CROATICA CHEMICA ACTA
CCACAA **79** (3) 483–488 (2006)
ISSN-0011-1643
CCA-3114
Original Scientific Paper

QSAR Modeling of Antifungal Activity of Some Heterocyclic Compounds

Oleg Ursu,[a,*] Adina Costescu,[a] Mircea V. Diudea,[a] and Bazil Parv[b]

[a]*Faculty of Chemistry and Chemical Engineering, Babeş-Bolyai University, 3400 Cluj, Romania*

[b]*Faculty of Mathematics and Computer Science, Babeş-Bolyai University, 3400, Cluj, Romania*

RECEIVED MAY 10, 2005; REVISED JUNE 9, 2006; ACCEPTED JUNE 28, 2006

QSAR analysis of a set of benzoxazoles, benzimidazoles, oxazolo[4,5-*b*]pyridines and benzothiazoles, showing growth inhibitory activity against *Candida albicans*, was performed using a multiple regression procedure. Topological indices (TIs) and principal component analysis (PCA) on TIs were used in modeling antifungal activity. Selection of TIs relevant to developing QSAR models was made using the largest PC factor loading scores. Correlation coefficient 0.97 obtained in the validation procedure indicated the excellent quality of the derived QSAR models.

Keywords
antifungal activity
QSAR/QSPR
PCA

INTRODUCTION

Mycosis diseases are affections caused by body infestat ion with pathogenic mycosis parasites of pathogenic potential. Some of these fungi live in nature as saprophytes (exogenetic saprophytes) or in the human body or animals (endogenetic saprophytes) and they accidentally become parasites.

Affections produced by mycosis parasites are widely spread worldwide, with predominance in some areas. Superficial mycosis affections are frequently encountered in our country. Candidiasis, an infection caused by *Candida*, a yeast-like fungus, is such an affection.

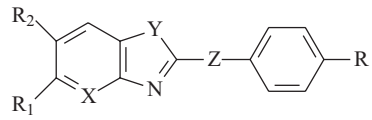
Candida species are ubiquitous fungi, existing as normal body flora. Candidiasis infection causes a variety of diseases, ranging from superficial disorders such as diaper rash to invasive, rapidly fatal infections in immunocompromized hosts. *Candida albicans* is commonly responsible for candidiasis, but others, such as *Candida tropicalis, Candida parapsilosis, Candida guilliermondi,* and *Torulopsis glabrata* are also causative organisms.[1]

Candidiasis affects a wide variety of organ systems. In immunocompetent persons, any warm, moist part of the body exposed to the environment is susceptible to infection. In immunocompromized patients, systemic diseases such as myocarditis and hepatosplenic abscess occur and antifungal therapy should be started immediately after necessary cultures have been obtained from all suspected sites of infection.

Antifungals — Amphotericin B, fluconazole, ketoconazole, and nystatin are the drugs most commonly used to treat candidiases. Most fungi are completely resistant to the action of antimicrobial drugs. Only a few substances have been discovered that exert an inhibitory effect

* Author to whom correspondence should be addressed. (E-mail: OUrsu@salud.unm.edu)
Current address: Division of Biocomputing, MSC11 6145, Research Incubator Building, Suite 190, 2703 Frontier, NE, 1 University of New Mexico, Albuquerque, NM 87131, USA

TABLE I. The structure and *in vitro* antifungal activity of compounds **1–68** against *C. albicans*



	X	Y	Z	R	R_1	R_2	log(1/*C*)
1	CH	O	–	H	H	H	3.892
2	CH	O	–	$C(CH_3)_3$	H	H	4.001
3	CH	O	–	NH_2	H	H	3.924
4	CH	O	–	$NHCH_3$	H	H	3.952
5	CH	O	–	C_2H_5	Cl	H	4.013
6	CH	O	–	$NHCOCH_3$	Cl	H	4.059
7	CH	O	–	$NHCH_3$	Cl	H	4.015
8	CH	O	–	Cl	Cl	H	4.024
9	CH	O	–	NO_2	Cl	H	4.04
10	CH	O	–	H	NO_2	H	4.282
11	CH	O	–	CH_3	NO_2	H	4.308
12	CH	O	–	$C(CH_3)_3$	NO_2	H	4.375
13	CH	O	–	NH_2	NO_2	H	4.31
14	CH	O	–	Cl	NO_2	H	4.342
15	CH	O	–	Br	NO_2	H	4.406
16	CH	O	–	C_2H_5	NH_2	H	3.979
17	CH	O	–	F	NH_2	H	3.96
18	CH	O	–	$N(CH_3)_2$	CH_3	H	4.005
19	CH	O	–	CH_3	CH_3	H	3.95
20	CH	O	–	C_2H_5	CH_3	H	3.977
21	CH	O	–	OCH_3	CH_3	H	3.98
22	CH	O	–	F	CH_3	H	3.958
23	CH	O	–	$NHCOCH_3$	CH_3	H	4.027
24	CH	O	–	$NHCH_3$	CH_3	H	3.979
25	CH	O	–	$N(CH_3)_2$	CH_3	H	4.004
26	N	O	–	CH_3	H	H	4.225
27	N	O	–	C_2H_5	H	H	4.253
28	N	O	–	OCH_3	H	H	4.257
29	N	O	–	OC_2H_5	H	H	4.283
30	N	O	–	NH_3	H	H	4.227
31	N	O	–	NO_2	H	H	4.285
32	CH	O	CH_2	H	H	H	4.223
33	CH	O	CH_2	OCH_3	H	H	4.282
34	CH	O	CH_2	Cl	H	H	4.29
35	CH	O	CH_2	NO_2	H	H	4.308
36	CH	O	CH_2	H	Cl	H	4.29
37	CH	O	CH_2	OCH_3	Cl	H	4.34
38	CH	O	CH_2	Br	Cl	H	4.41
39	CH	O	CH_2	NO_2	Cl	H	4.363
40	CH	O	CH_2	H	NO_2	H	4.609
41	CH	O	CH_2	OCH_3	NO_2	H	4.657
42	CH	O	CH_2	Br	NO_2	H	4.725
43	CH	O	CH_2	Cl	NO_2	H	4.664
44	CH	O	CH_2	NO_2	NO_2	H	4.68
45	CH	O	CH_2O	H	CH_3	H	3.98
46	CH	O	CH_2O	H	H	NO_2	3.732
47	CH	O	CH_2O	H	CL	NO_2	3.785
48	CH	O	CH_2O	Cl	CL	NO_2	3.831
49	CH	O	CH_2S	H	NO_2	H	4.359
50	CH	O	CH_2S	H	CH_3	H	4.009
51	N	O	CH_2O	H	H	H	4.26
52	N	O	CH_2O	Cl	H	H	4.319
53	CH	NH	CH_2O	Cl	CH_3	H	4.037
54	CH	NH	CH_2S	H	NO_2	H	4.358
55	CH	NH	CH_2S	H	CH_3	H	4.009
56	CH	O	CH_2O	H	$COOCH_3$	H	4.054
57	CH	O	CH_2O	Cl	$COOCH_3$	H	4.104
58	CH	NH	CH_2O	Cl	$COOCH_3$	H	4.102
59	CH	NH	CH_2S	H	$COOCH_3$	H	4.076
60	CH	O	C_2H_4	H	NO_2	H	4.331
61	N	O	C_2H_4	H	H	H	4.253
62	CH	O	–	Br	NH_2	H	4.11
63	CH	O	CH_2	Br	H	H	4.36
64	CH	O	CH_2O	Cl	H	H	4.016
65	CH	NH	CH_2O	H	NO_2	H	4.283
66	CH	NH	CH_2O	Cl	H	H	4.015
67	CH	NH	CH_2S	H	Cl	H	4.041
68	CH	NH	C_2H_4	H	H	H	4.078

on the fungi pathogenic to man, and most of these are relatively toxic.[2]

In this context, the synthesis of some novel derivatives of benzoxazoles, benzimidazoles, oxazolo[4,5-*b*]pyridines and benzothiazoles over the last few years[3–7] was welcome.

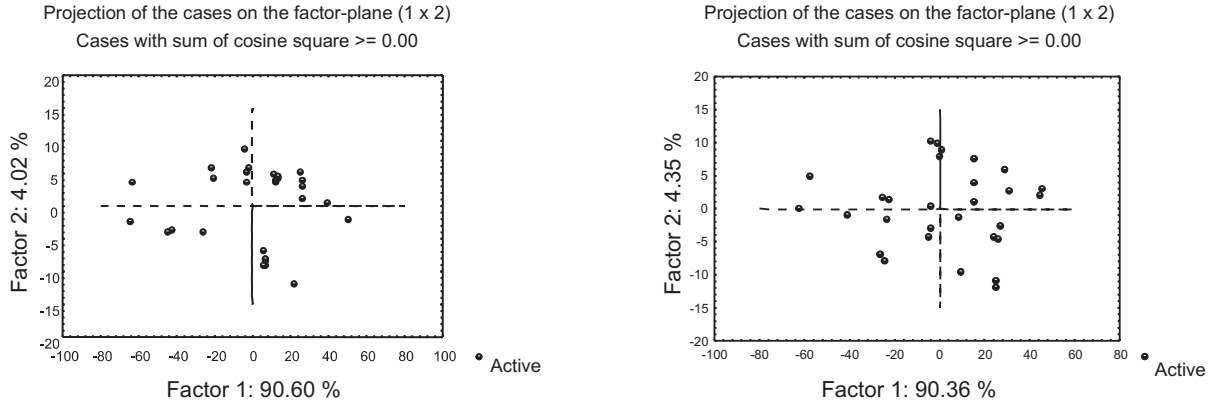
DATA SET

Compounds **1–68**, included in this QSAR analysis, have shown antifungal activity at MIC (minimum inhibitory concentration) values between 6.25–50 g/ml, against *Candida albicans*, comparable to the activity of commonly used antifungals, such as Clotrimazole, Oxiconazole and Haloprogin.[5]

In this study, QSAR analysis of antifungal benzoxazoles, benzimidazoles and oxazolo[4,5-*b*]pyridines **1–68** is presented. Table I lists the formulas and activity data, log(1/*C*), for the above mentioned structures.

DATA ANALYSIS

QSAR analysis consists of the following steps: (i) structure optimization using the semiempirical method PM3; (ii) calculation of molecular descriptors; (iii) classification using the Good similarity index; (iv) principal component analysis (PCA); (v) correlation analysis using step-forward selection of descriptors; (vi) multiple regression analysis by selected descriptors from PCA; (vii)



Subset 1
Quality of representation 100 %
No. of active cases 29
No. of components 33

Subset 2
Quality of representation 100 %
No. of active cases 30
No. of components 33

Figure 1. Principal component analysis, projection on factor coordinates.

evaluation of the significance level of the model; (viii) validation of the model, and (ix) interpretation of the results.

Our structures were optimized by using the semi empirical PM3 Hamiltonian, available in HyperChem.

A large pool of descriptors, provided by the TOPO-CLUJ software package,[8] was calculated for every molecule. The descriptors list includes: (i) topological descriptors (Wiener,[9] Randić,[10] *etc.*); (ii) constitutional indicator descriptors (*e.g.*, absence or presence of certain functional groups in the molecular structure); (iii) quantum calculated descriptors (HOMO energy – highest occupied molecular orbital, and LUMO energy – lowest occupied molecular orbital, dipole moment, total electronic energy); (iv) force field parameters – *F* (field effect) for electronic description and (v) steric descriptors (L and B4) for substituents R and R1.[11–14]

Similarity expresses the relatedness of two molecules, with a large number if their molecular descriptions are closely related and with a number approaching zero in case they are unrelated. The Good index is one of the most used coefficients of similarity:[13–16]

$$\mathrm{SIM(A,B)}_k = 1 - \left(\frac{|I_{Ak} - I_{Bk}|}{\max|I_{Ak}, I_{Bk}|} \right) \qquad (1)$$

where I_{Ak} refers to the k indices for compound A and I_{Bk} to the k indices for compound B. It takes values in the range [0, 1]. Considering all the n descriptors, the index becomes:

$$\mathrm{SIM(A,B)} = \frac{1}{n} \sum_{k=1}^{n} w_k \, \mathrm{SIM(A,B)}_k \qquad (2)$$

with w_k being the weighting factor for descriptor k.

The initial data set (68 antifungal compounds) was split into 3 subsets: two training subsets (n = 29 for subset 1 and n = 30 for subset 2), clustering the ordered BA values (for the regression calibration) and one external (randomly selected, n = 9) validation subset (see the left column of Table II). The quality of prediction depends on correct classification of presumably »unknown« compounds with respect to the training subsets.

The values in Table II are Good index frequencies, calculated according to Eq. (2), for each compound in the external validation subset with respect to each structure in both training subsets. Boldface values indicate the maximal frequency and thus the subset to which each compound of the validation subset belongs.

Principal components analysis, PCA, is a very powerful statistical technique useful for reducing the noise of the data set and eliminating uncorrelated variables. Due to the high complexity of interactions between the receptor molecule and potential inhibitor molecules, it is often difficult to model BA using simple linear regression mo-

TABLE II. Good frequencies for compounds in the external validation subset[(a),(b)]

Compd.	Subset 1	Subset 2
	Σ Good frequencies	Σ Good frequencies
a17	**0.7269**	0.6503
a27	0.6792	**0.6816**
a33	0.6175	**0.6222**
a35	0.6663	**0.7189**
a47	**0.6124**	0.5725
a53	**0.7599**	0.6821
a58	**0.6429**	0.5574
a65	0.7423	**0.7729**
a07	**0.7789**	0.7444

[(a)] The values are Good index frequencies, calculated according to Eq. (2).
[(b)] Boldface values indicate the maximal frequency.

TABLE III. Statistics for derived QSAR models for antifungal compounds in Table I[a]

	b	Std. error	*t*	*p*-level
		Subset 1		
Intercept	1.845E+00	2.237E-01	8.2468	2.407E-07
WkOp[Detour U Adjacency]	–6.402E-04	7.880E-05	–8.1252	2.953E-07
PDS[Sh[U]]	1.092E-03	2.092E-04	5.2182	6.959E-05
Distance	1.296E-01	1.011E-02	12.8173	3.649E-10
PDS4[Sh[Detour]]	–1.941E-04	2.597E-05	–7.4746	9.087E-07
PDS4[LM[Charge]]	–1.648E-02	2.389E-03	–6.8972	2.584E-06
PRD^2S[LM[Charge]]	–2.582E-02	4.235E-03	–6.0984	1.182E-05
PDS2[LM[Charge]]	1.063E-01	7.039E-03	15.1085	2.761E-11
		Subset 2		
Intercept	4.171E+00	1.224E-01	34.0747	7.744E-22
PRD^2S[Sh[CjMax[Covalent radius]]]	3.553E-03	5.744E-04	6.1859	2.161E-06
PDS8[LM[Charge]]	2.822E-02	2.651E-03	10.6459	1.428E-10
PDS3[Sh[Detour]]	–6.179E-04	8.158E–05	–7.5745	8.190E-08
PDS6[Sh[Detour]]	2.018E-04	9.255E-05	2.1802	3.928E-02

[a] *b* – regression coefficients of the variables, Std. error – standard deviation of a mean, *t* – test for independent samples, *p*-level – probability of error.

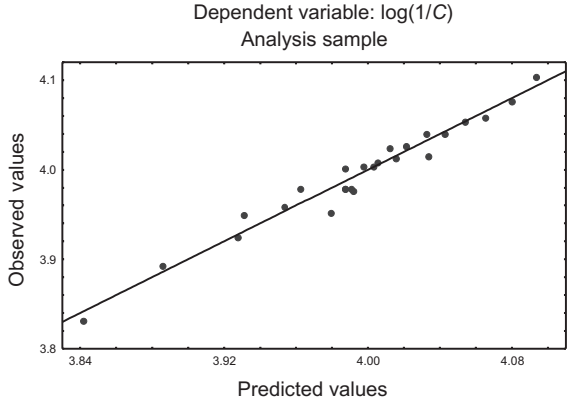
dels. In this study, we used a large pool of descriptors (864) to derive the QSAR model; PCA will help reduce the data drastically (see Figure 1).[17]

Since eigenvalues are sorted in descending order, the first PC will describe the greatest amount of data variance. The second PC will describe the greatest amount of data variance in a direction orthogonal to the first PC. The first two PCs account for more than 94 % (94.62 % for subset 1 and 94.71 % for subset 2, respectively) of data variance, indicating that data variance can be well described in only two dimensions. [17]

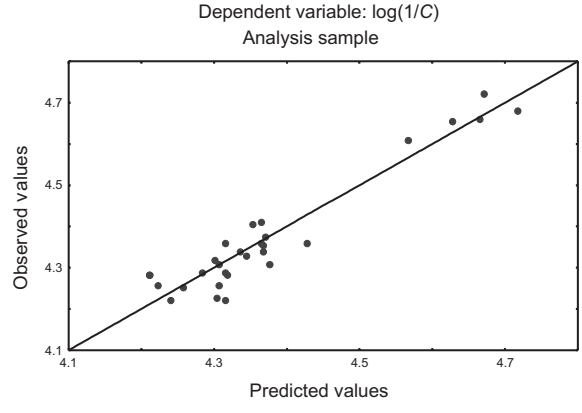
Loading plots can be used to evaluate the relevant descriptors. A high loading value indicates that the PC is aligned in a direction close to the original descriptor response. Selection of descriptors, contributing highly to data variance, can be accomplished by examining the loading plots or tables. Additionally, the relation of descriptors to each other (*i.e.*, co-linearity) can be explored.

We selected 6 PCs in each training subsets, for a representation quality of 99 %, and, for each PC, 5 descriptors with the greatest loadings as independent variables. Thus, we greatly reduced the searching space (descriptor's pool) by one order of magnitude (from 864 to 22 for subset 1 and from 864 to 24 for subset 2). In the next step, automated selection of descriptors was performed by the forward stepwise procedure.

The derived QSAR multivariate models are presented in Table III (b – regression coefficients of the vari-



R^2 = 0.9635; s = 0.0132; F = 64.222;
p = 5.49E–11; n = 29



R^2 = 0.9095; s = 0.1223; F = 60.129;
p = 3.71E–12; n = 30

Figure 2. Predicted *vs.* observed values for the training subsets 1 and 2 (s – square root of the mean square for error, *F* – the ratio of the model mean square to the error mean square, *p* – probability of error).

TABLE IV. Prediced and residual values in the validation set, $n = 9$

Compd.	$BA_{obs.}$	$BA_{pred.}$	Residual	CV/%
a17	4.253	4.219	0.034	0.799
a27	4.282	4.265	0.017	0.397
a33	4.283	4.217	0.066	1.541
a35	4.308	4.312	–0.004	0.093
a47	3.785	3.804	–0.019	0.502
a53	3.96	3.949	0.011	0.278
a58	4.015	4.024	–0.009	0.224
a65	4.037	4.075	–0.038	0.941
a07	4.102	4.138	–0.036	0.878

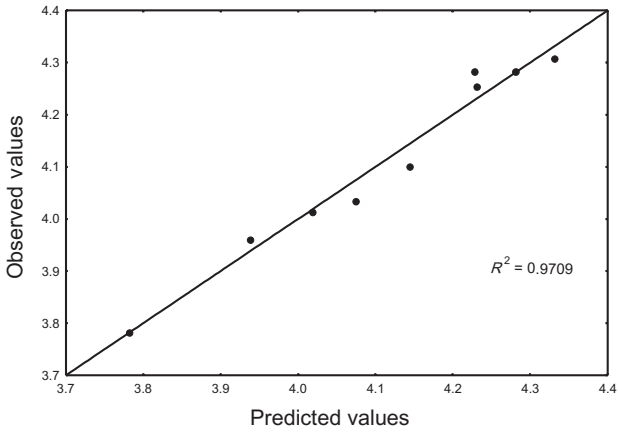


Figure 3. Graphical representation of data given in Table IV.

ables; Std. error – standard deviation of a mean; *p*-level – probability of error) and plots in Figure 2 (for subset 1 ($R^2 > 0.96$) and for subset 2 ($R^2 > 0.90$)). The square root of the mean square for error (*s*), the ratio of the model mean square to the error mean square (*F*), and probability of error (*p*) were also recorded.

Both models were tested for statistical significance, by a random exchange of BA column entries, within each subset; a dramatic drop of the correlation coefficient proved that no chance correlation occurred in our models.

Validation of the models, on the external validation dataset (Table IV, Figure 3), indicated a good prediction ability ($R^2 > 0.97$) of the observed BA. Thus, the derived models can be used in prediction of BA for new compounds, in homologous series.

Note that the published[18] QSAR models, based on certain physicochemical properties or structural indicator variables, have good correlation coefficients but lower prediction ability ($R^2 = 0.67$).

CONCLUSION

In this paper, we have presented a QSAR model for a set of antifungal compounds. Our attention was focused on model development and validation. Statistical tools such as PCA and stepwise regression analysis have been employed. Another important aspect taken into consideration was the correct splitting of the data set into training and validation subsets, which can considerably affect the results. The derived models indicated good prediction ability, thus being useful in evaluation of BA for new or not screened compounds, in homologous series. The obtained high quality QSAR models can be used in large database mining to find compounds with antifungal activity. Note that during the lead discovery processes, in high throughput screening of large databases, such QSAR models can be employed for data clustering, thus avoiding screening for antifungal activity of irrelevant structures. This could represent an early step in guiding the synthetic chemistry of a drug, notably of antifungal compounds.

We stress here that complementary studies on the toxicity of such compounds are needed in order to avoid potent antifungal but toxic structures, or at least to balance between the desired and undesired properties within a set of biologically active designed compounds. More elaborate modeling and drug design, aimed at gaining the specific antifungal activity (not easily achievable in others, because of the eukaryotic structure of *Candida*) are also required.

REFERENCES

1. http://www.emedicine.com/emerg/byname/candidiasis.htm
2. F. H. Meyers, E. Jawetz, and A. Goldfien, *Review of Medical Pharmacology*, Lange Medical Pub., California, 1976.
3. I. Yalçìn, I. Ören, E. Sener, A. Akìn, and N. Uçartürk, *Eur. J. Med. Chem.* **27** (1992) 401–406.
4. I. Yalçìn, E. Sener, T. Özden, S. Özden, and A. Akìn, *Eur. J. Med. Chem.* **25** (1990) 705–708.
5. I. Ören, Ph. D. Thesis, University of Ankara, 1995.
6. I. Ören, Ö. Temiz, I. Yalçìn, A. Sener, and N. Altanlar, *Eur. J. Pharm. Sci*. **7** (1998) 153–160.
7. A. Mrozek, H. Trzezwinska, J. Karolak-Wojciechowska, I. Yalçìn, and E. Tener, *Pol. J. Chem.* **73** (1999) 625–633.
8. M. V. Diudea and O. Ursu, TOPOCLUJ (Copyright Babeş-Bolyai University, Cluj), 2002.
9. H. Wiener, *J. Am. Chem. Soc.* **69** (1947) 17–20.
10. M. Randić, *J. Am. Chem. Soc.* **97** (1975) 6609–6615.
11. C. Hansch and T. Fujita, *J. Am. Chem. Soc.* **86** (1964) 1616.
12. C. Hansch and T. Fujita, *Acc. Chem. Rev.* **2** (1969) 232.
13. A. Verloop, W. Hoogenstraten, and J. Tipker, *Drug Design* **7** (1976) 165.
14. C. Hansch and A. Leo, *Substituent Constant for Correlation Analysis in Chemistry and Biology*, John Wiley and Sons, New York, 1979.
15. A. C. Good, *J. Mol. Graphics* **10** (1992) 114.
16. S. H. Son, Ch. K. Han, S. K Ahn, J. H. Yoon, and K. T. No, *J. Chem. Inf. Comput. Chem*. **39** (1999) 601– 609.
17. StatSoft, Inc., STATISTICA (data analysis software system), version 6, 2001. www.statsoft.com
18. I. Yalçìn, I. Ören, O. Temiz, and A. Sener, *Acta Biochim. Pol.* **47** (2000) 481–486.

SAŽETAK

QSAR-modeliranje protugljivičnoga djelovanja nekih heterocikličkih spojeva

Oleg Ursu, Adina Costescu, Mircea V. Diudea i Bazil Parv

Metodom višestruke regresije učinjena je QSAR-analiza niza benzoksazola, benzimidazola, oksazolo-[4,5-*b*]piridina i benzotiazola, čije je inhibitorno djelovanje na *Candida albicans* ispitivano. Topološki indeksi (TI) i analiza njihovih glavnih komponenata (PCA, *Principal Component Analysis*) uporabljeni su za modeliranje protugljivičnoga djelovanja. Izbor odgovarajućih topoloških indeksa za izradu QSAR-modela učinjen je temeljem najviših rezultata faktora glavnih komponenata. Korelacijski koeficijent 0,97 utvrđen u validacijskome postupku ukazuje na visoku kakvoću izrađenih QSAR-modela.