

Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds

Francisco J. Prado-Prado,^{a,b} Humberto González-Díaz,^{b,c,*} Octavio Martínez de la Vega,^a
Florencio M. Ubeira^b and Kuo-Chen Chou^c

^aDepartment of Bioinformatics, CINVESTAV, LANGEPIO, Irapuato 629 36500, Mexico

^bUnit for Bioinformatics & Connectivity Analysis (UBICA), Institute of Industrial Pharmacy,

Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Spain

^cGordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

Received 12 March 2008; revised 22 April 2008; accepted 25 April 2008

Available online 29 April 2008

Abstract—Several pathogen parasite species show different susceptibilities to different antiparasite drugs. Unfortunately, almost all structure-based methods are one-task or one-target Quantitative Structure–Activity Relationships (ot-QSAR) that predict the biological activity of drugs against only one parasite species. Consequently, multi-tasking learning to predict drugs activity against different species by a single model (mt-QSAR) is vitally important. In the two previous works of the present series we reported two single mt-QSAR models in order to predict the antimicrobial activity against different fungal (*Bioorg. Med. Chem.* **2006**, *14*, 5973–5980) or bacterial species (*Bioorg. Med. Chem.* **2007**, *15*, 897–902). These mt-QSARs offer a good opportunity (unpractical with ot-QSAR) to construct drug–drug similarity Complex Networks and to map the contribution of sub-structures to function for multiple species. These possibilities were unattended in our previous works. In the present work, we continue this series toward other important direction of chemotherapy (antiparasite drugs) with the development of an mt-QSAR for more than 500 drugs tested in the literature against different parasites. The data were processed by Linear Discriminant Analysis (LDA) classifying drugs as active or non-active against the different tested parasite species. The model correctly classifies 212 out of 244 (87.0%) cases in training series and 207 out of 243 compounds (85.4%) in external validation series. In order to illustrate the performance of the QSAR for the selection of active drugs we carried out an additional virtual screening of antiparasite compounds not used in training or predicting series; the model recognized 97 out of 114 (85.1%) of them. We also give the procedures to construct back-projection maps and to calculate sub-structures contribution to the biological activity. Finally, we used the outputs of the QSAR to construct, by the first time, a multi-species Complex Networks of antiparasite drugs. The network predicted has 380 nodes (compounds), 634 edges (pairs of compounds with similar activity). This network allows us to cluster different compounds and identify on average three known compounds similar to a new query compound according to their profile of biological activity. This is the first attempt to calculate probabilities of antiparasitic action of drugs against different parasites.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Infections caused by parasites have increased dramatically during the last decades and these diseases are

among the most contagious in the world. The most important is malaria, infection caused by *Plasmodium* spp.; near to 300 million people are infected each year, and 3 millions of them die each year.¹ Consequently, there is a high interest on the research of rational approaches for antiparasite drugs discovery. In this regard, various computational approaches,^{2,3} particularly QSAR studies^{4–6} may play an important role. Disappointingly, QSAR studies predict only structurally parent compounds acting against one single microbial species such as *Toxoplasma gondii*⁷ or *Leishmania* spp.⁸

Keywords: QSAR; Multi-tasking learning; Machine learning; Complex networks; Antiparasite drugs; Markov Chain Model; Malaria; *Plasmodium*; *Leishmania*; *Toxoplasma*; *Trypanosoma*; *Trichomonas*.

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912; e-mail: gonzalezdiazh@yahoo.es

for instance. More recently, very interesting models have been developed in order to predict antiparasite activity for heterogeneous series of compounds but without differentiating between various species with the same model.⁹ Models of this type have been developed for anti-toxoplasma,¹⁰ anti-entamoeba,¹¹ antimalarial,¹² trichomonacidal,¹³ antihelminthic,¹⁴ fasciolicides,¹⁵ anti-coccidials,^{16,17} and other antiparasite drugs. These models are a great step forward in antiparasite QSAR theory, but nevertheless a different QSAR model must be used for each and every one of the parasite species, in order to predict the antiparasite activity for a given series of compounds. Thence, the report of one single multi-target unified equation is very important for calculating the probability of activity of a given drug against different parasite species based on simple molecular descriptors.

Today there are near to 1600 molecular descriptors that may be in principle generalized and used to solve the former problem.¹⁸ In addition, other QSAR approaches for RNAs,^{19–21} and proteins^{20,22–24} have been recently introduced, with demonstrable utility, in medicinal chemistry including antimicrobial drugs research. 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, nucleotides, or amino acids labeled with physicochemical properties (mass, polarity, electronegativity, and charge). In a recent review our group have discussed latest advances on the field.²⁵ 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. Multi-tasking QSAR (mt-QSAR)²⁶ can be defined as the prediction of multiple outputs with a single model and is closely related to the more general term multi-tasking learning (used in cognitive sciences).^{27,28} This 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 mtQSAR models may be very useful to optimize important aspects such as activity, toxicity or pharmacokinetics using one single model. The first way to develop mtQSAR models is the consideration of an output variable for each activity or drug property we pretend to predict. This alternative usually leads to linear models based on the fit of a function by each one of the drug properties or non-linear models able to fit several outputs at time. For instance, in relation to antimicrobials QSAR research Marrero-Ponce et al.¹⁴ reported an LDA-QSAR for five mechanisms of actions of antihelminthic drugs; whereas Vilar et al.²⁹ reported a ANN-QSAR model able to predict up to four different mechanism of actions for HIV inhibitors. The first model is linear but bases on five classification functions while the second one is a non-linear model with a more hidden connection between the molecular descriptors and the properties predicted. The relative complexity of these models derives mainly from the high number of output variables we have to use in mtQSAR models learning. We called this alternative herein the Output-Coding multi-tasking learning.

Our group has recently proposed an alternative to Output-Coded mtQSAR models based on the codification of all the properties to be predicted in the input part of the model instead of in the output part. In this sense we could use dummy variables to encode each kind of property to be predicted in the input part of data but the problem remains the same. So, we decided to introduce information relative to the type of property to be predicted inside the molecular descriptors. We do not have any value of a free energy parameter for each molecule, for instance, with this kind of methodology. In our mt-QSAR approach we have different value of the CIs, or structural parameter, for the same molecule depending on the specific biphasic system, for example, where we want to estimate the partition coefficient of the drug³⁰ or the specific drug side effect we want to predict.^{31,32} We shall call this alternative from now on as the Input-Coded Multi-Tasking Learning approach and by extension we have Input-Coded mtQSAR models. The method is very flexible and can be extended to any type of molecular indices (CIs included).

2. Methods

2.1. Molecular descriptors

In this work, we focus on a QSAR method introduced elsewhere than the one used by a Markov Chain Model (MCM) to encode systems structural information using molecular, macro-molecular, supra-molecular CIs and 3D parameters as well. The method was first named as the **MARKovian CHemicals IN Silico D**esign approach (**MARCH-INSIDE**).^{33,34} This first name was oriented to describe the potentials applications in Medicinal Chemistry related to small-sized molecules.²¹ Currently, the method retains the acronym but with slight modification on the developed name: **MARCH-INSIDE: MARKov CHains INvariants for SIMulation & D**esign to better explain more broad applications.³⁵ We have recently reported two reviews: one in Current Topics Medicinal Chemistry (2007)³⁶ and the other in Proteomics (2008).³⁷ Both revisions made an in-depth review of 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 ${}^kC_s(\text{Set})$. These CIs can be interpreted as the average contributions ${}^kC_s(\text{Set})$ 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 parasite species (s). Figure 1 illustrates graphically 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 (Csat) or others. We derive these kC_s by summing up all the atomic contributions of each atom to the interaction ${}^0c_j(s)$ pre-multiplied by the probability of the atom distribution in the molecule ${}^A p_k(j,s)$. The ${}^0c_j(s)$ values depend both on the atom and on the parasite species while the ${}^A p_k(j,s)$ depends also on molecular

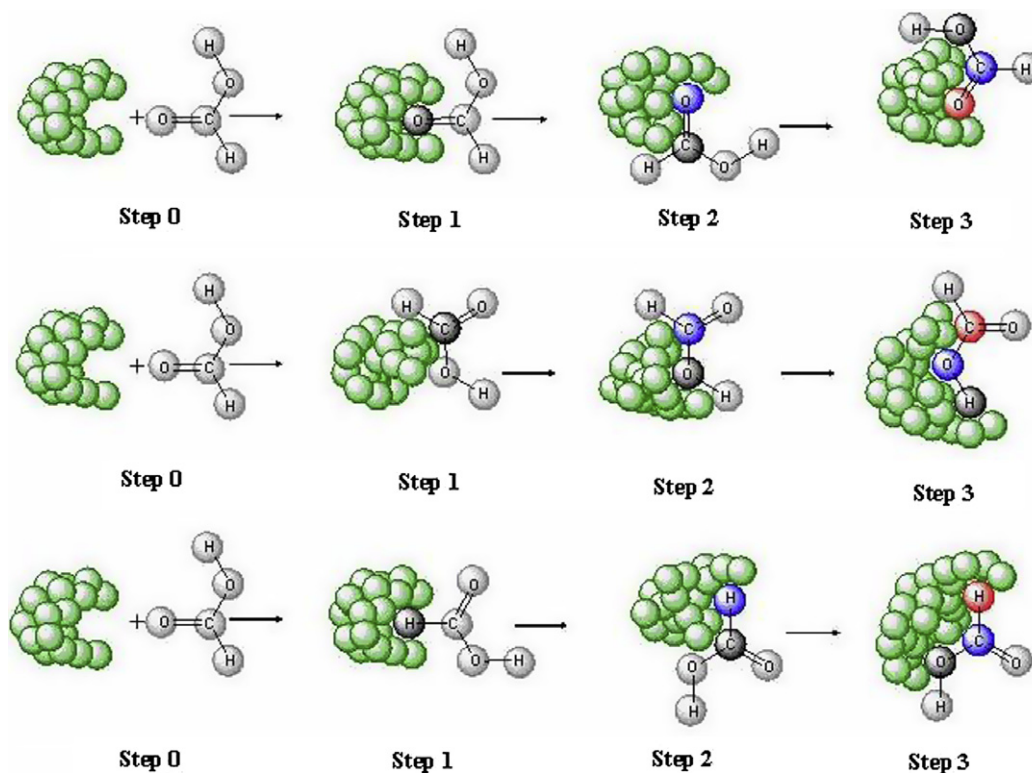


Figure 1. Stochastic drug-target step-by-step interaction.

topology or connectivity. The detailed calculation of these values using a MCM has been explained in our previous work of this series and others so we omit it for the sake of simplicity (see the two previous works of this series).^{38,39} These CIs were calculated using our software MARCH-INSIDE⁴⁰.

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

2.2. Statistical analysis

In order to continue the previous sections, we can attempt to develop a simple linear mt-QSAR using the general formula:

$$\text{Actv} = b_{0,G} \cdot {}^0 C_s(G) + b_{1,G} \cdot {}^1 C_s(G) + b_{2,G} \cdot {}^2 C_s(G) + b_{3,G} \cdot {}^3 C_s(G) \cdots + b_{k,G} \cdot {}^k C_s + b \quad (2)$$

Here ${}^k C_s(G)$ act as the microbial species dependent mt-descriptors. We selected Linear Discriminant Analysis (LDA)^{41,42} to fit the classification functions. The model deals with the classification of a compound set, that can be 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 the absence (Actv = -1) of antimicrobial 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.⁴³ 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 good classification, the ratios between the cases, variables in the equation and variables that have to be explored in order to avoid over-fitting or chance correlation. The model validation was corroborated by re-substitution of cases in four predicting series.^{44–47}

2.3. Data set

The data set includes marketed and/or very recently reported compounds with reported activity against different parasites; consult Table 1 SMA-b and Table 2 SM (online Supplementary material). In total, 500 different drugs experimentally tested against some species of a list of 16. Not all drugs were tested in the literature against all listed species so we were able to collect 694 cases (drug/species pairs) instead of 500×17 cases. The three data sets used were the following: training series: 115 active compounds plus 129 non-active compounds (244 in total); predicting series: $114 + 129 = 243$ in total; virtual screening 114 active compounds. The complete list of references used to collect the database is provided also at the end of the online Supplementary material.

2.4. Multi-species CNs based clustering

In order to perform the clustering analysis of anti-parasite multi-species activity with a CNs approach we carried out the following steps:

Table 1. Summary for the forward-stepwise analysis

CIs	F^a	p^b	Effect ^c	CIs	F^a	p^b	Effect ^c
⁰ C _s (Het)	103.797	0.001	In	⁰ C _s (X)	0.347	0.556	Out
⁰ C _s (Csat)	24.504	0.001	In	¹ C _s (X)	5.665	0.018	Out
⁰ C _s (Cinst)	74.247	0.001	In	² C _s (X)	5.477	0.020	Out
⁵ C _s (Cinst)	16.170	0.001	In	⁴ C _s (X)	0.498	0.480	Out
¹ C _s (T)	15.715	0.001	In	⁵ C _s (X)	1.671	0.197	Out
⁴ C _s (Het)	6.363	0.012	In	⁰ C _s (T)	2.197	0.139	Out
³ C _s (X)	4.249	0.040	Out	¹ C _s (Het)	5.683	0.017	Out
⁵ C _s (Csat)	1.568	0.211	Out	² C _s (Het)	1.055	0.305	Out
² C _s (X)	5.477	0.020	Out	³ C _s (Het)	0.034	0.852	Out
³ C _s (T)	6.734	0.010	Out	⁵ C _s (Het)	0.181	0.670	Out
¹ C _s (Csat)	3.241	0.073	Out	⁰ C _s (H-Het)	0.082	0.774	Out
² C _s (Csat)	0.070	0.790	Out	¹ C _s (H-Het)	0.942	0.332	Out
³ C _s (Csat)	2.712	0.100	Out	² C _s (H-Het)	0.017	0.895	Out
⁴ C _s (Csat)	3.190	0.075	Out	³ C _s (H-Het)	0.041	0.837	Out
⁵ C _s (T)	0.122	0.726	Out	⁴ C _s (H-Het)	34.197	0.836	Out
⁴ C _s (T)	2.728	0.099	Out	⁵ C _s (H-Het)	0.006	0.937	Out
¹ C _s (Cinst)	5.668	0.018	Out	³ C _s (Cinst)	0.974	0.324	Out
² C _s (Cinst)	1.252	0.264	Out	⁴ C _s (Cinst)	0.243	0.621	Out

^a Fisher ratio.^b p-level of error.^c In and Out indicates that the variable enter or not into the models after stepwise analysis.**Table 2.** Results of the model, analysis, validation and virtual-screening

	Percent	Antiparasite	Non-active
<i>Analysis</i>			
Antiparasite	82.6	95	20
Non-active	91.0	12	117
Total	87.0		
<i>Validation</i>			
Antiparasite	89.5	102	12
Non-active	81.4	24	105
Total	85.2		
<i>Virtual-screening</i>			
Antiparasite	85.1	97	17

Bold values denote the compounds predicted correctly.

- First, we calculated the molecular descriptors included in the mt-QSAR equation for 380 selected drugs using the MARCH-INSIDE software.
- We calculated the biological activity scores of every drug used against all the parasites species studied here by substituting the molecular descriptors into the mt-QSAR equation using the Microsoft Excel application.
- All the activity scores predicted were organized into a Table of drugs (rows) versus species (columns).
- This table was used as input for the software Statistica employed to calculate drug–drug multi-species correlations in the form of Euclidean distances.
- Using Microsoft Excel again we transformed the drug pair distances matrix derived of Statistica into a Boolean matrix. The elements of this matrix are equal to 1 if two drugs have a high or the same correlations are very close (short Euclidean distance). The threshold value used was a distance of 0.005. The line command used in Excel to transform the distance matrix into a Boolean matrix was $f = \text{if} (A\$1 = \$B2, 0, \text{if} (B2 > 0.0051, 0, 1))$. This 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 renaming the .txt file as a .mat file we read it with the software CentiBin.
- Using CentiBin we can not only represent the network but also highlight all drugs (nodes) connected to a specific drug and calculate many parameters, including node degree. The CentiBin software was used to calculate different measures of network topology such as the Diameter (D), Average distance (Av.Distance), Average node degrees (δ_{av}), Max node degree (δ_{max}).

2.5. Multi-species BPMs analysis

In order to illustrate the procedure of deriving antimicrobials multi-species activity by a back-projection approach we selected at random 4 drugs and 1 species (*Plasmodium falciparum*) and we created another back-projection, where we selected 1 drug and 4 species (*Leishmania donovani*, *Leishmania mexicana*, *Trypanosoma brucei* and *Plasmodium falciparum*) and carried out the following steps. These steps are the same reported for the construction of BPMs using classic ot-QSAR^{34,48,49} but we repeated them for each different parasite species using species-dependent atomic contributions:

- First, we calculated the species-dependent atomic descriptors included in the QSAR equation for selected drugs using the MARCH-INSIDE software.
- We calculated the contribution scores for each atom of the 4 drugs against a species studied by substituting the atomic descriptors into the QSAR equation using the Microsoft Excel application.
- Next, we repeated the same steps but calculated the contribution scores for each atom of the first drug against other species studied by substituting the atomic descriptors into the QSAR equation using the Microsoft Excel application.
- The contributions for each atom of the drug were scaled into a percentage value.

- The scaled atom contributions were grouped into different molecular fragments.
- These molecular fragment contributions to the biological activity were back-projected onto the molecular structure for obtaining a colour-scaled biological structure–activity BPMs.

3. Results and discussion

3.1. Multi-species QSAR model

One of the main advantages of the present stochastic approach is the possibility of deriving average thermodynamic parameters depending on the MM probabilities. The generalized parameters fit on more clearly physicochemical sense compared to our previous ones.^{22–24} In other words, this work introduces for the first time a single linear QSAR equation model for predicting the antiparasitic activity of drugs against different parasite species. Summary for the forward-stepwise analysis shows the variables that entire first in the model (Table 1). The final model selected was:

$$\begin{aligned} \text{actv} = & 4.15 \times 10^{-14} \cdot {}^1C_s(\text{T}) + 8.9 \times 10^{-14} \cdot {}^0C_s(C_{\text{sat}}) \\ & - 1.5 \times 10^{-13} \cdot {}^0C_s(C_{\text{unst}}) + 4.7 \times 10^{-7} \cdot {}^5C_s(C_{\text{unst}}) \\ & + 2 \times 10^{-7} \cdot {}^0C_s(\text{Het}) - 7.9 \times 10^{-7} \cdot {}^4C_s(\text{H-Het}) - 0.72 \\ Rc = & 0.75 \quad \lambda = 0.434 \quad F = 51.44 \quad p < 0.001 \end{aligned} \quad (3)$$

In this model Rc is the Canonical correlation coefficient.⁴⁴ It measures the power of the model to account for actual variability of data, or the same, how strong is the linear relationship assumed between the inputs and the classification of cases. It varies between 0 (poor) and 1 (perfect correlation) as is the case for other correlation coefficients. The coefficient λ is the Wilk's statistics for overall discrimination; F is the Fisher ratio, and p the error level.⁴⁴ In this equation, the ${}^kC_s(G)$ values were calculated on the totality (T) of the atoms in the molecule or for specific collections of atoms. These collections are atoms with a common characteristic as, for instance, are: heteroatoms (Het) or unsaturated Carbon atoms (C_{unst}) or hydrogen atoms bonded to heteroatoms (H-Het) or carbon atoms (C_{sat}). The model correctly classifies 95 out of 115 active compounds (86.9%) and 117 out of 139 non-active compounds (90.7%). Overall training predictability was 87.0% (212 out of 244 cases). Validation of the model was carried out by means of external predicting series, classifying the model 102 out of 114, 89.5% of compounds, see Table 2. The lift chart⁵⁰ gives a visual summary of the usefulness of the information provided by a QSAR model for predicting a binomial (categorical) outcome variable (dependent variable). Specifically, the chart summarizes the gain that can be expected by using the respective predictive model, as compared to using baseline information only. For instance, our model predicts the 35% of compounds from test list twice as well as they can be predicted using simple random selection (see Fig. 2). Taking into consideration previous reports on the use of LDA for ot-QSAR studies this result can be considered very good.^{51–55}

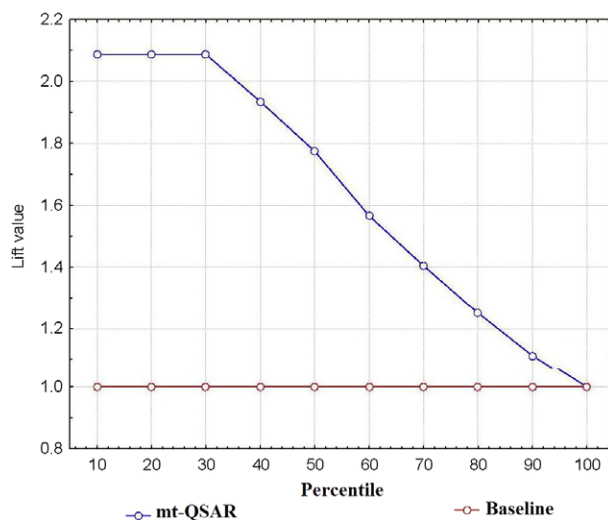


Figure 2. Lift Chart curve.

In order to calculate the different levels of the atomic contributions that guarantee the desired antiparasite activity we describe a desirability chart (see Fig. 3). This kind of graph projects the QSAR predicted value of the biological activity in a colour scale on a 2D Cartesian coordinates using different pairs of input variables in the axis. For example, one can map the tendency of a protein to act as anticancer agent against the electrostatic potentials in different regions of the protein⁵⁶ or the tendency of a micro-RNA to express at early stages against their folding thermodynamic parameters.⁵⁷ In our case, the higher C_{sat} is, the higher score for antiparasite activity can be predicted if the $C_{\text{H-Het}}$ values are low. We have previously used this technique for protein QSAR analysis.²⁷

In the common case of ot-QSAR the atomic contributions to the referenced above biological activity depend only on physicochemical atomic parameters such as atomic mass, polarizability, and charge,^{58–60} or electronegativity and/or chirality.⁴¹ The most remarkable characteristic of the present model is that the ${}^kC_s(G)$ parameters used as molecular descriptors depend not only on the molecular structure of the drug but interestingly they also depend on the parasite species that we have to control with the drug. The values of the corresponding atomic contributions ${}^0c(s)$ used to calculate ${}^kC_s(G)$ were reported herein for the first time for multi-species antiparasite action in Table 3 for some atoms and 16 selected parasite species. It was possible to model for the first time a very heterogeneous data for antiparasite drugs by using the above mentioned flexible definition of the present approach. The names of all the used drugs, the tested parasite species, and the results for training and validation by means of external predicting series are depicted in Table 1 SMa, b and c (see online Supplementary material). Finally, in order to show the good model functioning in practice, a virtual screening was carried out for recognizing the model as active 85.1%, 97 out of 114 antiparasitic compounds, not used in training or predicting series (see Table 2 SM).

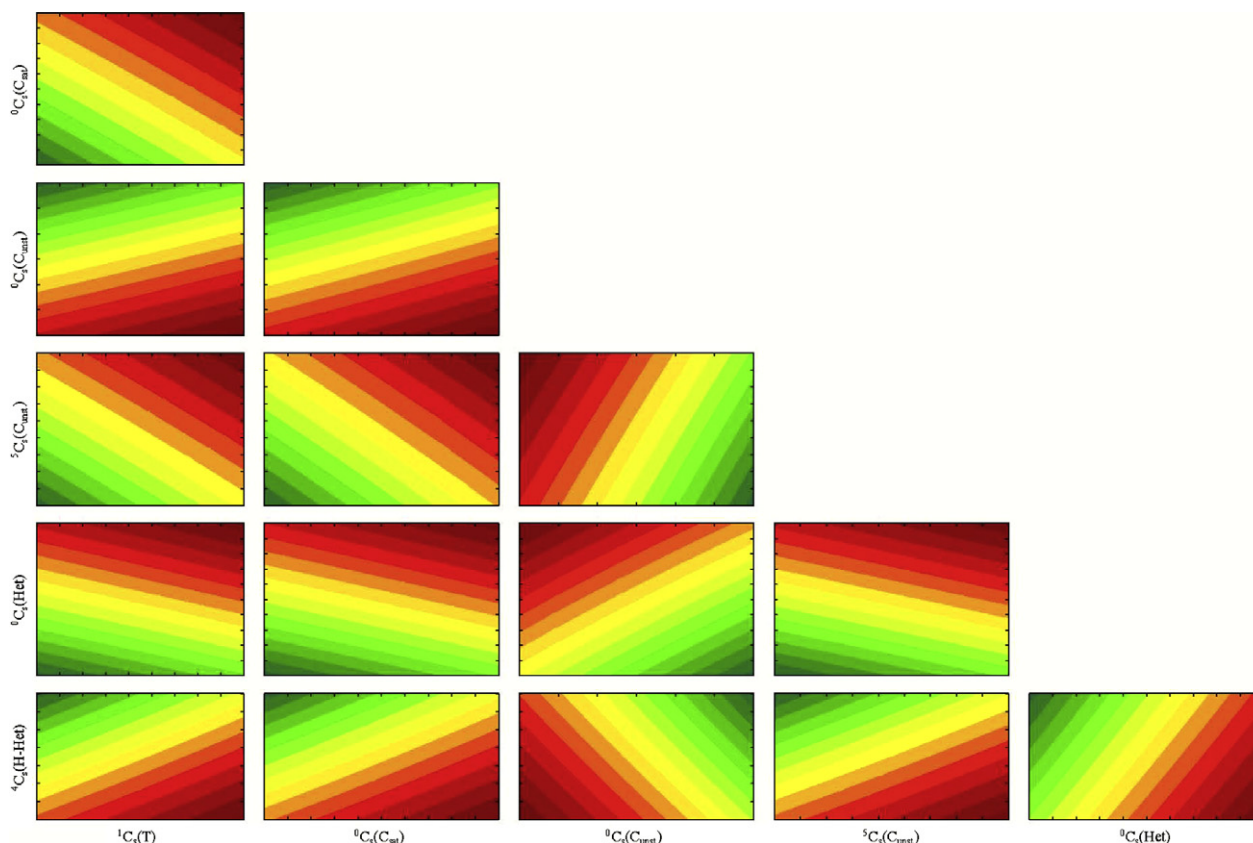


Figure 3. Desirability analysis.

3.2. Multi-species back-projection

In the two previous works of this series we applied the above mentioned Input-Coded mtQSAR philosophy to the **MARCH-INSIDE** method in order to extend it to predict antibacterial³⁹ and anti-fungal³⁸ activity of drugs 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.⁶¹ These mt-QSARs offer an unprecedented opportunity to construct Back-Projection Maps (BPMs)^{62,63} and drug–drug similarity Complex Networks (CNs). Back-Projection^{64,65} is very useful to map or project the predicted function backwards onto structure and determine the contribution of sub-structural molecular regions of the desired property. On the other hand, CNs study is a growing field in modern science with broad applications for studying connections between objects in very large databases ranging from drug, protein–protein interactions,⁶⁶ gen–gen⁶⁷ and tissue specific RNA–RNA co-expression⁶⁸ at the molecular level of the Brain cortex region co-activation,⁶⁹ Social networks,⁷⁰ and Internet⁷¹ or other systems at macroscopic level.⁷²

In addition, we constructed two back-projections: for the first one, we selected at random 4 drugs and 1 species (*Plasmodium falciparum*) whose results in our model are as follows: the hydroxylamines derivatives have differences in the activity (see Table 4) and concurrence in

the literature. The model predicted a 23% ring interaction of pentafluorobenzyl-*O*-hydroxylamine while predicted a 32% ring interaction of the phenyl-*O*-hydroxylamine that indicated the major interaction with the receptor, due to the ability of the parasite to absorb the compound. For the second one, we selected 1 drug and 4 species (*Leishmania donovani*, *Leishmania mexicana*, *Trypanosoma brucei* and *Plasmodium falciparum*) and predicted a 34% of interaction drug–receptor in every ring of pentamidine in all the species tested, that indicates the active part of the drug is the aromatic ring. That is why the species tested on the BPMs, except *Plasmodium falciparum*, belong to the same family. This result illustrates the advantages of extending from ot-QSAR⁷³ to mt-QSAR the BPMs study of the atom contribution, bonds, and sub-structures in general to the biological activity.

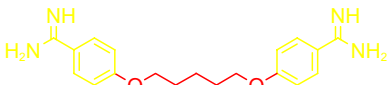
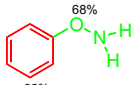
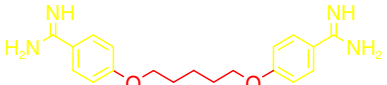
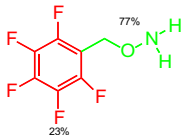
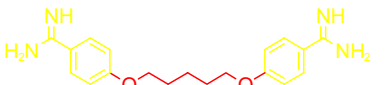
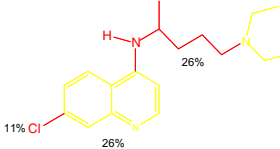

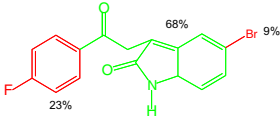
3.3. CNs based multi-species clustering of antiparasite compounds

In addition, antiparasite compounds were used to extend model validation for network and back-projection construction. In the present work, a multi-species complex networks were developed for the first time, the network have predicted 380 nodes (compounds), 634 edges (pairs of compound with similar activity) with a distribution closer to normal than to exponential (See Fig. 4). We measured 5 important network parameters: number of nodes ($n = 380$), number of edges ($m = 634$), Average Distance (Dist.av. = 10.6), Diameter or ($D = 33$), and

Table 3. Atomic contributions values for atom–receptor interactions

Parasite	C	H	N	O	S	Cl
<i>Cryptosporidium parvum</i>	0.2	0.202	0.195	0.183	0.095	0.253
<i>Entamoeba histolytica</i>	0.155	0.152	0.133	0.139	0	0.146
<i>Leishmania amazonensis</i>	0.176	0.182	0.176	0.176	0	0.176
<i>Leishmania donovani</i>	0.195	0.187	0.186	0.179	0.141	0.301
<i>Leishmania infantum</i>	0.192	0.195	0.184	0.194	0.146	0.163
<i>Leishmania major</i>	0.156	0.168	0.196	0.128	0	0.301
<i>Leishmania mexicana</i>	0.188	0.192	0.184	0.204	0	0
<i>Plasmodium falciparum</i>	0.213	0.213	0.217	0.173	0.156	0.258
<i>Pneumocystis carinii</i>	0.286	0.279	0.286	0.269	0	0
<i>Toxoplasma gondii</i>	0.221	0.219	0.235	0.215	0.156	0.186
<i>Trichomonas vaginalis</i>	0.15	0.153	0.116	0.189	0.13	0.222
<i>Trypanosoma brucei brucei</i>	0.174	0.17	0.156	0.186	0.208	0.222
<i>Trypanosoma brucei rhodesiense</i>	0.216	0.205	0.204	0.216	0.176	0
<i>Trypanosoma cruzi</i>	0.19	0.187	0.212	0.179	0.189	0.243

Table 4. Back-projection maps, atomic interaction drug–receptor

Species	BPMs ^a	Species	BPMs ^b
<i>Plasmodium falciparum</i>	 Pentamidine	<i>Plasmodium falciparum</i>	 Phenyl-O-hydroxylamine
<i>Leishmania donovani</i>	 Pentamidine	<i>Plasmodium falciparum</i>	 Pentafluorobenzyl-O-hydroxylamine
<i>Leishmania mexicana</i>	 Pentamidine	<i>Plasmodium falciparum</i>	 Chloroquine
<i>Trypanosoma brucei</i>	 Pentamidine	<i>Plasmodium falciparum</i>	 Oxinole 18

^a Red, 0–33%; yellow, 34–67% and green, 68–100% of contribution to the total activity.

^b Values contribution to activity in the map.

Average degree ($\delta_{av} = 3.34$), and Max degree ($\delta_{max} = 10$). This means that the network clusters 380 different drugs and identifies 634 pairs of similar drugs according to multi-species antiparasite activity. For a network with $n = 380$ nodes we can calculate a top number of $m_{max} = n! / [(n-2)! \cdot 2!] = 380! / [(380-2)! \cdot 2!] = (380 \cdot 379 \cdot 378!) / (378! \cdot 2) = 72,010$ edges (drug–drug pairs). Consequently, the drug–drug pair density of our network is $\rho = m / m_{max} = 634 / 72,010 = 0.0089$, which is a very low value. This fact indicates that our model does not make overestimated predictions for the similarity between pairs of drugs. It coincides with the large network Diameter and Average distance between two drugs, indicating a good

separation between pairs of drugs. The low average node degree indicates that if we use the network to predict a new drug we found on average 3 and no more than 10 similar drugs previously studied. This fact guarantees the practical use of the network for the fast retrieval search of compounds with similar multi-species profiles of antiparasite activity. Prediction of only one candidate may be uncertain and a very high number of candidates can be difficult to interpret. We can also give an example on the use of the network for the identification of trimethoprim derivatives with similar action mechanism. All the trimethoprim derivatives have good classification by the model and have been reported in the literature as active.

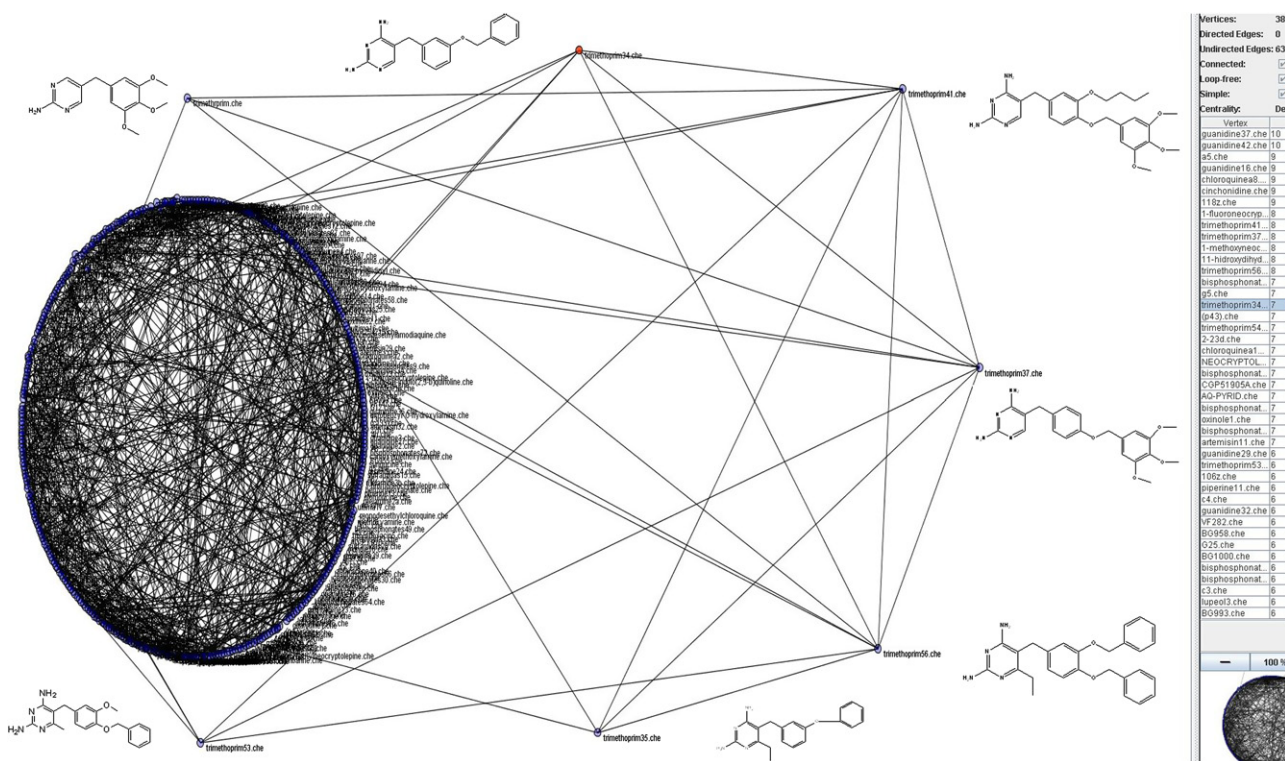


Figure 4. Graphic model of a multi-species complex network.

4. Concluding remarks

We devised the way of constructing BPMs and CNs with our Input-Coded mt-QSAR approach. These goals are unpractical or difficult with ot-QSAR due to the reasons, above given, on the higher number of models we have to use or with Output-Coded mt-QSAR because of the complexity of the QSAR models developed. The emerging possibilities of Input-Coded mt-QSAR were unattended in the two previous works of this series. In the present work, we explore these new possibilities on BPMs and CNs clustering and, at the same time, we continue this series towards other important direction of antiparasite chemotherapy drugs. We also give the procedure to construct multi-species antiparasite BPMs and to calculate the contribution of each sub-structure to the biological activity. Finally, we used the outputs of the Input-Coded mt-QSAR to construct, for the first time, multi-species CNs of antiparasite drugs. The CN predicted has 380 nodes (compounds), 634 edges (pairs of compounds with similar activity). This network allows us to cluster different compounds and identify on average three known compounds similar to a new query compound according to their profile of biological activity. This is the first attempt to calculate probabilities of antiparasitic action of drugs against different parasite using a single mt-QSAR.

In closing, we can add some brief concluding remarks and outline future research directions. The present work can be considered as the first unify QSAR model reported to predict antiparasitic activity of any organic compound against a very large diversity of parasitic pathogens. Back-projection and multi-species Complex networks were introduced for the first time increasing

the potential use of the model compared to previous works of this series. Consequently, if we also take into consideration the demonstrated success of the method for antibacterial and antifungal drugs, we can generally expect similar performance in antiviral activity and possibly in other not antimicrobial actions in the future.

Acknowledgments

We are grateful to the BMC Regional Editor Prof. Waldmann H. by his kind attention in the process of publication of this work. We also acknowledge sincerely the opinions of unknown referees that helped us to increase the final quality of the work. Prado-Prado, F. thanks 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 tenure-track research contract funded by the program Isidro Parga Pondal of the 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 Gordon Life Science Institute, San Diego, CA, USA. This study was supported in part by grants PGIDIT05RAG20301PR from the Regional Ministry of Innovation and Industry (Consellería de Innovación e Industria, Xunta de Galicia, Spain) and AGL2006-13936-C02-01 from the Ministry of Education and Science (Ministerio de Educación y Ciencia, Spain).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.04.068.

References and notes

- Cooke, B. M.; Mohandas, N.; Coppel, R. L. *Semin. Hematol.* **2004**, *41*, 173.
- Chou, K. C. *Curr. Med. Chem.* **2004**, *11*, 2105.
- Chou, K. C.; Wei, D. Q.; Du, Q. S.; Sirois, S.; Zhong, W. Z. *Curr. Med. Chem.* **2006**, *13*, 3263.
- Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. J. *Am. Chem. Soc.* **1997**, *119*, 10509.
- Pae, A. N.; Kim, S. Y.; Kim, H. Y.; Joo, H. J.; Cho, Y. S.; Choi, K. I.; Choi, J. H.; Koh, H. Y. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 2685.
- Du, Q. S.; Mezey, P. G.; Chou, K. C. *J. Comput. Chem.* **2005**, *26*, 461.
- Ling, Y.; Sahota, G.; Odeh, S.; Chan, J. M.; Araujo, F. G.; Moreno, S. N.; Oldfield, E. *J. Med. Chem.* **2005**, *48*, 3130.
- Van Miert, S.; Van Dyck, S.; Schmidt, T. J.; Brun, R.; Vlietinck, A.; Lemiere, G.; Pieters, L. *Bioorg. Med. Chem.* **2005**, *13*, 661.
- 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. Model.* **2005**, *45*, 1082.
- Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. *SAR QSAR Environ. Res.* **1999**, *10*, 47.
- Gangjee, A.; Lin, X. J. *Med. Chem.* **2005**, *48*, 1448.
- Marrero-Ponce, Y.; Montero-Torres, A.; Zaldivar, C. R.; Veitia, M. I.; Perez, M. M.; Sanchez, R. N. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
- Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Pereira, D. M.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Aran, V. J.; Martinez-Fernandez, A. R.; Garcia Sanchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838.
- Marrero-Ponce, Y.; Castillo-Garrit, 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.
- González-Díaz, H.; Olazabal, E.; Castanedo, N.; Sanchez, I. H.; Morales, A.; Serrano, H. S.; Gonzalez, J.; de Armas, R. R. *J. Mol. Model.* **2002**, *8*, 237.
- González-Díaz, H.; Bastida, I.; Castanedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; de Armas, R. R. *Bull. Math. Biol.* **2004**, *66*, 1285.
- González-Díaz, H.; Olazabal, E.; Santana, L.; Uriarte, E.; Castañedo, N. *Bioorg. Med. Chem.* **2007**, *15*, 962.
- Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, 2002.
- Bermudez, C. I.; Daza, E. E.; Andrade, E. *J. Theor. Biol.* **1999**, *197*, 193.
- Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Int. J. Mol. Sci.* **2004**, *5*, 276.
- González-Díaz, H.; Agüero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Gianni, P. *J. Comput. Chem.* **2007**, *28*, 1049.
- González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Gonzalez-Díaz, Y.; Sanchez-Gonzalez, A. *J. Comput. Chem.* **2007**, *28*, 1042.
- González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990.
- González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. *J. Proteome Res.* **2007**, *6*, 904.
- Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.
- Erhan, D.; L'Heureux, P. J.; Yue, S. Y.; Bengio, Y. *J. Chem. Inf. Model.* **2006**, *46*, 626.
- Damos, D.; Smist, T. *Aviat. Space Environ. Med.* **1982**, *53*, 1177.
- Maslovat, D.; Chus, R.; Lee, T. D.; Franks, I. M. *Motor Control* **2004**, *8*, 213.
- Vilar, S.; Santana, L.; Uriarte, E. *J. Med. Chem.* **2006**, *49*, 1118.
- Cruz-Monteagudo, M.; González-Díaz, H.; Agüero-Chapin, G.; Santana, L.; Borges, F.; Domínguez, R. E.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1909.
- Cruz-Monteagudo, M.; González-Díaz, H.; Uriarte, E. *Bull. Math. Biol.* **2006**, *68*, 1527.
- Cruz-Monteagudo, M.; González-Díaz, H. *Eur. J. Med. Chem.* **2005**, *40*, 1030.
- González-Díaz, H.; Sanchez, I. H.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
- González-Díaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model.* **2003**, *9*, 395.
- Cruz-Monteagudo, M.; González-Díaz, H.; Borges, F.; Domínguez, E. R.; Cordeiro, M. N. *Chem. Res. Toxicol.* **2008**, *21*, 619.
- González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7*, 1025.
- González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. *Proteomics* **2008**, *8*, 750.
- González-Díaz, H.; Prado-Prado, F. J.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2006**, *14*, 5973.
- Prado-Prado, F.; González-Díaz, H.; Santana, L.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15*, 897.
- González-Díaz, H.; Molina-Ruiz, R.; Hernandez, I. **MARCH-INSIDE** v3.0 (**MARKov CHains INvariants for SIMulation & DEsign**), 2007; Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.
- Castillo-Garrit, J. A.; Marrero-Ponce, Y.; Torrens, F.; Rotondo, R. *J. Mol. Graph. Model.* **2007**, *26*, 32.
- Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Sultan, S.; Torrens, F.; Rotondo, R. *Bioorg. Med. Chem.* **2007**, *15*, 1483.
- StatSoft Inc., STATISTICA (data analysis software system), 2002.
- Hill, T.; P. Lewicki STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining; StatSoft: Tulsa, 2006.
- Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Khan, K. M.; Torrens, F.; Rotondo, R. *Eur. J. Med. Chem.* **2007**, *42*, 1370.
- Alvarez-Ginarte, Y. M.; Marrero-Ponce, Y.; Ruiz-Garcia, J. A.; Montero-Cabrera, L. A.; Vega, J. M.; Noheda Marin, P.; Crespo-Otero, R.; Zaragoza, F. T.; Garcia-Domenech, R. *J. Comput. Chem.* **2007**.
- Castillo-Garrit, J. A.; Marrero-Ponce, Y.; Torrens, F. *Bioorg. Med. Chem.* **2006**, *14*, 2398.
- Saiz-Urra, L.; Gonzalez, M. P.; Collado, I. G.; Hernandez-Galan, R. *J. Mol. Graph. Model.* **2007**, *25*, 680.

49. Estrada, E.; González-Díaz, H. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75.
50. Yi, B.; Hughes-Oliver, J. M.; Zhu, L.; Young, S. S. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1221.
51. Helguera, A. M.; Cabrera Perez, M. A.; Gonzalez, M. P.; Ruiz, R. M.; González-Díaz, H. *Bioorg. Med. Chem.* **2005**, *13*, 2477.
52. Mattioni, B. E.; Jurs, P. C. *J. Mol. Graph. Model.* **2003**, *21*, 391.
53. Marrero-Ponce, Y.; Khan, M. T.; Casanola Martin, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F.; Rotondo, R. *ChemMedChem* **2007**, *2*, 449.
54. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borrás, R. *J. Antimicrob. Chemother.* **2004**, *53*, 65.
55. Saiz-Urra, L.; Gonzalez, M. P.; Fall, Y.; Gomez, G. *Eur. J. Med. Chem.* **2007**, *42*, 64.
56. González-Díaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Díaz, Y. *J. Inorg. Biochem.* **2006**, *100*, 1290.
57. Gonzalez-Diaz, H.; Vilar, S.; Santana, L.; Podda, G.; Uriarte, E. *Bioorg. Med. Chem.* **2007**, *15*, 2544.
58. Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. *J. Chem. Inf. Model.* **2005**, *45*, 502.
59. Perez Gonzalez, M.; Dias, L. C.; Helguera, A. M.; Rodriguez, Y. M.; de Oliveira, L. G.; Gomez, L. T.; González-Díaz, H. *Bioorg. Med. Chem.* **2004**, *12*, 4467.
60. Perez Gonzalez, M.; Morales Helguera, A. *J. Comput. Aided Mol. Des.* **2003**, *17*, 665.
61. González-Díaz, H.; Prado-Prado, F. *J. Comput. Chem.* **2008**, *29*, 656.
62. Gia, O.; Marciani Magno, S.; González-Díaz, H.; Quezada, E.; Santana, L.; Uriarte, E.; Dalla Via, L. *Bioorg. Med. Chem.* **2005**, *13*, 809.
63. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. *Comput. Chem.* **2002**, *26*, 583.
64. Stiefl, N.; Baumann, K. *J. Chem. Inf. Model.* **2005**, *45*, 739.
65. Stiefl, N.; Baumann, K. *J. Med. Chem.* **2003**, *46*, 1390.
66. Estrada, E. *Proteomics* **2006**, *6*, 35.
67. Gupta, A.; Maranas, C. D.; Albert, R. *Bioinformatics* **2006**, *22*, 209.
68. Yu, X.; Lin, J.; Zack, D. J.; Qian, J. *Nucleic Acids Res.* **2006**, *34*, 4925.
69. Honey, C. J.; Kotter, R.; Breakspear, M.; Sporns, O. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 10240.
70. Barabasi, A. L. *Science* **2005**, *308*, 639.
71. Yook, S. H.; Jeong, H.; Barabasi, A. L. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 13382.
72. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. *Phys. Rep.* **2006**, *424*, 175.
73. Estrada, E.; Quincoces, J. A.; Patlewicz, G. *Mol. Divers* **2004**, *8*, 21.