http://dx.doi.org/10.5935/0103-5053.20150087

J. Braz. Chem. Soc., Vol. 26, No. 6, 1218-1226, 2015.
Printed in Brazil - ©2015 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

In silico Antibacterial Activity Modeling Based on the TOMOCOMD-CARDD Approach

Juan A. Castillo-Garit,[a,b,c] Yovani Marrero-Ponce,[b,c,d,e] Stephen J. Barigye,*[f] Ricardo Medina-Marrero,[b,g] Milagros G. Bernal,[b,g] José M. G. de la Vega,[h] Francisco Torrens,[c] Vicente J. Arán,[i] Facundo Pérez-Giménez,[d] Ramón García-Domenech[d] and Rosa Acevedo-Barrios[e]

[a]*Centro de Estudio de Química Aplicada, Facultad de Química-Farmacia, Universidad Central "Marta Abreu" de Las Villas, 54830 Santa Clara, Villa Clara, Cuba*

[b]*Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research International Network (CAMD-BIR International Network), Los Laureles L76MD, Nuevo Bosque, 130015 Cartagena de Indias, Bolivar, Colombia*

[c]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, Poligon la Coma s/n, E-46980, Paterna, Spain*

[d]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, 46100 València, Spain*

[e]*Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar, 130010 Cartagena de Indias-Bolívar, Colombia*

[f]*Departamento de Química, Universidade Federal de Lavras, 3037, 37200-000 Lavras-MG, Brazil*

[g]*Chemical Bioactive Center, Department of Microbiology, Central University of Las Villas, 54830, Santa Clara-Villa Clara, Cuba*

[h]*Departamento de Química Física Aplicada, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049, Madrid, Spain*

[i]*Instituto de Química Médica, CSIC, c/ Juan de la Cierva 3, 28006, Madrid, Spain*

In the recent times, the race to cope with the increasing multidrug resistance of pathogenic bacteria has lost much of its momentum and health professionals are grasping for solutions to deal with the unprecedented resistance levels. As a result, there is an urgent need for a concerted effort towards the development of new antimicrobial drugs to stay ahead in the fight against the ever adapting bacteria. In the present report, antibacterial classification functions (models) based on the topological molecular computational design-computer aided ''rational'' drug design (TOMOCOMD-CARDD) atom-based non-stochastic and stochastic bilinear indices are presented. These models were built using the linear discriminant analysis (LDA) method over a balanced chemical compounds dataset of 2230 molecular structures, with a diverse range of structural and molecular mechanism modes. The results of this study indicated that the non-stochastic and stochastic bilinear indices provided excellent classification of the chemical compounds (with accuracies of 86.31% and 84.92%, respectively, in the training set). These models were further externally validated yielding correct classification percentages of 86.55% and 87.91% for the non-stochastic and stochastic bilinear models, respectively. Additionally, the obtained models were compared with those reported in the literature and demonstrated comparable results, although the latter were built over much smaller datasets and with much higher degrees of freedom. Finally, simulated ligand-based virtual screening of 116 compounds, recently identified as potential antibacterials, was performed yielding 86.21% and 83.62% of correct classification, respectively, and thus demonstrating the utility of the obtained TOMOCOMD-CARDD models in the search of novel compounds with desirable antibacterial activity.

Keywords: TOMOCOMD-CARDD software, atom-based bilinear index, linear discriminant analysis, antibacterial activity, QSAR, virtual screening

*e-mail: sjbarigye@gmail.com

Introduction

The race to cope with the increasing multidrug resistance of pathogenic bacteria has in the recent times lost much of its momentum, particularly due to a fundamental shift in the interest of the pharmaceutical companies. In the period 2010-2012, only one antibiotic was approved by the US food and drug administration, highlighting the tremendous fall in the development of novel antibiotics.[1] The argument put forward by pharmaceutical companies is that antibiotics are a non-viable investment given they are short course therapies compared to other chronic illnesses.[2] Therefore in the wake of the ever leaner antibiotic pipeline, antibiotic resistance has risen to unprecedented levels catching up with remedies long considered as "last resorts" such as vancomycin and carbapenem.[3] Indeed troublesome enterococci and staphylococci such as methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant enterococci (VRE) have found health professionals grasping for solutions without much in store to count on.[4-10]

Consequently, it is needless to say that there is an urgent need for concerted efforts in the development of new antimicrobial drugs (with more selectivity and less toxicity) to stay ahead in the battle against the ever adapting disease causing bacteria. During the past two decades, drug discovery research has been reoriented towards the development of theoretical and/or computational methods enabling the rational selection or design of novel agents with the desired properties, offering huge advantages in terms of the time and cost incurred in the search of novel lead compounds.[11] Recently, high throughput and ultrahigh throughput screening (HTS and uHTS) have been introduced to spearhead the identification of new lead compounds.[12] However, while these methods are rapid they are still relatively cost friendly, for many pharmacological activities, HTS endpoints may not be available.[13] In addition, because of the low number of high-quality leads derived from HTS tests (1 *per* 100 000 compounds),[14] several techniques for "recognizing drug-like molecules" have been introduced. Thus, virtual (computational) screening has emerged as an interesting alternative to HTS.[15] In this way, computational techniques are used to select a reduced number of potentially active compounds, from large available chemical or virtual combinatorial libraries. The main aim of this approach is to discriminate potential candidate molecules from inactive ones. Indeed, various *in silico* approaches have been employed in the construction of quantitative structure-activity relationship (QSAR) models for the prediction of antibacterial activity,[13,16-22] with greater emphasis on ligand-based classification methods.

The QSAR models can be categorized as local, global and universal models. Local models typically characterize a particular class of chemical compounds, usually congeneric in nature,[23,24] global models are based on a single chemical mechanism of action[25] and universal models are those based on structurally diverse datasets and corresponding to different mechanisms of action.[26-28] Although there many classification models developed in antimicrobial research so far that may be considered as universal QSARs, these were generally built on much smaller datasets explaining their narrow applicability domains (ADs), which in turn limits their usability in yielding new molecular entities (NMEs) of therapeutic interest.[13,16-21] Therefore, it is imperative to construct diverse chemical datasets (in structural and activity mechanisms terms) for QSAR modeling and to explore alternative molecular structural characterizing strategies/parameters capable of codifying orthogonal information to the existing ones, as a means of expanding the AD of the obtained models. In this context, a novel scheme to perform rational *in silico* molecular design (or selection/identification of lead drug-like chemicals) and quantitative structure activity/property relationship (QSAR/QSPR) studies, known as TOMOCOMD-CARDD (acronym of topological molecular computational design-computer aided ''rational'' drug design)[29] has recently introduced. This approach has been applied to the virtual screening of novel anthelmintic, antimalarial, antitrypanosomal, antiinflamatory and tyrosinase inhibiting compounds, which were then synthesized and evaluated *in vitro* and/or *in vivo* studies with successful results.[30-34]

The primary objective of the present report is to construct classification functions, based on the TOMOCOMD-CARDD atom-based non-stochastic and stochastic bilinear indices and a diverse chemical dataset of 2230 compounds, with the ultimate goal of screening for NMEs with possible broad spectrum antibacterial activity. The linear discriminant analysis (LDA) technique will be employed as the classification method. The key advantage of LDA is its inherent simplicity, both in terms of the lower computational cost involved with this method [compared to other non-linear methods such as artificial neural networks (ANN), support vector machine (SVM), random forest (RF), etc.] and the ease in the analysis of linear biological relationships. Comparisons with other approaches reported in the literature will be performed with the aim of assessing the performance of the built classification models. Finally, simulated virtual screening is carried out over 116 new compounds, recently reported in the literature as possessing antibacterial activity, to evaluate the true predictive ability of the TOMOCOMD–CARDD models.

Methodology

Construction of chemical dataset

It is known that greater structural variability of the training series dataset enhances the general performance of learning methods as it favors a broader AD for the classification models, critical in the successful screening the huge chemical compound databanks present today for NMEs with possibly novel modes of action. In this sense, we constructed a data set comprised of 2230 chemicals, with 1051 compounds reported in the literature as antibacterials[35-37] and the rest (1179 compounds) with other pharmacological uses.[36,37] The former were considered as active, while the latter as inactive, consistent with binary classification models. The dataset of active compounds was built considering representatives from most of the different structural patterns and action modes of antibacterial activity (see also Table S1 in Supplementary Information). 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, lincosamides), chemicals affecting DNA gyrase (quinolones), miscellaneous antibacterial agents (vancomycin, polymixim antibiotics, nitrovinylfurans, bacitracin) and many others.[38,39] Other compounds for which a specific mode of action has not been found or defined were also included.[36,37] Therefore great variability in structural and molecular mechanism terms was achieved.

Posteriorly, in order to divide the built chemical compound dataset into training and test sets, respectively, two k-means cluster analyses (k-MCAs) were performed for active and inactive series of chemicals, separately.[40,41] The main idea consists of carrying out a partition of the active and inactive series of chemicals in several statistically representative classes of chemicals. This procedure ensures that all chemical classes (as determined by the clusters derived from k-MCA) will be represented in both series of compounds.[40] Posteriorly, for each cluster, the selection of the training and test sets was performed following a random sampling procedure. As a result, the training set was composed of 799 antibacterials and 918 non-antibacterials out of a set of 1717 chemicals. The remaining group, composed of 252 antibacterials (active) and 261 compounds with different biological properties (considered as inactive in this context), was used as the test set for the validation of the models.

Finally, an external validation set comprised of 116 novel antimicrobial agents recently reported in the literature[42,43] was set aside to assess, in a simulate virtual screening experiment, the earnest predictive ability of the obtained classification models.

Molecular descriptor computation

The proper selection of molecular structural characterizing parameters for use in statistical modeling plays an important role in the quality of the models constructed thereof. While there is no general consensus on which family/class of theoretical molecular descriptors (MDs) should be preferred, orthogonality (in regard to the chemical information codified), simplicity (in terms of algorithms followed in their computation) and high discriminating power for structurally different chemicals constitute desirable characteristics to be taken into account.[44] In this sense, the so-called topological (and topo-chemical) indices (TIs) have gained increasing utility in QSPR/QSAR modeling as they generally approximate to these attributes.[44,45] The TIs are numbers that describe the structural information of molecules through graph-theoretical invariants and can be considered as structure-explicit descriptors.[46]

In previous reports, Marrero *et al.*[11,40] introduced a new family of TIs known as TOMOCOMD-CARDD bilinear indices. These MDs are based on the calculation of bilinear maps in $\Re^n$, in canonical basis sets.[47,48] The computation of the non-stochastic and stochastic bilinear indices is performed using the k^{th} "non-stochastic and stochastic graph-theoretical electronic-density matrices" denoted by M^k and S^k, respectively.[47,48] These matrix operators are graph-theoretical electronic-structural models, similar to the ''extended Hückel MO (molecular orbital) model''. The M^1 matrix considers all valence-bond electrons (σ- and π-networks) in one step, and their power k (k = 0, 1, 2, 3, ...) can be considered as an interacting-electronic chemical-network in k^{th} step. The present approach is based on a simple model for the intramolecular (stochastic) movement of all outer-shell electrons. Therefore, our approach describes changes in the electronic distribution throughout the molecular backbone with time.[47,48] These indices may be computed to characterize the totality of the entire molecular structure and/or determined fragments or characteristics of the molecule (local atom and atom-type indices). This is an essential attribute bearing in mind that the antibacterial activity may sometimes not necessarily depend on the entire molecular structure but on determined regions which interact with the inhibitory sites in the microorganism.

To automatize the computation of these indices, a free and interactive computational program denominated TOMOCOMD-CARDD was implemented. This software was in the Java programming language and is designed to operate in a parallel environment and thus maximizing the architecture of modern computers.[49] The following descriptors were calculated for this study: (*i*) k^{th} non-stochastic total bilinear indices, considering hydrogen suppressed and filled molecular pseudographs (G) denoted by $b_k(\overline{x}, \overline{y})$ and $b_k^H(\overline{x}, \overline{y})$, respectively; (*ii*) k^{th} non-stochastic local bilinear indices (for heteroatoms based on atom-types: S, N, O), for H filled and suppressed molecular pseudographs (G) denoted by $b_{kL}^H(\overline{x}_E, \overline{y}_E)$ and $b_{kL}(\overline{x}_E, \overline{y}_E)$, respectively. These local descriptors are putative H-bonding acceptors, charge and the dipole moment; (*iii*) k^{th} non-stochastic local (for atom groups based on H-atoms bonded to heteroatoms: S, N, O) bilinear indices, considering H atoms in the molecular pseudograph (G) $[b_{kL}^H(\overline{x}_{E\text{-}H}, \overline{y}_{E\text{-}H})]$. These local descriptors are putative H-bonding donors.

The k^{th} stochastic total $[{}^sb_k(\overline{x}, \overline{y})$ and $ {}^sb_k^H(\overline{x}, \overline{y})]$ and local $[{}^sb_{kL}(\overline{x}_E, \overline{y}_E)$, $ {}^sb_{kL}^H(\overline{x}_E, \overline{y}_E)$, and $ {}^sb_{kL}^H(\overline{x}_{E\text{-}H}, \overline{y}_{E\text{-}H})]$ bilinear indices were also computed. Additionally, the following properties were employed as weighting schemes for atoms in the molecular structures: atomic mass (M), atomic polarizability (P), atomic Mulliken electronegativity (K) and van der Waals volume (V).

Chemometric tools

The statistical software Statistica was used to perform the k-MCA.[50] The number of members in every cluster and the standard deviation of the variables in the cluster were taken into account (as low as possible) in order to have acceptable statistical quality of data partition in clusters. We also made an inspection of the standard deviation between and within the clusters, as well as the respective Fisher ratios and p-levels of significance, the latter was considered to be lower than 0.05.[41,51] Posteriorly, LDA-based classification models were built using the Statistica software.[50] The forward stepwise algorithm was used as the strategy for variable selection. The performance of the LDA models was assessed using the statistical parameters square Mahalanobis distance (D^2), Wilks' λ parameter (U-statistic), Fisher ratio (F), p-level [p(F)] and the percentage of correct classification in the training and test sets, respectively. The statistical robustness and predictive power of the obtained models were assessed using the test set of compounds. The binary classification variable CA with values 1 and −1 for active and inactive compounds, respectively. It follows that a compound is classified as active, if $\Delta P\% > 0$ with $\Delta P\% = [P(active) - P(inactive)] \times 100$, and as inactive otherwise [P(active) and P(inactive) are the probabilities with which the equations classify a compound as active and inactive, respectively]. 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'), Matthews correlation coefficient (MCC) and the receiver operating characteristic (ROC) curve analysis for the training (1717 compounds) and test (513 compounds) sets, respectively, permitted us to carry out the assessment of the models' robustness, predictive power and the statistical significance with respect a random classifier.[52,53] Posteriorly, external validation set (VS) of 116 novel antimicrobial agents, taken from recently published reports[42,43] were used to more vigorously assess the predictive ability of the obtained classification models in a simulated virtual screening experiment.

Results and Discussions

Development, comparison and validation of the obtained models

Development of the discriminant functions

For the training set, the best discriminant functions obtained by using atom-based non-stochastic and stochastic bilinear indices, are given below:

$$CA = -3.08 + 6.93 \times 10^{-3\,MK}b_{1L}^H(\overline{x}_{E\text{-}H}, \overline{y}_{E\text{-}H}) + 1.75 \times 10^{-3MK}b_{2L}^H(\overline{x}_E, \overline{y}_E) - 6.00 \times 10^{-3MP}b_{1L}^H(\overline{x}_E, \overline{y}_E) \quad (1)$$

$N = 1717$; $\lambda = 0.56$; $D^2 = 3.09$; $F = 440.61$; $p < 0.0001$; $MCC = 0.72$; $Q_{LS} = 86.31\%$; $Q_{PS} = 86.55\%$

$$CA = -2.96 + 14.92 \times 10^{-3\,VK}b_{7L}^H(\overline{x}_{E\text{-}H}, \overline{y}_{E\text{-}H}) + 28.14 \times 10^{-3VK}b_{1L}^H(\overline{x}_E, \overline{y}_E) - 15.07 \times 10^{-3VK}b_{1L}(\overline{x}_E, \overline{y}_E) \quad (2)$$

$N = 1717$; $\lambda = 0.57$; $D^2 = 2.95$; $F = 419.85$; $p < 0.0001$; $MCC = 0.70$; $Q_{LS} = 84.92\%$; $Q_{PS} = 87.91\%$

where N is the number of compounds, λ is Wilks' lambda, D^2 is the square Mahalanobis distance, F is the Fisher ratio, *p*-value is the significance level, MCC is the Matthews' correlation coefficient for the training set, Q_{LS} and Q_{PS} are the accuracy of the model for the training and prediction sets, respectively.

Equation 1, built from non-stochastic indices, has an accuracy of 86.31% for the training set. This model showed a good MCC of 0.72; MCC quantifies the strength of the linear relation between the MDs and the classifications, and it may often provide a much more balanced evaluation of the prediction than, for instance, the percentages (accuracy).[54] Nevertheless, the most important criterion,

Table 1. 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 MDs [eq. (1)]					
Training set	0.72	86.31	84.92	86.06	11.87
Test set	0.73	86.55	87.91	87.35	11.88
Stochastic MDs [eq. (2)]					
Training set	0.70	84.92	84.73	83.17	14.92
Test set	0.76	87.91	86.90	88.31	11.11

for the acceptance or not of a discriminant model, is based on the statistics for the test set. The non-stochastic model showed an accuracy of 86.55% (MCC = 0.73) for the compounds in the test set.

A rather similar behavior was obtained with the stochastic bilinear indices (equation 2). In this case, the model achieved an accuracy of 84.92% with a MCC of 0.70 for the training set, and for the test set an accuracy of 87.91% and MCC of 0.76; these values are similar-to-better than to those obtained with non-stochastic bilinear indices. The results of classification obtained with both models are shown in Table 1.

Figures 1 and 2 illustrate the ROC curves for the non-stochastic and stochastic bilinear index-based LDA models, respectively, for the training and test sets in each case. As can be observed, the values for the area under the curve (AUC) in the case of the former are of 0.931 and 0.938, while in the latter are 0.928 and 0.942 for the training and test sets, respectively. These scores ratify the inference that the obtained models are significantly different from a random classifier [AUC (random classifier) = 0.5].

All together, the statistical quality of the built models was satisfactory, validating the applicability of these models in virtual screening of chemical compounds. The complete set of compounds in the training and test sets, as well as their classification using both models, is given in Supplementary Information (for details see Tables S2-S4).

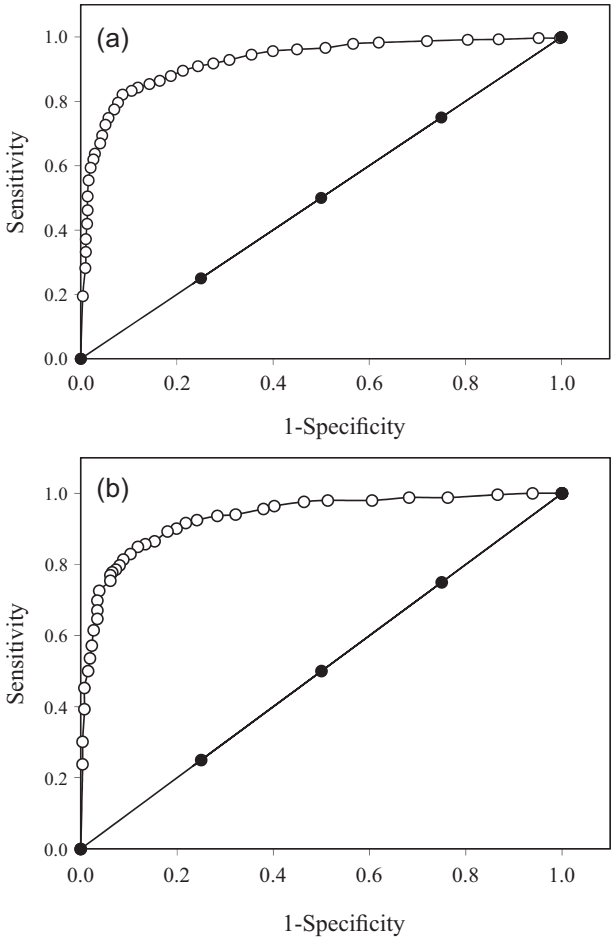

Figure 1. ROC curves for discriminant function based on the non-stochastic indices (equation 1) for the (a) training set (b) test set.

Comparison with other approaches for antibacterial activity prediction

The statistical parameters of the classification models obtained with the atom-based non-stochastic and stochastic bilinear indices were compared with those of other methods reported in the literature.[13,16-21] Nevertheless, straightforward comparisons are not possible, bearing in mind the differences in the chemometric methods and experimental data employed in the respective studies. Therefore this comparative study was based on the characteristics and statistics of the different studies such as: the fitting and validation methods for the different models and the corresponding statistics, the number and diversity of chemical compounds in the datasets and the percentages of correct classifications. The Table 2 shows a comparison of the results obtained using the bilinear indices with those reported in the literature.

As can be observed, in this study we used a much larger (2230 compounds) and more diverse (comprised of a broader range of antibacterial families) dataset, probably

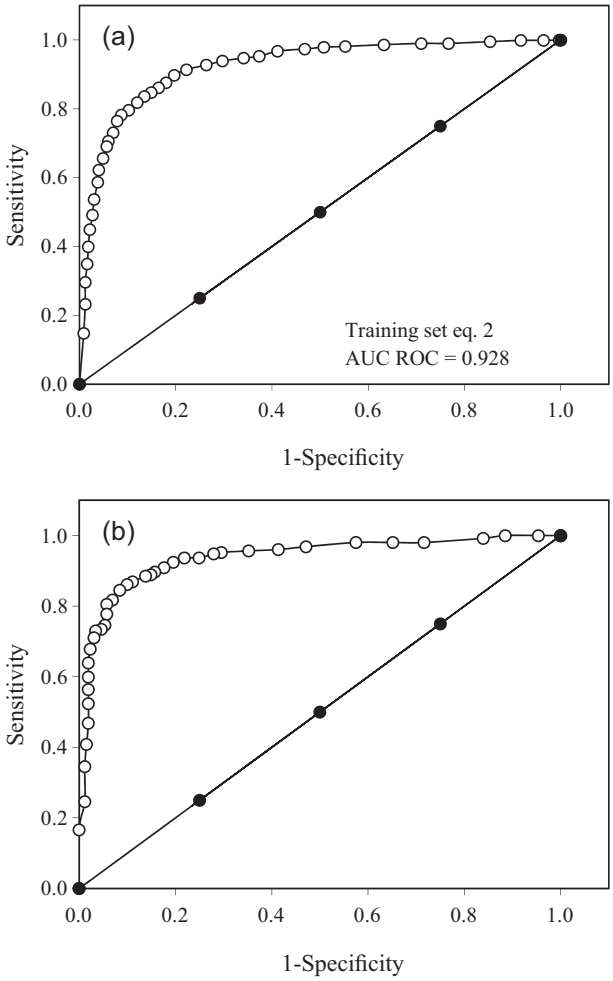

Figure 2. ROC curves for discriminant function based on the stochastic indices (equation 2) for the (a) training set (b) test set.

the largest and most diverse dataset used in antibacterial activity modeling, from the best of our knowledge. Therefore this dataset may be considered as a benchmark for posterior modeling tasks of the antibacterial activity (a complete list of the 2230 chemicals employed in the present report is available as Supplementary Information; see S2-S4). Other than a previous study performed by some of the authors of the present report,[11] the rest of the studies reported in the literature were carried out on datasets 3-20 times less than the dataset employed in the present study and with 3-8 families of antibacterial drugs.[13,16-21] The importance of the dataset structural pattern in QSAR modeling cannot be overemphasized. It is anticipated that models built with datasets with diverse families of antibiotics and mechanisms of action should increase their utility in the search of broad spectrum antibacterials, as a desirable characteristic of novel chemotherapeutic compounds. Therefore, although there are some models with superior statistics (for the training and test set) compared to those in the present report, the former are generally based on a narrower chemical space (in terms of the number and diversity of the compounds), which decreases their utility in present day virtual screening tasks.

Additionally, with the exception of the models reported by Domenech and de Julián-Ortiz,[17] the bilinear index-based models (equations 1 and 2) are much smaller size (3-variable models) compared to those reported in the literature (ranging from 6-62 variables). It is thus logical that the models in the present report possess superior Fisher ratio (F) values than the rest of the models (Table 2). This implies that the former are unlikely prone to overfitting and thus possess a greater generalizability. Also some studies reported in the literature used much more robust non-linear modeling methods [i.e., artificial neural networks (ANN) and binary logistic regression (BLR)] were used and these are known to generally yield better results (Table 2). Even then the statistics of the present models do not significantly differ from the ones obtained with these approaches.

It may therefore be concluded that the models obtained in the present report compare relatively well with those reported in the literature, bearing in mind the inherent dissimilarities in the different studies.

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

Several reports in the literature have pointed out that it is much more desirable that the obtained and validated QSAR models be further tested on chemical compounds that did not form part of the studied dataset, as this adds greater rigor to the external validation procedure.[57,58] However, due to the scarcity of experimental results on the biological activity for chemical compounds, the model building workflow is more often limited to the splitting of datasets into training and test sets, for fitting and external validation purposes, respectively, but rarely are models tested on an additional set of compounds not forming part of original datasets, in what may be denominated as "simulated" virtual screening. Therefore in this study, a search for compounds recently identified as possessing antibacterial activity was performed yielding 116 compounds[42,43] and these were evaluated using TOMOCOMD-CARDD models. The non-stochastic (eq. 1) and stochastic (eq. 2) bilinear models showed predictive accuracies of 86.21% and 83.62%, respectively (for details see, Supplementary Information for Tables S5 and S6). Considering that many of the 116 compounds in this validation set were recently identified as antimicrobial agents, this -*in silico*- evaluation may be viewed as an equivalent to the discovery of new lead compounds using the developed models. Therefore the utility, in particular

Table 2. Comparison between TOMOCOMD-CARDD discriminant functions and others chemoinformatic approaches

Models' features to be compared[a]	Ligand-based classification models of antibacterial activity												
	eq. 1	eq. 2	A	1	2	3	4	5	6	7	8	9	10
N total	2230	2230	2030	111	111	664	596	661	661	352	433	433	657
N antibacterials	1051	1051	1006	60	60	249	307	249	249	219	217	217	249
Technique[b]	LDA	LDA	LDA	LDA	ANN	ANN	LDA	LDA	BLR	LDA	LDA	ANN	ANN
Wilks'λ (U-statistics)	0.56	0.57	0.47	0.28	–	–	0.57	N. R.	–	0.45	–	–	–
F	440	419	191.03	20.9	–	–	116.6	N. R.	–	48.2	–	–	–
D^2	3.09	2.95	4.54	N. R.	–	–	N. R	N. R.	–	4.9	–	–	–
p-Level	0.00	0.00	0.00	0.00	–	–	N. R	N. R.	–	0.00	–	–	N. R.
Variables in the model	3	3	8	7	16	62	3	6	6	7	6	62	34
Training set													
N total	1717	1717	1525	64	64	465	463	661	661	289	305	305	592
N antibacterials	799	799	754	34	34	174	242	249	249	174	153	153	197
Accuracy / %	86.31	84.92	92.66	94.0	89.0	N.R.	–	92.6	94.7	91.0	ca. 85.7	ca. 98.7	93.3
Families of drugs[c]	broader range	broader range	broader range	3	3	8	–	8	8	8	8	8	8
Validation method													
Validation method[d]	i(iii)[e]	i(iii)[e]	i(iii)[f]	i	i	i	i	ii	ii	i	i	i	ii
N total	513	513	505	47	47	199	133	–	–	63	128	128	–
N antibacterials	252	252	252	26	26	75	65	–	–	45	64	64	–
Predictability / %	84.92	87.91	92.28	92	97.9	ca. 95	84	93.6	94.3	89.0	ca. 87.5	ca. 91.4	–
Families of drugs[c]	broader range	broader range	broader range	3	3	8	–	–	–	5	6	6	–

[a]Equations 1 and 2 are reported in this work, model A is reported by some of the present authors by using another dataset and different (quadratic) TOMOCOMD-CARDD MDs;[11] models 1 and 2 were reported by Domenech and de Julián-Ortiz;[17] model 3 was reported by Tomás-Vert *et al.*;[18] model 4 was reported by Mishra *et al.*,[16] models 5 and 6 are after Cronin *et al.*;[13] model 7 was reported by Molina *et al.*;[19] models 8 and 9 were reported by Murcia-Soler *et al.*;[20] model 10 was reported by Cherkasov,[21] models 11 and 12 were reported by Aptula *et al.*;[55] and model 13 was introduced by González-Díaz *et al.*;[56] [b]LDA refers to linear discriminant analysis, ANN to artificial neural network, and BLR to binary logistic regression; [c]only 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; [d]validation methods are: (*i*) test set, (*ii*) leave-many-out cross-validation (sub-sample) and (*iii*) external individual prediction set: [e]by using 116 compounds or [f]87 chemicals. N.R.: not reported.

the predictive power and generalizability, of the obtained QSAR classification models is demonstrated, which opens way for posterior virtual screening tasks of novel compounds with antibacterial activity.

Conclusions

The TOMOCOMD-CARDD based models obtained in the present study were statistically significant and compared satisfactorily with most of the ligand-based antimicrobial classification models reported up to date. Additionally, the simulated virtual screening experiment performed on the 116 compounds, recently reported as possessing antibacterial activity, revealed the predictive power and the ultimate usability of the models obtained using the TOMOCOMD-CARDD approach as a quicker and reliable alternative applicable in the virtual screening of novel antibacterial lead compounds.

On the other hand, the comprehensive chemical compounds dataset presented in the present report constitutes an important benchmark for posterior QSAR studies the modeling and/or virtual screening of novel compounds with antibacterial activity.

Supplementary Information

Supplementary information is available free of charge at http://jbcs.sbq.org.br as PDF file.

Acknowledgements

Marrero-Ponce, Y. thanks the program 'International Professor' for a fellowship to work at Universidad Tecnologica de Bolivar in 2014. Barigye, S. J. acknowledges financial support from CNPq.

References

1. Hede, K.; *Nature* **2014**, *509*, S2.
2. Projan, S. J.; Shlaes, D. M.; *Clin. Microbiol. Infect.* **2004**, *10*, 18.
3. McKenna, M.; *Nature* **2013**, *499*, 394.
4. Dimov, D.; Nedyalkova, Z.; Haladjova, S.; Schüürmann, G.; Mekenyan, O.; *QSAR Comb. Sci.* **2001**, *20*, 298.
5. Greenwood, D.; *J. Med. Microbiol.* **1998**, *47*, 751.
6. Tenover, F. C.; *Clin. Infect. Dis.* **2001**, *33*, S108.
7. Hooper, D. C.; *Clin. Infect. Dis.* **2001**, *33*, S157.
8. Collis, C.; Hall, R.; *Antimicrob. Agents Chemother.* **1995**, *39*, 155.
9. Xiong, Y.; Caillon, J.; Drugeon, H.; Potel, G.; Baron, D.; *Antimicrob. Agents Chemother.* **1996**, *40*, 35.
10. Maskell, J.; Sefton, A.; Hall, L.; *Antimicrob. Agents Chemother.* **1997**, *41*, 2121.
11. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A.; *Bioorg. Med. Chem.* **2005**, *13*, 2881.
12. Wölcke, J.; Ullmann, D.; *Drug Discovery Today* **2001**, *6*, 637.
13. Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schuurmann, G.; *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869.
14. Walters, W. P.; Murcko, A. A.; Murcko, M. A.; *Curr. Opin. Chem. Biol.* **1999**, *3*, 384.
15. Seifert, M. H. J.; Wolf, K.; Vitt, D.; *BIOSILICO* **2003**, *1*, 143.
16. Mishra, R. K.; Garcia-Domenech, R.; Galvez, J.; *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 387.
17. Garcia-Domenech, R.; de Julian-Ortiz, J. V.; *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445.
18. Tomás-Vert, F.; Pérez-Giménez, F.; Salabert-Salvador, M. T.; García-March, F. J.; Jaén-Oltra, J.; *J. Mol. Struct.: THEOCHEM* **2000**, *504*, 249.
19. Molina, E.; Diaz, H. G.; Gonzalez, M. P.; Rodriguez, E.; Uriarte, E.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
20. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A.; *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031.
21. Cherkasov, A.; *Int. J. Mol. Sci.* **2005**, *6*, 63.
22. Barigye, S. J.; Marrero-Ponce, Y.; López, Y. M.; Santiago, O. M.; Torrens, F.; Domenech, R. G.; Galvez, J.; *SAR QSAR Environ. Res.* **2013**, *24*, 3.
23. Hansch, C.; Fujita, T.; *J. Am. Chem. Soc.* **1964**, *86*, 1616.
24. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M.; *Nature* **1962**, *194*, 178.
25. Roberts, D. W.; Williams, D. L.; *J. Theor. Biol.* **1982**, *99*, 807.
26. Estrada, E.; Patlewicz, G.; Chamberlain, M.; Basketter, D.; Larbey, S.; *Chem. Res. Toxicol.* **2003**, *16*, 1226.
27. Miller, M. D.; Yourtee, D. M.; Glaros, A. G.; Chappelow, C. C.; Eick, J. D.; Holder, A. J.; *J. Chem. Inf. Model.* **2005**, *45*, 924.
28. Fedorowicz, A.; Zheng, L.; Singh, H.; Demchuk, E.; *Int. J. Mol. Sci.* **2004**, *5*, 56.
29. García-Jacas, C. R.; Marrero-Ponce, Y.; Acevedo-Martínez, L.; Barigye, S. J.; Valdés-Martiní, J. R.; Contreras-Torres, E.; *J. Comp. Chem.* **2014**, *35*, 1395.
30. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castanedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sanchez, A. M.; Torrens, F.; Castro, E. A.; *Bioorg. Med. Chem.* **2005**, *13*, 1005.
31. Marrero-Ponce, Y.; Iyarreta-Veitia, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Avila, P. E.; Kirchgatter, K.; Machado, Y.; *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 1082.
32. Marrero-Ponce, Y.; Meneses-Marcel, A.; Catillo-Garit, J. A.; Machado-Tugores, Y.; Escario, J. A.; Gómez-Barrio, A.; Montero Pereira, D.; Nogal-Ruiz, J. J.; Arán, V. J.; Martínez-Fernández, A. R.; Torrens, F.; Rotondo, R.; *Bioorg. Med. Chem.* **2006**, *14*, 6502.
33. Marrero-Ponce, Y.; Casanola-Martin, G. M.; Khan, M. T. H.; Torrens, F.; Rescigno, A.; Abad, C.; *Curr. Pharm. Des.* **2010**, *16*, 2601.
34. Marrero-Ponce, Y.; Siverio-Mota, D.; Gálvez-Llompart, M.; Recio, M. C.; Giner, R. M.; García-Domènech, R.; Torrens, F.; Arán, V. J.; Cordero-Maldonado, M. L.; Esguera, C. V.; *Eur. J. Med. Chem.* **2011**, *46*, 5736.
35. Glasby, J. S.; *Encyclopedia of Antibiotics*; Woodhouse: Manchester, 1978.
36. *The Merck Index*, 12th ed.; Chapman and Hall: London, 1996.
37. Negwer, M.; *Organic-Chemical Drugs and their Synonyms*; Akademie: Berlin, 1987.
38. Anderson, R.; Groundwater, P.; Todd, A.; Worsley, A.; *Antibacterial Agents: Chemistry, Mode of Action, Mechanisms of Resistance and Clinical Applications*; Wiley: Weinheim, 2012.
39. Flórez, J.; *Farmacología Humana,* 3ra ed.; Masson, S.A.: Barcelona, 1998.
40. Marrero-Ponce, Y.; Marrero, R. M.; Torrens, F.; Martinez, Y.; Bernal, M. G.; Zaldivar, V. R.; Castro, E. A.; Abalo, R. G.; *J. Mol. Model.* **2006**, *12*, 255.
41. Mc Farland, J. W.; Gans, D. J. In *Chemometric Methods in Molecular Design*; Waterbeemd, H., Ed.; VCH Publishers: New York, 1995; ch. 4.
42. Bryskier, A.; *Clin. Infect. Dis.* **1998**, *27*, 865.
43. Bryskier, A.; *Clin. Infect. Dis.* **2000**, *31*, 1423.
44. Karelson, M.; *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, 2000.
45. Katritzky, A. R.; Gordeeva, E. V.; *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835.
46. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach: Amsterdam, the Netherlands, 1999.

47. Marrero-Ponce, Y.; Khan, M. T. H.; Casañola-Martín, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F.; Rotondo, R.; *ChemMedChem* **2007**, *2*, 449
48. Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R.; Ortega-Broche, S. E.; Zaldivar, V. R.; *J. Math. Chem.* **2008**, *44*, 650.
49. Valdés-Martiní, J. R.; Marrero-Ponce, Y.; Barigye, S. J.; García-Jacas, C. R.; Almeida, Y. S. V. d.; Morrel, C. A.; *TOMOCOMD*; CAMD-BIT International Network; Cartagena de Indias, Bolivar, Colombia, 2014.
50. Statsoft Inc.; *STATISTICA* version 6*; Data Analysis Software System*; Tulsa, Oklahoma, 2001.
51. Johnson, R. A.; Wichern, D. W.; *Applied Multivariate Statistical Analysis*. Prentice-Hall: Englewood Cliffs, 1988.
52. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H.; *Bioinformatics* **2000**, *16*, 412.
53. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R.; *Bioinformatics* **2010**, *26*, 822.
54. Penney, K. B.; Smith, C. J.; Allen, J. C.; *J. Invest. Dermatol.* **1984**, *82*, 308.
55. Aptula, A.; Kühne, R.; Ebert, R.; Cronin, M. T. D.; Netzeva, T. I.; Schüürmann, G.; *QSAR Comb. Sci.* **2003**, *22*, 113.
56. González-Díaz, H.; Torres-Gómez, L.; Guevara, Y.; Almeida, M.; Molina, R.; Castañedo, N.; Santana, L.; Uriarte, E.; *J. Mol. Model.* **2005**, *11*, 116.
57. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A.; *J. Med. Chem.* **2013**, *57*, 4977.
58. Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V.; *J. Chem. Inf. Model.* **2008**, *48*, 766.

Submitted: October 21, 2014
Published online: April 10, 2015