Bioorganic & Medicinal Chemistry 17 (2009) 569–575

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug–drug complex networks

Francisco J. Prado-Prado [a,b,c], Octavio Martinez de la Vega [b], Eugenio Uriarte [c], Florencio M. Ubeira [a], Kuo-Chen Chou [d], Humberto González-Díaz [a,c,d,*]

[a] Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela (USC), Santiago de Compostela 15782, Spain
[b] Department of Bioinformatics, CINVESTAV, LANGEBIO, Irapuato 629 36500, Mexico
[c] Unit of Bioinformatics and Connectivity Analysis (UBICA), Institute of Industrial Pharmacy, Department of Organic Chemistry, Faculty of Pharmacy, USC, Santiago de Compostela 15782, Spain
[d] Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Received 27 June 2008
Revised 24 November 2008
Accepted 28 November 2008
Available online 6 December 2008

Keywords:
QSAR
Multi-target learning
Machine learning
Complex Networks
Antimicrobial drugs
Antiviral drugs
Markov Chain Model
Giant Component

ABSTRACT

One limitation of almost all antiviral Quantitative Structure–Activity Relationships (QSAR) models is that they predict the biological activity of drugs against only one species of virus. Consequently, the development of multi-tasking QSAR models (mt-QSAR) to predict drugs activity against different species of virus is of the major vitally important. These mt-QSARs offer also a good opportunity to construct drug–drug Complex Networks (CNs) that can be used to explore large and complex drug-viral species databases. It is known that in very large CNs we can use the Giant Component (GC) as a representative sub-set of nodes (drugs) and but the drug–drug similarity function selected may strongly determines the final network obtained. In the three previous works of the present series we reported mt-QSAR models to predict the antimicrobial activity against different fungi [Gonzalez-Diaz, H.; Prado-Prado, F. J.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2006**, *14*, 5973], bacteria [Prado-Prado, F. J.; Gonzalez-Diaz, H.; Santana, L.; Uriarte E. *Bioorg. Med. Chem.* **2007**, *15*, 897] or parasite species [Prado-Prado, F.J.; González-Díaz, H.; Martinez de la Vega, O.; Ubeira, F.M.; Chou K.C. *Bioorg. Med. Chem.* **2008**, *16*, 5871]. However, including these works, we do not found any report of mt-QSAR models for antivirals drug, or a comparative study of the different GC extracted from drug–drug CNs based on different similarity functions. In this work, we used Linear Discriminant Analysis (LDA) to fit a mt-QSAR model that classify 600 drugs as active or non-active against the 41 different tested species of virus. The model correctly classifies 143 of 169 active compounds (specificity = 84.62%) and 119 of 139 non-active compounds (sensitivity = 85.61%) and presents overall training accuracy of 85.1% (262 of 308 cases). Validation of the model was carried out by means of external predicting series, classifying the model 466 of 514, 90.7% of compounds. In order to illustrate the performance of the model in practice, we develop a virtual screening recognizing the model as active 92.7%, 102 of 110 antivirus compounds. These compounds were never use in training or predicting series. Next, we obtained and compared the topology of the CNs and their respective GCs based on Euclidean, Manhattan, Chebychey, Pearson and other similarity measures. The GC of the Manhattan network showed the more interesting features for drug–drug similarity search. We also give the procedure for the construction of Back-Projection Maps for the contribution of each drug sub-structure to the antiviral activity against different species.

1. Introduction

Examples of diseases caused by viruses include the common cold (produced by any one of a variety of related viruses), AIDS (caused by HIV), and cold sores (caused by herpes simplex); which produced some of the major health problems in the last 30 years.[1] Other relationships are being studied such as the connection of Human Herpesvirus 6 (HHV6), one of the eight known members of the human herpesvirus family, with organic neurological diseases such as multiple sclerosis and chronic fatigue syndrome. Recently, it has been shown that cervical cancer is caused, at least partially, by papillomavirus, representing the first significant evidence in humans for a link between cancer and an infective agent.[1–3] The rel-

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.
E-mail address: gonzalezdiazh@yahoo.es (H. González-Díaz).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.11.075

ative ability of viruses to cause disease is described in terms of virulence.

There is a high interest on the search of rational approaches for antiviral drugs discovery. In this sense, Quantitative Structure–Activity Relationships (QSAR) studies may play an important role. Disappointingly, QSAR studies are generally based on databases considering only structurally parent compounds acting against one single microbial species or target. We call these classic models the one-target QSAR approach (ot-QSAR). As a consequence, to predict the antiviral activity for a given series of compounds one have to use/seek as many QSAR models as microbial species drugs susceptibility is desirable to predict or rely upon the use of artificial neural networks (ANNs) with multiple outputs. For instance, Vilar et al.[4] reported an ANN-QSAR model able to predict up to four different mechanisms of actions for HIV inhibitors. The relative complexity of theses models derives mainly on the high number of output variables we have to use in the QSAR model training. In this sense, it is very important the report of one single multi-target QSAR equation (mt-QSAR) to calculate the probability of activity of a given drug against different antiviral species.

Up at the moment there are near to 1600 molecular descriptors that may be in principle generalized and used to solve the former problem. Many of these indices are known as topological Indices (TIs) or Connectivity Indices (CIs) or simply invariants of a molecular graph, whose vertices are atoms weighted with physicochemical properties (mass, polarity, electronegativity, or charge).[5] Multi-tasking QSAR (mt-QSAR)[6] 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).[7,8] 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. The mt-QSAR models may be very useful to optimize important aspects such as activity, toxicity or pharmacokinetics using one single model. The first way to develop mt-QSAR models is the consideration of one output variable for each activity or drug property we pretend to predict. This alternative usually leads to linear models based on the fit of one function by each one of the drug properties or non-linear models able to fit several outputs at time. These mt-QSARs offer an unprecedented opportunity to construct Back-Projection Maps (BPMs)[9] and drug–drug similarity Complex Networks (CNs). Back-Projection[10,11] is very useful to map or project the predicted function backwards onto structure and determine the contribution of sub-structural molecular regions to the desired property.

2. Methods

2.1. Molecular descriptors

In this work we are going to focus in a QSAR method introduced elsewhere that uses a Markov Chain Model (MCM) to encode systems structural information using molecular, macro-molecular, supra-molecular CIs and 3D parameters as well. The method named as the MARCH-INSIDE: MARkov CHains INvariants for SImulation & Design; Recently, we have reported two reviews with applications of this and other QSAR methods.[12,13] Both revisions made in depth review on the several applications in Chemistry and Biomedical Sciences of CIs and include several references to MARCH-INSIDE.

We used as input for the mt-QSAR analysis the CIs type molecular descriptors ${}^{k}C_s$(Set). These CIs can be interpreted as the average contributions ${}^{k}C_s$(Set) of a group of atoms (Set) in the molecule to the gradual step-by-step interaction (k) between the drug and the receptor with unknown structure for a given virus species (s). Figure 1 graphically illustrates the idea of this gradual interaction. A group of atoms may enclose the whole molecule (T), only halogens (X), heteroatoms (Het), heteroatom-bound hydrogen atoms (H-Het), Sp_3 carbon atoms (C_{sat}) or others. We derive these ${}^{k}C_s$ by summing up all the atomic contributions of each atom to the interaction ${}^{0}c_j(s)$ pre-multiplied by the probability of the distribution of the atom in the molecule ${}^{A}p_k(j,s)$. The ${}^{0}c_j(s)$ values depend both on the atom and the different virus species while the ${}^{A}p_k(j,s)$ depends also on molecular topology or connectivity. The detailed calculation of these values using a MCM have been explained in the previous work of this series and others so we omit it for the sake of simplicity (see the three previous works of this series).[14,15] These CIs were calculated with our software MARCH-INSIDE:[12,16]

$$ {}^{k}C_s = \sum_{j=1}^{n} {}^{A}p_k(j,s) \cdot {}^{0}c_j(s) \tag{1} $$

2.2. Statistical analysis

As a continuation of the previous sections, we can attempt to develop a simple linear mt-QSAR with the general formula:

$$ Act\,v = b_{0,G} \cdot {}^{0}C_s(G) + b_{1,G} \cdot {}^{1}C_s(G) + b_{2,G} \cdot {}^{2}C_s(G) + b_{3,G} \cdot {}^{3}C_s(G) \cdots + b_{k,G} \cdot {}^{k}C_s + b \tag{2} $$

Here, ${}^{k}C_s(G)$ act as the species specific molecular descriptors of TI/CI type for antiviral compounds. We selected Linear Discriminant Analysis (LDA)[17,18] 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 antiviral activity of the drug against the specific species. In Eq. (2), $b_{k,G}$ 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.[19] Forward stepwise was fixed as the strategy for variable selection. 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 antiviral (specificity), non-antiviral (sensitivity), and total cases (accuracy) correctly classified by the model 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 (interchange of training and validation cases) in order to construct four predicting series.[20–23]

2.3. Data set

The data set was conformed by a set of marketed and/or very recently reported antiviral drugs which low reported $MIC_{50} < 10$ μM against different virus. The three data sets used were as follows training series: 143 active compounds plus 119 non-active compounds (262 in total); predicting series: 206 + 260 = 466 in total; virtual screening 102 active compounds. The literature reports experimental test of each drug against some but not all species of a list of 40. In consequence, we were able to collect 843 cases (drug/species pairs). The names or codes for all compounds as well as the references consulted are depicting in Table 1SM of the Supplementary data by reasons of space. The complete list of references used to collect the database is given at the end of the online Supplementary data too.

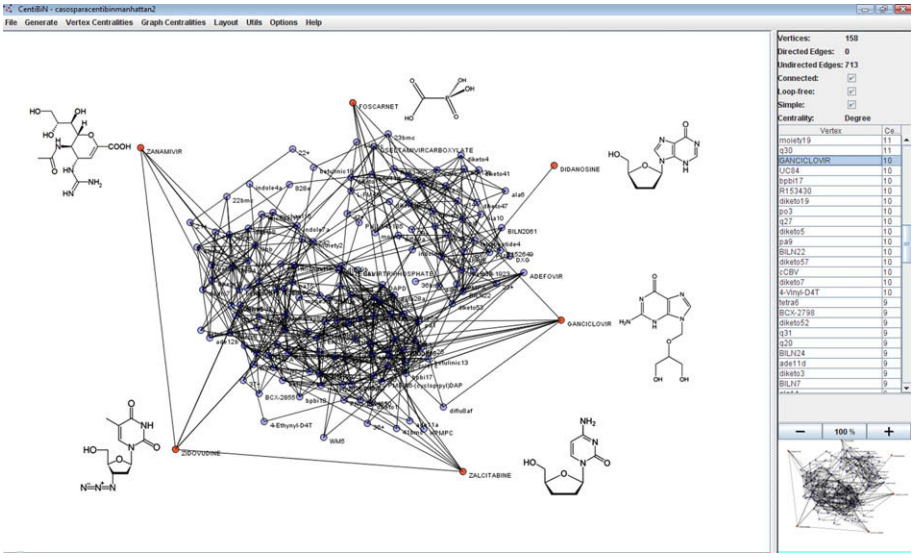


Figure 1. Visualization of GC(Manhattan) and some drugs (nodes) with the graphical interface of CentiBin.

2.4. Construction of Complex Networks

In order to approach the multi-specie activity of antiviral drugs with a CNs approach we calculated the molecular descriptors include in the QSAR equation for 600 selected drugs and 40 virus using the MARCH-INSIDE software.[24] We calculated the scores of biological activity of each one of the 600 drugs against all the virus species studied here by substituting the molecular descriptors into the QSAR equation using the Microsoft Excel application.[25] All the activity scores predicted were organized into a table of drugs (rows) versus species (columns), which was used as input for the software STATISTICA[19] employed to calculate drug–drug multispecies correlations in diverse coefficients. These correlations were represented actually as distances (1 − pearson r distance, Chebychey distance metric, Euclidean distance, City-block (Manhattan) distance, Percent disagreement distance, Power distance) between drugs pairs. All the obtained distances matrices were derived using the software package STATISTICA.[19] Using Microsoft excel[25] again we transformed the drug pair distances matrix derived with Statistic into a Boolean matrix. The elements of this matrix are equal to 1 if two drugs have a high correlation or the same are very close (short distance (1 − pearson r distance, Chebychey distance metric, Euclidean distance, City-block (Manhattan) distance, Percent disagreement distance, Power distance)). The following line command was used in Excel to transform the Distance matrix into a Boolean matrix was f = if ($A2 = B$1, 0, if (B2 > A1, 0, 1)). In this line command, $A2 is the relative reference to Distance matrix row labels; B$1 is the relative reference to Distance matrix column labels; B2 refers to the top-left cell of the Distance matrix and $A$is a total reference to the cut-off cell. The command runs over all cells of the Distance matrix (inputs) and gives a result the 1 or 0 values for the Bollean matrix. It allows transforming distance into Boolean values and equals the main diagonal elements to 0 avoiding loops in the future network. The Boolean matrix was saved as a txt format file. After, renamed the .txt file as a .mat file we read it with the software CentiBin[26] and save the file as a .net to read in Pajek Software.[27] Both CentiBin and Pajek software were used to calculating different TIs and/or CIs that describe numerically network topology such as number of nodes (n), number of edges (m), topological distance (D), average node degree (z), Zagreb group index 1 ($M1$), Zagreb group index 2 ($M2$), Randic connectivity index (X_r), Platt index (F), Index of relinking (P). The parameter n coincides with the number of drugs and m is the number of similar drug–drug pairs.

2.5. Multi-species BPMs analysis

In order to illustrate the procedure to derive antivirals multi-specie activity with a back-projection approach we selected at random three drugs and three virus species (HSV-1, HIV-1, VZV) and carried out the following steps. These steps are the same reported for a classic QSAR repeated here for each different virus species using species-dependent atomic contributions.

First, we calculated the species-dependent atomic descriptors included in the QSAR equation for three selected drugs using the MARCH-INSIDE software. We calculated the scores of contributions of each one of atoms of the three drugs against three species studied by substituting the atomic descriptors into the QSAR equation using the Microsoft Excel application. The contributions of each atom of the drug were scaled into a percentage value. The scaled contributions of atoms were grouped for different molecular fragments. These molecular fragment contributions to the biological activity were back-projected onto the molecular structure to obtain a color scaled biological structure–activity map.

3. Results and discussion

3.1. Multi-species mt-QSAR model for antiviral drugs

There are advantages of the present stochastic approach; an important one is the possibility of deriving average thermodynamic parameters depending on the probability of the states of the MM. The generalized parameters fit on more clearly physicochemical sense with respect to our previous ones. In specific, this work introduces by the first time a single linear QSAR equation model to predict the antiviral activity of drugs against different virus species. Summary for the forward-stepwise analysis shows the variables that entire first in the model (Table 1). The final model selected was

$$\begin{aligned} act\,v = {} & 1.90 \cdot {}^{0}C_S(C_{sat}) - 1.64 \cdot {}^{0}C_S(C_{uns}) + 1.02 \cdot {}^{2}C_S(C_{uns}) \\ & + 1.10 \cdot {}^{5}C_S(C_{uns}) + 0.73 \cdot {}^{1}C_S(X) + 1.08 \cdot {}^{1}C_S(\text{Het}) \\ & + 1.07 \cdot {}^{0}C_S(\text{H-Het}) - 0.75 \cdot {}^{4}C_s(\text{H-Het}) + 0.08 \\ \lambda = {} & 0.47 \quad Rc = 0.726 \quad p < 0.001 \end{aligned} \tag{3}$$

where λ is the Wilk's statistics, statistic for the overall discrimination, F is the Fisher ratio, and p the error level. In this equation,

Table 1
Summary for the forward-stepwise analysis and correlation matrix

Variable	*F*	*p*	Effect	Variable	*F*	*p*	Effect
$^{0}C_s$(Het)	5.0	0.02	Out	$^{5}C_s$(C_{ins})	23.18	0.00	In
$^{0}C_s$(C_{sat})	104.8	0.00	In	$^{0}C_s$(X)	10.21	0.00	Out
$^{3}C_s$(X)	0.3	0.55	Out	$^{1}C_s$(X)	23.03	0.00	In
$^{1}C_s$(T)	1.1	0.27	Out	$^{2}C_s$(X)	14.77	0.00	Out
$^{0}C_s$(C_{ins})	71.4	0.00	In	$^{4}C_s$(X)	9.91	0.00	Out
$^{5}C_s$(C_{sat})	0.2	0.62	Out	$^{5}C_s$(X)	1.98	0.15	Out
$^{2}C_s$(X)	14.7	0.00	Out	$^{0}C_s$(T)	0.06	0.80	Out
$^{3}C_s$(T)	8.4	0.00	Out	$^{1}C_s$(Het)	16.05	0.00	In
$^{1}C_s$(C_{sat})	0.8	0.35	Out	$^{2}C_s$(Het)	2.14	0.14	Out
$^{2}C_s$(C_{sat})	0.9	0.32	Out	$^{3}C_s$(Het)	0.14	0.70	Out
$^{3}C_s$(C_{sat})	3.2	0.07	Out	$^{4}C_s$(Het)	1.09	0.29	Out
$^{4}C_s$(C_{sat})	5.2	0.02	Out	$^{5}C_s$(Het)	1.20	0.27	Out
$^{5}C_s$(T)	0.1	0.70	Out	$^{0}C_s$(H-Het)	5.00	0.02	Out
$^{4}C_s$(T)	0.4	0.52	Out	$^{1}C_s$(H-Het)	7.09	0.00	Out
$^{1}C_s$(C_{ins})	0.2	0.63	Out	$^{2}C_s$(H-Het)	1.73	0.18	Out
$^{2}C_s$(C_{ins})	19.9	0.00	In	$^{3}C_s$(H-Het)	0.00	0.97	Out
$^{3}C_s$(C_{ins})	1.3	0.24	Out	$^{4}C_s$ (H-Het)	17.89	0.001	Entered
$^{4}C_s$(C_{ins})	1.7	0.18	Out	$^{5}C_s$(H-Het)	0.09	0.75	Out

Variable	Symbol	1	2	3	4	5	6	7	8
1	$^{0}C_s$(C_{sat})		−0.05	−0.1	−0.14	0.06	−0.01	−0.01	0.13
2	$^{0}C_s$(C_{ins})			0.21	0.20	0.14	−0.06	−0.20	−0.04
3	$^{2}C_s$(C_{ins})				0.6	0.03	0.21	−0.20	−0.07
4	$^{5}C_s$(C_{ins})					−0.01	0.14	−0.14	−0.05
5	$^{1}C_s$(X)						0.02	−0.07	−0.05
6	$^{1}C_s$(Het)							0.33	0.23
7	$^{0}C_s$(H-Het)								−0.02
8	$^{4}C_s$(H-Het)								

Table 2
Results of the model, training, validation and virtual screening

Parameter	%	Cases[a]	antiviral	Non-active
Original training series (T1)				
Specificity	81	Antiviral	**188**	44
Sensitivity	86.1	Non-active	39	**242**
Accuracy	83.8	Total		
First interchanged cross-validation series (CV1)				
Specificity	81.52	Antiviral	**150**	34
Sensitivity	86.28	Non-active	31	**195**
Accuracy	83.91	Total		
Second interchanged cross-validation series (CV2)				
Specificity	79.71	Antiviral	**110**	28
Sensitivity	86.98	Non-active	22	**147**
Accuracy	83.35	Total		
Third interchanged cross-validation series (CV3)				
Specificity	79.71	Antiviral	**110**	28
Sensitivity	80.59	Non-active	33	**137**
Accuracy	80.15	Total		
Four interchanged cross-validation series (CV4)				
Specificity	81.52	Antiviral	**150**	34
Sensitivity	86.28	Non-active	31	**195**
Accuracy	84.91	Total		
Average of interchanged cross-validation series				
Specificity	80.71	Antiviral	**130**	31
Sensitivity	85	Non-active	29.5	**168.5**
Accuracy	82.85	Total		
Virtual-screening series				
Specificity	Antiviral	93	**102**	8

[a] The number cases correctly classified by the model appear in boldface style.

$^{k}C_s$ where calculated for the totality (*T*) of the atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic, for example, halogens (X) or unsaturated carbon atoms (C) or heteroatom-bound hydrogen atoms (H-Het). We calculated the correlation matrix in order to check possible co-linearity between these variables. We found that the correlation between variables of the model was not significant, see Table 1.

The model correctly classifies 143 of 169 active antiviral compounds (specificity = 84.62%) and 119 of 139 non-active compounds (sensitivity = 85.61%). Overall training Accuracy was 85.1% (262 of 308 cases) see Table 2. On the other hand, validation of the model was carrying out by means of external predicting series, classifying the model 466 of 514 compounds (cv-accuracy = 90.7%). In order to illustrate the performance of the model in practice, we developed a virtual screening recognizing the model as active 92.7%, 102 of 110 antiviral compounds not used in training or predicting series. We performed a four-folded interchange of cases between training and cross-validation series (it means that we used four different training and cv series) to demonstrate the stability of the model. The results obtained are depicted in Table 2. The present is an attempt to calculate within a unify framework probabilities of antiviral action of drugs against many different species.

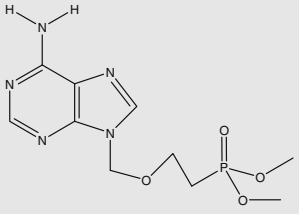
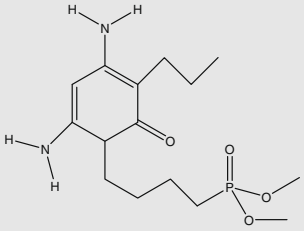
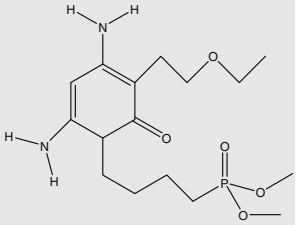
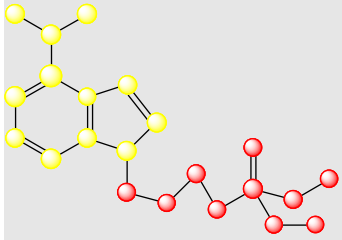
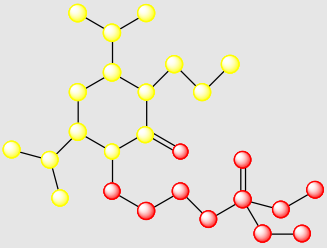
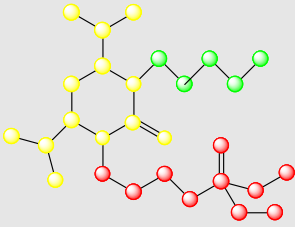
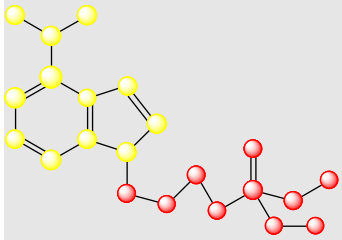
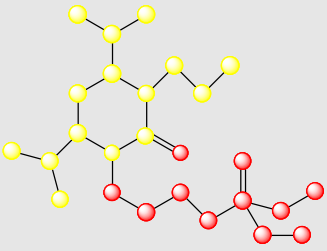
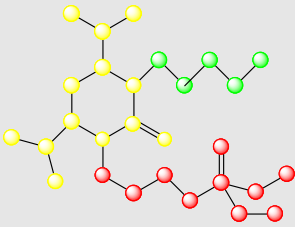
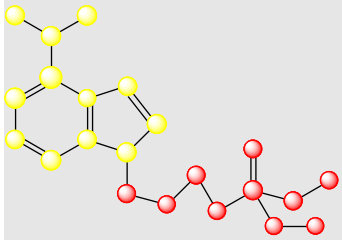
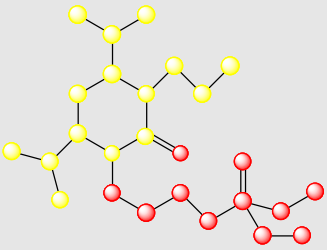
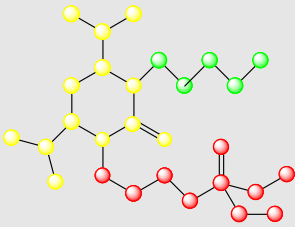
In the common case of ot-QSAR the atomic contributions to biological activity referenced above depend only on physicochemical atomic parameters such as atomic mass, polarizability, and charge,[28] or electronegativity and/or chirality.[17] The characteristic most remarkable of the present model is that the $^{k}C_s(G)$ parameters used as molecular descriptors depend not only on the molecular structure of the drug but interestingly they also depends on the species of virus we have to control with the drug. The named of all the drugs used, the tested virus species, and the results for training and validation by means of external predicting series are depicted in Table 1SM (see online Supplementary data). Last, in order to show how good the model function in practice a virtual screening was carried out recognizing the model as active 92.7%, 102 of 110 antiviral compounds not used in training or predicting series (see Table 2SM).

3.2. Multi-species back-projection analysis of antiviral drugs

In the previous works of this series we applied the above mentioned Input-Coded mt-QSAR philosophy to the MARCH-INSIDE method in order to extended it to predict antibacterial[15], antifungal[14] or antiprotozoal drugs[29] active against different species. In these works we have used CIs based on MCM that resemble Free Energy parameters. We also extended the study to the mt-QSAR prediction of antifungal activity using Absolute Probabilities; another type of CIs based on MCM.[30] These mt-QSARs offer an unprecedented opportunity to construct Back-Projection Maps (BPMs).[9,31] Using BPMs[10,11] is a very useful method to map or project the predicted function backwards onto structure and determine the contribution of sub-structural molecular regions to the desired property.[32,33]

As illustrative example we of construction of BPMs for antiviral drugs we selected at random three compounds (PMEA, 11a and 11d) and three virus species (Herpes virus simplex 1 (HVS-1), HIV-1 and Varicella zoster virus (VZV)) whose results in our model are as follows: the rings structures have differences in the activity (see Table 3) and concurrence with the literature. The model predicted a between 34% and 66% ring interaction of PMEA in the three species that indicated the major interaction with the receptor, because the ability of the virus to absorb the compound. The second one, the structure called 11A, the BPMs predicted a between 34% and 66% interaction drug–receptor in every ring of pentamidine in all the species tested; which indicates the active part of the drug is the aromatic ring. The 11D drug have an active part, the BPMs predicted between 67% and 99% interaction in every linear cetona in all virus species. This result illustrates the advantages of extending from ot-QSAR to mt-QSAR the BPMs study of the contribution of atoms, bonds, and sub-structures in general to the biological activity.

Table 3
BPMs of each molecule

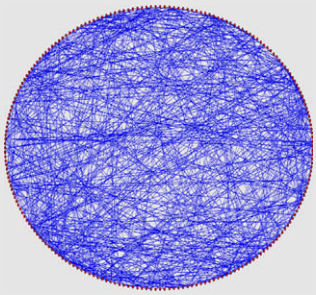
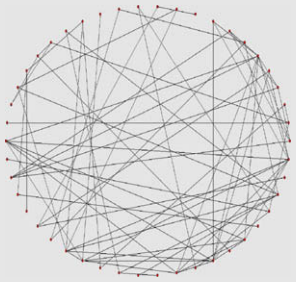
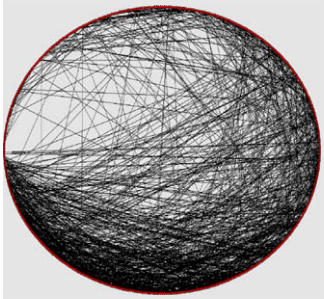
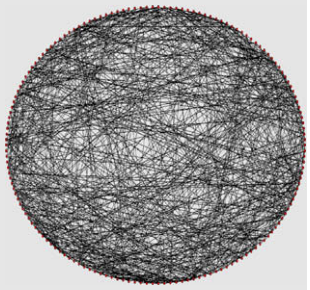
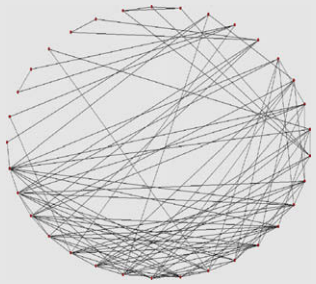
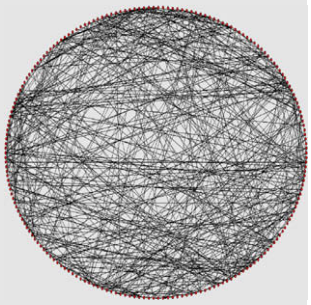
Compounds		
PMEA	11A	11D
		
BPMs[a] for strain[b] HVS-1		
		
BPMs[a] for strain[b] HIV-1		
		
BPMs[a] for strain[b] VZ		
		

[a] Red: 0–33% interaction. Yellow: 34–67% interaction. Green: 68–100% interaction.
[b] HVS is Herpes Virus Simplex, HIV-1 is Human Immunodeficiency Virus type 1; and VZ is Varicella Zoster Virus.

3.3. Comparative study Complex Networks of antiviral drugs

Haggarty, Clemons, and Schreiber published a very interesting works related to the chemical profiling of different fungi strains to construct CNs. In this case, the network derived based on only 23 drugs and 10 strains of one single species *Saccharomyces cerevisiae* (budding yeast). In any case, the method reported uses experimental values and is unable to add new drugs or strains to the network at least you measured it. In the previous work of this series we constructed a similar type of CNs for antiprotozoal compounds[29] and species of parasites based on an mt-QSAR. It allowed us to add growth the CN using the mt-QSAR to predict new nodes without to measure the results experimentally. In addition, we have reported very recently an mt-QSAR and CN for antifungal compounds.[34] In both cases the CNs used based on Euclidean distance but it is well known that different distance measure can be used to construct CNs obtaining different results.[35] Taking into consideration that we do not have an a priori reason to select one particular class of distance the study of the effect of distance selection in drug–drug ms-CNs becomes mandatory. The present work present the first comparative study of different distances to construct drug–drug ms-CNs for antiviral compounds using one mt-QSAR model. In Table 4 we generated six drug–drug networks using the following distances: 1 − (Pearson r) distance, Chebychey distance metric,

Table 4
Comparative study of the Giant Components for networks based on different distances

GC(1 − Pearson *r*)	Value	TIs[a]	Value	GC(Chebychey)
	211	*n*	46	
	629	*m*	152	
	18	*D*	12	
	6	*z*	5	
	2.84	*C*	14.7	
	11100	*M*1	2976	
	55669	*M*2	16320	
	99.1	X_r	21.36	
	9842	*F*	2672	
	1.98	*P*	2.28	
GC(Euclidean)	**Value**	**TIs[a]**	**Value**	**GC(Manhattan)**
	499	*n*	158	
	2452	*m*	713	
	20	*D*	**7***	
	8	*z*	**3***	
	1.97	*C*	5.75	
	75516	*M*1	16208	
	674428	*M*2	96420	
	231.3	X_r	74.73	
	70612	*F*	14728	
	3.86	*P*	3.5	
GC(Disagreement)	**Value**	**TIs[a]**	**Value**	**GC(Power)**
	33	*n*	198	
	102	*m*	476	
	5	*D*	20	
	2	*z*	8	
	19.32	*C*	2.44	
	1556	*M*1	6588	
	6173	*M*2	26155	
	15.9	X_r	92.6	
	1352	*F*	5636	
	2.06	*P*	1.4	

[a] The TIs used are number of nodes (*n*), number of edges (*m*), topological distance (*D*), average node degree (*z*), Zagreb group index 1 (*M*1), Zagreb group index 2 (*M*2), Randic connectivity index (X_r), Platt index (*F*), Index of relinking (*P*).

Euclidean distance, City-block (Manhattan) distance, Percent disagreement distance and Power distance.

In this work, we constructed the CNs also to explore globally and locally the similarity interrelationships between all drugs active against different species of virus. We this aim we focus the theoretical study on an special type of sub-network called the Giant Component of the CN for one type of drug–drug distance D_{ij}. We denoted these sub-networks as GC(D_{ij}). The Giant Components are very important because that contained a sub-set of nodes (drugs) that are near to (are similar to) all the nodes in the CN. For instance, Barabasi et al. studied very recently the GC of CN similar to the present; which describes multi-target drug–receptor interactions.[36] They used different TIs and/or CIs of these CNs to describe the network topology or connectivity numerically. Here, we calculated the number of nodes (*n*), number of edges (*m*), topological distance (*D*), average node degree (*z*), clustering coefficient (*C*), Zagreb group index 1 (*M*1), Zagreb group index 2 (*M*2), Randic connectivity index (X_r), Platt index (*F*), Index of relinking (*P*). The parameter *n* coincides with the number of drugs and *m* is the number of similar drug–drug pairs. Two drugs are connected according to the cut-off value. The threshold value used was a distance of 0.0051 for all networks. This cut-off value used was the fraction of the average distance of a node to all the others that guarantee that the node is connected to at least one other node. The description of these kind of parameters have been reported previously and the applications for small molecules, macromolecules, and networks reviewed.[34,37] The obtained topological values varied notably from one network to other. Considering that *C* was stimulated in the form of an edge density $C = 100 \times (m/m_{max})$ we can classify them approximately into two classes: CNs that resemble Small World Random networks and CNs similar to Low-Density Random networks.[38] In the first group we may put GC(Manhattan) and GC(Disagreement) and in the second group we can classify the other networks. In order to unravel local relationships of drug–drug similarity for instance the drugs predicted to be more similar to a new drug or the degree of similarity between two apparently unrelated drugs (drug–drug distance) we should select Small World networks. Because, we can select for each drug in average

a low number of similar drugs (z). In addition, the average distance between two drugs is unexpectedly short (D) so we can select on few relative drugs to see the change from one drug to other. In the present study we select GC(Manhattan) that presents a lower $z = 3$ and $D = 7$ avoiding the larger values of $z > 4$ and $D > 10$ of the other network. We do not preferred GC(Disagreement) because $z = 2$ is very small and we have to take into account the possibility of selecting false similar drugs due to mt-QSAR misclassifications. In Figure 1 we illustrate the picture of GC(Manhattan) visualized in the interface of CentiBin and exemplify the relationships between some known drugs.

4. Concluding remarks

The present mt-QSAR model is useful to predict antiviral activity of any organic compound against a large diversity of virus species, obtain BPMs and construct multi-species Complex Networks.

Acknowledgments

The authors thank the kind attention and useful comments of the Editor Professor Herbert Waldmann and the unknown reviewer. Prado-Prado, F. acknowledges financial support from Xunta the Galicia for a one-year post-doctoral position (Research Project IN89A 2007/101-0) under supervision of Dr. Martinez-de La Vega, O at CINVESTAV, Irapuato, Mexico. González-Díaz, H. acknowledges a partial financial support of program Isidro Parga Pondal, Xunta the Galicia. The same author thanks additional financial support from Xunta the Galicia for a one-year post-doctoral position (Research Project IN89A 2007/84-0) under supervision of Dr. Kuo-Chen Chou at the Gordon Life Science Institute, San Diego, CA, USA.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.11.075.

References and notes

1. Jaen, A.; Esteve, A.; Miro, J. M.; Tural, C.; Montoliu, A.; Ferrer, E.; Riera, M.; Segura, F.; Force, L.; Sued, O.; Vilaro, J.; Garcia, I.; Masabeu, A.; Altes, J.; Coltet, B.; Podzamczer, D.; Murillas, J.; Navarro, G.; Gatell, J. M.; Casabona, J. *J. Acq. Immun. Def. Synd.*, 2008. 212.
2. Tebourski, F.; Slim, A.; Elgaaied, A. *Diagn. Microbiol. Infect. Dis.* **2004**, *48*, 59.
3. Barin, F.; Cazein, F.; Lot, F.; Pillonel, J.; Brunet, S.; Thierry, D.; Damond, F.; Brun-Vezinet, F.; Desenclos, J. C.; Semaille, C. *Aids* **2007**, *21*, 2351.
4. Vilar, S.; Santana, L.; Uriarte, E. *J. Med. Chem.* **2006**, *49*, 1118.
5. Estrada, E.; Molina, E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 791.
6. Erhan, D.; L'Heureux, P. J.; Yue, S. Y.; Bengio, Y. *J. Chem. Inf. Model.* **2006**, *46*, 626.
7. Damos, D.; Smist, T. *Aviat. Space Environ. Med.* **1982**, *53*, 1177.
8. Maslovat, D.; Chus, R.; Lee, T. D.; Franks, I. M. *Motor Control* **2004**, *8*, 213.
9. Gia, O.; Marciani Magno, S.; Gonzalez-Diaz, H.; Quezada, E.; Santana, L.; Uriarte, E.; Dalla Via, L. *Bioorg. Med. Chem.* **2005**, *13*, 809.
10. Stiefl, N.; Baumann, K. *J. Chem. Inf. Model.* **2005**, *45*, 739.
11. Stiefl, N.; Baumann, K. *J. Med. Chem.* **2003**, *46*, 1390.
12. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7*, 1015.
13. Gonzalez-Diaz, H.; Gonzalez-Diaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. *Proteomics* **2008**, *8*, 750.
14. Gonzalez-Diaz, H.; Prado-Prado, F. J.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2006**, *14*, 5973.
15. Prado-Prado, F. J.; Gonzalez-Diaz, H.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15*, 897.
16. Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. *J. Comput. Chem.* **2007**, *28*, 1049.
17. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; Rotondo, R. *J. Mol. Graphics Modell.* **2007**, *26*, 32.
18. Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Sultan, S.; Torrens, F.; Rotondo, R. *Bioorg. Med. Chem.* **2007**, *15*, 1483.
19. StatSoft.Inc., 2002, pSTATISTICA (data analysis software system), version 6.0, www.statsoft.com.
20. Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. *J. Chem. Inf. Model.* **2006**, *46*, 1984.
21. Marrero-Ponce, Y.; Khan, M. T.; Casanola-Martin, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Garcia-Domenech, R.; Torrens, F.; Rotondo, R. *J. Comput. Aided Mol. Des.* **2007**, *21*, 167.
22. Alvarez-Ginarte, Y. M.; Marrero-Ponce, Y.; Ruiz-Garcia, J. A.; Montero-Cabrera, L. A.; Garcia de la Vega, J. M.; Noheda Marin, P.; Crespo-Otero, R.; Zaragoza, F. T.; Garcia-Domenech, R. *J. Comput. Chem.* **2008**, *29*, 317.
23. Marrero-Ponce, Y.; Meneses-Marcel, A.; Castillo-Garit, J. A.; Machado-Tugores, Y.; Escario, J. A.; Barrio, A. G.; Pereira, D. M.; Nogal-Ruiz, J. J.; Aran, V. J.; Martinez-Fernandez, A. R.; Torrens, F.; Rotondo, R.; Ibarra-Velarde, F.; Alvarado, Y. J. *Bioorg. Med. Chem.* **2006**, *14*, 6502.
24. González-Díaz, H.; Molina, R.; Hernandez, I.; Chemicals Bio-actives Center, Central University of Las Villas, Cuba, 2006, p MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design).
25. Gupta, A.; Maranas, C. D.; Albert, R. *Bioinformatics* **2006**, *22*, 209.
26. Koschützki, D.; Lehmann, K. A.; Peeters, L.; Richter, S.; Tenfelde-Podehl, D.; Zlotowski, O. In *Network Analysis: Methodological Foundations*; LNCS Tutorial Brandes, U., Erlebach, T., Eds.; Springer, 2005.
27. Mrvar, V. B. a. A. *p Pajek software 1.23*, 2008.
28. Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. *J. Chem. Inf. Model.* **2005**, *45*, 502.
29. Prado-Prado, F. J.; González-Díaz, H.; Martinez de la Vega, O.; Ubeira, F. M.; Chou, K. C. *Bioorg. Med. Chem.* **2008**, *16*, 5871.
30. Gonzalez-Diaz, H.; Prado-Prado, F. J. *J. Comput. Chem.* **2008**, *29*, 656.
31. Hasegawa, K.; Shindoh, H.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Ichihara, S.; Horii, I.; Shimma, N. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 968.
32. Molina, E.; González-Díaz, H.; Gonzalez, M. P.; Rodriguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
33. Estrada, E.; Molina, E. *J. Mol. Graphics Modell.* **2001**, *20*, 54.
34. González-Díaz, H.; Prado-Prado, F. *J. Comput. Chem.* **2008**, *29*, 656.
35. Zhang, W. *Environ. Monit. Assess.* **2007**, *124*, 253.
36. Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. *Nat. Biotechnol.* **2007**, *25*, 1119.
37. Junker, B. H.; Koschuetzki, D.; Schreiber, F. *BMC Bioinform.* **2006**, *7*, 219.
38. Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Complex Networks: From the Genome to the Internet*; WILEY-VCH GmbH & CO. KGa: Weinheim, 2003.