10th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-10). 1-30 November 2006. http://www.usc.es/congresos/ecsoc/10/ECSOC10.htm & http://www.mdpi.org/ecsoc-10/

[g003]

Unify QSAR approach to antibacterial activity of organic drugs against different species

Humberto González-Díaz,* Francisco J. Prado-Prado, Lourdes Santana, Eugenio Uriarte.

Department of Organic Chemistry, University of Santiago de Compostela 15782, Spain.

Abstract: There are many different kinds of pathogen bacteria species with very different susceptibility profile to different antibacterial drugs. One limitation of QSAR models are the biological activity of drugs against only one bacteria species. In previous paper we develop one unified Markov model to describe the biological activity of different drugs tested in the literature against some of the antimicrobial species. Consequently predicting the probability with which a drug is active against different bacteria species with a single unify model is a goal of the major importance. This work develops one unified Markov model to describe the biological activity of more than 70 drugs tested in the references against to 96 bacteria species. Linear Discriminant Analysis (LDA) classifying drugs as active or non-active against the different tested bacteria species processed the data. The model correctly classifies 199 out of 237 active compounds (83.9%) and 168 out of 200 non-active compounds (84%). Overall training predictability was 84% (367 out of 437 cases). Validation of the model was carring out by means of external predicting series, classifying the model 202 out 243, 83.13% of compounds. In order to show how the model function in practice a virtual screening was carring out recognizing the model as active 84.5%, 480 out of 568 antibacterial compounds not used in training or predicting series. The present is an attempt to calculate withing a unify framework probabilities of antibacterial action of drugs against many different species.

***corresponding author:** gonzalezdiazh@yahoo.es or qohumbe@usc.es

1. Introduction

With the increase in resistance of bacteria to antibiotic treatment, attention has focussed on developing novel means of anti-microbial therapies. One approach is to exploit natural mechanisms used by mammals including humans to combat microbial invaders. Modern rational drug design widely relies on building extensive QSAR (quantitative structure-activity relationships) models which represent a substantial part of the current *'in silico'* research. QSAR can then be utilized to optimizing both the activity profile for the molecule and its chemical synthesis.[1] Disappointingly; QSAR studies are generally based on databases considering only structurally parent compounds acting against one single microbial species. As a consequence, to predict the antimicrobial 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.[2] In previous paper, we develop one unified Markov model to describe the biological activity of different drugs tested in the literature against different antimicrobial species. In this sense, it is very important the report of one single unified equation to calculate the probability of activity of a given drug against different antimicrobial species.

Bacteria infections 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. Water purification, immunization (vaccination) and modern antibiotic treatment continueto 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 recognizing 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]

There are more than 1 600 molecular descriptors that may be in principle generalized and used to solve the former problem.[4-7] In addition other QSAR approaches have been introduced recently with demonstrated utility in medicinal chemistry.[8-11] In any case, no one of these indices have been extended yet to encode additional information to chemical structure. Our group has introduced elsewhere one Markov Model (MM) encoding molecular backbones information, with several applications in bioorganic medicinal chemistry. The method was named the MARCH-INSIDE approach, MARkovian CHemicals IN SIlico Design. It allowed us introducing matrix invariants such as stochastic entropies and spectral moments for the study of molecular properties. Specifically, the stochastic spectral moments introduced by our group have been largely used for small molecules QSAR problems including design of fluckicidal, anticancer and antihypertensive drugs. Applications to macromolecules have been restricted to the field of RNA without applications to proteins.[12-15] The entropy like molecular descriptors has demonstrated flexibility in many bioorganic and medicinal chemistry problems such as: estimation of anticoccidial activity, modeling the interaction between drugs and HIV-packaging-region RNA, and predicting proteins and virus activity.[16-22]

In recent studies, the MARCH-INSIDE method has been extended to encompass molecular environment interesting information in addition to molecular structure. This new interpretation allows calculating molecular thermodynamic free energy for many physicochemical and biological processes.[23,24] This approach is able to take into consideration for instance not only the molecular structure of the drug but the free energy of its interaction with the specific microbial organism the drug has to eliminate, too. The present study develops a single linear equation based on these previous ideas to predict the antibacterial activity of drugs against different species.

2. Methods

2.1. Markov model for drug-target step-by-step interaction

We will consider a hypothetical situation in which a drug molecule is free in the space at an arbitrary initial time (t_0). It is then interesting to develop a simple stochastic model for a step-by-step interaction between the atoms of a drug molecule and a molecular receptor in the time of beginning of the pharmacological effect. For the sake of simplicity, we consider a model in which unknown or not taken into consideration the chemical structure of the receptor.

Let be, the initial contribution of the j-th atom to the drug-receptor interaction is ${}^{0}c_j(s)$. In this symbol the c points to contribution, the 0 indicates that we refer to the initial interaction atom-receptor, and the s indicate that the contribution depends on the specific microbial species. Afterwards, we have to define the contribution ${}^{k}c_{ij}(s)$ of interaction between the j-th atom and the receptor given that i-th atom has been interacted at previous time t_k. With respect to ${}^{1}c_{ij}(s)$ we must taking 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. In particular, immediately after of the first interaction ($t_0 = 0$) takes place an interaction ${}^{1}c_{ij}(s)$ at time $t_1 = 1$ and so on. In consonance, we defined ${}^{1}c_{ij}(s) = \alpha_{ij} \cdot {}^{0}c_j(s)$, being $\alpha_{ij} = 1$ if the j-th atom is adjacent to the i-th one and $\alpha_{ij} = 0$ otherwise. So, one can suppose that, atoms binds to its receptor in discrete intervals of time t_k. There several alternative ways in which such step-by-step binding process may occur. Figure 1 illustrates this idea.

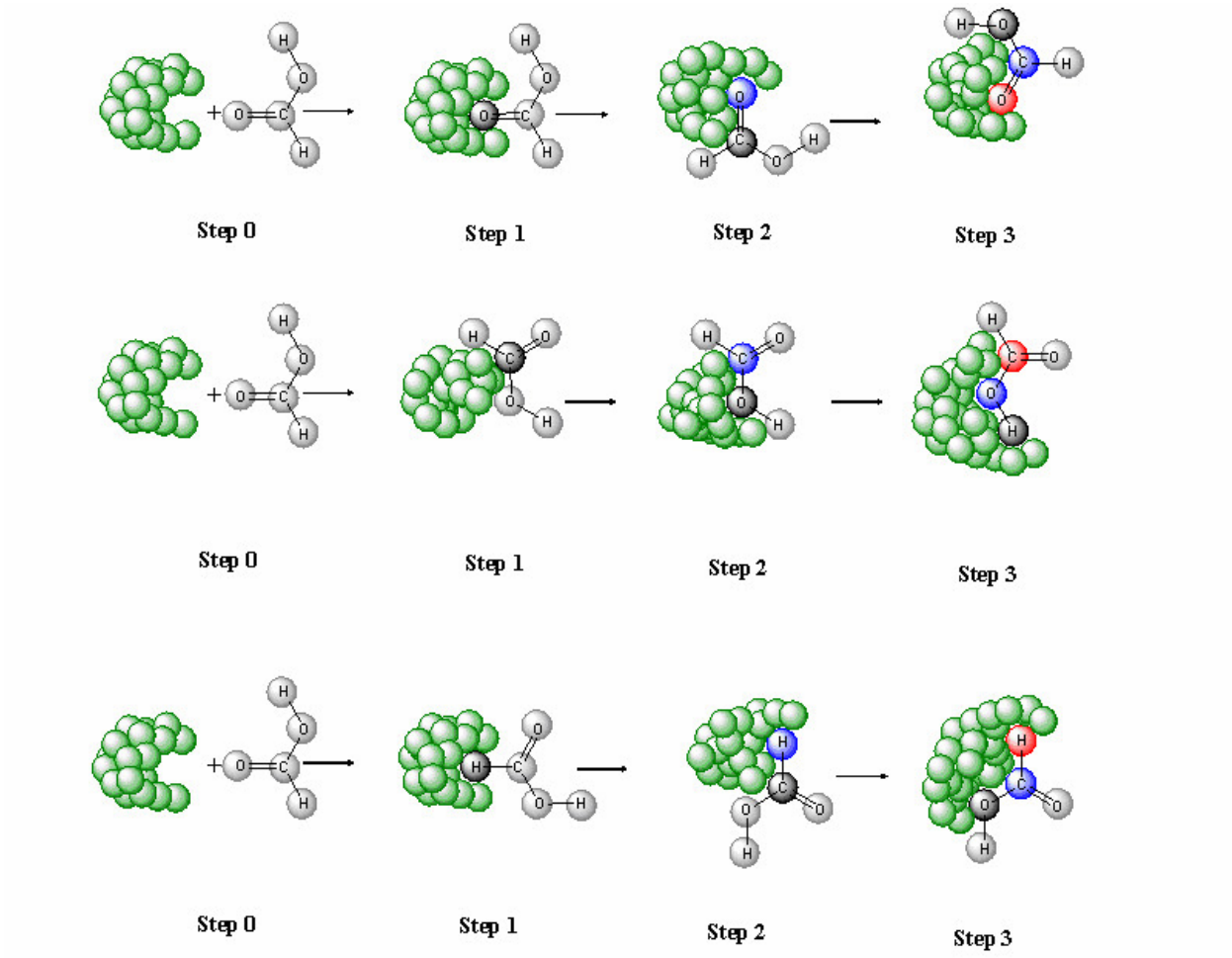

Figure1. Stochastic drug-target step-by-step interaction

Markov Model allowed us to derive the average contributions ${}^{k}C_s$ of the atoms in the molecule to the gradual interaction between the drug and the receptor at a specific time k in a given microbial species (s). We derive these ${}^{k}C_s$ by summing up all the atomic contributions of interaction ${}^{0}c_j(s)$ pre-multiplied by the absolute probabilities of drug-target interaction ${}^{A}p_k(j,s)$:[23-25]

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

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 markov model for drug-target step-by-step interaction method was describe in a previous paper.[26]

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:

$$ Actv = b_0 \cdot {}^{0}C_s + b_1 \cdot {}^{1}C_s + b_2 \cdot {}^{2}C_s + b_3 \cdot {}^{3}C_s ..... + b_k \cdot {}^{k}C_s + b \qquad (2) $$

Here, ${}^{k}C_s$ act as the microbial species specific molecule-target interaction descriptors. We selected Linear Discriminant Analysis (LDA)[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 antimicrobial activity of the drug against the specific species. In equation (8), b_k 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.[27] Forward stepwise was fixed as the strategy for variable selection.[19,20]

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.[26,27]

2.3. Data set

The data set was conformed by a set of marketed and/or very recently reported antibacter drugs with a $MIC_{50} \leq 10$ μM against different bacterias. The three data sets used were as follows training series: 199 active compounds plus 168 non-active compounds (367 in total); predicting series: 137 + 106 = 243 in total; virtuals screening 568 active compounds. The literature reports experimental test of each drug against some but not all species of a list of 137. In consequence, we were able to collect 1248 cases (drug/species pairs). The names or codes for all compounds as well as the references consulted can be obtained from the corresponding author upon request.

3. Results and discussion

The advantage of the present stochastic approach is the possibility of deriving average contributions to the biological activity 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.[23-25] In specific, this work is the first one that introduces a single linear QSAR equation model to predict the antibacterial activity of drugs against different species.

The best model found was:

$$\begin{aligned} Actv = & -1.12 \cdot {}^{1}C_S(T) + 1.34 \cdot {}^{3}C_S(T) \\ & + 1.84 \cdot {}^{0}\cdot C_S(C_{sat}) \\ & - 0.90 \cdot {}^{0}C_S(C_{uns}) \\ & + 0.88 \cdot {}^{5}C_S(X) \\ & - 1.27 \cdot {}^{0}C_S(H - Het) \\ & - 0.90 \cdot {}^{2}C_S(H - Het) \\ & + 0.698 \qquad (3) \end{aligned}$$

$$\lambda = 0.49 \qquad Rc = 0.715 \qquad p < 0.001$$

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, ${}^{k}C_s$ 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 as for instance are: halogens (X) or unsaturated Carbon atoms (C) or heteroatom-bound hydrogen atoms (H-Het). Summary for the forward-stepwise analysis shows the variables that enter first in the model (**Table1).**

Table1. Summary for the forward-stepwise analysis.

	F	*P*	*Effect*
${}^{0}C_s$(Het)	68.2	0.001	In
${}^{0}C_s$(Csat)	151.3	0.001	In
${}^{3}C_s$(X)	50	0.001	In
${}^{1}C_s$(T)	59.9	0.001	In
${}^{0}C_s$(Cinst)	50.4	0.001	In
${}^{5}C_s$ (Csat)	47.6	0.001	In
${}^{2}C_s$ (X)	24.3	0.001	Entered
${}^{3}C_s$ (T)	12.6	0	Out
${}^{1}C_s$ (Csat)	0.7	0.398	Out
${}^{2}C_s$ (Csat)	0.9	0.334	Out
${}^{3}C_s$ (Csat)	0.2	0.623	Out
${}^{4}C_s$ (Csat)	0.2	0.641	Out
${}^{5}C_s$ (T)	2.4	0.123	Out
${}^{4}C_s$ (T)	13.1	0	Out
${}^{1}C_s$ (Cinst)	0.8	0.38	Out
${}^{2}C_s$ (Cinst)	2.3	0.127	Out
${}^{3}C_s$ (Cinst)	2.1	0.145	Out

The model correctly classifies 798 out of 848 active compounds (94%) and 312 out of 400 non-active compounds (78%). Overall training predictability was 84.05% (1049 out of 1248 compounds). We validated the model by means of external predicting series, classifying the model 202 out 243, 83.13% of compounds (**see Table2**).

Table 2 Results of the model, analysis, validation and virtual-screening.

ANALYSIS			
	Percent	antibacterials	non-active
antibacterials	84.0	**199**	38
non-active	84.0	32	**168**
Total	84.0		
VALIDATION			
	Percent	antibacterials	non-active
antibacterials	83.8	**119**	23
non-active	82.2	18	**83**
Total	83.0		
VIRTUAL-SCREENING			
	Percent	antibacterials	non-active
antibacterials	84.5	**480**	88

In addition, we used a ROC curve (see **Figure 2**) to investigate the reability of the model, being the areas under curve equal to 0.86 for predicting series and 0.82 for training ones.

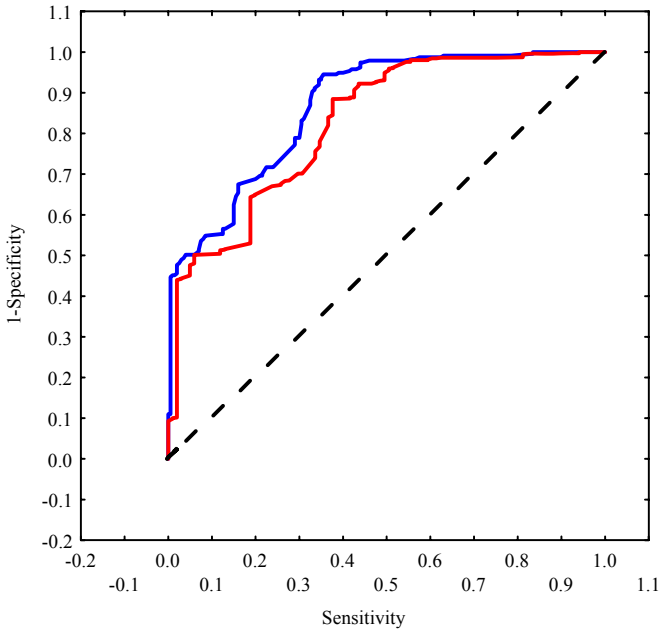

Figure 2. Results for the ROC curve.

It indicates that the present model give results statistically significant and clearly different from those obtained with a random classifier (area = 0.5). In order to show how to use the model in practice we carried out a virtual screening recognizing 480 out of 568 antibacterial compounds (84.5%). These compounds where never used in training or predicting series.

The more interesting characteristic of the present model is that the $^{k}C_s$ used as molecular descriptors depend both on the molecular structure of the drug and the bacterias species against the drug have to act. The codification of the molecular estructure is in first place due to the use of the adjacency factor α_{ij} to encode atom-atom bondig, molecular connectivity. The other aspect that allow encoding molecular structural changes is that the atomic contributions $^{0}c_j(s)$ are atom-class specific. Consequently, one change in the molecular structure of, e.g. F by O necessarily implies a change in the interaction. In any case, the more interesting fact is that $^{k}C_s$ are the first molecular descriptors reported for antimicrobial QSAR studies with the skill of discerning among a large number of bacterial species. This property is related to the definition of the $^{0}c_j(s)$. The values of these atomic contributions reported herein by the first time for antibacterial action are given in **Table 3** for some atoms and some selected species (email corresponding author for detailed compilation with more than 90 species).

Atomic contributions for antibacterial property can be ejecutate by the model, not only to distinguish different species (see **Table 3**), the model can be calculated the atomic contributions from different strains of the same species. One advantage of our model is to mark resistant strains of susceptible strains to a different drug. For instance, the **Table 3** shows the atomic contributions to antimicrobial action agaisnt susceptible and resistant strains of *Staphylococcus aureus* and *Staphylococcus epidermidis*. For the first of these two species, the regression coefficient between atomic contributions for resistant and susceptible strains is 0.51.

Conversely, the regression coefficient is 0.82 *Staphylococcus epidermidis*. This notable difference between both regression coefficients possibly refelects how large is the difference between the respective resitant and susceptible strains. In general, the atomic contributions of different atoms to the antibacterial property against all the studies species are connectied between them. The **Table 4** shows high regression coefficients for some of contributions.

Tabla4. Correlation values of atomic values.

	C	N	O	H	S	F	Cl
C	1.00	**0.95**	**0.96**	**0.98**	0.54	0.40	*0.33*
N		1.00	**0.97**	**0.96**	0.45	0.38	*0.23*
O			1.00	**0.98**	0.44	0.28	*0.33*
H				1.00	0.52	0.36	*0.32*
S					1.00	*0.26*	*0.43*
F						1.00	*-0.05*
Cl							1.00

Tabla 3. Some atomic contributions values for atom-receptor interactions.

Bacteria species	C	N	O	H	S	F	Cl
Acinetobacter baumanni	0.22	0.18	0.2	0.23	0.18	0	0
Bacteroides thetaiotaomicron	0.21	0.23	0.2	0.21	0.2	0.3	0
Clostridium perfringens	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Corynebacterium pseudodiphtheriticum	0.15	0.18	0.17	0.12	0.12	0	0.2
Chlamydia trachomatis	0.3	0.3	0.3	0.3	0	0.3	0
Citrobacter freundii	0.18	0.17	0.17	0.18	0.16	0.3	0
Clostridium difficile	0.25	0.25	0.25	0.25	0.17	0.27	0.2
Eikenella corrodens	0.15	0.21	0.12	0.12	0	0.3	0
Enterococcus faecium	0.24	0.22	0.24	0.24	0.19	0.26	0.3
Eubacterium lentum	0.22	0.2	0.21	0.22	0.22	0.3	0.3
Haemophilus influenzae	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Klebsiella oxytoca	0.24	0.23	0.23	0.24	0.22	0.3	0
Legionella pneumophila	0.3	0.3	0.3	0.3	0	0.3	0.3
Listeria monocytogenes	0.3	0.3	0.3	0.3	0.3	0	0.3
Mycobacterium avium	0.1	0.11	0.11	0.05	0	0.14	0
Micoplasma pneumoniae	0.3	0.3	0.3	0.3	0	0.3	0.3
Moraxella catarrhalis	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Morganella morganii	0.25	0.26	0.25	0.26	0.26	0.3	0
SAMR	0.14	0.15	0.14	0.13	0.08	0.22	0.2
SEMR	0.18	0.15	0.19	0.18	0.1	0	0.3
Staphylococcus aureus	0.24	0.25	0.24	0.24	0.23	0.31	0.2
SAMS	0.23	0.23	0.22	0.22	0.24	0.22	0
Staphylococcus epidermidis	0.29	0.27	0.28	0.29	0.25	0.3	0.3
SEMS	0.24	0.24	0.24	0.24	0.24	0.3	0

Please, email the corresponding author for details on the names of all the drugs used, the bacterias species tested, and detailed results for training and validation. The above-mentioned flexible definition of the present approach makes it possible to model by the first time the present very heterogeneous antibacterial activity data. In fact, the present is the first reported unify model that allow one predicting antibacterial activity of any organic compound against a very large diversity of bacterial pathogens. As a sort of concluding remark and future research outlook one may note that the present QSAR methodology may be able to predict biological activity of drugs in more general situacions than the tradional QSAR models may be.

Acknowledgments

Authors thank projects PXIB20304PR and BTF20302PR from Xunta de Galiza for partial financial support.

References and notes

1. Todeschini, R.; Consonni V. *Handbook of Molecular Descriptors*. Wiley VCH, Weinheim, Germany. 2000.
2. Fratev F, Benfenati E. *J Chem Inf Model.* **2005,** *45*, 634.
3. http//www.textbookofbacteriology.net
4. Kubinyi, H.; Taylor, J.; Ramdsen, C. Quantitative Drug Design. In, *Comprehensive Medicinal Chemistry*. Ed. C. Hansch. Pergamon. 1990, vol. 4, p. 589.
5. González, M. P.; Morales, A. H.; Molina R. *Polymer* **2004**, *45*, 2773.
6. Cabrera, M.A.; Bermejo, S. *Bioorg. Med .Chem.* **2004**. *22*, 5833.
7. Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez Y.; Romero-Zaldivar, V.; Castro, E.A. *Bioorg. Med. Chem.* **2005,** *13,* 2881.
8. Marrero-Ponce, Y.; Castillo-Garit, J.A.; Olazabal, E.; Serrano, H.S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sánchez, A.M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.,* **2005***, 13,* 1005.
9. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, M.; Mayón-Peréz, M.; García-Sánchez, R. N. *Bioorg. Med. Chem.* **2005,** *13,*1293.
10. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernádez, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de A., R. *J. Mol. Mod.* **2002**, *8*, 237.
11. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, 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. Mod.* **2003**, *9*, 395.
12. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
13. González-Díaz, H.; Ramos de A., R.; Molina, R. R. *Bull. Math. Biol.* **2003**, *65*, 991.
14. González-Díaz, H.; Uriarte, E.; Ramos de A. R. *Bioorg. Med. Chem.* **2004,** *13*, 323.
15. González-Díaz, H.; Molina, R. R.; Uriarte. E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691-4695.
16. González-Díaz, H.; Ramos de A., R.; Molina, R. R. *Bioinformatics* **2003**, *19*, 2079.
17. González-Díaz H, Molina RR, Uriarte E. *Polymer* **2004**, *45*, 3845.
18. Ramos de A., R.; González Díaz, H.; Molina, R.; González, M. P.; Uriarte, E. *Bioorg. Med. Chem.* **2004**, *12*, 4815.
19. Ramos de A, R.; González-Díaz, H.; Molina, R. R.; Uriarte, E. *Proteins, Struct. Func. and Bioinf.* **2004**, *56*, 715.
20. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de A., R. *Bull. Math. Biol.* **2004**, *66*, 1285.
21. González-Díaz, H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, I.; Nasco, O.; Uriarte, E.; Castañedo, N. C.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A; González, M. P. *Chem. Res. Tox.* **2003**, *16*, 1318.
22. González-Díaz, H.; Agüero, G., Cabrera, M.A., Molina, R, Santana, L., Uriarte, E., Delogu, G., Castañedo, N. *Bioorg. Med. Chem. Lett.* **2005**, *15,* 551.
23. González-Díaz, H.; Cruz-Monteagudo, M., Molina, R., Tenorio, E., Uriarte, E. *Bioorg. Med. Chem*. **2005**, *13,* 1119.
24. Cruz-Monteagudo, M.; González-Díaz, H. *Eur. J. Med. Chem.* **2005**, *doi:10.1016/j.ejmech.2005.04.012.*
25. Van Waterbeemd, H. Discriminant analysis for activity prediction. In, *Method and Principles in Medicinal Chemistry*, Ed, R. Manhnhold, Krogsgaard-Larsen, H. Timmerman, vol 2, Chemometric methods in molecular design. Ed, H. Van Waterbeemd, VCH, Weinhiem. 1995. pp 265-282.
26. Gonzalez-Diaz, H; Prado-Prado, F. J; Santana, Lourdes; Uriarte, Eugenio; *Bioorg. Med. Chem.* **2006,** 14, 5973.
27. STATISTICA for Windows release 6.0. Statsoft Inc., **2001**.