Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 16 (2008) 3039–3048
Bioorganic & Medicinal Chemistry

Molecular modeling studies of N-substituted pyrrole derivatives—Potential HIV-1 gp41 inhibitors

Cátia Teixeira,[a] Florent Barbault,[a,*] Joseph Rebehmed,[a] Kun Liu,[b] Lan Xie,[b] Hong Lu,[c] Shibo Jiang,[c] BoTao Fan[a,†] and François Maurel[a]

[a]*ITODYS, University Paris 7—CNRS UMR 7086, 1 rue Guy de la Brosse, 75005 Paris, France*
[b]*Beijing Institute of Pharmacology & Toxicology, 27 Taiping Road, Beijing 100850, China*
[c]*Lindsley F. Kimball Research Institute, New York Blood Center, 310 East 67th Street, New York, NY 10021, USA*

Received 10 September 2007; revised 12 December 2007; accepted 18 December 2007
Available online 28 January 2008

Abstract—2D-, 3D-QSAR and docking studies were carried out on 23 pyrrole derivatives, to model their HIV-1 gp41 inhibitory activities. The 2D, 3D-QSAR studies were performed using CODESSA software package and comparative molecular field analysis (CoMFA) technique, respectively. The CODESSA five-descriptor model has a correlation coefficient $R^2 = 0.825$ and a standard deviation $s^2 = 0.132$. The 3D-QSAR CoMFA study allowed to obtain a model showing a good correlative and predictive capability which statistical results, provided by a eight-component regression equation, are: $R^2 = 0.984$, $q^2 = 0.463$, $s = 0.119$. Docking studies were employed to determine probable binding conformation of these analogues into the gp41 active site using the AutoDock program whose results were found complementary with thus of 2D- and 3D-QSAR analysis. These findings provide guidance for the design and structural modifications of these derivatives for better anti-HIV-1 activity which is important for the development of a new class of entry inhibitors.

1. Introduction

It is well known that the human immunodeficiency virus (HIV), the origin of the acquired immunodeficiency syndrome (AIDS) still remains a major cause of death.[1] The infection with HIV is characterized clinically by an inexorable decline in immune functions, leading to fatal consequences. The ability to inhibit viral replication is currently achieved in most patients with the so-called highly active antiretroviral therapy,[2] a combination of viral protease inhibitors[3] and nucleosidic, non-nucleosidic or nucleotidic reverse transcriptase inhibitors.[4,5] This combination therapy has been remarkably successful in reducing viral load and has led to a decline in morbidity and mortality.[6,7] Unfortunately, most of these molecules present numerous shortcomings such as viral resistances[8] and adverse effects.[9,10] In addition, these drugs are involved to later stages of infection.[11] Therefore, it is necessary to develop new drugs which are able to block the first steps of viral cycle life.

The fusion and entry of HIV into susceptible cells are mediated by its envelope glycoproteins gp41 and gp120.[12] Gp120 directs the virus to the appropriate target cell by binding to the rHuman T-cell receptor (CD4) and chemokine co-receptor CXCR4 (also called fusin) or CCR5 (chemokine C–C motif receptor 5).[13,14] Upon gp120 binding, gp41 undergoes a series of conformational changes to convert from native, non-fusogenic conformation to fusogenic conformation mediating fusion of the HIV membrane with the human cell membrane, thereby allowing introduction of the viral genome into the target cell.[15,16] The biomolecules involved in the HIV entry process can serve as targets for development of anti-HIV drugs to block these early steps of viral cycle life. Entry inhibitors are a new family of antiretrovirals presently represented only by one drug, T-20 (Enfuvirtide, Fuzeon®) a peptidic drug targeting gp41.[17] This synthetic peptide of 36 amino acid based on the sequence of the C-terminal heptad repeat regions (CHR) helix of gp41 was found to bind to the N-terminal heptad repeat regions (NHR) coiled coil.[18] It prevents the peripheral attachment of the three

Keywords: HIV-1 fusion inhibitors; gp41; CODESSA; Docking; CoMFA.
* Corresponding author. E-mail: florent.barbault@univ-paris-diderot.fr
† Deceased on 22nd October 2006.

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.12.034

CHR helices to the NHR coiled coil, thereby inhibiting the formation of the six-helix bundle and consequent fusion of the viral and cell membranes.[19,20] However, T-20 as a relatively long peptide suffers from two crucial limitations: lack of oral availability and high cost of production. Furthermore, T-20 is vulnerable to proteolytic digestion because it is composed of L-amino acids.[16] As a result there is a considerable interest to develop small-molecules anti-HIV-1 compounds with a mechanism of action similar to that of C peptides but without the disadvantages of the peptidic drugs.

It was proposed that compounds binding to the gp41 NHR and CHR regions and blocking the six-helix bundle formation might have inhibitory activity on HIV mediated membrane fusion.[21] Effectively, each of the grooves on the surface of the N-helix trimer has a deep hydrophobic pocket that accommodates three conserved hydrophobic residues in the gp41 CHR region, suggesting that this pocket is an attractive target for designing new class of anti-HIV-1 drugs.[22,23] Based on this information, several studies have developed a series of high-throughput screening assays, which have used for screening chemical libraries consisting of 'drug-like' compounds. Two small-molecules, ADS-J1 and XTT formazan, were identified and presented anti-HIV activity by docking into the gp41 pocket, thereby interfering with the fusion viral and cell membranes.[24–26] More recently, two pyrrole derivatives, NB-2 and NB-64, were reported as novel HIV-1 fusion inhibitors at low micromolar levels, which may bind to the gp41 hydrophobic pocket via hydrophobic and ionic interactions and block the formation of the six-helix bundle.[24,27]

NB-2 and NB-64 are 'drug-like' and were used by Xie et al.[28] as leads for designing more pyrrole derivatives. These new pyrrole derivatives were then synthesized and the inhibitory activities determined (data to be published). After this, we search to gain an insight into the binding mode and interactions of these compounds with HIV-1 gp41 and, consequently, to improve the development of more efficient fusion inhibitors. The applicability of QSAR/QSPR (Quantitative Structure–Activity Relationship/Quantitative Structure–Property Relationship) methodology to various problematic has been convincingly demonstrated in a series of publications.[29–31] In this paper, we report the 2D- and 3D-QSAR analyses along with docking studies on 23 novel pyrrole derivatives as HIV-1 entry inhibitors. Moreover, the results obtained from QSAR analyses were superimposed on the gp41 active site and the main interactions were studied. These findings provide us very good advices for future structural modifications of this new class of entry inhibitors.

2. Materials and methods

2.1. Data set preparation

The data set consists in 23 N-substituted pyrrole compounds listed in Table 1 which were synthesized[28] and evaluated for their inhibitory activity on HIV-1 replication by ELISA for measuring HIV-1 gag protein p24 as previously described.[27] Here, all these values are expressed in terms of $-\log(EC_{50})$ where EC_{50} represents the concentration of a compound resulting in 50 % inhibition of p24 production.

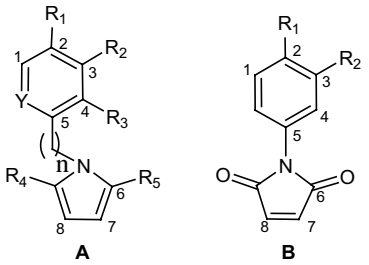
Since no structural information is available for the molecules of our data set, the molecular models were constructed by using standard geometries (standard bond lengths and angles) from the TRIPOS force field within the Sybyl 7.2 molecular modeling program.[32] Charges were assigned using the Gasteiger–Marsili method. Energy minimization was performed using 20 simplex iterations followed by 1000 steps of Powell minimization until the gradient norm 0.05 kcal/mol was achieved. The structures having a carboxylic acid moiety were constructed with a carboxylate group instead a carboxylic one, once they are expected to be ionized in physiological pH.

2.2. Codessa

The minimized structures were then subjected to the quantum-mechanical semi-empirical AM1 calculations[33] with MOPAC[34] in order to calculate the molecular characteristics. After this final optimization the ligands were used as an input data for the CODESSA[35] program where calculation and selection of descriptors for QSPR were carried out. CODESSA (comprehensive descriptors for structural and molecular analysis) uses statistical structure–property correlation techniques for the analysis of experimental data in combination with the calculated molecular descriptors, thus allowing the development of QSAR/QSPR models.

A pool of 554 molecular descriptors of different types as constitutional, topological, geometric, electrostatic and quantum-chemical was computed. After this calculation, the heuristic method was used to search the best set of descriptors for multi-linear correlations. The heuristic method can quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which are insignificant and which descriptors are highly inter-correlated. All this information is helpful in reducing the number of descriptors involved in the search for the best QSAR/QSPR model. Hence, this method proceeds to a pre-selection by sequentially eliminating descriptors which do not match any of the following criteria:[36] (a) Fisher *F*-criterion greater than 1; (b) R^2 value less than a value defined at the start (0.01); (c) Student's *t*-criterion less than that defined (0.1); and (d) duplicate descriptors having a higher squared inter-correlation coefficient than a predetermined thresholds (usually 0.8), retaining the descriptor with higher R^2 referred to the property, in order to avoid redundant information. The remaining descriptors are then listed in decreasing order of correlation coefficients when used in a global search for two-parameter correlations. Each significant two-parameter correlation by *F*-criterion is recursively expanded to an *n*-parameter correlation until the normalized *F*-criterion remains greater than the start-up value. The top *N*-correlations by R^2 as well as *F*-criterion are saved.

Table 1. Summary of experimental biological activities, $-\log(EC_{50})$, of 23 pyrrole derivatives as inhibitors of the HIV-1 gp41



Compound	R_1	R_2	R_3	R_4	R_5	Y	n	$-\log(EC_{50})$/M Experimental
$\mathbf{A}_1$	CH_2COOH	—	—	Me	Me	C	0	5.68
$\mathbf{A}_2$	Cl	COOH	—	Me	Me	C	0	5.82
$\mathbf{A}_3$	Cl	—	COOH	Me	Me	C	0	5.07
$\mathbf{A}_4$	$COOCH_3$	OH	—	—	—	C	0	4.45
$\mathbf{A}_5$	COOH	OH	—	—	—	C	0	3.89
$\mathbf{A}_6$	—	Tetrazole	—	Me	Me	C	0	5.11
$\mathbf{A}_7$	CH_2COOH	—	—	—	—	C	0	3.43
$\mathbf{A}_8$	—	Tetrazole	—	—	—	C	0	4.39
$\mathbf{A}_9$	Cl	COOH	—	—	—	C	0	5.62
$\mathbf{A}_{10}$	—	—	—	Me	Me	N	0	3.54
$\mathbf{A}_{11}$	SO_2NH_2	—	—	Me	Me	C	0	4.00
$\mathbf{A}_{12}$	CH_3	—	—	Me	Me	N	0	4.81
$\mathbf{A}_{13}$	COOH	—	—	Me	Me	C	1	4.81
$\mathbf{A}_{14}$	—	—	—	—	—	C	1	3.89
$\mathbf{A}_{15}$	COOH	—	—	—	—	C	1	4.19
$\mathbf{A}_{16}$	COOH	—	—	Me	Me	C	0	3.76
$\mathbf{A}_{17}$	COOH	—	—	—	—	C	0	4.22
$\mathbf{A}_{18}$	—	COOH	—	Me	Me	C	0	4.93
$\mathbf{A}_{19}$	—	COOH	—	—	—	C	0	4.35
$\mathbf{A}_{20}$	OH	COOH	—	Me	Me	C	0	6.16
$\mathbf{A}_{21}$	OH	COOH	—	—	—	C	0	5.02
$\mathbf{B}_1$	COOH	—	—	—	—	—	—	3.69
$\mathbf{B}_2$	$COOCH_3$	OH	—	—	—	—	—	4.40

In the QSAR/QSPR approach the choice of the number of descriptors for the model consists in a major step. Effectively, an excessive number of descriptors can lead to an over-correlated equation which is difficult to analyse in terms of interactions mechanism. A simple technique to control the model expansion is the so-called 'breaking point' rule. This procedure consists in the plot of the number of descriptors involved in the model versus the corresponding squared correlation coefficient. From the analyses of the plot it appears that the statistical improvement of the model is higher until one point and after that melioration is negligible. Consequently, the model corresponding to the breaking point is considered the best/optimum model.

2.3. CoMFA

This analysis was performed using CoMFA, a module of Sybyl. In this work, the most active molecule ($\mathbf{A}_{20}$) of the data set was used as a template for superimposition, assuming that its lowest-energy conformation represents the most probable bioactive conformation of the N-pyrrole substituted derivatives at the binding site of gp41. This assumption was made because the molecular mechanism of this new class of molecules on the inhibition of the HIV-1 glycoprotein is not completely elucidated yet. The alignment was manually carried out using the Database Align tool in Sybyl 7.2. The atoms of the common structure were selected and are numbered 1–8 in Table 1 and involve not only atoms from the pyrrole moiety but also from the phenyl ring. The remaining molecules were aligned to the compound $\mathbf{A}_{20}$.

The CoMFA fields were generated using C sp^3 atom with a + 1 charge as the probe. The region was created automatically, and the default grid spacing (2 Å) was employed, which extended 4.0 Å units in all directions beyond the dimensions of each molecule. The Van der Walls potential and Columbic terms, which represent steric and electrostatic fields, respectively, were calculated using the standard Tripos force field. The default cut-off (30 kcal/mol) was applied to both fields by considering the 'smooth' method. A distance dependent dielectric constant of 1.00 was used.

To quantify the relationship between the structural parameters and the biological activities, the PLS algorithm was used. The experimental EC_{50} (M) values were converted into $-\log(EC_{50})$ values and used as the dependent column. The cross-validation analysis was performed using leave-one-out (LOO) method wherein one compound is removed from the data set and its activity predicted with the model derived from the remaining compounds. The optimum number of

components (N) used in the model derivation was chosen from the analysis with the highest q^2 value and lowest-standard error of prediction. The column filtering was set at 2.0 kcal/mol to speed up the analysis and reduce noise. Finally, the non-cross-validation analysis was performed to calculate conventional R^2 using the optimum number of components obtained from the analysis above.

2.4. Molecular docking

Docking calculations were performed with version 3.0.5 of the program AutoDock.[37] It combines a rapid energy evaluation through pre-calculated grids of affinity potentials with a variety of search algorithms to find suitable binding positions for a ligand on a given macromolecule.

The structures were prepared for docking as follows: the X-ray crystal structure of the gp41 core was retrieved from the protein data bank (PDB code = 1aik).[23] The hydrophobic pocket in the gp41 core was used as the binding site for docking simulations. For that one of the C helices was removed from the six-helix bundle to expose the hydrophobic groove. Polar hydrogen atoms were added and Kollman charges, atomic solvation parameters and fragmental volumes were assigned to the protein. For all ligands, Gasteiger charges were assigned and non-polar hydrogen atoms merged. All torsions were allowed to rotate during docking.

The auxiliary program AutoGrid generated the grid maps. The grids (one for each atom type in the ligands, plus one for electrostatic interactions) were chosen to be sufficiently large to include significant portions of the hydrophobic pocket of gp41. The grid dimensions were thus $50 \times 50 \times 60$ Å^3, with points separated by 0.375 Å. Docking was carried out using the empirical free energy function and the Lamarckian genetic algorithm, applying a standard protocol, the energy evaluations were 250,000, the maximum number of iterations 27,000 for an initial population of 150 randomly placed individuals with a mutation rate of 0.02, a crossover rate of 0.80, and an elitism value of 1. The number of docking runs was 100 and, after docking, the 100 solutions were clustered into groups with the RMS deviations lower than 0.5 Å. The clusters were ranked by the lowest-energy representative of each bunch.

3. Results and discussion

The experimental results cover a wide range of EC_{50} from 0.69 μM to 371.24 μM giving us information useful to understand the different activity profiles of the molecules of our data set.

3.1. Codessa analysis

The optimum number of descriptors for the best statistical model describing $-\log(EC_{50})$ for the current set of compounds, was selected according to the breaking point rule for the improvement of R^2 as demonstrated in Figure 1.

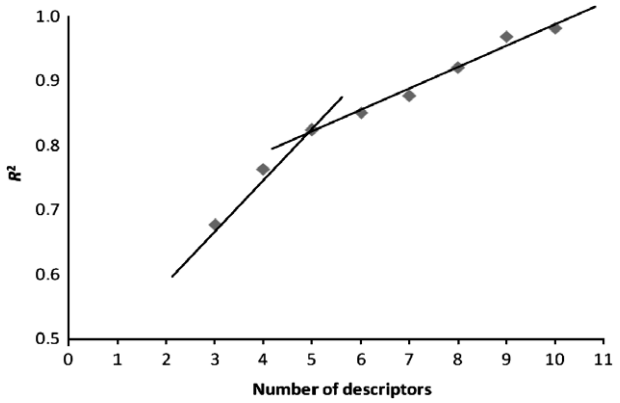


Figure 1. Plot of correlation coefficients R^2 versus number of descriptors (breaking point rule).

Here, we can see that R^2 rises steeply as the number of parameters increases from 3–10 and the breaking point occurs at the fifth descriptor. Therefore, we used the best correlation equation with five descriptors, shown in Table 2, for the basic analysis. The respective plot of calculated versus experimental $-\log(EC_{50})$, as molar, data is presented in Figure 2.

Molecular descriptors collected from the derived QSAR equation consist in one constitutional (relative number of Cl atoms, D_3), two geometric (YZ shadow, D_1, YZ shadow/YZ rectangle, D_4) and two quantum-chemical (minimum 1-electron reactivity index for a C atom, D_2, average 1-electron reactivity index for a C atom, D_5). Even if it is often difficult to identify the physical meaning of each molecular descriptor that appears in a QSAR/QSPR model, an analysis of equation in Table 2 reveals some interesting features. According to the t-test values ($|t|$), the importance of the descriptors involved in the model decreases in the following order: $D_1 > D_2 > D_3 > D_4 > D_5$.

3.1.1. Geometrical descriptors. The most statistically significant descriptor is the YZ shadow (D_1), which has a geometrical origin and can be related to the size and shape of molecules. This descriptor is easily explained when we assumed that the orientation of the compound in the space is done according to its principal moments of inertia and that the molecule is viewed from three orthogonal directions defined by the X, Y and Z coordinate axes.[38] Thus, for the perspective along the X axis, the X coordinates would be disregarded and the molecule projected onto the YZ plane, illustrated in Figure 3.

The area corresponding to this projection composes our geometrical index, i.e., the YZ shadow (natural shadow). Also, this area can be normalized by dividing the value of the shadow's area by the area of the rectangle defined by the maximum dimensions of the projection on the plane. This new value consists in the normalized shadow or, in this case, YZ shadow/YZ rectangle which is another geometrical descriptor. Therefore, we can say these indices reflect the size (natural shadow indices) and geometrical shape (normalized shadow indices) of the molecule. Our QSAR model also

Table 2. The best five-descriptor model for $-\log(EC_{50})$: $N = 23$, $R^2 = 0.825$, $R^2_{CV} = 0.715$, $F = 16.05$, $s^2 = 0.132$

Number	X	$\pm\Delta X$	t-test	Descriptor
0	−7.192	2.909	−2.476	Intercept
1	0.119	0.021	5.627	YZ shadow, D_1
2	−62.25	11.27	−5.545	Min 1-electron reactivity index for a C atom, D_2
3	26.74	6.268	4.275	Relative number of Cl atoms, D_3
4	9.987	3.282	3.047	YZ shadow/YZ rectangle, D_4
5	−625.2	255.4	−2.451	Average 1-electron reactivity index for a C atom, D_5

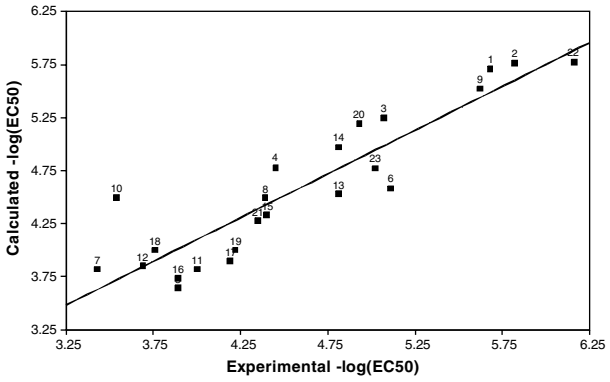


Figure 2. Calculated versus experimental $-\log(EC_{50})$ based on the five-parameter correlation equation (with a $R^2 = 0.825$) from 2D-QSAR model for the full data set.

presents this descriptor, D_4, which appears here to be inter-correlated ($R^2 = 0.744$) to the YZ shadow.

The positive coefficient in the regression equation for both descriptors reported above indicates that we can reach better inhibitory activities with the increase of these two indices values. A better regard provides us some interesting aspects about the shape of the compounds of our data set. Indeed, we observed that a carboxylate group in *meta* position on the phenyl ring tends to enhance the activity of molecules. This fact can be observed if we compare the compounds $\mathbf{A}_{16}$ an $\mathbf{A}_{18}$ with $-\log(EC_{50})$ values of 3.76 and 4.93, respectively. The only difference between these two molecules is the position of the carboxylate group and its evident that $\mathbf{A}_{16}$, with the substituent in *para*, reveals lower inhibitory activity than $\mathbf{A}_{18}$. This observation allows the thinking that a *meta*-substitution by an electron withdrawing group on the phenyl ring may be important for a better orientation of the molecule in the binding site of the protein. Also, the molecules with higher inhibitory activities show another geometrical aspect that must be pointed out. This one consists in the tendency of the pyrrole ring to be placed perpendicular to the aromatic. The molecules $\mathbf{A}_{20}$ and $\mathbf{A}_{17}$ present different $-\log(EC_{50})$ values, 6.16 and 4.22, respectively, and can illustrate this feature (Fig. 3). Effectively, the observation of their YZ shadow permits us to distinct the pyrrole perpendicular to phenyl ring for $\mathbf{A}_{20}$, contrary to $\mathbf{A}_{17}$.

These two geometric descriptors inform about the structural characteristics related to size and shape of compounds, proving that the hydrophobic and steric interactions are important for binding between the ligands and protein.

3.1.2. Quantum-chemical descriptors. The minimum 1-electron reactivity index for a C atom ($RI^{1e}_{min,C}$) is defined as:

$$RI^{1e} = \sum_{i\in A}\sum_{j\in A} c_{i\mathrm{HOMO}}c_{j\mathrm{LUMO}}/(\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}})$$

where the summations are performed over all atomic orbitals i, j at the given atom (A), $c_{i\mathrm{HOMO}}$ and $c_{j\mathrm{LUMO}}$ denote the i-th and j-th AO coefficients on the highest occupied molecular orbital (HOMO) and the lowest-unoccupied molecular orbital (LUMO), respectively, and $\varepsilon_{\mathrm{LUMO}}$ and $\varepsilon_{\mathrm{HOMO}}$ are the energies of these orbitals.

The difference of E_{HOMO} and E_{LUMO}, termed the *band gap* (or HOMO–LUMO gap), usually serves as a measure of the reactivity of the molecule: the smaller the

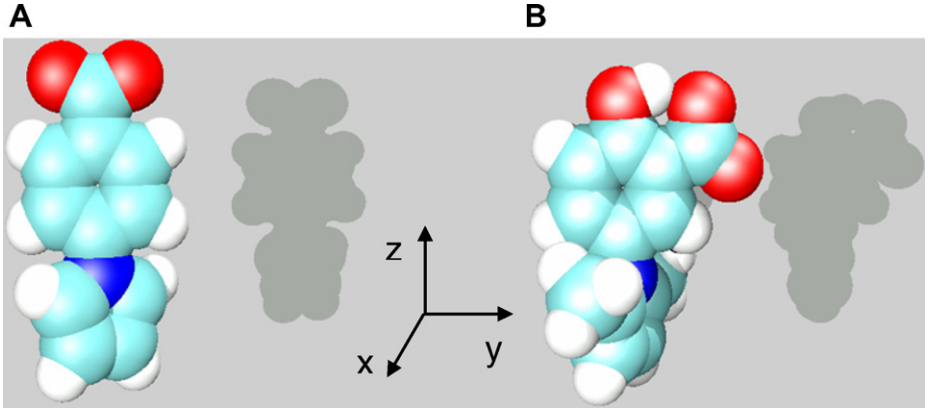


Figure 3. Representation of the projection of the YZ-shadow of a less and the most active compound $\mathbf{A}_{17}$ (A) and $\mathbf{A}_{20}$ (B), respectively.

energy, the more reactive it will be. Reactivity indices are essentially Fukui functions[39] and are implicitly designed to describe chemical reactions. However, we infer that they serve, here, to quantify interactions between the charged groups present in the ligands and protein. The negative coefficient in the model implies that a more negative value of this descriptor in the model leads to the increase of the inhibitory activity. Effectively, this is clearly demonstrated if we compare the molecules $\mathbf{A}_{20}$ and $\mathbf{B}_1$ with values of $-\log(EC_{50})$ of 6.16 and 3.69, respectively. The first one, which presents a more negative value for this descriptor (i.e., smaller *band gap*) is more active than $\mathbf{B}_1$.

One other quantum-chemical descriptor that appears in our model is the average 1-electron reactivity index for a C atom. Also the negative coefficient in the model suggests that more negative values of this descriptor induce higher values for activity. This enhances the importance of the reactivity of C atoms for the increase of inhibitory activity.

3.1.3. Constitutional descriptor. Relative numbers of Cl atoms (RNcA is the ratio between the number of Cl atoms and the total number of atoms), a constitutional descriptor, is related to the polarizability of the compounds. Presence of chlorine atoms enhances the polarity of the molecules. It is important to notice that all the ligands in the data set having chlorine atoms show good activities. The positive coefficient in the regression equation indicates that increasing the number of Cl atoms leads to higher inhibitory activity. For example, the molecules $\mathbf{A}_9$ and $\mathbf{A}_{19}$ are the same with the only difference that the first one presents a Cl atom in position *para* on the phenyl ring. The presence of the chlorine is sufficient to induce divergent values of activity for the related compounds. Indeed, the values of $-\log(EC_{50})$ are 5.62 and 4.35, respectively. The same observation can be done if we consider the molecules $\mathbf{A}_2$ and $\mathbf{A}_{18}$. Here, the variation between the $-\log(EC_{50})$ values, 5.82 and 4.93, respectively, are smaller than the previous example but still remains significative. Once the chlorinated compounds always present Cl atom in *para*, we cannot deduce about the importance of the location of the halogen in the phenyl ring. For that, it will be interesting to test some molecules having the Cl in *ortho* and *meta* so we can conclude about a strategic positioning.

3.1.4. The model validation. The first technique applied for the validation of the proposed five-parameter model was based on the leave-one-out algorithm (LOO). This method gives a squared cross-validated correlation coefficient (R^2_{CV}) value 0.715.

For further internal validation,[40] the data set was divided into three subsets A, B and C (the first, fourth, seventh,... entries go into subset A, the second, fifth, eighth,... into subset B, and the third, sixth, ninth,... into the subset C). For each of three combinations, two of the sets are combined into one and the correlation equation was derived with the same descriptors. The obtained equation was used to predict data for the remaining subset. The results of the internal validation are listed in Table 3 and show that the predicted R^2 values are in good agreement with our original QSAR model.

3.2. CoMFA analysis

The CoMFA model was developed using the whole data set listed in Table 1. The biological activities of the molecules, expressed as $-\log(EC_{50})$, were related to the independent variable using the PLS methodology. For an optimal number of components (N) of 8 the best 3D-QSAR model, obtained with combined steric/electrostatic fields that yields a leave-one-out cross-validated correlation coefficient q^2 of 0.463, a non-cross-validated R^2 of 0.984, an estimated F value of 110.472 and a standard error of estimate s of 0.119.

The results obtained for the different statistical parameters of CoMFA indicate the statistical validity and stability of the developed model. The contribution of steric to electrostatic fields was found to be 42:58. The plot of experimental values versus predicted values of $-\log(EC_{50})$ is shown in Figure 4 and the values of each compound are in Table 4.

The information content in the CoMFA model is usefully portrayed as three-dimensional coefficient contour maps.[41] Mainly, these polyhedra surround all lattice points where variability in molecules' fields is able to explain changes in experimental biological property. More specifically, the contour maps were generated by interpolating the products between the 3D-QSAR coefficients and their associated standard deviations. The CoMFA contour maps were obtained from the analysis based on molecular alignment for the best model derived by steric/electrostatic field combination. One of the most active compound ($\mathbf{A}_{20}$) and the least one ($\mathbf{A}_7$) of the data

Table 3. Internal validation of the five-parameter 2D-QSAR model for 23 inhibitors

Training set	R^2 (fit)	s^2 (fit)	Test set	R^2 (pred)	s^2 (pred)
A + B	0.849	0.155	C	0.713	0.112
A + C	0.754	0.201	B	0.959	0.032
B + C	0.933	0.053	A	0.733	0.181

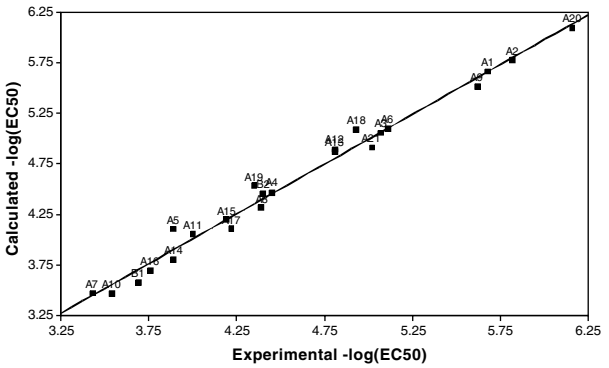


Figure 4. Plot of predicted versus experimental values of $-\log(EC_{50})$ for best CoMFA model with a $R^2 = 0.984$.

Table 4. Experimental and predicted values of $-\log(EC_{50})$ obtained with the best CoMFA model

Compound	$-\log(EC_{50})$ experimental	$-\log(EC_{50})$ calculated
A_1	5.68	5.66
A_2	5.82	5.78
A_3	5.07	5.06
A_4	4.45	4.47
A_5	3.89	4.11
A_6	5.11	5.10
A_7	3.43	3.47
A_8	4.39	4.32
A_9	5.62	5.51
A_{10}	3.54	3.47
A_{11}	4.00	4.05
A_{12}	4.81	4.89
A_{13}	4.81	4.87
A_{14}	3.89	3.80
A_{15}	4.19	4.20
A_{16}	3.76	3.70
A_{17}	4.22	4.11
A_{18}	4.93	5.09
A_{19}	4.35	4.54
A_{20}	6.16	6.10
A_{21}	5.02	4.91
B_1	3.69	3.58
B_2	4.40	4.45

set were superimposed on both maps in order to understand the different activity profiles of these molecules.

The representation illustrated in Figure 5 shows the electrostatic maps contribution of the best model generated by CoMFA, in which the blue and red contours correspond to regions where an increase in positive or negative charge, respectively, will enhance the bioactivity profile. The blue and red contour maps are mainly located around the upper moieties attached to the phenyl ring, suggesting their importance for the electrostatic interactions with the presumed receptor. When we analyze deeply the maps around compound A_7 and compare them to those around compound A_{20}, it is clear that the carboxylate of the phenyl ring of A_7 falls in a region where an increase in positive charge is activity enhancing. This may be one of the structural reasons for a low activity of this molecule once there are no electron-rich groups in the molecular structure of A_7 pointing toward this region. Additionally, the hydroxyl group of A_{20} falls in the red region of the contour maps and the carboxylate group is pointed toward a red one too.

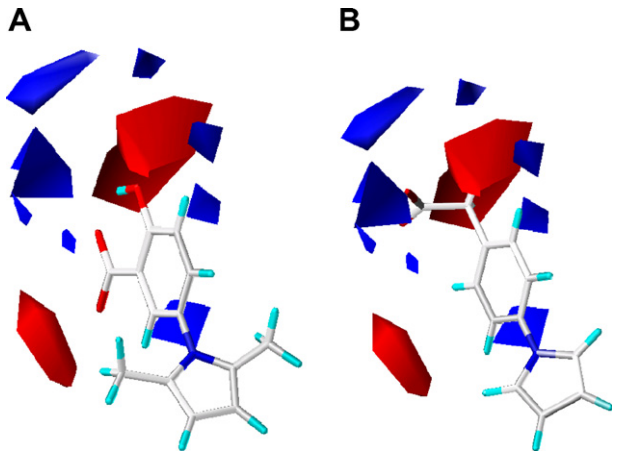


Figure 5. Electrostatic contour maps of the most active compound and a less one A_{20} (A) and A_7 (B), respectively.

These features are probably responsible for the good activity showed by this derivative. Once our data set mainly presents electronegative groups as substituents on the phenyl ring (at both *para* and *meta* position) it will be interesting to test some positive groups. By this way we will be able to infer about their positioning toward the blue regions and their capability in enhancing the biological activity.

Figure 6 shows the steric maps contribution of the best 3D-QSAR model. Green contours indicate regions where an increase in steric bulk will enhance activity, and yellow contours indicate regions where an increase in steric bulk reduces activity. The steric contour maps are situated near the adjacent substituents of pyrrole and around the upper moieties attached to the phenyl ring, telling that the addition of bulky groups in this region will cause a reduction in activity. The analysis of these maps around compound A_7 permitted us to visualize that the carboxylate group of the phenyl ring falls in a sterically unfavoured region, which should contribute destructively for the inhibitory activity. Once A_{20} does not present any bulky groups that are in a steric clash this could be a probable raison for its high activity. Also it should be noticed the absence of green contour maps. This may be due to the fact that CoMFA methodology is logically quite insensitive to subunits which do not present substituents.

However the lack of contour maps in these areas does not necessarily mean that the occupation of such regions in space will not result in steric or electrostatic terms significantly correlated to activity. So, suggested by the position of the yellow maps, it will be interesting to increase the molecular structure where there are no unfavourable steric regions.[42]

Finally the data are indicative of the specific electrostatic/steric requirements of the binding pocket of the hypothetical target bioreceptor.

3.3. Molecular docking analysis

In order to link the information obtained by QSAR studies to the position of pyrrole derivatives into the active site of gp41 we performed docking calculations on 23 molecules using AutoDock program. The region of gp41 chosen for these calculations was suggested by Jiang and collaborators.[27] Effectively, they demonstrated that two related compounds of our data set fit into a hydrophobic groove and how these molecules interact with the amino acids residues within gp41 pocket and surrounding area. Successively, we docked our inhibitors into the same region of gp41. We know that the selection of docked conformations is often complicated as the results are particularly sensitive to the

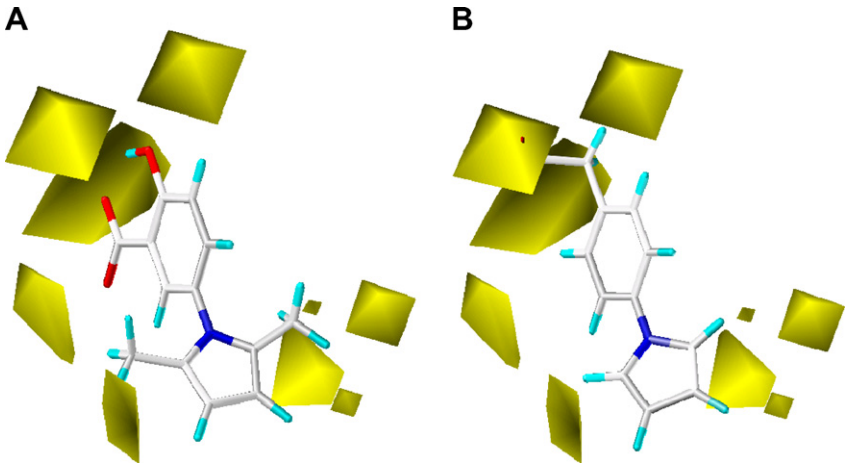


Figure 6. Steric contour maps, generated by CoMFA, of the most active compound and a less one $\mathbf{A}_{20}$ (A) and $\mathbf{A}_{7}$ (B).

scoring function. Here, the conformation selected was the one which present the lowest-docking energy in the most populated cluster.

The poses in the cluster for all analogues have similar binding positions and the same orientation with each other in the docking complex. Also most of compounds of our data set present very similar conformation to that derived by using minimization in Sybyl. This fact did not surprise us once the ligands do not have more than three dihedral angles. The binding conformation of most active $\mathbf{A}_{20}$ inside the active site of gp41 is shown in Figure 7.

When we look deeply, we can observe that the docked conformations are located in a groove defined by the following residues of gp41: Trp571, Gln575, Arg579, Leu581, Gln577, Lys574 and Ile573. At this position we are able to distinguish which are the main interactions between the binding pocket and the docked ligands. In effect, the phenyl ring is placed in the groove perpendicularly to the protein surface (Figs. 7 and 8) making T-shape interactions with the aromatic ring of Trp571 whereas the pyrrole ring is parallel to the molecular surface forming hydrophobic interactions.

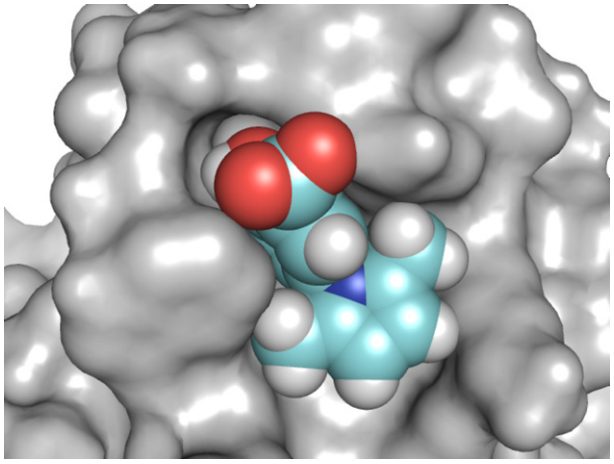


Figure 8. Surface representation of the hydrophobic pocket of gp41 with the docked compound $\mathbf{A}_{20}$.

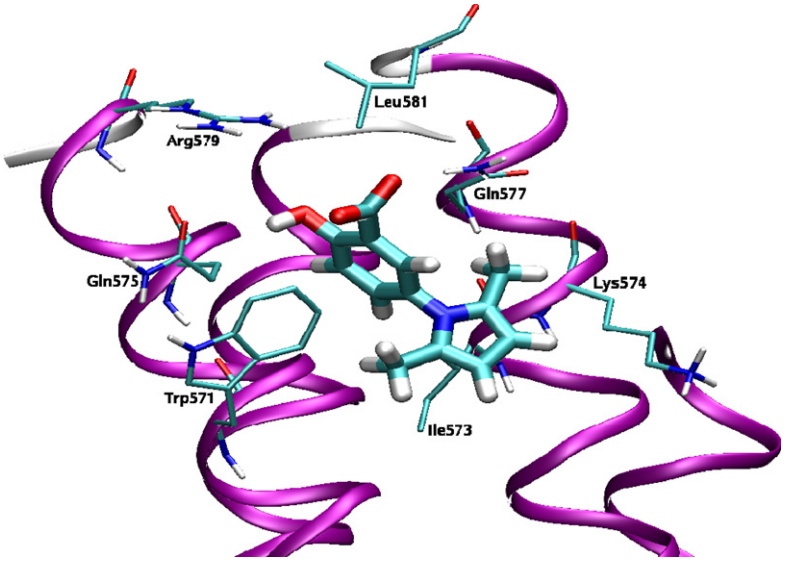


Figure 7. The docking conformation of most active compound $\mathbf{A}_{20}$ inside the hydrophobic pocket of gp41.

These two features are complementary with the results obtained with 2D-QSAR once the *YZ* shadow also indicated a preference by the perpendicularity of the two aromatic rings with each other. We can still observe electrostatic interactions between carboxylate (Fig. 7) group of $\mathbf{A}_{20}$ and Arg579 as well hydrogen bond between the hydroxyl group and Gln575. Once again these results are in agreement with those obtained from 3D-QSAR analysis. We clearly remark that the R_1 and R_2 groups oriented towards the electropositive polar region of the binding site. Therefore, as also suggested by the electrostatic contour maps, substituting R_1 and R_2 positions with electronegative groups may allow them to interact with the arginine residue, which is favourable for the electrostatic interactions and consequently for the activity. However it should be noticed that this substitution is limited by the distance of ligands to molecular surface. Thus the substituents R_1 and R_2 should not be too large so that they do not make disfavoured steric interactions as also recommended by steric contour maps from CoMFA analysis.

4. Conclusions

Here, we report the systematic 2D, 3D-QSAR and docking studies on pyrrole derivatives as HIV-1 gp41 inhibitors. The activity of the ligands was discussed by analyzing the physico-chemical meaning of the descriptors involved in the QSAR model. These descriptors could be essentially related to the shape (*YZ* shadow) and electron reactivity for a C atom concluding that an increase of the first property and a decrease of the last one are contributing for the activity. The consistency of the 2D-QSAR model was verified using the R^2_{CV} leave-one-out and internal cross-validation.

CoMFA analysis was successfully applied to the set of 23 structures. Reasonable q^2 and good R^2 values were obtained for the 3D-QSAR indicating that the model constructed has a good correlative and predictive property. The docking analysis provides a qualitative representation of ligand and protein interactions, which are complementary with CoMFA maps and the 2D-QSAR model. Even if the CoMFA and docked conformations do not really correspond, they characterize fundamental features (areas of the active site) of ligand–protein interactions. Thus, both QSAR and docking studies indicate that the substitution of electron-rich groups on the R_1 and R_2 position of phenyl ring may lead to improved biological activity of pyrrole derivatives. So we can see the significance that it is to verify the ligand-based contours consistent with the binding site feature of their receptor is important. Thus when the correct binding mode is gotten from the docking simulation and correlated with the results from ligand-based drug design, it will enhance the approach and give a better chance to find leads and improve activity.

In summary, the present work, based on 2D, 3D-QSAR and docking simulations lies in the first study in this area for HIV-1 gp41 inhibitors. The proposed models and the characterization of some ligand–protein interactions provides the first step toward the prediction of novel active compound and suggestions to modify/increase ligands molecular structure in order to enhance their biological activity. Some of the new suggested compounds are already being synthesized.

Acknowledgments

We are grateful for the financial support from Fundação para a Ciência e a Tecnologia (Portugal) for the PhD fellowship SFRH/BD/22190/2005 to Cátia Teixeira, from the French Ministry of Research and Technology for the PhD fellowship to Joseph Rebehmed, from Ministry of Science and Technology in China (2006DFA3356 and 2006AA02Z319) to Lan Xie and the US NIH grant (RO1AZ46221) to Shibo Jiang.

References and notes

1. http://www.unaids.org/en/HIV_data/default.asp, 2007.
2. Fauci, A. S. *Nat. Med.* **2003**, *9*, 839.
3. De Clercq, E. *J. Clin. Virol.* **2004**, *30*, 115.
4. De Clercq, E. *J. Clin. Virol.* **2001**, *22*, 73.
5. Cohen, J. *Science (New York, NY)* **2002**, *296*, 2322.
6. Detels, R.; Munoz, A.; McFarlane, G.; Kingsley, L. A.; Margolick, J. B.; Giorgi, J.; Schrager, L. K.; Phair, J. P. *JAMA* **1998**, *280*, 1497.
7. Hogg, R. S.; Heath, K. V.; Yip, B.; Craib, K. J.; O'Shaughnessy, M. V.; Schechter, M. T.; Montaner, J. S. *JAMA* **1998**, *279*, 450.
8. Rousseau, M. N.; Vergne, L.; Montes, B.; Peeters, M.; Reynes, J.; Delaporte, E.; Segondy, M. *J. Acquir. Immune Defic. Syndr.* **2001**, *26*, 36.
9. Viraben, R.; Aquilina, C. *AIDS (London, England)* **1998**, *12*, F37.
10. Carr, A.; Samaras, K.; Burton, S.; Law, M.; Freund, J.; Chisholm, D. J.; Cooper, D. A. *AIDS (London, England)* **1998**, *12*, F51.
11. Nisole, S.; Saib, A. *Retrovirology* **2004**, *1*, 9.
12. Freed, E. O.; Martin, M. A. *J. Biol. Chem.* **1995**, *270*, 23883.
13. Berger, E. A.; Murphy, P. M.; Farber, J. M. *Annu. Rev. Immunol.* **1999**, *17*, 657.
14. Doms, R. W.; Moore, J. P. *J. Cell Biol.* **2000**, *151*, F9.
15. Kwong, P. D.; Wyatt, R.; Robinson, J.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A. *Nature* **1998**, *393*, 648.
16. Eckert, D. M.; Kim, P. S. *Annu. Rev. Biochem.* **2001**, *70*, 777.
17. Robertson, D. *Nat. Biotechnol.* **2003**, *21*, 470.
18. Wild, C. T.; Shugars, D. C.; Greenwell, T. K.; McDanal, C. B.; Matthews, T. J. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 9770.
19. Furuta, R. A.; Wild, C. T.; Weng, Y.; Weiss, C. D. *Nat. Struct. Biol.* **1998**, *5*, 276.
20. Jiang, S.; Lin, K.; Strick, N.; Neurath, A. R. *Nature* **1993**, *365*, 113.
21. Jiang, S.; Lin, K.; Zhang, L.; Debnath, A. K. *J. Virol. Meth.* **1999**, *80*, 85.
22. Chan, D. C.; Chutkowski, C. T.; Kim, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15613.
23. Chan, D. C.; Fass, D.; Berger, J. M.; Kim, P. S. *Cell* **1997**, *89*, 263.
24. Debnath, A. K.; Radigan, L.; Jiang, S. *J. Med. Chem.* **1999**, *42*, 3203.
25. Zhao, Q.; Ernst, J. T.; Hamilton, A. D.; Debnath, A. K.; Jiang, S. *AIDS Res. Hum. Retroviruses* **2002**, *18*, 989.

26. Jiang, S.; Debnath, A. K. *Biochem. Biophys. Res. Commun.* **2000**, *270*, 153.
27. Jiang, S.; Lu, H.; Liu, S.; Zhao, Q.; He, Y.; Debnath, A. K. *Antimicrob. Agents Chemother.* **2004**, *48*, 4349.
28. Xie, L., et al., unpublished.
29. Sztandera, L.; Garg, A.; Hayik, S.; Krishna, L. B.; Bock, C. W. *Dyes Pigments* **2003**, *59*, 117.
30. Bauvais, C.; Barbault, F.; Zhu, Y.; Petitjean, M.; Fan, B. T. *SAR QSAR Environ. Res.* **2006**, *17*, 253.
31. Sova, M.; Perdih, A.; Kotnik, M.; Kristan, K.; Rizner, T. L.; Solmajer, T.; Gobec, S. *Bioorg. Med. Chem.* **2006**, *14*, 7404.
32. Sybyl 7.3, Tripos Software, St. Louis, USA.
33. Dewar, M. J. S.; Zoebish, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
34. Steward, J. J. P., MOPAC 93 annual revision Number 2, Fujitsu, 1993.
35. CODESSA Software, University of Florida, 1996.
36. Kratritzky, A. R.; Lobanov, V. S.; Karelson, M., CODESSA Reference Manual, University of Florida, Gainesville, 1996.
37. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998**, *19*, 1639.
38. Rohrbaugh, R. H.; Jurs, P. C. *Anal. Chim. Acta* **1987**, *199*, 99.
39. Fukui, K.; Yonezawa, T.; Haruo, S. *J. Chem. Phys.* **1952**, *20*, 722.
40. Xu, Q.-S.; Liang, Y.-Z. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1.
41. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
42. Oprea, T. I.; Waller, C. L.; Marshall, G. R. *Drug. Des. Discovery* **1994**, *12*, 29.