Int. J. Mol. Sci. **2003**, *4,* 512-536

International Journal of
Molecular Sciences
ISSN 1422-0067
© 2003 by MDPI
www.mdpi.org/ijms/

Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2 Permeability of Drugs

Yovani Marrero Ponce,[1, 2*], Miguel Angel Cabrera Pérez,[2] Vicente Romero Zaldivar[3], Ernest Ofori[4] and Luis A. Montero[5]

[1]Department of Pharmacy, Faculty of Chemical-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.
[2]Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.
[3]Faculty of Informatics. University of Cienfuegos, Cienfuegos, Cuba
[4]Faculty of Electrical Engineering. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.
[5]Bioinformatic Virtual Center. University of the Habana. Ciudad Habana, Cuba.
[*]Corresponding author. Fax: 53-42-281130, 281455.Tel.: 53-42-281192, 281473.
E-mail: yovanimp@qf.uclv.edu.cu, ymarrero77@yahoo.es

Received: 16 May 2003 / Accepted: 17 June 2003 / Published: 30 August 2003

Abstract: The high interest in the prediction of the intestinal absorption for New Chemical Entities (NCEs) is generated by the increasing rate in the synthesis of compounds by combinatorial chemistry and the extensive cost of the traditional evaluation methods. Quantitative Structure–Permeability Relationships (QSPerR) of the intestinal permeability across the Caco-2 cells monolayer (P_{Caco-2}) could be obtained by the application of new molecular descriptors. In this sense, quadratic indices of the "molecular pseudograph's atom adjacency matrix" and multiple linear regression analysis were used to obtain good quantitative models to determine the P_{Caco-2}. QSPerR models found are significant from a statistical point of view. The total and local quadratic indices were calculated with the *TOMO-COMD* software. A leave-*one*-out cross-validation procedure (internal validation) and the evaluation of external test set of 20 drugs (external validation) revealed that regression models had a good predictive power. A comparison with results derived from other theoretical studies shown a quite satisfactory behavior of the present method. The descriptors included in the prediction models permitted the interpretation in structural terms of the permeability process, evidencing the main role of H-bonding and size properties. The models found were used in virtual screening of drug intestinal permeability and a relationship between P_{Caco-2} calculated and percentage of human intestinal absorption for the 72 compounds was established. These results suggest that the proposed method is able to predict P_{Caco-2}**,** being a good tool for screening of P_{Caco-2} for large sets of NCEs synthesized via combinatorial chemistry approach.

Keywords: Drug Absorption; Caco-2 cell, *TOMO-COMD* Software, Quadratic Indices, QSPerR.

Introduction

The oral administration is one of the most important routes due to its convenience, low cost and high patient compliance rates. The prediction of human oral drug absorption for new drug candidates is of considerable utility in the early stage of drug discovery process [1, 2]. As a rapid way to predict the human intestinal absorption during the high throughput screening (HTS) [3], many *in vitro* cell culture models, has been investigated as potential tool for drug absorption and metabolism studies [4-5]. The most widely used *in vitro* model is a Caco-2 cell line. The permeability coefficient across Caco-2 cell monolayer (P_{caco-2}) has been used to estimate oral absorption of New Chemical Entities (NCEs) [6-9]. Artursson and Karlsson have obtained a good correlation between human oral drug absorption and permeability coefficient, determined through the Caco-2 cell monolayer, which suggest that the human absorption can be predicted by this *in vitro* model [7]. However, inter-laboratory differences of Caco-2 cell permeability have been demonstrated by several researchers [10, 11].

The wide use of Caco-2 cell screening for oral absorption is based on the biological membrane properties expressed by these cells, such as: the brush borders at their apical surface and the expression of carrier-mediated transport systems and typical small intestinal enzymes [5-7, 12]. These properties permit the use of this cell culture for understanding the mechanism of cellular permeability and identify which of the drug's properties are responsible for cellular permeation. Nevertheless, this cell line presents several disadvantages including: a) the permeability of compounds that are transported via carrier-mediated absorption is lower than obtained in the human small intestine. Besides, the hydrophilic compounds with paracellular transport have a poor permeability [10], b) Globet, endocrine and M cells are not expressed in this cell lines, c) the cancer origin of this cell line produce and overexpression of P-glycoprotein with the consequently lower permeabilities in the absorptive direction [13], d) the lack of standardization in cell culture and experimental procedures and e) the long culture periods (21-24 day culture times), being the last one the major practical shortcoming of this approximation, with consequently extensive cost.

Several molecular interactions have been proposed to explain the oral absorption for a great diversity of substrates. The lipophilicity, among these interactions, is considered as the most significant driving forces for permeability through Caco-2 monolayer cell cultures [6, 8, 14, 15] besides the role of hydrogen bonding capacity or the molecule net charge [2, 4, 8, 14]. Waterbeemd *et al.* have proposed a function, where these interactions are considered [14]:

$$\text{Permeability}= f\,(\text{lipophilicity, molecular size, H-bonding capacity, charge}) \qquad (1)$$

where, for each property there are limited ranges as have been established in the *Rule- of- 5* [1], but none of them are independent [16].

At present, is known that theoretical approach appears to be a good alternative for "*in silico*" prediction of human absorption for NCEs obtained by combinatorial chemistry methodologies [17-25]. The significant failure rate of drug candidates, in late stage of drug development, suggest the use of good predictive tools able to eliminate inappropriate compounds before substantial time and money are invested in testing [26].

Several methods have been developed in order to explain the drug-membrane interactions and among them appear computational chemistry and QSAR/QSPR techniques such as: linear regression, [10, 14, 15, 27, 28] partial least square, [14] artificial neural networks [27] and no linear relationship [6, 8, 14]. In some of these papers traditional QSPR analysis were applied to derive quantitative relationships between the $P_{Caco\text{-}2}$ and molecular structures. Some kinds of molecular descriptors have been introduced, including size and hydrogen-bonding descriptors [14], polar surface area (PSA) [10, 29, 30], Molsurf-derived descriptors [31], MO-calculation [27] and membrane-interaction analysis [28]. These QSPR models have predicted the Caco-2 cell permeability with a reasonable accuracy, although the number of compounds used in the data sets is limited.

Recently, several molecular descriptors based on the two–dimensional topological structure of molecules have been defined and tested in QSAR models [32-46]. In this sense, two of the present authors have developed a novel method called *TOMO-COMD* (acronym of *TO*pological *MO*lecular *COM*puter *D*esign) [47]. It calculates several families of topologic molecular descriptors. One of these families has been defined as quadratic indices in analogy to the quadratic mathematical forms. Several works have been conducted with the use of these topological indices and they will be published elsewhere.

The purpose of this study was to develop a quantitative model that permits the prediction of Caco-2 cell permeability from the molecular structure using a combinatorial approach of quadratic indices and multiple linear regression method; in a second place, to compare the results obtained with other methodologies in order to assess it. Furtherly, to evaluate the relationships between the structures, expressed by the quadratic indices and the permeability coefficients of the data set split in anionic, neutral and cationic compounds. In addition, to corroborate the predictive power of the models found, using an external prediction set of 20 drugs and by a cross-validation procedure (leave-*one*-out) of the original data set. Finally, a virtual screening of drug intestinal permeability was carried out.

Materials and Methods

Mathematical Definition of the Calculated Molecular Descriptors

Molecular vector space

Each element of the periodic table has inherent atomic properties, such as electronegativity, density, atomic radii and so on. Each one of these properties numerically characterizes each kind of atom taking values in the real set ($\Re$). For example, the Mulliken electronegativity (X_A) [48] of the atom A take the values $X_H = 2.2$ for Hydrogen, $X_C = 2.63$ for Carbon, $X_N = 2.33$ for Nitrogen, $X_O = 3.17$ for Oxygen, $X_{Cl} = 3.0$ for Chlorine and so on.

Let be a molecular vector whose elements are the atomic properties of the atoms in the molecule, for instance X_A. Thus, a molecule having 2, 3, 4,…, n atoms can be "represented" by means of vectors, with 2, 3, 4,...., n components, belonging to the spaces $\Re^2$, $\Re^3$, $\Re^4$,..., $\Re^n$, respectively. Where n is the dimension of these real subsets ($\Re^n$).

This approach allows us to express compounds such as: benzene, cyclohexane, hexane and all the constitutional and geometric isomers of hexane through a general kind of vector X= (X_C, X_C, X_C, X_C, X_C, X_C). On the other hand, *n*-propanol, *iso*-propanol, propanal, and acetone may be represented by (X_C, X_C, X_C, X_O) or any permutation of the components of this vector. All these vectors belong to the products space $\Re^6$ and $\Re^4$, respectively. It must be noted that the order of the vector components is meaningless here. This fact, not common in classical vector spaces, will be explained elsewhere. Besides, in this example were not considered the hydrogen atoms.

By taking into consideration all the universe of organic molecules, a molecular vector space (E) could be defined:

$$E= \Re \oplus \Re^2 \oplus \Re^3 \oplus ... \oplus \Re^n = \bigoplus_{i=1}^{n} \Re^i \tag{2}$$

where, i=1, 2, 3,…n; $\Re^k \cap \Re^l = \{0\}$: $k \neq l$ and the dimension of E is the sum of the dimensions of each one of the $\Re^i$ spaces. Therefore, this dimension is $n(n+1)/2$.

This space includes all the possible molecules having n atoms as vectors of the $\Re^n$ spaces. The present mathematical formalism makes possible to represent any drug or organic molecule into a vector space and then, to use the well-known applications of this algebraic construction to codify molecular structure in a timely but mathematically rigorous way.

Total quadratic indices; [$q_k(x)$]

If a molecule is consists of n atoms (*vector of* $\Re^n$) then the k-th quadratic indices $q_k(x)$ are defined like q application (q: $\Re^n \rightarrow \Re$) where X can be expressed by a linear combination with a base belonging to the vector sub-space $\Re^n$ ($X=x_1a_1+...+x_na_n$, where $(a_i)_{1\leq i\leq n}$ is a base of $\Re^n$). Taking into consideration the conditions mentioned above q is a quadratic form if Eq. 3 is considered.

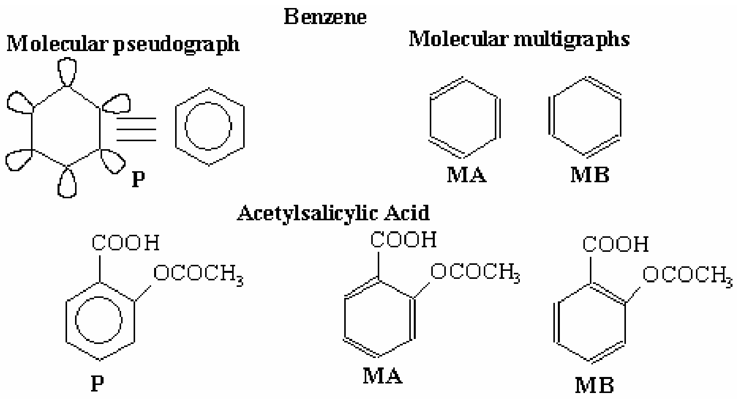
$$q_k(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}a_{ij} X_i X_j \tag{3}$$

where, $a_{ij} = a_{ji}$ and n is the number of atoms of the molecule. The coefficients ${}^{k}a_{ij}$ are the elements a_{ij} of the k-th power of the matrix M of the molecular pseudograph (G). Then, M (G) = M = $[a_{ij}]$, where n is the number of vertices and the elements a_{ij} are defined as follows:

$$\begin{aligned} a_{ij} &= P_{ij} \text{ if } i \neq j \text{ y } \exists\, e_k \in E \,/\, e_k \sim v_i, v_j \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \tag{4}$$

where, P_{ij} is the number of edges that comply with $e_k \sim v_i, v_j$ among the vertices v_i *y* v_j. L_{ii} is the number of loops in v_i.

The elements a_{ij} (if $a_{ij} = P_{ij}$) of this matrix represents the bonds between an atom "*i*" and other "*j*." The matrix M^k provides the number of paths of length k that links the vertices v_i and v_j. For this reason each edge represents 2 electrons of the covalent bond between 2 atoms v_i and v_j, and it is appreciated in the M (k=1) matrix input that v_{ij} and v_{ji} is equal to 1. In this way, the benzene molecules can be represented for two different multigraphs, where each multigraph is related with one of the Kekulé structures. Taken into consideration that mentioned above, it is necessary the use of a pseudograph to avoid this situation in compounds with more than one canonical structure. It happened for substituted aromatic compounds such as pirydine, naphthalene, quinoline, etc., where the electrons of PI(π)-orbitals are represented as loops of all ring atoms. Aromatic rings with only one canonical structure, such as furan, thiophene, pyrrol etc. are represented like a multigraph. This explanation is represented, in an easy way, in Scheme 1 and in Table 1. As can be observed, for benzene molecule the total quadratic indices (without considering hydrogen atoms) calculated using the multigraph matrices (connectivity matrices) had the same values. However, single molecules like acetylsalicylic acid show differences in the total and local (heteroatoms and H-bonding heteroatoms) quadratic indices obtained from each multigraph (MKA and MKB). The representation numbers, like a multigraph, are higher when the number of rings with more than one canonical structure is increased.



Scheme 1. Graphical representation of benzene and acetylsalicylic acid using "multigraph (MA and MB) and pseudograph (P)".

Table 1. Total [$q_0(x)$] and Local [$^{E}q_0(x)$ and $^{H}q_0(x)$] Quadratic Indices Calculated for Multigraphs (MA, MB) and Pseudographs (P).

				Benzene				
	$q_0(x)$	$q_1(x)$	$q_2(x)$	$q_3(x)$	$q_4(x)$	$q_5(x)$	$q_6(x)$	$q_7(x)$
P	41.5014	124.5042	373.5126	1120.5378	3361.6134	10084.8402	30254.5206	90763.5618
MA	41.5014	124.5042	373.5126	1120.5378	3361.6134	10084.8402	30254.5206	90763.5618
MB	41.5014	124.5042	373.5126	1120.5378	3361.6134	10084.8402	30254.5206	90763.5618
				Acetylsalicylic acid				
				Total				
	$q_0(x)$	$q_1(x)$	$q_2(x)$	$q_3(x)$	$q_4(x)$	$q_5(x)$	$q_6(x)$	$q_7(x)$
P	102.4477	268.8912	873.5982	2566.8034	8381.4114	25593.6122	83330.7872	260026.931
MA	102.4477	268.8912	873.5982	2549.8376	8284.7898	25063.374	81351.7828	250745.988
MB	102.4477	268.8912	873.5982	2566.5118	8389.425	25513.2092	83389.772	258104.308
				Local				
	$^{E}q_0(x)$	$^{E}q_1(x)$	$^{E}q_2(x)$	$^{E}q_3(x)$	$^{E}q_4(x)$	$^{E}q_5(x)$	$^{E}q_6(x)$	$^{E}q_7(x)$
P	40.1956	58.3597	265.963	510.2749	2171.4817	4947.1654	19328.9482	49869.8377
MA	40.1956	58.3597	265.963	500.226	2133.2198	4618.7534	18773.2472	44486.7656
MB	40.1956	58.3597	265.963	508.5631	2201.8503	4802.1696	19870.6695	47162.9747
	$^{H}q_0(x)$	$^{H}q_1(x)$	$^{H}q_2(x)$	$^{H}q_3(x)$	$^{H}q_4(x)$	$^{H}q_5(x)$	$^{H}q_6(x)$	$^{H}q_7(x)$
P	4.84	6.974	10.626	33.682	67.54	270.578	670.604	2600.972
MA	4.84	6.974	10.626	33.682	67.54	269.632	647.306	2589.686
MB	4.84	6.974	10.626	33.682	67.54	271.766	653.092	2639.868

On the other hand, we can obtain $q_k(x)$ by means of the matrix expression $q_k(x)$ =X^tM^kX ($k\geq 10$), being X the column vector of the coordinates in the base a_i. In this case as we work with the canonical base, the coordinates of any vector X, coincide with the components of this vector. For that reason, such coordinates can be considered as weights of the vertices of the pseudograph, because the components of the vectors are values of some atomic property that characterizes each kind atom. In Table 2 the calculation of five quadratic indices for acetylsalicylic acid is exemplified.

As can be seen in the Eq. 3 the products appear between each other, for even pairs, of the different coordinates of X, which gives it a quadratic aspect. As $^{k}a_{ij}$ = $^{k}a_{ji}$ (*the matrix is symmetric*) and x_ix_j=x_jx_i, we can rewrite the $q_k(x)$ expression in the form:

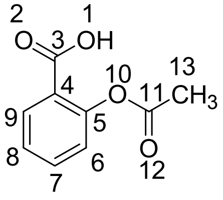
$$q_k(x)=\sum_i {}^{k}a_{ii}X_i^2+2\sum_{(i,j)} {}^{k}a_{ij}X_iX_j \qquad (5)$$

Local approach (local invariant) of the quadratic indices; [$q_{kL}(x)$]

In the case of the quadratic indices it is possible to define analogs to the total quadratic indices that possess similar properties and which are defined as local quadratic indices of the "molecular pseudograph's atom adjacency matrix". The definition of this descriptor (invariant theoretical-graph for a given fragment F$_i$ within a specific pseudograph G) is the following:

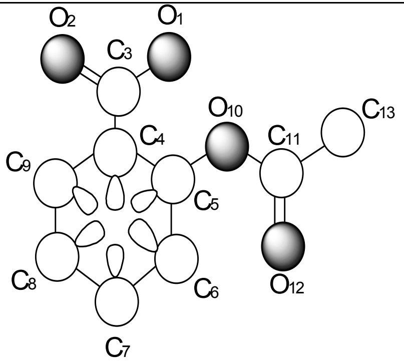
$$q_{kL}(x)=\sum_{i=1}^{m}\sum_{j=1}^{m} {}^{k}a_{ijL}X_iX_j \qquad (6)$$

Table 2. Definition and calculation of six (k=0-5) total quadratic indices of the “Molecular Pseudograph’s Atom Adjacency Matrix of the molecule of Acetylsalicylic Acid.



Acetylsalicylic Acid

Molecular structure



Molecular Pseudograph (G) (Suppressed Hydrogen Atom)

X=[O_1 O_2 C_3 C_4 C_5 C_6 C_7 C_8 C_9 O_{10} C_{11} O_{12} C_{13}]
Molecular Vector: $X \in \Re^{13}$ and $\Re^{13} \in E$;
E: Molecular Vectorial Space

In the definition of the X, as molecular vector, the chemical symbol of the element is used to indicate the corresponding electronegativity value. That is: if we write O it means $\chi(O)$, oxygen Mulliken electronegativity or some atomic property, which characterizes each atom in the molecule. So, if we use the canonic bases of R^{13}, the coordinates of any vector X coincide with the components of that molecular vector

X^t =[3.17 3.17 2.63 2.63 2.63 2.63 2.63 2.63 2.63 3.17 2.63 3.17 2.63]

X^t = transposed of X and it means the vector of the coordinates of X in the Canonical base of R^{13} (a row Matrix)
X: vector of coordinates of X in the Canonical base of R^{13} (a columns matrix)

$$q_0(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{0}a_{ij} X_i X_j = X^t M^0 X = 102.4472$$

$$q_1(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{1}a_{ij} X_i X_j = X^t M^1 X = 268.8912$$

$$q_2(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{2}a_{ij} X_i X_j = X^t M^2 X = 373.5982$$

$$q_3(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{3}a_{ij} X_i X_j = X^t M^3 X = 2566.8034$$

$$q_4(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{4}a_{ij} X_i X_j = X^t M^4 X = 8381.1414$$

$$q_5(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{5}a_{ij} X_i X_j = X^t M^5 X = 25593.612$$

$$\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{bmatrix}$$

M^1 (G) : Molecular Pseudograph`s Atom Adjacency Matrix where *m* is the number of atoms of the fragment of interest and ${}^{k}a_{ijL}$ is the element of the file "*i*" and column "*j*" of the matrix $\mathrm{M}^{k}{}_{L}=\mathrm{M}^{k}(\mathrm{G}, \mathrm{F}i)$ $[q_{kL}(x) = q_{k}(x, \mathrm{F}i)]$. This matrix is extracted from the M^{k} matrix and it contains the information referred to the vertices of the specific fragments (F*i*) and also of the molecular environment.

The matrix $\mathrm{M}^{k}{}_{L}=[{}^{k}a_{ijL}]$ and the elements ${}^{k}a_{ijL}$ is defined as follows:

${}^{k}a_{ijL}={}^{k}a_{ij}$ if both v_i and v_j are vertices contained in the specific fragment. (7)

$={}^{1}/{}_{2}\,{}^{k}a_{ij}$ if either v_i or v_j is contained in the specific fragment but not both at the same time

=0 otherwise

being ${}^{k}a_{ij}$ the elements of the *k*-th power of M. These local analogs can also be expressed in matrix form by the expression:

$$q_{kL}(x) = \mathrm{X}^{t}\,\mathrm{M}^{k}{}_{L}\,\mathrm{X}\text{: } \mathrm{M}^{k}{}_{L}\text{:it is extract of the } \mathrm{M}^{k} \qquad (8)$$

As can be seen if a molecule is partitioned in Z molecular fragments, the matrix M^{k} can be partitioned in Z local matrices $\mathrm{M}^{k}{}_{L}$, L=1,... Z. The *k*-th power of matrix M is exactly the sum of the *k*-th power of the local Z matrixes:

$$\mathrm{M}^{k}=\sum_{L=1}^{Z} M^{k}_{L} \qquad (9)$$

or in the same way as $\mathrm{M}^{k}=[{}^{k}a_{ij}]$ where:

$${}^{k}a_{ij}=\sum_{L=1}^{Z} {}^{k}a_{ijL} \qquad (10)$$

and the total quadratic indices is the sum in the quadratic indices of the Z fragments:

$$q_{k}(x)=\sum_{L=1}^{Z} {}^{e}q_{kL}(x) \qquad (11)$$

Any local quadratic index has a particular meaning, especially for the first values of *k*, where the information about the structure of the fragment F*i* is contained. High values of *k* are in relation with the environment information of the fragment F*i* considered inside the molecular pseudograph (G).

In any case, whether a complete series of indices is considered, a specific characterization of the chemical structure is obtained (whole structure or fragment), which is not repeated in any other molecule. The generalization of the matrices and descriptors to "superior analogs" is necessary for the evaluation of situations where only one descriptor is unable to bring a good structural characterization [49]. These local indices can also be used together with total indices as variables of QSAR and QSPR models for properties or activities that depend more on a region or fragment than on the whole molecule.

The TOMO-COMD Software

The calculation of total and local quadratic indices for any organic molecule was implemented in the software *TOMO-COMD* [47]. This software has a graphical interface that becomes it user friendly for medicinal chemists. The main steps to conducted for the application of this method to QSAR/QSPR can be briefly resumed as follows:

1. Draw the molecular pseudographs for each molecule of the data set, using the software drawing mode. This procedure is carried out by a selection of the active atom symbol belonging to different groups of the periodic table. The multiples edges and loops are edited with a right mouse click,
2. Use appropriated atom weights in order to differentiate the molecular atoms. In this work, we used as atomic property the electronegativity of Mulliken [48] for each kind atom,
3. Compute the total and local quadratic indices of the molecular pseudograph's atom adjacency matrix. They can be carried out in the software calculation mode, which you can select the atomic properties and the family descriptor previously to calculate the molecular indices. This software generate a table in which the rows correspond to the compounds and columns correspond to the total and local quadratic indices or any others family molecular descriptors implemented in this program,
4. Find a QSPR/QSAR equation by using statistical techniques, such as multilinear regression analysis (MRA), Neural networks, linear discrimination analysis, and so on. That is to say, we can find a quantitative relation between a property *P* and the quadratic indices having, for instance, the following appearance:

   $$P=a_0q_0(\mathrm{x}) + a_1q_1(\mathrm{x}) + a_2q_2(\mathrm{x}) +....+ a_kq_k(\mathrm{x}) + \mathrm{c} \qquad (12)$$

   where *P* is the measurement of the property, $q_k(\mathrm{x})$ [or $q_{kL}(\mathrm{x})$] is the *k*th total [or local] quadratic indices, an the *a*$_k$*'s* are the coefficients obtained by the linear regression analysis.
5. Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques,
6. Develop a structural interpretation of obtained QSAR/QSPR model using quadratic indices as molecular descriptors.

The descriptors found in the whole models obtained were the following:

(1) $q_\mathrm{k}(\mathrm{x})$ and $q_\mathrm{k}{}^\mathrm{H}(\mathrm{x})$ are the *k*-th total quadratic indices calculated using the *k*-th power of the matrices [$\mathrm{M}^\mathrm{k}(\mathrm{G})$] of the molecular pseudograph (G) considering and not considering hydrogen atoms, respectively.

(2) ${}^\mathrm{E}q_{k\mathrm{L}}(\mathrm{x})$ [or ${}^\mathrm{E}q_{k\mathrm{L}}{}^\mathrm{H}(\mathrm{x})$] and ${}^\mathrm{H}\ q_{k\mathrm{k}}(\mathrm{x})$ are the *k*-th local quadratic indices calculated using a *k*-th power of the local matrices [$\mathrm{M}^\mathrm{k}{}_\mathrm{L}(\mathrm{G}, \mathrm{F}i)$] of the molecular pseudograph (G) not considering (or considering) hydrogen atoms for heteroatoms (S,N,O) and hydrogen bonding heteroatoms (S,N,O), respectively.

Caco-2 Cell Permeation Coefficients

The 17 structurally diverse compounds used in the present study were taken from the literature [14]. The experimental values of Log $\mathrm{P}_{\mathrm{Caco\text{-}2}}$ (AP→BL) are illustrated in Table 3. The data set used for '*in* *silico*' permeability studies included compounds with a diverse molecular weight and their net charge is variable at pH 7.4 [14].

Statistical Analysis

The statistical analyses were carried out with the software STATISTICA [50]. The linear multiple regression analysis (*LMR*) was used to obtain quantitative models between structure and Caco-2 cell permeability coefficients. The quality of the models was determined examining the statistics parameter of multivariable comparison of the regression and the cross-validation (leave-*one*-out) procedures. In addition, to assess the predictive power of the model an external prediction set of 20 drugs was used [8].

Results and Discussion

Quantitative Structure Permeability Relationships

In order to test the applicability of quadratic indices on structure-permeability correlations and with the aim of predicting the Caco-2 cell permeability, 17 diverse structurally drugs were selected. Two quantitative models were developed. The values of Log $P_{\mathrm{Caco\text{-}2}}$ (AP→BL) were described by multivariate linear regression analysis using a stepwise procedure. The best models obtained together with its statistical parameters are given below:

$$Log\ P_{\mathrm{caco\text{-}2}} = -4.61426\ (\pm\ 0.486) - 0.00245(\pm\ 0.301x10^{-3})\cdot{}^{\mathrm{H}}q_{5\mathrm{L}}(\mathrm{x}) + 0.004175\ (\pm\ 1.618x10^{-3})\cdot q_0{}^{\mathrm{H}}(\mathrm{x}) \qquad (13)$$

N=17 R=0.93 R_{CV}= 0.86 s=0.43 $RMSE_{CV}$=0.52 F(2,14)= 39.968 p<0.0000

$$Log\ P_{\mathrm{caco\text{-}2}} = -3.16658\ (\pm\ 0.194) - 0.00291(\pm\ 0.238x10^{-4})\cdot{}^{\mathrm{H}}q_{5\mathrm{L}}(\mathrm{x}) \qquad (14)$$

N=16 R=0.96 R_{CV}=0.93 s=0.32 $RMSE_{CV}$=0.35 F(1,14)=149.45 p<0.0000

where, R is the multiple regression coefficient, R_{CV} is the regression coefficient for the leave-*one*-out cross-validation procedure, s the standard error of estimated, $RMSE_{CV}$ is the root of the mean square error of the cross-validation, F is the Fisher ratio at the 95% confidence level and p-value is the significance level. This regression models are significant at p-value < 0.001 using the F statistics. The p-value is the observed significance probability of obtaining a greater F value by chance alone if a model fits no better than the over-all response mean.

In the Table 3 are depicted the values of experimental and calculated permeability coefficients for data set (both models), and in Figure 1 and 2 are illustrated the linear relationships between them. In the development of the first quantitative model for description of Log $P_{\mathrm{Caco\text{-}2}}$ (AP→BL) (Eq. 13), the acetylsalicylic acid was detected as statistical outliers. Outliers detection was carried out using the following standard statistical test: residual, standardized residuals, studentized residual and Cooks' distance [51]. Once rejected the statistical outlier, the Eq. 14 was obtained with better statistical

Table 3. Experimental and calculated values for Caco-2 cell permeability coefficients by equations 13 and 14.

Compounds*	Obs.[a]	Cal.[b]	Res.[c]	CV-res[d]	Cal.[e]	Res.[c]	CV-res[d]	$q_0^H(x)$[f]	$^Hq_{5L}(x)$[g]
				Training set					
Alprenolol	-4.378	-4.795	0.417	0.458	-4.550	0.172	0.191	235.7602	475.2
Testosterone	-4.286	-3.973	-0.313	-0.384	-3.829	-0.457	-0.574	287.0389	227.458
Metoprolol	-4.569	-4.675	0.106	0.116	-4.550	-0.019	-0.021	264.4829	475.2
Salicylic acid	-4.924	-5.598	0.674	1.028	-4.868	-0.056	-0.061	107.605	584.386
Propranol	-4.378	-4.905	0.527	0.572	-4.691	0.313	0.343	237.8371	523.732
Corticosterone	-4.263	-4.185	-0.078	-0.096	-4.297	0.034	0.039	330.6505	388.212
Warfarin	-4.417	-4.437	0.020	0.023	-4.191	-0.226	-0.264	249.0567	351.758
Hydrocortisone	-4.668	-4.830	0.162	0.199	-5.112	0.444	0.475	340.6994	668.118
Dexamethasone	-4.903	-4.793	-0.110	-0.144	-5.154	0.251	0.268	358.0644	682.638
Acetylsalicilic acid*	-5.62	-4.688	-0.932	-1.409	-	-	-	141.1677	270.578
Atenolol	-6.7	-6.076	-0.624	-0.693	-6.089	-0.611	-0.678	239.4811	1003.904
Terbutaline	-6.42	-6.641	0.221	0.268	-6.535	0.115	0.136	193.9415	1156.98
Mannitol	-6.744	-6.693	-0.051	-0.064	-6.476	-0.269	-0.315	169.5548	1136.696
Sulphasalasine	-6.886	-6.936	0.050	0.073	-7.262	0.376	0.536	270.0324	1406.702
Practolol	-6.046	-6.073	0.027	0.030	-6.086	0.040	0.045	239.4811	1002.892
Olsalazine	-6.959	-6.577	-0.382	-0.457	-6.569	-0.390	-0.464	216.3878	1168.772
Felodipine	-4.644	-4.929	0.285	0.310	-4.929	0.285	0.307	280.0887	605.462
				External test set					
Compounds	Obs.[h]	Cal.[b]	Res.[c]		Cal.[e]	Res.[c]		$q_0^H(x)$[f]	$^Hq_{5L}(x)$[g]
Cumarin	-4.11	-4.149	0.039		-3.167	-0.943		111.3899	0
Theophyline	-4.35	-5.328	0.978		-4.653	0.303		128.9517	510.576
Epinephrine	-6.02	-6.699	0.679		-6.438	0.418		160.7477	1123.914
Guanoxan*	-4.71	-6.876	2.166		-6.687	1.977		168.4735	1209.406
Guanabenz*	-4.14	-7.011	2.871		-6.675	2.535		133.7708	1205.226
Lidocaine	-4.21	-5.081	0.871		-4.832	0.622		224.2233	572.088
Tiacrilast	-4.90	-4.482	-0.418		-3.894	-1.006		178.2154	249.766
Imipramine	-4.26	-3.535	-0.725		-3.167	-1.093		258.4389	0
Furosemide	-6.09	-8.424	2.334		-8.741	2.651		212.1532	1914.814
Sulpiride	-6.16	-7.328	1.168		-7.763	1.603		277.3639	1578.896
Nitrendipine	-4.77	-4.850	0.080		-4.873	0.103		287.6154	586.102
Fleroxacin	-4.81	-4.055	-0.755		-3.951	-0.859		292.165	269.5
Diltiazem	-4.31	-3.236	-1.074		-3.167	-1.143		330.0333	0
Verapamil	-4.58	-2.853	-1.727		-3.167	-1.413		421.7297	0
Mibefradil	-4.87	-4.150	-0.720		-4.828	-0.042		446.2316	570.702
Bosentan	-5.98	-5.270	-0.710		-6.029	0.049		420.3623	983.422
Proscillaridin*	-6.20	-4.662	-1.538		-5.634	-0.566		486.3382	847.66
Ceftriaxone	-6.88	-7.368	0.488		-8.030	1.150		321.3851	1670.482
Remikiren	-6.13	-6.651	0.521		-8.327	2.197		553.2348	1772.76
Squinavir	-6.26	-8.734	2.474		-9.320	3.060		254.6519	2113.892

*Detected outlier. [a]Permeability (cm/s) througth cultured Caco-2 monolayers; from Ref. [14]. [b]Calculated with the equation 13. [c]Residual, defined as Log P_{Caco-2}.(obsd)- Log P_{Caco-2} (calc). [d]Residual of the Cross-Validation. [e]Calculated with the equation 14. [f]Total quadratic indice of zero order, calculated considering atom hydrogen in the pseudograph. [g]Local quadratic indice of five order, calculated considering atom hydrogen in the pseudograph. [h]Permeability (cm/s) througth cultured Caco-2 monolayers; from Ref. [8].

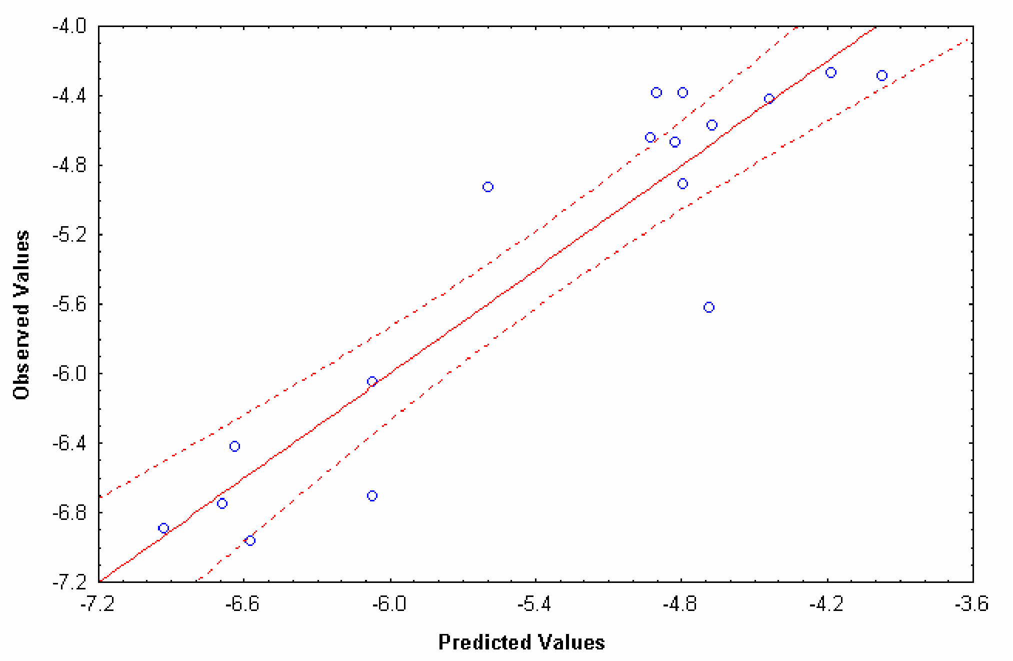


Figure 1. Correlation between experimental and calculated (by Eq. 13) permeability coefficients Log P_{Caco-2} of 17 compounds of the data set.

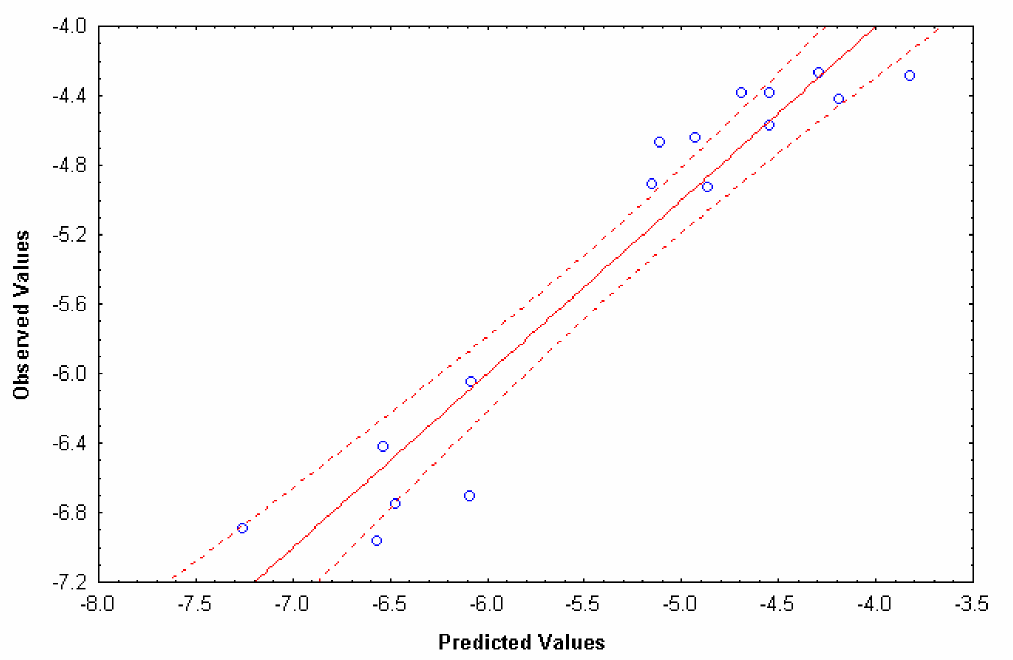


Figure 2. Correlation between experimental and calculated (by Eq. 14) permeability coefficients Log P_{Caco-2} of 16 compounds of the data set.

parameters. The outlier in linear model (Eq. 13; acetylsalicylic acid) has been detected as outliers in other work. In this sense, Waterbeemd *et al.* [14] developed relationship between permeability (in Caco-2) and several physicochemical properties such as lipophilicity, H-bonding capability, etc. Among the compounds used in this study, the acetylsalicylic acid was detected as outlier with diverges from the curve obtained for two properties above mentioned. The quadratic indices included in the Eq. 13 have structural information of molecular features in relationship with lipophilicity and hydrogen-bonding property. This can be explained the outlier behavior of the acetylsalicylic acid. Besides, outlier from linear relationship have been explained in terms of active transport, molecular size, diffusion limitation though aqueous stagnant layers at the membrane, or solubility of the drug that produced a sigmoid relationship [14].

The correlation coefficient (R^2), for equations 13 and 14 were 0.86 and 0.92 respectively, so these models explained the 86% and 92% of the variance for the experimental values of Log Caco-2 permeability [52, 53].

Validation is a crucial aspect of any QSAR/QSPR modeling [54]. One of the most popular validation criteria is leave-*one*-out cross-validated R^2 (LOO q^2; internal validation). For this reason, in order to assess the predictability of the model found, a LOO q^2 was carried out. This methodology systematically removed one data point at a time from the data set. A QSPerR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data point. This procedure was repeated until a complete set of predicted was obtained. Using this approach, the model 13 and 14 had a LOO q^2 of 0.73 and a 0.88, respectively. These values of q^2 (q^2>0.5) can be considered as a proof of the high predictive ability of the models. However, this assumption is generally incorrect and can be that exist the lack of the correlation between the high LOO q^2 and the high predictive ability of QSAR/QSPR models has been established and corroborated recently [54]. Thus, the high value of LOO q^2 appears to be the necessary but not the sufficient condition for the models to have a high predictive power. In this sense, Golbraikh and Tropsha [54] emphasize that the predictive ability of a QSAR/QSPR model can only be estimated using an external test set (external validation) of compounds that was not used for building the model and formulated a set of criteria for evaluation of predictive ability of QSAR/QSPR model. For this reason and as a second corroboration of the predictive power of the model, an external prediction set of 20 drugs was used. These compounds were also experimentally studied by Camenisch *et al.* [8]. The Caco-2 experiment was designed based on the work of Artursson's group [7], and with the objective of combine your data with previously measured compounds in order to obtain a large data set. The comparison of permeability in Caco-2 monolayers in these two different laboratories using the same experimental conditions, only Mannitol, shown inter-laboratory variations. As in the original paper, the compounds used by Waterbeemd *et al.* were taken from the same literature [7], we has selected the Caco-2 cell permeability coefficient data set development by Camenisch *et al.* [8] as one way to validate the predictive power of our models. The permeability coefficients of the drugs included in the external test set were predicted with the same accuracy that compound in the data set, if taken into consideration that these compounds were study in a different laboratory. Considering the *full* set (data and test sets) the correlation coefficients were 0.80 (s=0.64) and 0.82 (s=0.56) for Eq. 13 and 14, respectively. As can be seen, in both series, the predictability and robustness of the theoretical model was demonstrated. From the *full* data set only 3 compounds were outliers for both equations.

Waterbeemd *et al.* obtained, for this data set, a correlation coefficient (R^2) of 0.77 (s = 0.52) where two components principal, as variables in linear regression models, were used [14]. In this study, a principal component analysis, to visualize the relationships between the 26 descriptors, was developed. The three principal components used in the regression analysis explained 86.9% of the variance. The first component (43.4%) contains information about the H-bonding potential and the second one (34.2%) encodes for molecular size. This correlation coincides with the current paradigm of structure-permeability expressed in Eq. 1. Besides, these authors using a representative set of molecular weight and various H-bonding descriptors obtained 12 models applying MLR and one equation using a PLS analysis. The QSPR models developed for 17 compounds had a correlation coefficient less than 0.89. In our approach, if the statistical parameters are considered, the obtained model appears to be better than previously reported [14].

Other researchers have explored QSPerR involving Caco-2 cell permeability. For example, correlation coefficients of 0.74 and 0.76 have been obtained, considering quadratic and interactive terms [27]. In the previous paper, these authors obtained a regression coefficient of 0.79 using neural network. In other published work, Ren and Lien [15] developed a QSAR analysis for a data set of 51 compounds, where an adequate regression coefficient value, was obtained (0.79). Finally, a recently study about prediction of Caco-2 cell permeation coefficients was carried out by Kulkarni *et al.* [28] where 6 predictive models were obtained using Multidimensional Linear Regression (MLR) and the R values were between 0.86 and 0.92, but in this case only 74% from the original data set [6] was selected.

In order to understand the individual contribution of several properties and thus their effect on permeation, the compounds were considered separately according their net charge. As can be seen in equation 1 the charge of molecules has a special effect on the drug permeability, which is in relation with the negative charge of the biological membrane [55]. In the obtained models (Eq. 13 and 14) although there is not a specific variable for heteroatoms, the charge effect on lipophilicity of the compounds is already taken into account by the use of included descriptors (${}^{H}q_{5L}(x)$ and $q_0{}^{H}(x)$). However, when the whole data set was correlationed with the quadratic indices calculated (${}^{E}q_{0L}(x)$), over heteroatoms (O, N, S), the correlation coefficient was 0.546 (data not shown), which indicate the influence of this indices to describe the charge effect over the permeability.

When 17 compounds were divided into three subsets, namely anionic, neutral and cationic compounds, the following equations were obtained:

$$Log\ P_{\text{caco-2 (Anionic)}} = -3.21294\ (\pm\ 0.722) - 0.049911(\pm\ 0.013)\cdot{}^{E}q_{0L}(x) \qquad (15)$$

$$N=5 \quad R=0.91 \quad F(1, 3)= 14.079 \quad s=0.55 \quad p<0.0000$$

$$Log\ P_{\text{caco-2 (Neutral)}} =-3.32888\ (\pm\ 0.289)-0.00773\ (\pm\ 1.257x10^{-3})^{\cdot H}q_{4L}(x) \tag{16}$$

$$N=6 \quad R=0.95 \quad F(1, 4)= 37.784 \quad s=0.32 \quad p<0.0000$$

$$Log\ P_{\text{caco-2 (Cationic)}} =-2.10671(\pm\ 0.199)-0.03143\ (\pm\ 1.812x10^{-3})^{\cdot H}q_{3L}(x) \tag{17}$$

$$N=6 \quad R=0.99 \quad F(1, 4)= 300.81 \quad s=0.14 \quad p<0.0000$$

Observed and calculated values of log P_{Caco-2} as well as the residuals and cross-validation residuals for permeability in Caco-2 cell are given in Table 4. In the same Table, are depicted the values of local quadratic indices included as variables in the models.

These models explained more than 82, 90 and 98% (R^2=0.824, 0.904 and 0.986) of the variance in the experimental values of permeability coefficient for anionic, neutral and cationic compounds, respectively.

Table 4. Experimental and calculated values for the Log Caco-2 cell permeability coefficients of anionic, neutral and cationic compounds by equation 15, 16 and 17 respectively.

Compounds	Obs.	Cal.	Res.	CV-res	$^{E}q_{0L}(x)$	$^{H}q_{3L}(x)$	$^{H}q_{4L}(x)$
				Anionic compounds (-)			
Salicylic acid	-4.924	-4.693	-0.231	-0.431	30.1467	64.988	158.466
Warfarin	-4.417	-5.187	0.770	1.064	40.1956	31.306	90.684
Acetylsalicylic acid	-5.62	-5.187	-0.433	-0.599	40.1956	33.682	67.54
Sulphasalazine	-6.886	-7.032	0.146	0.343	77.7682	130.746	332.046
Olsalazine	-6.959	-6.707	-0.252	-0.425	71.1512	129.976	316.932
				Neutral compounds (0)			
Testosterone	-4.286	-3.881	-0.405	-0.727	20.0978	30.36	71.434
Corticosterone	-4.263	-4.325	0.062	0.084	40.1956	59.774	128.832
Hydrocortisone	-4.668	-5.030	0.362	0.436	50.2445	91.08	220
Dexamethasone	-4.903	-5.066	0.163	0.197	65.5326	91.08	224.708
Mannitol	-6.7445	-6.466	-0.278	-1.281	60.2934	180.268	405.768
Felodipine	-4.644	-4.739	0.095	0.116	63.6245	50.094	182.424
				Cationic compounds (+)			
Alprenol	-4.378	-4.394	0.016	0.024	25.5267	72.776	190.036
Metoprol	-4.569	-4.394	-0.175	-0.267	35.5756	72.776	190.036
Propranol	-4.378	-4.546	0.168	0.239	25.5267	77.616	200.662
Atenolol	-6.7	-6.601	-0.099	-0.167	41.0045	143	371.624
Terbutaline	-6.42	-6.514	0.094	0.149	35.5756	140.228	381.326
Practolol	-6.046	-6.043	-0.003	-0.004	41.0045	125.246	349.602

Interpretation of QSPerR Models

For a better statistical interpretation of the models built, where inter-related indices are considered (such as topological indices or topologic and topographic indices based on the same graph-theoretical invariant), the inclusion in the model of strongly interrelated variables should be avoided. It is necessary to consider the above-mentioned criterion because of the interrelation among different descriptors produce a highly unstable regression coefficients and makes difficult to know the real contribution of each variable included in the model [52]. For this reason, an interrelation study between the quadratic indices used in the equation 13 was carried out. In the Table 5, the correlation matrix for this equation shows that there is not collinearity among these variables. In the same Table other useful parameters to detect the existence of multicollinear variables (partial correlation and tolerance) are depicted. In this sense, the tolerance represents the unexplained variability for the other variables and the partial correlation coefficient explain the correlation between the property and a specific variable when the linear effects of other independent variables have been eliminated. From the Eq. 14 to 17 the tolerance value is 1 and the partial correlation coincides with the correlation coefficient.

At present, it is known the absorption is influenced by a different kind of interactions. In the equation 1 the permeability is represented like function of several interaction properties. However, Waterbeemd *et al*. expressed that charge is included in lipophilicity when distribution coefficient (Log D) instead of partition coefficient (Log P) is used; also the molecular size and H-bonding are components of lipophilicity. Thus, these authors also wrote this relationship as: [14]

$$\text{Permeability} = f\,(\text{molecular size, H-bonding capacity}) \qquad (18)$$

As can be observed in the regression models, the included variables are related with the factors that influence on the permeability values and these one with the structural features of molecules. For example, in the equations 13, the variables ${}^{H}q_{5L}(x)$ and $q_0^{H}(x)$ are in relation with the hydrogen atoms as donors and with the molecular weight (size of molecules), respectively. Both variables are identified with the paradigm of structure-permeability relationship (Eq.1 and 18). Besides, this result coincides with the information contained in the two principal components using by Waterbeemd *et al*. (Eq.4, ref.14). The coefficient of the “protonic” variable in the equation 13 is negative, which is a logical result due to when the number of the hydrogen atom bonding to heteroatoms in the molecules is increased then the permeability across the biological membrane decrease. This effect is also related with the decreasing of molecules lipophilicity and the possibility of ionization and to obtain a charge.

Table 5. The squared correlation matrix and several parameters of the quadratic indices used in the regression analysis (Eq. 13) for 17 compounds.

	Tolerance	Corr. Partial
${}^{He}q_{5L}(x)$	0.981901	-0.90851
${}^{e}q_0^{H}(x)$	0.981901	0.56783
	Correlation Matrix	
	${}^{He}q_{5L}(x)$	${}^{e}q_0^{H}(x)$
${}^{He}q_{5L}(x)$	1	0.134534
${}^{e}q_0^{H}(x)$	0.134534	1

On the other hand, the $q_0^H(x)$, with a positive contribution, is related with the possible effect of this variable on lipophilicity of compounds. That is to say, transcellular lipid permeation depends both on molecular size via lipophilicity and the diffusion coefficient through the membrane, while paracellular pore permeation depends on molecular size via the sieving effect and on diffusion in water. In the equation 14, only was included as variable the H-bonding capacity and in this case the model shown a better description that obtained in the Eq. 13. This result coincides with described by Waterbeemd *et al.* [16], due to the variable ${}^{H}q_{5L}(x)$ take into consideration not only the hydrogen atoms bond to heteroatoms but also the molecular environment, being this variable the better choice to describe the physicochemical space defined by the combination of molecular weight and H-bonding capacity.

The results obtained in the equations 15, 16 and 17 evidenced the role of total H-bonding capacity (in Eq. 15 as acceptor and in Eq. 16, 17 as a donor of hydrogen atoms). The negative contribution of the included variables may result in less membrane permeability. However, oral absorption will nevertheless be limited because of the high H-bonding capacity.

Virtual Screening and relationship of human intestinal absorption and Caco-2 cell permeability

The virtual screening has emerged as an interesting alternative to the screening of large database in order to find a set of potential new drug candidates [56-58]. In the present study we simulated a virtual search of P_{Caco-2} values by using the regression equations (Eq. 13 and 14) obtained. In Table 6, the Caco-2 cell permeability data for 72 structurally diverse compounds, obtained from different source (2, 7, 24-29) and the evaluation results of these compounds, are summarized.

As can be seen in the Table 6, existing significant variability in P_{Caco-2} experimental values obtained from two or more source. The differences in the permeability coefficients reported from various laboratories might be due to variations in cell culture conditions such as passage number, type of medium, day in culture, as well as the experimental conditions used for their measurement. Taken into consideration the inter-laboratories variability, most of 72 compounds evaluated are predicted adequately using the models obtained. It is obvious that from these results the quality of the predictions corroborates the predictive power of the models found and justified their use in the prediction of this important property. Besides, the '*in silico*' estimated intestinal permeability could be used as a predictor of the true fraction of the drug absorbed (Fa) using the theoretical relationship described by Amidon et al. [59]:

$$Fa=(1-e^{-Apeff\,x10-6})100[\%] \tag{19}$$

In these sense, the literature analysis demonstrates that the range selection for permeability coefficient in Caco-2 cells is a bottleneck whether a correlation with the human absorption is searched. Several classifications methods have been described in the literature [6, 60-62], where the inter-laboratory and experimental variability is considered. Nevertheless, if all the approaches reported in the literature are analyzed, we can state that a value of permeability coefficient greater than $10x10^{-6}$ cm/s will classify well-absorption compounds (70-100%). A second group ($P_{Caco-2}<10x10^{-6}$ cm/s),

Table 6. Caco–2 cell permeability coefficients calculated using Eq. 13 or 14; percent absorption in human and observed Caco–2 cells permeability coefficients from the different reports.

Compounds	Cal.	Obs.[c]	% Absorbed	Ref.	$^{H}q_{5L}(x)$	$q_0{}^{H}(x)$
Acebutolol	0.53[a]	0.51	90	6	1067.836	311.0776
Acetylsalicylic acid	20.50[b]	30.67	68	62	270.578	141.1677
		9.09	100	6		
		2.40	100	7		
Alprenolol	28.19[a]	40.50	93	7	475.2	235.7602
		25.30	93	6		
Aminopyrine	163.99[b]	36.50	100	6	0	198.5353
Atenolol	0.84[b]	4.00	50	60	1003.904	239.4811
		1.16	40-70, 50	63		
		0.53	50	6		
		0.20	50	7		
		0.13	40	64		
Penicilin G	5.40[b]	1.96	30	62	700.128	254.6519
Caffeine	98.53[b]	84.29	100	63	0	145.5486
		50.50	100	62		
		30.80	100	6		
		21.40	100	60		
Chloramphenicol	2.49[a]	20.60	90	62	837.386	213.2682
Cimetidine	0.12[b]	3.06	62	62	1251.47	184.9905
		1.37	95	6		
Clonidine	3.13[a]	30.10	95	62	803.264	140.0988
		21.80	100	6		
Corticosterone	50.50[a]	21.20	100	6	388.212	330.6505
Desipramine	48.65[b]	24.40	95	6	288.97	241.842
		21.60	100	62		
Dexamethasone	16.11[b]	23.40	92	62	682.638	358.0644
		12.50	100	7		
		12.20	100	6		
Diazepam	172.00[b]	70.97	100	62	0	203.4971
		33.40	100	6		
Felodipine	11.77[b]	22.70	100	7	605.462	280.0887
Fluconazole	46.07[b]	29.80	100	62	263.45	221.1982
Ganciclovir	0.07[b]	2.67	8	61	1361.932	192.5122
		0.38	3	6		
Hydrocortisone	14.80[b]	44.67	95	65	668.118	340.6994
		35.40	80	61		
		21.50	89	7		
		14.00	89	6		
		12.19	80, 89, 95	63		
Ibuprofen	39.41[b]	52.50	100	62	250.14	197.1375
Imipramine	291.70[b]	14.10	100	62	0	258.4389
Indomethacin	80.96[b]	20.40	100	6	235.62	263.4856
Labetalol	0.08[b]	9.31	90	6	1494.878	288.5856

Table 6. *Cont.*

Compounds	Cal.	Obs.[c]	% Absorbed	Ref.	$^{H}q_{5L}(x)$	$q_0^H(x)$
Mannitol	0.33[a]	3.23	17	61	1136.696	169.5548
		1.17	5, 16, 17	63		
		0.83	5	65		
		0.65	16	62		
		0.50	16	60		
		0.38	16	6		
		0.18	16	7		
Meloxicam	2.34[a]	19.50	90	6	846.714	227.8551
Metoprolol	21.13[b]	27.00	95	7	475.2	264.4829
		23.70	95	6		
		18.00	95	63		
Nadolol	2.37[b]	4.50	35	60	904.75	289.0518
Noloxone	13.21[b]	28.20	91	62	582.604	278.6856
Naproxen	38.51[b]	74.17	100	61	250.14	194.7433
Nevirapine	12.27[a]	30.10	90	6	599.236	203.278
Nicotine	100.67[b]	19.40	100	6	0	147.7868
Phenytoin	0.61[a]	89.83	100	65	1046.672	192.7891
		26.70	90	6		
Pindolol	0.46[b]	16.70	95	6	1085.788	224.5922
Piroxicam	1.44[b]	35.60	100	6	890.208	228.9639
Practolol	0.84[b]	0.90	100	7	1002.892	239.4811
Progesterone	481.44[b]	78.93	100	61	0	310.5527
Propranolol	20.36[a]	41.90	90	7	523.732	237.8371
		34.43	90	63		
		27.50	90	62		
		21.80	90	6		
		14.80	90	60		
Salicylic acid	13.56[a]	41.90	100	60	584.386	107.605
		22.00	100	6		
		11.90	100	7		
Sucrose	0.07[b]	0.71	42	63	1557.072	300.0207
Sumatriptan	0.24[b]	3.00	55	62	1225.466	240.6692
Telmisartan	112.00[a]	15.10	90	6	269.39	415.2711
Tenidap	17.85[a]	51.20	90	62	543.4	196.2092
Terbutaline	0.29[a]	1.04	25-80, 73	63	1156.98	193.9415
		0.47	73	6		
		0.38	73	7		
Testosterone	106.32[b]	72.27	100	62	227.458	287.0389
		51.80	100	7		
		44.50	100	63		
		24.90	100	6		
Timolol	14.01[b]	12.80	72	6	546.766	263.7501
Valproic acid	25.89[b]	48.00	100	62	249.194	152.873
Warfarin	36.58[b]	38.30	98	7	351.758	249.0567
		21.10	98	6		

Table 6. *Cont.*

Compounds	Cal.	Obs.[c]	% Absorbed	Ref.	$^{H}q_{5L}(x)$	$q_0^{H}(x)$
Ziprasidone	15.53[a]	12.30	60	62	564.168	293.4675
Cephalexin	0.23[b]	2.69	100	63	1261.81	255.2408
		0.18	95	64		
		0.50	100	60		
L-Phenylalanine	6.23[a]	29.50	100	65	700.348	141.0188
		6.91	100	63		
Antipyrine	107.98[b]	49.01	97	63	0	155.0726
Guanabenz	0.21[a]	20.90	79	60	1205.226	133.7708
Glycine	30.93[a]	80.00	100	62	461.362	63.5605
D-Phe-L-Pro	5.62[a]	44.30	100	62	715.704	224.9611
Gabapentin	5.17[b]	4.33	74	65	563.728	170.0588
		1.50	36	65		
BVaraU	0.43[a]	4.00	82	60	1099.494	217.7747
Pravastatin	2.19[a]	2.30	34	60	856.548	398.831
Amoxicillin	0.06[b]	0.80	100	60	1534.962	274.9697
		0.33	100	63		
SQ-29852	2.84[a]	0.02	60	60	817.476	374.5622
Trovafloxacin	5.40[b]	30.23	88	62	783.772	303.8246
Scopolamine	94.77[b]	11.80	100	6	181.786	248.2549
Ziduvudine	0.00[b]	6.93	100	6	1937.496	204.2691
Taurocholic acid	0.43[b]	4.02	100	63	1499.63	459.8587
Acyclovir	0.25[a]	2.00	30	62	1179.332	165.8664
		0.25	20	6		
Methotrexate	0.00[b]	1.20	20	62	2153.426	338.4937
Glutamine	0.19[a]	0.85	60-90	63	1221.484	123.989
Enaprilate	9.42[a]	0.62	10	63	638.748	330.1203
Hidrochlorothiazide	0.00[b]	0.51	90	6	2336.84	164.2368
Ranitidine	2.65[b]	0.49	50	6	824.978	254.0701
Sulphasalazine	0.12[b]	0.30	13	6	1406.702	270.0324
		0.13	13	7		
Doxorubicin	0.08[b]	0.16	5	62	1761.694	433.4031
Olsalazine	0.27[b]	0.11	2	7	1168.772	216.3878
Lisinopril	0.14[a]	0.05	25	60	1271.886	356.9861

[a] $P_{caco\text{-}2}$ *x* 10^{-6} (cm/s) calculated from Eq. 13; [b] $P_{caco\text{-}2}$ *x* 10^{-6} (cm/s) calculated from Eq. 14; [c] $P_{caco\text{-}2}$ *x* 10^{-6} (cm/s) from several references (see Ref. columns).

with moderate-poor absorption can be considered, although in this range a high variability is appreciated when the human oral absorption is predicted from the $P_{Caco\text{-}2}$ values. From a practical perspective the established boundary assure that classified compounds have a good absorption profile. However, the general form, when the absorbed dose fraction, from human studies, is compared with the predictive P(AP→BL) Caco-2 cell, a good relation between the theoretical and observed values is obtained. The following Table 6 demonstrates this relation for the compounds used in the study.

Conclusions

The total and local quadratic indices appear to be a very promising structural invariant. Using these molecular indices and multiple regressions, two QSPerR models were obtained for the description and determination of AP→BL transportation across monolayer of intestinal epithelial (Caco-2) cell. The results derived from the comparison with other theoretical studies shown a quite satisfactory behavior of the proposed method. The statistical quality of the models was demonstrated by evaluation of the statistical parameter of regression and those obtained by the cross-validation procedure. Besides, a test set of 20 drugs also assessed the predictive power of these models. Furthermore, this approximation permits us to obtain significant interpretation of the experimental results in terms of the structural features of molecules. This molecular descriptor is suitable for screening and *a priori* determination, during the early drug discovery, of cellular permeability coefficient for large sets of new chemical entities synthesized via combinatorial chemistry approach.

References

1. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3-25.
2. Anonymous. Waiver of in vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System. **2000**. Available from http://www.fda.gov/cder/OPS/BCS_guidance.htm.
3. Rodríguez, A. D.; Rational High Throughput Screening in Preclinical Drug Metabolism. *Med Chem Res*. **1998**, *8*, 422-433.
4. Artusson, P. Cell Cultures as Models for Drug Absorption Across the Intestinal Mucosa. *Cri. Rev. Ther. Carrier Syst*. **1991**, *8*, 305-330.
5. Quaroni, A.; Hochman, J. Development of Intestinal Cell Culture Models for Drug Transport and Metabolism Studies. *Adv. Drug Del. Rev*. **1996**, *22*, 3-52.
6. Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating Partitioning and Caco-2 Cell Permeability of Structurally Diverse Small Molecular Weight Compounds. *Pharm. Res*. **1998**, *15*, 1490-1494.
7. Artursson, P.; Karlsson, J. Correlation between Oral Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cells. *Biochem. Biophys Res. Com*. **1991**, *175*, 880-885.
8. Camenisch, G.; Alsenz, J.; Van de Waterbeemd, H.; Folkers, G. Estimation of Permeability by Passive Diffusion Through Caco-2 Cell Monolayer Using the Drugs′ Lipophilicity and Molecular Weight. *Eur. J. Pharm. Sci*. **1998**, *6*, 313-319.
9. Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating Partitioning and Caco-2 Cell Permeability of Structurally Diverse Small Molecular Weight Compounds. *Pharm. Res*. **1998**, *15*, 1490-1494.

10. Artursson, P.; Palm, K.; Luthman, K. Caco-2 Monolayer in Experimental and Theoretical Predictions of Drug Transport. *Adv. Drug Del. Rev*. **1996**, *22*, 67-84.
11. Delie, F.; Rubas, W. A. Human Colonic Cell Line Sharing Similarities with Enterocytes as a Model to Examine Oral Absorption. *Crit. Rev. Ther. Drug Carrier Syst*. **1997**, *14*, 221-286.
12. Hidalgo, I. J.; Raub, T. J.; Borchardt, R. T. Characterization of the Human Colon Carcinoma Cell Line (Caco-2) as a Model System for Intestinal Permeability. *Gastroenterology*. **1989,** *96*, 736-749.
13. Anderle, P.; Niederer, E.; Rubas, W.; Hilgendorf, C.; Spahn-Langguth, P. P-glicoprotein (P-gp) mediated efluxx in caco-2 cell monolayers: The Influence of Culturing Condition and Drug Exposure on P-gp Expression Levels. *J. Pharm. Sci*. **1998,** *87*, 757-762.
14. Van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Raevsky, Estimation of Caco-2 Cell Permeability Using Calculated Molecular Descriptors. *Quant. Struct-Act. Relat*. **1996**, *15*, 480-490.
15. Ren, S.; Lien, E. J. Caco-2 Cell Permeability *vs* Human Gastro-Intestinal Absorption: QSPR Analysis. *Prog. Drug Res*. **2000**, *54*, 3-23.
16. Van der Waterbeemd, H.; Smith, D. A.; Jones, B. C. Lipophilicity in PK Desing: Methyl, Ethyl, Futile. *J. Comp-Aided Mol. Des.* **2001**, *15*, 273-286.
17. Dressman, J. B.; Amidon, G. L.; Fleisher, D. Absorption Potential: Estimating the Fraction Absorbed for Orally Administered Compounds. *J. Pharm. Sci.* **1985**, *74,* 588-589.
18. Hamilton, H. W.; Steinbaugh, B. A.; Stewart, B. H.; Chan, O. H.; Schmid, H. L.; Schroeder, R.; Ryan, M. J.; Keiser, J.; Taylor, M. D.; Blankley, C. J.; Kalttenbronn, J. S.; Wright, J.; Hicks, J. Evaluation of Physicochemical Parameters Important to the Oral Bioavailability of Peptide-Like Compounds: Implications for the Synthesis of Renin Inhibitors. *J. Med. Chem.* **1995**, *38,* 1446-1455.
19. Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of Drug Absorption with Molecular Surface Properties. *J Pharm Sci*. **1996**, *85,* 32-39.
20. Abraham, M. H.; Chadha, H. S.; Mitchell, R. C. Hydrogen Bonding. 33. Factors that Influence the Distribution of Solutes between Blood and Brain. *J. Pharm. Sci.* **1994**, *83*, 1257-1268.
21. Basak, S. C.; Gute, B. D.; Drewes, E. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm. Res.* **1996**, *13,* 775-778.
22. Potts, R. O.; Guy, R. H. A Predictive Algorithm for Skin Perme-ability: The Effects of Molecular Size and Hydrogen Bond Activity. *Pharm. Res.* **1995***, 12,* 1628-1663.
23. Yoshida, F.; Topliss, J. G. Unified Model for the Corneal Permeability of Related and Diverse Compounds with Respect to Their Physico-chemical Properties. *J. Pharm. Sci.* **1996**, *8*5, 819-823.
24. Gobburu, J. V. S.; Shelver, W. H. Quantitative Structure-Pharmaco-kinetic Relationships (QSPR) of Beta Blockers Derived Using Neural Networks. *J. Pharm. Sci*. **1995**, *84*, 862-865.
25. Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 726-735.
26. Clark, D. E.; Pickett, D. S. Computational Methods for the Prediction of 'Drug-Likeness'. *Drug Disc. Today*. **2000**, *5*, 49-58.

27. Fujiwara. S-I.; Yamashita, F.; Hashida, M. Prediction of Caco-2 Cell Permeability Using a Combination of MO-Calculation and Neural Network. *Int. J. Pharm*. **2002**, *237*, 95-105.
28. Kulkarni, A.; Han, Y.; Hopfinger, J. Predicting Caco-2 Cell Permeation Coefficients of Organic Molecules Using Membrane-Interaction QSAR Analysis. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 331-342.
29. Van der Waterbeemd, H.; Kansy, M. Hydrogen-bonding Capacity Permeability Using Calculated Molecular Descriptors. *Quant. Struct.-Act Relat*. **1992***, 15*, 480-490.
30. Krarup, H.; Christensen, T. I.; Hovgaard, L.; Frokjaer, S. Predicting Drug Absortion From Molecular Surface Properties Based on Molecular Dynamics Simulations. *Pharm. Res*. **1998**, *15*, 972-978.
31. Norinder, U.; Osterber, T.; Artursson, P. Theoretical Calculation and Prediction of Caco-2 Cell Permeability Using MolSurf Parameterization and PLS Statistics. *Pharm. Res*. **1997**, *14*, 1786-1791.
32. Diudea, M. V. (Ed.), *QSPR/QSAR Studies by Molecular Descriptor*s, Nova Science, Huntington, N.Y., **2001**.
33. Ivanciuc, O.; Ivanciuc, T.; Cabrol–Bass, D.; Balaban, A. T. Evaluation in Quantitative Structure–Property Relationship Models of Structural Descriptors Derived from Information–Theory Operators. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 631–643.
34. Balaban, T.; Mills, D.; Ivanciuc, O.; Basak, S. C. Reverse Wiener Indices. *Croat. Chem. Acta***.** **2000**, *73,* 923.
35. Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Wiener Index Extension by Counting Even/Odd Graph Distances. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 536–549.
36. Ivanciuc, O. Building–Block Computation of the Ivanciuc–Balaban Indices for the Virtual Screening of Combinatorial Libraries. *Internet Electron. J. Mol. Des.* **2002**, *1*, 1–9, http://www.biochempress.com.
37. Rios–Santamarina, I.; García–Doménech, R.; Cortijo, J.; Santamaría, P.; Morcillo, E. J.; Gálvez, J. Natural Compounds with Bronchodilator Activity Selected by Molecular Topology. *Internet Electron. J. Mol. Des*. **2002**, *1*, 70–79, http://www.biochempress.com.
38. Marino, D. J. G.; Peruzzo, P. J.; Castro, E. A.; Toropov, A. A. QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants. *Internet Electron. J. Mol. Des*. **2002**, *1*, 115–133, http://www.biochempress.com.
39. Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1412–1422.
40. Estrada, E. Spectral Moment of Edge Adjacency Matrix in Molecular Graphs.1. Definition and Application to the Prediction of Physical Properties of Alkanes. *J. Chem. Inf. Comp. Sci*. **1996**, *36*, 846-849.
41. Toropov, A. A.; Toropova, A. P. QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals. *Internet Electron. J. Mol. Des*. **2002**, *1*, 108–114, http://www.biochempress.com.

42. Liu, S. S.; Liu, H. L.; Shi, Y. Y.; Wang, L. S. QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3-Diarylcyclopentenones Based on MEDV–13. *Internet Electron. J. Mol. Des.* **2002**, *1*, 310–318, http://www.biochempress.com.

43. Estrada, E. The Structural Interpretation of the Randiæ Index. *Internet Electron. J. Mol. Des.* **2002**, *1*, 360–366, http://www.biochempress.com.

44. Gute, B. D.; Basak, S. C.; Mills, D.; Hawkins, D. M. Tailored Similarity Spaces for the Prediction of Physicochemical Properties. *Internet Electron. J. Mol. Des.* **2002**, *1*, 374–387, http://www.biochempress.com.

45. Lukovits, I.; Milièevic, A.; Nikolic, S.; Trinajstic, N. On Walk Counts and Complexity of General Graphs. *Internet Electron. J. Mol. Des.* **2002**, *1*, 388–400, http://www.biochempress.com.

46. Cao, C.; Yuan, H. A. Modified Distance Matrix to Distinguish *Cis*/*Trans* Isomers of Cycloalkanes. *Internet Electron. J. Mol. Des.* **2002**, *1*, 401–409, http://www.biochempress.com.

47. Marrero Y, Romero V. TOMO-COMD sotfware. Central University of Las Villas, 2002.

48. Cotton, F. A. *Advanced Inorganic Chemistry*, Ed. Revolucionaria, Cuba, pp. 103.

49. Randić, M. Generalized Molecular Descriptors. *J. Math. Chem.* **1991**, *7*, 155-168.

50. STATISTICA ver. 5.5, Statsoft, Inc. **1999.**

51. Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*, Wiley, New York, **1980**.

52. Needham, D. E.; Chien, W. I.; Seybold, P. G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1998**, *110*, 4186-4194.

53. Alzina, R. B. *Introduccion conceptual al análisis multivariable. Un enfoque informatico con los paquetes SPSS-X, BMDP, LISREL Y SPAD*. PPU, SA Barcelona, **1989**, charter 8, Vol. I, p.202.

54. Golbraikh, A.; Tropsha, A. Beware of q^2!. *J. Mol. Graphic Modell.* **2002**, *20*, 269-276.

55. Conradi, R. A.; Burton, P. S.; Borchardt, R. T. in: Pliska, B. Testa, H. van de Waterbeemd, Eds. *Lipophilicity in Drug Action and Toxicology*, VCH, Weinheim, **1996**, pp. 233-252.

56. Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening-an Overview, *Drug Disc Today*. **1998**, *3,* 160-178.

57. Drie, J. H. V.; Lajinees, M. S. Approaches to Virtual Library Design. *Drug Disc Today*. **1998**, *3*, 274-283.

58. Julian-Ortiz, J. V.; Gálvez, J.; Muños-Collado, C.; García- Doménech, R.; Gimeno-Cardona, C. Virtual Combinatorial Synthesis and Computational Screening of New Potential Anti-Herpes Compounds. *J Med Chem.* **1999**, *42*, 3308-3314.

59. G. L. Amidon, P. J. Sinko and D. Fleisher, Estimating Human Oral Fraction Dose Absorbed: A Correlation Using Rat Intestinal Membrane Permeability for Passive and Carrier-Mediated Compounds, *Pharm Res.* **1988**, *5*, 651-654.

60. Chong, S.; Dando, S. A.; Morrison, R. Evaluation of Biocoat Intestinal Epithelium Differentiation Environment (Accelerated Cultured Caco-2 Cells) as an Absorption-Screening Model with Improved Productivity. *Pharm Res.* **1997**, *14*, 1835-1837.

61. Rubas, W.; Jezyk, N.; Grass, G. M. Comparison of the Permeability Characteristics of a Human Colonic Epithelial (Caco-2) Cell Line to Colon of Rabbit, Monkey and Dog Intestine and Human Drug Absorption. *Pharm Res*. **1993**, *10*, 113-118.

62. Yee, S. In vitro Permeability Across Caco-2 Cells (colonic) Can Predict in vivo (Small Intestinal) Absorption in Man-Fact or Myth. *Pharm Res*. **1997**, *14*, 763-766.
63. Grès, M.; Julian, B.; Bourrié, M.; Meunier, V.; Roques, C.; Berger, M.; Boulenc, X.; Berger, Y.; Fabre, G. Correlation Between Oral Drug Absorption in Humans and Apparent Drug Permeability in TC-7 Cells, a Human Epithelial Intestinal Cell Line: Comparison with the Parental Caco-2 Cell Line. *Pharm Res*. **1998**, *15*, 726-733.
64. Walter, E.; Janich, S.; Roessler, B.J.; Hilfinger, J. M.; Amidon, G. L. HT29-MTX/Caco-2 Cocultures as an in vitro Model for the Intestinal Epithelium: In vitro-in vivo Correlation with Permeability Data From Rats and Humans. *J Pharm Sci.* **1996**, *85*, 1070-1076.
65. Stewar, B. H.; Chan, O. H.; Lu, R. H.; Reyner, E. L.; Schmid, H. L.; Hamilton, H. W.; Steinbaugh, B. A.; Taylor, M. D. Comparison of Intestinal Permeabilities Determined in Multiple in vitro and in situ Models: Relationships to Absorption in Humans. *Pharm Res*. **1995,** *12*, 693-699.