

Original article

Quantitative structure–activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors

Liane Saíz-Urra^a, Maykel Pérez González^{a,b,c,*}, Yagamare Fall^b, Generosa Gómez^b^a Chemical Bioactive Center, Central University of Las Villas, Santa Clara, Villa Clara, C.P. 54830, Cuba^b Department of Organic Chemistry, Vigo University, C.P. 36200, Vigo, Spain^c Service Unit, Experimental Sugar Cane Station “Villa Clara-Cienfuegos”, Ranchuelo, Villa Clara, C.P. 53100, Cuba

Received 20 February 2006; received in revised form 11 August 2006; accepted 14 August 2006

Available online 9 October 2006

Abstract

The GEometry, Topology, and Atom-Weights Assembly (GETAWAY) approach has been applied to the study of the HIV-1 integrase inhibition of 172 compounds that belong to 11 different chemistry families. A model able to describe more than 68.5% of the variance in the experimental activity was developed with the use of the mentioned approach. In contrast, none of the five different approaches, including the use of Randić Molecular Profiles, Geometrical, RDF, 3D-MORSE and WHIM descriptors was able to explain more than 62.4% of the variance in the mentioned property with the same number of variables in the equation. Finally, after extracting five compounds considered by us as outliers the model was able to describe more than 72.5% of the variance in the experimental activity.

© 2006 Elsevier Masson SAS. All rights reserved.

Keywords: QSAR; HIV-1 integrase inhibitors; GETAWAY descriptors

1. Introduction

The human immunodeficiency virus (HIV) is the causative agent of the acquired immunodeficiency syndrome (AIDS). An estimated 36 million people worldwide are currently living with HIV, and some 20 million people having already died, giving a cumulative total number of HIV infections to be 56 million [1], however, there is still no known cure or vaccination against it.

There are three viral enzymes encoded by the pool gene of HIV-1, namely reverse transcriptase (RT), protease (PR), and integrase (IN). Currently, the combination of RT and PR inhibitors, although highly effective, produces unwanted side effects like toxicity and patient adherence, among others [2].

HIV-1 IN functions in a two-step manner by initially removing a dinucleotide unit from the 3'-ends of the viral

DNA (termed 3'-processing). The 3'-processed strands are then transferred from the cytoplasm to the nucleus where they are introduced into the host DNA following 5 base-pair offset cleavages of opposing host strands (termed 3'-strand transfer or end joining). There is obviously a requirement for a functional integrase in HIV-1 replication and has no cellular homologue [3]. This is one of the reasons why IN represents an attractive and validated target for chemotherapeutic intervention and has become a focus of anti-AIDS drug design efforts.

Through the years QSAR studies about activity of the inhibition of the HIV-1 integrase or the synthesis of new compounds with this activity have been carried out. Different descriptors and variable selection techniques were used to develop models capable of relating the activity to the structure.

For example, Aiello et al. [4] carried out QSAR studies on novel thiazolothiazepine based HIV-1 integrase inhibitors. In this work, in an attempt to establish a coherent structure–activity relationship amongst thiazolothiazepines they investigated whether the sulfur atom could be replaced by an oxygen and whether a naphthalene ring could be replaced by fused

* Corresponding author. Service Unit, Experimental Sugar Cane Station “Villa Clara-Cienfuegos”, Ranchuelo, Villa Clara, C. P. 53100, Cuba.

E-mail address: mpgonzalez76@yahoo.es (M.P. González).

rings and employed structure-based optimization methods by determining all the amino acid residues on the active site of IN that are important for ligands' binding. In order to identify the biologically active conformation of the compounds, they docked all the compounds onto the active site of IN using GOLD.

Another interesting work that employs docking technique is the work of Dayam and Neamati about active site binding modes of the β -diketoacids like a multi-active site approach in HIV-1 integrase inhibitor design [5].

Makhija et al. [6] investigated about design and synthesis of HIV-1 integrase inhibitors for 3'-processing and 3'-strand transfer activities. They performed a 3D-QSAR through comparative molecular field analysis (CoMFA) for different classes of compounds using a novel alignment technique based on molecular electrostatic potentials (MEPs) [3,7–9]. Partial least square (PLS) fitting was applied to derive the 3D-QSAR models.

Costi et al. investigated about 2,6-bis(3,4,5-trihydroxybenzylidene) derivatives of cyclohexanone like novel potent HIV-1 integrase inhibitors that prevent HIV-1 multiplication in cell-based assays. In order to develop a model capable of predicting the anti-HIV-IN activity and to design useful novel derivatives they performed a comparative molecular field analysis (CoMFA) like 3D-QSAR [10].

On the other hand, Ma et al. performed a comparative molecular field analysis (CoMFA) like 3D-QSAR and docking studies with the objective of understanding pharmacophore properties of styrylquinoline derivatives and to design inhibitors of HIV-1 integrase [11].

Makhija and Kulkarni performed a QSAR study using genetic function approximation (GFA) technique to examine the correlations between the calculated physicochemical descriptors and the in vitro activities, 3'-processing and 3'-strand transfer inhibition of a series of human immunodeficiency virus type 1 (HIV-1) integrase inhibitors [3].

In the literature no reports on GEometry, Topology, and Atom-Weights Assembly (GETAWAY) descriptors for QSAR studies about HIV-1 integrase inhibitors can be found, which is why we have, for the first time, applied it to model this biological property using these descriptors and make a comparison with other approaches and 3D-descriptors.

2. Materials and methods

2.1. Data set

In the present study we used a data set of 172 HIV-1 integrase inhibitors for which the IC_{50} for 3'-processing were reported. These compounds are shown in the literature through several works [12–22]. Among this data set are included 10 tyrphostins [12], 28 coumarins [13], 15 aromatic sulfonamides [14], 19 chicoric acids [15], eight tetracyclines [16], seven arylamides and naphthalene-based compounds [17], 20 thiazolo-thiazepines [18], six curcumins [19], 20 salicylhydrazines [20], 10 styrylquinolines [21] and 11 depsides and depsidones [22]. Determination of IC_{50} values was achieved by plotting

drug concentration versus percentage of inhibition and by measuring the concentration at which 50% inhibition occurred. In order to guarantee the linear distribution of the dependent variable we calculated the natural logarithm of the IC_{50} values and obtained the model using these values. The structures of compounds and their biological activities are provided as [Supplementary data](#).

2.2. Molecular descriptors and geometry optimization

The Dragon computer software [23] was employed to calculate the molecular descriptors. Besides, we carried out a preliminary MM+ geometry optimization calculations for each compound of this study, and then using the quantum chemical semi-empirical method AM1 [24] included in MOPAC 6.0 [25] determined the (x,y,z)-atomic coordinates of the minimal energy conformations for each one. Six models were developed using the descriptors generated with the Dragon computer software such as GETAWAY, the Randić Molecular Profiles, Geometrical, WHIM, RDF and 3D-MORSE descriptors [26]. Descriptors with constant values inside each group of descriptors were discarded. Pairwise correlation analysis was performed for the remaining descriptors. This descriptors exclusion method was used to reduce the collinearity and correlation between them.

2.3. Statistical methods

The statistical processing to obtain the QSAR models was carried out by using Genetic Algorithm (GA) analysis using the Statistic 6.0 software [27]. The first step is to create a population of linear regression models. These regression models mate with each other, mutate, crossover, reproduce, and then evolve through successive generations towards an optimum solution. The GA simulation conditions were 10 000 generations, number of crossovers were 5000, smoothness factor was 1, mutation probability for adding new term was 50% and 300 model populations. The GA procedure was repeated n -times to confirm that the selected descriptors are the most optimal descriptor set for describing the biological property. Analysis of residuals was done and deleted residuals from the regression equations were used to identify outliers. The statistical significance of the models was determined by examining the regression coefficient, the standard deviation, the number of variables, and the proportion between the cases and variables in the equation.

2.4. Validation of the models

To validate a QSAR model, most of the researchers apply the leave-one-out (LOO) or leave-group-out (LGO) cross-validation procedures. Frequently, q^2 is used as a criterion of both robustness and predictive ability of the model. Many authors consider high q^2 (for instance, $q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of the QSAR model. They do not test the models for their ability to predict the activity of compounds of an external test set (i.e.

compounds, which have not been used in the QSAR model development). Although it is important to highlight that some QSAR studies are carried out with a reduced data set with the aim to study one or some specific characteristics of their biological activity. That is why the selection of the test set is very difficult and even impossible [28–31].

Nevertheless, several authors, including us, have suggested that when the number of compounds available is enough, a good way to estimate the true predictive power of a QSAR model is to compare the predicted and observed activities of an (sufficiently large) external test set of compounds that were not used in the development of the model [32–37]. In this connection, here we used an external predicted set for validating the models obtained.

In order to obtain validated QSAR models the data set should be divided into training and test sets. Ideally, this division must be performed such that points representing both training and test sets are distributed within the whole descriptor space occupied by the entire data set, and each point of the test set is close to at least one point of the training set. This approach ensures that a similar principle can be employed for the activity prediction of the test set. For this reason, using k-Means Cluster Analysis (k-MCA) for splitting the set of compounds we ensure the validity of the model.

2.5. k-Means Cluster Analysis

The k-MCA may be used in training and test sets' design[38]. The idea is to carry out a partition of the set of compounds under study into several statistically representative classes of chemicals. Then, one may select the training and test sets from the members of all these classes. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both compound series (training and test). It permits to design both training and test sets, which are representative of the entire “experimental universe”. Fig. 1 illustrates graphically the above-described procedure.

The k-MCA splits the chemical compounds into four clusters with 57, 28, 51 and 36 members and standard deviations of 0.08, 0.11, 0.23 and 0.25, respectively. Selection of the test set was carried out by taking the compounds with the minor Euclidean distance belonging to each cluster to guarantee the representative of the biological data in this set. We took

into account the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible) to ensure a statistically acceptable data partition into several clusters. We also made an inspection among and within clusters where the respective Fisher's ratio and their *p*-level of significance were considered to be lower than 0.05. The variables, which were finally, used in the analysis showed *p*-levels < 0.05 for Fisher's test. The results are depicted in Table 1.

The main conclusion should be achieved from the k-MCA: the structural diversity of several up-to-date known compounds (as codified by the descriptors) may be described at the least by four statistically homogeneous clusters of chemicals. Anyhow, further conclusions about the mechanistic and molecular significance of these clusters seem to be speculative. Mainly, if it is considered that k-MCA based partitions of data, which consider not only four but also five or six clusters are statistically significant too. However, the use of the k-MCA analysis points out a structurally representative distribution of chemicals into training and predicting series as shown in the following dendrogram (Fig. 2).

2.6. Comparison with other approaches

The use of GETAWAY descriptors for the prediction of HIV-1 integrase inhibition was compared with other methodologies (Randić Molecular Profiles, Geometrical, RDF, 3D-MORSE and WHIM). The development of these five models involved the use of the same data set that was used in developing the model of GETAWAY descriptors. The first comparison was based on the quality of the statistical parameters of the regression as well as the predictive capability of the models generated and the second one was based on the same number of variables in the models for demonstrating the superiority of our approximation.

In addition, further criteria exist to compare the quality of the models obtained due to the correlation coefficient *R* which is clearly the worst criterion because it tends to select as many variables as possible, the standard deviation *S* value tends to include too many variables and the *F* value sometimes selects fewer variables than usually accepted by a QSAR practitioner because large *F* values are often achieved by including only one or two variables in the model.

Table 1

Main results of the k-Means Cluster Analysis for the compounds included in the current study

Variance analysis				
Descriptors	Between SS ^a	Within SS ^b	Fisher's ratio (<i>F</i>)	<i>p</i> -Level ^c
H2u	54.11	10.14	298.72	10 ^{−7}
H5v	9.91	2.19	253.78	10 ^{−7}
H7e	34.09	8.70	219.45	10 ^{−7}
R1u	3.28	1.60	114.64	10 ^{−7}

^a Variability between groups.

^b Variability within groups.

^c Level of significance.

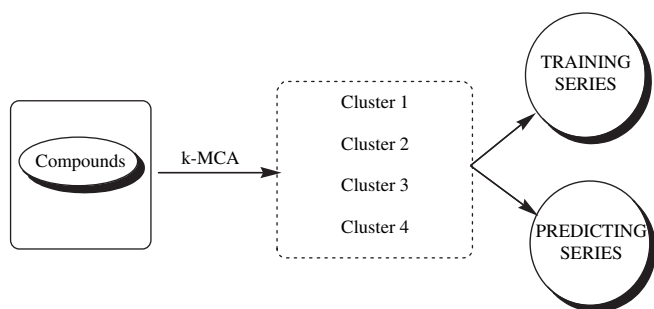


Fig. 1. Training and predicting series design through k-MCA.

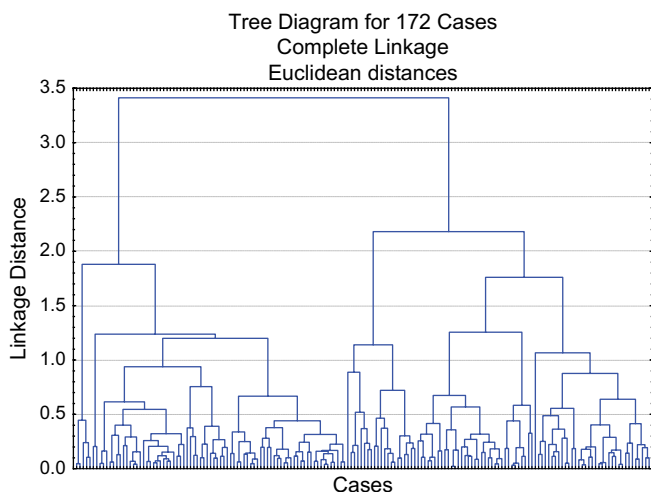


Fig. 2. Tree diagram for 172 compounds used in this study according to the variables selected for the analysis.

One of these criteria is the FIT Kubinyi function (Eq. (1)), being closely related to the F value, which was created and proved to be useful [39,40].

$$\text{FIT} = \frac{R^2(n-k-1)}{(n+k^2)(1-R^2)} \quad (1)$$

where n is the number of compounds in the training set and k is the number of variables in the equation.

The main disadvantage of the F value is its sensitivity to changes in k , if k is small, and its lower sensitivity if k is large. The FIT criterion has a low sensitivity towards changes in k values, as long as they are small numbers, and a substantially increasing sensitivity for large k values [39,40]. The best model will be the present one due to the high value of this function.

Other of these criteria was formulated by Akaike sometime ago [41,42]. Akaike's information criteria (AIC) take into account the statistical goodness of fit and the number of parameters that have to be estimated to achieve that degree of fit. This criterion is calculated using the following equation:

$$\text{AIC} = \text{RSS} \frac{(n+p')}{(n-p')^2} \quad (2)$$

where RSS is the sum of squared differences between the observed (y) and estimated response (\hat{y}); n is the number of compounds in the training set; and p' is the number of adjustable parameters in the model. When comparing the models, the model that produces the minimum value of these statistics should be considered potentially the most useful. We calculated the Akaike and FIT Kubinyi function values for the five approaches and included them in the comparison.

3. Results and discussion

In this work, the model selection was subjected to the principle of parsimony [43]. Then, we chose a function with

higher statistical significance but having as few parameters as possible for every type of descriptor. The best GETAWAY QSAR models with eight, nine and 10 variables are reported in Table 2 with the aim to choose the best among them.

According to the statistical results of the models we chose the model with nine variables since this one showed high FIT and the other statistical parameters do not change significantly with the addition of a new descriptor in the equation. In this connection, the addition of the descriptor number 10 in the equation was not justified from the statistical point of view. The equation with nine variables is given below together with the statistical parameters of the regression:

$$\begin{aligned} -\log(\text{IC}_{50}) = & 5.44 \times \text{H5v} - 13.74 \times \text{R5p}^+ + 39.06 \times \text{R7v}^+ \\ & - 1.67 \times \text{H2u} - 19.23 \times \text{HATS1v} - 1.55 \\ & \times \text{H7e} + 2.84 \times \text{R1u} - 2.17 \times \text{R7m}^+ + 14.46 \\ & \times \text{R2v}^+ - 3.51 \end{aligned} \quad (3)$$

$N = 139$, $R^2 = 0.688$, $S = 0.520$, $F(9,129) = 31.549$, $p < 10^{-7}$, $\text{AIC} = 0.312$, $\text{FIT} = 1.293$

where N is the number of compounds included in the model, R^2 is the square of correlation coefficient, S is the standard deviation of the regression, F is the Fisher's ratio, p is the significance of the variables in the model, AIC is the Akaike information criteria and FIT is the Kubinyi function.

On the other hand, the results obtained with the use of the other kinds of descriptors are given in Table 3.

As can be seen, there are remarkable differences concerning the explanation of the experimental variance given by all the models reported here compared to the GETAWAY one. While the GETAWAY QSAR model explains more than 68% of the activity, the rest of the models are unable to explain more than 61.5% of such variance; all these models also have important statistic parameters of a lower quality compared to the GETAWAY approach, such as the Fisher's ratio (F), the standard deviation (S), the Akaike information criteria (AIC) and the Kubinyi function (FIT).

On the other hand, all models require validation i.e., they can be used to make predictions. If a QSAR cannot be used to make predictions, then it is of no practical use. Statistical fit should not be confused with the ability of a model to make predictions.

In this sense also, the GETAWAY descriptors present the best R^2 (0.669) and S (0.529) predicting the external test set

Table 2

The statistical parameters of the linear regression models obtained for eight, nine and 10 variables for the GETAWAY descriptors

Variables	R^2	S	F	$p <$	AIC	FIT
8	0.642	0.554	29.131	10^{-7}	0.350	1.148
9	0.688	0.520	31.549	10^{-7}	0.312	1.293
10	0.696	0.516	29.337	10^{-7}	0.310	1.226

Table 3

The statistical parameters of the linear regression models obtained for the six kinds of descriptors

Descriptors	Variables	R^2	S	F	AIC	FIT	$R^2_{\text{Ext. Val.}}$	$S_{\text{Ext. Val.}}$
Randić Molecular Profiles	9	0.493	0.662	13.931	0.507	0.570	0.388	0.789
Geometrical	9	0.570	0.610	18.996	0.430	0.777	0.574	0.626
RDF	9	0.624	0.570	23.809	0.376	0.973	0.352	0.840
3D-MORSE	9	0.621	0.573	23.494	0.379	0.961	0.434	0.737
WHIM	9	0.612	0.579	22.649	0.387	0.925	0.444	0.721
GETAWAY	9	0.688	0.520	31.549	0.312	1.293	0.669	0.529

regarding the better predictive power of the methodologies used. For these reasons, we considered that these descriptors can be very useful tools for the prediction of HIV-1 integrase inhibition.

Once we have demonstrated the superiority of the GETAWAY descriptors to other methodologies in this test set for this biological property, we will consider the outliers present in Eq. (3).

A step-by-step outlier extraction procedure led to different models with better statistical profiles. In this study the outliers' numbers were continuously extracted from 0 to 3, considering that a number of outliers lower 10% of the general data are classically accepted in the literature as threshold [44]. In our case the higher extracted outlier number represented a 2.16% of the whole data. The structure of these outliers is shown below and the new statistic parameters for each successive extraction in Table 4.

The following equation was obtained without the outliers.

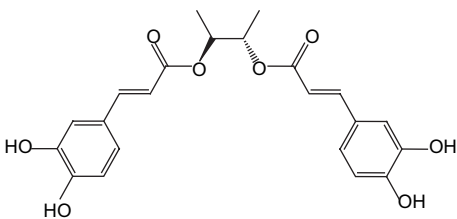
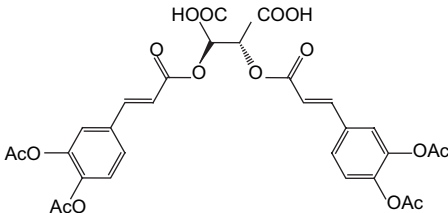
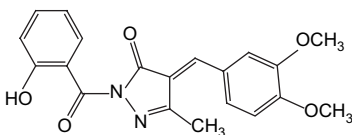
$$\begin{aligned}
 -\log(\text{IC}_{50}) = & 5.69 \times \text{H5v} - 15.49 \times \text{R5p}^+ + 37.66 \times \text{R7v}^+ \\
 & - 1.68 \times \text{H2u} - 18.02 \times \text{HATS1v} - 1.60 \\
 & \times \text{H7e} + 2.81 \times \text{R1u} - 1.90 \times \text{R7m}^+ + 14.99 \\
 & \times \text{R2v}^+ - 3.50
 \end{aligned} \quad (4)$$

$$\begin{aligned}
 N = 136, R^2 = 0.730, S = 0.484, F(9, 126) = 37.781, \\
 p < 10^{-7}, \text{AIC} = 0.301, \text{FIT} = 1.338, R^2_{\text{Ext. Val.}} = 0.724, \\
 S_{\text{Ext. Val.}} = 0.480
 \end{aligned}$$

The variables in the model (Eq. (4)) encoded specific structure information. As can be seen the variables in this model are related to the polarizabilities, van der Waals volumes, electro-negativities and the atomic mass.

Table 4

Structures of the outliers and statistic parameters of the new equations found

Compounds	Structure	R^2	S	F	AIC	FIT
(58)		0.703	0.510	33.316	0.301	1.338
(66)		0.714	0.501	35.179	0.293	1.454
(140)		0.730	0.484	37.781	0.301	1.338

In this equation, there are four variables with a positive influence in the studied biological property and five variables with negative ones.

Based on this classification we can observe how the weighted by atomic van der Waals volumes have a positive influence on both types of descriptors H-GETAWAY and R-GETAWAY not being in that same way for the weighted by atomic Sanderson electronegativities.

From these facts we may think that drugs with high van der Waals volume values might be inhibitors of the 3'-processing, if we take into account only the possible interactions between atoms at the topological distance, in the molecules. Furthermore the largest values derive from the external atoms and simultaneously next to each other in the molecular space.

A previous study has implied that salicylhydrazines inhibit HIV-1 integrase by chelating to the metal at the active site as they are active only when Mn^{2+} is used as a cofactor [20]. However, thiazolothiazepines showed equal activities in the presence of Mg^{2+} or Mn^{2+} , thus indicating that they differ from salicylhydrazines and perhaps act at a different site on HIV-1 integrase [18]. The aromatic moiety common to many inhibitors has been proposed to interact with the divalent cation in a "cation- π " type interaction [14]. There is also a possibility of a typical charge-charge interaction between the metal ions and ionic or partial charges of the ligands [14,20]. It has been shown that both types of interactions can co-exist in a binding site [45]. We can think that electron withdrawing substituents on the aromatic ring exert an unfavourable effect in relation with the electronegativity. It is expected that with an increase of the electronegativity of the substituents, the $-I$ inductive effect increases too and this could impoverish the aromatic ring as electrons make the interaction between the aromatic ring and the cation less probable. It is very important to take into account the resonance because there are many compounds that inhibit this enzyme and have many OH as substituents, these compounds have a $-I$ inductive effect. Previous studies suggest that these inhibitors could block the reaction through inhibiting the glycerolysis, hydrolysis, and circular nucleotide formation that are involved in the 3'-processing step [12].

On the other hand, as can be seen our model explains more than 72.5% of activity after the extraction of the outliers. In spite of using better statistic parameters, the models are not sufficiently good to explain the property. This fact can be explained by the complexity of the enzyme as we have shown above. Furthermore previous crystallographic studies on three HIV and ASV integrase inhibitors occupying three distinct binding sites in the two enzymes indicated the possibility of more than one binding site in HIV-1 integrase [46–48]. These results provide support for the possibility that structurally different inhibitors interact at different sites. Maybe because of that our model does not explain successfully the activity in question. Another aspect to take into account may be the presence of compounds' target in different moieties of the protein. Work towards the theoretical solution of this prediction problem is currently under way in our laboratories.

4. Conclusions

In this work we model the IC_{50} for 3'-processing taking into account GETAWAY descriptors and other approaches and 3D-descriptors and reflect some conclusions.

The best model obtained was using the GETAWAY descriptors, explaining more than 72.5% of activity after the extraction of the outliers, the other 3D-descriptors are less successful than the GETAWAY descriptors. The variables in this model are related to the polarizabilities, van der Waals volumes, electronegativities and the atomic mass. The most important variables are related to the van der Waals volumes and electronegativities, favourable being, for this activity, high van der Waals volume values, and high electronegativity values being unfavourable for the substituents taking into account that the aromatic moiety common to many inhibitors has been proposed to interact with the divalent cation in a "cation- π " type interaction and there is also a possibility of a typical charge-charge interaction between the metal ions and ionic or partial charges of the ligands. It is very important to take into account the presence of compounds' target in different moieties of the protein for a better analysis of the models of this activity.

Acknowledgements

Maykel Pérez González thanks the Universidad de Vigo for its kind hospitality. We thank the Xunta de Galicia (PGIDT04BTF301031PR) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at [doi:10.1016/j.ejmech.2006.08.005](https://doi.org/10.1016/j.ejmech.2006.08.005).

References

- [1] J. Stover, S. Bertozzi, J.P. Gutierrez, N. Walker, K.A. Stanecki, R. Greener, E. Gouws, C. Hankins, G.P. Garnett, J.A. Salomon, et al., *Science* (2006).
- [2] D.D. Richman, *Nature* 410 (2001) 995–1001.
- [3] M.T. Makhija, V.M. Kulkarni, *Bioorg. Med. Chem.* 10 (2002) 1483–1497.
- [4] F. Aiello, A. Brizzi, A. Garofalo, F. Grande, G. Ragno, R. Dayam, N. Neamati, *Bioorg. Med. Chem.* 12 (2004) 4459–4466.
- [5] R. Dayam, N. Neamati, *Bioorg. Med. Chem.* 12 (2004) 6371–6381.
- [6] M.T. Makhija, R.T. Kasliwal, V.M. Kulkarni, N. Neamati, *Bioorg. Med. Chem.* 12 (2004) 2317–2333.
- [7] M.T. Makhija, V.M. Kulkarni, *J. Comput. Aided Mol. Des.* 15 (2001) 961–978.
- [8] M.T. Makhija, V.M. Kulkarni, *J. Chem. Inf. Comput. Sci.* 41 (2001) 1569–1577.
- [9] M.T. Makhija, V.M. Kulkarni, *J. Comput. Aided Mol. Des.* 16 (2002) 181–200.
- [10] R. Costi, R.D. Santo, M. Artico, S. Massa, R. Ragno, R. Loddo, M. La Colla, E. Tramontano, P. La Colla, A. Pani, *Bioorg. Med. Chem.* 12 (2004) 199–215.

- [11] X.H. Ma, X.Y. Zhang, J.J. Tan, W.Z. Chen, C.X. Wang, *Acta Pharmacol. Sin.* 25 (2004) 950–958.
- [12] A. Mazumder, A. Gazit, A. Levitzki, M. Nicklaus, J. Yung, G. Kohlhausen, Y. Pommier, *Biochemistry* 34 (1995) 15111–15122.
- [13] H. Zhao, N. Neamati, H. Hong, A. Mazumder, S. Wang, S. Sunder, G.W. Milne, Y. Pommier, T.R. Burke Jr., *J. Med. Chem.* 40 (1997) 242–249.
- [14] M.C. Nicklaus, N. Neamati, H. Hong, A. Mazumder, S. Sunder, J. Chen, G.W. Milne, Y. Pommier, *J. Med. Chem.* 40 (1997) 920–929.
- [15] Z. Lin, N. Neamati, H. Zhao, Y. Kiryu, J.A. Turpin, C. Aberham, K. Strebel, K. Kohn, M. Witvrouw, C. Pannecouque, et al., *J. Med. Chem.* 42 (1999) 1401–1414.
- [16] N. Neamati, H. Hong, S. Sunder, G.W. Milne, Y. Pommier, *Mol. Pharmacol.* 52 (1997) 1041–1055.
- [17] H. Zhao, N. Neamati, A. Mazumder, S. Sunder, Y. Pommier, T.R. Burke Jr., *J. Med. Chem.* 40 (1997) 1186–1194.
- [18] N. Neamati, J.A. Turpin, H.E. Winslow, J.L. Christensen, K. Williamson, A. Orr, W.G. Rice, Y. Pommier, A. Garofalo, A. Brizzi, et al., *J. Med. Chem.* 42 (1999) 3334–3341.
- [19] A. Mazumder, N. Neamati, S. Sunder, J. Schulz, H. Pertz, E. Eich, Y. Pommier, *J. Med. Chem.* 40 (1997) 3057–3063.
- [20] N. Neamati, H. Hong, J.M. Owen, S. Sunder, H.E. Winslow, J.L. Christensen, H. Zhao, T.R. Burke Jr., G.W. Milne, Y. Pommier, *J. Med. Chem.* 41 (1998) 3202–3209.
- [21] F. Zouhiri, J.F. Mouscadet, K. Mekouar, D. Desmaele, D. Savoure, H. Leh, F. Subra, M. Le Bret, C. Auclair, J. d'Angelo, *J. Med. Chem.* 43 (2000) 1533–1540.
- [22] N. Neamati, H. Hong, A. Mazumder, S. Wang, S. Sunder, M.C. Nicklaus, G.W. Milne, B. Proksa, Y. Pommier, *J. Med. Chem.* 40 (1997) 942–951.
- [23] R. Todeschini, V. Consonni, M. Pavan version 2.1, Dragon Software (2002).
- [24] M.J.S. Dewar, E.G. Zebisch, E.F. Healy, J.J.P. Stewart, *J. Am. Chem. Soc.* 107 (1985) 3902–3909.
- [25] J. Frank, MOPAC version 6.0, Seiler Research Laboratory, US Air Force Academy, Colorado Springs CO, 1993.
- [26] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, first ed. Wiley-VCH, Mannheim, 2000.
- [27] Statsoft Inc., STATISTICA (Data Analysis Software System) version 6.0, Statsoft Inc., 2002.
- [28] M.P. González, M. Moldes del Carmen Teran, *Bioorg. Med. Chem. Lett.* 14 (2004) 3077–3079.
- [29] M.P. Gonzalez, C. Teran, M. Teijeira, *Bioorg. Med. Chem. Lett.* in press.
- [30] M.P. González, C. Teran, M. Teijeira, M.J. Gonzalez-Moa, *Eur. J. Med. Chem.* 40 (2005) 1080–1086.
- [31] M.P. González, C. Teran Moldes Mdel, *Bull. Math. Biol.* 66 (2004) 907–920.
- [32] M.P. González, L.C. Dias, A.M. Helguera, Y.M. Rodriguez, L.G. de Oliveira, L.T. Gomez, H.G. Diaz, *Bioorg. Med. Chem.* 12 (2004) 4467–4475.
- [33] M.P. González, H. Gonzalez Diaz, R. Molina Ruiz, M.A. Cabrera, R. Ramos de Armas, *J. Chem. Inf. Comput. Sci.* 43 (2003) 1192–1199.
- [34] A.H. Morales, M.A. Cabrera Perez, M.P. González, R.M. Ruiz, H. Gonzalez-Diaz, *Bioorg. Med. Chem.* 13 (2005) 2477–2488.
- [35] A. Golbraikh, M. Shen, Z. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* 17 (2003) 241–253.
- [36] A. Golbraikh, A. Tropsha, *J. Mol. Graphic. Model.* 20 (2002) 269–276.
- [37] A. Golbraikh, A. Tropsha, *J. Comput. Aided Mol. Des.* 16 (2002) 357–369.
- [38] W.R. Dillon, M. Goldstein, *Multivariate Analysis: Methods and Applications*, Wiley, New York, 1984.
- [39] H. Kubinyi, *Quant. Struct.-Act. Relat.* 13 (1994) 393–401.
- [40] H. Kubinyi, *Quant. Struct.-Act. Relat.* 13 (1994) 285–294.
- [41] H. Akaike, *IEEE Trans. Autom. Control* AC-19 (1974) 713–716.
- [42] H. Akaike, : Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csaki (Eds.), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 1973, pp. 267–281.
- [43] D.M. Hawkins, *J. Chem. Inf. Comput. Sci.* 44 (2004) 1–12.
- [44] R.L. Lipnick, *Sci. Total Environ.*(109–110) (1991) 131–153.
- [45] D.A. Dougherty, *Science* 271 (1996) 163–168.
- [46] Y. Goldgur, R. Craigie, G.H. Cohen, T. Fujiwara, T. Yoshinaga, T. Fujishita, H. Sugimoto, T. Endo, H. Murai, D.R. Davies, *Proc. Natl. Acad. Sci. U.S.A.*(96) (1999) 13040–13043.
- [47] J. Lubkowski, F. Yang, J. Alexandratos, A. Wlodawer, H. Zhao, T.R. Burke Jr., N. Neamati, Y. Pommier, G. Merkel, A.M. Skalka, *Proc. Natl. Acad. Sci. U.S.A.*(95) (1998) 4831–4836.
- [48] V. Molteni, J. Greenwald, D. Rhodes, Y. Hwang, W. Kwiatkowski, F.D. Bushman, J.S. Siegel, S. Choe, *Acta Crystallogr., Sect D: Biol. Crystallogr.* 57 (2001) 536–544.