

Yovani Marrero-Ponce · Ricardo Medina Marrero
Francisco Torrens · Yamile Martinez
Milagros García Bernal · Vicente Romero Zaldivar
Eduardo A. Castro · Ricardo Grau Abalo

Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and “rational” selection of new lead antibacterial agents

Received: 16 September 2004 / Accepted: 20 June 2005 / Published online: 4 November 2005
© Springer-Verlag 2005

Abstract A novel approach (*TOMOCOMD-CARDD*) to computer-aided “rational” drug design is illustrated. This approach is based on the calculation of the non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix representing molecular structures. These *TOMOCOMD-CARDD* descriptors are introduced for the computational (virtual) screening and “rational” selection of new lead antibacterial agents using

linear discrimination analysis. The two structure-based antibacterial-activity classification models, including non-stochastic and stochastic indices, classify correctly 91.61% and 90.75%, respectively, of 1525 chemicals in training sets. These models show high Matthews correlation coefficients ($MCC=0.84$ and 0.82). An external validation process was carried out to assess the robustness and predictive power of the model obtained. These QSAR models permit the correct classification of 91.49% and 89.31% of 505 compounds in an external test set, yielding MCCs of 0.84 and 0.79, respectively. The *TOMOCOMD-CARDD* approach compares satisfactorily with respect to nine of the most useful models for antimicrobial selection reported to date. Finally, an in silico screening of 87 new chemicals reported in the anti-infective field with antibacterial activities is developed showing the ability of the *TOMOCOMD-CARDD* models to identify new lead antibacterial compounds.

Electronic Supplementary Material Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00894-005-0024-8>

Y. Marrero-Ponce (✉) · Y. Martinez
Department of Pharmacy, Faculty of Chemical-Pharmacy,
Central University of Las Villas, Santa Clara,
54830 Villa Clara, Cuba
E-mail: yovanimp@qf.uclv.edu.cu
Tel.: + 53-42-281192
Fax: + 53-42-281130

R. M. Marrero · M. G. Bernal · Y. Marrero Ponce
Department of Drug Design, Chemical Bioactive Center,
Central University of Las Villas, Santa Clara,
54830 Villa Clara, Cuba

F. Torrens · Y. Marrero-Ponce
Institut Universitari de Ciència Molecular,
Universitat de València, Dr. Moliner 50,
E-46100 Burjassot (València), Spain

V. R. Zaldivar
Faculty of Informatics, University of Cienfuegos,
Cienfuegos, Cuba

E. A. Castro
INIFTA, División Química Teórica,
Suc. 4, C.C. 16, La Plata, 1900 Buenos Aires, Argentina

R. G. Abalo
Center of Studies on Informatics,
Central University of Las Villas, Santa Clara,
54830 Villa Clara, Cuba

Keywords TOMOCOMD-CARDD software ·
Non-stochastic and stochastic linear index ·
Classification model · LDA-based QSAR ·
Antibacterial activity

Background

Drug discovery and development are highly complex processes requiring the generation of large amounts of data and information [1]. In the last decade, the pharmaceutical giants believed they could sustain growth indefinitely by dramatically increasing the rate of bringing new medicines to market simply by increasing spending and using the same research philosophies that worked in the past. The discovery of new drugs by this “old equation” is becoming less favorable because of the rise in expenditure [1, 2].

At present, however, we are on the verge of an exciting new age of drug discovery through cheminformatics, in which large amounts of data are generated using a variety of innovative technologies and the limiting step is accessing, searching and integrating this data [1–4]. The promise of cheminformatics is to reduce development times by becoming more efficient in managing the large amounts of data generated during a long drug-discovery program. Further, with managed access to all of the data, information and experience, discoveries are more likely and the expectation is that the probability of technical success will increase [1–8].

There has already been quite a change in the way in which drugs are discovered [5–11]. Particularly, the search for antibacterial compounds has always been on the desktop of molecular-modeling and drug-design specialists. In spite of this intensive search, the discovery of selective antibacterial agents has remained a largely elusive goal of antimicrobial research. Subsequently, new approaches are needed in order to make an efficient search for candidates to be assayed as antibacterial drugs. In this sense, several *in silico* methods have been used to develop QSARs on antimicrobial activity [12–17]. The effort in this area has been placed mainly into the development of structure-based classification methods, utilizing pattern-recognition techniques (such as the linear discriminant analysis (LDA), binary logistic regression (BLR) analysis and artificial neural networks (ANNs)) to predict biologically active molecules. Many 2D-physicochemical and structural descriptors were calculated in these studies, to classify the compounds into active (antibacterial) or inactive ones [12–17]. However, in all cases, the spectrum of structural patterns (diversity of chemical families) considered was small.

On the other hand, due to the widespread use and misuse of antibiotics, bacterial resistance to them has become a serious public-health problem. Some of these resistant strains, such as vancomycin-resistant enterococci (VRE) and multidrug resistant *Staphylococcus aureus* (MRSA), are capable of surviving the effects of most, if not all, antibiotics currently in use [18–28]. This recent increase in resistant bacterial infections has created a critical need to develop novel antibacterial drugs that elude existing mechanisms of resistance. For this reason, many researchers worldwide have been interested in the search and evaluation of novel lead antibacterial compounds [29–40].

In this context, our research group has recently introduced a novel scheme to perform rational *in silico* molecular designs (or selection/identification of lead drug-like chemicals) and QSAR/QSPR studies, known as *TOMOCOMD-CARDD* (acronym of *TOPological MOlecular COMputer Design-Computer Aided “Rational” Drug Design*) [41].

This method has been developed to generate molecular fingerprints based on the application of discrete mathematics and linear algebra theory to chemistry. In this sense, atomic, atom-type and total linear and quadratic molecular fingerprints have been defined in anal-

ogy to linear and quadratic mathematical maps [42, 43]. This *in silico* method has been applied successfully to the prediction of several physical, physicochemical and chemical properties of organic compounds [42–45]. In addition, *TOMOCOMD-CARDD* has been extended to consider three-dimensional features of small/medium-sized molecules based on the trigonometric-3D-chirality-correction factor approach [46].

The latter opportunity has allowed the description of significance-interpretation and comparison to other molecular descriptors [43, 44]. The approach describes changes in the electronic distribution with time throughout the molecular backbone. Specifically, the features of the *k*th total and local linear and quadratic indices were illustrated by examples of various types of molecular structures, including chain length and branching, as well as content of heteroatoms and multiple bonds [43, 44]. Additionally, the linear independence of the atom-type linear and quadratic fingerprints to other 229 0D-3D “DRAGON” molecular descriptors was demonstrated. That is to say, it was concluded that the local fingerprints are independent indices containing important structural information to be used in QSPR/QSAR and drug design studies [43, 44].

The prediction of pharmacokinetic properties of organic compounds is a problem that can also be addressed using this approach. In this sense, this method has been used to estimate the intestinal–epithelial transport of drugs in human adenocarcinoma of colon cell line type 2 (Caco-2) cultures of a heterogeneous series of drug-like compounds [47–49]. The results obtained suggest that the *TOMOCOMD-CARDD* method is able to predict the permeability values and it proved to be a good tool for studying the oral absorption of drug candidates during the drug development process.

The *TOMOCOMD-CARDD* strategy has also been useful for selecting of novel subsystems of compounds with a desired property/activity. In this sense, it was applied successfully to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and evaluated *in vivo* on *Fasciola hepatica* [50, 51].

Studies for the fast-track discovery of novel paramphistomocides and antimalarial compounds were also conducted with this theoretical approach [52, 53].

Later, promising results were obtained in modeling the interaction between drugs and the HIV Ψ -RNA packaging-region in the field of bioinformatics using the *TOMOCOMD-CANAR* (Computed-Aided Nucleic Acid Research) approach [54]. Finally, an alternative formulation of our approach for structural characterization of proteins was carried out recently [55, 56]. This extended method [*TOMOCOMD-CAMPS* (Computed-Aided Modeling in Protein Science)] was used to encompass protein stability studies, specifically how alanine substitution mutation on arc repressor wild-type protein affects protein stability, by means of a combination of protein linear or quadratic indices (macromolecular fingerprints) and statistical (linear and non-linear model) methods [55, 56].

The main objectives of this paper are, first, to find rationality in the search of novel antibacterial drug-like compounds using non-stochastic and stochastic linear indices, and second, but not less important, to continue the validation of the method for describing the biological activity of a heterogeneous series of compounds.

Theoretical approach

The theoretical framework of a *TOMOCOMD-CARD-Ds* molecular descriptor family was split into two parts; one for describing the mathematical features of non-stochastic and the other for stochastic linear indices.

Atomic, atom-type, and total non-stochastic linear indices

The atomic, atom-type and total linear indices (non-stochastic) of the “molecular pseudograph’s atom-adjacent matrix” for small-to-medium-sized organic compounds have been explained elsewhere in some detail [43]. However, an overview of this approach will be given.

For a given molecule composed of n atoms, the “molecular vector” (X) is constructed and the k th atomic linear indices, $f_k(x_i)$ are calculated as linear maps on \mathfrak{R}^n [$f_k(x_i): \mathfrak{R}^n \rightarrow \mathfrak{R}^n$; thus $f_k(x_i): \text{Endomorphism on } \mathfrak{R}^n$] in a canonical basis as shown in Eq. 1:

$$f_k(x_i) = \sum_{j=1}^n {}^k a_{ij} X_j \quad (1)$$

where, ${}^k a_{ij} = {}^k a_{ji}$ (symmetrical square matrix), n is the number of atoms in the molecule, and X_1, \dots, X_n are the coordinates or components of the “molecular vector” (X) in a canonical basis set of \mathfrak{R}^n . The components of the “molecular” vector are numerical values, which can be considered as weights (atom-labels) for the vertices of the pseudograph. Different weighting schemes can be used with this purpose, such as: (1) the atomic masses, (2) the van-der-Waals volumes, (3) the Pauling atomic electronegativities, (4) the atomic polarizabilities, and so on [57]. In this work, the Pauling electronegativities were selected as atom weights because they take into account the electronic features of each atom in the molecule, and permit adequately differentiating among atoms [58].

The coefficients ${}^k a_{ij}$ are the elements of the k th power of the symmetrical square matrix $\mathbf{M}(G)$ of the molecular pseudograph (G), and are defined as follows [42–53]:

$$\begin{aligned} a_{ij} &= P_{ij} \text{ if } i \neq j \quad \text{and } \exists e_k \in E(G) \\ &= L_{ii} \quad \text{if } i = j \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where $E(G)$ represents the set of edges of G . P_{ij} is the number of edges (bonds) between vertices (atoms) v_i and v_j , and L_{ii} is the number of loops in v_i (see Table 1).

Notice that atomic linear indices are defined as a linear transformation $f_k(x_i)$ on a molecular vector space \mathfrak{R}^n . This map is a correspondence that assigns a vector $f_k(x)$ to every vector X in \mathfrak{R}^n , in such a way that

$$f(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 f(X_1) + \lambda_2 f(X_2) \quad (3)$$

for any scalar pair (λ_1, λ_2) and any vector pair (X_1, X_2) in \mathfrak{R}^n . The defining Eq. 1 for $f_k(x_i)$ may be written as the single matrix equation:

$$f_k(x_i) = [X]^k = \mathbf{M}^k[X] \quad (4)$$

where $[X]$ is a column vector (a $n \times 1$ matrix) of the coordinates of X in the canonical basis of \mathfrak{R}^n and \mathbf{M}^k , the k th power of the matrix \mathbf{M} of the molecular pseudograph (map’s matrix).

Notice that this approach is rather similar to the LCAO-MO (Linear Combination of Atomic Orbitals-Molecular Orbital) method. Really, our approach (for $k=1$) is a quite similar approximation to the Hückel MO method due to the fact that, in our formalism, each MO ψ_i consists of n valence atomic orbitals (AOs) in the molecule.

The main idea of the LCAO-MO method is that the electrons in a molecule are accommodated in definite MOs, just as those in an atom are accommodated in definite AOs. Normally, MOs are made up as LCAO of the atoms composing the system, i.e., it can be written in the form:

$$\psi_i = \sum_{j=1}^n c_{ij} \phi_j \quad (5)$$

where i is the number of the MO ψ [in our case, $f_i(x_i)$]; j is the numbers of the atomic ϕ 1-orbitals (in our case, X_j); c_{ij} (in our case, ${}^1 a_{ij}$) are the numerical coefficients defining the contributions of individual AOs into the given MO. Such a way of constructing an MO is based on the assumption that an atom, represented by a definite set of orbitals, remains distinctive in the molecule.

Total (whole-molecule) linear indices are linear functionals (some mathematicians use the synonym linear form) on \mathfrak{R}^n . That is to say, the k th total linear index is a linear map from \mathfrak{R}^n to the scalar \mathfrak{R} [$f_k(x): \mathfrak{R}^n \rightarrow \mathfrak{R}$]. The mathematical definition of these molecular descriptors is:

$$f_k(x) = \sum_{i=1}^n f_k(x_i) \quad (6)$$

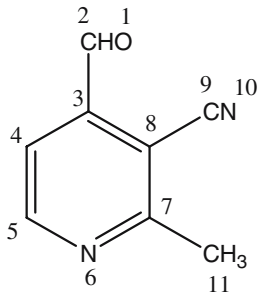
where n is the number of atoms, and $f_k(x_i)$ are the atomic linear indices (linear maps) obtained by Eq. 1. Then, a linear form $f_k(x)$ can be written in matrix form:

$$f_k(x) = [u]^t [X]^k \quad (7)$$

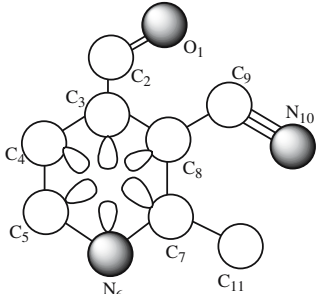
or

$$f_k(x) = [u]^t \mathbf{M}^k [X] \quad (8)$$

Table 1 Calculation of M^k (G) and S^k (G) for 2-formyl-6-methyl-benzonitrile, when k varies from 0 to 2, and i is a specific atom in the molecule



Molecular Structure



Molecular Pseudograph (G)

a_{ij}	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁	$k\delta_l$	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁	
M^0 (G)												S^0 (G)												
O ₁	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
C ₂	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
C ₃	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
C ₄	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
C ₅	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
N ₆	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
C ₇	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
C ₈	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
C ₉	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
N ₁₀	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
C ₁₁	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1
M^1 (G)												S^1 (G)												
O ₁	0	2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0
C ₂	2	0	1	0	0	0	0	0	0	0	0	3	0.66	0	0.33	0	0	0	0	0	0	0	0	0
C ₃	0	1	1	1	0	0	0	1	0	0	0	4	0	0.25	0.25	0.25	0	0	0	0.25	0	0	0	0
C ₄	0	0	1	1	1	0	0	0	0	0	0	3	0	0	0.33	0.33	0.33	0	0	0	0	0	0	0
C ₅	0	0	0	1	1	1	0	0	0	0	0	3	0	0	0	0.33	0.33	0.33	0	0	0	0	0	0
N ₆	0	0	0	0	1	1	1	0	0	0	0	3	0	0	0	0	0.33	0.33	0.33	0	0	0	0	0
C ₇	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0	0	0.25	0
C ₈	0	0	1	0	0	0	1	1	1	0	0	4	0	0	0.25	0	0	0	0.25	0.25	0.25	0	0	0
C ₉	0	0	0	0	0	0	0	1	0	3	0	4	0	0	0	0	0	0	0	0.25	0.25	0.75	0	0
N ₁₀	0	0	0	0	0	0	0	0	3	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0
C ₁₁	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
M^2 (G)												S^2 (G)												
O ₁	4	0	2	0	0	0	0	0	0	0	0	6	0.66	0	0.33	0	0	0	0	0	0	0	0	0
C ₂	0	5	1	1	0	0	0	1	0	0	0	8	0	0.625	0.125	0.125	0	0	0	0.125	0	0	0	0
C ₃	2	1	4	2	1	0	1	2	1	0	0	14	0.143	0.071	0.287	0.143	0.071	0	0.071	0.143	0.071	0	0	0
C ₄	0	1	2	3	2	1	0	1	0	0	0	10	0	0.1	0.2	0.3	0.2	0.1	0	0.1	0	0	0	0
C ₅	0	0	1	2	3	2	1	0	0	0	0	9	0	0	0.111	0.222	0.333	0.222	0.111	0	0	0	0	0
N ₆	0	0	0	1	2	3	2	1	0	0	1	10	0	0	0	0.1	0.2	0.3	0.2	0.1	0	0	0	0.1
C ₇	0	0	1	0	1	2	4	2	1	0	1	12	0	0	0.083	0	0.083	0.166	0.333	0.166	0.083	0	0.083	
C ₈	0	1	2	1	0	1	2	4	1	3	1	16	0	0.063	0.125	0.063	0	0.063	0.125	0.25	0.063	0.188	0.063	
C ₉	0	0	1	0	0	0	1	1	10	0	0	13	0	0	0.077	0	0	0	0.077	0.077	0.769	0	0	
N ₁₀	0	0	0	0	0	0	0	3	0	9	0	12	0	0	0	0	0	0	0	0.25	0	0.75	0	
C ₁₁	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0	0	0.25	0

for each molecular vector $X \in \mathbb{R}^n$. Vector $[u]^l$ is an n -dimensional unitary row vector. As can be seen, the k th total linear index is calculated by adding the local (atomic) linear indices for all atoms in the molecule.

In addition to atomic linear indices computed for each atom in the molecule, a local-fragment (atom-type) formalism can be developed. The k th atom-type linear index of the molecular pseudograph's atom-adjacency matrix is calculated by adding the k th linear indices of all atoms of the same type in the molecule. Consequently, if a molecule is partitioned into Z molecular fragments, the total

linear indices can be partitioned into Z local linear indices $f_{kL}(x)$, $L = 1, \dots, Z$. That is to say, the total linear indices of order k can be expressed as the sum of the local linear indices of the Z fragments of the same order:

$$f_k(x) = \sum_{L=1}^Z f_{kL}(x) \quad (9)$$

In the atom-type linear indices formalism, each atom in the molecule is classified into an atom-type (frag-

ment), such as heteroatoms (O, N and S), hydrogen bonding (H-bonding) to heteroatoms, halogen atoms, aliphatic carbon chain, aromatic atoms (aromatic rings), and so on. For all data sets, including those with a common molecular scaffold as well as those with diverse structure, the k th fragment (atom-type) linear indices provide much useful information.

Atomic, atom-type, and total stochastic linear indices

Notice that the linear indices matrices, \mathbf{M}^k , are graph-theoretical electronic-structure models, like the “extended Hückel MO model”. The \mathbf{M}^1 matrix considers all valence-bond electrons (σ - and π -networks) in one step, and their power k ($k=0, 1, 2, 3, \dots$) can be considered as an interacting-electronic chemical-network in step k . This model can be seen as an intermediate one between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas [59].

The present approach is based on a simple model for the intramolecular (stochastic) movement of all outer-shell electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time (t_0). In this time, the electrons are distributed around atomic nuclei. Alternatively, these electrons can be distributed around cores in discrete intervals of time t_k . In this sense, the electron in an arbitrary atom i can move to other atoms at different discrete time periods t_k ($k=0, 1, 2, 3, \dots$) throughout the chemical-bonding network.

The k th stochastic molecular pseudograph’s atom adjacency matrix [$\mathbf{S}^k(\mathbf{G})$] can be obtained from \mathbf{M}^k . Here, $\mathbf{S}^k(\mathbf{G}) = \mathbf{S}^k = [{}^k s_{ij}]$ is a squared table of order n (n = number of atoms), and the elements ${}^k s_{ij}$ are defined as follows:

$${}^k s_{ij} = \frac{{}^k a_{ij}}{{}^k \text{SUM}_i} = \frac{{}^k a_{ij}}{{}^k \delta_i} \quad (10)$$

where ${}^k a_{ij}$ are the elements of the k th power of \mathbf{M} , and the SUM of the i th row of \mathbf{M}^k are named the k -order vertex degree of atom i , ${}^k \delta_i$. The ${}^k s_{ij}$ elements are the transition probabilities to which electrons moving from atom i to j in the discrete time period t_k . Notice that the k th elements s_{ij} take into account the molecular topology in step k throughout the chemical-bonding (σ - and π -network). For instance, the ${}^2 s_{ij}$ values can distinguish between hybrid states of atoms in bonds. In this sense, it can clearly be seen from Table 1 that electrons will have a higher probability of returning to the sp N atom [$p(\text{N}_{10})=0.75$] than to the sp² N atom [$p(\text{N}_6)=0.33$] in t_2 . A similar behavior can be observed among the different hybrid states of C atoms in the molecule of 2-formyl-6-methyl-benzonitrile (see Table 1): Csp³ [$p(\text{C}_{11})=0.25$]; Csp² [$p(\text{C}_2)=0.625$]; Csp² _{arom} [$p(\text{C}_3)=0.285$, $p(\text{C}_4)=0.3$, $p(\text{C}_5)=0.33$, $p(\text{C}_7)=0.33$, $p(\text{C}_8)=0.25$]; and Csp [$p(\text{C}_9)=0.769$]. This is a logical result as the electronegativity scale of these hybrid states is taken

into account. The k th total [and local (atomic and atom-type) stochastic linear indices], ${}^s f_{\mathbf{k}}(x)$ [${}^s f_{\mathbf{k}}(x_i)$] are calculated in the same way that the linear indices (non-stochastic), but using the k th stochastic molecular pseudograph’s atom adjacency matrix, $\mathbf{S}^k(\mathbf{G})$, as mathematical linear maps’ matrices.

Materials and methods

Computational methods: TOMOCOMD-CARDD approach

TOMOCOMD is an interactive program for molecular design and bioinformatic research [41]. It is composed by four subprograms’ each of which allows drawing the structures (drawing mode) and calculating molecular 2D/3D (calculation mode) descriptors. The modules are named Computed-aided “Rational” drug design (CARDD), Computed-aided modeling in protein science (CAMPS), Computed-aided nucleic acid research (CANAR) and Computed-aided bio-polymers docking (CABPD). In the present report, we outline salient features concerned with only one of these subprograms: CARDD.

The calculation of the total and local linear indices of any organic molecule was implemented in the *TOMOCOMD-CARDD* software [41]. The main steps for the application of this method in QSAR/QSPR and drug design can be summarized briefly as follows.

1. Draw the molecular pseudographs for each molecule of the data set, using the software drawing mode. This procedure is performed by a selection of the active atomic symbol belonging to the different groups in the periodic table of the elements.
2. Use appropriate weights in order to differentiate the molecular atoms. In this study, we used the Pauling electronegativity [58] as atomic property for each kind of atom.
3. Compute the total and local (atomic and atom-type) linear indices of the molecular pseudograph’s atom-adjacency matrix. This can be carried out in the software calculation mode, where the user can select the atomic properties and the descriptor family prior to calculating the molecular indices. This software generates a table in which the rows correspond to the compounds, and columns correspond to the total and local linear indices or other molecular-descriptors family implemented in this program.
4. Find a QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis (MRA), neural networks (NN), linear discrimination analysis (LDA), and so on. That is to say, we can find a quantitative relation between an activity \mathbf{A} and the linear indices having, for instance, the following appearance:

$$\mathbf{A} = a_0 f_0(x) + a_1 f_1(x) + a_2 f_2(x) + \dots + a_k f_k(x) + c \quad (11)$$

where \mathbf{A} is the measured activity, $f_k(x)$ are the k th total linear indices, and 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 (leave-one-out cross-validation) and external (using a test set and an external predicting set) validation techniques.

The following descriptors were calculated in this work.

1. $f_k(x)$ and $f_k^H(x)$ are the k th total linear indices not considering and considering H-atoms in the molecular pseudograph (G), respectively.
2. $f_{kL}(x_E)$ and $f_{kL}^H(x_E)$ are the k th local (atom-type = heteroatoms: S, N, and O) linear indices not considering and considering H-atoms in the molecular pseudograph (G), respectively. These local descriptors are putative H-bonding acceptors.
3. $f_{kL}^H(x_{E-H})$ are the k th local (atom-type = H-atoms bonding to heteroatoms: S, N, and O) linear indices considering H-atoms in the molecular pseudograph (G). These local descriptors are putative H-bonding donors.

The k th stochastic total [$f_k(x)$ and $f_k^H(x)$] and local [$f_{kL}(x_E)$, $f_{kL}^H(x_E)$ and $f_{kL}^H(x_{E-H})$] linear indices were also computed.

Data set selection

The general performance of the current method depends decisively on the selection of compounds for the training series used to build the classifier model. The most critical aspect for constructing the training set is to guarantee wide molecular diversity in this data set. With this aim, we selected a large data set of 2,030 chemicals having great structural variability; 1,006 of them are active (antibacterial agents) and the others are non-antibacterial (1,024 compounds with other clinical uses, such as antivirals, sedative/hypnotics, diuretics, anticonvulsants, haemostatics, oral hypoglycemics, antihypertensives, antihelminthics, anticancer compounds and so on) [60–62]. The classification of these compounds as “inactive” (without antibacterial activity) does not guarantee that any of these compounds show undetected antimicrobial activity.

On the other hand, the data set of active compounds was selected by considering representatives of most of the different structural patterns and action modes of antibacterial activity. For instance, it includes antimicrobial agents that interfere with the synthesis or action of folate (sulphonamides and dihydrofolate reductase inhibitors such as trimethoprim), β -lactam antibiotics (cephalosporins, cephamycins, penicillins, monobactams and carbapenems), antimicrobial agents

affecting bacterial protein synthesis (tetracyclines, phenicols, aminoglycosides, macrolides, and lincosamides), chemicals affecting DNA girase (quinolones), miscellaneous antibacterial agents (vancomycin, polymyxin antibiotics, nitrovinylfurans, and bacitracin) and so forth. Other compounds that have no specific mode of action, but have been reported as antibacterial agents, were also included [60–62]. Figure 1 shows a representative sample of such compounds.

Later, two k-means cluster analyses (k-MCA) were performed for active and inactive series of chemicals, which allowed the dataset (2030 chemicals) to be split into training and predicting series [63, 64]. That is to say, all cases were processed using k-MCA in order to design training and predicting data series in a “rational” way. The main idea consists of carrying out a partition of either active or inactive series of chemicals in several statistically representative classes of chemicals. Thence, one may select from the members of all these classes of training and predicting series. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both series of compounds. Finally, an external cross-validation set of 87 novel antimicrobial agents was taken from recent Refs. [65, 66].

Chemometric method

The statistical software package STATISTICA was used to develop the k-MCA [67]. The number of members in each cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taken into account, to have an acceptable statistical quality of data partition in clusters. We also inspected of the standard deviation (SS) between and within clusters, of the Fisher ratios and their p -levels of significance, which were considered if lower than 0.05 [63, 64].

Afterward, a simple linear QSAR using the TOMOCOMD-CARDD method which the general formula depicted in Eq. 11 was developed. The statistical analysis was also carried out with the STATISTICA software [67]. The tolerance parameter (proportion of variance that is unique to the respective variable) or the default value for minimum acceptable tolerance was taken as 0.01. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as a strategy for model selection. In this connection, we selected the model with a high statistical significance but as few parameters (a_k) as possible and the maximizes the degrees of freedom. In Eq. 11, a_k are the coefficients of the classification function, determined by the least squares method as implemented in LDA modulus of STATISTICA [67].

The quality of the models were determined by examining Wilks' λ parameter (U -statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding p -level ($p(F)$), as well as the percentage of good

otherwise. $P(\text{active})$ and $P(\text{inactive})$ are the probabilities to which the equations classify a compound as active and inactive, respectively.

On the other hand, validation is a crucial aspect of any QSAR/QSPR modeling [68, 69]. One of the most popular validation criteria is the leave-one-out (LOO) cross-validation method (internal validation). This method systematically removes one data point at a time from the data set. A QSAR/QSPR model is then constructed based on this reduced data set and subsequently used to predict the removed data point. This procedure is repeated until a complete predicted set is obtained. Good results in this experiment can be considered as a proof of the high predictive ability of the models. However, this assumption is generally incorrect, as there may be a lack of correlation between good LOO results and high predictive ability of QSAR/QSPR models [68, 69]. Thus, the good behavior of models in an LOO procedure appears to be a necessary but not sufficient condition for models to have a high predictive power. In this sense, Golbraikh and Tropsha [69] emphasized that the predictive ability of a QSAR/QSPR model can be estimated using only a test set (external validation) of compounds that were not used for building the model, and formulated a set of criteria for the evaluation of the predictive ability of a QSAR/QSPR model. For this reason, in order to assess the predictability of the model obtained, external validation procedures were carried out. In this sense, the statistical robustness and predictive power of the model obtained was assessed using a prediction (test) set. Later, an external test set of 87 compounds was also used, in order to assess the predictive ability of the LDA models obtained [65, 66].

Finally, the calculation of percentages of global good classification (accuracy), sensibility, specificity (also

known as “hit rate”), false positive rate (also known as “false alarm rate”), and Matthews correlation coefficient (MCC) in the training and test sets allows assessment of the model [70].

Results and discussions

Training and test sets design through k-means cluster analysis

The first step in this study was the design of the training and predicting series to prevent a non-random distribution of chemicals between the two sets. This was achieved using k-MCA [63, 64]. This “rational” design of training and predicting series allowed us to design the two sets that are representative of the entire “experimental universe.”

We first carried out a k-MCA with active compounds and afterwards with inactive ones. A first k-MCA (I) split antibacterials into 20 clusters with 22, 55, 45, 56, 23, 55, 92, 21, 61, 97, 38, 4, 72, 66, 9, 76, 67, 81, 47, and 19 members. On the other hand, the inactive compound series was also partitioned into 20 clusters (k-MCA II) with 39, 58, 37, 41, 53, 105, 72, 30, 85, 74, 17, 69, 16, 38, 61, 75, 41, 4, 98, and 19 members.

Then, selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. In this sense, the training set was composed by 754 antibacterials and 771 non-antibacterials from a set of 2030 chemicals. The remaining subseries were used as test series, containing 252 active and 253 inactive compounds. Figure 2 shows above-described procedure graphically, where two independent cluster analyses (one for active and the other for inactive

Fig. 2 General algorithm used to design training and test sets throughout k-MCA

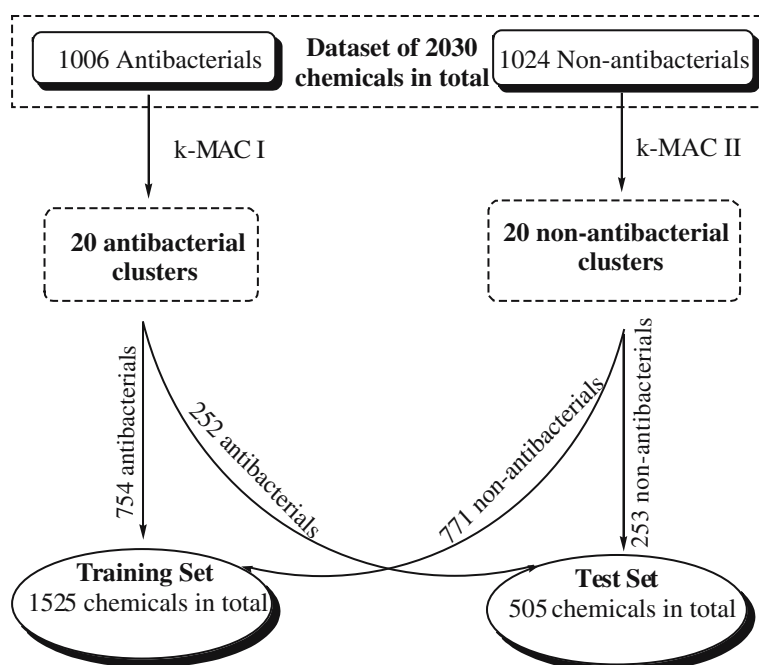


Table 2 Main results of the *k*-means cluster analysis, for antibacterial and non-antibacterial chemicals

Analysis of variance				
Total and atom-type linear indices	Between SS ^a	Within SS ^b	Fisher ratio (F)	<i>p</i> -Level ^c
Antibacterial agents clusters (k-MCA I)				
$f_6^H(x)$	862.53	163.83	273.21	0.00
$f_{11L}^H(x_E)$	1467.19	86.25	882.77	0.00
$f_{12L}^H(x_E)$	1475.90	85.83	892.38	0.00
$f_{10L}(x_{E-H})$	1036.05	346.56	155.14	0.00
Non-antibacterial agents clusters (k-MCA II)				
$f_6^H(x)$	512.90	38.14	716.29	0.00
$f_{11L}^H(x_E)$	132.12	17.34	405.88	0.00
$f_{12L}^H(x_E)$	131.99	16.23	433.18	0.00
$F_{10L}(x_{E-H})$	174.72	27.29	340.98	0.00

^aVariability between groups^bVariability within groups^cLevel of significance

compounds) were performed to select a representative sample for the training and test sets.

The *k*th total and atom-type non-stochastic linear indices were used with all variables showing *p*-levels of <0.05 for the Fisher's test. The results are depicted in Table 2.

From the k-MCA, it can be concluded that the structural diversity of several up-to-date known antibacterials (as codified by TOMOCOMD-CARDD descriptors) may be described at least by 20 statistically homogeneous clusters of chemicals.

Development of the discrimination functions

The best discrimination functions obtained, using non-stochastic and stochastic linear indices, for the training set are given below:

$$\begin{aligned} \text{Class} = & -3.5624 + 0.02f_2(x) + 2.5243 \times 10^{-4}f_5^H(x) \\ & - 1.7735 \times 10^{-4}f_6^H(x) + 6.9103 \times 10^{-6}f_{10L}(x_{E-H}) \\ & + 0.0934f_{0L}^H(x_E) - 0.028f_2^H(x_E) \\ & + 3.6536 \times 10^{-6}f_{11L}^H(x_E) - 8.4662 \times 10^{-7}f_{12L}^H(x_E) \end{aligned} \quad (12)$$

$$\begin{aligned} \text{Class} = & -3.5371 - 2.9254^s f_{11}(x) + 3.0342^s f_{15}(x) \\ & + 0.6676^s f_0^H(x) - 1.3212^s f_2^H(x) + 0.6092^s f_3^H(x) \\ & - 0.72^s f_{12L}^H(x_E) - 0.0640^s f_{14L}(x_E) \\ & + 0.9476^s f_{15L}^H(x_E) \\ N = & 1525, \quad \lambda = 0.47, \quad D^2 = 4.53, \\ F(8.1516) = & 214.82, \quad p < 0.0001 \end{aligned} \quad (13)$$

where *N* is the number of compounds, λ is the Wilks' coefficient, *F* is the Fisher's ratio, D^2 is the squared Mahalanobis distance and *p*-value is the significance level.

In Table 3, we show the results obtained in the classification of compounds of the training set by both equations. Here, we illustrate only a small quantity of the 1525 chemicals (754 antibacterial and 771 non-antibacterial) that were used in the development of the discrimination functions. The complete set of compounds in these series, as well as their classification using both models is given as "Electronic supplementary material". In these sets, 2.75% (42/1525) and 4.26% (65/1525) of compounds were classified as false antibacterials and the 5.64% (86/1525) and 4.98% (76/1525) as false inactives, by Eqs. 12 and 13, respectively. False active and false inactive compounds are those that the model predicts as active or inactive, and they are inactive and active, correspondingly. The overall accuracy of the models (12) and (13) are respecting 91.61% (1397/1525) and 90.75% (1384/1525) for the training sets. Table 4 summarizes the results of the classifications for both models in the training sets.

One of the most important aspects of any quantitative structure-property model is its ability to predict the studied property for compounds not included in the training set. When the discrimination functions (Eqs. 12 and 13) are applied to the test sets of 505 (252 antibacterial and 253 non-antibacterial) chemicals we obtained the following results. The percentage of false actives and false inactives obtained by Eq. 12 Eq. 13 were 1.98% (10/505) [4.75% (24/505)] and 6.54% (33/505) [5.94% (30/505)], respectively. The overall accuracy of the models (12) and (13) in the test sets were 91.49% (462/505) and 89.31% (451/505), correspondingly. In Table 5, we give the classification of some compounds in the prediction sets obtained by Eqs. 12 and 13. The complete set of chemicals in these series as well as their classification using both models is also given as "Electronic supplementary material". Table 4 summarizes the results of the classifications for both models in the test sets. This table also list most parameters commonly used in medical statistics (accuracy, sensitivity, specificity and false positive rate) and the MCC for both models obtained [70]. While the sensitivity is the probability of

Table 3 Results of the classification of compounds in the training sets

Active compounds name	ΔP^a (%)		Inactive compounds name	ΔP^a (%)	
	Non-Stoch	Stoch		Non-Stoch	Stoch
Mefuralazine	84.60	65.84	Amantadine	-87.46	-94.14
Sulfathiadiazole	96.97	40.83	Cetohexazine	-57.14	-46.04
Glycylsulfanilamide	70.72	23.48	Paraldehyde	-89.53	-93.21
Septosil	94.76	50.80	Ethchlorvynol	-86.78	-79.57
Mepartricin A	97.54	49.91	Thiacetazone	-20.17	-18.74
Rifabutin	99.85	99.71	Ectylurea	-83.79	-80.08
Furidiazina	79.86	79.41	Mtrafazoline	-93.90	-92.43
Myxin	98.01	64.21	Bromobutanol	-77.42	0.82
Demethylthiolutin	67.73	47.52	Trichlorourethan	-76.41	7.92
Cefazolin	98.38	99.62	Isopral	-80.73	-48.91
Aldanil	72.20	45.64	Vernelan	-62.96	-9.62
Bluensomycin	97.81	99.32	Colestipol	-92.08	-90.51
Nitrofurantoin	68.90	72.30	Alcabrol	-81.96	-69.47
Furalazine	87.34	76.74	Oxazidione	-84.26	-83.78
Melarsenoxyd	99.29	84.32	Beclamide	-88.85	-93.34
Tetracycline	95.44	97.34	Buramate	-90.10	-96.44
Melarsen	99.06	94.57	Pheneturide	-87.73	-88.08
Chlorozotocin	11.17	86.76	Primidone	-86.08	-84.62
Dipyrrithione	95.65	24.18	Ferrosi fumaras	-62.88	-37.88
Akritoin	63.41	64.67	Iron aspartate	-62.96	-9.62
Amikacin sulfate	97.56	99.01	Clocapramine	-57.86	-88.65
Rifordin	99.79	99.71	Fructosum Ferricum	-51.60	-49.44
Coumamyacin	100.00	100.00	Diciferron	-84.92	-75.79
Esperine	99.56	99.42	Assedil	-87.88	-95.61
Nifurdazil	45.74	37.23	Besunide	66.86	71.38
Alfasol	95.61	75.74	Canrenone	-81.40	-69.24
Cordycepin	87.46	34.47	Acustasin	-66.76	-45.84
Carbadox	98.39	72.36	Merbiurelidin	-78.89	5.67
Tevenel	85.88	81.00	Pallirad	-87.38	-75.25
Azotomycin	93.48	98.84	Peucedanin	-53.54	-15.63
Bemural	93.41	78.65	Etomoxir	-46.96	-59.54
Actinomycin D	99.99	100.00	Guaifenesin	-74.81	-81.19
Dapsone	86.20	34.30	Tiforminhydrochloride	-73.14	-66.50
Ciprofloxacin hydrochloride	-54.67	11.54	Amformin	-73.58	-85.04
Temodox	97.11	53.90	Etoformin hydrochloride	-79.64	-87.46
Thiamphenicol	72.55	74.54	Clonidine hydrochloride	-3.87	-37.14
Isoniazid sodium glucuronate	75.83	60.80	Olmidine	-14.25	-30.89
Acrotiazol	-15.17	18.15	Triacetonamine	-78.83	-65.06
Dirithromycin	84.56	93.38	Dipropamine	-93.81	-99.11
Astreonam	91.36	98.05	Metadiphenii bromidum	-95.74	-97.58
Glucose-INH	58.25	26.41	Tolonidine nitrate	-46.34	-64.51
Rokitamycin	88.90	83.55	Stilonium iodide	-96.94	-98.09
Neamine	42.28	66.53	Quateron	-83.71	-87.57
Lenigron	86.48	87.86	Roflurane	-76.04	-34.82
Clobromsalan	40.80	29.50	Benzochinoniumchlorid	-41.06	-91.20

^aClassification of compounds by both models, Eq. 12 (non-Stoch.) and Eq. 13 (Stoch.): $\Delta P\% = [P(\text{active}) - (\text{inactive})] \times 100$

Table 4 Global results of the classification of compounds in the training and test sets

	Matthews corr. coefficient	Accuracy " Q_{Total} " (%)	Sensitivity "hit rate" (%)	Specificity (%)	False positive rate "false alarm rate" (%)
Non-stochastic descriptors (Eq. 12)					
Training set	0.84	91.61	88.59	94.08	5.44
Test set	0.84	91.49	86.90	95.63	3.95
Stochastic descriptors (Eq. 13)					
Training set	0.82	90.75	89.92	91.25	8.43
Test set	0.79	89.31	88.10	90.24	9.48

predicting a positive example correctly, the specificity is the probability that a positive prediction is correct. On the other hand, MCC quantifies the strength of the lin-

ear relation between the molecular descriptors and the classifications, and it may often provide a much more balanced evaluation of the prediction than, for instance,

Table 5 Results of the classification of compounds in the test sets

Active compounds name	$\Delta P\%$ ^a		Inactive compounds name	$\Delta P\%$ ^a	
	Non-Stoch	Stoch		Non-Stoch	Stoch
Tio-Urasin	96.63	59.32	PALA	-38.85	-30.35
Chloramphenicol	29.74	47.26	Foscarnet	-30.00	38.31
Furazonal	66.56	44.94	Moroxidine	-67.85	-82.02
Solupront	96.32	59.84	Urethane	-91.87	-94.20
Sulfamethoxy pyridazine	98.90	61.56	Methenamine	-98.21	-97.36
MSD-819	98.66	83.08	Amylurea	-90.70	-86.46
Chiniofon	83.32	51.79	Pentrichloral	-65.42	-39.36
Thiazosulfone	98.00	77.46	MECap	-89.28	-85.21
Sulfamethizole	92.01	93.25	Norantoin	-80.75	-82.37
Nifurprazine	74.57	59.63	Mephentoin	-77.33	-76.66
Cinerubin A	98.34	99.81	Promoxolane	-84.45	-78.16
FCE 22101	42.60	79.26	Sodium dipantoylferrate	-72.47	-37.35
Furamizole	87.44	88.13	Prorenone	-80.43	-67.28
Pyrimethamine	59.76	41.60	Pamabron	-88.89	-75.88
Bicozamycin	89.64	87.93	Propamin"soviet	-92.02	-88.78
Erythromycin C	91.13	93.60	Dopamine	-62.90	-56.41
Cefmetazole	97.42	99.60	BAEA	-82.31	-69.01
Diploicin	96.94	99.04	Pentacynium chloride	-97.79	-98.67
Cefadroxil	84.65	86.17	Oxaditon	-96.59	-98.15
Ampicillin	51.88	53.93	Tiamethonium iodide	-97.11	-96.66
Baludon	99.85	86.24	Penhexamine	-94.18	-91.26
Azoseptyl-T	95.93	93.03	Teflurane	-70.39	-33.63
Azosulfanilamide	100.00	99.92	Neothyl	-94.46	-95.82
Arsutyl	99.97	88.67	Anatiroidol	-90.07	-85.85
Fluoropolyoxin L	99.72	99.69	Cathine	-88.85	-86.55
Picloxydine	92.89	60.16	Cyclocumarol	-83.78	-28.72
Flucloxacillin	96.91	99.01	Carbimazole	-84.34	-74.78
Streptothricin F	94.60	95.74	Auxinutril	-87.50	-88.13
Novobiocin	88.92	98.79	Nafetolol	-59.32	-59.42
Streptomycin	98.98	99.38	Pentritinol	-52.48	57.92
Metacycline	93.37	95.69	Molsidomine	-11.38	-36.13
Chlortetracycline	98.20	99.03	Berberine	-61.32	-50.31
Habekacin	93.90	97.47	Punicine	-86.78	-89.22
Nocardicin A	96.98	90.93	Antafenite	-81.76	-85.08
Hygromycin	92.56	68.53	Cetovex	-94.54	-88.33
Blastmycin	90.76	65.81	Noxiptiline	-92.48	-91.48
Gentamicin X	94.64	95.77	Metamfetamine	-95.33	-95.17
Maridomycin	94.89	96.90	Closiramine aceturate	-90.04	-91.25
Tylosin	98.21	83.12	Octastine	-89.63	-94.05
Antibiotic SF-1623	97.25	97.79	Estradiol	-86.48	-84.31
Carumonan	99.27	99.72	Tiadenol	-89.82	-95.13
YM-13115	99.85	99.99	Metiapine	-75.10	-87.32
Actinomycin C3	99.99	100.00	Azabuperone	-73.68	-80.47
Antibiotic Ro 21-6150	92.23	87.33	Dienestrol	-90.31	-61.77
Antibiotic LL-BM123 alpha	100.00	99.96	Lost	-90.38	-80.18

^aClassification of compounds by both models, Eq. 12 (non-Stoch.) and Eq. 13 (Stoch.): $\Delta P\% = [P(\text{Active}) - (\text{Inactive})] \times 100$

the percentages [70]. The models obtained, Eqs. 12 and 13, showed a high MCC of 0.84 (0.84) and 0.82 (0.79) in training and test sets, respectively.

TOMOCOMD-CARDD method *versus* other cheminformatic approaches

Recently, several *in silico* methods have been used to develop structure-based classification models of antimicrobial activity, which give rise to good discrimination of this activity in large and heterogeneous series of organic compounds [12–17]. However, because of differences in the composition of experimental data and

chemometric methods used to carry out the QSAR, it is not feasible to perform a comparison among the models reported in the literature for the selection of antibacterial agents. For this reason, a “strict” comparison between the methodologies is not possible. Thus, a relative comparison will be based on the kind of method used to derive the QSARs and their statistical parameters, the explored molecular descriptors, the number and diversity of chemical structural patterns contained in the data, the overall accuracy (%) and the validation method used. Table 6 shows the comparison between the *TOMOCOMD-CARDD* method and other reported approaches for antimicrobial activity.

Table 6 Comparison between TOMOCOMD-CARDD method and other chminformatic approaches, for antimicrobial activity

Models' features to be compared ^a	Structure-Based Classification Models of Antibacterial Activity										
	Eq. 12	Eq. 13	1	2	3	4	5	6	7	8	9
<i>N</i> total	2030	2030	111	111	664	596	661	661	352	433	433
<i>N</i> antibacterials	1006	1006	60	60	249	307	249	249	219	217	217
Technique ^b	LDA	LDA	LDA	ANN	ANN	LDA	LDA	BLR	LDA	LDA	ANN
Wilks' λ (U-statistics)	0.46	0.47	0.28	–	–	0.57	N. R.	–	0.45	–	–
F	226.61	214.82	20.9	–	–	116.6	N. R.	–	48.2	–	–
<i>D</i> ²	4.78	4.53	N. R.	–	–	N. R.	N. R.	–	4.9	–	–
<i>p</i> -Level	0.00	0.00	0.00	–	–	N. R.	N. R.	–	0.00	–	–
Explored variables	75	75	16	16	62	N. R.	167	167	15	62	62
Variables in the model	8	8	7	16	62	3	6	6	7	6	62
Training set											
<i>N</i> total	1525	1525	64	64	465	463	661	661	289	305	305
<i>N</i> antibacterials	754	754	34	34	174	242	249	249	174	153	153
Accuracy (%)	91.61	90.75	94.0	89.0	N.R.	–	92.6	94.7	91.0	~85.7	~98.7
Families of drugs ^c	Broader range	Broader range	3	3	8	–	8	8	8	8	8
Validation method											
Validation method ^d	i	i	i	i	i	i	ii	ii	i	i	i
<i>N</i> total	505	505	47	47	199	133	–	–	63	128	128
<i>N</i> antibacterials	252	252	26	26	75	65	–	–	45	64	64
Predictability (%)	91.49	89.31	92	97.9	~95	84	93.6	94.3	89.0	~87.5	~91.4
Families of drugs ^c	Broader range	Broader range	3	3	8	–	–	–	5	6	6

^aEquations 12 and 13 are reported in this work, models 1 and 2 were reported by Domenech and de Julián-Ortiz [12], model 3 was reported by Tomás-Vert et al. [13], model 4 was reported by Mishra et al. [14], models 5 and 6 are after Cronin et al. [15], model 7 was reported by Molina et al. [16] and models 8 and 9 were reported by Murcia-Soler et al. [17]

^bLDA linear discriminant analysis, ANN artificial neural network and BLR binary logistic regression

^cOnly largely represented families were considered, e.g., methods 1 and 2 used 3 in training quinolones, sulphonamides, and cephalosporins but add only diaminopyridine (1 compound), cephamicins (2), oxacephems (1) and sulfones (1) to predicting series.

^dValidation methods are: (i) test set, and (ii) leave-30%-out

First, *TOMOCOMD-CARDD* data set has more than 18 (17), 3 (4), 3 (3), 3 (4), 6 (4), and 5(4) times the number of chemicals (antibacterial compounds) with respect to the models reported by Domenech and de Julián-Ortiz [12], Tomás-Vert et al. [13], Mishra et al. [14], Cronin et al. [15], Molina et al. [16] and Murcia-Soler et al. [17], respectively. Furthermore, all models

recognized the existence of antibacterial and non-antibacterial groups of compounds significantly.

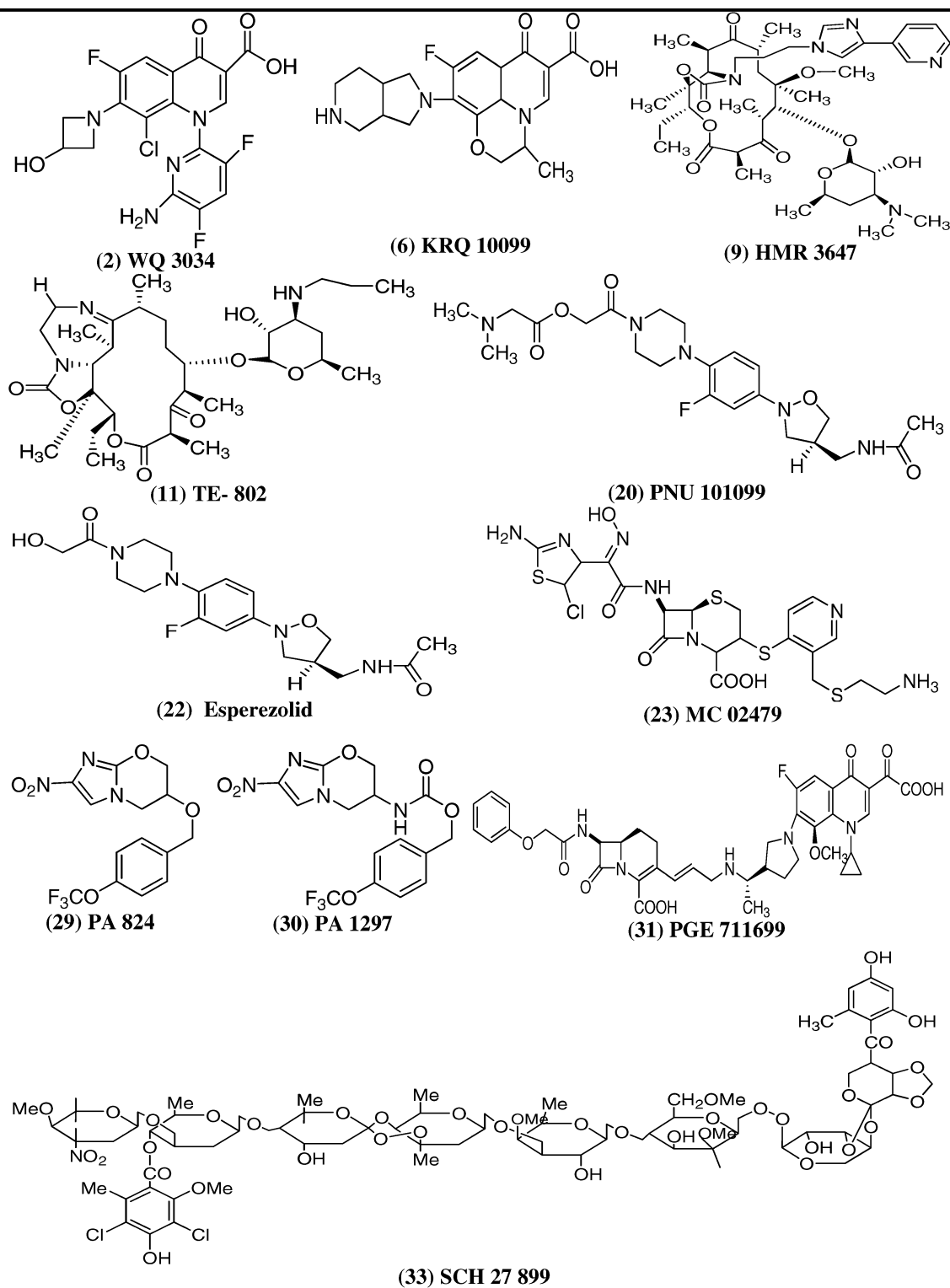
The global good classification in the training set of *TOMOCOMD-CARDD* models (Eq. 12=91.61% and Eq. 13=90.75%) was better than most reported LDA equations (see Table 6). Conversely, a connectivity function [12], the BLR model [15], and an ANN model

Table 7 Results of the virtual screening simulation of novel antimicrobial agents

Chemicals ^a	$\Delta P\%$ ^b		Chemicals ^a	$\Delta P\%$ ^b	
	Non-Stoch	Stoch		Non-Stoch	Stoch
2 WQ 3034	99.18	99.82	54 ABT 773	65.51	90.41
6 KRQ 10099	33.56	59.01	64 DK-35C	53.81	74.98
9 HMR 3647	85.02	94.68	67 J 111, 225	–7.42	–10.01
11 TE-802	20.81	68.93	69 LB 10827	100.00	99.99
20 PNU 101099	29.15	14.09	75 Psammaplin A	97.79	98.49
22 Esperezolid	3.79	–31.73	76 Bisbenzylamide eromomycin	100.00	100.00
23 MC 02479	99.13	99.73	77 HKI 9724037	99.99	99.99
29 PA 824	72.05	17.88	78	62.93	38.80
30 PA 1297	82.77	38.14	79 SEP 137199	27.83	36.40
31 PGE 711699	99.80	99.46	80 SEP 32196	74.58	65.96
33 SCH 27 899	100.00	100.00	82 KY-9	63.72	68.94
39 PGE 4175997	–12.45	40.85	83 Ro 62-6091	84.86	77.85
41 NFSQ	91.79	91.65	84 Ro 64-5781	92.69	89.09
47 KB 5290	50.43	75.02	85 VRC 483	–0.47	24.70
52 RU 79115	81.81	84.89	86 9567 567	99.63	94.24

^aChemicals 1–33 and 34–87 were taken from Refs. 65, 66, respectively. The molecular structures of these compounds are illustrated in Table 8 (see “Electronic Supplementary Material” to obtain the complete set of chemicals in this set)

^bClassification of compounds by both models, Eq. 12 (non-Stoch.) and Eq. 13 (Stoch.): $\Delta P\% = [P(\text{active}) - (\text{inactive})] \times 100$

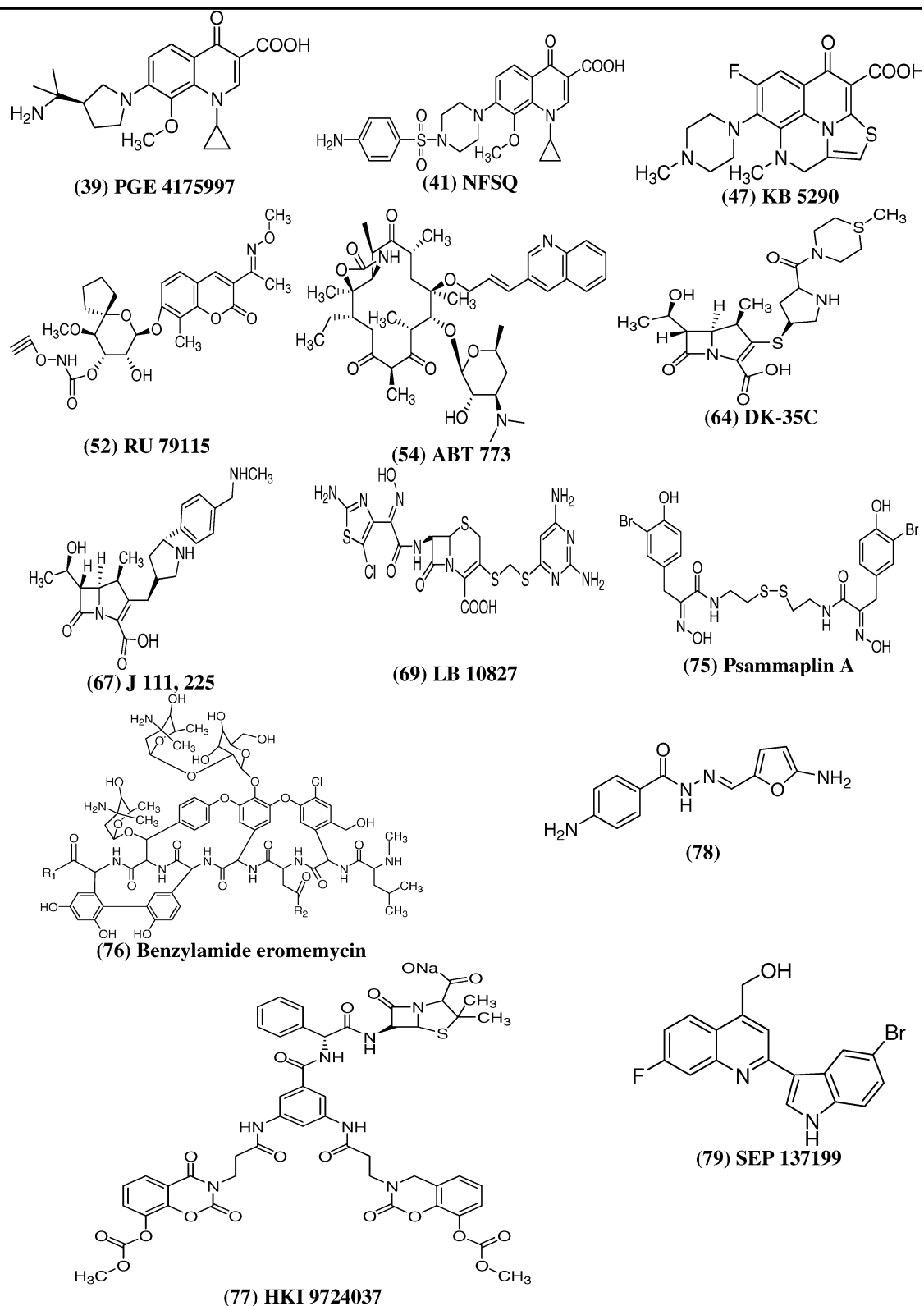
Table 8 Structure of new compounds reported in the anti-infective field with antibacterial activity

[17] gave an overall predictability of 94%, 94.7% and 98.7%, respectively, which seem to be larger than the *TOMOCOMD-CARDD* functions' predictability. Nevertheless, it is remarkable that the *TOMOCOMD-CARDD* models were derived from training series 23 (1525/64), 2 (1525/661), and 5 (1525/305) times larger

than the series used by Domenech and de Julián-Ortiz [12], Cronin et al. [15] and Murcia-Soler et al. [17], respectively.

On the other hand, Golbraikh and Tropsha established a set of criteria to assess the predictive ability of QSAR models, emphasizing that it can only be esti-

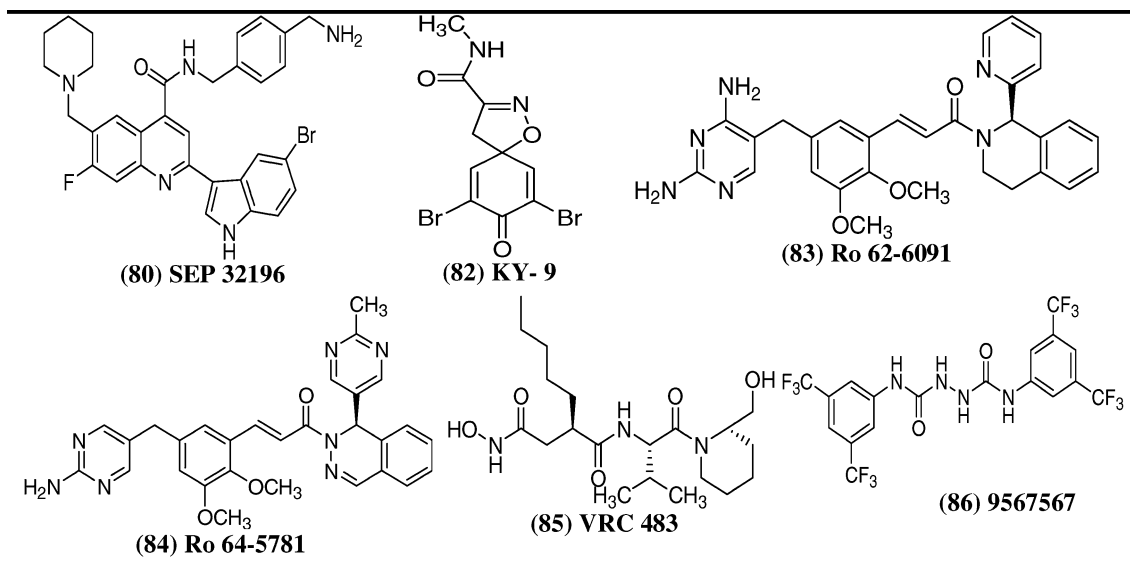
Table 8 (Contd.)



mated using an external test set (external validation) of compounds [69]. However, the model predictability obtained by Cronin et al. [15], was assessed and validated

by the random removal of 30% of the compounds to form a test set, for which predictions were made from the model. Conversely, the rest of the models reported

Table 8 (Contd.)



were validated successfully by means of external prediction series.

In this sense, the overall accuracy in test sets of *TOMOCOMD-CARDD* models (Eq. 12 = 91.49% and Eq. 13 = 89.31%) was higher than those in the rest of the reported LDA equations (see Table 6). Only two non-linear (ANN) models (Eqs 2 [12] and (3) [13] in Table 6) have larger predictabilities, but use a much-reduced number of antibacterial compounds (26 and 75, respectively) than the *TOMOCOMD-CARDD* approach (252 antibacterial agents).

Another remarkable problem, especially in the case of classification of heterogeneous series of chemicals, is the spectrum of structural patterns considered. Without doubt, the *TOMOCOMD-CARDD* models reported consider a great diversity of antimicrobial families (see Tables S3 and S5 of supporting information to obtain the complete list of 2030 compounds used in the training and test sets), taking into account that all previous studies added just a few compounds to only three to seven families in the predicting series.

Computational screening of new compounds reported in the anti-infective field with antibacterial activity

The massive cost of developing new drugs, coupled with candidate-attrition rates during the discovery and development processes, highlights the need for a “see change” in the drug discovery paradigm. Predictive in silico models could be used for identifying the desired activity, accelerating the selection process of lead compounds [71]. One of the most important features of any QSAR model is its ability to predict the desired activity, for new compounds from databases of chemicals [5–17]. Computational in silico screening (based on QSAR

techniques) of large databases, considering the use of such models, has emerged as an interesting alternative to high-throughput screening (HTS) and an important drug-discovery tool [5–11, 72, 73]. With the aim of proving the possibilities of the *TOMOCOMD-CARDD* approach to detect new lead compounds, we performed a simulated virtual screening of 87 new organic-chemical drugs reported in the anti-infective field with antibacterial activity [65, 66]. These chemicals were evaluated by *TOMOCOMD-CARDD* models (Eqs. 12 and 13) as active/inactive. The ability of the models to classify these compounds and their molecular structures are recorded in Tables 7 and 8, respectively. Here we illustrate only a small quantity of these chemicals. The complete set of compounds in this set, their molecular structures, as well as their classification using both models is given as “Electronic supplementary material.”

As can be seen, both models (Eqs. 12 and 13) classify most of the 87 selected compounds correctly, showing an overall accuracy of 90.81% and 96.55%, respectively.

Some of these chemicals are new lead antibacterial agents. That is to say, no compound with this kind of structure was included in the training data set for developing models (12) and (13). This in silico evaluation is equivalent to the discovery of new lead compounds using the models developed. In this way, new lead compounds could be designed using the *TOMOCOMD-CARDD* method described in the present report.

Concluding remarks

This study has examined a large dataset of compounds with considerable structural variability, which has been classified according to their antibacterial activity. In this

sense, the collected data of antibacterial chemicals used in this computational screening is an important tool, not only for the theoretical research but also for the general scientific work in this area.

In addition, the *TOMOCOMD-CARDD* approach (atom-type and total non-stochastic and stochastic linear indices) was used to obtain quantitative QSAR-LDA models that discriminate antibacterial compounds from inactive ones. The models obtained were significant from the statistical point of view and compare satisfactorily with respect to nine of the most useful structure-based classification equations for antimicrobial selection reported to date. Computational in silico screening of 87 drug-like compounds with antibacterial activity was carried out by us to prove the usefulness of the present approach to discover new antibacterial agents from 2D-structural chemical databases or combinatorial libraries.

Supplementary materials

The complete list of compounds used in the training and prediction sets, as well as their a posteriori classification according to models (12) and (13) are available as Supplementary materials.

Acknowledgments The authors thank both referees for their critical opinions about the manuscript, which have significantly contributed to improving its presentation and quality. F.T. acknowledges financial support from the Spanish MCT (Plan Nacional I+D+I, Project No. BQU2001-2935-C02-01) and Generalitat Valenciana (DGEUI INF01-051 and INFRA03-047, and OCYT GRUPOS03-173). Finally, Marrero-Ponce is also indebted to the editorial assistant, Isabelle Bundesmann for her kind attention. One of the authors (M-P. Y) thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University.

References

- Claus BL, Underwood DJ (2002) *Drug Disc Today* 7:957–966
- Mosqueira A (ed) (1994) *Diseño de Medicamentos*. Farmaindustria, Madrid
- Hass LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC (2000) *IBM Syst J* 40:489–511 (available online at: <http://www.research.ibm.com/journal/sj/402/haas.html>)
- Ricadela A (2001) *Information Week*. March 1 (available online at: <http://www.informationweek.com/828/research.htm>)
- Estrada E, Uriarte E, Montero A, Teijeira M, Santana L, De Clercq E (2000) *J Med Chem* 43:1975–1985
- Julián-Ortiz JV, Gálvez J, Muñoz-Collado C, García-Domenech R, Gimeneo-Cardona C (1999) *J Med Chem* 42:3308–3314
- González-Díaz H, Marrero-Ponce Y, Hernández I, Bastida I, Tenorio E, Nasco O, Uriarte U, Castañedo N, Cabrera MA, Aguila E, Marrero O, Morales A, Pérez M (2003) *Chem Res Toxicol* 16:1318–1327
- García-García A, Gálvez J, de Julián-Ortiz J-V, García-Domenech R, Muñoz C, Guna R, Borrás R (2004) *J Antimicrob Chemother* 53:65–73
- Estrada E, Peña A (2000) *Bioorg Med Chem* 8:2755–2770
- Gozalbes R, Galvez J, Moreno A, Garcia-Domenech R (1999) *J Pharm Pharmacol* 51:111–117
- González H, Olazabal E, Castañedo N, Hernández I, Morales A, Serrano HS, González J, Ramos R (2002) *J Mol Mod* 8:237–245
- García-Domenech R, de Julián-Ortiz JV (1998) *J Chem Inf Comput Sci* 38:445–449
- Tomás-Vert F, Pérez-Giménez F, Salabert-Salvador MT, García-March FJ, Jaén-Oltra J (2000) *J Mol Struct (Theochem)* 504:249–259
- Mishra RK, Garcia-Domenech R, Galvez J (2001) *J Chem Inf Comput Sci* 41:387–393
- Cronin MTD, Aptula AO, Dearden JC, Duffy JC, Netzeva TI, Patel H, Rowe PH, Schultz TW, Worth AP, Voutzoulidis K, Schüürmann G (2002) *J Chem Inf Comput Sci* 42:869–878
- Molina E, González-Díaz H, Pérez-González M, Rodríguez E, Uriarte E (2004) *J Chem Inf Comput Sci* 44:515–521
- Murcia-Soler M, Pérez-Giménez F, García-March J, Salabert-Salvador Ma T, Díaz-Villanueva W, Castro-Bleda M, Villanueva-Pareja A (2004) *J Chem Inf Comput Sci* 44:1031–1041
- Galimand M, Courvalin P, Lambert T (2003) 47:2565–2571
- Tenover FC (2001) *Clin Infect Dis* 33:S108–S115
- Murray BE (2000) 342:710–721
- Williams RJ, Heymann DL (1998) *Science* 279:1153–1154
- Livermore DM (2000) *Int J Antimicrob Agents* 16:S3–S10
- Murray BE (1998) *Emerg Infect Dis* 4:37–47
- Cetinkaya Y, Falk P, Mayhall CG (2000) *Clin Microb Rev* 13:686–707
- Amyes SGB, Towner KJ, Carter GI, Thomson CJ, Young HK (1989) *J Antimicrob Chemother* 24:111–119
- Bambeke FV, Glupczynski Y, Plésiat P, Pechère JC, Tulkens PM (2003) *J Antimicrob Chemother* 51:1055–1065
- Wylie BA, Amyes SGB, Young HK, Koornof HJ (1988) *J Antimicrob Chemother* 22:429–435
- Nagai K, Davies TA, Jacobs M R, Appelbaum PC (2002) *Antimicrob Agents Chemother* 46:1273–1280
- Jung F, Delvare C, Boucherot D, Hamon A (1991) *J Med Chem* 34:1110–1116
- Fung-Tomc JC, Clark J, Minassian B, Pucci M, Tsai YH, Gradelski E, Lamb L, Medina I, Huczko E, Kolek B, Chaniecki S, Ferraro C, Washo T, Bonner DP (2002) *Antimicrob Agents Chemother* 46:971–976
- Macchia M, Menchini E, Orlandini E, Rossello A, Broccoli G, Visconti M (1995) *Farmaco* 50:713–718
- Choi KH, Hong JS, Kim SK, Lee DK, Yoon SJ, Choi EC (1997) *J Antimicrob Chemother* 39:509–514
- Sum PE, Petersen P (1999) *Bioorg Med Chem Lett* 17:1459–1462
- Chopra I (2001) *Curr Opin Pharmacol* 1:464–469
- Colca JR, McDonal WG, Waldon DJ, Thomasco LM, Gadwood RC, Lund ET, Cavey GS, Mathews WR, Adams LD, Cecil ET, Pearson JD, Bock JH, Mott JE, Shinabarger DL, Xiong L, Mankin AS (2003) *J Biol Chem* 278:21972–21979
- Domagala JM, Bridges AJ, Culbertson TP, Gambino L, Hagen SE, Karrick G (1991) *J Med Chem* 34:1142–1154
- Kus C, Göker H, Ayhan G, Ertan R, Altanlar N, Akin A (1996) *Farmaco* 51:413–417
- Payne DJ, Miller WH, Berry V, Brosky J, Burgess WJ, Chen E, DeWolf WE, Fosberry AP, Greenwood R, Head MS, Heerding DA, Janson CA, Jaworski DD, Keller PM, Manley PJ, Moore TD, Newlander C, Pearson S, Polizzi BJ, Qiu X, Rittenhouse SF, Radosti S, Salyers KL, Seefeld MA, Smyth MG, Takata DT, Uzinskas IN, Vaidya K, Wallis NG, Winram SB, Yuan CCK, Huffman WF (2002) *Antimicrob Agents Chemother* 46:3118–3124
- Chopra I, Hodgson J, Metcalf B, Poste G (1997) *Antimicrob Agents Chemother* 41:497–503
- García-Garrote F, Cercenado E, Martín-Pedroviejo J, Cuevas O, Bouza E (2001) *J Antimicrob Chemother* 47:681–684
- Marrero-Ponce Y, Romero V (2002) *TOMOCOMD* software. Central University of Las Villas. *TOMOCOMD (TOPOlogical MOlecular COMputer Design)* for Windows, Version 1.0 is a

- preliminary experimental version; in future a professional version will be obtained upon request to Marrero Y: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es
42. Marrero-Ponce Y (2003) *Molecules* 8:687–726
 43. Marrero-Ponce Y (2004) *J Chem Inf Comput Sci* 44:2010–2026
 44. Marrero-Ponce Y (2004) *Bioorg Med Chem* 12:6351–6369
 45. Marrero-Ponce Y, Castillo-Garit JA, Torrens F, Romero-Zaldivar V, Castro E (2004) *Molecules* 9:1100–1123
 46. Marrero-Ponce Y, González-Díaz H, Romero-Zaldivar V, Torrens F, Castro EA (2004) *Bioorg Med Chem* 12:5331–5342
 47. Marrero-Ponce Y, Cabrera MA, Romero V, Ofori E, Montero LA (2003) *Int J Mol Sci* 4:512–536
 48. Marrero-Ponce Y, Cabrera MA, Romero V, González DH, Torrens F (2004) *J Pharm Pharm Sci* 7:186–199
 49. Marrero-Ponce Y, Cabrera MA, Romero-Zaldivar V, Bermejo M, Siverio D, Torrens F (2005) *Internet Electronic J Mol Des* (accepted for publication)
 50. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castañedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, Sánchez AM, Torrens F, Castro EA (2005) *Bioorg Med Chem* 13:1005–1020
 51. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castañedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, del Valle A, Torrens F, Castro EA (2004) *J Comput Aided Mol Des* 18:615–633
 52. Marrero-Ponce Y, Huesca-Guillen A, Ibarra-Velarde F (2004) *J Theor Chem (THEOCHEM)* DOI: 10.1016/j.theochem.2004.11.027
 53. Marrero-Ponce Y, Montero-Torres A, Romero-Zaldivar C, Iyarreta-Veitía I, Mayón Pérez M, García Sánchez R (2005) *Bioorg Med Chem* 13:1293–1304
 54. Marrero-Ponce Y, Nodarse D, González-Díaz H, Ramos de Armas R, Romero-Zaldivar V, Torrens F, Castro E (2004) *Int J Mol Sci* 5:276–293
 55. Marrero-Ponce Y, Medina R, Castro EA, de Armas R, González H, Romero V, Torrens F (2004) *Molecules* 9:1124–1147
 56. Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA (2005) *Bioorg Med Chem* (accepted for publication)
 57. Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis*. Research Studies Press, Letchworth, UK
 58. Pauling L (1939) *The nature of chemical bond*. Cornell University Press, New York, pp 2–60
 59. Klein DJ (2003) *Internet Electron. J Mol Des* 2:814–834
 60. Negwer M (1987) *Organic-chemical drugs and their synonyms*. Akademie-Verlag, Berlin
 61. Chapman & Hall (1996) *The merck index*, 12th edn
 62. Glasby JS (1978) *Encyclopedia of antibiotics*. Woodhouse, Manchester
 63. Mc Farland JW, Gans DJ (1995) Cluster significance analysis. In: Manhnhold R, Krogsgaard-Larsen P, Timmerman H (eds) *Method and principles in medicinal chemistry*, vol 2. Chemometric methods in molecular design. van Waterbeemd H (ed) VCH, Weinheim, pp 295–307
 64. Johnson RA, Wichern DW (1988) *Applied multivariate statistical analysis*. Prentice Hall, NJ
 65. Bryskier A (1997) *Clin Infect Dis* 27:865–883
 66. Bryskier A (2000) *Clin Infect Dis* 31:1423–1466
 67. STATISTICA, Version 5.5 (1999) Statsoft Inc.
 68. Wold S, Erikson L (1995) Statistical validation of QSAR results. Validation tools. In: van de Waterbeemd H (ed) *Chemometric methods in molecular design*. VCH Publishers, New York, pp 309–318
 69. Golbraikh A, Tropsha A (2002) *J Mol Graphic Modell* 20:269–276
 70. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) *Bioinformatics* 16:412–424
 71. Watson C (2003) *Biosilico* 1:83–85
 72. Drie JHV, Lajiness MS (1998) *Drug Disc Today* 3:274–283
 73. Lajiness M (1990) Molecular similarity-based methods for selecting compounds for screening. In: Rouvray DH (ed) *Computational chemical graph theory*. Nova Science, New York