MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **60** (2008) 977-984

ISSN 0340 - 6253

A novel QSAR approach in modeling antifungal activity of some 5- or 6-methyl-2-substituded benzoxazoles/ benzimidazoles against *C. albicans* using molecular descriptors

Cristina D. Moldovan*, Adina Costescu, Gabriel Katona and Mircea V. Diudea

**** Faculty of Chemistry and Chemical Engineering, Babes-Bolyai University, Cluj-Napoca, Romania***

(Received April 12, 2008)

Abstract. A quantitative structure-activity relationship (QSAR) is a mathematical model that relates a series of molecular structures to a biological activity. A new approach, called "direct prediction", is used to model the antifungal activity against *C. albicans* of a set of 25 molecules, derivatives of 5-or 6-methyl-2-substituded benzoxazole/benzimidazole. The direct prediction approach, here proposed, builds up similarity clusters for each molecule in the test set and makes predictions in the so obtained more congeneric subsets. The results are compared with those reported in literature**.**

1. Introduction

Mycotic illnesses in humans are divided into three groups: contagious skin and hair infections, noncontagious soilborne or airborne systemic infections and noncontagious foodborne toxemias. The responsible organisms and methods of prevention and treatment differ with each group. The prevalence of systemic fungal infections has increased significantly during the past decade. This increase is due to greater use of broad-spectrum antibiotics, immunosuppressive agents, central venous catheters, intensive care low birth weight infants, organ transplantation and the acquired immunodeficiency syndrome (AIDS) epidemic. Biologically active benzoxazoles are known for a long time and it was seen that position 2 is decisive for the biological activity, whereas position 5 determines the intensity of their activity.

Quantitative structure-activity relationships (QSAR) mathematically relate descriptors of a chemical structure to a biological activity and therefore can also be used to relate descriptors to antifungal effects. A structure-activity study can indicate which features of a given molecule are responsible for its activity, thus making possible to synthesize new and more potent compounds with enhanced biological activity. QSAR analysis is based on the assumption that the activity of compounds is correlated to the characteristics of their structure [3].

Molecular similarity was extensively and successfully used in drug discovery, often to compare molecules in the absence of other mechanistic information [4-6]. Reasons for the increasing popularity of similarity based methods include technological advances in high throughput screening and synthesis in the last decade and the need of applications of computer based methods in compound selection and activity evaluation to a much more extent than before [7]. Similarity searching [8,9] and clustering methods [9,10] can be used to classify compounds into structural groups [11] and for the prediction of biological activities as well [12]. The basic idea underlying on similarity-based QSAR approaches was enounced explicitly by Johnson and Maggiora [13]: "*molecules that are structurally similar likely will have similar properties*". Thus, when the activity of a given molecule is unknown, we can predict it by taking into account similarity values between the molecules under study and the molecules of a data set whose activities are known. Studies of similarity in chemical structures can also be overtaken by using molecular graphs [14,15].

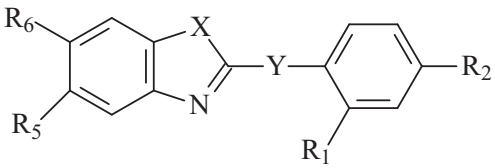
The article proposes a new approach called "direct prediction" which develops clusters of similar structures aimed to be more congeneric subset in predicting of a biological activity.

2. Materials and Methods

Recently, the synthesis and in vitro antifungal activities of different 5- or 6-methyl-2-substitutedbenzoxazoles/benzimidazoles against the fungus *Candida albicans* [2] have been reported (Figure 1 and Table 1).

Topological molecular descriptors are used in QSAR studies because of their accessibility, being easily computed by available software programs. The set of molecular descriptors used in this study is calculated by DRAGON software package

[16]. The structures were optimized by using the semi-empirical PM3 Hamiltonian, available in HyperChem. The regression equations were calculated by STATISTICA 6.0 software package.[17]



R_1= H, Cl, F, NO_2, CH_3, OCH_3 ; Y= —, CH_2, CH_2O, CH_2S

R_2= H, Cl, Br, NH_2, CH_3, OCH_3 ; X= O, NH

R_5=H, CH_3; R_6= H, CH_3

Figure 1. Antifungal active benzoxazoles and benzimidazoles against *C. albicans*

Table 1. Structure and antifungal activity of benzoxazole and benzimidazole derivatives against *C. albicans*

Struct.	R_1	R_2	R_5	R_6	X	Y	Obs. log 1/C
1	Cl	H	CH_3	H	O	-	3.989
2	OCH_3	H	CH_3	H	O	-	3.980
3	NO_2	H	CH_3	H	O	-	4.007
4	Cl	Cl	CH_3	H	O	-	4.046
5	CH_3	CH_3	CH_3	H	O	-	3.977
6	OCH_3	OCH_3	CH_3	H	O	-	4.032
7	Cl	H	H	CH_3	O	-	3.989
8	OCH_3	H	H	CH_3	O	-	3.980
9	F	H	H	CH_3	O	-	3.958
10	NO_2	H	H	CH_3	O	-	4.007
11	Cl	Cl	H	CH_3	O	-	4.046
12	CH_3	CH_3	H	CH_3	O	-	3.977
13	OCH_3	OCH_3	H	CH_3	O	-	4.032
14	H	H	CH_3	H	O	CH_2	4.251
15	H	Br	CH_3	H	O	CH_2	4.383
16	H	NH_2	CH_3	H	O	CH_2	4.280
17	H	H	H	CH_3	O	CH_2	4.251
18	H	H	CH_3	H	NH	CH_2	4.249
19	H	Cl	CH_3	H	NH	CH_2	4.312
20	H	Br	CH_3	H	NH	CH_2	4.382
21	H	NH_2	CH_3	H	NH	CH_2	4.278
22	H	H	CH_3	H	O	CH_2O	3.980
23	H	H	CH_3	H	O	CH_2S	4.009
24	H	H	CH_3	H	NH	CH_2S	4.007
25	H	Cl	CH_3	H	NH	CH_2O	4.037

Once the desired set of descriptors had been calculated and stored, the process of descriptor analysis is started. It is important to examine the pool of descriptors in an objective manner and to remove from further consideration those descriptors which are redundant or do not contain enough discriminatory information to be of any significant value. All descriptors containing identical values for 90% or more of the compounds in the data set, including both zero and non-zero values, were removed. All possible pairs of remaining descriptors were examined to identify those pairs which are highly correlated. As a rule, a critical value of 0.950 for the correlation coefficient (*R*) was used. If two descriptors were correlated at or above the critical value, one descriptor was discarded. The decision of which one is retained was based on the possible physical interpretation of the descriptor, ease of calculation, or usefulness in the past studies. The result of this analysis is a reduced pool of information-rich descriptors which can then be screened by using multiple linear regression analysis. After all of these procedures we reduced the searching space from 1600 to 680 descriptors.

Linear regression models were developed by multiple regressions with stepwise addition of descriptors, when the inclusion of a given term is based on the F statistic values. A further reduction of descriptors was based on the *leave-one-out* procedure, but here used to eliminate non-significant descriptors.

The quality of models is given by the squared regression coefficient (R^2), Fisher-ratio (F) and standard error of estimate (s). The final data reduction was performed by the aid of principal components analysis PCA.

Molecular similarity calculations were performed by Cluj-Simil software.[18]

3. Direct Prediction Procedure

Once the appropriate descriptors are selected, the data are submitted to the "direct prediction" procedure, which consists of the following steps:

(i) Take any molecule of the studied set as the leader in the similarity test.

(ii) Calculate the similarity scores with respect to the leader and cluster the structures up to a chosen threshold of similarity.

(iii) Leave out the leader and make the best regression model of the cluster (by using eventually, previous models).

(iv) Predict the activity of the leader by the best model derived on its own cluster.

(v) Validate the predicted data, when the values of the leader activities are known.

This procedure does not need the split of the set of molecules under study into training and predicting sets, as usually in QSAR studies. Each molecule activity is predicted by its own training set derived model and no general model is needed. Similarity clusters ensure a more congeneric set of molecules which is expected to provide more structurally-related calculated activities, thus allowing a direct interpretation of the results. The descriptors used in the various clusters, of various leaders, can be identical or may differ, the frequency of which could tell about particular structural features that compose the global molecular activity.

The similarity, calculated in this case by the Cluj-Simil software, is based on similarity of molecular descriptions (generalized Randić and DS indices). The models in all clusters were derived by using *MW* – molecular weight; *nCs* – number of secondary carbon atoms C(sp^3) as descriptors in bivariate regressions (Table 2).

Table 2. Molecular descriptors, observed and calculated log 1/C values by the bivariate equations: $\log(1/C) = a + b \times MW + c \times nCs$ derived from each similarity cluster, and the corresponding similarity threshold, for the molecules in Table 1.

Struct.	*MW*	*nCs*	Obs. log 1/C	Bivariate eqs.* used in prediction			Calc. log 1/C	No. structures in clusters	Similarity threshold
				a	b	c			
1	243.7	0	3.989	3.529	0.0018	0.300	3.989	19	0.94
2	239.29	0	3.980	3.578	0.0016	0.298	3.980	23	0.94
3	254.26	0	4.007	3.577	0.0016	0.298	4.005	19	0.95
4	278.14	0	4.046	3.577	0.0016	0.298	4.046	23	0.94
5	237.32	0	3.977	3.577	0.0016	0.298	3.977	19	0.95
6	269.32	0	4.032	3.586	0.0016	0.299	4.030	19	0.90
7	243.7	0	3.989	3.577	0.0016	0.299	3.988	19	0.94
8	239.29	0	3.980	3.578	0.0016	0.298	3.980	23	0.94
9	227.25	0	3.958	3.579	0.0016	0.298	3.960	19	0.94
10	254.26	0	4.007	3.577	0.0016	0.298	4.005	19	0.95
11	278.14	0	4.046	3.586	0.0016	0.299	4.045	19	0.93
12	237.32	0	3.977	3.586	0.0016	0.299	3.978	19	0.94
13	269.32	0	4.032	3.586	0.0016	0.299	4.030	19	0.93

14	223.29	1	4.251	3.529	0.0018	0.300	4.250	19	0.94
15	302.18	1	4.383	3.582	0.0016	0.299	4.385	19	0.94
16	238.31	1	4.280	3.585	0.0016	0.299	4.279	19	0.94
17	223.29	1	4.251	3.529	0.0018	0.300	4.250	19	0.70
18	222.31	1	4.249	3.529	0.0018	0.300	4.249	19	0.71
19	256.75	1	4.312	3.585	0.0016	0.298	4.309	19	0.94
20	301.2	1	4.382	3.583	0.0016	0.299	4.383	19	0.94
21	237.33	1	4.278	3.585	0.0016	0.299	4.277	19	0.94
22	239.29	0	3.980	3.587	0.0016	0.299	3.981	19	0.93
23	255.36	0	4.009	3.585	0.0016	0.299	4.007	19	0.94
24	254.38	0	4.007	3.585	0.0016	0.299	4.006	19	0.94
25	272.75	0	4.037	3.586	0.0016	0.299	4.036	19	0.96

For each of the 25 structure, taken as leader in the similarity test, the training equations showed R^2=0.9999 and the numbers varied at the 5th decimal. The final validation of predictions showed a good coefficient of correlation: R^2=0.999 (Figure 2).

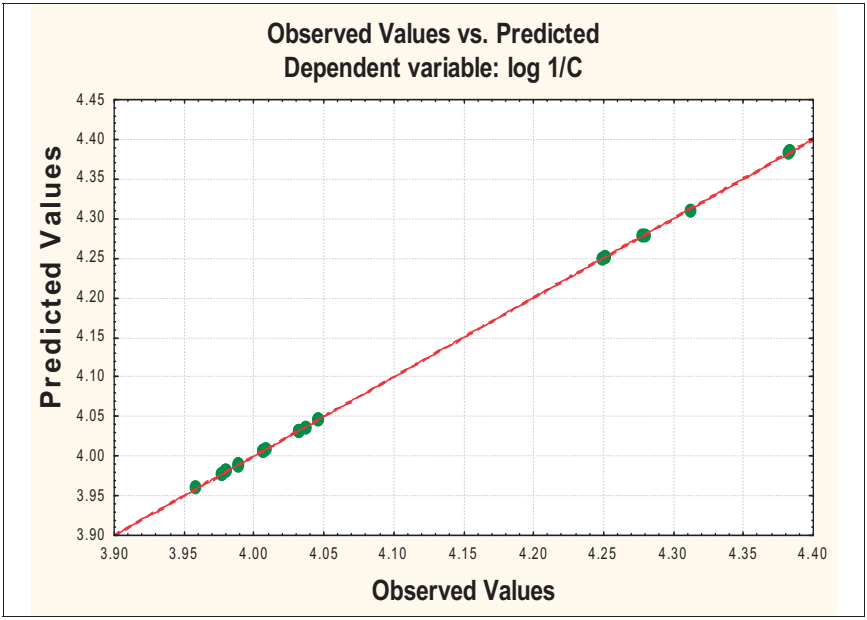


Figure 2. Plots of observed vs. predicted property for similarity test

This correlation is better than that previously reported[2] (R^2=0.94) one.

4. Conclusions

A set of 5- or 6-methyl-2-substituded benzoxazoles/ benzimidazoles derivatives previously tested for their antibacterial activities against *C. albicans* was analyzed by quantitative structure-activity relationship and the activity was modeled by using multiple linear regressions. The resulting QSAR reveled that, for the antifungal activity, the contribution of fragments at position Y (descriptor *nCs*) plays an important role.

The newly proposed "direct prediction" procedure, working on similarity clusters, ensures a more congeneric set of molecules on which the prediction is made, contrarily to the procedures in use so far. It provides, thus, more structurally-related calculated activities, which allow a deeper understanding of local contributions to the global molecular activity. The procedure needs, however, supplementary testing.

Acknowledgement. The authors are grateful to the Ministry of Education and Research for financial support from CEEX CHEMMOD 233/2006 Project.

References

1. W.E. Dismukes, Introduction to Antifungal Drugs, *Clin. Infect. Dis.*, **2000**, *30*, 653-657.

2. Özlem TEMÜZ ARPACI, QSARs of Some 5- or 6-Methyl-2-Substituted Benzoxazoles/Benzimidazoles against *Candida albicans*, *Turk. J. Med. Sci.*, **2001,** *31*, 493-497

3. D. Rogers, and A. J. Hopfinger. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*, 854-866.

4. R.P. Sheridan, S.K. Kearsley, Why do we need so many chemical similarity search methods? *Drug Discov. Today,* **2002***, 7*, 903-911.

5. H.-J Bohm, G. Schneider, Virtual Screening for Bioactive Molecules; *Wiley-VCH*, Weinheim, **2000**.

6. G. Schneider, H.J. Bohm, Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002***, 7*, 64-70.

7. J.A. DiMasi, R.W. Hansen, H.G. Grabowski, The price of innovation: new estimates of drug development costs, *J. Health Econ.,* **2003***, 22*, 151-185.

8. P.M. Dean, Molecular Similarity in Drug Design, *Blackie Academic*, London, **1995**

9. P. Willet, Similarity and Clustering in Chemical Information Systems, *Research Studies Press* Ltd: Letchworth, U.K., **1987**.

10. J.M. Barnard,; G.M. Downs, Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures, *J. Chem. Inf. Comput. Sci.,* **1992**, *32*, 644-649

11. D.J. Wild, C.J. Blankey, Comparison of 2D Fingerprints Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering, *J. Chem. Inf. Comput. Sci.*, **2000**, *40,* 155-162.

12. R.D. Brown, Y.C. Martin, Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection, *J. Chem. Inf. Comput. Sci.,* **1996**, *36*, 572-584.

13. M.A. Johnson, G.M. Maggiora, Eds., Concepts and Applications of Molecular Similarity, *John Wiley & Sons*, New York, **1990**.

14. P. Willett, Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

15. N. Nikolova, J. Jaworska, Approaches to measure chemical similarity – a review. *QSAR & Combin. Sci.*, **2004**, *22*, 1006-1026.

16. Dragon Profesional version 5.4, **2006** (http://www.talete.mi.it/dragon_exp.htm).

17. StatSoft, Inc. STATISTICA (data analysis software system), version 6, **2001** (www.statsoft.com).

18. G. Katona and M. V. Diudea, CLUJSIMIL software, "Babes-Bolyai" University, 2001, Cluj, Romania.