"BABEŞ-BOLYAI" UNIVERSITY CLUJ-NAPOCA
FACULTATE OF CHEMISTRY AND CHEMICAL ENGINEERING

RESUME

OF THE PhD THESIS

TOPOLOGICAL SIMILARITY STUDIES

SCIENTIFIC COORDINATOR
Prof.Dr. MIRCEA V. DIUDEA

PhD student
CRISTINA DORINA MOLDOVAN

2012

TABLE OF CONTENTS

Keywords: matrix, index, topological descriptors, biological activity, QSAR, data analyses, PCA, statistic, correlation coefficient, similarity, androstan, benzimidazol, benzoxazol, SIMIL, TOPOCLUJ, HYPERCHEM, DRAGON

INTRODUCTION

The mathematization of chemistry has a long, and colorful history extending back well over two centuries. At any period in the development of chemistry the extent of the mathematization process roughly parallels the progress of chemistry as a while. Thus, in 1874 one of the great pioneers of chemical structures theory, Alexander Crum Brown prophesied that "... chemistry will become a branch of applied mathematics; but it will not cease to be an experimental science. Mathematics may enable us retrospectively to justify results obtained by experiment, may point out useful lines of research and even sometimes predict entirely novel discoveries. We do not know when the change will take place, or whether it will be gradual or sudden...". This prophecy was soon to be fulfilled.

Molecular design of structures with biological or physicochemical properties desired is one of the main objectives of different industries (chemical, pharmaceutical, etc..) Creating a new product with a specific chemical property or biological activity involve great expense both material and human. Drug design is an iterative process that starts with a compound that displays an interesting biological profile and ends with optimized activity profile for molecule synthesis. The process is initiated when the chemist designing a hypothesis linking chemical characteristics of the molecule (or number of molecules), the biological activity. Without a detailed understanding of the biochemical process responsible for biological activity, in general, the hypothesis is refined by examining similarities and structural differences for active and inactive molecules.

In this sense, this thesis develops the concepts used to determine different QSPR models / QSAR and design of new structures on them. It is divided into two distinct parts. The first part refers to the presentation of the molecular topology, methods and techniques used, and the second part refers to the personal contributions in shaping the structure-property relationships, and structure-biological activity:

Chapter I "Topological similarity" shows concept and methods of molecular similarity.
Chapter II "Topological Matrices", present in addition to traditional arrays, new matrix type Szeged and Cluj.

Chapter III "topological indices" deals with topological descriptors derived based on topological matrices. Are presented the main topological indices used and new topological indices proposed, SP, type indices Szeged and Cluj.

Chapter IV "Quantitative Structure-Activity Relationships biological" deals with different methods and models used to describe the physicochemical and biological properties of compounds.

Chapter V "data analysis" refers to statistical processing methods used in designing models QSPR / QSAR. Also are presented a few applications on different sets of structures.
In the own contribution presents the results obtained from the analysis of sets of chemical structures by using methods of similarity and predict the properties of new compounds with potential biological activity.

CHAPTER 1. MOLECULAR SIMILARITY

Similarity of molecular structures expresses the existence of common features in a set of molecules. The similarity is defined based on various criteria and / or procedures also generate intermolecular equivalence classes in the set of molecules. Molecular similarity, as branching is intuitive notion, so you can not define a unique and non-ambiguous measure of similarity. Many areas of science such as synthetic organic chemistry, structural biology, pharmacology and toxicology are required in the drug development process. Very high costs and long lead times characteristic of this process, emphasizes the need to invest in technologies that accelerate the design of new chemical structures with desired biological effects, and to shorten the time until they are put on the market (the same is true for pesticides, fungicides, etc.., chemicals for agriculture). These investments have led to the development of highly sophisticated systems for storing, searching and processing of a diverse range of chemical information.

Molecular description used in molecular similarity analysis is performed using molecular descriptors. Any description of molecular induces a partition of the set of equivalence classes of molecules. We need to define this equivalence relation:
Let S be a set of molecular structures and R a binary relation on S connecting pairs of molecules. If $x, y \in S$ are so linked, then we can write xRy. The relation R is an equivalence relation if it satisfies the following properties:

xRx , for every $x \in S$	(reflexivity)
if xRy, then yRx	(simetry)
if xRy şi yRz, then xRz	(tranzitivity)

Subset of elements $y \in S$, in relation xRy, is the equivalence class of x. Imposing equivalence relation R on set S results in a partitioning of S into disjoint subsets called equivalence classes in R. Such a subset, denoted S / R, it writes and S modulo R.[1]
Let f be a function mapping (coverage, fitness, association, etc..) Separate elements of a set S over some set Y. That is, for any $x \in S$, f assign an appropriate value $y = f(x)$ în Y. This corespondence that be write as $f: S \rightarrow Y$. If Y is the set descriptions, overlap function associated molecular description of each molecule in S. Those molecules of S are equivalent with the same molecular description. Such a function f can be a numbering code or simply a measurement process. It can be shown that different molecular descriptions with their algebraic representation form a group.[1,2]

A matching can be achieved by overlapping the two molecules. Such an operation may indicate aspects of the two molecules or their descriptions. A partial ordering covers some local ordering induced by a cover (matching) partial (ie Substructure matching) between molecules of a set. Mathematical relationship involves ordering antisymmetry property:

if xRy şi yRx than x = y,

Randić[3] reported a partial ordering of alkanes isomers induced airway number (path numbers) p_2 and p_3. Any other graph-theoretical descriptors (eg, topological indices, sequence lengths, and so on) can be used for the characterization and partial ordering and grouping (clustering) molecular structures.
Compounds positioned closer in sequence, is expected to show closer values of some properties (ie, properties similar). Proximity expresses essentially two categories: similarity and dissimilarity. Similarity of two molecules expressed by a large number, if their molecular descriptions are similar and that through a number tending to zero for these descriptions have nothing in common.[4] For example, the superposition of two molecules, the ratio of the number of atoms and bonds that overlap and their total number in the whole molecule, the

molecule multiplied by the same ratio that compares[5] was proposed as a measure of similarity between two moleculesOne such measure is the correlation property (non-zero-one correlation to full correlation).

Dissimilarity expresses the similarity of two molecules, with a near zero when their molecular descriptions are similar and that in a large number if these descriptions are different.[4] In the example above superposing the number of atoms and bonds that do not overlap can be taken as a measure of dissimilarity between two molecules. This particulary caseof disimilarity[6] is know as chemical distance.[7,8, 9, 10]

Recent research conducted in the design of bioactive substances ("drug design"), especially drugs and insecticides, herbicides, etc.., Revealed importance of similarity[1] compounds involved, which I will refer to effectors (E), for their interaction with a biological receptor (R), whatever its nature. Because in drug design is to achieve a quantitative relations between chemical structure and their biological activity effectors E - usually measured as log (1 / C), where C is the molar concentration (dose) that determines a biological response became constant quantifying similarity, respectively chemical dissimilarity. This involves finding appropriate quantitative descriptors can be used in linear relationship (or non-linear) chemical structure - biological activity (QSAR = Quantitative Structure - Activity Relationship) and / or QSAR series design.

Although intuitive idea of similarity, as it was used by chemists - the "measured" by the number of structural features common to both compounds and their mutual arrangement - seems simple, in reality there are several approaches due to multiple possible representations of compounds organic chemicals (E) by the molecular formula, topological - constitutional molecular graphs using planar (2D) or, more precisely, by introducing a Euclidean metric over molecular structures represented in Cartesian space (3D) - Configuration and the conformational formulas, or based on molecular shape as it surfaces described by van der Waals (VDW) or by means of molecular force fields.

The notion of similarity, although it is intuitively simple, depends essentially on the point of view from which it is addressed. For these reasons were proposed for each of the ways of describing the molecular structure above different measures of similarity, able to assess quantitatively the vague (in the mathematical sense). Should be noted that these quantitative indicators of similarity were used almost entirely in the design of bioactive substances. Besides similarity, but closely related to it, there are two notions, however different - and complementary, dissimilarity - also commonly used in QSAR. Each of these concepts should be defined unequivocally the mathematical point of view to avoid any ambiguity. The notion of complementarities was proposed at the end of the last century by E. Fischer, who founded the protein and carbohydrate chemistry to explain the specificity of biological action of molecules is, in theory "key in the lock." Since then the concept of complementarities as a basis for interaction RE, has been continuously developed and refined based on them are proposed a series of QSAR methods (MSA[2], MSD[3], MTD[3], etc.) and molecular modeling (CoMFA[4], etc).

The three concepts mentioned above – similarity, dissimilarity and complementarities - could be placed on a scale similar to that of simple linear correlation a correlation coefficient r = +1 indicates a perfect linear correlation y = f (x), all points are arranged on a straight line with slope strictly positive (overall similarity between molecules Y described by y_i, i=1,N and X characterized by x_i, i=1,N, x_i and y_i are structural descriptors), r = -1 corresponds to a perfect inverse correlation (Y and X are perfect complementare). Considering the scale above, it appears likely that two molecules are dissimilar in terms of, for example, the distribution of atoms in 3D space, topological, 2D, with a correlation coefficient $r \rightarrow 0$ defined from the appropriate metrics and present an advanced similarity with r close to 1 ($r \rightarrow 1$), din punctul de vedere al formei, volumului sau suprafeţei, etc.

Similarity of 2D and 3D molecular structures

Molecular graph is a representation (typically 2D, but this does not diminish the generality of the method) how binding of atoms between them (molecular connectivity). Atoms of a given molecular system formed nodes (peaks) and molecular graphics that chemical bonds between them are similar edges. The molecular graph can be decomposed successively into fragments becoming smaller, ie subgraphs (elements of theory on this issue can be found in Chapter 1 of the first part). Once defined subgraphs of a certain type, comparing two molecular structures to determine the degree of similarity or dissimilarity can be achieved by finding the maximum size common subgraph (MCS) or by considering the type of atoms and their connectivity. This last mode of measurement values obtained may provide greater physical significance, a fact often neglected in QSAR studies.
MCS method can be extended to 3D space by considering the distance matrix. Thus, if each fragment considered distance matrix is associated with appropriate confidence limits, defined following similarity measure, R_{AB}, between molecules A and B:

$$r_{AB} = \frac{MCS_{A,B}}{N_A + N_B - MCS_{A,B}} \qquad (1)$$

In equation (1) NA and NB is the number of atoms of, respectively, the molecules A and B. The main use of this quantitative indicator is related to the search in large data bases containing bioactive molecules, to extract molecules with similar structures.

Concluzion about description ways of similarity

Finding a method of quantitative assessment of similarity is just one of the problems to be solved by the "designer" of bioactive substances. Once established the calculation algorithm is required in order to compare the two molecules, A and B, whose similarity or dissimilarity to be assessed quantitatively. Usually, one of the molecules is kept fixed and the other is rotated to align certain molecular characteristics in order to maximize the similarity of structures A and B.
A difficult problem is to construct an overlay algorithm. One method is based on the rotation of a molecule, B, around Euler axes of the other molecule, A. The method was developed by Oxford Molecular Ltd. under the "ANACONDA". In this case, the method does not give good results when molecular structures are very dissimilar forms. The main difficulties are related to how it is set the center of rotation of the molecule B. The center of inertia is appropriate for a spherical molecule but not for a molecule with ellipsoidal structure. In addition, if the molecule is flexible, difficulties soar as the center of inertia must be continually shifted. In conclusion, the best method of superposition of two dissimilar molecules is to develop methods invariant to rotation and translation. In this respect, methods using distance matrix (or with different topological metrics) are most appropriate to compare points distributed in space because in this case, it is necessary to compare the molecular structures are identical centered and also not required transformations of points in the process of superposition.

CHAPTER 2. TOPOLOGICAL MATRICES

A molecular graph can be represented by: a number, a sequence of numbers, a matrix or a polynomial [11]. These representations are intended to be unique for a given structure. Randić believe[12] that topological matrices can be accepted as a rational basis for the development of topological indices are useful in correlation studies or similarity.

Adiacency matrix

In 1874 Sylvester[3] showed that an suitable numbered organic molecule, can be represented by the adjacency matrix A (G). This is a square table of size N∗**N**, whose elements $\mathbf{[A]_{ij}}$ are defined as follows:

$$[A]_{ij} = \begin{cases} 1 & \text{dacã} \quad i \neq j \quad \text{şi} \quad (i,j) \in E(G) \\ 0 & \text{dacã} \quad i = j \quad \text{sau} \quad (i,j) \notin E(G \end{cases} \tag{2}$$

And the matrix **A(G)**:

$$A(G) = \{[A]_{ij};\ \ i, j \in V(G)\} \tag{3}$$

A(G) characterizes the graph up to isomorphism, of it being able to reconstruct **G**. The matrix **A(G)** is symmetrical about the main diagonal, so its transpose, $\mathbf{A^T(G)}$**,** let adjacency matrix unchanged[13]:

$$\mathbf{A^T(G) = A(G)} \tag{4}$$

Connectivity matrix

To indicate the type of bonding is used for connectivity matrix **C(G)** defined by the relations:

$$[C]_{ij} = \begin{cases} b_{ij} & \text{dacă} \quad i \neq j \quad şi \quad (i,j) \in E(G) \\ 0 & \text{dacă} \quad i = j \quad \text{sau} \quad (i,j) \notin E(G \end{cases} \tag{5}$$

where $\mathbf{b_{ij}}$ is the conventional order of bond: 0; 1; 2; 3; 1.5 for nonbond, simple, duble, triple and aromatic bond. **C(G)** matrix will be:

$$C(G) = \{[C]_{ij};\ \ i, j \in V(G)\} \tag{6}$$

Distance matrix

Distance matrix, **D**(G), is a square sized matrix **N∗N**, whose elements, $\mathbf{[D]_{ij}}$, are defined follow:

$$[D]_{ij} = \begin{cases} numarul \quad de \quad arce \quad pe \quad drumul \\ cel \quad mai \quad scurt \quad (i,j), \quad dacă \quad i \neq j \\ \\ 0 \quad dacă \quad i = j \end{cases} \tag{7}$$

and the matrix **D(G)** will be:

$$\mathbf{D(G) = \{[D]_{ij};\ \ i,j \in V(G)\}} \tag{8}$$

CHAPTER 3. TOPOLOGICAL INDICES

A number representing a chemical structure in graph-theoretic terms is called topological descriptor. Being a structural invariant, it is independent of the numbering of atoms or pictorial representation of molecular graphics. Despite the considerable loss of information by "design" as the number one structure, such invariants have found wide applications in the correlation and prediction of many molecular properties[14,15] and also in the simmilarity and isomorphism tests[16,12]. When a topological descriptor correlates with molecular property it can be called index or molecular topological index (***IT***).

Indices based on adiacency matrix

Indicele RANDIC

Indices which operatedin chain (in particular arcs) are named connectivity indices. χ index (connectivity, simmilar with M_2) was introduced by Randic to characterizing bond in the graphs[19]:

$$\chi = \sum_{(ij)\in E(G)} (k_i * k_j)^{-1/2} \quad (9)$$

Diudea[6] defined index χ on vertices:

$$\chi_i = \sum_{j:(ij)\in E(G)} (k_i * k_j)^{-1/2} \quad (10)$$

$$\chi = \frac{1}{2}\sum_i \chi_i \quad (11)$$

Indices based on distance matrix

WIENER index

Wiener[30] defined W as "the sum of the distances (number of carbon-carbon links) between any two carbon atoms in the molecule." In acyclic structures, the author calculated as the sum of contributions W "per bond" ("bond contributions" that correlated with the thermodynamic properties of acyclic hydrocarbons)

$$W = W(G) = \sum_e W_e = \sum_e N_{L,e} * N_{R,e} \quad (11)$$

where:

$$N_{L,e} + N_{R,e} = N(G) \quad (12)$$

N_L, şi N_R are the number of vertices left and right of the edge **e**, summation being made by all peaks in **G**.

Molecular simmilarity indices

The chemical structure of each molecule is decoded into a set of n structural descriptors (SD) collected in an array of typeX=X(A),

$$X(A) = \{SD_1, SD_2, SD_3, ...SD_n\} \quad (13)$$

For a set of molecules M, M = {A, B, C, ...} all structural descriptors are collected in a matrix of size mxn where each row and each column corresponds to a molecule corresponds to a particular structural descriptor. For the calculation of similarity indices, structural descriptors can be scaled Z. Score method (autoscaling) which gives values that have mean zero and are scaled to unit variance.

Cosine simmilarity index

Cosine index, C_s are the similarity between two molecules A and B, and is given by:

$$C_S(A,B) = \frac{\sum_{i=1}^{n} X(A)_i X(B)_i}{\left[\sum_{i=1}^{n} X(A)_i^{\,2} \sum_{i=1}^{n} X(B)_i^{\,2}\right]^{1/2}} \quad (14)$$

C_S take value between [-1,1]. Carbo[17] used a form of cosine similarity index defined on integral electron density throughout space.

Dice simmilarity index

Dice index, D_S for the similarity between two vectors of two structural descriptors X (A) and X (B) is given by:

$$D_S(A,B) = \frac{2\sum_{i=1}^{n} X(A)_i X(B)_i}{\sum_{i=1}^{n} X(A)_i^2 + \sum_{i=1}^{n} X(B)_i^2} \quad (15)$$

with properties that $-1 \le D_S \le 1$.

Richards simmilarity index

Richards simmilarity index are defined as:

$$R_S(A,B) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|X(A)_i - X(B)_i|}{\max(|X(A)_i|, |X(B)_i|)}\right) \quad (16)$$

CHAPTER 4. QUANTITATIVE STRUCTURES-BIOLOGICAL ACTIVITY RELATIONSHIP

Modern methods used to "design" of molecular structures with biological activity specified, eg drugs, insecticides, herbicides and fungicides are based on quantification bioactivity as a function of molecular structure[18]. This approach has its origins in the work of Meyer and Overton[19] of the last century and early this century. Thus, they successfully demonstrated for the first time a dependence of bioactivity of physico-chemical parameter, partition coefficient, which is a function of molecular structure.

Introducing the concept of biological receptor situs was a vital element of exceptional importance to this area of research, it was inferred from Langley,[20] but founded and developed by Ehrlich,[21] the father of chemotherapy. According to this model, biological activity depends on bioactive substrate recognition (effect) by receptor active site, this is followed by effectors binding in situ.

Dependence of bioactivity and configuration discovery[22] led to the recognition that steric effects, regardless of the type or nature, play an essential role in receptor-effector interactions, conditioning and modeling biological potency of effector.

Advaced QSAR models

A bioactive compound introduced into a living organism, induce a biological response, a specific response from the body. The answer is conditioned by the structure and chemical identity of bioactive compounds.

Bioactive compound interaction with the body at the molecular level is the so-called biological receptors. They are situs active sites of protein macromolecules located within the living cell or cell membranes.

Biological receptors[23,24,25,26] have the following characteristics:

(i) specificity: receptors recognize active enantiomers ("eutomeri" unlike "distomeri" not biologically active) or diastereoisomers agonist or antagonist;

(ii) saturability: number of active site on a cell formation is finite;

(iii) generally is in cell receptors that generate biological response.

E. FISCHER (1894) formulated the first theory of the interaction of bioactive compounds (effects) to the handset. According to this, the handset is seen as a rigid cavity in which effector must "fit" like "key in the lock" ("key in lock"). Later, he admitted that the receptors are semi-rigid, they seek mutual optimization (albeit limited) with its effectors. Both partners effector-receptor complex deviate from the minimum energy conformational state to achieve the most stable complex (eg induced biological response).[149,150,151].

Biological response caused by a chemical penetrated the body is proportional to the concentration of compound **E. .. R** (effector-receptor), resulting in the reaction:

$$E + R \xleftrightarrow{K} E...R \xrightarrow{k} P \tag{17}$$

While **E. .. R** is complex, there is a specific response of the body, called biological response. The complex can dissociate into components (equilibrium characterized by the constant **K**) or shape (characterized by constant speed **k**) **P** product concentration while it varies according to the relation:

$$\frac{d[P]}{dt} = k[E...R] \tag{18}$$

complex concentration can be approximated in equilibrium relationship:

$$[E...R] = [E][R]K = [E][R]\exp(-\Delta G / RT) \tag{19}$$

CHAPTER 5. DATA ANALYSES

Correlation coefficient. Theoretical correlation coefficient ρ_{xy} of two variables **x** and **y** are the corresponding normalized covariance variables:

$$\rho_{xy} = \operatorname{cov}\left(\frac{x-\bar{x}}{\sigma_x}, \frac{y-\bar{y}}{\sigma_y}\right) = \operatorname{Med}\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\right] = \frac{Med\left[(x-\bar{x})(y-\bar{y})\right]}{\sigma_x\sigma_y} = \frac{\operatorname{cov}(x, y)}{\sigma_x\sigma_y} \tag{20}$$

Valorile lui fiind cuprinse în domeniul: $-1 \leq \rho_{xy} \leq 1$. În cazul unei selecţii, relaţia (20) se va scrie:

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{S_x S_y} = r_{xy} \tag{21}$$

where r_{xy} are the empirical correlation coeficient. When variables x and y are independent, cov (x, y) = 0 and rxy = 0. The converse is not true. We emphasize that ρ (theoretical correlation coefficient) refers to the entire population and r (correlation coefficient empirically) refers to a selection.

Correlation coefficient (regression) globally, R^2, is given by the ratio dispersion values calculated from the average variance from the mean empirical values:

$$R^2 = 1 - \frac{\sum_i (y_i - y_{i,calc})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (y_{i,calc} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{22}$$

R^2 ia valori în domeniul [0, 1]. Valori apropiate de 1 arată că dependenţa liniară este potrivită pentru descrierea relaţiei dintre y şi x.

Liniar regressions

Let the set of chemical structures C_1, C_2,. . ., Cn and observed values of molecular properties y_1, y_2,. . ., Y_n (y_i corresponds compound C_i). Estimation of y_i with independent

variables x_{ij} (also called predictor or explanatory variables) can be made according to the relation:

$$y_{i,calc} = b_0 + \sum_{j=1}^{m} b_j x_{ij} \quad (23)$$

x_{ij} variables numeric encoding structural features (topological) or physico-chemical j - present in the compound C_i.

Fischer F statistics, indicates the significance of the regression equation. F is calculated by the formula:

$$F = \frac{SS_{reg}}{SS_e /(n-k)} \quad (24)$$

where SS_{reg} is the sum of squared errors attributed to regression:

$$SS_{reg} = \sum_i (y_{i,calc} - \bar{y})^2 \quad (25)$$

Student t statistics. t indicates the significance of the coefficients b_j, and it is calculated by the formula:

$$t_j = \frac{|b_j|}{\sigma_{bj}} \quad (26)$$

where σ_{bj} is the standard error of regression coefficient b_j. From comparing the calculated values tabulated t_j (for materiality required, which is a function of the degrees of freedom of regression - see estimator F) is validated or not a variable x_{ij} for contribution to global correlation.[27,28].

Clusters analyses

Objectives of clusters analyses

The notion of cluster analysis (CA) falls into a family of methods that is mainly used for finding and highlighting structures within data. In this way cluster analysis can be seen as a method called knowledge model. It used to it a name synonymous with auto numbers and taxonomic classification. If the purpose of data analysis or questions to be put on them is clearly established and leads to achieve results, the next question is: what causes the structure found?

Suppose we have a reasonable number of objects, n, and their properties have been selected and arranged in matrix X as in the example below:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \quad (27)$$

Similar units and preprocessing information

In order to find data structures or discovery group similarities samples, bodies ... which below will be called objects, first of all we need similar units. The simple geometry can be derived from similar units. Without demonstrate that intuitive concepts of similarity and distance are complementary in nature and remembering the law of Pythagoras, the distance d between two points O1şi O2 in a rectangular two axes x and y is:

$$d(O_1, O_2) = \sqrt{(y_1 - y_2)^2 + (x_1 - x_2)^2} \quad (28)$$

That are represented in figure 1.

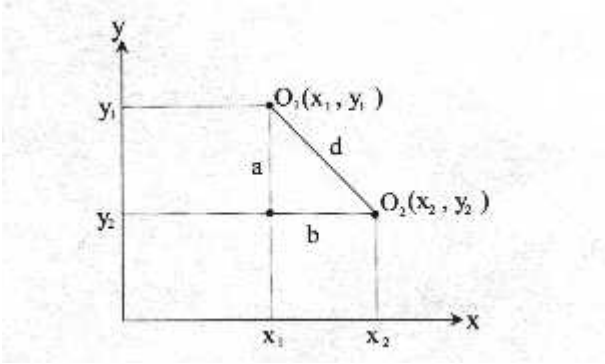


Figure 1. The distance between two objects in space plan (the law of Pythagoras)

Extent of the law over more than two dimensions of "space Pythagoras' lead to Euclidean distance of any two objects O_i and O_k that below will be written simply as D(i, k):

$$d(i,k) = \sqrt{\sum_{j=1}^{m} \left(x_{ij} x_{kj}\right)^2} \tag{29}$$

where: *m*- number of features.

Factors analyses

Factor analysis has emerged as a method of reducing herd of variables proposed to describe an area by the construction of new variables (called factors) in much smaller numbers, and determining mathematical relationships specifying the relationship between the original variables and factors so that by these new variables to reproduce, to the greatest extent, the information provided by the original variables. The method dates back to early last century and is due to K. Pearson (1901) and C. Spearman (1904).
Statistical techniques known under that name have the common goal to reduce the number of variables that characterize a lot of items to a smaller number of variables, usually different from the original.

Base equation of the factor analyses
From the matrix of standardized variables can be inferred correlation matrix of the original variables:

$$R = \frac{1}{n} Z \cdot Z^T, \; r_{ij} = \frac{1}{n}\left(z_{i1} z_{j1} + z_{i2} z_{j2} + ... + z_{in} z_{jn}\right) \tag{30}$$

Number of factors extracted

One way of empirical mode decomposition with k factors sufficiently defined by Malinowki (1977) by IND function. Minimum IND probably indicates the number of relevant factors.

$$IND = \frac{RE}{\left(m-k\right)^2} \tag{31}$$

$$RE^2 = \frac{1}{n\left(m-k\right)} \sum_{j=k+1}^{m} \lambda_i \tag{32}$$

where m – number of variables
n – object numbers
k – number of factor extracted
RE – „real error”

PERSONAL CONTRIBUTIONS

Techniques of quantitative structure-activity relationships (QSAR) are essential in all aspects of research on molecular interpretation of biological properties.[29] It is understood that the physical, chemical or biological properties of compounds depend on the arrangement of 3D (three-dimensional) of atoms in the molecule. Ability to produce quantitative correlation between 3D molecular structure and biological activity is important in choosing the synthesis of biologically active substances.[30]

The biological activity of steroids varies considerably with very slight modification of the structure. These families of molecules important characteristics very changeable for any method of prediction, due to low relative flexibility cyclopentanoperhidrofenantrene skeleton. For this reason, many QSAR models based on 2D properties such as topological descriptors have quality comparable models from complex 3D methods.[31,32]

In this part are identified and presented aspects of molecular structure are particularly relevant for biological activity (receptor binding affinity) for different classes of substances with biological activity.

Steroidic compounds

Set of 31 steroid structures (androstane) AS (Figure 2, Table 1), with affinity corticosteroid-binding globulin (CBG) was taken from publications of Dunn *et al*.[33] and Tuppurainen, *et al*.[34] This set of structures was often used for performance evaluation of new methods of QSAR analysis. However, citing other authors, we can say that many publications that have used this set included errors in the structure of steroids. Structures used in this study were checked very carefully in order to avoid further propagation of errors. Quality, molecules with substituents such as oxygen or hydroxyl in position 17 of the steroid structure have increased CBG activity, while the presence of a large chain like-$COCH_2OH$ leads to decrease this activity.

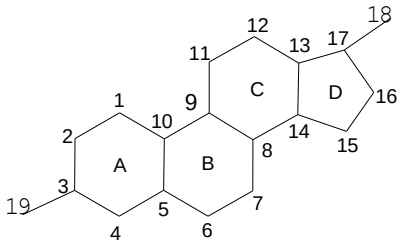


Figure 2. Androstan structure

Table 1. Set of AS structures:

	Compound	Activity		Compound	Activity
1	aldosterone	6.279	17	pregnenolone	5.255
2	androstanediol	5.000	18	17-hydroxypregnenolone	5.000
3	androstenediol	5.000	19	progesterone	7.380
4	androstenedione	5.763	20	17-hydroxyprogesterone	7.740
5	androsterone	5.613	21	testosterone	6.724
6	corticosterone	7.881	22	prednisolone	7.512
7	cortisol	7.881	23	cortisol 21-acetate	7.553
8	cortisone	6.892	24	4-pregnene-3,11,20-trione	6.779
9	dehydroepiandrosterone	5.000	25	epicorticosterone	7.200
10	deoxycorticosterone	7.653	26	19-nortestosterone	6.144
11	deoxycortisol	7.881	27	16R,17-dihydroxy-4-pregnene-3,20-dione	6.247
12	dihydrotestosterone	5.919	28	16-methyl-4-pregnene-3,20-dione	7.120
13	estradiol	5.000	29	19-norprogesterone	6.817
14	estriol	5.000	30	11β,17,21-trihydroxy-2R-methyl-4-pregnene-3,20-dione	7.688
15	estrone	5.000	31	11β,17,21-trihydroxy-2R-methyl-9R-fluoro-4-pregnene-3,20-dione	5.797
16	etiocholanolone	5.255			

Subset electronic parameters include molecular descriptors derived from partial atomic tasks. With TOPOCLUJ program, partial charges Ch_i were calculated as follows:

$$Ch_{i,j} = \log(S_j / S_i)^{1/(d_{i,j})^2} \quad (33)$$

$$Ch_i = \sum_j ch_{i,j} \quad (34)$$

In both cases $\boldsymbol{S_i}$, $\boldsymbol{S_j}$ is calculated group electronegativity Sanderson hybrid groups (eg heavy atoms are surrounded by hydrogen atoms) in the molecule, while $\boldsymbol{d_{ij}}$ is the Euclidean distance separating atoms i and j in chemical structure optimized minimum energy (HyperChem).[137] $\boldsymbol{Ch_{i,j}}$ is the atomic electronegativity perturbation caused by atom *i* to any atom *j* of molecule while $\boldsymbol{Ch_i}$ este is the result of these disturbances on atom *i*.[35,36]

Any steroid compound can be described by the partial load characteristic positions substituted or unsubstituted and heteroatoms. Based on this we defined a new flexible global descriptor (***CD***) can be defined as a function of additive weighted autocorrelation of atom *j* corresponding partial tasks considered:

$$CD = \sum_j c_j \cdot Ch_j \quad (35)$$

where $\boldsymbol{c_j}$ is the regression coefficient given by multivariate regression $\mathbf{log(A_i obs) = f(Ch_j)}$. This "ad-hoc" weighting depends on the set of molecules considered and also using local descriptors. Part load ($\boldsymbol{Ch_j}$) corresponds to the following positions on the basic structure: 3, 10, 11, 13, 17, 18, 19 (Figure 2). *Dragon 2.1*[136] software was used to calculate 1600 molecular descriptors for the compounds studied. The most relevant of these descriptors used here are those radial distribution functions (RDF) autocorrelation indices and geometrical descriptors.

Descriptors belonging to the class of radial distribution function are based on distance distribution of the geometric representation of the molecule. In addition to the interatomic distances across the molecule, RDS provide information about bond distances, ring types,

planar or planar systems, types of atoms and other important structural changes. By using different weighting schemes, including types of atoms, electronegativity, atomic mass (*RDF090m*) or range van der Walls, RDF can be adjusted to give the important developments of descriptors in QSAR studies.
The second group of descriptors is obtained by applying the two-dimensional autocorrelation function of the molecular graph. Such descriptors express the correlation between numerical values that are statistically weighted using atomic properties at equal intervals of time value.[37] For example, *MATS1p*-Moran-lag auto-correlation 1/pondered with atomic polarizability; *MATS4e*-lag auto-correlation Moran atomic Sanderson electronegativity 4/pondered with. Sanderson electronegativity applications as weights taken in this regard in some cases changing distribution within the molecule.
Geometric descriptors indicating the size of the molecule, they are derived from three-dimensional coordinates of the atomic nuclei and atomic masses and / or atomic distance in the molecule.

Data processing
Given the complexity of interactions between receptor molecules and molecules with potential inhibitor is difficult to model the set of structures using only simple linear regression models. In order to analyze the obtained data was proposed the following model:

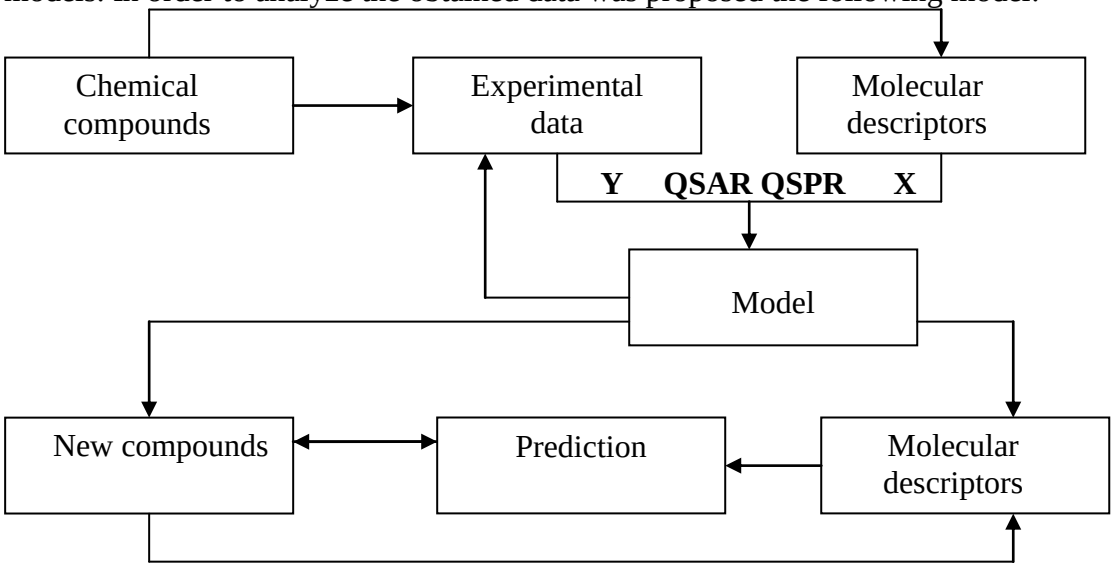


Figure 3. Schematic representation of the model construction.

QSAR analysis consists of the following steps:

(i) structure optimization using semi empirical PM3;
(ii) calculation of molecular descriptors;
(iii) multivariate regression to find the autocorrelation coefficients;
(iv) dividing the data set in one school (regression calibration) and a prediction (for model validation);
(v) testing the ability to predict model;
(vi) principal component analysis (PCA);
(vii) finding a regression model features;
(viii) testing the predictive ability of the model;
(ix) interpretation model.

In both regressions monovariable and bivariable (Table 2) compound 13 appears to be outlier. This compound was not included in further analyzes. With this outlier excluded, we observed an improvement in correlation with the descriptors given below.

Table 2. Models used and results of the cross-validation

Data sets	**Number of observable (*n*)**	**Model**	**Correlation coefficient (r^2) (before LOO)**	**Outlier structures**	**Correlation coefficient (r^2) (after LOO)**
AS	31	$\log(P_i\text{calc}) = f(CD)$	0.891	13	0.920
		$\log(P_i\text{calc}) = f(CD, L/Bw)$	0.931	13	0.939

Trivariable regression gave similar results but did not bring essential improvements. The results will be presented on.

AS structurs set

We divided the set of AS structures in two sets: set school (n = 20) and prediction set (validation) (n = 11) as shown in Table 3.

(a) School set (n = 20)

CD electronic descriptor, current ***CDP***, was calculated de nuovo on the school set given by equation 6.7, it's because correlation weights c_j according to the selected property and only given set (in this case school kit with 20 structures). Table 3 presents the most relevant descriptors for the set of data structures. The best models for the set of AS structures are shown below.

Monovariable regression

$$\log P_i\text{calc} = 7.236 + 1.033 \cdot \text{CDP}_i \quad (36)$$

n = 20 $R^2 = 0.903$ s = 0.10 F = 159.99

Bivariable regression (37)

$$\log P_i\text{calc} = 5.737 + 0.914 \cdot \text{CDP}_i + 0.194 \cdot \text{L/Bw}_i$$

n = 20 $R^2 = 0.931$ s = 0.31 F = 114.12

Multivariable regression (38)

$$\log P_i\text{calc} = 6.268 + 0.17 \cdot \text{L/Bw}_i - 0.166 \cdot \text{RDF090m}_i + 0.904 \cdot \text{CDP}_i$$

n = 20 $R^2 = 0.962$ s = 0.24 F = 129.73

Table 3. Topological descriptors and logP partition coefficient for the set of structures observed AS

Structures	L/Bw	RDF090m	CDP	log *P* obs.
School set				
1	7.3	3.17	-1.722	5
2	6.5	3.127	-2.101	5
4	6.5	3.505	-2.062	5
5	6.9	1.755	-2.456	5
7	6.3	2.803	-2.188	5
8	4.3	0.191	-1.950	5.255
9	7.1	1.91	-2.087	5.255
12	7	4.01	-0.880	5.797
13	6.1	3.497	-0.187	5.919
14	7.2	0.61	-1.212	6.144
17	6.9	2.729	-0.336	6.724
18	6	3.288	0.492	6.779
20	7.1	3.976	-0.243	6.892
22	6.6	2.525	-0.046	7.2
23	7.1	0.942	0.085	7.38
25	9.5	1.704	0.205	7.553
26	9.1	0.673	0.006	7.653
28	7.8	1.122	0.317	7.74
29	8.7	2.31	0.040	7.881
30	9	1.077	0.309	7.881
Prediction set (validation)				
3	6.7	2.669	-2.350	5
6	6.6	0.811	-2.117	5
10	6.3	1.338	-2.184	5.613
11	6.2	1.272	-0.479	5.763
15	6.9	1.986	0.400	6.247
16	6.4	1.759	-0.716	6.279
19	9.1	0.406	-0.810	6.817
21	6.9	2.014	0.046	7.12
24	5.9	2.777	0.110	7.512
27	6.9	3.512	0.675	7.688
31	8.7	1.423	0.225	7.881

(b) Prediction set (validation) (n = 11)

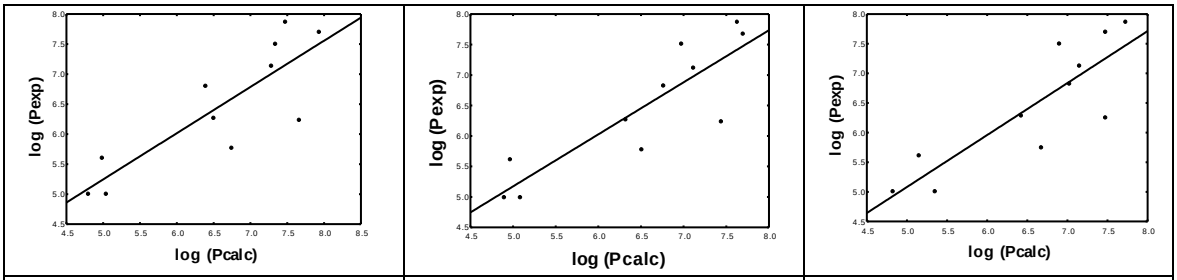
Any QSAR model must be validated with a set of external prediction. *CDP* descriptor calculation of the standard prediction was based on partial loads generated for the set of parameters c'_j school (see Table 4.). We assume that a set of predictive biological activity is unknown.

Table 4. Prediction set for AS structures (n = 11)

Structures	log $P_i$obs	log $P_i$calc (eq. 6)	log $P_i$calc (eq. 7)	log $P_i$calc (eq. 8)
3	5.000	4.808	4.890	4.836
6	5.000	5.049	5.083	5.339
10	5.613	4.980	4.964	5.140
11	5.763	6.741	6.504	6.676
15	6.247	7.650	7.444	7.472
16	6.279	6.496	6.326	6.415
19	6.817	6.400	6.765	7.015
21	7.120	7.284	7.121	7.147
24	7.512	7.350	6.984	6.907
27	7.688	7.934	7.696	7.466
31	7.881	7.468	7.633	7.713
R^2		**0.716**	**0.766**	**0.728**
CV%		7.089	5.010	6.586

To set the predictive ability of AS structures seems to be much better for bivariable regression (Table 4).

Figure 4. a-c presents the experimental vs. calculated values for binding affinity receptor structures AS: (a) values calculated according to equation 36 (b) values calculated according to equation 37, (c) values calculated according to equation 38.



a b c

(a) log ($P_i$obs) = 1.3932 + 0.77043 × log ($P_i$calc); R = 0.84636;
(b) log ($P_i$obs) = 0.90106 + .85433 × log ($P_i$calc);R = 0.87565;
(c) log ($P_i$obs) = 0.69650 + 0.87704 × log ($P_i$calc); R = 0.85352

Figure 4. Graphical representation of experimental values versus calculated values of binding affinity receptor structures AS set

In order to explain the contribution of each substituent of the graph from the set of structures AS to receptor binding affinity (CBG), I created a simple electronic descriptor based on partial atomic tasks, linked in part with the property studied.

Modelele QSAR descrise în acest studiu indică faptul că acest descriptor global este unul QSAR models described in this study indicates that the global descriptor is one of the most significant in predicting the activity of our compounds. It can indicate the most important positions substitutes. Thus, **CD** calculated for atoms in positions above without substituents in position 17 of the steroid structure, leading to a variance of 25% of CBG activity in this set of structures including **AS** while position 17 increases to 89%.

The model was validated with an external set of prediction. The molecular activity using descriptor derived for **CD** and descriptors obtained by PCA's factor loading is comparable with other models described in the literature, with a good predictive ability. Note that the

simple 2D like this that we've developed is comparable to results obtained with complex 3D models (COMF, COMSA, GRIND, Eeva etc.)[27,38,39,] which require more resources computing.

Antibacterial compounds

Set of 38 2-furilethilene derivatives, with antibacterial activity was taken from Miguel Angel Cabrera Pérez[40] and Yovani Marrero Ponce[41] publications. 2-furiletilenă derivatives are biologically active broad-spectrum antimicrobial, antispasmoic, but that in some cases cytotoxic and carcinogenic mutagenetic activities (Yahagi *et al*, 1974; Dore şi Viel, 1975[42]; Miyaji, 1976; McCalla, 1979; McCalla, 1983; Kelloval *et al*, 1984; Estrada, 1998). The interest in studying derivatives of 2-furiletilenă increased in recent years as a consequence of the discovery of new compounds with potential microcidal with this chemical structure (McCoy and Thornburgh, 1992; Castañedo *et al*, 1994[43]; Blondeau *et al*, 1999[44]).
The model used here is shown schematically in Figure 5.

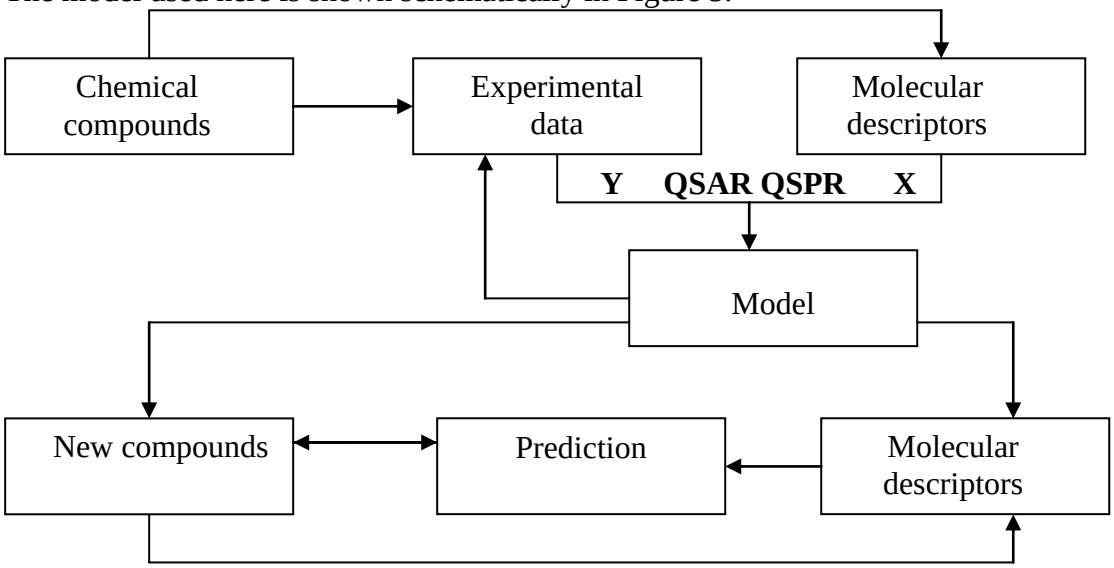
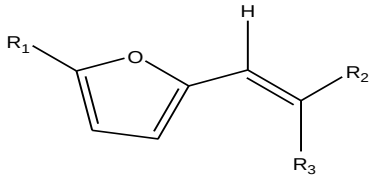


Figure 5. Schematic representation of the model.

Dragon 2.1 software was used to calculate the compounds studied 1,600 molecular descriptors. What most relevant of these descriptors used here are the constitutional (MW - molecular weight), connectivity (X4v - the valence connectivity index chi-4) and geometrical (G (N.. A) - the sum of geometric distances between N and O). TOPOCLUJ topological descriptors calculated by the program are derived from topological matrices or polynomials describing molecular graphs. Algorithms include some weighting schemes molecular graphs with partial atomic tasks, electronegativity and masses of molecular fragments. Topological matrices calculated on the base include: adjacency, distance, Detour, connectivity, Wiener, etc. and TOPO Group Cluj matrices developed including: Cluj Cluj fragments, matrix W, LM, MS, matrix operator W (M1, M2, M3). The most relevant of these descriptors are:
- *IE[CfMax[Density]]*,
- *VAA1*,
- *VAD1*,
- *CS[LM[Electronegativity]]*,
- *CS[Sh[W4[Charge_Adjacency]]]*.

Table 5. 2-furilethilene derivatives.



	R_1	**R_2**	**R_3**	**log P**
1	H	NO_2	$COOCH_3$	1.879
2	CH_3	NO_2	$COOCH_3$	2.439
3	Br	NO_2	$COOCH_3$	2.739
4	$COOCH_3$	NO_2	$COOCH_3$	1.869
5	NO_2	NO_2	$COOCH_3$	1.599
6	NO_2	$COOC_2H_5$	$COOC_2H_5$	2.504
7	NO_2	H	NO_2	1.303
8	H	H	NO_2	1.583
9	NO_2	H	$CONHC_2H_5$	1.386
10	NO_2	H	$CONH(CH_2)_2CH_3$	1.86
11	NO_2	H	$CONHCH(CH_3)_2$	1.803
12	NO_2	H	$CONH(CH_2)_3CH_3$	2.356
13	NO_2	H	$CONHCH_2CH(CH_3)_2$	2.225
14	NO_2	H	$CONHCH(CH_3)C_2H_5$	2.284
15	NO_2	H	$CONHC(CH_3)_3$	2.333
16	NO_2	H	$CONHCH_2C(CH_3)_3$	2.605
17	NO_2	H	$COOCH_3$	1.652
18	NO_2	H	$COOC_2H_5$	2.098
19	NO_2	H	$COO(CH_2)_2CH_3$	2.673
20	NO_2	H	$COOCH(CH_3)_2$	2.641
21	NO_2	H	$COO(CH_2)_3CH_3$	2.827
22	NO_2	H	$COOCH_2CH(CH_3)_2$	3.135
23	NO_2	H	$COOCH(CH_3)C_2H_5$	3.091
24	NO_2	H	$COOC(CH_3)_3$	3.06
25	NO_2	H	$COO(CH_2)_4CH_3$	3.404
26	NO_2	H	Br	2.447
27	NO_2	H	CN	1.05
28	NO_2	H	OCH_3	1.591
29	NO_2	H	H	1.611
30	NO_2	CN	$COOCH_3$	1.488
31	I	NO_2	$COOCH_3$	2.999
32	NO_2	H	$CONH_2$	0.649
33	NO_2	H	$CONHCH_3$	0.984
34	NO_2	H	$CON(CH3)_2$	0.819
35	Br	NO_2	Br	2.820
36	Br	NO_2	CH_3	2.730
37	H	NO_2	H	1.290
38	H	NO_2	CH_3	1.940

Analyses and data processing

N-octanol/water partition coefficient (log P) plays an important role in understanding the biological behavior of these derivatives of 2-furilethilene. Given the complexity of interactions between receptor molecules and molecules with potential inhibitor is difficult to model the set of structures using only simple linear regression models. In SIMIL software package filtering procedures are implemented using sequences valence peaks used clusteration chemical systems.

QSAR analysis consists of the following steps:

(i) structure optimization using semiempirical PM3;
(ii) calculation of molecular descriptors;
(iii) dividing the data set in one school (regression calibration) and a prediction (for model validation) based on molecular similarity;
(iv) principal component analysis (PCA);
(v) testing the ability to predict model;
(vi) finding a regression model features;
(vii) testing the predictive ability of the model;
(viii) the performance model.

Tabelul 6. Similarity for the set of 2-furilethilene derivatives to elected leadership structure prediction set and their n-octanol/water partition coefficient (log P).

Number of the structures	35	36	37	38	log P
1	0.59524	0.72024	0.71429	0.78571	1.879
2	0.55556	0.67222	0.66667	0.73333	2.439
3	0.67222	0.8	0.66667	0.73333	2.739
4	0.462963	0.56019	0.55556	0.61111	1.869
5	0.490196	0.59314	0.58824	0.64706	1.599
6	0.459375	0.48151	0.45125	0.50114	2.504
7	0.64103	0.64103	0.76923	0.6993	1.303
8	0.83333	0.83333	1	0.90909	1.583
9	0.50139	0.6125	0.60167	0.66818	1.386
10	0.47005	0.57422	0.56406	0.62642	1.86
11	0.47005	0.57422	0.56406	0.62642	1.803
12	0.490196	0.54044	0.53088	0.58957	2.356
13	0.490196	0.54044	0.53088	0.58957	2.225
14	0.490196	0.54044	0.53088	0.58957	2.284
15	0.490196	0.54044	0.53088	0.58957	2.333
16	0.462963	0.51042	0.50139	0.55682	2.605
17	0.5372	0.65625	0.64464	0.71591	1.652
18	0.50139	0.6125	0.60167	0.66818	2.098
19	0.47005	0.57422	0.56406	0.62642	2.673
20	0.47005	0.57422	0.56406	0.62642	2.641
21	0.490196	0.54044	0.53088	0.58957	2.827
22	0.490196	0.54044	0.53088	0.58957	3.135
23	0.490196	0.54044	0.53088	0.58957	3.091
24	0.490196	0.54044	0.53088	0.58957	3.06

25	0.462963	0.51042	0.50139	0.55682	3.404
26	0.83523	0.68371	0.82045	0.74587	2.447
27	0.62674	0.76563	0.75208	0.83523	1.05
28	0.62674	0.62674	0.75208	0.68371	1.591
29	0.75208	0.75208	0.9025	0.82045	1.611
30	0.47005	0.57422	0.56406	0.62642	1.488
31	0.55556	0.67222	0.66667	0.73333	2.999
32	0.57853	0.70673	0.69423	0.77098	0.649
33	0.5372	0.65625	0.64464	0.71591	0.984
34	0.50139	0.6125	0.60167	0.66818	0.819
35	1	0.84028	0.83333	0.75758	2.820
36	0.84028	1	0.83333	0.91667	2.730
37	0.83333	0.83333	1	0.90909	1.290
38	0.75758	0.91667	0.90909	1	1.940

Because the correlation coefficient is subject to fluctuations selection, high r value should be treated with caution if the number of observations is small and, moreover, it cannot be used as a comparison for equations with several different data.
In Table 6.6 are observed molecular fingerprint similarity chosen set of predictive molecules from other structures in the base set. We can see that the four structures can be part of a cluster as the similarity between them is very high. We also set school especially those structures that have similarity coefficient greater than 0.70 for at least one of these structures.

School and prediction set
School set structures are presented in Table 7, structures with similarity coefficients up to the leader structures (Table 8) of the set of prediction, is presented here and the most significant descriptors calculated by Dragon software with n-octanol partition coefficient / water logP.

Table 7. Topological descriptors and observed partition coefficient log*P* for the school set.

Number of the structures	MW	X4v	G(N..O)	logP
1	197.16	0.799	9.619	1.879
2	211.19	0.956	9.613	2.439
3	276.05	1.15	9.626	2.739
8	139.12	0.52	4.386	1.583
17	197.16	0.748	14.97	1.652
26	218.01	0.853	0	2.447
27	164.13	0.649	22.162	1.05
28	169.15	0.665	6.598	1.591
29	139.12	0.55	0	1.611
31	323.05	1.265	9.641	2.999
32	182.15	0.694	29.088	0.649
33	196.18	0.771	29.063	0.984

Table 8. Topological descriptors and observed partition coefficient log*P* for the prediction set.

Number of the structures	MW	X4v	G(N..O)	logP
35	296.9	1.121	4.405	2.49
36	232.04	0.984	4.516	2.37
37	139.12	0.52	4.386	1.56
38	153.15	0.632	4.517	1.92

Used equation for prediction is from calibrated cluster school (their shapes normalized in Matlab). This equation show good correlation coefficient (R^2 = 0.9843 for bivariable regression, eq. 39 and R^2 = 0.98653 for multivariable regression, with 3 descriptors, eq. 40). To note is that in previous studies on the best model obtained in the prediction of log P were used multiple regression equations 7 descriptors (eq. 17 from Yovani Marrero Ponce et all.) with R^2 = 0.968 and prediction ability R^2 = 0.938, much lower than our results..

Bivariable regression

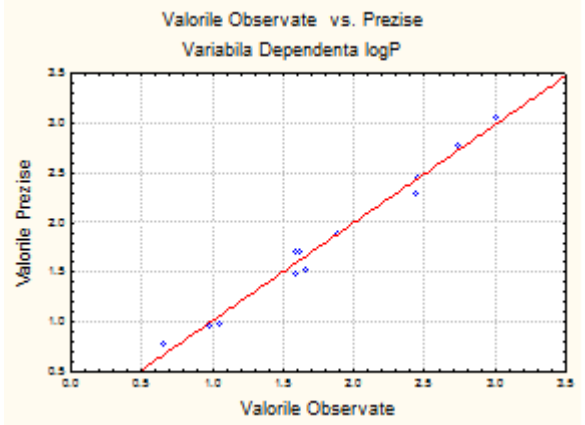
logP= 0.086403+3.442395· **X4v**+2.163165· **G(N..O)** (39)

n = 12 R^2 = 0.9843 s = 0.091837 F = 282.3465

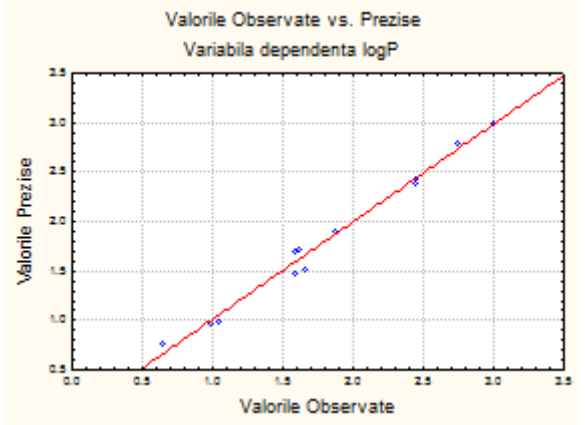
Multivariable regression

logP= 0.391261-0.003500· **MW** +3.287415· **X4v**-0.043587· **G(N..O)** (40)

n = 12 R^2 = 0.98653 s = 0.07885 F = 195.3105



a) R^2 = 0.9843, eq. 6.16



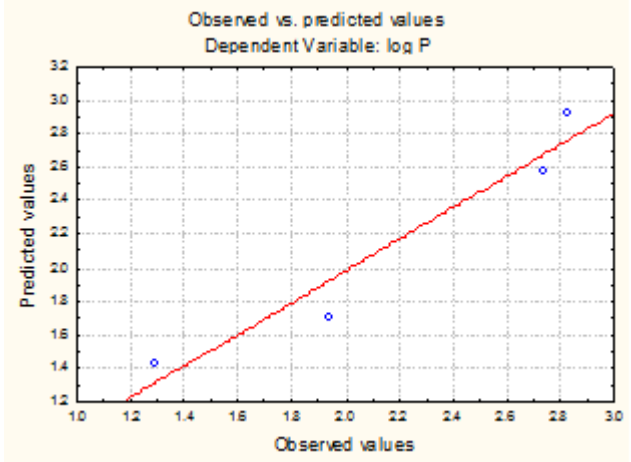
b) R^2 = 0.98653 eq. 6.17

Figure 6. Plot of the observed vs. Predicted values for eq. 39 and 40 for school set.

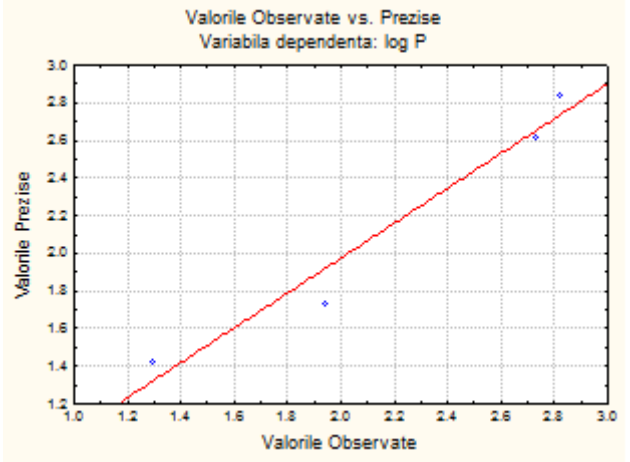
Any QSAR model must be validated with a set of external prediction. In this case the prediction set consists of the 4 derivatives of 2-furiletene that predict the n-octanol/water partition coefficient logP values (Table 9).

Table 9. Observed and predicted logP or validation set.

Number of the structures	log *P* **obs.**	log *P* **predicted eq. 39**	log *P* **predicted eq. 40**
35	2.820	2.924846	2.845186
36	2.730	2.580922	2.617008
37	1.290	1.438395	1.422570
38	1.940	1.709787	1.735940
R^2		**0.9333**	**0.9585**



a) **R^2** = 0.9333
log *P*(obs) = 0.9425*logP(calc) + 0.0948
log P (calc) – eq. 39



b) **R^2 =0.9585**
log *P*(obs) = 0.9278* logP(calc) + 0.1186
log P (calc) – eq. 40

Figure 7. Plot of the observed vs. Predicted values for eq. 39 and 40 for external validation set.

Equation explains 95.85% of the variance of log P, this value indicates increased ability of the prediction model (equation 39). In Figure 7.b are presented the relationship between observed and predicted values of log of P. In this sense, the equation obtained with descriptors MW X4v, G (N.. A) is the best property to predict log P.

Going through the same steps I realized the prediction for the school set and validation for the prediction set for n-octanol/water partition coefficient log P, but this time using descriptors calculated with TOPOCLUJ software. The results are presented below. The most significant descriptors calculated with TOPOCLUJ software are presented in Table 10 along with the property log P for the set school based on molecular similarity of the leader on external prediction set.

Table 10. Topological descriptors and partition coefficients log*P*
For prediction set.

Number of the structures	CS[LM [Electronegativity]]	CS[Sh[W4 [Charge_Adjacency]]]	IE[CfMax [Density]]	VAA1	VAD1	Pobs
35	410	22	190	7.1	36	2.49
36	410	19	170	7.1	36	2.37
37	290	11	62	6.1	27	1.56
38	340	17	120	6.6	31	1.92

Used equation for prediction is from calibrated cluster school (their shapes normalized in Matlab). This equation show good correlation coefficient bun (R^2 = 0.986337 - eqc. 41, R^2 = 0.960186 –eq. 42, R^2 = 0.911236 - ec.43) (Figure 8).

Bivariable regression

logP= 3.562116+0.015455· **IE[CfMax[Density]]** -0.109763· **VAD1** (41)
n = 12 R^2 = 0.911236 s = 0.519627 F = 46.19599

Multivariable regression

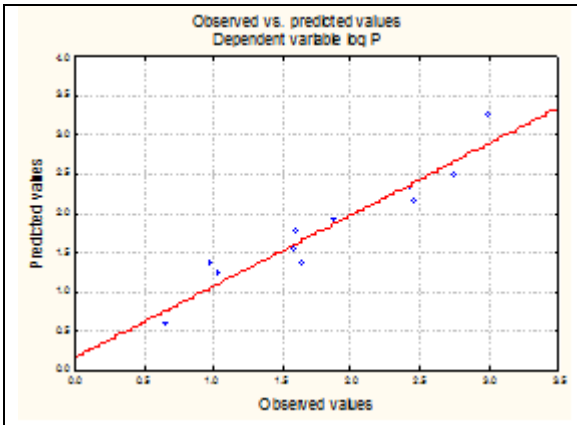
logP= 7.99004+0.05992· **CS[Sh[W4[Charge_Adjacency]]]** +0.01853· **IE[CfMax[Density]]** -1.45456· **VAA1** (42)
n = 12 R^2 = 0.960186 s = 0.233074 F = 64.31076

logP= 12.43950+0.01138· **CS[LM[Electronegativity]]**+ 0.06207· **CS[Sh[W4[Charge_Adjacency]]]**+0.01658· **IE[CfMax[Density]]** -2.70391· **VAA1** (43)
n = 12 R^2 = 0.986337 s = 0.079982 F = 126.3344

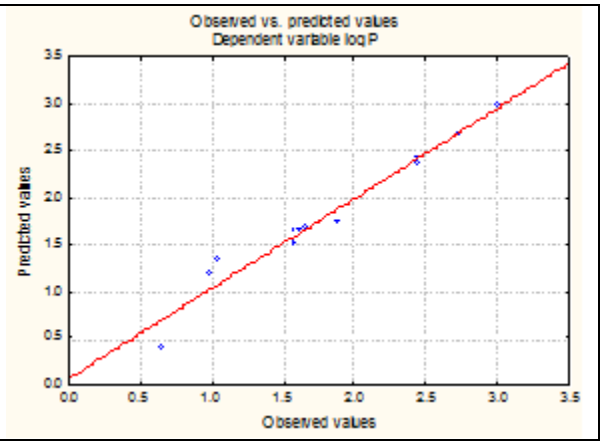
QSAR models must be validated by external prediction set. Prediction set consists of the 4 derivatives of 2-furiletene that predict the n-octanol/water partition coefficient logP values (Table 12).

Table 12. Observed and predicted log P for validation set

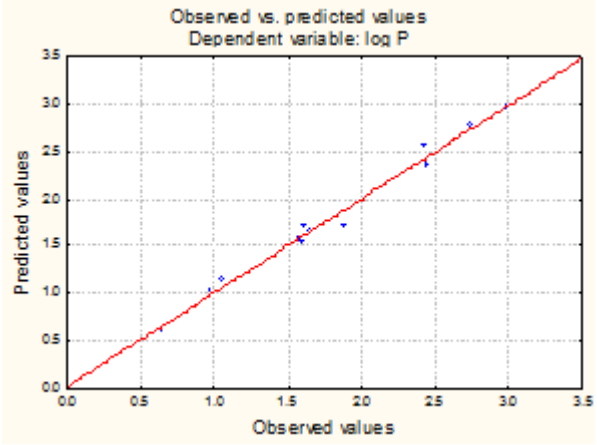
Number of the structures	log *P* obs.	log *P* predicted eq 47	log *P* predicted eq. 48	log *P* predicted eq. 49
35	2.820	2.547183	2.501652	2.424993
36	2.730	2.238073	1.951290	1.907205
37	1.290	1.556749	0.925237	0.957875
38	1.940	2.014115	1.632225	1.509150
R^2		**0.924**	**0.9037**	**0.9143**



a) R^2 = 0.911236, ecuaţia nr 41



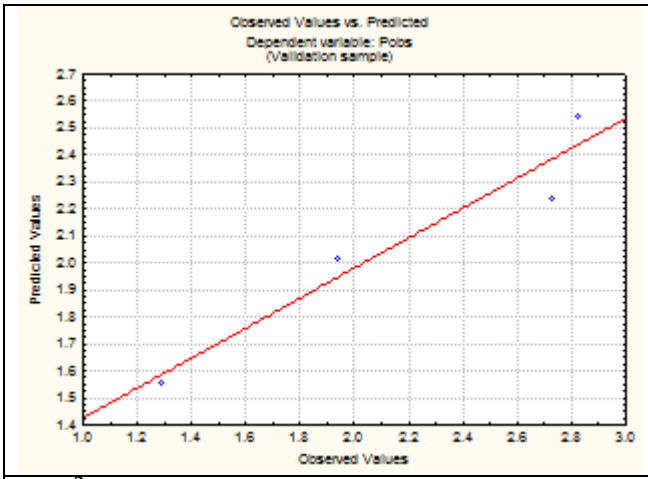
b) R^2 = 0.960186, ecuaţia nr 42



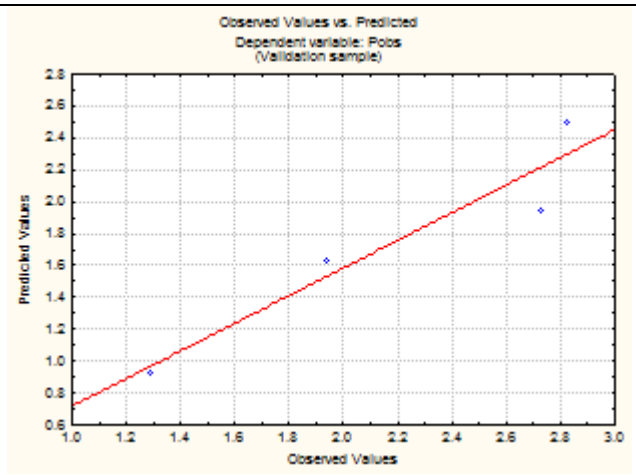
c) R^2 = 0.986337, eq. nr 43

Figure 8. Plot of the observed vs. Predicted properties for eq. 41, 42 and 43 for school set.

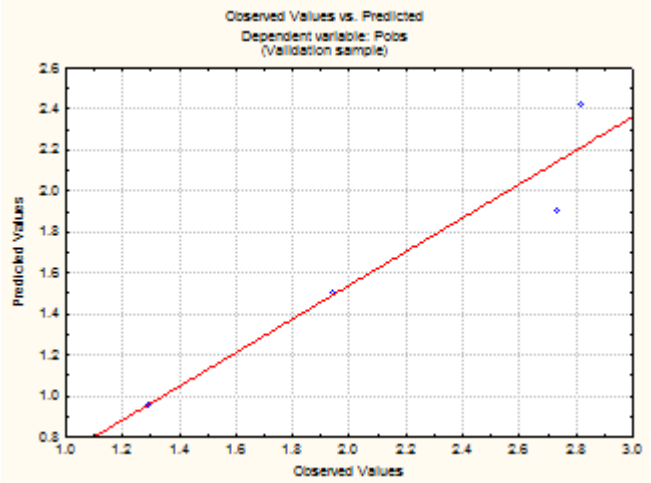
Topological descriptors involved in this equation are IE [CfMax [Density]] and VAD1. Graphical representation of this is shown in Figure 9.



a) $\mathbf{R^2 = 0.924}$
log *P(*obs) = 0.5554*logP(calc)+0.87
log P (calc) – ecuaţia nr. 44



b) $\mathbf{R^2 = 0.9037}$
log *P(*obs) = 0.8674* logP(calc) - 0.1513
log P (calc) – ecuaţia nr. 45



c) $\mathbf{R^2 = 0.9143}$
log *P(*obs) = 0.8228*logP(calc) - 0.1062
log P (calc) – ecuaţia nr. 46

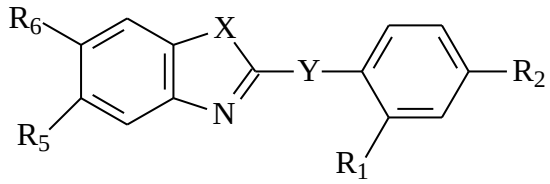
Figura 9. Plot of the observed vs. Predicted properties for validation set.

New QSAR models for predicting the biological activity of benzoxazol/benzimidazole derivatives

Fungal diseases are caused by microorganisms classified in the systematically **Fungi regnum**. They are found in large numbers into the environment. Most have adopted saprobiot living environment, but some of them have adapted to parasitic life. It cites that over 300 species have been found to be pathogenic to animals. Obligatory parasitic fungi belong, for the most part, dermatomicete category. They do not thrive in the environment can survive only in living organisms can be transmitted by direct or indirect contagion.

Obtaining molecular descriptors

Recently was reported the synthesis and activity of some benzoxazole/benzimidazole derivatives against C. albicans species. Analysis of quantitative structure activity relationships (QSAR) has been more widespread and effective design used in theoretical studies of drugs. QSAR models offer researchers in this field are focused, allowing savings of time, money and energy, improving pharmaceutical research (Figure 10, Table 13).



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	

Figura 10. Antimycotic activity benzoxazol and benzimidazole derivatives against species C.Albicans

Table 13. Observed log 1/C values benzoxazole/benzimidazole derivatives.

	R_1	**R_2**	**R_5**	**R_6**	**X**	**Y**	**log 1/C Obs.**
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

Topological descriptors used in QSAR studies are accessible and can be easily calculated using the software. Set of molecular descriptors used in this study was calculated with DRAGON software package.[45] Structures were optimized using the semi empirical PM3 Hamiltonian option, available in HyperChem software.[137]

Data processing and analysis

The purpose of this study is to develop a new QSAR model, more efficient than other previously obtained for predicting antifungal activity of bezoxazole/benzimidazole derivatives substituted in position 2 and methyl group in position 5 or 6.
Topological descriptors obtained from Dragon software and biological activity (antifungal) log 1 / C for this set of structures are presented in Table 14.

Tabelul 14. Topologici descriptors obtained with Dragon software, experimental and predicted log 1/C values for benzoxazole/benzimidazole derivatives.

Nr.	MW	nCs	nHDon	Obs. log 1/C	Calc. log 1/C
1	243.7	0	0	3.989	3.988
2	239.29	0	0	3.980	3.981
3	254.26	0	0	4.007	4.006
4	278.14	0	0	4.046	4.046
5	237.32	0	0	3.977	3.977
6	269.32	0	0	4.032	4.031
7	243.7	0	0	3.989	3.988
8	239.29	0	0	3.980	3.981
9	227.25	0	0	3.958	3.960
10	254.26	0	0	4.007	4.006
11	278.14	0	0	4.046	4.046
12	237.32	0	0	3.977	3.977
13	269.32	0	0	4.032	4.031
14	223.29	1	0	4.251	4.251
15	302.18	1	0	4.383	4.384
16	238.31	1	2	4.280	4.279
17	223.29	1	0	4.251	4.251
18	222.31	1	1	4.249	4.251
19	256.75	1	1	4.312	4.309
20	301.2	1	1	4.382	4.384
21	237.33	1	3	4.278	4.278
22	239.29	0	0	3.980	3.981
23	255.36	0	0	4.009	4.008
24	254.38	0	1	4.007	4.007
25	272.75	0	1	4.037	4.038

Semnificative molecular descriptors calculated with Dragon software are: MW- molecular weight, nCs- secundar C atoms (sp^3) and nHDon- number of the H bonds donor atoms H (N and O).
QSAR equation was found using multivariable regression analysis. Quality model is given by the square of the regression coefficient (R^2), Fischer report, the estimated standard error (s)

and leave-one-aut procedure (LOO) as cross-validation procedure. Equation has a good value of the correlation coefficient R^2 = 0.999 in multivariable regression with 3 descriptors. This correlation is significantly better than previously reported (R^2 = 0.94). Experimental and calculated values of log 1 / C for this set are plotted in Figure 11.

log 1/C= 3.577225 + 0.001686* **MW** + 0.297347 * **nCs** + 0.00122 * **nHDon** (44)
n=25 R2=0.999932 s=0.0029 F= 103092.3

Pobs Std.Dev.= 0.145464
Pobs Std.Err = 0.029093

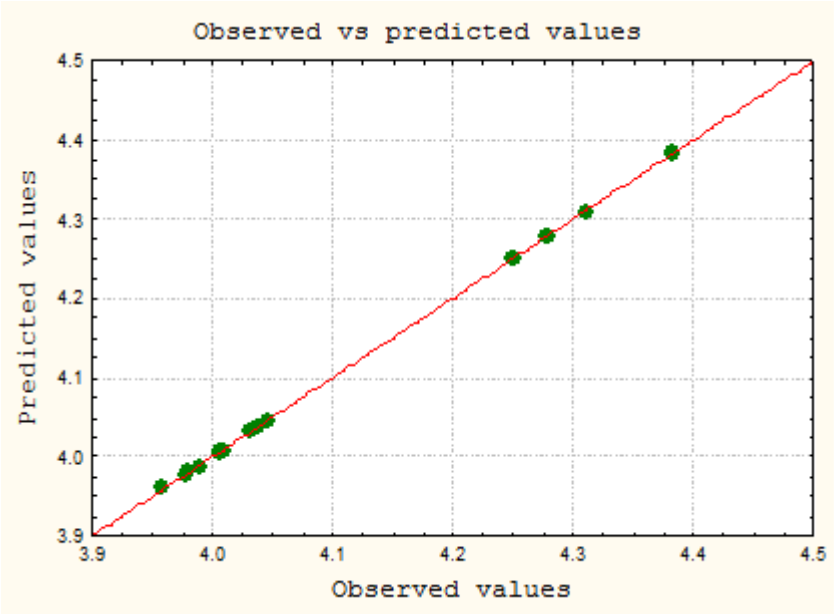


Figure 11. Plot observed vs. Predicted values for eq. 44.

Based on data proposed a new prediction model developed for direct biological activity. Were constructed clusters of similarity for each molecule in the set school and have made predictions on subsets of congeners obtained. Significant molecular descriptors are MW and NCS. QSAR equations were found using multivariate regression analysis (Table 15). The purpose of this method is to obtain the best model for predicting future antifungal activity of other new structures. For each structure, taken as a leader in test similarity was calculated equation of bivariable prediction for each subset of similarity. Each of the 25 learning equations has R^2 = 0.9999 and this figure varies only from the 5 th decimal place. Prediction equation obtained has a very good correlation coefficient, R^2 = 0.999, as a bivariable regression (Figure 12).
Correlation is also much better than previously reported (R^2 = 0.94).

Table 15. Regression equation parameters, observed and calculated values for log 1/C, ec. Y_{calc}=a + b*MW +c*nCs.

Nr.	MW	nCs	log 1/C Obs.	Equation parameters			log 1/C Calc.
				a	b	c	
1	243.70	0	3.989	3.529574	0.001885	0.300421	3.989006
2	239.29	0	3.980	3.578017	0.001684	0.298368	3.980871
3	254.26	0	4.007	3.577699	0.001684	0.298482	4.005963
4	278.14	0	4.046	3.577586	0.001685	0.298405	4.046282
5	237.32	0	3.977	3.577923	0.001684	0.298388	3.97753
6	269.32	0	4.032	3.586488	0.001650	0.299659	4.030916
7	243.70	0	3.989	3.577964	0.001683	0.299659	3.988105
8	239.29	0	3.980	3.578017	0.001684	0.298368	3.980871
9	227.25	0	3.958	3.579723	0.001677	0.298309	3.960813
10	254.26	0	4.007	3.577699	0.001684	0.298482	4.005963
11	278.14	0	4.046	3.586385	0.001651	0.299609	4.045527
12	237.32	0	3.977	3.586968	0.001649	0.299502	3.978302
13	269.32	0	4.032	3.586488	0.001650	0.299659	4.030916
14	223.29	1	4.251	3.529649	0.001885	0.300392	4.250924
15	302.18	1	4.383	3.582466	0.001666	0.299986	4.385951
16	238.31	1	4.280	3.585095	0.001656	0.299324	4.279045
17	223.29	1	4.251	3.529649	0.001885	0.30032	4.250924
18	222.31	1	4.249	3.529748	0.001884	0.300463	4.249153
19	256.75	1	4.312	3.585044	0.001656	0.298955	4.309212
20	301.20	1	4.382	3.583753	0.001661	0.299843	4.383946
21	237.33	1	4.278	3.585631	0.001654	0.299458	4.277595
22	239.29	0	3.980	3.587052	0.001649	0.299486	3.981574
23	255.36	0	4.009	3.585960	0.001652	0.299638	4.007876
24	254.38	0	4.007	3.585968	0.001652	0.299608	4.006286
25	272.75	0	4.037	3.586233	0.001651	0.299593	4.036649

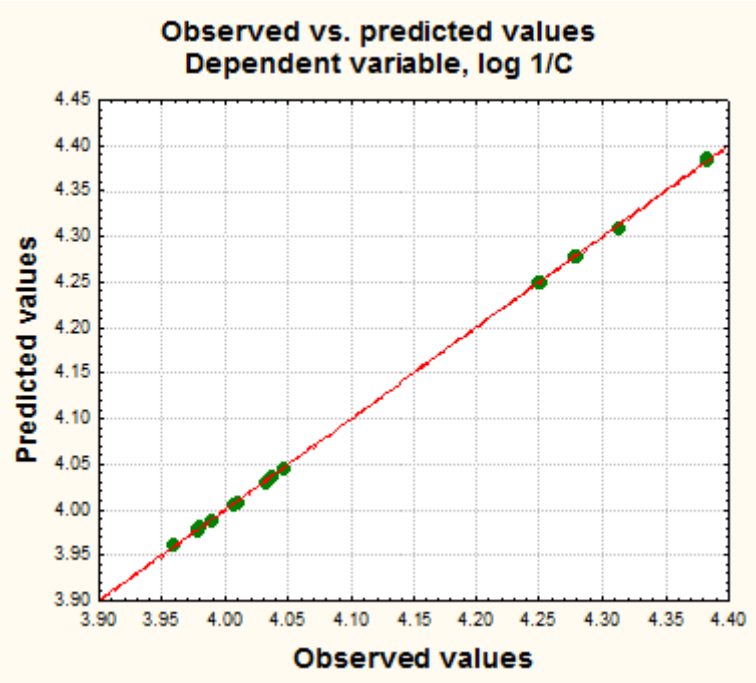


Figure 12. Plot observed vs. calculated log 1/C.

The set of benzoxazole/benzimidazole derivatives substituted in position 2 and with the methyl group in position 5 or 6, previously tested for antibacterial activity against C. albicans was analyzed by quantitative structure-activity relationship and contributions to biological activity and structural effects function were determined using multiple regression method. QSAR results obtained show that the substituent at position Y (NCS descriptor) plays an important role and makes an important contribution to the antibacterial activity.

CONCLUSIONS

Synthesis and implementation on the market of powerful drugs to meet a higher degree at solving acute society health problems is an important goal of the pharmaceutical industry in special, and for researchers in particular.

Finding theoretical methods which reduce the cost and time required finding and synthesis of biologically active compounds with practical application are the goal of any researcher. One of these methods is provided by mathematical chemistry by using different techniques and models, in which structures are assessed and quantified in numbers.

Thesis follows the treatment and obtaining various theoretical models applicable in the design of bioactive molecules, especially in terms of structural similarity. Obtaining quantitative evaluation methods similarity is just one of the problems to be solved by the "designer" of bioactive substances. Central concepts of any intermolecular similarity retrieval system are similarity based criterion and the measure used to quantify the similarity.

Thus, the problem is treated molecular similarity, following and developing various aspects of this theory. In this regard were presented in detail:

- Quantitative description of similarity
- Molecular similarity indices
- Graph-molecular descriptors
- TOPOCLUJ-SIMIL software

Were analyzed different techniques and methods overlap molecular similarity applied to classes of structures with biological activity.

These methods include:

- Cluster analysis
- Principal component analysis
- Factor analysis

Have been proposed and developed different cluster models based on similarity useful for modeling and predicting molecular properties (biological or physicochemical), with a major contribution to quantitative studies of structure - activity relationship QSAR for different biological molecules.

Proposed QSAR model results in excellent follow biological properties of classes proposed structures. Thus, the proposed model with 3 independent variables, MW, X4v and G (N. .. O), with $R^2 = 0.98653$, the advanced capability of prediction of log P in the validation set with $R^2 = 0.9585$.

Obtained results ahead similar results from the literature with a better predictive ability and at the same time a small number of descriptors used in the regression equation is useful in predicting physico-chemical and biological properties of new derivatives of 2-furilethenes.

Successful QSAR models obtained shows a statistically significant correlation between the chemical characteristics of compounds (descriptors) and biological activity. Following the analysis equations derived from modeling sets of compounds, it was concluded that it is necessary in all cases to avoid predicting the biological activity of compounds that have very different structure from the set school compounds. However, any of these procedures can be used separately as part of the multi-purpose computational studies to achieve. These analytical techniques are required increasingly more generation, interpretation and rendering more about the potential biological molecules and beyond.

It was also noted that only a subset of descriptors of molecular structures that are the most important and statistically significant, are selected to describe a biological activity chosen.

The results were published in the following journals:

1. Costescu, A., Moldovan, C.D., Diudea, M.V., QSAR modeling of steroid hormones, *Match* 55 (2), pp. 315-329, **2006**.
2. Moldovan, C.D., Costescu, A., Katona, G., Diudea, M.V., A novel QSAR approach in modeling antifungal activity of some 5-or 6-methyl-2-substituded benzoxazoles/benzimidazoles against C. albicans using molecular descriptors, *Match* 60 (3), pp. 977-984, **2008**.
3. Costescu, A., Moldovan, C.D., Katona, G., Diudea, M.V., QSAR modeling of human catechol O-methyltransferase enzyme kinetics, *Journal of Mathematical Chemistry* 45 (2), pp. 287-294, **2009**.
4. Moldovan, C.D., Costescu, A., Katona, G., Diudea, M.V., Application to QSAR studies of 2-furylethylene derivatives, *Journal of Mathematical Chemistry* 45 (2), pp. 442-451, **2009**.

References - selected

1 Rosen, R. in: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 12, 369-382.

2 Mezey, P. G. Three-Dimensional Topological Aspects of Molecular Similarity. In: Johnson, M.A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 11, 321-368.

3 Randić, M. Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization. In: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 5, 77-145.

4 Maggiora, G. M.; Johnson, M. A. Introduction to Similarity în Chemistry. In: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 1, 1-13.

5 Tsai, C. -c.; Johnson, M. A.; Nicholson, V.; Naim, M. Eds., Graph Theory and Topology in Chemistry, Elsevier, Amsterdam, **1987**, 231.

6 Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. Steric Fit în QSAR (Lecture Notes în Chemistry, Vol. 15), Springer, Berlin, **1980**, Chap. 6.

7 Kvasnička, V.; Pospichal, J. Fast Evaluation of Chemical Distance by Tabu Search Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 1109-1112.

8 Diudea, M.V. Layer Matrices în Molecular Graphs, *J. Chem. Inf. Comput. Sci.* **1994**, 34, 1064-1071.

9 Ugi, I.; Wochner, M.A.; Fontain, E.; Bauer, J.; Gruber, B.; Karl, R. Chemical Similarity, Chemical Distance, and Computer Assisted Formalized Reasoning by Analogy, in: Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 9, 239-288.

10 Basak, S.C.; Magnusson, V.R.; Niemi, G.J.; Regal, R.R., Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices, *Discr. Appl. Math.* **1988**, 19, 17-44.

11 M. Randić, W.L. Woodworth, A. Graovac, Unusual random Walks, *Int. J. Quantum Chem.*, **1983**, 24, 435-452.

12 M. Randić, Generalized molecular descriptors, *J. Math. Chem.*, **1991**, 7, 155-168.

13 N. Trinajstić, Chemical Graph Theory, CRC Press. Inc., Boca Raton Florida, **1983**.

14 A.T. Balaban, I. Moţoc, D. Bonchev, O. Mekenyan, Topological Indices for Structure - Acvtivity Correlations, *Top. Curr. Chem.*, **1993**, 114, 21-55.

15 D.H. Rouvray, The challange of characterizing branching in molecular species, *Discr. Appl. Math.*, **1988**, 19, 317-338.

16 M. Randić, Design of molecules with desired properties. A molecular similarity approach to property optimization, in "Concepts and Applications of Molecular Similarity, M.A. Johnson and G.M. Maggiora, Eds., John Wiley & Sons, Inc., **1990**.

17 R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.*, **1980**, 17, 1185–1189.

18 D. Ciubotariu, S. Mureşan, V. Gogonea, M. Medeleanu, D. Dragoş, Relaţii Cantitative Structură Chimică-Activitate Biologică (QSAR), Ed.Mirton, Timişoara, **1996**

19 E. Overton, Studien über die Narkose, Fischer, Jena, **1901**; A. Meyer, *Arch. Exptl. Pathol. Pharmakol.*, **1899**, 42, 110.

20 J.N. Langley, *J. Physiol.(*London), **1908**, 1, 339.

21 P. Ehrlich, *Ber. Dtsch. Chem. Ges.*, **1909**, 42, 17.

22 A. R. Cushny, Biological Relations of Optically Isomeric Substances, Balliere, Tindall and Cox, London, **1926**.

23 I. Moţoc, Structura moleculelor şi activitatea biologică, Ed. Facla, Timişoara, 1980.

24 E.J. Ariëns, Stereochemistry: A Source of Problems in Medicinal Chemistry, *Med.Res.Rev.*, **1986**, 6, 451-466.

25 E.J. Ariëns, Stereochemistry in the Analysis of Drug-Action, Part II. *Med. Res. Rev.*, **1987**, 7, 367-387.
26 E.J. Ariëns, Stereochemical Implications of Hybrid and Pseudohybrid Drugs, Part III. *Med. Res. Rev.*, **1988**, 8, 309-320.
27 I.D. Resa, S. Petrescu, M. Precupas, A. Câra, Probleme de statistica rezolvate pe calculator, Ed. Facla, Timisoara, **1984**.
28 D. McCormick, A. Roach, Measurement, Statistics and Computation, John Wiley & Sons, London, **1987**.
29 Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Perspective: Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1-18.
30 Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure-Activity Studies. *J. Med. Chem.* **1999**, 42, 573-583.
31 Cramer, R. D., I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959-5967.
32 Coats, E. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. Perspect. *Drug DiscoV. Design.* **1998**, 12/13/14, 199-213.
33 Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogeneous Steroids to Both Testosterone- Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, 53, 58-68.
34 Tuppurainen K, Viisas M, Laatikainen R, Perakyla M.: Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 607-613
35 Rios-Santamaria I, Garcia-Domenech R, Cortijo J, Santamaria P, Morcillo E J and Galvez J, *Internet Electronic Journal of Molecular Design*, **2002**, 1, 70.
36 Galvez J, Garcia-Domenech R, Salabert M T and Soler R, *J Chem Inf Comput Sci*, **1994**, 34, 520.
37 Todeschini, R.; Consonni, V. Handbook of molecular descriptors. Wiley-VCH: Weinheim, Germany, **2000**.
38 Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (COMSA): a novel Tool for Molecular Design. *Comput. Chem.* **2000**, 24, 615-625.
39 M. Pastor, G. Cruciani, I. McLay,S. Pickett and Sergio Clementi: Grid Independent Descriptors (GRIND): A Novel Class of Alignment- Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, 43, 3233-3243.
40 M.A. Cabrea Pérez et al.: Experimental and theoretical determination of physicochemical properties in a novel family of microcidal compounds. *European Bulletin of Drug Research*, **2001**, 9, 1.
41 Yovani Marrero Ponce et al.: Atom, Atom-Type, and Total Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Application to QSPR/QSAR Studies of Organic Compounds, *Molecules* **2004**, 9, 1100-1123.
42 Dore, J. Ch.; Viel, C. Antitumoral Chemoterapy. X. Cytotoxic and Antitumoral Activity of β-Nitrostyrenes and Nitrovinyl Derivatives. *Farmaco.* **1975**, 30, 81-109.
43 Castañedo, N.; Goizueta, R.; Perez, J.; Gonzalez, J.; Silveira, E. Cuesta, M.; Martinez, A.; Lugo,E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. Cuban Pat. 22446, 1994; Can. Pat. 2,147,594,**1999**.
44 Blondeau, J. M.; Castañedo, N.; Gonzalez, O.; Medina, R.; Silveira, E. In Vitro Evaluation of G-1: A Novel Antimicrobial Compound. Antimicrob. Agents *Chemother*. **1999**, 11, 1663-1669.

45 DragonProfessional version 5.4, software -**2006** (http://www.talete.mi.it/dragon_net.htm)