European Journal of Medicinal Chemistry 44 (2009) 3577–3583

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

Multivariate QSAR study of 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase inhibitors

Eduardo Borges de Melo [a,b,*], Márcia Miguel Castro Ferreira [b]

[a] Curso de Farmácia, Centro de Ciências Médicas e Farmacêuticas, Universidade Estadual do Oeste do Paraná, Rua Universitária 2069, 85819-110 Cascavel, PR, Brazil
[b] Laboratório de Quimiometria Teórica e Aplicada[1], Instituto de Química, Departamento de Físico-Química, Universidade Estadual de Campinas, Campinas, SP, Brazil

ARTICLE INFO

Article history:
Received 23 January 2009
Accepted 2 March 2009
Available online 9 March 2009

Keywords:
QSAR
4,5-Dihydroxypyrimidine carboxamides
HIV-1 integrase
OPS
PLS

ABSTRACT

A multivariate QSAR study of thirty-three 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase (HIV-1 IN) inhibitors was performed employing Ordered Predictors Selection (OPS) algorithm and PLS regression for variable selection and model construction, respectively. Four descriptors were chosen and a reasonable model ($n = 30$; $R^2 = 0.68$; SEC = 0.57; $PRESS_{cal} = 8.72$; $F_{(2,27)} = 28.97$; $Q^2_{LOO} = 0.58$; SEV = 0.62; $PRESS_{val} = 11.62$; $R^2_{pred} = 0.87$; SEP = 0.29; $ARE_{pred} = 4.37\%$; $k = 0.99$; $k' = 1.01$; $|r^2_0 - r^2_0'| = -0.18$) was built with two latent variables (59.54% of the information). Leave-*N*-out (LNO) and *Y*-randomization methods confirmed the model robustness. The descriptors indicated that the HIV-1 IN inhibition depends on the electronic distribution of the investigated compounds. The interpretation of the model is related to the most accepted mechanism of action.

1. Introduction

Human immunodeficiency virus (HIV), a retrovirus, is the primary cause of AIDS (acquired immunodeficiency syndrome), and one of the main medical and social problems nowadays. In 2007, approximately 33 million people in the world are infected by HIV, and the number of deaths so far totaled about two million. Seventy-two percent of AIDS deaths occurred in the Sub-Saharan Africa [1]. In the last two decades, approximately twenty anti-HIV drugs have been developed, but most of them act against the viral enzymes HIV-reverse transcriptase and HIV-protease [2].

In spite of the development of the highly active anti-retroviral therapy (HAART) [3], there is an emergent need to search for new anti-HIV agents. The main reasons are the serious adverse side effects of the available drugs and the emergence of drug-resistance (including cross-resistance) [4]. Attention has been given to the development of drugs that act on new targets, such as the host protein cell [5–8] and other viral structures, such as the enzyme HIV-1 integrase (HIV-1 IN) [9,10].

The HIV-1 IN is, actually, a major breakthrough in AIDS research. However, despite studies in this field, only recently the first inhibitor of HIV-1 IN, raltegravir (Isentress®; Merck Co.), has been approved by the FDA [11]. This drug, an *N*-Me pyrimidone, is a derivative of the 4,5-dihydroxypyrimidine carboxamides (Fig. 1). Studies have indicated that it is well tolerated, and has not shown serious drug-related adverse events [12–14].

HIV-1 IN displays a conserved catalytic triad of metal-coordinating carboxylates, which catalyzes two reactions: the 3′-processing (3′P) that occurs in the cellular cytoplasm and processes the retrotranscribed viral cDNA, and the strand transfer reaction (ST), which catalyzes the initial joining of the processed 3′-ends to the 5′-ends of the host-cell DNA [15,16]. The raltegravir and the 4,5-dihydroxypyrimidine carboxamide derivatives inhibit the ST reaction and are classified as integrase strand transfer inhibitors (INSTIs) [15].

Quantitative structure–activity relationship (QSAR) describes how a given biological activity can vary as a function of molecular descriptors derived from the chemical structure of a set of molecules. Thus, a model containing those calculated descriptors can be used to predict responses of new compounds [17,18]. Only a few studies involving computer aided-drug design (CADD) of HIV-1 IN inhibitors were performed employing the 2D-QSAR approach

Abbreviations: 3′P, 3′-processing; 5CITEP, 1-(5-chloroindol-3-yl)-3-hydroxy-3-(2*H*-tetrazol-5-yl)-propenone; AIDS, acquired immunodeficiency syndrome; AM1, Austin Model 1; ARE, average relative error; B3LYP, Becke, three-parameter, Lee–Yang–Parr; cDNA, complementary deoxyribonucleic acid; DFT, density functional theory; DKA, diketo acid; FDA, Food and Drug Administration; HAART, highly active anti-retroviral therapy; HF, Hartree–Fock; HIV, human immunodeficiency virus; HOMO, highest occupied molecular orbital; IN, integrase; INSTI, integrase strand transfer inhibitors; OPS, ordered predictors selection; PDB, Protein Data Bank; PLS, partial least squares; QSAR, quantitative structure–activity relationship.

* Corresponding author. Curso de Farmácia, Centro de Ciências Médicas e Farmacêuticas, Universidade Estadual do Oeste do Paranà, Rua Universitária 2069, 85819-110 Cascavel, PR, Brazil. Tel.: +55 45 3320 3156; fax: +55 45 3320 3290.
E-mail address: ebmelo@unioeste.br (E.B. de Melo).
[1] http://lqta.iqm.unicamp.br

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.03.001

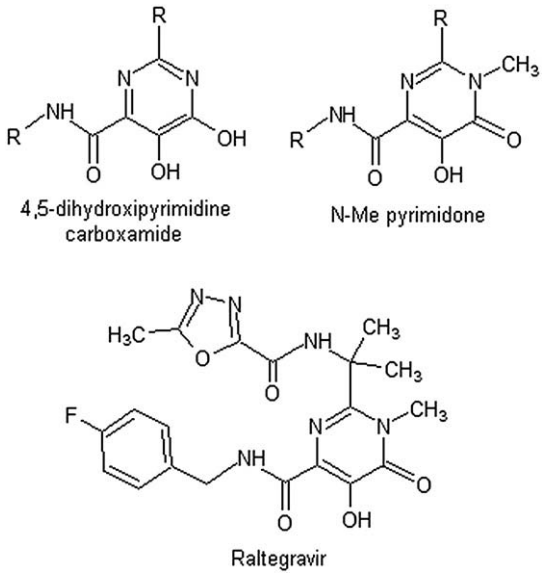

Fig. 1. Basic structures of 4,5-dihydroxypyrimidine carboxamides, *N*-Me pyrimidone and the structure of raltegravir. Adapted from references [11,12].

(see Ref. [19] as a review). Therefore, in this paper, a multivariate QSAR study of thirty-three 4,5-dihydroxypyrimidine carboxamide derivatives acting as HIV-1 IN inhibitors was carried out. The findings can be helpful for designing new active derivatives and better understanding the inhibition of the ST reaction.

2. Chemistry

The training set containing thirty-three 4,5-dihydroxypyrimidine carboxamide derivatives (Fig. 2) was selected from Ref. [20]. These compounds present the β-diketo acid (DKA) substructure, which is the pharmacophore of the INSTIs.

3. Pharmacology

The biological activity of the investigated compounds was measured according to the concentration required for 50% inhibition of the ST reaction, IC_{50} (nM), using the methodology described by Hazuda and co-workers [21]. The experimental IC_{50} values were converted into their corresponding pIC_{50} ($-\log IC_{50}$) measurements and are listed in Fig. 2.

4. Results

The selected model ($n = 33$; $S_{press} = 0.52$) obtained by OPS–PLS methodology presented six descriptors. Three outliers were detected (**8**, **21** and **34**) through the leverage *versus* studentized residuals plot (Pirouette version 4) [22]. For an easier physicochemical interpretation, two more descriptors were eliminated. The PLS model (Equation (1)) obtained with two latent variables cumulated 59.54% from the original information. The selected descriptors were the energy of the highest occupied molecular orbital (E_{HOMO}), the component vector to the overall polarizability in the *Y* plane (α_{yy}), the total energy (E_T), and the sum of the bond electrotopological values of carbon–carbon aromatic bonds in which the carbons are not substituted (or bond-type E-state index SeaC2C2aa) (Table 1). These properties were capable of elucidating 68.10% and predicting 57.67% of total variance. The *F-test* result was much higher than the tabled *critical*-F (cF) with 95% confidence interval and the $PRESS_{cal}$ and $PRESS_{val}$ were smaller than SS_Y.

$$pIC_{50} = +127.17(E_{HOMO}) + 0.06(\alpha_{yy}) - 0.001(E_T) + 0.10(SeaC2C2aa) + 28.67 \quad (1)$$

$n = 30$; outliers: 3; LVs = 2; Cumulated information = 59.54% (LV1: 38.41%; LV2: 21.13%); $R^2 = 0.68$; SEC = 0.57; $PRESS_{cal} = 8.72$; $F_{(2,27)} = 28.97$ (cF = 3.35); $Q^2_{LOO} = 0.58$; SEV = 0.62; $PRESS_{val} = 11.62$ ($SS_Y = 27.46$).

LNO and *Y*-randomization results are shown in Fig. 3. The model presented high average Q^2_{LNO}, small fluctuations of the standard deviations for each LNO point and small variations related to the Q^2_{LNO} value. Furthermore, the highest average value was found for L7O (0.61), and the lowest average value was found for L6O (0.53). *Y*-randomization results are in agreement with the suggested limits [23]. This indicates that the explained variance by the model is not due to chance correlation. Thus, considering all these tests, the model selected as the optimum is robust.

To verify the external predictability, five compounds (**10**, **12**, **22**, **26** and **36**) having low leverage values and presenting biological activities covering the entire range of the training set were selected.

The model without the test set presented similar statistical parameters to those found for model (**1**) ($R^2 = 0.66$, SEC = 0.61, $PRESS_{cal} = 8.30$, $F = 21.71$, $Q^2_{LOO} = 0.53$; SEV = 0.68, and $PRESS_{val} = 11.70$), therefore, they can be considered equivalent. Many authors argue that only externally validated models, after the internal validation, may be considered realistic and applicable for drug design or regulatory purposes [24,25]. Studies such as those reported by Golbraikh and Tropsha [26], and Aptula and co-workers [27], support this assumption. The results (Table 2) has shown that the selected model presented high external predictability, considering the proposed limits [26]. SEP and ARE_{pred} values were also considerably low, which might indicate low prediction errors (low deviations compared to a real value) for a synthesized derivative based on this model, for example.

5. Model discussion

Three out of the four selected descriptors influence positively the pIC_{50} values. Considering the standardized coefficients ($+0.40E_{HOMO}$, $+0.31\alpha_{yy}$, $-0.25E_T$ and $+0.36$SeaC2C2aa) given by PLS model (**1**), all descriptors are significant to the model. The model presented statistical quality, good prediction power and robustness. But in QSAR studies it is desirable to obtain a model where the physicochemical properties, represented by the molecular descriptors, can be interpreted and a parallel with the mechanism of action under study (when available) can be traced [28].

There are several works in literature speculating the action mechanism of INSTIs. In this sense, the three atoms with lone electron pairs (oxygen and/or nitrogen) in the pharmacophoric structure of these compounds (DKA substructure) present a required distance for binding two metallic ions at the same time (probably Mg^{+2} or Mn^{+2}) coordinated by the catalytic amino acid residues (D64, D116 and E152) [29]. The crystallographic core containing the *in vitro* inhibitor 5CITEP (PDB 1QS4) [30] shows this site with just one metallic ion. Pharmacological studies confirmed the metal-dependent enzymatic activity. However, additional bonds probably are also necessary, mainly that one formed by an aromatic side chain and a hydrophobic environment in the active site located in the disorganized loop formed by the amino acid residues 140–145 (Fig. 4) [31–35].

The optimum model presented high influence of properties related to electronic distribution. The E_{HOMO} values are normally related to the molecular reactivity, ionization potential and the capacity of a molecule to perform a nucleophilic attack, and

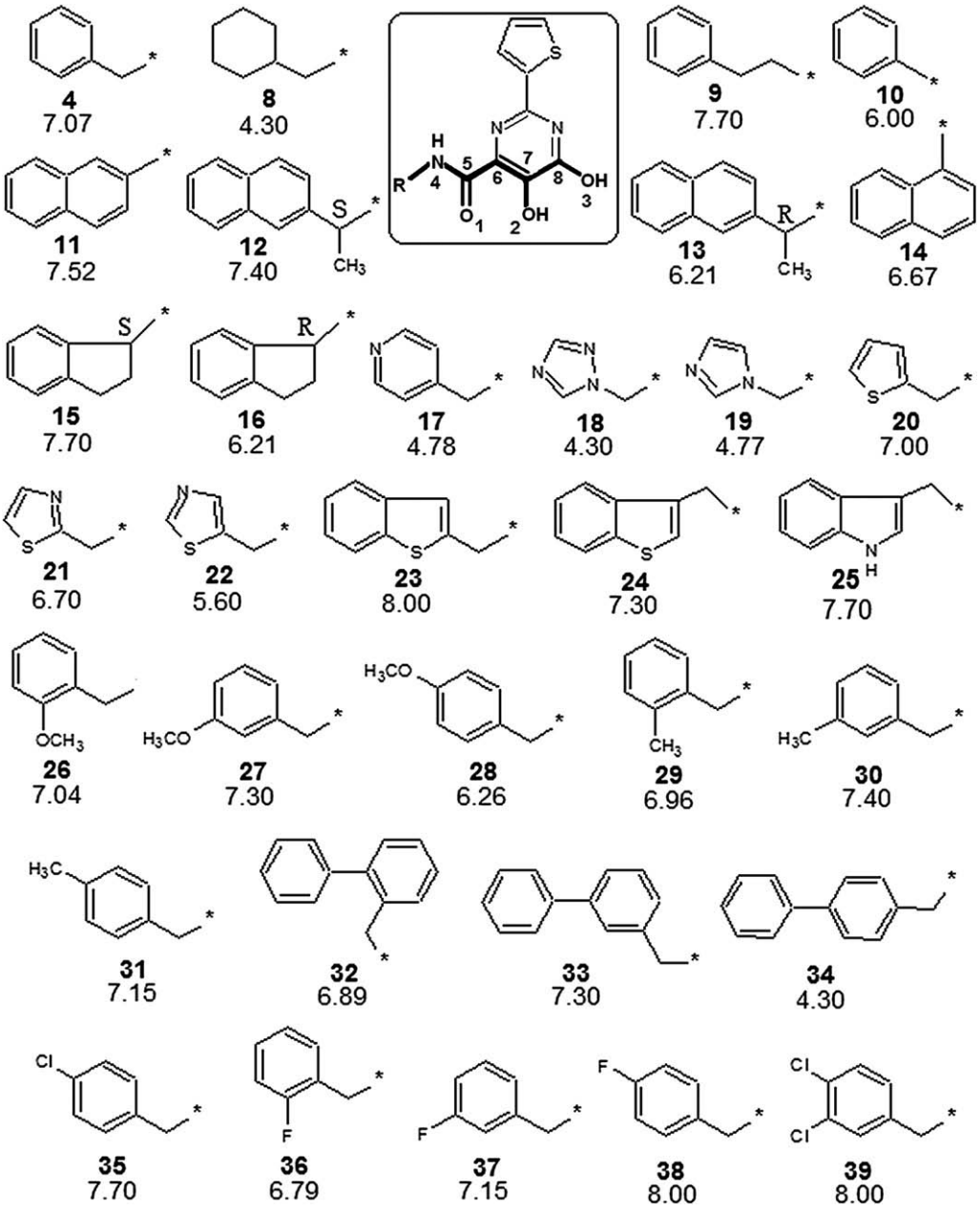

Fig. 2. Training set of thirty-three 4,5-dihydroxypyrimidine carboxamides. The identification code (in bold) used in the original reference [20] was kept. Only the partial charges of indicated atoms (1–8) were utilized as descriptors, because they correspond to DKA pharmacophoric structure.

establish charge-transfer complexes [36]. The molecular polarizability is described as the ease of a molecule to have its electronic cloud distorted by an external electric field [37]. In molecular modeling the total energy, a thermodynamic descriptor, is calculated through the electronic distribution of a molecule [36]. Finally, in the E-state formalism, each atom or bond is seen as having an intrinsic state which is perturbed by every other atom or bond in the molecule. This state encodes information regarding electronic distribution (variation caused by all atoms of a structure) and topologic aspects (greater/minor accessibility of atoms and bonds to the external environment), and as those information can influence the intermolecular interactions [38].

Fig. 5 shows that the HOMO is located in the basic structure included the DKA substructure. Considering the most probable mechanism of action, the presence of E_{HOMO} in the selected model was expected, because the inhibitors possibly act as Lewis base, donating the electrons supplied from the HOMO to the formation of bonds with the metal ions. How readily this occurs is reflected in the E_{HOMO}. Molecules with high E_{HOMO} values present a higher tendency to donate their electrons and hence are relatively reactive compared to the molecules with low-lying E_{HOMO}. In this case, compounds **19** and **20**, the less active molecules, presented the lowest E_{HOMO} values, while **25**, one of the most potent inhibitors, presented high E_{HOMO} value. The positive coefficient supports this relationship. Since the electron metal affinity (Mg^{+2} or Mn^{+2}) is constant, only the E_{HOMO} values of the 4,5-dihydroxypyrimidine carboxamide derivatives should be examined as a factor that influences on the reactivity between the training set and the ions. Other QSAR studies where this descriptor appears support this hypothesis [39–41]. Despite the small range of values found for this

Table 1
Descriptors used for the formulation of model, observed activities and LOO predicted values (three outliers were excluded).

Compound	E_{HOMO}[a]	α_{yy}[b]	E_T[c]	SeaC2C2aa[d]	pIC$_{50}$ obs	pIC$_{50}$ pred	Residues
4	−0.214	39.000	−1405.705	11.938	7.070	6.760626	0.309374
9	−0.215	47.319	−1445.023	12.137	7.700	7.179344	0.520656
10	−0.217	37.740	−1366.392	11.619	6.000	6.279759	−0.279759
11	−0.213	50.826	−1520.039	14.159	7.520	8.125476	−0.605476
12	−0.214	42.003	−1598.670	12.465	7.400	7.253083	0.146917
13	−0.214	44.465	−1598.666	12.465	6.210	7.484566	−1.274566
14	−0.211	30.463	−1559.327	14.306	6.670	7.363976	−0.693976
15	−0.214	39.024	−1483.133	10.451	7.700	6.680517	1.019483
16	−0.214	39.031	−1483.133	10.451	6.210	6.761083	−0.551083
17	−0.218	38.296	−1421.739	7.604	4.780	5.935599	−1.155599
18	−0.222	33.891	−1415.701	3.940	4.300	4.83099	−0.53099
19	−0.222	34.629	−1399.664	5.626	4.770	4.89445	−0.12445
20	−0.215	36.725	−1726.455	8.111	7.000	6.461845	0.538155
22	−0.217	38.019	−1742.498	3.972	5.600	6.043526	−0.443526
23	−0.215	49.411	−1880.110	10.300	8.000	7.659255	0.340745
24	−0.217	51.029	−1880.110	10.301	7.300	7.624567	−0.324566
25	−0.206	45.101	−1537.278	10.149	7.700	8.388641	−0.688642
26	−0.212	44.565	−1520.232	9.817	7.040	7.332997	−0.292997
27	−0.213	44.845	−1520.230	7.868	7.300	6.979424	0.320577
28	−0.213	40.498	−1520.230	7.866	6.260	6.771919	−0.511919
29	−0.214	42.056	−1445.023	10.123	6.960	6.83435	0.12565
30	−0.214	42.880	−1445.026	8.095	7.400	6.636039	0.763961
31	−0.214	41.426	−1445.026	8.143	7.150	6.567787	0.582213
32	−0.214	47.645	−1636.765	7.896	6.890	7.239001	−0.349001
33	−0.214	43.005	−1636.768	9.366	7.300	7.046585	0.253415
35	−0.217	41.587	−1865.299	7.441	7.700	6.539001	1.160998
36	−0.212	37.662	−1504.938	5.911	6.790	6.416278	0.373722
37	−0.216	39.381	−1504.936	18.471	7.150	7.410614	−0.260614
38	−0.216	38.968	−1504.936	16.427	8.000	6.888638	1.111362
39	−0.219	44.985	−2324.888	16.505	8.000	8.402181	−0.402181

[a] In eV; calculated at B3LYP/6-31G** level in the Gaussian 03 software.
[b] Calculated with the Charge Plugin of the Marvin 4.1.8 software.
[c] In Hartree; calculated at B3LYP/6-31G** level in the Gaussian 03 software.
[d] Calculated with E-state implemented by PClient Interface (http://www.vcclab.org).

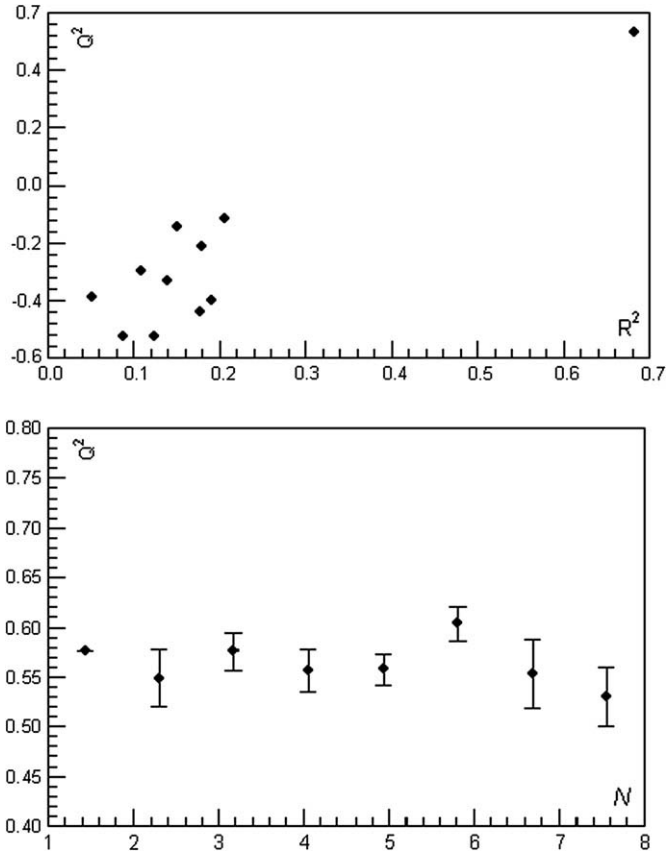

Fig. 3. Plots of *Y*-randomization (A) test and LNO validation (B) plots.

descriptor (−0.21 to −0.22 eV), it is enough to describe this characteristic over the training set. Such tendency can be observed by the linear correlation coefficient to the pIC$_{50}$ without the outliers ($r = 0.50$).

The α_{yy}, calculated by a method based on the empiric model proposed by Miller and Savchik [42,43], describes the ability of a molecule to be polarized in the *Y* direction. Despite some displacement of the center of mass, the component *Y* in the compounds of training set crosses through the DKA substructure, which is the binding site for the ions (Fig. 6). Thus, this descriptor can also be related to a possible nucleophilic attack to the metallic ions and the resultant change in the charge distribution. The signal of the coefficient is positive, indicating that the improvement of the polarization in this plane is favorable to the activity. In fact, the tendency of the compounds in a range of pIC$_{50}$ 4.30–7.07 (half of training set without outliers) presented low values for α_{yy}.

Table 2
Predicted values of the test set for HIV-IN inhibition and the statistics parameters.

Compound	pIC$_{50}$ obs	pIC$_{50}$ pred	Residues
10p	6.000	6.263667	−0.263667
12p	7.400	7.267147	0.132853
22p	5.600	6.006945	−0.406945
26p	7.040	7.308198	−0.268198
36p	6.790	6.480028	0.309972
R^2_{pred}			0.8736
SEP			0.2901
ARE$_{pred}$			4.3663%
k			0.9862
k'			1.0122
$\lvert r^2_0 - r^2_{0'}\rvert$			−0.1754

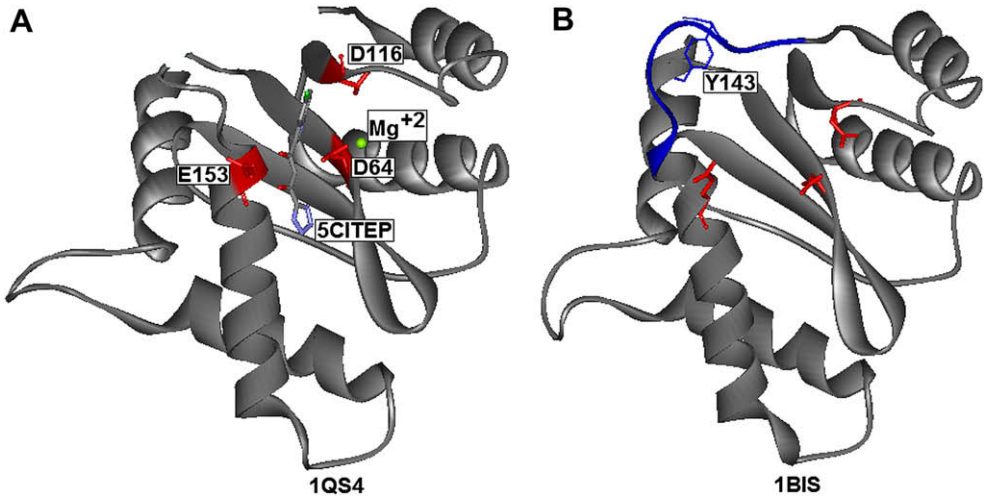

Fig. 4. Crystallographic structures of the core of HIV-1 IN. Catalytic triad in red tubes. Inhibitor 5CITEP in color-by-atoms tubes. Mg^{+2} in green. Flexible loop and Y143 in blue. Square: β-diketo acid, DKA, pharmacophoric group, where Ar are aromatic chains. Figure built in ViewerLite 4.2 software (Accelrys Inc, http://www.accelrys.com). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

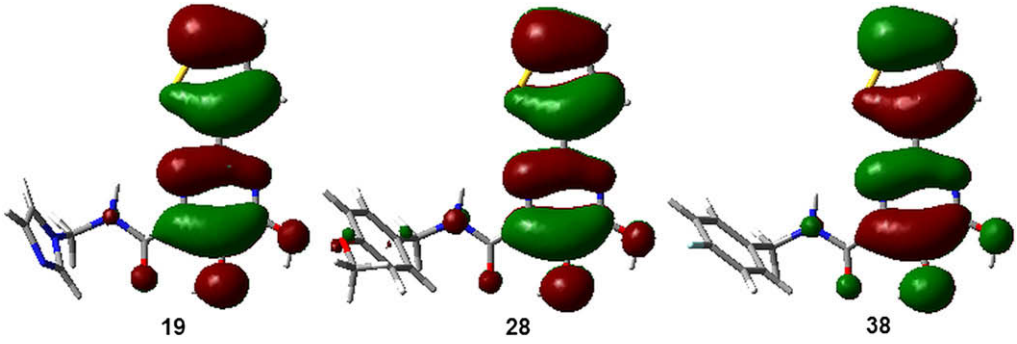

Fig. 5. Highest Occupied Molecular Orbitals (HOMO) of compounds **19**, **28** and **38**. Figure built in GaussView 3.0 software (Gaussian Inc, http://www.gaussian.com).

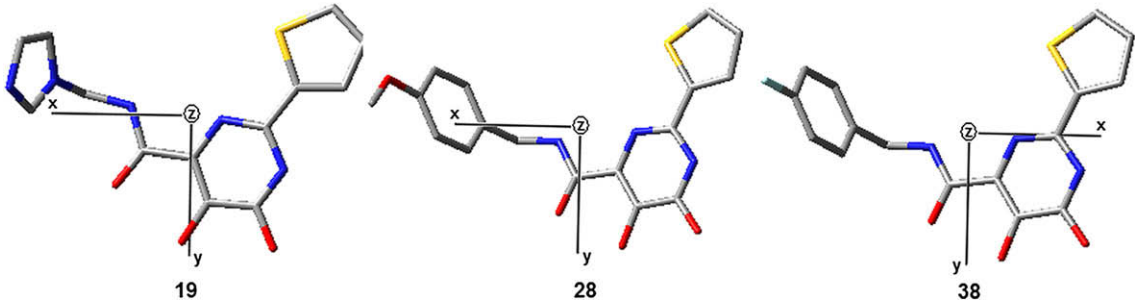

Fig. 6. Cartesian axes' representation for compounds **19**, **28** and **38**. Despite the little displacement of center of mass, *Y* axis always crosses DKA substructure. Figure built in GaussView 3.0 software (Gaussian Inc, http://www.gaussian.com).

In QSAR studies, the descriptor E_T (and other thermodynamic descriptors obtained by molecular modeling) has been suggested to be related to molecular stability [44–46]. In the molecular structures of the training set, the region which presents more conformational freedom corresponds to the aromatic side chain, and it was reported that this substructure binds to the HIV-1 IN establishing a π-stacking interaction with the Y143 residue, located in the flexible loop [32,33,47]. Thus, the negative signal of the standardized coefficient may denote that the binding of the less stable inhibitors (small absolute value for E_T) is not favored. Thus, an increase of this property can improve the binding with the specific active site, promoting the increase of activity.

The descriptor SeaC2C2aa is probably related to the binding between the aromatic side chain and the residue Y143, regarding a π-stacking interaction, because its values varied only as a function of the aromatic side chain. The positive coefficient indicates that the increase of this parameter contributes favorably to the activity. Some of the less active compounds – **8** (outlier), **18**, **19** and **22** – have lower values of this parameter, and some of the most active – **11**, **37**, **38** and **39** – have higher values. This descriptor considers the environment of each bond, and the electronegativity of the substituents is also important for the result. For instance, the SeaC2C2aa values found for compounds **4** and **37** were 11.94 and 18.47, respectively, but the only difference between these two

Table 3
Statistical parameters analyzed and correspondent equations.

Parameter	Definition	Equation
R^2	Squared correlation coefficient of calibration model	$1 - [\sum_i (y_{obsi} - y_{ci})^2 / \sum_i (y_{obsi} - \bar{y}_{obsi})^2]$
SEC	Standard deviation of calibration model	$[\sum_i (y_{obsi} - y_{ci})^2 / n - p - 1]^{1/2}$
$PRESS_{cal}$	Predictive Residual Sum of Squares of Calibration	$\sum_i (y_{obsi} - y_{ci})^2$
F	F-test (with 95% confidence interval)	$[\sum_i (y_{obsi} - y_{ci})^2 / k] / [\sum_i (y_{obsi} - \bar{y}_{obs})^2 / n - p - 1]$
Q^2_{LNO}	Squared correlation coefficient of cross validation (“leave-N-out”, LNO).	$1 - [\sum_i (y_{obsi} - y_{vi})^2 / \sum_i (y_{obsi} - \bar{y}_{obsi})^2]$
SEV	Standard error of cross validation	$[\sum_i (y_{obsi} - y_{vi})^2 / n]^{1/2}$
$PRESS_{val}$	Predictive Residual Sum of Squares of Calibration of Validation	$\sum_i (y_{obsi} - y_{vi})^2$
R^2_{pred}	Squared correlation coefficient of prediction	$1 - [\sum_i (y_{obsi} - y_{evi})^2 / \sum_i (y_{obsi} - \bar{y}_{obsi})^2]$
ARE_{pred}	Average relative error of prediction	$[\sum_i (\lvert y_{obsi} - y_{evi} \rvert) / y_{obsi}] \times 100/n$
SEP	Standard error of prediction	$[\sum_i (y_{obsi} - y_{evi})^2 / n]^{1/2}$
k and k'	Slopes of the linear prediction regression lines	$\sum_i (y_{obsi} - y_{evi}) / \sum_i y_{evi}$ and $\sum_i (y_{obsi} - y_{evi}) / \sum_i y_{obsi}$

y: biological activity; $\bar{y}$: average observed biological activity; obs: experimental values; c: estimated activity in the regression model; v: estimated activity in the cross-validation; ev: estimated activity in the external validation; n: number of samples of training set; p: number of latent variables; $\bar{y}_{obs}$: average observed activity for the complete training set; test: test set.

compounds is a fluorine atom in the ring. Thus, as already mentioned, the bond-type E-state index encodes the electron accessibility of a specific bond type [48]. This information can indicate the relationship between the SeaC2C2aa descriptor and the importance of the π-stacking aromatic interaction in the active site of HIV-1 IN [32,43,47].

The SeaC2C2aa descriptor also should be the reason for compounds **8** and **34** being outliers. The first compound does not have aromatic bonds in the side chain. The second has the same SeaC2C2aa value found for compounds **32** and **33**, its isomer, but it is much less active. Compound **21** could be an outlier because its descriptor values were quite similar to its isomer (compound **22**), but it is more potent than **22**.

6. Conclusion

In this study it was possible to obtain a multivariate QSAR model for a set of thirty-three 4,5-dihydroxypyrimidine carboxamides that have the capability of inhibiting the *in vitro* ST reaction catalyzed by HIV-1 IN. The LOO and LNO cross-validation methods, the *Y*-randomization technique, and the external validation indicated that the model is significant, robust and has good internal and external predictability. The quality of the selected model is strengthened by the physicochemical interpretation, which found very satisfactory support in the literature for all descriptors of the best model. The inhibitory activity of the investigated compounds was described based on the E_{HOMO}, E_T, α_{yy} and SeaC2C2aa values, all related to the electronic distribution. It was possible a relationship between all descriptors and the most accepted hypothesis regarding the mechanism of action, especially for the E_{HOMO} and SeaC2C2aa descriptors, which were the most important parameters considering the standardized coefficients. It is interesting to notice that the OPS algorithm was able to select a combination of descriptors related to the mechanism of action. This could have happened because OPS considers the biological activity information through the informative vectors for variable selection. The resulting findings can be helpful in the development and optimization of new HIV-1 IN inhibitors.

7. Methodology

Three-dimensional structures were assembled based on similar crystallographic fragment (code DOTRUZ) retrieved from the Cambridge Structural Database [49]. The molecular modifications and geometry optimization by molecular mechanics (MM + force field) were carried out using HyperChem 7 [50]. Geometry optimizations were performed in the following sequence: AM1 → HF/6-31G* → B3LYP/6-31G** at Gaussian 03 [51]. The DFT/B3LYP functional was chosen because it is reported that this method leads to satisfactory results when molecular geometries and energies are considered [18,52]. The minimun energy structures were used to obtain the electronic descriptors (Gaussian 03). Other descriptors (steric, topological, solubility, constitutional) were calculated employing other chemical representations and using different software (see Supplementary material), giving a total of 162 molecular descriptors.

To obtain the best model, a three-step procedure was employed. The number of descriptors was reduced to sixty-three, eliminating those in which the absolute value of the linear correlation coefficient ($|r|$) to the pIC_{50} was lesser than 0.3. It was considered that below this threshold, no useful statistical information would be provided to the model. The remaining descriptors were further analyzed employing the Ordered Predictors Selection (OPS) algorithm [53]. In this algorithm, the descriptors are selected in three steps: (i) obtaining an informative vector; (ii) ordering the variables into decreasing order by this vector; and (iii) investigating the ordered variables. Three informative vectors were used in this work: regression vector, the correlation vector and the product between the elements of both vectors. Then, the OPS method builds models using Partial Least Squares [54]. Following a suggestion from Wold and co-workers [55], the obtained models were sort by the S_{press}, $(PRESS_{val})^{1/2}/n - p - 1$, value. This parameter can penalize the model with larger number of latent variables, which seems to be preferable to encourage model parsimony.

The best reduced combination of descriptors was refined using the software Pirouette 4 [22], by removing outliers and some more descriptors, seeking to obtain a statistically significant, robust and interpretative model.

Several statistical parameters, listed in Table 3, were used to evaluate the quality of the model. For the internal quality, the recommended limits are $R^2 \geq 0.6$ and $Q^2_{LOO} \geq 0.5$ [26,56], the SEC and SEV should be as low as possible and the $PRESS_{cal}$ and $PRESS_{val}$ values should be lower than the sum of squares of the response values (SS_Y) [57]. The *F-test* value should be higher than the tabled *critical-F* (cF) and, the higher the difference between them is, the more statistically significant the model will be [58].

The robustness of the model was examined by the leave-*N*-out cross-validation (LNO, $N = 1$ at 8) and *Y*-randomization test [56]. The LNO employs smaller training sets than the LOO procedure, and QSAR models with a high average Q^2_{LNO} and small oscillations can be considered robust [24]. The LNO was performed in three steps, where in each a pre-randomization of all rows of data matrices (**X** and **Y**) was performed before the LNO process, in order to decrease the impact that the withdrawal of sets of samples in some combinations could have in the result of Q^2_{LNO}. Results with an average for each point close to Q^2_{LOO} with standard deviations are expected near zero [26]. For the *Y*-randomization test, performed ten times [57], $R^2 \leq 0.3$ and $Q^2_{LOO} \leq 0.05$ for all results were considered acceptable. These limits were selected based on Eriksson and coworkers' suggestions [23]. The *Y*-randomization test is capable of verifying if models with high values of R^2 and Q^2_{LOO} present chance correlation [25]. Both tests were performed in Matlab 7 [59] and the plots built in the DataFit 9 [60].

After internal evaluations, a set for external validation (test set), having a representative pIC_{50} range as well as structural variations, was selected from the training set and a new model was built. The statistical quality of the new model cannot be much different from the model generated with all compounds. A QSAR model can be considered predictive when presenting $R^2_{pred} \geq 0.6$ [26]. But this is not enough. It is also suggested to check the following measures: (i) the slopes k or k' of the linear regression lines (equations in Table 3) between the observed activity (y_{obsi}) and the predicted activity from the external validation set (y_{evi}), where at least one slope should be in the range $0.85 \leq x \leq 1.15$ ($x = k$ or k'); and (ii) the absolute value of the difference between the determination coefficient between y_{obsi} and y_{evi}, r^2_0, and the determination coefficient between y_{evi} and y_{obsi}, r'^2_0, that result should be lesser than 0.3 [24–26]. The SEP and ARE_{pred} values must be as low as possible.

Acknowledgements

EBM thanks the Universidade Estadual do Oeste do Paraná (http://www.unioeste.br) for support to the author's doctoral thesis, and the Instituto de Química of Universidade Estadual de Campinas (http://www.iqm.unicamp.br). which made possible the development of this research. MMCF thanks FAPESP for financial support (2004/04686-5).

Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejmech.2009.03.001.

References

[1] Unaids/WHO (Ed.), AIDS Epidemic Update: December 2007, Unaids, Geneva, 2006.
[2] T. Matthews, M. Salgo, M. Greenberg, J. Chung, R. DeMasi, D. Bolognesi, Nat. Rev. Drug. Discov. 3 (2004) 215–225.
[3] M. Hagmann, Bull. World Health Organ. 81 (2003) 918–919.
[4] H.C. Castro, N.I. Loureiro, M. Pujol-Luz, A.M.T. Souza, M.G. Albuquerque, D.O. Santos, L.M. Cabral, I.C. Frugulhetti, C.R. Rodrigues, Curr. Med. Chem. 13 (2006) 313–324.
[5] E.B. Melo, A.G. Silveira, I. Carvalho, Tetrahedron 62 (2006) 10277–10302.
[6] E.B. Melo, I. Carvalho, Quim. Nova 29 (2006) 840–843.
[7] M.M. Rosenkilde, L. Gerlach, J.S. Jacobsen, R.T. Skerlj, G.J. Bridger, T.W. Schwartz, J. Biol. Chem. 279 (2004) 3033–3041.
[8] S. Hadlington, Chem. World 4 (2007) 14.
[9] M.V.N. Souza, M.V. Almeida, Quim. Nova 26 (2003) 366–372.
[10] E.B. Melo, A.T. Bruni, M.M.C. Ferreira, Quim. Nova 29 (2006) 555–562.
[11] A. Opar, Nat. Rev. Drug. Discov. 6 (2007) 258–259.
[12] V. Summa, A. Petrocchi, V.G. Matassa, C. Gardelli, E. Muraglia, M. Rowley, O.G. Paz, R. Laufer, E. Monteagudo, P. Pace, J. Med. Chem. 49 (2006) 6646–6649.
[13] Y. Wang, N. Serradell, J. Bolos, E. Rosa, M.D. Frederick, Drugs Future 32 (2007) 118–122.
[14] J. Cohen, Science 311 (2006) 943.
[15] A. Savarino, Retrovirology 4 (2007). http://www.retrovirology.com/content/pdf/1742-4690-4-21.pdf (accessed march 2009).
[16] E. Zeinalipour-Loizidou, C. Nicolaou, A. Nicolaides, L.G. Kostrikis, Curr. HIV Res. 5 (2007) 365–388.
[17] F.A.L. Ribeiro, M.M.C. Ferreira, J. Mol. Struct. Theochem 663 (2003) 109–126.
[18] F.A. Molfetta, A.T. Bruni, F.P. Rosseli, A.B.F. Silva, Struct. Chem. 18 (2007) 49–57.
[19] N. Nunthaboot, S. Pianwanit, V. Parasuk, S. Kokpol, J.M. Briggs, Curr. Comput. Aided Drug Des. 3 (2007) 160–190.
[20] A. Petrocchi, U. Koch, V.G. Matassa, B. Pacini, K.A. Stillmock, V. Summa, Bioorg. Med. Chem. Lett. 17 (2007) 350–353.
[21] D.J. Hazuda, P.J. Felock, J.C. Hastings, B. Pramanik, A.L. Wolfe, J. Virol. 71 (1997) 7005–7011.
[22] Pirouette Software Version 4, Infometrix Inc., USA, 2007.
[23] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramática, Environ. Health Perspect. 111 (2003) 1361–1375.
[24] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, J. Markopolus, O. Igglessi-Markopoulou, J. Comput. Aided Mol. Des. 21 (2007) 251–267.
[25] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[26] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[27] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz, M.T.D. Cronin, QSAR Comb. Sci. 24 (2005) 385–396.
[28] Organisation for Economic Co-Operation and Development, Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models, OECD, Paris, 2007, Available from: http://www.oecd.org/ehs.
[29] T. Kawasuji, M. Fuji, T. Yoshinaga, A. Sato, T. Fujiwara, R. Kiyama, Bioorg. Med. Chem. 14 (2006) 8420–8429.
[30] 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.
[31] R. Dayam, L.Q. Al-Mawsawi, N. Neamati, Drugs R.D. 8 (2007) 155–168.
[32] M.L. Barreca, L. De Luca, S. Ferro, A. Rao, A. Monforte, A. Chimirri, ARKIVOC 7 (2006) 224–244.
[33] G.C.G. Pais, X. Zhang, C. Marchand, N. Neamati, K. Cowansage, E.S. Svarovskaia, V.K. Pathak, Y. Tang, M. Nicklaus, Y. Pommier, T.R. Burke Jr., J. Med. Chem. 45 (2002) 3184–3194.
[34] L. De Luca, G. Vistoli, A. Pedretti, M.L. Barreca, A. Chimirri, Biochem. Biophys. Res. Commun. 336 (2005) 1010–1016.
[35] J.A. Grobler, K. Stillmock, B. Hu, M. Witmer, P. Felock, A.S. Espeseth, A. Wolfe, M. Egbertson, M. Bourgeois, J. Melamed, J.S. Wai, S. Young, J. Vacca, D.J. Hazuda, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 6661–6666.
[36] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Rev. 96 (1996) 1027–1043.
[37] M.M. Gonçalves, L.F. Fraceto, M.M.D.C. Vila, R.V.M. Oliveira, Quim. Nova 29 (2006) 1072–1077.
[38] L.B. Kier, L.H. Hall, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, United Kingdom, 1999, pp. 491–562.
[39] J.O. Morley, T.P. Matthews, Bioorg. Med. Chem. 14 (2006) 8099–8108.
[40] K.M. Nikolic, QSAR Comb. Sci. 26 (2007) 358–367.
[41] P. Silakari, S.D. Shrivastava, G. Silakari, D.V. Kohli, G. Rambabu, S. Srivastava, S.K. Shrivastava, O. Silakari, Eur. J. Med. Chem. 43 (2008) 1559–1569.
[42] K.J. Miller, J.A. Savchik, J. Am. Chem. Soc. 101 (1979) 7206–7213.
[43] Marvin User's Guide, Calculator Plugins, Charge Plugin (2008). http://www.chemaxon.com/marvin/help/calculations/chargegroup.html (accessed September 2008).
[44] O.A. Philips, E.E. Udo, S.M. Samuel, Eur. J. Med. Chem. 43 (2008) 1095–1104.
[45] B.B. Lohray, N. Gandhi, B.K. Srivastava, V.B. Lohray, Bioorg. Med. Chem. Lett. 16 (2006) 3817–3823.
[46] K. Toit, E.E. Elgorashi, S.F. Malan, S.E. Drewes, J. van Staden, N.R. Croueh, D.A. Mulholland, Bioorg. Med. Chem. 13 (2005) 2561–2568.
[47] A.L. Parril, Curr. Med. Chem. 10 (2003) 1875–1888.
[48] J.R. Votano, M. Parham, L.H. Hall, L.B. Kier, S. Oloff, A. Tropsha, Q. Xie, W. Tong, Mutagenesis 19 (2004) 365–377.
[49] Cambridge Structural Database Software Version 5.29 (November 2007) + 1 Update, Cambridge Crystallographic Data Centre, England, 2007.
[50] HyperChem Software Version 7.1, Hyper Co., USA, 2002.
[51] Gaussian 03W Software Version 6.0, Gaussian Inc., USA, 2003.
[52] J. Lameira, I.G. Medeiros, M. Reis, A.S. Santos, C.N. Alves, Bioorg. Med. Chem. 14 (2006) 7105–7112.
[53] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemom., J. Chemom. 23 (2009) 32–48.
[54] M.M.C. Ferreira, J. Braz. Chem. Soc. 13 (2002) 742–753.
[55] S. Wold, E. Johansson, M. Cocchi, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, Kluwer/Escom, Dordrecht, 2000, pp. 523–550.
[56] C. Rücker, G. Rücker, M. Meringer, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[57] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.
[58] A.C. Gaudio, E. Zandonade, Quim. Nova 24 (2001) 658–671.
[59] Matlab Software Version 7, MathWorks Inc., USA, 2006.
[60] DataFit Software 9, Oakdale Engineering, USA, 2008.