Computational Biology and Bioinformatics
2014; 2(2): 25-32
Published online May 30, 2014 (http://www.sciencepublishinggroup.com/j/cbb)
doi: 10.11648/j.cbb.20140202.12

QSAR studies, design, synthesis and antimicrobial evaluation of azole derivatives

Vasyl Kovalishyn[1, *], Iryna Kopernyk[1], Svitlana Chumachenko[2], Oleg Shablykin[2], Kostyantyn Kondratyuk[2], Stepan Pil'o[2], Volodymyr Prokopenko[2], Volodymyr Brovarets[2], Larysa Metelytsia[1]

[1]Department of medical and biological researches, Institute of Bioorganic Chemistry and Petrochemistry, NAS of Ukraine, Kyiv, Ukraine
[2]Department for Chemistry of Bioactive Nitrogen-Containing Heterocyclic Compounds, Institute of Bioorganic Chemistry and Petrochemistry, NAS of Ukraine, Kyiv, Ukraine

Email address:

vkovalishyn@yahoo.com (V. Kovalishyn), kopernik@bpci.kiev.ua (I. Kopernyk), sv.chumachenko@ukr.net (S. Chumachenko), shablykin@bpci.kiev.ua (O. Shablykin), kondratyuk@bpci.kiev.ua (K. Kondratyuk), brovarets@bpci.kiev.ua (S. Pil'o), velizariy@gmail.com (V. Prokopenko), brovarets@bpci.kiev.ua (V. Brovarets), metelitsa@bpci.kiev.ua (L. Metelytsia)

To cite this article:

Vasyl Kovalishyn, Iryna Kopernyk, Svitlana Chumachenko, Oleg Shablykin, Kostyantyn Kondratyuk, Stepan Pil'o, Volodymyr Prokopenko, Volodymyr Brovarets, Larysa Metelytsia. QSAR Studies, Design, Synthesis and Antimicrobial Evaluation of Azole Derivatives. *Computational Biology and Bioinformatics.* Vol. 2, No. 2, 2014, pp. 25-32. doi: 10.11648/j.cbb.20140202.12

Abstract: QSAR analysis of a set of previously synthesized azole derivatives tested for growth inhibitory activity against *Candida albicans* was performed by using Associative Neural Network. To overcome the problem of overfitting due to descriptor selection, 5-fold cross-validation with variable selection in each step of the analysis was used. The predictive ability of the models was tested through leave-one-out cross-validation, giving a Q^2 = 0.77 - 0.79 for regression models. Predictions for the external evaluation sets obtained accuracies in the range of 0.70 - 0.80 for regressions. Biological testing of compounds was performed by disco-diffusion method on solid medium culture versus strain *C. albicans* ATCC 10231 M885. Most of compounds demonstrated high antifungal activity. Five synthesized compounds also showed activity against clinical isolate strain of *C. albicans* received from a biological material and resistant to fluconazole.

Keywords: QSAR, Artificial Neural Networks, Candida Albicans, Drug Design

1. Introduction

The frequency of infections caused by pathogenic microorganisms has increased worldwide, becoming an important cause of morbidity in many countries. *Candida albicans* is the most common human fungal pathogen, and mortality from *C. albicans* infection is still unacceptably high [1]. *C. albicans* is able to colonize nearly every part of the gastrointestinal tract, from the mouth cavity up to the perianal tissues, epidermis and the vulvovaginal region [1]. For example, about 75% of adult women have at least one episode of vulvovaginal candidiasis during their life, with predominance of *C. albicans* in 70–90% [2]. *C. albicans* is notorious for causing candidiasis, it can affect the oesophagus with the potential of becoming systemic, provoking a much more serious condition, a fungemia named candidemia [3]. Therefore, *C. albicans* infections detect a number of problems including limited number of effective antifungal agents, resistance of Candida to commonly used antifungals, toxicity of the available antifungal agents, relapse of Candida infections and inexpensive effective antifungal agents. Thus, the search for new and effective synthetic inhibitors of *C. albicans* is an actual and important task for basic science and clinical medicine.

The experimental measurement of bioactivity of compounds are difficult, more expensive and time-consuming, thus a great deal of effort has been done into attempting to predict activity through statistical modeling. In order to design new inhibitors with higher inhibitory activity based on the inhibitors with known activity, it is quite necessary to study the quantitative structure-activity relationship (QSAR) of these compounds. In practice, QSAR models provide valuable tools for automated virtual screening, combinatorial library design, and data mining [4]. The Multiple Linear Regression analysis (MLRA), Support Vector Machines (SVM), Random Forests, Partial Least Squares (PLS) Regression, Artificial Neural Networks (ANNs), Bayesian Neural Networks, etc., are widely used techniques to discover structure-property relationships [5-8].

In recent years, substantial progress has been made in the application of *in silico* computational methods to predict *Candida albicans* inhibition activities of some chemicals [9-12]. However, many of these QSAR models were designed for specific classes of chemicals using small number of compounds and could not be used for virtual screening of big compound libraries aimed at identifying potential *C. albicans* inhibitors. Within our research group, a database has been assembled with more than 1878 azole derivatives which are potential *C. albicans* inhibitors, their respective biological activity expressed in terms of minimum inhibitory concentration (MIC), as well as, a large set of molecular descriptors and properties (geometrical and electronic parameters, etc.). The azole has been an important pharmacophore and privileged structure in medicinal chemistry. Many azoles are known as antifungal drugs, inhibiting the fungal enzyme 14α-demethylase which produces ergosterol (an important component of the fungal plasma membrane) [13]. However, the emergence of azole resistant strains has stimulated the search for new antimycotic compounds. This study describes the (1) creation of QSAR models to identify new potential azole inhibitors with above-stated activity using Artificial Neural Network approach, (2) methods that were employed to optimize the predictive performance of these models and (3) synthesis and biological testing a series of azole derivatives as new potential antifungal compounds.

2. Material and Methods

2.1. Experimental Data and Descriptor Generation

The data for our analysis were obtained from many publications and stored in the ChEMBL database [14]. The detailed structures and the corresponding bioactivities of the compounds and full list of publications are listed in *Supplementary Materials*. The biological data obtained as minimum inhibitory concentration (MIC) were converted into log (1/MIC) values and used as dependent variable in the following QSAR analyses. The dataset consisted of 1878 *C. albicans* inhibitors. The range of MIC values of the 1878 compounds was from 1.4 nM to 2.7 mM. All molecules were “transformed” using the ChemAxon standardizer in order to identify their standardized forms [15]. The 2D coordinates of atoms were recalculated, counter ions and salts were removed from molecular structures, molecules were neutralized, mesomerized, aromatized. Data sets were filtered to remove duplicates. The 3D structures were calculated using the ChemAxon standardizer from the SMILES notation available for each compound, and stored in SDF format [15]. Then, the resulted geometries were input into DRAGON software to calculate molecular descriptors [16]. The Dragon program provides many types of molecular descriptors such as numbers of hydrogen bond donors and acceptors, topological polar surface area, RDF, WHIM descriptors and many others. Finally, constant or near constant values and descriptors found to be correlated pairwise (one of any two descriptors with a correlation coefficient greater than 0.99 was removed to reduce redundant and useless information) were excluded in a preliminary step [16].

2.2. Associative Neural Networks

An Associative Neural Network (ASNN) combines an ensemble of Feed-Forward Neural Networks (FFNNs) with the method of *k*-Nearest Neighbours (*k*-NN) [17]. FFNNs represent supervised regression methods that are trained using a data set in which the property to be modeled is known. The traditional FFNN represents a memoryless approach, i.e. after training, the initial data are no longer needed and all the information necessary for predictions is stored within the neural network weights [18]. To the contrary, such methods as *k*-Nearest Neighbor Method represent the memory-based approach [19]. The *k*-NN keeps in memory the input data and their predictions are based on some local approximation of the stored examples. ASNNs use the *k*-NN method in the space of ensemble residuals. All compounds are represented as vectors of neural network predictions by the neural network ensemble. Correlation between such vectors is used by the nearest neighbor method as a measure of distance between the analyzed cases. Therefore, ASNN perform *k*-NN in the space of ensemble residuals. So the ASNN improves prediction by the bias correction of the neural network ensemble [17].

We have used the neural networks algorithm trained by SuperSAB [20]. The neural networks had a number of inputs equal to the number of descriptors. One hidden layer with five neurons was used in the calculations. Weights were initialized with random numbers. A bias neuron was also presented in both the input and hidden layers. The FFNNs used one output neuron for regression tasks, and the output values were linearly scaled between 0.1 and 0.9 [21]. All neural networks had the same architecture. Cross-validation techniques were used to strictly control the possibility of over-fitting the data. Each FFNN ensemble included M=200 networks. More details of the algorithm can be found in earlier publications [21, 22].

2.3. Search of an Optimal Descriptor Number

Usually, the initial data consist of big quantity of descriptors many of them not directly related to the solved problem. The selection methods can optimize number of descriptors. In the past years, several methods for the selection of molecular descriptors have been developed [23, 24]. In this work, the descriptor’s importance measure was obtained by “pruning methods” implemented in ASNN software. Pruning algorithms introduce some measure of importance of the ASNN matrix weights by the so called "sensitivities". These algorithms work similarly to a stepwise multiple regression analysis, whereby one input parameter considered to be non-significant is excluded at each step. Another word at each step, the model sensitivities to all weights and input nodes are evaluated and the descriptor associated with the input neuron showing the smallest sensitivities is deleted [25, 26]. For these studies, the descriptors obtained from previous stage were used. We analyzed the influence of the number of selected descriptors on the ASNN model quality (on the basis of the leave-one-out results for the training sets). Detailed explanations of the various sensitivity methods can be found elsewhere in literature [25, 26].

2.4. Validation of QSAR Models

To overcome the problem of overfitting given by descriptor selection (see section *Search of an optimal descriptor number*) we performed 5-fold cross-validation with variable selection in each step of the analysis [6]. In the 5-fold cross-validation the original data set was divided into 5-subsets of approximately equal size. Out of 5-subsets a single sub sample was retain as validation data for testing the model, and remaining four subsamples were used as training data. For each subset, we first selected descriptors using the corresponding training set, developed the model and then applied it to predict the molecules which were excluded from the training set. This procedure was sequentially repeated five times producing five different external validation data sets and corresponding training set molecules [27]. Then the average statistical coefficients for all 5-test sets were computed. Therefore we developed five predictive models by ASNN and a united model. The prediction statistics for QSAR models are given in Tables 1 (see section *Results and Discussion*).

The conventional way of summarizing "lack of fit" in QSAR models is the Root-Mean-Square Error (RMSE) and Mean Absolute Error (MAE) between observed and predicted activities. Another is the cross-validation coefficient (Q^2) [27, 28].

It was defined as:

$$Q^2 = (\text{SD-PRESS})/\text{SD} \qquad (1)$$

Here, the SD is the sum of the squared deviations of the target variable values from their mean, and PRESS is the prediction error sum of squares obtained from the leave-one-out cross-validation procedure. Use of the cross-validation coefficient Q^2 makes redundant the analysis of residuals by means of standard deviation, because both coefficients are interrelated and can be derived one from another. Here, we considered a QSAR model to have an acceptable predictive power if $Q^2 > 0.5$ [27].

2.5. Applicability Domain

The dataset used for QSAR analysis only covers limited chemical space. Therefore, QSAR models should have a well-defined applicability domain (AD) within which reliable predictions can be made [29]. The AD of a QSAR model is partly a function of the molecular coverage of the test molecule relative to the molecules in the training data set. If a test molecule is very different from the other compounds in the training set, the prediction of its activity is unreliable. A concept of the AD was created and used to avoid such an incorrect extrapolation of activity predictions. Thereby it is possible to predict the model accuracy for a particular compound and to select a subset of much more confident predictions. AD approaches rely on finding a measure, which correlates with the accuracy of predictions. The "distance to a model" can be defined as a metric that defines the similarity between the training set molecules and the test set compound for the given property in the context of a specific model [30].

We used approach based on similarity analysis for estimation of "distance to a model" [29]. The AD of the QSAR models was calculated from the distribution of similarities between each compound and its *k* nearest neighbors in the training sets [30]. Similarity of each molecule in test set was defined as the Dice Index (DI) between this molecule and the training set. They were computed as the average DI to the *k* nearest neighbors of this molecule in the training set [30]. Average DI values were calculated using *k*=10, which was the optimal number of nearest neighbors for the models. Thus, if the similarity of the external compound from all its nearest neighbors in the training set less this cutoff value, the prediction is considered unreliable. Our results have demonstrated that molecules with higher similarity are better predicted (see sections *Results and Discussion*).

3. Results and Discussion

3.1. Results of QSAR Modeling for C. Albicans Inhibitor Set

578 descriptors calculated by the DRAGON software were used. In the first stage, ASNN models were developed using total set of descriptors. Then, the number of descriptors was reduced for each set by ASNN pruning methods, keeping basically the same accuracy for training and test sets. In total, five models differing in the types of descriptors for each subset have been developed. Table 1 summarizes the statistical parameters for all models.

Table 1. *Statistical parameters for ASNN models for Candida albicans inhibitory activity.*

Name	Number of descriptors	Training set		Test set	
		Q^2	MAE[a]	Q^2	MAE
Set 1	41	0.77	0.38	0.70	0.41
Set 2	63	0.78	0.36	0.76	0.38
Set 3	64	0.78	0.36	0.77	0.37
Set 4	68	0.77	0.37	0.80	0.34
Set 5	90	0.79	0.35	0.70	0.42
		5-fold validation			
Total set				0.75	0.39

[a]MAE – mean absolute error

All QSAR models were first developed based on the training sets only and their accuracy was estimated using the Leave-One-Out (LOO) cross-validation [28]. The Q^2 coefficients for the training sets were 0.77 - 0.79. The compounds in the external test sets were predicted with the accuracy, Q^2 = 0.70-0.80. The total accuracy Q^2 calculated using the 5-fold validation was about 0.75, MAE = 0.39.

The model is stable and predictive both internally, as can be verified by the statistical parameters (high value of cross-validation parameters Q^2 and low MAE), and externally (similar high value of Q^2); the small values of MAE and standard deviation errors (not shown in table).

The quality of the results is given graphically in Fig. 1, showing the good correlation between observed and predicted values for the total data set (validation set 1-5) using 5-fold validation method. After analyzing the prediction results for all compounds of dataset, we found that most compounds (namely 1728 out of 1878) are well predicted with smaller residues lower than 1 log unit. Only 9 chemicals have residuals between the experimental and predicted log(1/MIC) higher than 2 log units. Slightly worse predictions for some compounds from the test set 1 and 5 (Table 1, Fig 1) can be explained by the fact that a number of compounds of these sets are more dissimilar from the molecules of training sets.

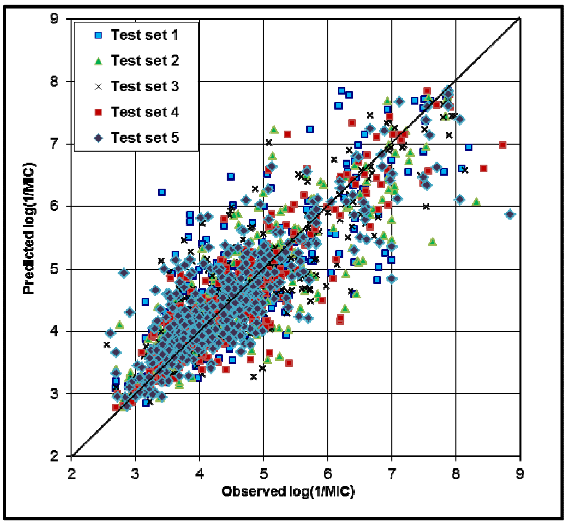


Figure 1. *Plots of experimental versus predicted values for the total QSAR model data set; MIC is the minimal inhibitory concentration.*

3.2. The Influence of the Applicability Domain

The AD defines the area of the descriptor space in which QSAR models can more accurately predict the target properties. If a compound is "too dissimilar" (beyond the defined distance cutoff value) to all compounds of the modeling set in the descriptor space then we assume that we cannot predict its activity reliably. The AD for models derived was calculated from the distribution of similarities between each compound and its *k* nearest neighbors in the training sets. The similarities were defined as Dice index between a molecule *i* and a training set. They were computed as the average DI to the *k* nearest neighbors of this molecule in the training set [30]. The average DI were calculated using *k*=10, which was an optimal number of nearest neighbors for the models.

We have investigated the effect of varying the threshold value of the AD on the interplay between chemical space coverage and prediction accuracy. Without applying the AD for the data set, the best overall accuracy was 0.75 (see Table 2).

Table 2. *Influence of Dice index cutoff on total model quality.*

Step	Dice index	Number of molecules	Statistical coefficients	
			Q^2	MAE[a]
1	-	1878	0.75	0.39
2	0.5	1860	0.75	0.38
3	0.6	1835	0.75	0.38
4	0.7	1762	0.76	0.37
5	0.8	1608	0.78	0.36
6	0.9	1064	0.81	0.33

[a] MAE – mean absolute error

It was found that for total data set, the Q^2 increases (MAE decreases) when DI cutoff increases from 0.6 to 0.9 (Table 2). However, it is generally hard to assign a standard AD threshold that should be used in all cases. The increased accuracy came at the expense of reducing the number of compounds for which the prediction could be made. For example, if DI cutoff equals 0.9, less than 60% of all compounds were within the AD for the total data set.

Typically, we tend to use a conservative DI cutoff of 0.6 to ensure high prediction accuracy for all (eligible) compounds in the external sets. Thus, results of this study illustrate that the prediction accuracy was not increased significantly by increasing the AD, i.e., DI cutoff, up to values as high as 0.9, which allowed making accurate predictions for more than half of the compounds in both data sets.

3.3. Prediction Activity of New Compounds

In the present study, we generated virtual set of drug-like molecules in order to screen potential inhibitors against *Candida albicans*. The 2D structures of new compounds were built using MarvinSketch template libraries [15] and consisted of azole analogues. Finally, the activity of all virtual compounds was predicted using the proposed QSAR models. The compounds that were most similar to the training set (DI similarity cutoff of 0.6) and predicted as being the most active were selected for biological testing (see Table 3).

Table 3. *Predicted activity of the azole study derivatives, against C. albicans.*

Number	M1[a]	M2	M3	M4	M5	Mean[b]
1	4.63	4.05	4.72	4.31	5.00	4.54±0.33
2	3.71	4.04	3.82	3.38	4.41	3.87±0.34

Number	M1[a]	M2	M3	M4	M5	Mean[b]
3	3.82	4.11	4.05	3.66	4.35	3.99±0.24
4	3.80	5.05	4.47	3.86	5.29	4.50±0.60
5	4.22	5.37	4.30	4.37	5.63	4.78±0.59
6	4.68	3.63	4.39	3.88	5.02	4.32±0.51
7	4.73	3.70	4.22	3.77	4.77	4.24±0.45
8	4.30	3.99	4.52	3.75	4.89	4.29±0.40
9	3.83	4.79	4.16	4.33	4.29	4.28±0.31
10	4.69	3.85	4.10	3.81	4.16	4.12±0.31
11	4.60	3.94	3.99	3.77	4.06	4.07±0.28

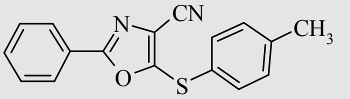
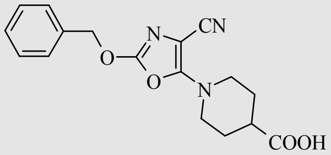
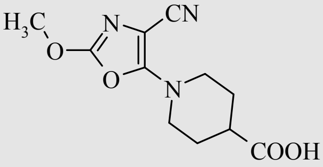
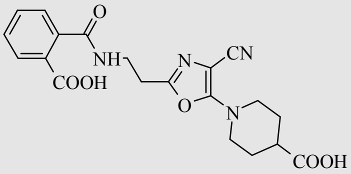
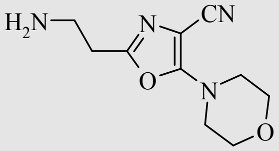
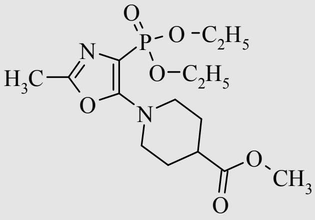
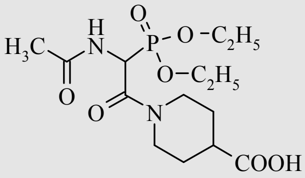
[a]M1- QSAR model number 1; [b]Mean value of log(1/MIC)

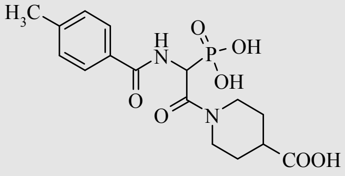
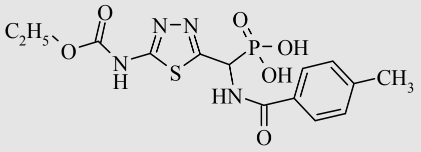
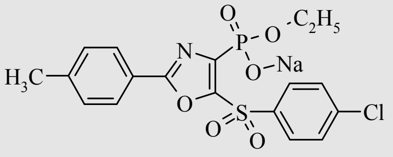
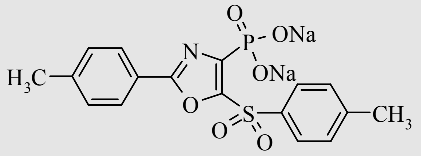
Compounds 1, 4 and 5 were predicted to be the most active. The activity value of compound 5 was predicted to be the highest at log (1/MIC) = 4.78±0.59, followed by compounds 1 and 4 with forecasted activities of log (1/MIC) = 4.54±0.33 and 4.5±0.6, respectively. The average Dice Index values ranged between 0.61-0.69 for these compounds. The results of the biological screening of the proposed azole derivatives confirmed the QSAR predictions for compounds 1, 4 and 5 (see section *Biology*, Table 5).

4. Chemistry

Compounds 1-11 were synthesized starting from available N-(2,3,3,3-tetrachloroethyl) carboxylic acid amides (see Table 4) [31]. For preparation 4-cyanoxazoles this reagent was treated with potassium cyanide and then with excess of aliphatic amine as described in [32].

Table 4. *Chemical structures of compounds 1-11.*

Number	Molecular weight	Chemical structure	Chemical name
1	292,36		2-Phenyl-5-tolylsulfanyl-4-cyano-1.3-oxazole
2	327,34		1-(2-Benzyloxy-4-cyano-1,3-oxazol-5-yl)piperidine-4-carboxylic acid
3	251,24		1-(4-Cyano-2-methoxy-1,3-oxazol-5-yl)piperidine-4-carboxylic acid
4	412,41		1-{2-[2-(2-Carboxy-benzoylamino)ethyl]-4-cyano-1,3-oxazol-5-yl}piperidine-4- carboxylic acid
5	222,25		2-(2-Aminoethyl)-5-(morpholin-4-yl)-1,3-oxazole-4-carbonitrile
6	360,35		Methyl N-[4-(diethoxy-phosphoryl)-2-methyl-1,3-oxazol-5-yl]isonipecotinate
7	364,34		1-[2-Acetylamino-2-(diethoxyphoshoryl)acetyl]-piperidin-4-ylcarboxylic acid

Number	Molecular weight	Chemical structure	Chemical name
8	384,33		1-{2-(Dihydroxyphosphoryl)-2-[(4-methylphenyl)-formamido]acetyl}piperidin-4-ylcarboxylic acid
9	400,35		(5-Ethyloxycarbonylamino-1,3,4-thiadiazol-2-yl)-[(4-methylbenzoyl)-amino]methyl-phosphonic acid
10	463,81		Sodium monoethyl 2-(4-methylphenyl)-5-[(4-chlorophenyl)sulphonyl]-1,3-oxazol-4-ylphosphonate
11	437,32		Disodium-2-(4-methylphenyl)-5-[(4-methylphenyl)-sulphonyl]-1,3-oxazol-4-ylphosphonate

4-(Diethoxyphosphoryl) oxazoles were obtained via treatment N-tetrachlorethylamides with ethylphosphite and then with excess of aliphatic amine [33]. Compound 1 was synthesised in reaction of 2-acylamino-3,3-dichloroacrylonitriles with thiophenol, followed by refluxion with Argentum carbonate [34]. Other compounds 2-11 was synthesised from 4-cyanoxazoles and 4-(diethoxyphosphoryl)oxazoles by known methods described in literature [35].

5. Biology

We investigated the activity of new 11 azole derivatives shown in Table 4. Fluconazole, a known effective antimycotic drug, was used as positive control. The fungistatic activity of each compound was assessed using *C. albicans* standard strain ATCC 10231 M 885 and its clinical isolate by the use of an established standard Kirby-Bauer disk diffusion method [36]. Seaboard's agar (JSC «Research center of pharmacotherapy» Saint-Petersburg) was prepared according to the manufacturer's instructions, then dispended into glass bottles and autoclaved at 121°C and 15 psi for 15 min. The microbial culture was evenly poured onto the surface of agar plates into a volume of 0.2 ml of sterile saline solution to produce end concentrations of $1{\cdot}10^5$, $1{\cdot}10^6$, and $1{\cdot}10^7$ colony forming units (CFU) in 1 ml. A sterile 6 mm paper disc was placed on each agar plate and the test compound then inoculated onto the disk in a volume of 20 µl. The plates were then incubated at 37 °C for 24 h. All compounds were tested in triplicate at a concentration of $1.3{\cdot}10^{-7}$ M. The standard disks (JSC «Research center of pharmacotherapy» Saint-Petersburg) of reference-preparation contained 40 µg fluconazole, corresponding to $1.3{\cdot}10^{-7}$M.

The activities of all compounds are shown in Table 5.

Table 5. *Growth inhibition of C. albicans strain ATCC 10231 M 885 by a set of azole derivatives. The diameters of inhibition zones are given in millimeters.*

Number	Microbial loading, CFU in 1 ml					
	$1\ 10^5$		$1\ 10^6$		$1\ 10^7$	
	standard[a]	isolate[b]	standard	isolate	standard	isolate
1	23	10	17	9	14	9
2	15	na	11	na	na	na
3	13	na	na	na	na	na
4	20	9	16	8	13	na
5	22	10	20	9	17	na
6	17	na	13	na	9	na
7	15	na	12	na	9	na
8	14	na	11	na	8	na
9	16	na	11	na	8	na
10	21	10	15	na	11	na
11	22	8	19	na	14	na
Fluconazole	21	na	22	na	20	na

[a]*C. albicans* standard strain ATCC 10231 M 885; [b]*C. albicans* strain isolated from biomaterial; [c]na, not active.

The data presented in Table 5 show that compounds 1, 4, 5, 10 and 11 exhibit high fungistatic activity against *C. albicans* strain ATCC 10231 M 885, that is in good agreement with the QSAR predictions. Zones of inhibition formed by these compounds under conditions of high microbial loading ($1{\cdot}10^5$ CFU in 1 ml) exceeded those obtained using Fluconazole. The increase of microbial loading ($1{\cdot}10^6$ and $1{\cdot}10^7$ CFU in 1 ml) led to gradual decrease of antimycotic activity of all compounds in comparison with the fluconazole. It should be noted that the compounds 1, 4, 5, 10 and 11 also showed activity against clinical isolate strain of *C. albicans* received from a biological material and resistant to fluconazole.

6. Conclusion

In summary, global ASNN model was built to study the quantitative structure-activity relationship for a series of selective *Candida albicans* inhibitors. DRAGON software was used to calculate the molecular descriptors. The proposed QSAR model have good stability, robustness and predictive power when verified by internal validation (cross-validation by LOO) and also external validation. An application of pruning methods was able to select subsets of most relevant input descriptors determining the molecular inhibitory activity. Integrating the Artificial Neural Network method with pruning algorithms and n-fold validation approach it was possible to build models with high predictive ability. To have "external" chemicals not used in the model development, the original data set is split into training and prediction sets randomly in order to avoid the bias of structural similarity. Experimental results also indicated that a DI similarity value ≥ 0.60 could be used to estimate the reliability of the predictions. This means that the QSAR models presented can be reliably used to predict the *C. albicans* inhibitor activity of new azole derivatives. Our results demonstrate that proposed azole derivatives show significant activity against *C. albicans*. Compounds 1, 4, 5, 10 and 11 appeared to have potential for the treatment of candidiasis. The proposed QSAR models can be applied as tools for finding new potential *C. albicans* inhibitors.

Acknowledgements

We would like to thank the Science and Technology Center in the Ukraine (STCU no. 5215) and the NATO Science for Peace and Security Programme (grant EAP.SFPP 984401) for funding this research.

References

[1] M. A. Pfaller, D. J. Diekema, "Epidemiology of invasive candidiasis: a persistent public health problem," Clin Microbiol Rev, vol. 20, pp. 133-163, 2007.

[2] P. P. Chong, Y. L. Lee, B. C. Ian, K. P. Ng, "Genetic relatedness of Candida strains isolated from women with vaginal candidiasis in Malaysia," J Med Microbiol, vol. 52, pp. 657–666, 2003.

[3] S. Rekha, G. M. Visyasagar, "Anti-Candida activity of medicinal plants," Int J Pharm Pharm Sci, vol. 5, pp. 9-16, 2013.

[4] A Tropsha, "Best Practices for QSAR Model Development, Validation, and Exploitation," Mol Inf, vol. 29, pp. 476-488, 2010.

[5] C. Ventura, F. Martins, "Application of Quantitative Structure Activity Relationships to the Modeling of Antitubercular Compounds. 1. The Hydrazide Family," J Med Chem, vol. 51, pp. 612–624, 2008.

[6] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, "Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," J Chem Inf Model, vol. 48, pp. 1733-1746, 2008.

[7] V. Kovalishyn, V. Tanchuk, L. Charochkina, I. Semenuta, V. Prokopenko, "Predictive QSAR modeling of phosphodiesterase 4 inhibitors," J Mol Graph Model, vol. 32, pp. 32-38, 2012.

[8] P. Prathipati, N. L. Ma, and T. H. Keller, "Global Bayesian Models for the Prioritization of Antitubercular Agents," J Chem Inf Model, vol. 48, pp. 2362–2370, 2008.

[9] I. Yalçin, I. Oren, O. Temiz, E. A. Sener, "QSARs of some novel isosteric heterocyclics with antifungal activity," Acta Biochim Pol, vol. 47(2), pp. 481-486, 2000.

[10] S. M. Tana, J. Jiaoa, X. L. Zhua, Y. P. Zhoua, D. D. Songa, H. Gongb, R. Q. Yuc, "QSAR studies of a diverse series of antimicrobial agents against Candida albicans by classification and regression trees," Chemometrics and Intelligent Laboratory Systems, vol. 103, pp. 184–190, 2010.

[11] A. Tafi, R. Costi, M. Botta, R. Di Santo, F. Corelli, S. Massa, A. Ciacci, F. Manetti, M. Artico, "Antifungal agents. 10. New derivatives of 1-[(aryl)[4-aryl-1H-pyrrol-3-yl] methyl]-1H-imidazole, synthesis, anti-candida activity, and quantitative structure-analysis relationship studies," J Med Chem., vol. 45(13), pp. 2720-2732, 2002.

[12] C. Rami, L. Patel, C.N. Patel, J.P. Parmar, "Synthesis, antifungal activity, and QSAR studies of 1,6-dihydropyrimidine derivatives," J Pharm Bioallied Sci. vol. 5, pp. 277-289, 2013.

[13] D. Zampieri, M. G. Mamolo, L. Vio, E. Banfi, G. Scialino, M. Fermeglia, M. Ferrone, S. Pricl, "Synthesis, antifungal and antimycobacterial activities of new bis-imidazole derivatives, and prediction of their binding to P450(14DM) by molecular docking and MM/PBSA method," Bioorg Med Chem., vol. 15(23), pp. 7444-7458, 2007.

[14] https://www.ebi.ac.uk/chembl/, (accessed in January, 2014).

[15] https://www.chemaxon.com/, (accessed in January, 2014).

[16] http://www.talete.mi.it/products/dragon_description.htm (accessed in March 2014).

[17] I. V. Tetko, "Neural network studies. 4. Introduction to associative neural network," J Chem Inf Comput Sci, vol. 42, pp. 717-728, 2002.

[18] I. W. Sandberg, J. T. Lo, C. L. Fancourt, J. C. Principe, S. Katagiri, S. Haykin, Nonlinear Dynamical Systems: Feed forward Neural Network Perspectives. John Wiley & Sons, CA, 2001, 312 p.

[19] B. V. Dasarathy, Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press., Washington, DC, 1991, 447 p.

[20] T. Tollenaere, "SuperSAB: Fast Adaptive Back Propagation with Good Scaling Properties. Neural Networks," vol. 3, pp. 561-573, 1990.

[21] I. V. Tetko, D. J. Livingstone, A. I. Luik, "Neural Network Studies. 1. Comparison of Overfitting and Overtraining," J Chem Inf Comput Sci, vol. 35, pp. 826-833, 1995.

[22] I. V. Tetko, A. E. P. Villa, “Efficient Partition of Learning Data Sets for Neural Network Training,” Neural Networks, vol. 10, pp. 1361-1374, 1997.

[23] Y. LeCun, J. S. Dencer, S. A. Solla, “Optimal Brain Damage”, in NIPS*2, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1990, pp. 598-605.

[24] Y. A. Chauvin, “Back-Propagation Algorithm with Optimal Use of Hidden Units,” in NIPS*1, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1989, pp. 519-526.

[25] I. V. Tetko, A. E. P. Villa, and D. J. Livingstone, “Neural network studies. 2. Variable selection,” J Chem Inf Comput Sci, vol. 36, pp. 794-803, 1996.

[26] V. V. Kovalishyn, I. V. Tetko, A. I. Luik, V. V. Kholodovych, A. E. P. Villa, and D. J. Livingstone, “Neural network studies. 3. Variable selection in the cascade-correlation learning architecture,” J Chem Inf Comput Sci, vol. 38, pp. 651-659, 1998.

[27] N. K. Mahobia, R. D. Patel, N. W. Sheikh, S. K. Singh, A. Mishra, R. Dhardubey, “Validation Method Used In Quantitative Structure Activity Relationship,” Der Pharma Chemica, vol. 2, pp. 260-271, 2010.

[28] R. D. Cramer III, D. E. Patterson, J. D. Bunce, “Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins,” J Am Chem Soc, vol. 110, pp. 5959-5967, 1998.

[29] I. Sushko, S. Novotarskyi, R. Körner, A. K. Pandey, V. V. Kovalishyn, V. V. Prokopenko, I. V. Tetko, “Applicability domain for in silico models to achieve accuracy of experimental measurements,” J Chemometrics, vol. 24, pp. 202–208, 2010.

[30] R. P. Sheridan, B. P. Feuston, V. N. Maiorov, S. K. Kearsley, “Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR,” J Chem Inf Comput Sci, vol. 44, pp.1912-1928, 2004.

[31] B. S. Drach, E. P. Sviridov, A. A. Kisilenko, A. V. Kirsanov “Interaction of secondary amines with N-acyl-2,2-dichlorvinylamines and N-acyl-1-cyano-2,2-dichlorvinylamines,” J Gen Chem, vol. 9(9), pp. 1818-1824, 1974.

[32] S. A. Chumachenko, O. V. Shablykin, A. N. Vasilenko, V. S. Brovarets, “Synthesis and properties of 5-alkylamino-2-(phtalimidoalkyl)-1,3-oxazol-4-nitriles,” Chem. Het. Comp., vol. 8, pp. 1238-1248, 2011.

[33] B. S. Drach, E. P. Sviridov, Y. P. Shatursky, “Interaction of diethyleters of 1-acylamino-2,2-dichlorvinylphosphonic acids with primary and secondary amines,” J Gen Chem, vol. 44(8), pp. 1712-1715, 1974.

[34] S. Pil’o, V. Brovarets, T. Vinogradova, A. Golovchenko, B. Drach, “Synthesis of new 5-mercapto-1,3-oxazole derivatives on the basis of 2-acylamino-3,3-dichloroacrylonitriles and their analogs,” Russ J Gen Chem, vol. 72(11), pp. 1714-1723, 2002.

[35] I. M. Kopernik, V. M. Blagodatnyj, O. V. Petrenko, L. E. Kalashnikova, V. V. Prokopenko, K. M. Kondratyuk, O. I. Lukashuk, O. V. Golovchenko, S. A. Chumachenko, O. V. Shablykin, L. O. Metelitsa, V.S. Brovarets, “Study in vitro o for antimicrobic activity of new oxazole derivatives and products of its transformations,” Ukr Bioorg Acta, vol. 2, pp. 57-68, 2011.

[36] A. W. Bauer, W. M. Kirby, J. C. Sherris, M. Turck, “Antibiotic susceptibility testing by a standardized single disk method,” Am J Clin Pathol, vol. 45, pp. 493-496, 1966.