Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 2680–2689
Bioorganic & Medicinal Chemistry

QSAR analysis for heterocyclic antifungals

Pablo R. Duchowicz,[a,*] Martín G. Vitale,[a] Eduardo A. Castro,[a] Michael Fernández[b] and Julio Caballero[b]

[a]*Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, 1900 La Plata, Argentina*
[b]*Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, 44740 Matanzas, Cuba*

Received 12 December 2006; revised 14 January 2007; accepted 19 January 2007
Available online 24 January 2007

Abstract—We perform linear regression analyses on 1202 numerical descriptors that encode the various aspects of the topological, geometrical and electronic molecular structure with the aim of achieving the best QSAR relationship between the antifungal potencies against the *Candida albicans* strain and the structure of 96 heterocyclic ring derivatives. As a realistic application we employ the model found to predict the biological activity for 60 non-yet measured compounds.

1. Introduction

Last decades have witnessed increasing efforts for developing new antifungal drugs that are capable to inhibiting various diseases related to *Candida albicans* species, considered to be responsible for the most common types of human pathogenic fungi.[1] The highly resistant fungi organisms produce a broad range of infections ranging from non-life-threatening mucocutaneous illnesses to invasive processes that may involve virtually any organ,[2] and represent an important cause of morbidity and mortality in hospitalized patients.[3] Obviously, the design of new pharmacological drugs possessing novel modes of action is required in order to reduce the dramatic increase in frequency of systemic infections along with the newly appearing fungal species, the development of resistance to the present azole therapies and also for diminishing the high toxicity of polyenes.[4–6] A wide number of known effective antimicrobial remedies include heterocyclic systems in their structure, like imidazoles, quinazolines, benzazoles and oxazolo (4,5-b) pyridines,[7–12] although none of these substances exhibit simultaneously an optimally desired spectrum, potency, pharmacological properties, etc.

The experimental search for better activities in drug discovery is commonly carried out in the laboratory by optimizing the structure–activity relationship (SAR) of the functional groups present in a leading structure in terms of their biological endpoint. However, an interesting alternative to this trial-error based procedure that constitutes an active field in complex biochemical phenomena are the analyses through Quantitative Structure–Activity Relationships (QSAR).[13] Nowadays, the QSAR theory is extensively applied for studying the effects and antifungal potencies of compounds.[12,14–16]

A recent study of M. Fernández et al.[17] employs the Genetics Algorithms (GA)[18] and the Bayesian-Regularized Neural Networks hybridised with GA (BRANN-GA)[19] variable selection approaches to establish new models for the antifungal activity (*C. albicans* strain), in a set composed of 96 heterocyclic compounds. These models are achieved by treating only 721 geometry-dependent descriptors from the software Dragon,[20] and pairs of variables with a correlation coefficient (R) greater than 0.9500 are omitted from the analysis, therefore leading to a total pool containing $D = 322$ descriptors to be investigated. An appropriate partition of the set of compounds into a training set ($N = 80$) and an external test set ($N = 16$) is performed by means of the k-Means Clustering technique.[21] The linear GA model found shows to fit the training set with $R = 0.8733$, meanwhile the non-linear BRANN-GA

Keywords: QSAR theory; Molecular descriptors; Replacement method; Antifungal potency.
* Corresponding author. Fax: +54 221 425 4642; e-mail addresses: prduchowicz@yahoo.com.ar; duchow@inifta.unlp.edu.ar

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.01.039

exhibits a higher value of $R = 0.9680$. The authors of this work concluded that the distribution of atomic mass, volume and polarizability are relevant factors influencing the antifungal potency of the compounds studied.

The purpose of the present research is to analyse the same molecular set of 96 heterogeneous heterocyclic compounds as studied previously[17] in order to propose an alternative connection between the antifungal activity and the structure of these organic substances. To this end, we simultaneously explore a greater number of molecular descriptors ($D = 1202$) including descriptor definitions of all classes and not limited only to the conformation-dependent types, by means of two alternative computational strategies: the Forward Stepwise Regression (FSR)[22] and the Replacement Method (RM).[23–26] Finally, we draw conclusions by contrasting our results with the previously reported ones.

2. Results and discussion

We use the same training set ($N = 80$)-test set ($N = 16$) partition as reported in the study of M. Fernández et al.[17] The application of RM to the available pool containing $D = 1202$ descriptors leads to an optimized relationship that, in terms of the best predictive power of the equation (given by the $l - 10\% - o$ parameters) and the least number of parameters involved, has six molecular descriptors:

$$\begin{aligned} \mathrm{p}MIC[\mathrm{M}] = {} & 4.362(\pm 0.1) + 0.0954(\pm 0.02)HATSp \\ & + 1.172(\pm 0.1)Mor13v + 1.245(\pm 0.2)H8u \\ & + 0.374(\pm 0.06)MATS6e \\ & + 2.409(\pm 0.2)Mor29p + 1.095(\pm 0.2)Mor27v \end{aligned} \quad (1)$$

$$N = 80, R = 0.8859, S = 0.106, F = 44.386,$$
$$AIC = 0.0134, FIT = 2.295, p < 10^{-4}, R_{loo} = 0.8617,$$
$$S_{loo} = 0.111, R_{l-10\%-o} = 0.8316, S_{l-10\%-o} = 0.123.$$

Here, the absolute errors of the regression coefficients are given in parentheses and R is the correlation coefficient of the model, F is the Fisher ratio, p is the significance of the model, AIC is the Akaike's information criterion and FIT is the Kubinyi function. The FSR procedure does not improve the quality of the previous relationship:

$$\begin{aligned} \mathrm{p}MIC[\mathrm{M}] = {} & 5.386(\pm 0.2) + 0.108(\pm 0.01)nDB \\ & + 1.058(\pm 0.1)Mor13v - 0.153(\pm 0.04)L2s \\ & + 0.0419(\pm 0.008)RDF060m \\ & + 0.295(\pm 0.06)Mor03p - 0.287(\pm 0.09)GATS4e \end{aligned} \quad (2)$$

$$N = 80, R = 0.8472, S = 0.121, F = 30.950, AIC = 0.0176,$$
$$FIT = 1.601, p < 10^{-4},$$
$$R_{loo} = 0.8272, S_{loo} = 0.135, R_{l-10\%-o} = 0.7800, S_{l-10\%-o} = 0.138$$

and so we will not consider the FSR result from now on. The details for the descriptors involved in all the proposed models of present analysis are given in Table 2. Table 3 summarizes the statistics found in Ref. 17 by resorting to the GA and BRANN-GA approaches, and includes the structural parameters that appeared in these models. As can be easily observed, a comparison between the linear GA results and those achieved with RM in Eq. 1 leads to good agreements in both the calibration and validation (leave-one-out cross validation) parameters.

Table 4 shows the predicted values for all the 96 antifungal potencies and their corresponding residuals between parentheses. The plot of predicted versus experimental activities shown in Figure 2 suggests that the 80 training compounds follow a straight line. Figure 3 plots the residuals in terms of the experimental data and demonstrates that the best molecular descriptors given by Eq. 1 lead to a model that follows a normal distribution and that does not obey any kind of undesired pattern, that would probably suggest the presence of non-modelled factors contributing to the biological activity. This figure includes three outlier heterocyclics with a residual exceeding the value $2 \cdot S$: compounds **37**, **40** and **75**. Also, the correlation matrix for the descriptors of the model in Table 5 reveals that the variables are not seriously intercorrelated and so justifying the presence of all the variables in the model. The application of the linear QSAR model (4) on the external test set of 16 heterocyclics leads to a root mean square residual (rms) of 0.015, comparable to the GA value of 0.013 (see Table 3). These results confirm once again the RM technique as a simple and useful tool for variable subset selection. On the contrary, the highly non-linear Bayesian-Regularized Neural Network hybridised with Genetics Algorithm (BRANN-GA) approach performs much better than both the GA and RM linear approaches in the present data set.

The QSAR model obtained with RM confirms the reported GA result[17] that various geometry-dependent descriptors are relevant to the antifungal potencies, as the model includes five 3D and one 2D structural variable. These structural parameters can be classified as follows: (i) three 3D-MoRSE: *Mor13v*, *Mor29p*, *Mor27v*; (ii) two GETAWAY: Leverage-weighted total index/weighted by atomic polarizabilities, *HATSp* and H autocorrelation of lag 8/unweighted, *H8u*; and (iii) one 2D-Autocorrelation: Moran autocorrelation lag-6/weighted by atomic Sanderson electronegativities, *MATS6e*. This subset of descriptors assumes that the distribution of atomic Sanderson electronegativity, volume and polariz-

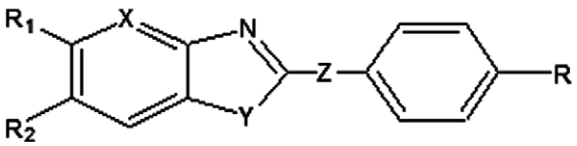

Figure 1. General structure of heterocyclic ring derivatives under study.

Table 1. Structures and in vitro antifungal activities against *Candida Albicans*

Compound	X	Y	Z	R	R_1	R_2	p*MIC*[M]
1[a]	CH	O	—	H	H	H	3.892
2[a]	CH	O	—	$C(CH_3)_3$	H	H	4.001
3[a]	CH	O	—	NH_2	H	H	3.924
4[a]	CH	O	—	$NHCOCH_3$	Cl	H	4.059
5[a]	CH	O	—	Cl	Cl	H	4.024
6[a]	CH	O	—	NO_2	Cl	H	4.040
7[a]	CH	O	—	H	NO_2	H	4.282
8[a]	CH	O	—	CH3	NO_2	H	4.308
9[a]	CH	O	—	$C(CH_3)_3$	NO_2	H	4.375
10[a]	CH	O	—	NH_2	NO_2	H	4.310
11[a]	CH	O	—	Cl	NO_2	H	4.342
12[a]	CH	O	—	Br	NO_2	H	4.406
13[a]	CH	O	—	C_2H_5	NH_2	H	3.979
14[a]	CH	O	—	F	NH_2	H	3.960
15[a]	CH	O	—	$N(CH_3)_2$	NH_2	H	4.005
16[a]	CH	O	—	CH_3	CH_3	H	3.950
17[a]	CH	O	—	C_2H_5	CH_3	H	3.977
18[a]	CH	O	—	OCH_3	CH_3	H	3.980
19[a]	CH	O	—	F	CH_3	H	3.958
20[a]	CH	O	—	$NHCOCH_3$	CH_3	H	4.027
21[a]	CH	O	—	$NHCH_3$	CH_3	H	3.979
22[a]	CH	O	—	$N(CH_3)_2$	CH_3	H	4.004
23[a]	N	O	—	CH_3	H	H	4.225
24[a]	N	O	—	C_2H_5	H	H	4.253
25[a]	N	O	—	OCH_3	H	H	4.257
26[a]	N	O	—	OC_2H_5	H	H	4.283
27[a]	N	O	—	NH_2	H	H	4.227
28[a]	N	O	—	NO_2	H	H	4.285
29[a]	CH	O	—	Br	NH_2	H	4.110
30[a]	CH	O	CH_2	OCH_3	H	H	4.282
31[a]	CH	O	CH_2	NO_2	H	H	4.308
32[a]	CH	O	CH_2	H	Cl	H	4.290
33[a]	CH	O	CH_2	OCH_3	Cl	H	4.340
34[a]	CH	O	CH_2	Br	Cl	H	4.410
35[a]	CH	O	CH_2	NO_2	Cl	H	4.363
36[a]	CH	O	CH_2	H	NO_2	H	4.609
37[a]	CH	O	CH_2	OCH_3	NO_2	H	4.657
38[a]	CH	O	CH_2	Br	NO_2	H	4.725
39[a]	CH	O	CH_2	Cl	NO_2	H	4.664
40[a]	CH	O	CH_2O	H	H	NO_2	3.732
41[a]	CH	O	CH_2O	Cl	Cl	NO_2	3.831
42[a]	CH	O	CH_2S	H	NO_2	H	4.359
43[a]	CH	O	CH_2S	H	CH_3	H	4.009
44[a]	N	O	CH_2O	H	H	H	4.26
45[a]	N	O	CH_2O	Cl	H	H	4.319
46[a]	CH	NH	CH_2O	Cl	CH_3	H	4.037
47[a]	CH	NH	CH_2S	H	NO_2	H	4.358
48[a]	CH	NH	CH_2S	H	CH_3	H	4.009
49[a]	CH	O	CH_2O	H	$COOCH_3$	H	4.054
50[a]	CH	O	CH_2O	Cl	$COOCH_3$	H	4.104
51[a]	CH	NH	CH_2O	Cl	$COOCH_3$	H	4.102
52[a]	CH	NH	CH_2S	H	$COOCH_3$	H	4.076
53[a]	CH	O	C_2H_4	H	NO_2	H	4.331
54[a]	N	O	C_2H_4	H	H	H	4.253
55[a]	CH	NH	CH_2O	H	NO_2	H	4.283
56[a]	CH	NH	CH_2O	Cl	H	H	4.015
57[a]	CH	NH	CH_2S	H	Cl	H	4.041
58[a]	CH	NH	C_2H_4	H	H	H	4.078
59[a]	CH	O	CH_2O	H	H	CH_3	3.981
60[a]	CH	O	CH_2O	Cl	Cl	H	4.071
61[a]	CH	O	CH_2O	Cl	CH_3	H	3.738
62[a]	CH	O	CH_2O	Cl	H	CH_3	3.738
63[a]	CH	O	CH_2O	H	Cl	H	4.344
64[a]	CH	O	CH_2S	H		CH_3	4.009
65[a]	CH	O	CH_2O	H	H	H	3.955

Table 1 (*continued*)

Compound	X	Y	Z	R	R_1	R_2	p*MIC*[M]
66[a]	CH	O	CH_2O	H	NO_2	H	4.034
67[a]	CH	O	CH_2O	H	Cl	H	4.017
68[a]	CH	O	CH_2O	Cl	NO_2	H	4.086
69[a]	CH	O	CH_2S	H	H	H	4.286
70[a]	CH	O	CH_2S	H	Cl	NO_2	4.409
71[a]	CH	O	CH_2S	H	$COOCH_3$	H	4.379
72[a]	CH	S	CH_2O	H	H	H	3.684
73[a]	CH	S	CH_2O	Cl	H	H	3.742
74[a]	CH	S	CH_2S	H	H	H	4.013
75[a]	CH	NH	CH_2O	H	Cl	H	4.316
76[a]	CH	NH	CH_2O	H	$COOCH_3$	H	4.053
77[a]	CH	NH	CH_2O	Cl	Cl	H	4.370
78[a]	CH	NH	CH_2NH	H	H	H	3.951
79[a]	CH	NH	CH_2NH	H	CH_3	H	3.977
80[a]	CH	NH	C_2H_4	H	Cl	H	4.012
81[b]	CH	O	-	$NHCH_3$	H	H	3.952
82[b]	CH	O	—	C_2H_5	Cl	H	4.013
83[b]	CH	O	—	$NHCH_3$	Cl	H	4.025
84[b]	CH	O	CH_2	H	H	H	4.223
85[b]	CH	O	CH_2	Cl	H	H	4.290
86[b]	CH	O	CH_2	NO_2	NO_2	H	4.680
87[b]	CH	O	CH_2	Br	H	H	4.360
88[b]	CH	O	CH_2O	H	CH_3	H	3.980
89[b]	CH	O	CH_2O	H	Cl	NO_2	3.785
90[b]	CH	O	CH_2O	Cl	H	H	4.016
91[b]	CH	O	CH_2O	Cl	H	NO_2	3.785
92[b]	CH	O	CH_2S	H	H	NO_2	4.360
93[b]	CH	NH	CH_2O	H	H	H	3.953
94[b]	CH	NH	CH_2O	H	CH_3	H	3.979
95[b]	CH	NH	CH_2S	H	H	H	4.284
96[b]	CH	NH	C_2H_4	H	CH_3	H	4.277

[a] Member of the training set.
[b] Member of the external test set.

ability influences the variation of the antifungal activity in the training set.

The orthogonalization and standardization[22] of the regression coefficients in Eq. 1 allows assigning more importance to those variables exhibiting larger absolute standardized orthogonal coefficients (shown in parentheses), therefore leading to the following ranking of contributions to p*MIC*:

$$\begin{array}{cccccc} H8u & Mor13v & MATS6e & Mor29p & Mor27v & HATSp \\ (0.46) & (0.42) & (0.41) & (0.33) & (0.26) & (0.23) \end{array} \quad (3)$$

It has to be noticed that the observed activities are positive quantities, although this is not the case for the molecular descriptors appearing in Eq. 1 as these are real numbers that can take positive or negative values. In example, from the standardization result of (6) it is expected that higher values of descriptors such as *H*8*u*, *Mor13v* or *MATS6e*, which tend to contribute most to the antifungal potency, would cause an increment in the p*MIC* predictions while assuming that the remaining variables of the equation are held invariant.

The 3D-MoRSE type of descriptor is obtained considering a molecular transform derived from an equation used in electron diffraction studies.[38] The electron diffraction does not directly yield atomic coordinates, but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. These codes are defined in order to reflect the contribution at a prescribed scattering angle of an atomic property such as mass (*m*), polarizability (*p*), electronegativity (*e*) or volume (*v*) to the property under investigation, and so enabling to differentiate the nature of atoms. For example, in the case of *Mor13v*, the scattering angle is of 13 $Å^{-1}$ and the atomic volumes are employed as weighting scheme.

All GETAWAY descriptors[39] match the 3D-molecular geometry and are derived from the elements h_{ij} of the molecular influence matrix (H), obtained through the numerical values of atomic Cartesian coordinates. The diagonal elements of H (h_{ii}) are called leverages and are considered to represent the influence of each molecule atom in determining the whole shape of the molecule. In example, the mantle atoms always have higher h_{ii} values than atoms near the molecule centre. Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interactions with the *i*th atom. Both GETAWAY variables appearing in Eq. 6 contemplate in their calculation a sum of atomic contributions: *HATSp* considers the atomic polarizabilities and the diagonal elements h_{ii}, while *H*8*u* remains unweighted and assigns equal importance to each atom employing the off-diagonal elements h_{ij}.

Table 2. Symbols for molecular descriptors involved in the QSAR models

Molecular descriptor	Type	Description
HATSp	GETAWAY	Leverage-weighted total index/weighted by atomic polarizabilities
Mor13v	3D-MoRSE	3D-MoRSE—signal 13/weighted by atomic van der Waals volumes
H8u	GETAWAY	H autocorrelation of lag 8/unweighted
MATS6e	2D-Autocorrelations	Moran autocorrelation lag-6/weighted by atomic Sanderson electronegativities
Mor29p	3D-MoRSE	3D-MoRSE—signal 29/weighted by atomic polarizabilities
Mor27v	3D-MoRSE	3D-MoRSE—signal 27/weighted by atomic van der Waals volumes
nDB	Constitutional	Number of double bonds
L2s	WHIM	Second component size directional WHIM index/weighted by atomic electrotopological states
RDF060m	RDF[a]	Radial distribution function—6.0 weighted by atomic masses
Mor03p	3D-MoRSE	3D-MoRSE—signal 03/weighted by atomic polarizabilities
GATS4e	2D-Autocorrelations	Geary Autocorrelation-lag 4/weighted by atomic Sanderson electronegativities
Mor19v	3D-MoRSE	3D-MoRSE—signal 19/weighted by atomic van der Waals volumes
Mor29v	3D-MoRSE	3D-MoRSE—signal 29/weighted by atomic van der Waals volumes
H5m	GETAWAY	H Autocorrelation of lag 5/weighted by atomic masses
RDF055u	RDF	Radial distribution function—5.5 unweighted
RDF085m	RDF	Radial distribution function—8.5 weighted by atomic masses
Mor10u	3D-MoRSE	3D-MoRSE—signal 10/unweighted
Mor25p	3D-MoRSE	3D-MoRSE—signal 25/weighted by atomic polarizabilities
L2e	WHIM	Second component size directional WHIM index/weighted by atomic Sanderson
R5e	GETAWAY	R autocorrelation of lag 5/weighted by atomic Sanderson electronegativities
R7e	GETAWAY	R autocorrelation of lag 7/weighted by atomic Sanderson electronegativities
Mor18p	3D-MoRSE	3D-MoRSE—signal 18/weighted by atomic polarizabilities
G1e	WHIM	First component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities
BEHm6	BCUT	Highest eigenvalue no. 6 of Burden matriz/weighted by atomic masses
Mor18e	3D-MoRSE	3D-MoRSE—signal 18/weighted by atomic Sanderson electronegativities
E2u	WHIM	Second component accessibility directional WHIM index/unweighted
nO	Constitutional	Number of oxygen atoms
R8u	GETAWAY	R autocorrelation of lag 8/unweighted
PW4	Topological	Path/walk 4—Randic shape index
MWC08	MWC[b]	Molecular walk count of order 08
Mor19e	3D-MoRSE	3D-MoRSE—signal 19/weighted by atomic Sanderson electronegativities
nBnz	Constitutional	Number of benzene-like rings

[a] Radial distribution function.
[b] Molecular walk count.

Table 3. Statistical parameters for linear and non-linear QSAR models of antifungal activities reported17 and obtained in the present study ($N = 80$)

Model	Variables	R	R_{loo}	rms test set[a]
MLR-GA	Mor13v, Mor19v, Mor27v, Mor29v, H8u, H5m	0.8733	0.8450	0.0130
BRANN	*Mor13v, Mor19v, Mor27v, Mor29v, H8u, H5m*	0.9429	0.7900	0.0120
BRANN-GA	*RDF055u, RDF085m, Mor10u, Mor25p, E1u, H8v*	0.9680	0.8300	0.0083
RM	***HATSp, Mor13v, H8u, MATS6e, Mor29p, Mor27v***	**0.8859**	**0.8617**	**0.0150**
RM-CP	***HATSp, Mor13v, H8u, MATS6e, Mor29p, Mor27v***	**0.9555**	**0.9466**	**0.0118**

[a] rms, root mean square residual.

The structural variables introduced by Moran[40] correspond to bi-dimensional autocorrelations between pairs of atoms in the molecule and are also defined to quantify the contribution of a considered atomic property to the analysed property. These can be readily calculated, that is, by summing products of terms including the atomic weights for the terminal atoms in all the paths of a prescribed length. For the case of *MATS6e*, the path connecting a pair of atoms has length 6 and involves the atomic Sanderson electronegativities as weighting scheme.

The employment of cross-products among the descriptors of Eq. 1 of the form $X_i \cdot X_j \cdot X_k$ generates a linear RM-CP model that satisfies the following equation:

$$\begin{aligned} \mathrm{p}MIC[\mathrm{M}] = {} & 4.430(\pm 0.06) - 0.155(\pm 0.03)HATSp \cdot R7e \cdot Mor18p \\ & + 2.857(\pm 0.2)Mor13v \cdot G1e \cdot BEHm6 \\ & - 2.209(\pm 0.1)H8u \cdot Mor18e \cdot E2u \\ & + 0.495(\pm 0.04)MATS6e \cdot nO \cdot R8u \\ & + 35.060(\pm 2)Mor29p \cdot PW4 \cdot MWC08 \\ & + 0.548(\pm 0.05)Mor27v \cdot Mor19e \cdot nBnz \end{aligned} \quad (4)$$

$$N = 80, R = 0.9555, S = 0.0674, F = 127.600,$$

$$AIC = 0.00542, FIT = 6.604, p < 10^{-4},$$

$$R_{loo} = 0.9466, S_{loo} = 0.0704, R_{l-10\%-o} = 0.9302, S_{l-10\%-o} = 0.0770$$

and, as can be immediately noticed from Table 3, it yields a statistics of similar quality to that provided by the more computational demanding Bayesian-Regular-

Table 4. Predicted in vitro antifungal activities by QSAR models

Compound	p*MIC*[M]		
	Exp.	Pred. Eq. 4	Pred. Eq. 7
1[a]	3.892	3.867 (0.025)	3.936 (−0.044)
2[a]	4.001	4.133 (−0.132)	4.036 (−0.036)
3[a]	3.924	4.083 (−0.159)	3.997 (−0.073)
4[a]	4.059	4.046 (0.013)	4.084 (−0.026)
5[a]	4.024	4.131 (−0.107)	4.038 (−0.014)
6[a]	4.040	4.197 (−0.157)	4.107 (−0.067)
7[a]	4.282	4.186 (0.096)	4.255 (0.027)
8[a]	4.308	4.296 (0.012)	4.310 (−0.002)
9[a]	4.375	4.440 (−0.065)	4.390 (−0.015)
10[a]	4.310	4.189 (0.121)	4.266 (0.043)
11[a]	4.342	4.228 (0.114)	4.380 (−0.038)
12[a]	4.406	4.259 (0.147)	4.404 (0.001)
13[a]	3.979	3.949 (0.030)	4.028 (−0.050)
14[a]	3.960	4.050 (−0.090)	3.979 (−0.020)
15[a]	4.005	4.011 (−0.006)	4.028 (−0.024)
16[a]	3.950	3.880 (0.070)	3.933 (0.017)
17[a]	3.977	3.871 (0.106)	3.862 (0.115)
18[a]	3.980	3.999 (−0.019)	4.002 (−0.023)
19[a]	3.958	4.032 (−0.074)	3.905 (0.053)
20[a]	4.027	3.984 (0.043)	4.024 (0.003)
21[a]	3.979	4.014 (−0.035)	3.978 (0.001)
22[a]	4.004	3.969 (0.035)	3.929 (0.074)
23[a]	4.225	4.175 (0.050)	4.255 (−0.030)
24[a]	4.253	4.190 (0.063)	4.200 (0.052)
25[a]	4.257	4.230 (0.027)	4.270 (−0.014)
26[a]	4.283	4.271 (0.012)	4.324 (−0.042)
27[a]	4.227	4.219 (0.008)	4.148 (0.079)
28[a]	4.285	4.309 (−0.024)	4.338 (−0.054)
29[a]	4.110	4.129 (−0.019)	4.063 (0.047)
30[a]	4.282	4.134 (0.148)	4.229 (0.052)
31[a]	4.308	4.385 (−0.077)	4.296 (0.012)
32[a]	4.290	4.342 (−0.052)	4.364 (−0.074)
33[a]	4.340	4.209 (0.131)	4.314 (0.026)
34[a]	4.410	4.547 (−0.137)	4.461 (−0.051)
35[a]	4.363	4.435 (−0.072)	4.399 (−0.037)
36[a]	4.609	4.620 (−0.011)	4.623 (−0.015)
37[a]	4.657	4.435 (0.222)	4.637 (0.020)
38[a]	4.725	4.798 (−0.073)	4.788 (−0.064)
39[a]	4.664	4.515 (0.149)	4.525 (0.138)
40[a]	3.732	4.043 (−0.311)	3.853 (−0.122)
41[a]	3.831	3.881 (−0.050)	3.773 (0.058)
42[a]	4.359	4.431 (−0.072)	4.386 (−0.027)
43[a]	4.009	4.073 (−0.064)	4.102 (−0.093)
44[a]	4.260	4.151 (0.109)	4.283 (−0.024)
45[a]	4.319	4.144 (0.175)	4.221 (0.098)
46[a]	4.037	4.106 (−0.069)	4.067 (−0.031)
47[a]	4.358	4.445 (−0.087)	4.275 (0.083)
48[a]	4.009	3.927 (0.082)	3.989 (0.019)
49[a]	4.054	4.000 (0.054)	4.040 (0.013)
50[a]	4.104	4.019 (0.085)	4.017 (0.087)
51[a]	4.102	4.211 (−0.109)	4.223 (−0.122)
52[a]	4.076	4.125 (−0.049)	4.064 (0.012)
53[a]	4.331	4.370 (−0.039)	4.300 (0.031)
54[a]	4.253	4.248 (0.005)	4.319 (−0.066)
55[a]	4.283	4.326 (−0.043)	4.282 (0.000)
56[a]	4.015	4.176 (−0.161)	4.148 (−0.133)
57[a]	4.041	4.035 (0.006)	4.127 (−0.087)
58[a]	4.078	4.095 (−0.017)	4.032 (0.045)
59[a]	3.981	3.903 (0.078)	3.846 (0.134)
60[a]	4.071	3.933 (0.138)	3.952 (0.119)
61[a]	3.738	3.895 (−0.157)	3.827 (−0.090)
62[a]	3.738	3.921 (−0.183)	3.854 (−0.116)
63[a]	4.344	4.291 (0.053)	4.312 (0.032)
64[a]	4.009	4.107 (−0.098)	4.049 (−0.041)
65[a]	3.955	3.958 (−0.003)	3.935 (0.020)
66[a]	4.034	4.097 (−0.063)	4.040 (−0.007)
67[a]	4.017	4.047 (−0.030)	4.036 (−0.019)
68[a]	4.086	4.102 (−0.016)	4.106 (−0.021)
69[a]	4.286	4.167 (0.119)	4.182 (0.103)
70[a]	4.409	4.279 (0.130)	4.320 (0.088)
71[a]	4.379	4.311 (0.068)	4.403 (−0.025)
72[a]	3.684	3.857 (−0.173)	3.839 (−0.155)
73[a]	3.742	3.769 (−0.027)	3.770 (−0.029)
74[a]	4.013	3.972 (0.041)	3.961 (0.052)
75[a]	4.316	4.095 (0.221)	4.202 (0.113)
76[a]	4.053	4.092 (−0.039)	4.079 (−0.026)
77[a]	4.370	4.213 (0.157)	4.261 (0.108)
78[a]	3.951	3.946 (0.005)	3.933 (0.017)
79[a]	3.977	3.867 (0.110)	3.920 (0.057)
80[a]	4.012	4.099 (−0.087)	4.012 (0.000)
81[b]	3.952	4.109 (−0.157)	3.974 (−0.022)
82[b]	4.013	4.114 (−0.101)	4.091 (−0.078)
83[b]	4.025	4.078 (−0.053)	3.942 (0.082)
84[b]	4.223	4.245 (−0.022)	4.316 (−0.093)
85[b]	4.290	4.208 (0.082)	4.270 (0.020)
86[b]	4.680	4.566 (0.114)	4.575 (0.104)
87[b]	4.360	4.378 (−0.018)	4.305 (0.055)
88[b]	3.980	3.898 (0.082)	3.861 (0.119)
89[b]	3.785	3.890 (−0.105)	3.879 (−0.095)
90[b]	4.016	3.908 (0.108)	3.948 (0.067)
91[b]	3.785	4.009 (−0.224)	3.931 (−0.147)
92[b]	4.360	4.459 (−0.099)	4.449 (−0.090)
93[b]	3.953	4.057 (−0.104)	4.088 (−0.135)
94[b]	3.979	4.019 (−0.040)	4.070 (−0.091)
95[b]	4.284	4.268 (0.016)	4.237 (0.046)
96[b]	4.277	3.988 (0.289)	4.019 (0.258)

[a] Member of the training set.
[b] Member of the external test set.

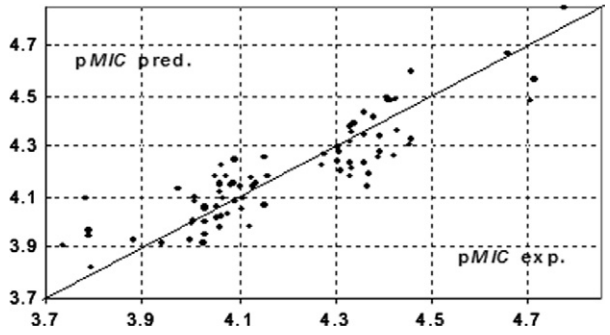

Figure 2. Predicted and experimental antifungal activities for Eq. 4 ($N = 80$).

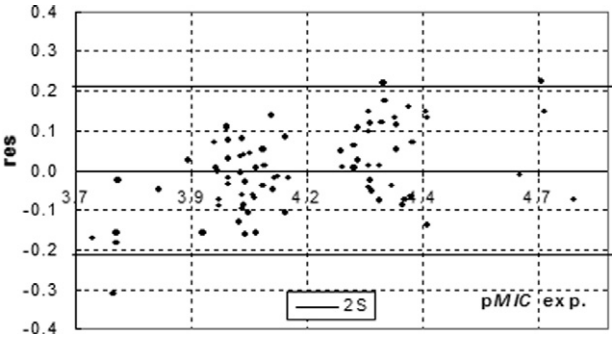

Figure 3. Dispersion plot of the residuals for Eq. 4 ($N = 80$).

Table 5. Correlation matrix for descriptors of Eq. 4 ($N = 80$)

Symbol	*HATSp*	*Mor13v*	*H8u*	*MATS6e*	*Mor29p*	*Mor27v*
HATSp	1	0.4618	0.3606	0.3793	0.2471	0.5000
Mor13v		1	0.2557	0.3829	0.0821	0.4361
H8u			1	0.2422	0.5089	0.4834
MATS6e				1	0.0542	0.3969
Mor29p					1	0.4127
Mor27v						1

ized Neural Networks hybridised with GA approach. From the non-linear point of view, Eq. 4 can be seen as a non-linear model for the optimal descriptors *H8u*, *Mor13v*, *MATS6e*, *Mor29p*, *Mor27v* and *HATSp* found by RM, with non-linear coefficients calculated as products of the remaining two descriptors in each term of the equation. The correlation matrix for this model is presented in Table 6 displaying low intercorrelations. The predictions shown in Table 4 lead to the straight line plotted in Figure 4 and reveal the presence of only one outlier heterocyclic in Figure 5 exceeding the value $2 \cdot S$ (compounds **39**).

As a final practical application of the non-linear relationship derived we predict the antifungal activity for some non-yet measured structures. To our knowledge, the 60 molecular structures listed in Table 7 have no measured values of p*MIC*. These structures were proposed by modifying the functional groups of some of the heterocyclics exhibiting the highest potency in the set, such as compounds **34**, **36**, **69**, **70** and **75**.

3. Conclusions

We derived an alternative structure–antifungal activity relationship employing linear and nonlinear regression models possessing a predictive performance comparable to those of Genetics Algorithms and Bayesian-Regularized Neural Networks hybridised with GA approaches. The best molecular descriptors appearing in the QSAR equation are able to reflect the distribution of the atomic Sanderson electronegativities, volumes and polarizabilities influencing the variation of the antifungal activity in the 96 heterocyclics under investigation.

4. Method

4.1. Data set

The 96 heterocyclic ring derivatives collected[11] from the literature consist of 2,5,6-trisubstituted benzoxazoles, 2,5-disubstituted benzimidazoles, 2-substituted benzothiazoles and 2-substituted oxazolo(4,5-b)pyridines. The general structure for these compounds is displayed in Figure 1. All in vitro inhibitory activities against *Candida albicans* species are expressed as p*MIC*[M] = −log(*MIC*[M]), with the quantity *MIC*(M) representing the minimum inhibitory concentration in molar units. A twofold serial dilution technique was employed to carry out the activity essays. Table 1 includes the experimental potencies for both the training and test sets of heterocyclics as obtained from the application of the k-Means Clustering technique.

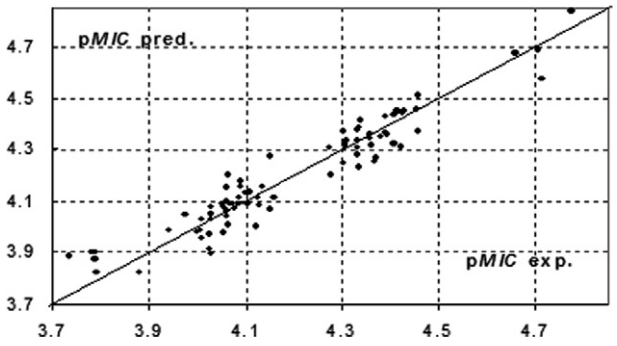

Figure 4. Predicted and experimental antifungal activities for Eq. 7 ($N = 80$).

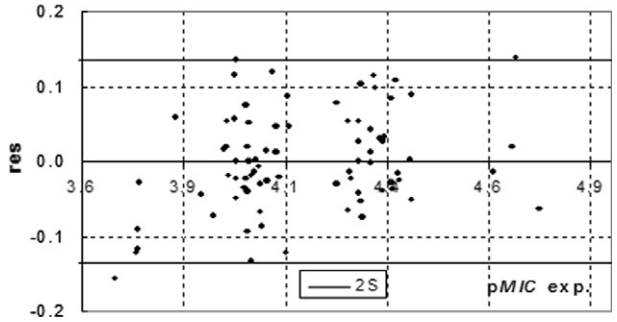

Figure 5. Dispersion plot of the residuals for Eq. 7 ($N = 80$).

4.2. Descriptors calculation

The structures of the compounds are first pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in Hyperchem 6.03,[27] and the resulting geometries are further refined by means of the semiempirical method PM3 (Parametric Method-3). We chose a gradient norm limit of 0.01 kcal/Å for the geometry optimization. We derived a pool of $D = 1202$ theoretical descriptors from the software Dragon 5.0[20] including several types of variables: constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology

Table 6. Correlation matrix for descriptors of the RM-CP QSAR model of Eq. 7 ($N = 80$)

symbol	*HATSp*	*Mor13v*	*H8u*	*MATS6e*	*Mor29p*	*Mor27v*
HATSp	1	0.3419	0.3636	0.2656	0.2755	0.4637
Mor13v		1	0.2634	0.2437	0.004898	0.3061
H8u			1	0.004254	0.5048	0.2906
MATS6e				1	0.1745	0.1597
Mor29p					1	0.3496
Mor27v						1

Table 7. Predicted in vitro antifungal activities for unknown compounds

Compound	X	Y	Z	R	R_1	R_2	p*MIC*[M]
1	CH	O	CH_2	F	NO_2	H	4.491
2	CH	O	CH_2	NH_2	NO_2	H	4.618
3	CH	O	CH_2	$C(CH_3)_3$	NO_2	H	4.266
4	CH	O	CH_2	$NHCOCH_3$	NO_2	H	4.586
5	CH	O	CH_2	CH_3	NO_2	H	4.688
6	CH	O	CH_2	C_2H_5	NO_2	H	4.626
7	CH	O	CH_2	$N(CH_3)_2$	NO_2	H	4.465
8	CH	O	CH_2	$NHCH_3$	NO_2	H	4.525
9	CH	O	CH_2	OC_2H_5	NO_2	H	4.500
10	CH	O	CH_2	$COOCH_3$	NO_2	H	4.559
11	CH	O	CH_2S	H	F	NO_2	4.173
12	CH	O	CH_2S	H	NH_2	NO_2	4.243
13	CH	O	CH_2S	H	$C(CH_3)_3$	NO_2	3.839
14	CH	O	CH_2S	H	$NHCOCH_3$	NO_2	3.939
15	CH	O	CH_2S	H	CH_3	NO_2	4.238
16	CH	O	CH_2S	H	C_2H_5	NO_2	4.079
17	CH	O	CH_2S	H	$N(CH_3)_2$	NO_2	4.100
18	CH	O	CH_2S	H	$NHCH_3$	NO_2	3.993
19	CH	O	CH_2S	H	OC_2H_5	NO_2	4.233
20	CH	O	CH_2S	H	H	NO_2	4.231
21	CH	O	CH_2S	H	OCH_3	NO_2	4.286
22	CH	O	CH_2S	H	Br	NO_2	4.088
23	CH	O	CH_2S	H	NO_2	NO_2	4.596
24	CH	O	CH_2S	H	$COOCH_3$	NO_2	4.551
25	CH	O	CH_2S	H	F	H	4.151
26	CH	O	CH_2S	H	NH_2	H	4.126
27	CH	O	CH_2S	H	$C(CH_3)_3$	H	4.111
28	CH	O	CH_2S	H	$NHCOCH_3$	H	4.231
29	CH	O	CH_2S	H	Cl	H	4.166
30	CH	O	CH_2S	H	C_2H_5	H	4.114
31	CH	O	CH_2S	H	$N(CH_3)_2$	H	3.850
32	CH	O	CH_2S	H	$NHCH_3$	H	3.996
33	CH	O	CH_2S	H	OC_2H_5	H	4.022
34	CH	O	CH_2S	H	OCH_3	H	4.062
35	CH	O	CH_2S	H	Br	H	4.344
36	CH	O	CH_2S	H	$COOCH_3$	H	4.269
37	CH	NH	CH_2O	F	Cl	H	3.847
38	CH	NH	CH_2O	NH_2	Cl	H	4.185
39	CH	NH	CH_2O	$C(CH_3)_3$	Cl	H	4.366
40	CH	NH	CH_2O	$NHCOCH_3$	Cl	H	3.673
41	CH	NH	CH_2O	CH_3	Cl	H	3.947
42	CH	NH	CH_2O	C_2H_5	Cl	H	4.216
43	CH	NH	CH_2O	$N(CH_3)_2$	Cl	H	4.103
44	CH	NH	CH_2O	$NHCH_3$	Cl	H	3.815
45	CH	NH	CH_2O	OC_2H_5	Cl	H	3.821
46	CH	NH	CH_2O	OCH_3	Cl	H	3.858
47	CH	NH	CH_2O	Br	Cl	H	3.895
48	CH	NH	CH_2O	NO_2	Cl	H	4.108
49	CH	NH	CH_2O	$COOCH_3$	Cl	H	4.034
50	CH	O	CH_2	F	Cl	H	3.898
51	CH	O	CH_2	NH_2	Cl	H	4.010
52	CH	O	CH_2	$C(CH_3)_3$	Cl	H	4.335
53	CH	O	CH_2	$NHCOCH_3$	Cl	H	4.223
54	CH	O	CH_2	CH_3	Cl	H	3.997
55	CH	O	CH_2	Cl	Cl	H	3.900
56	CH	O	CH_2	C_2H_5	Cl	H	3.934
57	CH	O	CH_2	$N(CH_3)_2$	Cl	H	4.123
58	CH	O	CH_2	$NHCH_3$	Cl	H	4.236
59	CH	O	CH_2	OC_2H_5	Cl	H	4.040
60	CH	O	CH_2	$COOCH_3$	Cl	H	4.017

and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk counts, BCUT descriptors, 2D-Autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups and atom-centred fragments. We excluded from our calculations the empirical and

property-based descriptors provided by the software, and also added more quantum-chemical type of descriptors to the pool such as homo and lumo energies, and homo–lumo gap.

4.3. Model search

In our calculations we employ the computer algebra systems Derive[28] and Maple.[29] It is our purpose to search a large set **D** containing D descriptors for an optimal subset of d ones that minimize the standard deviation S. More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where **d** is a point in a space of $D!/[d!(D-d)!]$ ones. A full search (FS) of optimal variables requires $D!/[d!(D-d)!]$ linear regressions. The Forward Stepwise Regression[22] (FSR) consists of a step by step addition of descriptors to the model, initially without any independent variable present in the regression, until there is no variable left outside the equation that minimizes its S. Here we define the standard deviation as follows:

$$S = \frac{1}{(N-d-1)} \sum_{i=1}^{N} \mathrm{res}_i^2, \tag{5}$$

where N is the number of molecules in the training set and res_i is the residual for molecule i (difference between the experimental and predicted property). The FSR sacrifices accuracy for a much smaller number of linear regressions than a FS.

Some time ago we proposed the Replacement Method (RM)[23–26] that produces linear QSPR models that are quite close to the FS ones with much less computational work. The RM gives models with better statistical parameters than the FSR and similar ones to the more elaborated Genetics Algorithms.[18] This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$. The RM consists of the following simple steps: we first choose a set **d** at random and perform a linear regression. Then we choose one of the descriptors of this set, say X_i, and replace it with each of the D descriptors of the pool $\mathbf{D} = \{X_1, X_2, \ldots, X_D\}$, $D \ll d$ (except itself) keeping the best resulting set (i.e., that with smallest S). Since one can start replacing any of the d descriptors in the initial model, then there will be d possible paths. Choose the variable in the resulting model with the greatest relative error in its coefficient (omitting the one replaced in the previous step) and replace it with all the D descriptors (except itself) keeping again the best set. Replace all the remaining variables in the same way by omitting those replaced in previous steps. When finishing, start again with the variable having the greatest relative error in the coefficient and repeat the whole process. Repeat this process as many times as necessary until the set of descriptors remains unchanged. At the end, we have the best model for the path i. Proceed in exactly the same way for all possible paths $i = 1, 2, \ldots, d$, compare the resulting models and keep the best one. Our numerical experiments show that in this way one obtains a model almost as good as the best FS one with much less than $D!/[d!(D-d)!]$ linear regressions when this combinatorial number is large. We may carry out this calculation for $d = 1, 2, \ldots$ in order to obtain the overall best model.

The theoretical validation practiced in all our models is based on the leave-more-out cross validation procedure $(l - n\% - o)$,[30] with $n\%$ representing the percentage of molecules removed from the training set. This percentage depends simultaneously upon the number of compounds N (one cannot remove many molecules from the training set if a small sample is analysed as the normality condition of the fitted data has to be obeyed) and their structural diversity (if the molecules are structurally very different, more compounds would have to be removed from the set for checking the predictive performance of the model). In this study, we consider $n\% = 10\%$ for analysing $N = 80$ derivatives that include a common pattern (see Fig. 1). The number of cases for random removal in $l - n\% - o$ is 90,000.

The quality of the final optimized equations obtained via the two approaches FSR and RM is compared by means of two different criteria: the Akaike criterion and the Kubinyi function. These represent alternative criteria that are able to compare the quality of two different models and take into account in a single number the number of variables of the equation, providing further statistical details. Akaike's information criterion (AIC)[31,32] considers the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit:

$$AIC = \left(\sum_{i=1}^{N} \mathrm{res}_i^2 \right) \cdot \frac{(N+d+1)}{(N-d-1)^2}. \tag{6}$$

Therefore, the model that produces the minimum AIC value should be considered potentially the most useful. The Kubinyi function (FIT)[33,34] closely relates to the Fisher ratio (F), although the main disadvantage of F is its sensitivity to changes in d, if d is small, and its lower sensitivity if d is large. The FIT criterion has a low sensitivity towards changes in d values, as long as they are small numbers, and a substantially increasing sensitivity for large d values. The following equation is employed:

$$FIT = \frac{R^2 \cdot (N-d-1)}{(N+d^2) \cdot (1-R^2)}, \tag{7}$$

where R is the correlation coefficient. The best model will present the highest value of the FIT function.

In order to improve the predictive performance of the linear regression model found that includes the best subset **d**, we derive new QSAR relationships as follows: we add to the pool **D** with $D = 1202$ new variables (of the form $X_i \cdot X_j$) calculated by performing cross-products between each of the d optimal descriptors of the model and descriptors from the set **D**. This leads to a pool **DD** composed of $DD = D \cdot d + D$ descriptors (in the present study, $DD = 8414$). After that, we apply the RM on **DD** and find a new model with an optimal new subset $\mathbf{d}_{DD}$. As a further step, we also consider a

pool **DDD** obtained by adding to the pool **D** the subset $\mathbf{d}_{DD}$ found previously and new variables (now of the form $X_i \cdot X_j \cdot X_k$) calculated by performing cross-products between descriptors from $\mathbf{d}_{DD}$ and descriptors from **D**, thus leading to $DDD = D \cdot d + D + d$ descriptors ($DDD = 8420$ here). Therefore, we apply RM on **DDD** and find a new model with the optimal subset $\mathbf{d}_{DDD}$. This type of model is denoted as RM-CP throughout the article, meaning that optimal cross-products among descriptors are selected with the Replacement Method optimization technique. In this way we expect that the RM-CP algorithm allows establishing non-linear relationships between the optimal (linear) descriptors **d** searched with the RM algorithm and the property under investigation.

We employ the stepwise orthogonalization procedure introduced several years ago by Randic[35–37] as a way of improving the statistical interpretation of the model built by interrelated indices. From our point of view, the co-linearity of the molecular descriptors should be as low as possible, because the interrelatedness among the different descriptors can lead to highly unstable regression coefficients, which makes it impossible to know the relative importance of an index and underestimates the utility of the regression coefficients of the model. The crucial step of the orthogonalization process is the choice of an appropriate order of orthogonalization, which in the present analysis is the order that maximises the correlation between each orthogonal descriptor and the observed antifungal activity.

Acknowledgments

P.R.D. thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) institution. We also thank Professor Gustavo P. Romanelli for his helpful comments on this work.

References and notes

1. St-Georgiev, V. *Curr. Drug Targets* **2000**, *1*, 261.
2. Rex, J. H.; Walsh, T. J.; Sobel, J. D.; Filler, S. G.; Pappas, P. G.; Dismukes, W. E.; Edwards, J. E. *Clin. Infect. Dis.* **2000**, *30*, 662.
3. Georgopapadakou, N. H. *Curr. Opin. Microbiol.* **1998**, *1*, 547.
4. Rees, J. R.; Pinner, R. W.; Hajjeh, R. A. *Clin. Infect. Dis.* **1998**, *27*, 1138.
5. Polak, A. *Mycoses* **1999**, *42*, 355.
6. Fostel, J. M.; Lartey, P. A. *Drug Discov. Today* **2000**, *5*, 25.
7. Tafi, A.; Costi, R.; Botta, M.; Di Santo, R.; Corelli, F.; Massa, S.; Ciacci, A.; Manetti, F.; Artico, M. *J. Med. Chem.* **2002**, *45*, 2720.
8. Chan, J. H.; Hong, J. S.; Kuyper, L. F.; Baccanari, D. P.; Joyner, S. S.; Tansik, R. L.; Boytos, C. M.; Rudolph, S. K. *J. Med. Chem.* **1995**, *38*, 3608.
9. Elnima, E. I.; Zubair, M. U.; Al-Badr, A. A. *Antimicrob. Agents Chemother.* **1981**, *19*, 29.
10. Göker, H.; Kus, C.; Boykin, D. W.; Yildizc, S.; Altanlarc, N. *Bioorg. Med. Chem.* **2002**, *10*, 2589.
11. Yildiz-Oren, I.; Yalcin, I.; Aki-Sener, E.; Ucarturk, N. *Eur. J. Med. Chem.* **2004**, *39*, 291.
12. Yalcin, I.; Oren, I.; Temiz, O.; Sener, E. A. *Acta Biochim. Pol.* **2000**, *47*, 481.
13. Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
14. García-Domenech, R.; Ríos-Santamarina, I.; Catalá, A.; Calabuig, C.; del Castillo, L.; Gálvez, J. *J. Mol. Struct. Theochem.* **2003**, *624*, 97.
15. Hasegawa, K.; Deushi, T.; Yaegashi, O.; Miyashita, Y.; Sasaki, S. *Eur. J. Med. Chem.* **1995**, *30*, 569.
16. Mghazli, S.; Jaouad, A.; Mansour, M.; Villemin, D.; Cherqaoui, D. *Chemosphere* **2001**, *43*, 385.
17. Caballero, J.; Fernández, M. *J. Mol. Model.* **2006**, *12*, 168.
18. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
19. Mackay, D. J. C. *Neural Comput.* **1992**, *4*, 415.
20. DRAGON 5.0 Evaluation Version, <http://www.disat.unimib.it/chm>.
21. McFarland, J. W.; Gans, D. J. In *Methods and Principles in Medicinal Chemistry*; Mahnnhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; Weinheim: VCH, 1995; Vol 2.
22. Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1981.
23. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. *Chem. Phys. Lett.* **2005**, *412*, 376.
24. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179.
25. Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. *Bioorg. Med. Chem.* **2006**, *16*, 5876.
26. Helguera, A. M.; Duchowicz, P. R.; Cabrera Pérez, M. A.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. *Chemometr. Intell. Lab.* **2006**, *81*, 180.
27. HYPERCHEM 6.03 Hypercube, <http://www.hyper.com>.
28. Derive 5: http://education.ti.com/us/product/software/derive/features/features.html.
29. Maple 7: http://www.maplesoft.com.
30. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Model.* **2003**, *43*, 579.
31. Akaike, H. In *Second International Symposium on Information Theory*; Petrov, B. N., Csáki, F., Ed.; Akademiai Kiado: Budapest, 1973.
32. Akaike, H. *IEEE Trans. Automat. Control* **1974**, *AC-19*, 716.
33. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393.
34. Kubinyi, H. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285.
35. Klein, D. J.; Randic, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quant. Chem.* **1991**, *63*, 215.
36. Randic, M. *J. Chem. Inf. Model.* **1991**, *31*, 311.
37. Randic, M. *New J. Chem.* **1991**, *15*, 517.
38. Schuur, J.; Selzer, P.; Gasteiger, J. *J. Chem. Inf. Model.* **1996**, *36*, 334.
39. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Model.* **2002**, *42*, 693.
40. Moran, P. A. P. *Biometrika* **1950**, *37*, 17.