *QTS1* sum of absolute values of the atomic charges, *VOLS2* volume of the substituent at region B, *HOMO−1* energy of the molecular orbital below HOMO

Compound	T_5 (°)	QTS1	VOLS2 (Å^3)	HOMO−1 (eV)	Activity
1	106.53	0.021	243.56	−0.3545	Inactive
2	106.47	0.002	243.09	−0.3522	Inactive
3	106.46	0.014	242.91	−0.3520	Inactive
4	112.93	−0.018	241.33	−0.3541	Inactive
5	112.84	−0.049	241.43	−0.3449	Inactive
6	112.96	−0.049	242.02	−0.3501	Inactive
7	113.17	−0.046	242.46	−0.3404	Inactive
8	113.07	−0.074	242.31	−0.3421	Inactive
9	112.80	−0.075	241.44	−0.3449	Inactive
10	120.80	−0.001	248.43	−0.3488	Inactive
11	118.35	−0.056	354.71	−0.3557	Active
12	112.84	−0.097	241.98	−0.3711	Active
13	132.55	−0.097	261.14	−0.3614	Active
14	132.66	−0.161	373.62	−0.3591	Active
15	132.60	−0.149	418.22	−0.3588	Active
16	121.75	−0.090	200.43	−0.3570	Active
17	123.09	−0.018	450.33	−0.3495	Active
18	123.80	−0.034	454.67	−0.3525	Active
19	123.82	−0.025	405.35	−0.3526	Active
20	118.19	−0.130	420.39	−0.3528	Active
21	122.40	−0.093	187.59	−0.3600	Active
22	121.49	−0.053	200.99	−0.3566	Active
23	123.24	−0.102	410.63	−0.3570	Active
24	123.21	−0.048	466.21	−0.3630	Active
25	123.31	−0.180	540.66	−0.3562	Active
26	122.93	−0.1150	411.17	−0.3423	Inactive
27	120.07	−0.0480	465.14	−0.3622	Active
28	128.48	−0.1820	380.90	−0.3576	Active

be labeled as (0 1) and this is an indication that this compound was incorrectly classified. Thus, through MLP feedforward neural networks, trained by back-propagation (BP), only one compound was classified as an outlier.

Kohonen network (self-organizing map)

Cluster analysis is often used to justify a chemistry space: if compounds with similar biological behavior are grouped in the proposed space, then it seems reasonable to conclude that the chemistry space is good. In order to settle structural similarities among the quinone compounds, a self-organizing map (SOM) was built for these compounds [8].

The Kohonen network or SOM can be used to study data of high-dimensional spaces by projection into a two-dimensional plane [22]. The Kohonen architecture is based on a single layer of neurons that are arranged in a box having on its top a two dimensional grid of responses. During the training phase, each input vector x_S is presented to the network, and only the neuron whose weight vector is most similar to this input is stimulated (competitive learning). In more mathematical terms, this so-called winning neuron (c) is selected as the one providing the minimal Euclidean distance to the pattern vector x_S:

$$c \leftarrow \min_j \left\{ \sum \left(x_{si} - w_{ji}\right)^2 \right\}, j = 1, 2, 3, \ldots, N \times N \qquad (5)$$

where x_{Si} and w_{ji} are the ith coordinate of the input vector x_S and the ith weight level of neuron j, respectively, and N × N is the number of neurons in the Kohonen layer. After the winning neuron in the Kohonen layer is selected, the weights of each neuron j (w_{ji}) in the Kohonen layer are updated in order to make the weights closer to the input vector according to following equation:

$$\Delta w_{ji} = \eta\left(1 - \frac{d_r}{d_{\max} + 1}\right) \times \left(x_{Si} - w_{ji}^{old}\right) \text{for } d = 0, 1, \ldots, d_{\max} \qquad (6)$$

where η is the learning rate, w_{ji}^{old}denotes the numerical value of the weight w_{ji} at the previous iteration, Δw_{ji} the weight update, x_{Si} the input vector, and the error is scaled

Table 4 Models with four descriptors, with rightness rate, and training (RMS1) and test (RMS2) errors

	Compounds of test set	Rightness rate (%)	RMS1	RMS2
Model 1	9, 10, 22, 23, 24, 25, 27, 28	100	0.0107	0
Model 2	1, 10, 22, 23, 24, 25, 27, 28	100	0.0058	0
Model 3	9, 2, 22, 23, 24, 25, 27, 28	100	0.0060	0
Model 4	9, 10, 3, 23, 24, 25, 27, 28	100	0.0044	0
Model 5	9, 10, 22, 4, 24, 25, 27, 28	100	0.0045	0
Model 6	5, 9, 10, 22, 23, 25, 26, 27	87.5	0.0068	0.125
Model 7	6, 9, 10, 22, 23, 25, 27, 28	100	0.0042	0
Model 8	1, 2, 7, 22, 23, 24, 25, 26	87.5	0.0063	0.125

Table 5 Results obtained with model 6 for training test

Compound	Activity	Class of compound		Predicted class	
5	Inactive	0	1	0.0171	0.9707
9	Inactive	0	1	0.0154	0.9626
10	Inactive	0	1	0.0412	0.9357
22	Active	1	0	0.8508	0.1025
23	Active	1	0	0.9967	0.0172
25	Active	1	0	0.9927	0.0192
26	Inactive	0	1	0.9198	0.0104
27	Active	1	0	0.9927	0.0269

according to the topological distance d_r from the winner. The topological distance d_r is defined as the number of neurons separating the neuron j from the winning neuron. The size of the neighborhood $d_{\max}$, which at the beginning of learning covers the entire network, decreases during the training phase and eventually it is limited only to the winning neuron.

Additionally, the learning rate constant η also changes during the training phase up to a minimum value, which is reached when the number of training epochs (n_{epoch}) equals a pre-specified maximum value (n_{tot}) [23]:

$$\eta = \left(\eta^{start} - \eta^{final}\right)\left(1 - \frac{\eta_{epoch}}{\eta_{tot}}\right) + \eta^{final} \tag{7}$$

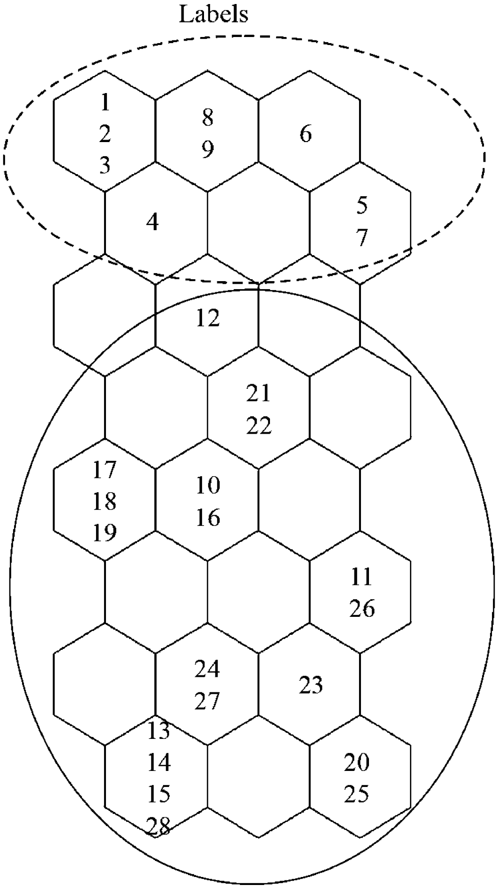


Fig. 2 Compounds mapping onto a bi-dimensional space for 12 descriptors selected by Fisher's weights. All compounds were assigned to one of two groups according to their activity. *Continuous* and *dotted lines* represent active and inactive groups, respectively

The advantage of the SOM, compared with some other projection methods is that the algorithm is very simple, straightforward to implement and fast to compute. In the field of pharmaceutical sciences, the SOM has been applied to search for useful drugs [24].

SOM toolbox version 2 was used for clustering the quinone compounds according to their activity. The SOM toolbox, developed by the Laboratory of Computer and Information Science, Helsinki University of Technology, Finland, is available free of charge as a Matlab toolbox at the website http://www.cis.hut.fi/projects/somtoolbox/. The computation environment used was Matlab version 6.5, developed by Mathworks (http://www.mathworks.com).

In order to improve data interpretation, three SOMs were built with the 28 compounds. The first map was built with 12 descriptors selected by Fisher's weights. The second was built by the five PCs obtained with 12 descriptors, and the third was built with the four descriptors that were responsible for the separation between the active and inactive compounds in previous work [20].

Figure 2, also named Labels, corresponds to the mapping of the 28 samples from a space of X-dimensional onto a bi-dimensional region. In this figure, each sample is represented by one neuron (the winner neuron). The SOM made

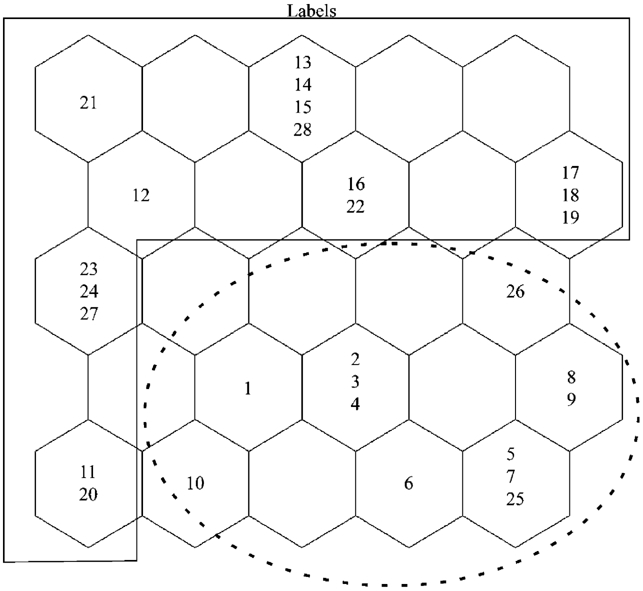


Fig. 3 Compounds mapping onto a bidimensional space for five principal components (PCs). All compounds were assigned to one of two groups according to their activity. *Continuous* and *dotted lines* represent active and inactive groups, respectively

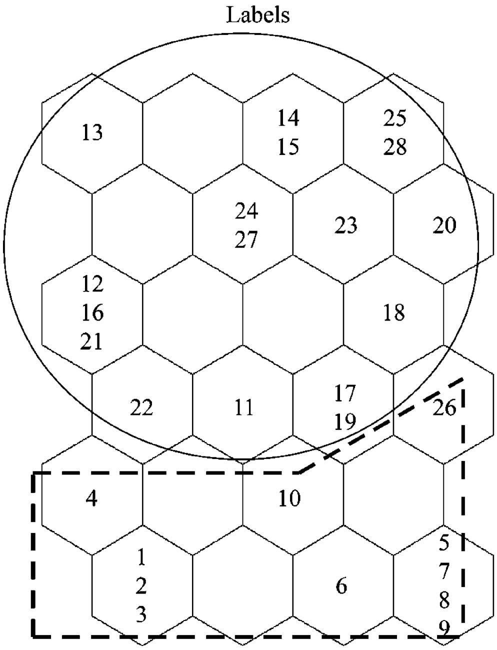


Fig. 4 Compounds mapping onto a bi-dimensional space for four-descriptors. All compounds were assigned to one of two groups according to their activity. *Continuous* and *dotted lines* represent active and inactive groups, respectively

possible separation of each group into two regions: one for active compounds and the other for inactive ones.

From Fig. 2, with 12 descriptors selected by Fisher's weights, we can see that compounds 10 and 26 are defined in a region of the map of active compounds. As illustrated in Fig. 2, these compounds are defined by a unique neuron with compounds 16 and 11, and thus are considered as one sub-group.

With five PCs obtained with 12 descriptors, we can see according to Fig. 3 that compound 25 was defined in a region of the map for inactive compounds. This compound is also defined by a unique neuron with compounds 5 and 7. We can see also that compounds 10 and 26 are close within a region of the map for active compounds. This result is in agreement with MLP results, where compound 25 was also classified as inactive and compounds 10 and 26 as active.

From the four descriptors (T_5, QTS1, VOLS2 and HOMO−1) it was possible to reduce high dimensionality data, thus making data interpretation and visualization easier than with the original data set. The classification of two different classes is also possible. Again, according to Fig. 4 we can see that compound 26 was defined in a close region of the map of active compounds. This result was similar to that generated by MLP neural networks. Therefore, this compound was also classified as an outlier through a Kohonen neural network.

Figure 5 shows the component maps, each one corresponding to the compounds as well as those variables determined before. The value of the variable can be read on the scale (right side of each map). Considering both Figs. 4

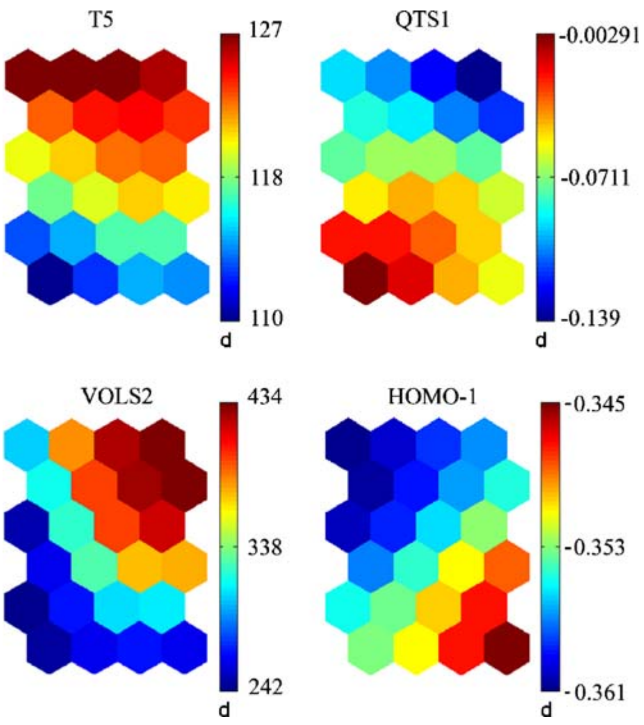


Fig. 5 Compound component maps, where each map shows the calculated descriptors

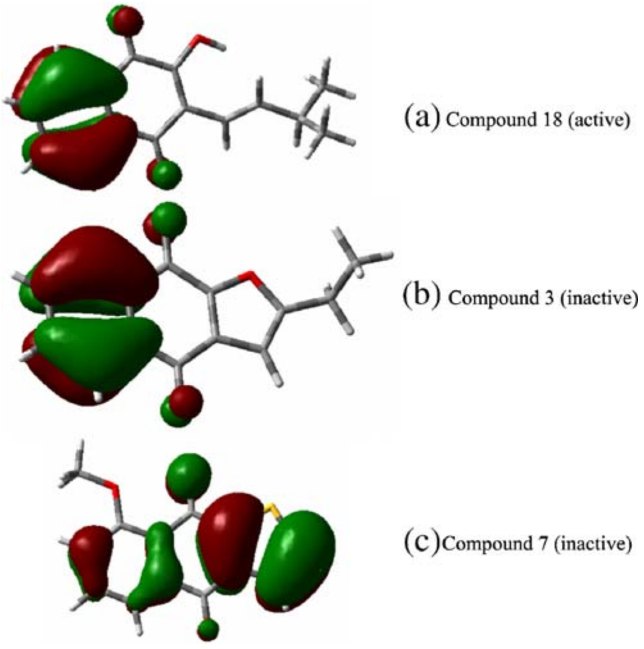


Fig. 6 Examples of HOMO−1 contributions for three of the quinone compounds studied. **a** Compound 18, representing active compounds (note that all active compounds have no methoxy groups in their structures); **b** compound 3, representing inactive compounds that have no methoxy groups in their structures; **c** compound 7, representing inactive compounds that have methoxy groups in their structures

and 5, the value of each variable for each sample can be recognized. The position occupied by a sample in the *Label* map (Fig. 4) corresponds to the same position in the component maps (Fig. 5).

From Fig. 5 we can see that the variable T_5 has higher values for active compound than inactive ones, and this indicates that active compounds need to have a suitable conformation, as determined by the torsion angle formed by the atoms C1, C2, C3 and C4 (T_5), so that they may effectively interact with the biological receptor.

Regarding the sum of the charges of atoms C1, C2, C17, C18, and the substituents in region A (QTS1), we can see from Fig. 5 that the active compounds need to have electron-acceptor atoms in region A.

According to Fig. 5, the active compounds need to have high values of VOLS2 (the volume formed by the atoms C15, O16, C17, C18 and the substituents at region B). The variable VOLS2 can be related to the "fit" between the compound and the receptor.

The energy of the frontier orbitals is an important property in several chemical and pharmacological processes, and the reason for this is the fact that these properties provide information on the electron-donating and electron-accepting character of a compound, i.e., on the formation of a charge transfer complex. From Fig. 5 we can see that the energy of HOMO−1 for the active compounds must be lower than the value for inactive compounds. This means that active compounds are not good electron-donor molecules when compared to inactive ones, i.e., perhaps the inactive compounds interact through a charge transfer mechanism before reaching the biological receptor, causing the loss of anti-trypanocidal activity of these compounds.

Here, it is interesting to make some comments on the influence of HOMO−1 on the trypanocidal activity of the quinone compounds studied in this work, since HOMO−1 has also been found to be an important variable in previous studies [25–28]. In order to illustrate the importance of HOMO−1 in the discrimination between the active and inactive quinone compounds studied here, we show in Fig. 6 where HOMO−1 has its main contributions (the three cases for the HOMO−1 behavior shown in Fig. 6 are those observed in most of the 28 quinone compounds studied).

Figure 6a and b display the HOMO−1 contributions for active and inactive compounds without methoxy groups in their structures, respectively; Fig. 6c shows the HOMO−1 contributions for inactive compounds that have methoxy groups in their structures. From Fig. 6, we can draw two important conclusions: (1) all active, and some inactive, quinone compounds have their main HOMO−1 contributions located in atoms of ring A (see Fig. 1); (2) the main HOMO−1 contributions in inactive compounds that have a methoxy group, are located in atoms of region A (see Fig. 1). So, only in inactive compounds is the HOMO−1 significantly affected by the presence or absence of methoxy groups.

In concluding, we can state that electronic and structural properties are important factors in understanding the interaction between quinone compounds with anti-trypanocidal activity and their biological receptors. The electronic properties are related to the strength of a molecular association by electronic interaction, and structural properties are related to the positioning of the molecule during its interaction with the biological receptor. In fact, the characteristics of the quinone compounds revealed in this work could be useful in the design of new quinone molecules with trypanocidal activity.

Conclusions

In this study, theoretical calculations and artificial neural networks (ANN) were used for modeling and predicting the behavior of 28 quinone compounds regarding their trypanocidal activity.

Both multilayer perceptrons (MLP) feedforward neural networks trained by back-propagation (BP) and Kohonen neural networks showed similar results when we used different values as input to the neural networks. The results reveal that the reduction of variables is important for the improvement of the correctness rate of active/inactive compounds. With four variables, two electronic (QTS1 and HOMO−1) and two structural properties (T_5 and VOLS2), the rightness rate of the net increased by 87.5% for the test set and only one compound was classified incorrectly, which is an indication of the goodness of the fit.

From the four variables, we can conclude that electronic and structural properties are important factors in determining the interaction between quinone compounds with trypanocidal activity and their biological receptors. Electronic properties are related to the strength of a molecular association by electronic interaction, and structural properties are related to the positioning of the molecule during the interaction with the biological receptor.

Acknowledgments The authors would like to acknowledge CAPES, CNPq and FAPESP (Brazilian Agencies) for the financial support given to this research.

References

1. Siles R, Chen S, Zhou M, Pinney KG, Trawick ML (2006) Bioorg Med Chem Lett 16:4405–4409
2. Batista R, Humberto JL, Chiari E, Oliveira AB (2007) Bioorg Med Chem Lett 15:381–391
3. Bauer H, Massey V, Arscott LD, Schirmer RH, Ballou DP, Williams CH (2003) J Biol Chem 278:33020–33028

4. Li ZL, Fennie MW, Ganem B, Hancock MT, Kobaslija M, Rattendi D, Bacchi CJ, O'Sullivan M (2001) Bioorg Med Chem Lett 11:251–254
5. del Corral JMM, Castro MA, Oliveira AB, Gualberto SA, Cuevas C, San Feliciano A (2006) Bioorg Med Chem 14:7231–7240
6. Winkler DA, Burden FR (2004) Drug Discov Today 2:104–111
7. Katritzky AR, Pacureanu LM, Slavov S, Dobchev DA, Karelson M (2006) Bioorg Med Chem 14:6933–6939
8. Fernández M, Caballero J (2006) Bioorg Med Chem 14:280–294
9. Selzer P, Ertl PJ (2006) Chem Inf Model 46:2319–2323
10. Goulart MOF, Zani CL, Tonholo J, Freitas LR, de Abreu FC, Oliveira AB, Raslan DS, Starling S, Chiari E (1997) Bioorg Med Chem Lett 7:2043–2048
11. Becke AD (1993) J Chem Phys 98:5648–5652
12. El-Azhary AA, Sutter HU (1996) J Phys Chem 100:15056–15063
13. Turecek F (1998) J Phys Chem A 102:4703–4713
14. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA, Vreyen Jr T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Octhterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al- Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA (2004) Gaussian 03, revision C.02. Gaussian, Pittsburgh PA
15. Breneman CM, Wiberg KB (1990) J Comput Chem 11:361–373
16. Wang R, Fu Y, Lai L (1997) J Inf Comput Sci 37:615–621
17. Todeschini R, Gramatica P (1998) Perspect Drug Discov Des 9:355–380
18. Skrobot VL, Castro EVR, Pereira RCC, Pasa VMD, Fortes ICP (2007) Energy Fuels 21:3394–3400
19. González-Arjona D, López-Pérez A, González AG (2002) Talanta 56:79–90
20. Molfetta FA, Bruni AT, Honório KM, da Silva ABF (2005) Eur J Med Chem 40:329–338
21. Bykov VA, Popov PI, Pleteneva TV, Anisimova IE, Syroeshkin AV (2004) Pharm Chem J 38:243–249
22. Kubinyi H, Folkers G, Martin YC (2002) 3D QSAR in drug design, vol 2. Kluwer, New York
23. Marini F, Zupan J, Magri AL (2005) Anal Chim Acta 544:306–314
24. Kawakami J, Hoshi K, Ishiyama A, Miyagishima S, Sato K (2004) Chem Pharm Bull 52:751–755
25. Santo LLD, Galvão DS (1999) J Mol Struct (THEOCHEM) 464:273–279
26. Barone PMVB, Camilo A, Galvão DS (1996) Phys Rev Lett 77:1186–1189
27. Braga RS, Barone PMVB, Galvão DS (1999) J Mol Struct (THEOCHEM) 464:257–266
28. Rothenberg G, Sasson Y (1999) Tetrahedron 55:561–568