



Contents lists available at ScienceDirect

## European Journal of Medicinal Chemistry

journal homepage: <http://www.elsevier.com/locate/ejmech>

## Original article

## Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs

Francisco J. Prado-Prado<sup>a,b</sup>, Eugenio Uriarte<sup>b</sup>, Fernanda Borges<sup>c</sup>, Humberto González-Díaz<sup>a,\*</sup><sup>a</sup> Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain<sup>b</sup> Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain<sup>c</sup> Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal

## ARTICLE INFO

## Article history:

Received 15 April 2009

Received in revised form

15 June 2009

Accepted 16 June 2009

Available online 24 June 2009

## Keywords:

Antibacterial drugs

Molecular descriptor

Markov model

Complex Networks

QSAR

## ABSTRACT

There are many of pathogen bacteria species which very different susceptibility profile to different antibacterial drugs. There are many drugs described with very different affinity to a large number of receptors. In this work, we selected Drug–Bacteria Pairs (DBPs) of affinity/non-affinity drugs with similar/dissimilar bacteria and represented it as a large network, which may be used to identify drugs that can act on bacteria. Computational chemistry prediction of the biological activity based on one-target Quantitative Structure–Activity Relationship (ot-QSAR) studies substantially increases the potentialities of this kind of networks avoiding time and resource consuming experiments. Unfortunately almost all ot-QSAR models predict the biological activity of drugs against only one bacterial species. Consequently, multi-tasking learning to predict drug's activity against different species with a single model (mt-QSAR) is a goal of major importance. These mt-QSARs offer a good opportunity to construct drug–drug similarity Complex Networks. Unfortunately, almost QSAR models are unspecific or predict activity against only one receptor. To solve this problem, we developed here a multi-bacteria QSAR classification model. The model correctly classifies 202 out of 241 active compounds (83.8%) and 169 out of 200 non-active cases (84.5%). Overall training predictability was 84.13% (371 out of 441 cases). The validation of the model was carried out by means of external predicting series, classifying the model 197 out of 221 (89.4%) cases. In order to show how the model functions in practice a virtual screening was carried out recognizing the model as active 86.7%, 520 out of 600 cases not used in training or predicting series. Outputs of this QSAR model were used as inputs to construct a network. The observed network has 1242 nodes (DBPs), 772,736 edges or DBPs with similar activity (sDBPs). The network predicted has 1031 nodes, 641,377 sDBPs. After edge-to-edge comparison, we have demonstrated that the predicted network is significantly similar to the observed one and both have distribution closer to exponential than to normal.

© 2009 Elsevier Masson SAS. All rights reserved.

## 1. Introduction

Infections caused by bacteria have increased dramatically during the past years. The bacteria have been the cause of some of the most deadly diseases and widespread epidemics of human civilization. Bacterial diseases such as tuberculosis, typhus, plague, diphtheria, typhoid fever, cholera, dysentery, and pneumonia have taken a mighty toll on humanity [1,2]. Water purification, immunization (vaccination) and modern antibiotic treatment continue to reduce the morbidity and the mortality of bacterial disease in the twenty-first century, at least in the developed world where these

are acceptable cultural practices. However, many new bacterial pathogens have been recognized in the past 25 years and many bacterial pathogens, such as *Staphylococcus aureus* and *Streptococcus pneumoniae*, have emerged with new forms of virulence and new patterns of resistance to antimicrobial agents [3]. At risk are patients after organ transplantation treated with immunosuppressive or those suffering with a weakened immune system, for example patients with AIDS.

Quantitative Structure–Activity Relationship (QSAR) studies may play an important role. See for instance the works of Roy et al. [4–9] using QSAR like predictive tools in this sense. Especially, QSAR models have been successfully applied to predict antibacterial drugs [10–13]. Unfortunately almost all QSAR models predict the biological activity of drugs against only one bacteria species; which means that they are one-target QSAR models (ot-QSAR).

\* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.

E-mail address: [gonzalezdiazh@yahoo.es](mailto:gonzalezdiazh@yahoo.es) (H. González-Díaz).

Consequently, predicting the probability with which a drug is active against different bacterial species with a single unify model is a goal of major importance (mt-QSAR) [14]. Till today there are nearly 1600 molecular descriptors that may be in principle generalized and used to solve the former problem [15]. Many of these indices are known as Topological Indices (TIs) or Connectivity Indices (CIs) or simply invariants of a molecular graph. In recent reviews our group has discussed advances and trends in the field including applications to study antimicrobial drugs and targets [16–18]. However, in spite of its great potential, in general, CIs and other indices have not been extended to allow Multi-tasking (mt) prediction of biological properties [19]. Multi-tasking QSAR (mt-QSAR) [19] can be defined as the prediction of multiple outputs with a single model and it is closely related to the more general term multi-tasking learning (used in cognitive sciences) [20,21]. It means that we can predict, for instance, several mechanisms of actions, partition coefficient in different biphasic systems, inhibition of different cancer lines, or activity against different microbial species to any drug using a single model. We decided to introduce information relative to the type of property to be predicted inside the molecular descriptors like spectral moments. In our mt-QSAR approach we have different value of the TIs, or structural parameter, for the same molecule depending on, for example, the specific biphasic system in which we want to estimate the partition coefficient of the drug [22] or the specific drug side effect we want to predict. This very high number of possible Drug–Bacteria Pairs (DBPs) may be investigated using Complex Networks (CNs) to regroup or cluster drugs with similar multi-bacterial affinity profile. In previous works, we constructed for the first time pairs of antiparasite [14], antifungal [23,24] and antiviral drugs [25] with multi-species predicted-activity profile and represented it as CNs. We can construct a CN of DBPs (DBP-CN) accounting for drug affinity by multiple targets. In DBP-CN the DBPs are nodes, the DBPs in the DBP-CN are interconnected by an edge if they have similar drug–bacteria, and consequently affinity, in the DBP-CNs two bacteria are interconnected if they have similar susceptibility and consequently function. In DRP-CN we need to measure the affinity of the drug on different bacterial (DBPs stability) if we cannot predict it.

We propose to construct, for the first time, a DBP-CN taking into consideration only the number of bacteria in the model and the susceptibility of the drug. The present method is also interesting because it may fail if there is no similar susceptibility-sensibility in the database. Consequently, if we propose to use the QSAR method to construct the DBP-CN we have to develop a QSAR able to predict DBPs. First, we developed the DBPs-QSAR classification model and later used the outputs of the model to construct a DBP-CN. The QSAR model proposed here is the first able to discriminate pairs of DBPs (PDBPs) with similar activity (sPDBPs) from non-similar ones (nsPDBPs). Finally, we compare both the DBP-CN predicted with a DBP-CN constructed here based on measured values of DBP affinity.

## 2. Methods

### 2.1. Computational methods

The free energy of a drug atom–receptor interaction  ${}^k g_{ij}(s)$  is considered as a state function. In this symbol  $g$  points to Gibbs energy,  $k$  indicates that we refer to the atoms interaction and  $s$  indicate that this energy depends on the specific microbial species. Afterwards, we have to define the free energy of interaction  ${}^k g_{ij}(s)$  between the  $j$ -th atom and the receptor for a specific microbial species ( $s$ ) given that  $i$ -th atom has been interacted at previous time  $t_k$ . So, one can suppose that atoms begin binding to this receptor in

discrete intervals of time  $t_k$ . However, there are several alternative ways in which such step-by-step binding process may occur.

The free energy  ${}^1 g_{ij}(s)$  can be defined by analogy as  $\Gamma_{ij}(s)$  [23,26]:

$${}^1 g_{ij}(s) = -RT \log {}^1 \Gamma_{ij}(s) \quad (1)$$

The present approach to antimicrobial–receptor interaction has two main drawbacks. The first is the difficulty on the definition of the constants. In this work, we solve the first question estimating  ${}^0 \Gamma_j(s)$  as the rate of occurrence  $n_j(s)$  of the  $j$ -th atom on active molecules against a given species with respect to the number of atoms of the  $j$ -th class in the molecules tested against the same species  $n_T(s)$ . Consequently, one of the more important steps is the change in the value of the atomic weights used  ${}^k g_{ij}(s)$  for different pathogen species.

With respect to  ${}^1 \Gamma_{ij}(s)$  we must take into consideration that once the  $j$ -th atom have interacted the preferred candidates for the next interaction are such  $i$ -th atoms bound to  $j$  by a chemical bond. Both constants can be then written down as [26]:

$${}^1 \Gamma_{ij}(s) = \left( \alpha_{ij} \frac{n_j(s)}{n_T(s)} + 1 \right) = e^{\frac{{}^1 g_{ij}(s)}{-RT}} \quad (2)$$

where  $\alpha_{ij}$  are the elements of the atom adjacency matrix,  $n_j(s)$ ,  $n_T(s)$ , and  ${}^1 g_{ij}(s)$  have been defined in the paragraph above,  $R$  is the universal gas constant, and  $T$  the absolute temperature. The number 1 is added to avoid forbidden negative values as inputs for the logarithmic function. The second problem relates to the description of the interaction process at higher times  $t_k > t_1$ . Therefore, MM theory enables a simple calculation of the probabilities with which the drug–receptor interaction takes place in the time until the antimicrobial effect is achieved. This model deals with the calculation of the probabilities ( ${}^k p_{ij}$ ) with which any arbitrary molecular atom  $j$ -th bind to the structure-less molecular receptor given that other atom  $i$ -th has been bound before; along discrete time periods  $t_k$  ( $k = 1, 2, 3, \dots$ ); ( $k = 1$  in grey), ( $k = 2$  in blue) and ( $k = 3$  in red) throughout the chemical bonding system. The procedure described here considers as states of the MM the atoms of the molecule. The method arranges all the  ${}^1 g_{ij}(s)$  free energies of interaction as a squared table of  $n \times n$  dimension. After normalization of both the vector and matrix we can build up the corresponding absolute initial probability vector  $\pi(s)$  and the stochastic matrix  ${}^1 \Pi(s)$ , which has the elements  ${}^1 \pi_{ij}(s)$  respectively; where  $m$  represents all the atoms in the molecule including the  $j$ -th,  $n_a$  is the rate of occurrence of any atom and including the  $j$ -th with value  $n_j$ . On the other hand, the matrix is called the 1-step drug–target interaction stochastic matrix.  ${}^1 \Pi(s)$  is built too as a squared table of order  $n$ , where  $n$  represents the number of atoms in the molecule. The elements  ${}^1 \pi_{ij}(s)$  of the 1-step drug–target interaction stochastic matrix are the binding probabilities with which a  $j$ -th atom binds to a structure-less molecular receptor given that other  $i$ -th atoms have been interacted before at time  $t_1 = 1$ : [26]

$$\begin{aligned} {}^1 \pi_{ij}(s) &= \frac{{}^1 g_{ij}(s)}{\sum_{a=1}^n {}^1 g_{ia}(s)} = \frac{\alpha_{ij}(-RT) \log \left( \frac{n_j(s)}{n_T(s)} + 1 \right)}{\sum_{a=1}^n \alpha_{ia}(-RT) \log \left( \frac{n_a(s)}{n_T(s)} + 1 \right)} \\ &= \frac{\alpha_{ij} \log \left( \frac{n_j(s)}{n_T(s)} + 1 \right)}{\sum_{a=1}^n \alpha_{ia} \log \left( \frac{n_j(s)}{n_T(s)} + 1 \right)} \end{aligned} \quad (3)$$

By using  ${}^0 \pi(s)$ ,  ${}^1 \Pi(s)$  and Chapman–Kolmogorov equations one can calculate the probabilities in many steps for interaction of different atoms with the receptors as the elements of the vectors  ${}^k \pi(s)$  [23,26].

$${}^k\pi_s = \text{Tr}([{}^1\Pi(s)]^k) = \text{Tr}\left(\begin{bmatrix} {}^1\pi_{11}(s) & {}^1\pi_{12}(s) & \cdot & \cdot & {}^1\pi_{1n}(s) \\ {}^1\pi_{21}(s) & {}^1\pi_{22}(s) & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ {}^1\pi(s) & \cdot & \cdot & \cdot & {}^1\pi_{nn}(s) \end{bmatrix}\right)^k \quad (4)$$

Such a model is stochastic *per se* (probabilistic step-by-step atom–receptor interaction in time) but also considers molecular connectivity (the step-by-step atom union in space throughout the chemical bonding system). The spectral moments  $\pi_k(j,s)$  for the drug–target interaction in different species are normalized so its sum for a drug molecule is equal to 1 and has no variance in a data set [27]. However, the  $\pi_k(j,s)$  can be summed for specific atom sets (AS) to create local molecular descriptors  $\pi_k(\text{AS})$  for the drug–target interaction. Herein the AS used were: halogens (X), unsaturated carbons ( $C_{\text{ins}}$ ), saturated carbons ( $C_{\text{sat}}$ ), heteroatom (Het), and hydrogen's bound to heteroatom's (H–Het). The corresponding symbols of the local spectral moments for these AS are:  $\pi_k(\text{X},s)$ ,  $\pi_k(C_{\text{ins}},s)$ ,  $\pi_k(C_{\text{sat}},s)$ ,  $\pi_k(\text{Het},s)$ ,  $\pi_k(\text{H–Het},s)$ .

## 2.2. Statistical analysis

As a continuation of the previous sections, we can attempt to develop a simple linear QSAR using the MARCH-INSIDE methodology, as defined previously, with the general formula:

$$\text{Actv} = b_0(\text{AS}) \cdot \pi_0(\text{AS}) + b_1(\text{AS}) \cdot \pi_1(\text{AS}) + b_2(\text{AS}) \cdot \pi_2(\text{AS}) + b_3(\text{AS}) \cdot \pi_3(\text{AS}) \dots + b_k(\text{AS}) \cdot \pi_k(\text{AS}) + b \quad (5)$$

Here,  $\pi_k(\text{AS})$  act as the microbial species specific molecule–target interaction descriptors. We selected Linear Discriminant Analysis (LDA) [28] to fit the classification functions. The model deals with the classification of a set of compounds as active or not against different microbial species. A dummy variable (Actv) was used to codify the antimicrobial activity. This variable indicates either the presence (Actv = 1) or absence (Actv = −1) of antimicrobial activity of the drug against the specific species. In Eq. (5),  $b_k(\text{AS})$  represents the coefficients of the classification function, determined by the least square method as implemented in the LDA module of the STATISTICA 6.0 software package. Forward stepwise was fixed as the strategy for variable selection [28–30]. The quality of LDA models was determined by examining Wilk's *U* statistic, Fisher ratio (*F*), and the *p*-level (*p*). We also inspected the percentage of good classification and the ratios between the cases and variables in the equation and variables to be explored in order to avoid over-fitting or chance correlation. Validation of the model was corroborated by re-substitution of cases in four predicting series [30]. We also inspected the specificity, sensitivity, and total accuracy to determine the quality of fit to data in training.

## 2.3. Data set

The data set includes a set of marketed and/or reported drug–bacteria pairs where affinity/non-affinity of drugs with the bacteria was established taking into consideration the  $\text{IC}_{50}$  values. The names or codes for all compounds as well as the references consulted to compile the data, which appear in this table, are depicted in as Tables of [Supplementary material](#) (Table 1SM and Table 2SM), available upon author's request due to space reasons.

## 2.4. Drug–Bacteria Pair (DBP) Complex Network (CN) construction

In order to achieve the drug–bacteria affinity with a network approach where one node represents a DBP and the edges show similarity between two nodes related to the activity (sDBPs or nDBPs), we carried out the following steps:

1. First, we calculated two types of activity Z-scores (drug score and bacteria score) for both experimental and QSAR-predicted values:

$$z_{\text{obs}}(d) = \frac{\log \text{MIC}_i}{\log \text{MIC}_{\text{max}}} \quad (6a)$$

$$z_{\text{pred}}(d) = p(-)_i \quad (6b)$$

where *d* is the score affinity, either observed score ( $s_{\text{obs}}$ ) or predicted score ( $s_{\text{pred}}$ ).  $s_{\text{obs}}$  were calculated on the experimental data ( $\text{IC}_{50}$ ). We calculated the  $s_{\text{pred}}$  of each one of the 1031 drugs with all the bacteria studied here by substituting the molecular descriptors into the QSAR equation using the Microsoft Excel application [31]. *Mean* is the average either of  $s_{\text{obs}}$  or  $s_{\text{pred}}$  for the DBP. We calculate the distance matrix between all DBP using a Euclidean distance:

$${}^{\text{obs}}d_{ij} = \frac{1}{\log \text{MIC}_{\text{max}}} \cdot |\log \text{MIC}_i - \log \text{MIC}_j| \quad (7a)$$

$${}^{\text{pred}}d_{ij} = |p(-)_i - p(-)_j| \quad (7b)$$

2. Using Microsoft excel [31] again we transformed the DBPs distance matrices derived into Boolean matrices. The elements of this type of matrix are equal to 1 if two DBPs have a Euclidean distance  $d_{ij} < a$  cut-off value. We explore the threshold values in a range from  $\log \text{MIC}_{\text{obs}}$  until  $\text{MIC}_{\text{pred}}$  trying to obtain average DBP node degree equal to 1 and minimizing the number of disconnected DBPs. The line command used in Excel to transform the distance matrix into a Boolean matrix was  $f = \text{SI}(\$N11 = \text{P\$9}, 0, \text{SI}(\text{ABS}(\$O11 - \text{P\$10}) < \text{O\$9}, \text{cut-off}, 1, 0))$ . It allows transforming distance into Boolean values and equals the main diagonal elements to 0 avoiding loops in the future network [24].
3. We compare the observed and predicted DBP–DBP networks pair-to-pair using a ROC curve test. Therefore, we used a contingency table where *a*, *b*, *c* and *d* are the observed frequencies in our networks. These frequencies were calculated as follows:  $f = \text{if}((\text{obs B2!} = 1, \text{pred B2!} = 1), 1, \text{if}((\text{obs B2!} = 1, \text{pred B2!} = 0), 2, \text{if}((\text{obs B2!} = 0, \text{pred B2!} = 1), 3, 4)))$ . Then:
  - “nn” is the number of DBP–DBP pairs unconnected in did not observe nor in predicted networks (the elements in the observed and predicted matrices are equal to 0).
  - “en” is the number of DBP–DBP pairs not connected in observed but connected in predicted networks (observed is 0 and predicted is 1).

**Table 1**  
Results of the model, analysis, validation and virtual screening.

Parameter	%	Group	Antibacterials	Non-active
<i>Training series</i>				
Sensitivity	83.82	Antibacterials	202	39
Specificity	84.50	Non-active	31	169
Accuracy	84.13	Total		
<i>External validation series</i>				
Sensitivity	86.78	Antibacterials	105	16
Specificity	92.0	Non-active	8	92
Accuracy	89.39	Total		
<i>Virtual screening</i>				
Sensitivity	86.67	Antibacterials	520	80

- “ne” is the number of DBP–DBP pairs connected in observed but not connected in predicted networks (observed is 1 and predicted is 0).
  - “ee” is the number of DBP–DBP pairs connected in observed and in predicted networks (observed and predicted is 1).
4. The Boolean matrix was saved as a text format file. After, renamed the .txt file as a .mat file we read it with the software CentiBin [32,33] and renamed the .mat file as a .net file and read it with Pajek software. Using Pajek we can not only represent the network but also highlight all DBP (nodes) connected to a specific DBP and calculating connectivity parameters [34].
  5. Pajek software was used to generate the observed network using a MIC data and generate the predicted network using the antibacterial model.

### 3. Results and discussion

Among the QSAR techniques applied in this field, molecular descriptors based on spectral moments are important because of their broad range of applicability. In general, moment analysis has been widely used in many other different contexts of solid, theoretical, and bioorganic chemistry. Estrada and Molina [35] carried out local structure–property studies and used a Markov approach, with total indices, to highlight interesting applications for moments in chemistry. Gutman used this type of technique to study the structure of benzenoids [36]. Molina et al. employed moments in the design of antibiotics [37]. Lastly, Vilar et al. reported interesting QSAR approaches to elucidate the action mechanism of anti-HIV drugs using spectral moments [38].

#### 3.1. Training and validation of the QSAR model

This work introduces for the first time a single linear mt-QSAR equation model to predict the activity of more than 90 antibacterial drugs against 90 bacteria species. The best model found was:

$$\begin{aligned} \text{Actv} = & -3.5\pi_1(C_{\text{sat}}) + 3\pi_0(C_{\text{sat}}) + 1.76\pi_2(C_{\text{uns}}) - 1.77\pi_3(\text{Het}) \\ & + 2.54\pi_5(\text{H–Het}) + 2.4\pi_3(\text{Het–Het}) \\ & - 5.42\pi_2(\text{H–Het}) + 0.74 \\ \lambda = & 0.49 \quad R_c = 0.718 \quad p < 0.001 \end{aligned}$$

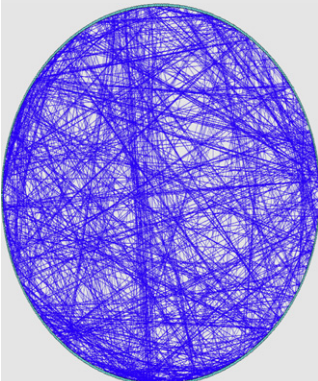
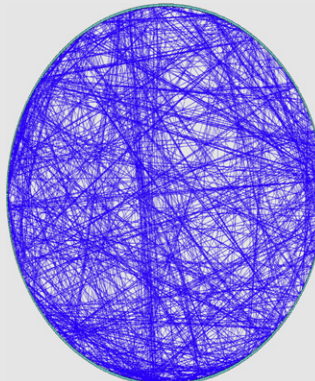
where  $\lambda$  is the Wilk's statistics, statistic for the overall discrimination,  $R_c$  is canonical index, and  $p$  the error level. In this equation,  $\pi_k(\text{AS})$  were calculated for the totality ( $T$ ) of the atoms in the molecule or for specific AS. The model correctly classifies 202 out of 241 DBPs of active compounds (83.82%) and 169 out of 200 DBPs of non-active cases (84.5%). Overall training predictability was 84.13% (371 out of 441 DBPs). The validation of the model was carried out by means of external predicting series and the model classified 197 out of 221, 89.4% of cases (DBPs). In order to show how the model works in practice, a virtual screening was carried out, recognizing the model as active (86.7%, 520 out of 600 cases); (DBPs) were not used in training or predicting series, see Table 1. The present work is an attempt to calculate within a unified framework probabilities of antibacterial action of drugs against many different bacterial species.

Two possible applications for the present model are the biomolecular screening of affinity of drugs with different receptors and the construction of multi-receptor affinity profile networks for drugs. In both cases, receptor susceptibility identification is imperative. This mt-QSAR offer a good opportunity to construct drug–drug similarity Complex Networks [39].

#### 3.2. Comparison of the observed vs. predicted DBP Complex Networks

Molecular CNs are being used to study large databases and/or complex systems [40–42]. Proteins, nucleic acids, and small molecules form a dense network of molecular interactions in a cell [43]. A DBP-CN may help us to detect the most similar cases (sDBPs) in the observed database and to study the complexity of drug–bacteria interaction [44]. We propose herein a different

**Table 2**  
Comparison of observed (MIC50) vs. predicted (QSAR) network of antibacterial drugs.

Observed network	Value	TIs <sup>a</sup>	Value	Predicted network
	1244	$n$	1244	
	772,736	$m$	641,377	
	2	Ad	2	
	1242	$z$	1031	
	154,7112	$W$	1,809,830	
	1.0005	$D$	1.1704	

<sup>a</sup> The TIs used are: number of nodes ( $n$ ), number of edges ( $m$ ), topological distance ( $D$ ), average node degree ( $z$ ), average distance (Ad), Wiener Index group ( $W$ ).



**Table 3**

Comparing sensitivity and specificity values of predicted vs. observed networks with different cut-off.

Cut-off MIC	Accuracy	Sensitivity	Specificity
<i>Cut-off prob = 0.4</i>			
0.0	45.4	79.1	42.5
0.1	61.0	78.5	51.4
0.2	69.4	76.4	61.1
0.3	73.5	73.8	72.8
0.4	71.2	69.3	78.0
0.5	65.4	63.9	78.0
0.6	62.2	61.2	80.8
0.7	59.9	59.7	85.0
0.8	59.5	59.4	88.3
0.9	59.2	59.2	86.9
<i>Cut-off prob = 0.5</i>			
0.0	33.9	90.6	29.0
0.1	55.5	89.7	36.9
0.2	68.3	87.9	45.4
0.3	75.9	85.3	55.4
0.4	76.6	81.2	59.5
0.5	74.4	76.4	58.4
0.6	74.0	74.4	64.6
0.7	73.0	73.0	73.5
0.8	72.8	72.8	80.8
0.9	72.6	72.6	77.5
<i>Cut-off prob = 0.6</i>			
0.0	24.7	98.6	18.4
0.1	50.9	97.9	25.2
0.2	67.0	96.3	32.6
0.3	77.8	94.3	41.7
0.4	81.1	90.7	45.7
0.5	81.3	86.2	42.8
0.6	82.8	84.6	48.5
0.7	83.2	83.4	60.1
0.8	83.1	83.2	72.0
0.9	83.0	83.0	73.9

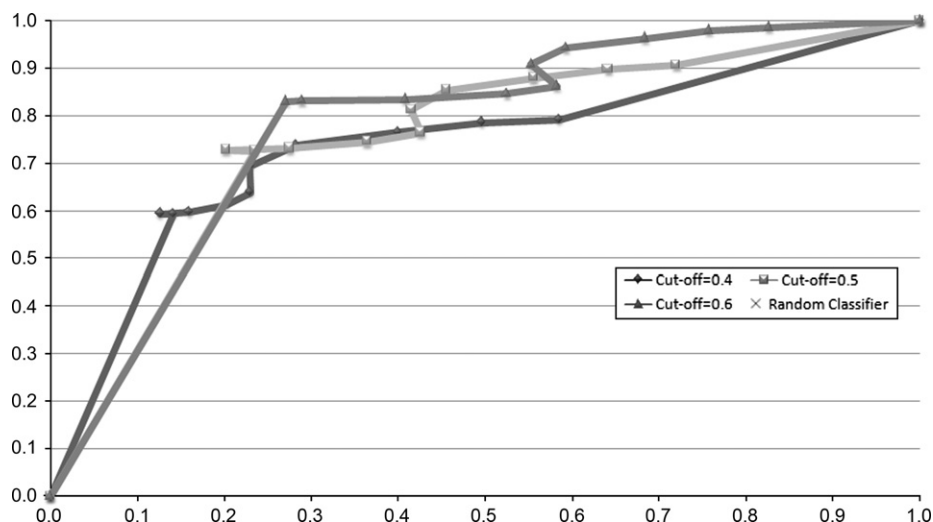
representation for the DBP-CN; it is not bipartite but considers only one type of node. As referred above, each node represents here both a drug and a bacteria (DBPs) [45]. The importance of the present study is that we cannot only construct the CN from experimental outcomes but also predict the CN for new DBPs not determined experimentally using the mt-QSAR. It allows us to add computationally new DBPs describing new potential drugs or targets that have not been described before [46]. In order to recall the capacity of the mt-QSAR to predict new CNs, we selected the

same data employed for the mt-QSAR equation. We first constructed a new observed DBP-CN considering the experimental MIC in the data and used a log of the MIC over log of MIC<sub>max</sub>; these values were used to obtain in the data some values in a range from 0 to 1, whose results are statistically more correct. Finally, we tested different cut-offs for the observed activity; cut-off = 0.94 was selected to obtain average sDBPs node degree equal to 1242 and 772,736 edges or sDBPs. Next, we used the mt-QSAR equation to predict the scores of biological activity for 1243 sDBPs including all the 96 bacteria species studied. As before, we explored the threshold values in a range from 0 to 1, obtaining average degree from 0 to 1243 respectively. We selected a different cut-off = 0.6; which leads to an average degree of 1031, and 641,377 edges or sDBPs see Table 2. In addition, with this threshold, the number of nodes was 1244 for the observed network and 1244 for the predicted network.

In order to compare the observed and predicted networks, we used an ROC curve analysis; the ROC curve can also be represented equivalently by plotting the fraction of true positives (sensitivity) vs. the fraction of false positives (1-specificity). As a result, we obtained three ROC curve analysis values using the observed network with the cut-off = (0.4; 0.5; 0.6) and predicted network with cut-off values in a range from 0 to 1, obtaining different LDA; in which “ee” was sensitivity and “nn” was specificity, see Table 3. The best match of the different analysis was obtained with the observed network cut-off = 0.6 and the predicted network cut-off = 0.94 is statistically more stable. Using these results, we constructed an ROC curve.

The results of ROC curve values (AUC) were as follows: for cut-off = 0.4 was 0.76; for cut-off = 0.5 was 0.82; finally cut-off = 0.6 was 0.84. It indicates that the present model gives statistically significant results and clearly different from those obtained with a random classifier (area = 0.5), see Fig. 1.

We can conclude that there is a statistically significant similarity between the observed and predicted networks as we saw in the ROC curve analysis where the AUC was considerably above average. In Table 2 we illustrate the contingency network with sDBPs for both CN (observed and predicted). It indicates that the predicted DBP-CN reproduces correctly the high specificity of the drugs studied against bacteria species. Finally, the DBP-CNs predicted with this procedure could become a useful tool to identify potential antibacterial drugs and/or susceptible bacteria species incorporating them to previous DT-CN [47].

**Fig. 1.**

#### 4. Conclusions

Using the MARCH-INSIDE approach, it is possible to seek for an mt-QSAR classifier to predict the probability of drugs with antibacterial activity of more than 90 different bacteria species. The model can be used as a tool for preliminary screening of drugs without relaying upon geometrical optimization of drug, receptor, and drug–receptor complex structure and avoiding receptor alignment as well. This mt-QSAR was also demonstrated to be an efficient tool for computational assembly of Drug–Bacterial Complex Networks that accurately reproduce the network based on experimental findings. This kind of Complex Networks could become a valuable approach to explore large drug–bacteria data of high complexity.

#### Acknowledgments

The authors sincerely thank the kind attention as well as useful comments from the editor Prof. A. Monge-Vega and three unknown referees. Prado-Prado acknowledges the financial support from the Xunta de Galicia for a one-year post-doctoral position (Research Project IN89A 2008/75-0). González-Díaz acknowledges the financial support of the Programme Isidro Parga Pondal and one-year post-doctoral position (Research Project IN89A 2008/117-0), both funded by the Xunta de Galicia and the European Research funds from European Social Fund (ESF).

#### Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.ejmech.2009.06.018.

#### References

- [1] A.C. Rodloff, R. Leclercq, E.A. Debbia, R. Canton, B.A. Oppenheim, M.J. Dowzicky, Clin. Microbiol. Infect. 14 (2008) 307–314.
- [2] R. Ergova, E. K'Oleian, I. Kharalambieva, I. Mitov, Iu Docheva, Vutr. Boles. 32 (2000) 13–17.
- [3] T. Lorand, B. Kocsis, P. Sohar, G. Nagy, P. Jozsef, G. Kispal, et al., Eur. J. Med. Chem. 37 (2002) 803–812.
- [4] K. Roy, A.S. Mandal, J. Enz. Inhib. Med. Chem. 24 (2009) 205–223.
- [5] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 13 (2005) 2967–2973.
- [6] K. Roy, J.T. Leonard, J. Chem. Inf. Model. 45 (2005) 1352–1368.
- [7] S. Mukherjee, A. Saha, K. Roy, Bioorg. Med. Chem. Lett. 15 (2005) 957–961.
- [8] K. Roy, J.T. Leonard, C. Sengupta, Bioorg. Med. Chem. Lett. 14 (2004) 3705–3709.
- [9] K. Roy, J.T. Leonard, Bioorg. Med. Chem. 12 (2004) 745–754.
- [10] Y. Marrero-Ponce, R.M. Marrero, F. Torrens, Y. Martinez, M.G. Bernal, V.R. Zaldivar, et al., J. Mol. Model. 12 (2006) 255–271.
- [11] M. Murcia-Soler, F. Perez-Gimenez, F.J. Garcia-March, M.T. Salabert-Salvador, W. Diaz-Villanueva, P. Medina-Casamayor, J. Mol. Graph. Model. 21 (2003) 375–390.
- [12] M.T. Cronin, A.O. Aptula, J.C. Dearden, J.C. Duffy, T.I. Netzeva, H. Patel, et al., J. Chem. Inf. Comput. Sci. 42 (2002) 869–878.
- [13] Y. Marrero-Ponce, R. Medina-Marrero, F. Torrens, Y. Martinez, V. Romero-Zaldivar, E.A. Castro, Bioorg. Med. Chem. 13 (2005) 2881–2899.
- [14] F.J. Prado-Prado, H. González-Díaz, O. Martinez de la Vega, F.M. Ubeira, K.C. Chou, Bioorg. Med. Chem. 16 (2008) 5871–5880.
- [15] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, 2002.
- [16] H. Gonzalez-Díaz, F. Prado-Prado, F.M. Ubeira, Curr. Top. Med. Chem. 8 (2008) 1676–1690.
- [17] H. González-Díaz, Y. González-Díaz, L. Santana, F.M. Ubeira, E. Uriarte, Proteomics 8 (2008) 750–778.
- [18] H. González-Díaz, S. Vilar, L. Santana, E. Uriarte, Curr. Top. Med. Chem. 7 (2007) 1025–1039.
- [19] D. Erhan, J. L'Heureux, S.Y. Yue, Y. Bengio, J. Chem. Inf. Model. 46 (2006) 626–635.
- [20] D. Damos, T. Smist, Aviat. Space Environ. Med. 53 (1982) 1177–1181.
- [21] D. Maslovat, R. Chus, T.D. Lee, I.M. Franks, Motor Control 8 (2004) 213–233.
- [22] M. Cruz-Monteagudo, H. González-Díaz, G. Agüero-Chapin, L. Santana, F. Borges, R.E. Domínguez, et al., J. Comput. Chem. 28 (2007) 1909–1922.
- [23] H. González-Díaz, F.J. Prado-Prado, L. Santana, E. Uriarte, Bioorg. Med. Chem. 14 (2006) 5973–5980.
- [24] H. González-Díaz, F. Prado-Prado, J. Comput. Chem. 29 (2008) 656–657.
- [25] F.J. Prado-Prado, O.M. de la Vega, E. Uriarte, F.M. Ubeira, K.C. Chou, H. Gonzalez-Díaz, Bioorg. Med. Chem. 17 (2009) 569–575.
- [26] F. Prado-Prado, H. González-Díaz, L. Santana, E. Uriarte, Bioorg. Med. Chem. 15 (2007) 897–902.
- [27] H. González-Díaz, M. Cruz-Monteagudo, D. Vina, L. Santana, E. Uriarte, E. De Clercq, Bioorg. Med. Chem. Lett. 15 (2005) 1651–1657.
- [28] H. González-Díaz, I.H. Sanchez, E. Uriarte, L. Santana, Comput. Biol. Chem. 27 (2003) 217–227.
- [29] R. Ramos de Armas, H. González-Díaz, R. Molina, E. Uriarte, Biopolymers 77 (2005) 247–256.
- [30] StatSoft, Inc., STATISTICA (data analysis software system), version 6.0, [www.statsoft.com](http://www.statsoft.com), Statsoft, Inc. 6.0 ed, 2002.
- [31] Microsoft Excel, Microsoft Corporation, 2002.
- [32] Dirk Koschützki, CentiBiN Version 1.4.2, Centralities in Biological Networks © 2004–2006 Dirk Koschützki Research Group Network Analysis, IPK Gate-sleben, Germany, 2006.
- [33] B.H. Junker, D. Koschuetzki, F. Schreiber, BMC Bioinformatics 7 (2006) 219.
- [34] V. Batagelj, A. Mrvar, Pajek 1.15 ed., 2006.
- [35] E. Estrada, E. Molina, J. Mol. Graph. Model. 20 (2001) 54–64.
- [36] S. Markovic, Z. Markovic, R.I. McCrindle, J. Chem. Inf. Comput. Sci. 41 (2001) 112–119.
- [37] H. Gonzalez-Díaz, R.R. de Armas, R. Molina, Bull. Math. Biol. 65 (2003) 991–1002.
- [38] S. Vilar, L. Santana, E. Uriarte, J. Med. Chem. 49 (2006) 1118–1124.
- [39] H. González-Díaz, L. Saiz-Urra, R. Molina, L. Santana, E. Uriarte, J. Proteome Res. 6 (2007) 904–908.
- [40] D. Bonchev, G.A. Buck, J. Chem. Inf. Model. 47 (2007) 909–917.
- [41] D. Bonchev, SAR QSAR Environ. Res. 14 (2003) 199–214.
- [42] J. Park, A.L. Barabasi, Proc. Natl. Acad. Sci. USA 104 (2007) 17916–17920.
- [43] V. Spirin, L.A. Mirny, Proc. Natl. Acad. Sci. USA 100 (2003) 12123–12128.
- [44] M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, M. Vidal, Nat. Biotechnol. 25 (2007) 1119–1126.
- [45] E. Estrada, N. Guevara, I. Gutman, J. Chem. Inf. Comput. Sci. 38 (1998) 428–431.
- [46] L. Han, J. Cui, H. Lin, Z. Ji, Z. Cao, Y. Li, et al., Proteomics 6 (2006) 4023–4037.
- [47] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabasi, Proc. Natl. Acad. Sci. USA 104 (2007) 8685–8690.