Bioorganic & Medicinal Chemistry 17 (2009) 2400–2409

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Receptor- and ligand-based 3D-QSAR study for a series of non-nucleoside HIV-1 reverse transcriptase inhibitors

Rongjing Hu [a,b], Florent Barbault [b], Michel Delamar [b,*], Ruisheng Zhang [a,c,*]

[a] Department of Chemistry, Lanzhou University, Lanzhou, Gansu 730000, PR China
[b] Interfaces, Traitement, Organisation et Dynamique des Systèmes (ITODYS), Paris-Diderot (Paris 7) University, CNRS UMR 7086, Bâtiment Lavoisier, 15 rue Jean Antoine de Baïf, 75205 Paris Cedex 13, France
[c] School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, PR China

ARTICLE INFO

Article history:
Received 3 October 2008
Revised 16 January 2009
Accepted 5 February 2009
Available online 8 February 2009

Keywords:
QSAR
CoMFA
CoMSIA
Docking
NNRTIs
HIV-1 reverse transcriptase

ABSTRACT

Molecular modeling of a series of HIV reverse transcriptase (RT) non-nucleoside inhibitors (2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners) was carried out by comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) approaches. Docking simulations were employed to position the inhibitors into RT active site to determine the most probable binding mode and most reliable conformations. The study was conducted using a complex receptor-based and ligand-based alignment procedure and different alignment modes were studied to obtain highly reliable and predictive CoMFA and CoMSIA models with cross-validated q^2 value of 0.723 and 0.760, respectively. Furthermore, the CoMFA and CoMSIA contour maps with the 3D structure of the target (the binding site of RT) inlaid were obtained to better understand the interaction between the RT protein and the inhibitors and the structural requirements for inhibitory activity against HIV-1. We show that for 2-amino-6-arylsulfonylbenzonitriles inhibitors to have appreciable inhibitory activity, bulky and hydrophobic groups in 3- and 5-position of the B ring are required. Moreover, H-bond donor groups in 2-position of the A ring to build up H-bonding with the Lys101 residue of the RT protein are also favorable to activity.

1. Introduction

The HIV type 1 reverse transcriptase (HIV-1 RT) is an essential viral enzyme in HIV-1 virus. After HIV-1 infects a cell, RT plays a central role in the viral replication cycle by catalyzing the conversion of genomic single-stranded RNA into double-stranded proviral DNA. The viral DNA is then integrated into the host chromosomal DNA which then allows host cellular processes, such as transcription and translation to reproduce the virus. HIV-1 RT is an asymmetric heterodimer consisting of two polypeptides, p66 and p51.[1]

Due to its importance, HIV-1 RT has been one of the major targets of the antiretroviral drugs. Nucleoside and non-nucleoside RT inhibitors (NRTIs and NNRTIs) are two different categories:[2–4] the former is an analogue of the natural deoxynucleotides that compete with the natural deoxynucleotides for incorporation into the growing viral DNA chain; the latter noncompetitively inhibits the movement of protein domains of RT that are needed to carry out the process of DNA synthesis. Nevirapine is a NNRTI used to treat HIV-1 infection and AIDS, which was discovered by Hargrave et al.[5] Later, different types of NNRTIs with similar interaction between inhibitors and RT protein have been discovered or designed.[6,7] The high bioactivity and effectivity of nevirapine and other inhibitors proved that RT is an important and significant target to design anti-HIV drugs.

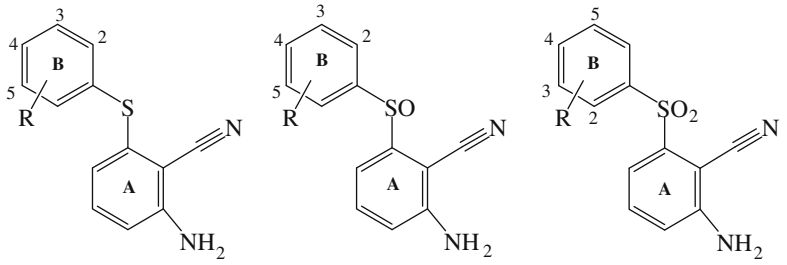
A series of NNRTIs (2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners) (Table 1) were designed by Chan et al.[8] Because of their high activity and low toxicity, researchers have studied them in order to get more information about the activity and to design more potential anti-HIV-1 drugs.[9–11] In the present work, a receptor-guided and ligand-based three-dimensional quantitative structure–activity relationship (3D-QSAR) study was carried out for the first time for this series of inhibitors. Receptor-guided QSAR is only available when the 3D structures of a target protein or its homologue bound to the active compound have been experimentally solved. We thus performed a docking study to search for reasonable conformations and alignments for CoMFA and CoMSIA studies and gain an insight into the interaction between ligands and RT protein.

CoMFA relates the bioactivity of inhibitors with their steric and electrostatic fields sampled at grid points defining a 3D box around molecules.[12] CoMSIA is most commonly applied in drug discovery

* Corresponding authors. Tel.: +33 1 57 27 54 32; fax: +33 1 57 27 72 63 (M.D.), Tel.: +86 931 8914000 8421 (R.Z.).
E-mail addresses: michel.delamar@univ-paris-diderot.fr (M. Delamar), zhangrs@lzu.edu.cn (R. Zhang).

0968-0896/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2009.02.003

Table 1
Observed and predicted anti-HIV-1 activity of 2-amino-6-arylsulfonylbenzonitriles and their congeners by CoMFA and CoMSIA



Sulfide family: 1-23 Sulfoxide family: 24-38 Sulfone family: 39-68

No.	R substituent	Observed pIC_{50}	Predicted	
			CoMFA	CoMSIA
1	H	1.836	1.469	1.847
2	2-OCH_3	2.367	1.898	2.128
3	3-OCH_3	2.222	1.704	2.37
4	2-CH_3	1.796	1.794	1.915
5	3-CH_3	2.215	2.102	1.879
6	4-CH_3	0.939	1.632	1.112
7*	2-Cl	2.387	1.792	1.459
8*	3-Cl	2.131	1.923	1.715
9#	4-Cl	–	1.530	–
10	2-Br	1.523	1.768	1.39
11*	3-Br	2.292	2.019	1.41
12	3-F	2.009	2.376	1.873
13#	2-CN	–	1.860	–
14	3-CN	2.762	2.111	2.888
15	4-CN	1.359	1.454	1.495
16	3-CF_3	1.893	1.992	1.563
17	3-NH_2	1.502	1.987	1.53
18#	2,5-Cl_2	–	2.752	–
19	3,5-$(CH3)_2$	3.367	3.178	3.388
20#	3,5-Cl_2	–	2.979	–
21	3-Cl,5-CH_3	2.754	3.177	3.082
20#	3-OCH3,5-CH_3	2.699	2.744	3.751
23	3-OCH3,5-CF_3	2.292	2.134	2.403
24	2-OCH_3	2.319	2.447	2.368
25	3-OCH_3	1.796	1.631	1.833
26*	2-CH_3	1.032	1.574	1.567
27	3-CH_3	1.534	2.149	1.902
28	4-CH_3	1.310	1.191	1.275
29*	2-Br	1.407	2.305	1.206
30*	3-Br	4.097	2.671	2.78
31*	4-Br	1.694	0.958	1.283
32	2-CN	2.409	1.826	2.254
33*	3-CN	1.848	2.902	1.881
34	3-CF_3	1.398	1.315	1.687
35	3,5-$(CH_3)_2$	3.469	3.116	3.353
36	2,5-Cl_2	2.007	2.617	2.124
37	3-Cl,5-CH_3	3.495	3.127	3.027
38	3-OCH_3,5-CF_3	2.684	3.237	2.569
39	H	2.699	2.526	2.403
40*	2-OCH_3	3.222	2.639	2.849
41	3-OCH_3	3.046	2.975	2.787
42	4-OCH_3	1.602	1.134	1.613
43	2-CH_3	2.638	2.553	2.525
44*	3-CH_3	3.398	3.881	4.071
45	4-CH_3	2.022	1.793	1.852
46	2-Cl	2.387	2.572	2.398
47	3-Cl	3.229	3.561	3.511
48*	4-Cl	2.523	1.781	1.973
49	2-Br	2.301	2.548	2.385
50	3-Br	3.268	2.885	3.108
51	4-Br	1.699	1.77	2.021
52	2-F	2.523	2.514	2.343
53	3-F	2.523	3.064	2.776
54	2-CN	2.268	2.477	2.436
55	3-CN	2.62	2.815	2.547
56	4-CN	1.097	1.398	0.959
57	3-CF_3	2.456	2.743	2.746
58	2,5-Cl_2	3.523	3.276	3.54

(*continued on next page*)

Table 1 (*continued*)

No.	R substituent	Observed pIC_{50}	Predicted	
			CoMFA	CoMSIA
59	3,5-Cl_2	4.155	3.913	4.11
60	(739W94)3,5-$(CH_3)_2$	5.000	4.369	4.694
61	3-Br,5-CH_3	4.699	4.206	4.815
62*	3-Cl,5-CH_3	4.523	4.165	4.598
63	3-OCH_3,5-CH_3	4.301	4.334	4.295
64	3-OCH_3,5-CF_3	4.046	4.082	3.992
65	3-OH,5-CH_3	3.367	3.607	3.393
66	3-OCH_2CH_3,5-CH_3	4.222	4.322	4.434
67	3-$O(CH_2)_2CH_3$,5-CH_3	4.222	3.929	4.025
68	3-$O(CH_2)_3CH_3$,5-CH_3	3.222	3.594	3.427

Compounds labeled with '*' are the test set; Compounds labeled with '#' are the predicted set; other compounds are the training set.

to find the common features that are important in binding to the relevant biological receptor, where steric and electrostatic features, hydrogen bond donor and acceptor and hydrophobic fields are considered.[13] Based on docking, a complex molecular alignment procedure makes CoMFA and CoMSIA models more reliable.

CoMFA and CoMSIA contour maps were extracted and superimposed to the 3D protein structure to get more interactional information of substituents and the bioactivity of corresponding molecules. Our final high quality correlation between actual activity and predicted values and the similarity in ligand–protein interactions, either from docking results or from 3D-QSAR results, proved that the receptor-guided and ligand-based 3D-QSAR method is a powerful tool to design more potent HIV-1 RT inhibitors.

2. Methods

2.1. Data

The 2-amino-6-arylsulfonylbenzonitriles and congeners data, represented by anti-HIV-1 activity IC_{50} (μM), were obtained from published data.[8] The structures and bioactivity values of potential inhibitors are presented in Table 1. The pIC_{50} ($-\log IC_{50}$) values were used to derive 3D-QSAR models. The whole data set of 68 compounds was separated into two groups in the approximate ratio 4:1: a training set with 51 compounds, a test set with 13 compounds and a prediction set with compounds **9**, **13**, 18 and **20** for which no experimental data are available (Table 1). The selection of the training and test sets was done manually such that low, moderate and high anti-HIV activity compounds were present in roughly equal proportions in both sets. The training set was used to build predictive models, while the test set was used to validate the predictive ability of the models. The activities of molecules in the prediction set were predicted in the present work.

2.2. Preparation of ligands and protein

Among the inhibitors, the crystal structure of compound **60** complexed with HIV-1 RT is known. Therefore, the complex of HIV-1 RT receptor and 739W94 (compound **60**) was extracted from the protein data bank (PDB code: 1JLQ).[14] 739W94 and RT receptor were isolated from the complex.

Compound **60** was chosen to define the most likely binding conformation in molecular docking. It was modified to add hydrogen atoms without any change of conformation and was minimized using the molecular modeling software Sybyl 8.0 with the following steps: (i) optimization by Steepest Descent with initial optimization of 20 simplex iterations using Tripos force field[15] and Gasteiger–Marsili charges;[16] (ii) optimization by conjugate gradient (iii) optimization by BFGS.

Three-dimensional structures of the other molecules were constructed from compound **60**. Energy minimizations of each compound were performed according to the above procedure. The receptor isolated from the crystal structure was treated by adding H and Gasteiger–Marsili charges and optimized using Powell conjugate gradient algorithm and the AMBER7 FF99 force field with a distance-dependent dielectric function.

2.3. Molecular docking

AutoDock4 was used to perform automated docking of ligands to their macromolecular protein receptor. A grid with spacing of 0.375 Å and $60 \times 60 \times 60$ points was created. Energy grid maps for all possible ligand atom types were calculated before performing docking. While keeping the protein structure rigid, we selected Lamarckian genetic algorithm (LGA) to search the conformational and orientational space of the ligands, because LGA was proved[17] to be the most efficient, reliable and successful technique as compared to Monte Carlo simulated annealing, or a traditional genetic algorithm. Compound **60**, which conformation was generated from 739W94, was first docked into the binding pocket of RT in order to compare docking results with the crystal structure and check the validity of the docking procedure. As this comparison was successful, the other compounds were then docked to RT. 100 Independent docking runs were performed, then the 100 solutions were clustered into groups with RMS deviations lower than 0.5 Å. The clusters were ranked according to the lowest energy representative of each group.

2.4. 3D-QSAR: CoMFA and CoMSIA

Several factors influence the modeling results of CoMFA and CoMSIA: among them, the most important are alignments and fields. Molecular alignment is a prelude to CoMFA and CoMSIA. It is necessary to align molecules in a common orientation relative to a template compound in order to compare the different features of analogues. A CoMFA field is then generated by creating a grid around molecules and calculating the steric and electrostatic potential at each point on the grid using a charged probe atom.

2.4.1. Alignments for CoMFA and CoMSIA

In CoMFA and CoMSIA studies, the position of a molecule is important because the descriptors are calculated based on the coordinates of atoms, hence, different methods of alignment will give different results. There are three main different procedures proposed for aligning molecules for QSAR: substructure overlap, pharmacophore overlap and docking-based alignment.

In the present work, first, alignment based on the geometries obtained from docking was applied to build a receptor-based model. Second, a complex of docking-based alignment and maximum

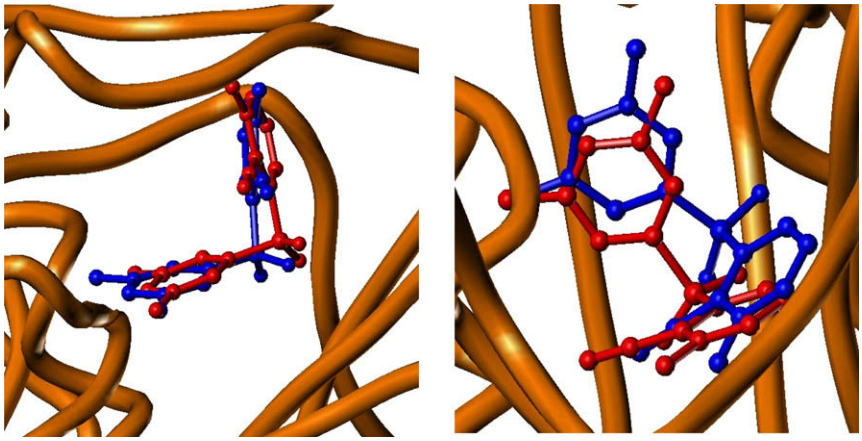


Figure 1. Comparison between X-ray crystal structure and structure from docking for compound **60** (red: docking; blue: crystal structure). For the sake of clarity, two different viewing angles are provided.

common substructures (MCSS) overlap approach was used[18] to build a receptor and ligand-based model. We have three families of molecules which might have different interaction modes with protein according to docking results. The second method was performed in two steps. In the first step, for each family, the molecules, which conformations were derived from docking, were aligned through MCSS overlap. Compounds **19**, 35 and **60** were chosen as templates for molecular alignment of the three types, respectively. In the second step, all the molecules of the three families were entered into the same data base.

2.4.2. Regions and fields of CoMFA and CoMSIA

In CoMFA, the steric fields were calculated using a Lennard–Jones potential, while the electrostatic fields were calculated using a Coulombic potential.[19] In CoMSIA, five different similarity fields (steric, electrostatic, hydrophobic, H-bond donor, and H-bond acceptor) were calculated.[13]

To calculate the CoMFA fields, a 3D cubic lattice with grid spacing of 2.0 Å in *X*, *Y* and *Z* directions was created automatically by SYBYL. The grid pattern extended 4.0 Å units in all directions beyond the dimensions of each molecule. The steric and electrostatic probe–ligand interaction energies were calculated by using a carbon probe atom with Van der Waals radius of 1.52 Å and a +1.0 charge with a distance-dependent dielectric function at each lattice point. The cut-off for energies was set to ±30 kcal/mol and the electrostatic contributions were ignored at lattice points with maximal steric interactions. CoMSIA models were also derived with the same lattice box and all five fields were calculated using a probe of charge +1, a radius of 1, hydrophobicity and hydrogen bonding properties of +1, and an attenuation factor of 0.3 for the Gaussian distance-dependent function.[13]

2.4.3. Statistical analysis

In order to derive CoMFA and CoMSIA models, CoMFA and CoMSIA descriptors were used as independent variables and the values as the dependent variables. PLS method with cross-validation (leave-one-out) was used in SYBYL 8.0 to determine the optimal numbers of components using cross-validated coefficient q^2 (r^2_{cv}).

After obtaining the optimal numbers of components, a PLS analysis was performed with no validation and column filtering 2.0 to generate the final model with the training set. The final correlation coefficient (r^2) is a measure of the quality of the model. A bootstrapping[20,21] procedure was applied to validate the final model.

The predictive capability of the 3D-QSAR models was determined from the predictive correlation r^2_{pred}. The predicted activities for the test set were obtained from the model produced by the training set.

3. Results and discussion

3.1. Docking

3.1.1. Comparison between X-ray crystal structure and structure from docking

To ensure the validity of docking calculations and of the conditions and parameters of docking, we performed a test. 739W94 (compound **60**), isolated from the complex crystal structure (PDB code: 1JLQ) was docked into the RT receptor, isolated from the complex. The docking procedure (100 runs) yielded only one cluster with conformations differing by less than 2 Å. The docked conformation of compound **60** which we choose for this study was the lowest energy one. It was close to the crystal structure since the RMSD between the two conformations was only 1.1 Å, which is quite satisfactory. Some conformations of higher energy had a slightly smaller RMSD relative to the crystal structure, but not less than 1.0 Å, except the highest energy one (0.96 Å). Relative to the crystal conformation, compound **60** docked thus correctly to the active site with only a slightly different conformation. Figure 1 shows the comparison between the X-ray crystal structure (in blue color) and cluster conformation (in red color).

3.1.2. Cluster selections

Clustering analysis of each molecule was done after docking calculations. For most molecules there are one, two or three clusters. These are sorted by energy (from the lowest to the highest) and their height is proportional to the number of conformations in each cluster. In order to choose the most reliable docked conformation for each compound, three strategies should be followed: (i) compare the population of each cluster and choose the one which is obviously the highest; (ii) if some clusters have similar populations, choose the one with the lowest energy; (iii) if the energies are close, choose the conformation similar to that of other molecules.

3.1.3. Binding mode of inhibitors

We studied three types of inhibitors: 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners. Two different interactional modes with the RT receptor were generated from docking: those of the sulfide and sulfone families were analogous because their substituents on ring A point in the same direc-

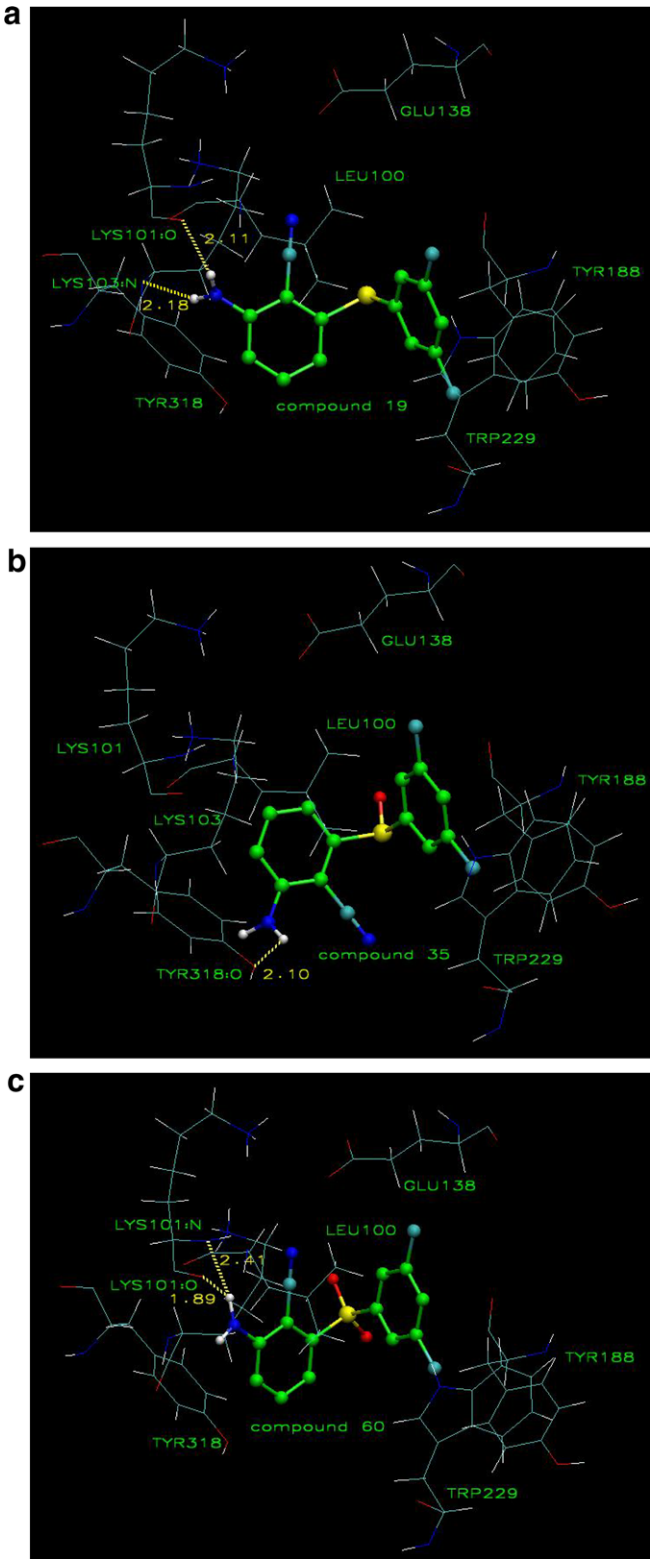


Figure 2. Interaction modes between RT receptors and compounds **19** (a), **35** (b), **60** (c).

tion. One compound of each family (compounds **19**, 35 and **60**) with the same substituents on ring B (3,5-$(CH_3)_2$) was selected to study the interaction with the protein.

Figure 2 shows the interactions and H-bonds between the RT receptor and the three families of molecules. Every compound has at least one strong H-bond between the amino group in the A ring and a functional group of a residue in the protein. Some of them have only one H-bond. A few of them have two H-bonds. As a representative compound for the sulfide family, compound **19** Figure 2a has two H-bonds: one between the carbonyl oxygen of Lys 101 and one amino hydrogen (>C═O···H–N) with a O···H distance of 2.11 Å and an angle (deviation from linearity) of 43°; the other one between the amino nitrogen of Lys 103 and the second amino hydrogen (–C–N···H–N) with a N···H distance of 2.18 Å and an angle of 26°. It should be noted here that compound **19** has a relatively high activity (pIC_{50} = 3.367).

The conformation of compound **35** in Figure 2b changes a lot when docked into RT. The amino group located in the opposite direction as compared to the amino group of compound **60** has an H-bond interaction with the hydroxyl oxygen of Tyr 318. The stability of this H-bond is probably less because of its larger angle (57°; distance: 2.10 Å). Compound **35** has also a relatively high activity and is involved in other interactions, for instance, steric effects.

For compound **60**, there exists a competition in forming one H-bond. One of the hydrogens in the amino group can connect to Lys 101 nitrogen by a –C–N···H–N bond with a N···H distance of 2.41 Å and an angle of 28° and can also connect to Lys 101 oxygen by a –C═O···H–N bond with a O···H distance of 1.89 Å and an angle of 60° (Fig. 2c). It should be noted that compound **60** has the highest bioactivity among all the compounds.

3.2. CoMFA and CoMSIA statistical results

We first used a receptor-based alignment procedure. All the molecules were aligned based on their docking conformations without a separation of training and test sets. For the 64 compounds, the CoMFA model yielded q^2 = 0.580, which is rather high and r^2 = 0.851, which is less satisfactory. All statistical data appear in Table 2. The CoMSIA model with all the five above-mentioned fields yielded q^2 of 0.488 and r^2 of 0.793. Although the correlations are not as satisfactory as we expected, the q^2 values much higher than 0.3[22] indicate that it is proper to use a receptor-based QSAR procedure as a first approach.

Next, receptor- and ligand-based alignment was performed without separating training and test sets. We obtained better models (q^2 = 0.674 and r^2 = 0.806 for CoMFA; q^2 = 0.695 and r^2 = 0.860 for CoMSIA). Either from the q^2 or from r^2 results, this second method of alignment is obviously better that the first one. This indicates not only that a complex alignment approach is powerful enough but also that the models are affected by the different alignment protocols. We thus used the complex alignment technique in what follows, with a separation of training and test sets.

An important step was to determine the optimum number of components with the training set. Figure 3 is the plot of q^2 and SEP (standard error of prediction) versus the number of components in CoMFA and CoMSIA. Plot (a) in Figure 3 shows that q^2 is increasing and SEP is decreasing when the number of components increases, up to 4 components. Then, q^2 remains constant while SEP slowly increases. Therefore, the optimal number of components for CoMFA is 4. With the model derived from the training set, we obtained the following statistical results by PLS: q^2 = 0.723, SEP = 0.532, r^2 = 0.868, and SEE = 0.368, where SEE means the standard error of estimation. q^2 > 0.6 means the model is fairly good.

The predicted values of activity are listed in Table 1. The distributed proportion of steric field is 47% and that of electrostatic field is 53%, emerging as a relative balance. Further, the robustness and statistical confidence of the models was verified by bootstrapping analysis. We performed 100 bootstrap samplings: a bootstrapped r^2_{boot} of 0.912 and a standard deviation (StdDev) of 0.294 were ob-

Table 2
Statistical data for CoMFA and CoMSIA models

Statistical parameters	Receptor-based models		Receptor- and ligand-based models				
	64-Compound model		64-Compound model		51-Compound model		
	CoMFA	CoMSIA	CoMFA	CoMSIA	CoMFA	CoMSIA (SEHDA)	CoMSIA (SHDA)
No. of components	5	5	4	5	4	9	8
q^2	0.580	0.488	0.674	0.695	0.723	0.741	0.760
SEP	0.657	0.724	0.575	0.561	0.532	0.545	0.519
r^2	0.851	0.793	0.806	0.860	0.868	0.969	0.959
SEE	0.391	0.461	0.443	0.380	0.368	0.190	0.214
r^2_{boot}					0.912	0.984	0.976
StdDev					0.294	0.008	0.158
r^2_{pred}					0.483	0.431	0.520
Contribution %:							
Steric					47	18.0	26.9
Electrostatic					53	23.7	
Hydrophobic						32.0	38.0
H-bond donor						11.2	13.7
H-bond acceptor						15.0	21.4

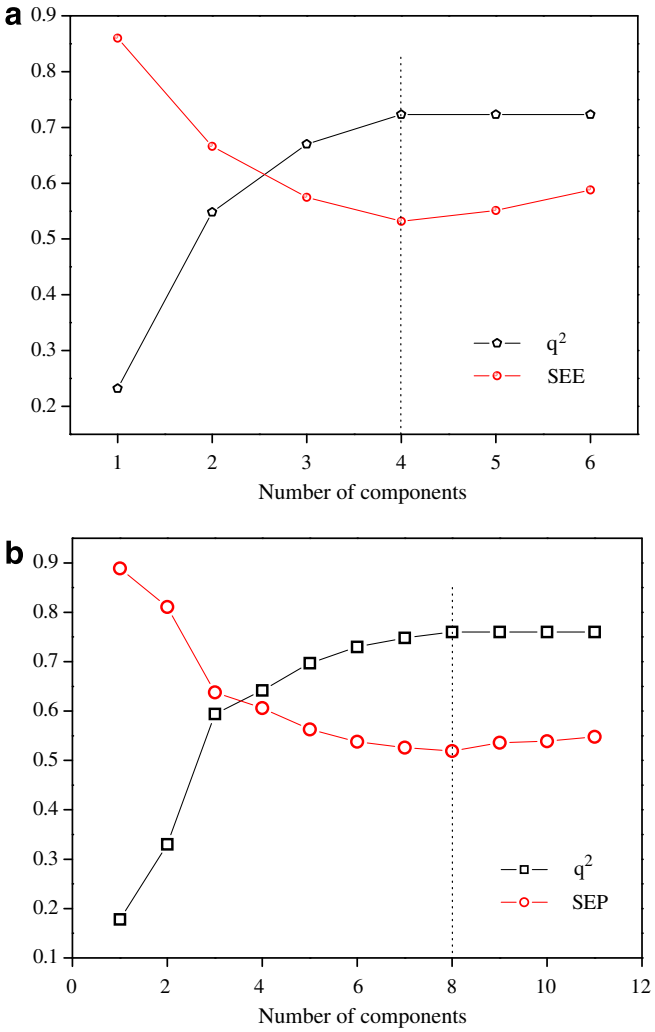


Figure 3. Plots of q^2 and SEE versus number of components in CoMFA (a) and CoMSIA (b) with 4 descriptors (SHDA).

tained, indicating a good internal consistency in the underlying data set.

In CoMSIA studies, all five fields were calculated, thus, five different indices were obtained at the same time: steric (S) and electrostatic (E), hydrogen bond donor (D), hydrogen bond acceptor (A) and hydrophobic (H) descriptors. Their relative contributions are presented in Table 2. With 9 components, we obtained a global descriptors CoMSIA model, with $q^2 = 0.741$; SEP = 0.545; $r^2 = 0.969$; SEE = 0.190.

It is true that CoMSIA model is superior to CoMFA model, however, the global descriptors model is not the best model in all probability. Some papers[23,24] have discussed whether the five different descriptor fields in CoMSIA are totally independent of each other. The dependencies of the individual fields usually decrease the signal-to-noise ratio in the data[24] and lower the statistical significance of the results. Hereby, an optimization of 31 possible combinations of different CoMSIA fields was evaluated from the values of the q^2 parameter (tested by leave-one-out cross-validation of PLS method) (Fig. 4). The higher the value of q^2, the better of the model. With the highest q^2 of 0.760, CoMSIA model with SHDA was finally chosen as the most predictive one. It should be noted that the electrostatic descriptor was eliminated from the combination. This indicates that the electrostatic fields are not totally independent from the others.

Figure 3b shows that 8 is the optimal number of components SEP on LOO, r^2 and SEE on no validation are, respectively 0.519, 0.959 and 0.214. The r^2_{boot} and StdDev of bootstrapping (100 runs) are 0.976 and 0.158, respectively, indicating that the model is stable and has high internal predictive ability. The distribution of the four descriptors is as follows: steric field is 26.9%; hydrophobic field is 38.0%; H-bond donor field is 13.7%; and H-bond acceptor field is 21.4%.

The fact that the correlation for CoMSIA models is much higher than that of CoMFA indicates that the steric and electrostatic effects are not enough to describe the accurate relationships between structures and activity. Hydrophobic effects and H-bond interactions already detected in our docking study are important too.

3.3. Validation of 3D-QSAR models

Although the bootstrapping analysis is a way to validate the accuracy of models, a predictive correlation coefficient r^2_{pred} was used to determine the predictive abilities of the CoMFA and CoMSIA models from the 13 compounds (test set) which were not included in the generation of the models. The predicted values for the test set are listed in Table 1 (compounds labeled with '*'). The final results appear in Table 2. r^2_{pred} is 0.483 for CoMFA (complex alignment with training set) and 0.520 for CoMSIA (complex alignment with training set on SHDA). Because of the high values of r^2_{pred}, either CoMFA models or CoMSIA models are thus shown

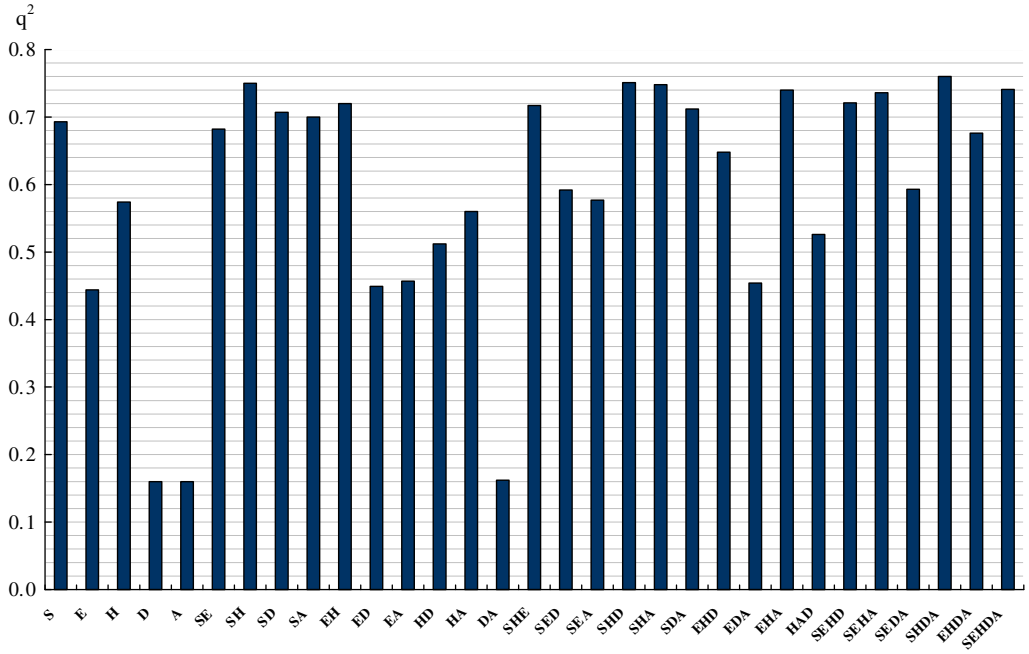


Figure 4. The histogram of 31 possibilities of the CoMSIA field combinations (S = steric, E = electrostatic, D = hydrogen bond donor, A = hydrogen bond acceptor, H = hydrophobic).

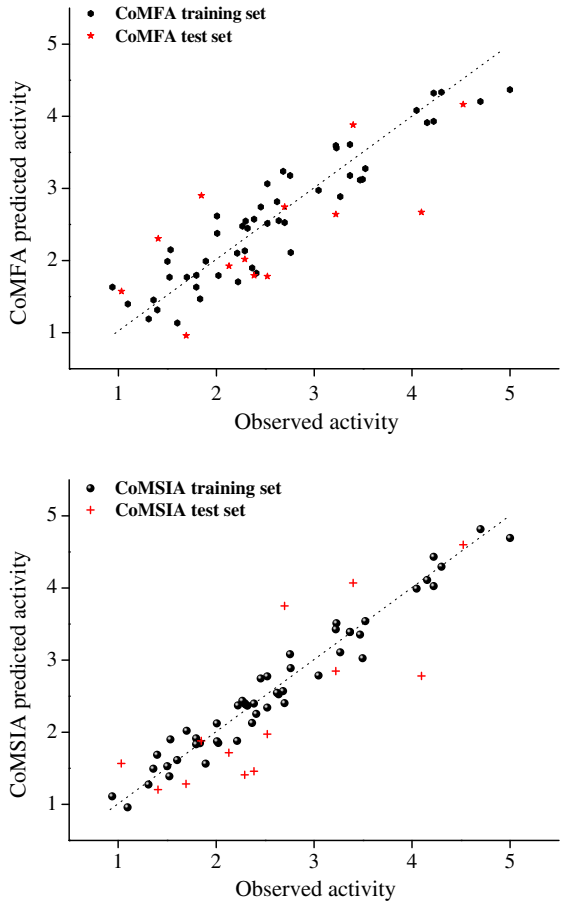


Figure 5. Plots of the predicted versus observed activity data of 3D-QSAR from both CoMFA and CoMSIA (SHDA) for the training set and test sets (the dotted line represents $y = x$).

to have a strong capability in predicting bioactivities. Correlation between predicted and observed activities is presented in Figure 5. The best 3D-QSAR model derived from CoMSIA with SHDA is quite satisfactory with respect to statistical significance and actual predictive ability.

3.4. CoMFA contour maps

3.4.1. Steric fields

The 3D contour maps of CoMFA steric fields are shown in Figure 6. They are represented by green colored contours (sterically favorable) and yellow colored contours (sterically unfavorable), with 80% and 20% level contributions, respectively. Three main contour areas (green one and two yellow ones) appear in Figure 6a.

In order to aid the visualization, the most potent compound (number 60) is overlaid on the map. The green contour located near a methyl group in the 3-position of B ring indicates that a bulkier group in this position is favorable to activity. Indeed, most molecules with a large group in 3-position have higher bioactivity compared with molecules without any substituent in that position such as compounds **11**, **14**, **19**, **21**, 22 and **23** in the sulfide family, compounds **30**, 35 and **37** in the sulfoxide family, and compounds **61**, **62**, 66 and **67** in the sulfone family.

A series of compounds (numbers **61–63** and **65–68)** were extracted as an example. They have the same methyl substituent in the 5-position and varying substituents in the 3-position. The activities of the seven molecules can be sorted from the highest activity to the lowest activity according to the substituents in the 3-position: $-Br > -Cl > -O(CH_2)_2CH_3 = -OCH_2CH_3 > -OCH_3 > -OH > -O(CH_2)_3CH_3$. Except the molecule with $-O(CH_2)_3CH_3$ in 3-position, the calculational results match the experimental results. The reason for this exception is that $-O(CH_2)_3CH_3$ is too large and does not easily enter the pocket of active site. In fact, compound **60** (with only $-CH_3$ groups in 3 and 5-positions) is also an exception with the highest bioactivity. This phenomenon is not haphazard,

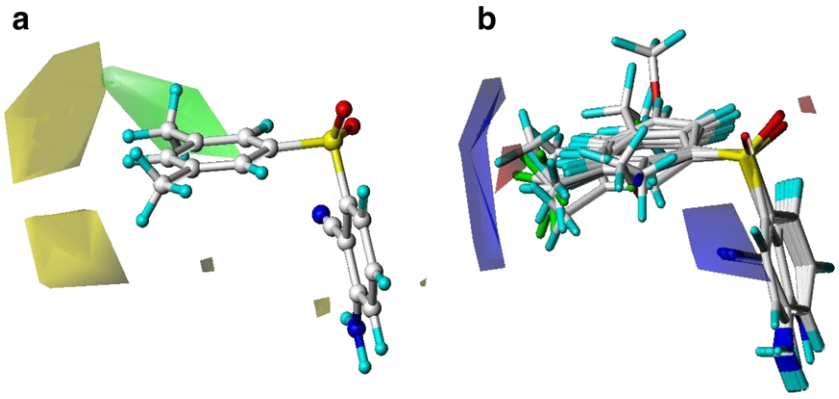


Figure 6. 3D contour maps (standard deviation × coefficient) for CoMFA. (a) Steric field. (Green: sterically favorable; yellow: sterically unfavorable); (b) electrostatic field. (Red: negative potential favorable; blue: negative potential unfavorable).

and shows that the steric effect is not the only factor to describe the interaction with protein.

The yellow contour surrounding the 4-position of B ring is a region where a bulkier group is not favorable to activity. This is confirmed by the low activity of compounds **6** with a $-CH_3$ group, **15** with a –CN group, **31** with –Br, 42 with $-OCH_3$ and **56** with –CN.

3.4.2. Electrostatic field

The electrostatic field contour is presented in Figure 6b; with a contribution of 53% it holds an important position in CoMFA. Compound **60** and the molecules of the sulfoxide family also represented for comparison.

The blue areas are the regions where negative potential is unfavorable to activity, while in the red areas a negative potential is favorable. In the 2-position and 3-position of B ring, the activity can be enhanced if an electronegative group is present. In the 4-position, the situation is quite the contrary since compounds with an electronegative group in this position are less active, which is confirmed by the low activity values of compounds **15**, **42**, **31**, 56 and **61**. As an example, compound **30** with a –Br in the 3-position has a more active inhibitory activity than compound **31** with a –Br in 4-position. A further example is given by compounds **40**–42 which have the same $-OCH_3$ group in 2-, 3- and 4-position, respectively. The value of pIC_{50} of compound **42** is obviously less that the other two ones. Similar results will be found in the series of compounds **49**–51 and in the series **54**–**56**.

3.5. CoMSIA contour maps

The steric and hydrophobic, H-bond donor and acceptor fields are represented as 3D contour plots in Figure 7. For each field, the favorable and disfavored contours represent 80% and 20% level contributions, respectively.

3.5.1. Steric field

From Figure 7a, where green regions are sterically favorable and yellow areas regions are sterically unfavorable, the distribution of steric field in CoMSIA is seen to be totally consistent with CoMFA results. The steric field is related to the structure of the RT protein pocket. Figure 7f is the surface of the RT protein. Compound **60** is located in the pocket of the active site. In this pocket, there is no free space to accept bulkier groups in the 4-position of the B ring. This ring is constrained between the two protuberant parts of the protein and steric hindrance arises consequently. Two hollow regions exist on both sides of the protuberant parts in the direction of the 3- and 5-position of the B ring. The lengths of substituents in the two positions can thus be increased to a certain extent. This is why the two areas are sterically favorable.

3.5.2. Hydrophobic field

Figure 7b is the hydrophobic contour map. Two large yellow areas in 3-position and 5-position of the B ring indicate that the two positions favor hydrophobic groups. This is supported by several examples. Compounds **19**, 35 and **60 are,** respectively almost the most active inhibitors in their families. They possess $-CH_3$ groups in 3- and 5-positions. Comparing **19** with **21** and **22**, or **60** with **62** and **63** it is clear that compounds with $-CH_3$ in 5-position are more active. An hydrophobic pocket indeed exists in these areas, composed of several amino acid residues (Leu100, Glu138, Val179, Trp229) of the protein. Comparing compounds **22** and **23**, or compounds **63** and **64** with $-OCH_3$ in the 3-position, it also appears that molecules with a $-CH_3$ group in the 3-position are more active.

There are also two large regions, one above the B ring and the other one below the ring (in white in Fig. 7b), where hydrophilic groups would be favorable.

3.5.3. H-bond donor field

The H-bond donor contour maps are presented in Figure 7c. Three pieces of contours in cyan appear in the map pointing out three possible H-bond donors. In order to aid visualization, compound **35** and compound **60** were overlaid into the fields. The largest region in the donor field corresponds to one of the amino hydrogen atoms offered by the sulfide family (represented by compound **60**) and the sulfone family. The second large region corresponds to the hydrogen atoms offered by the sulfoxide family (represented by compound **35**).

There is still a smaller region corresponding to the other amino hydrogen atom, located not far from the largest region. As we mentioned in the discussion of docking results, the sulfide family (represented by compound **19**, in Fig. 2a, has two hydrogen bonds interacting with the receptor. The CoMSIA results are thus consistent with the results from docking.

Figure 7e is the H-bond donor field with the RT protein superimposed. The favorable H-bond donor contours are exactly overlapping the residues of Lys 101 and Tyr 318, which are H-bond acceptors in the protein. Figure 2 offers more information about H-bond interactions between inhibitors and the protein.

3.5.4. H-bond acceptor field

The acceptor field contains information about where hydrogen bond donating groups should be located on the receptor. The –OH, $-CF_3$, –CN, $-NH_2$, –OH and >C═O groups can be H-bond donors

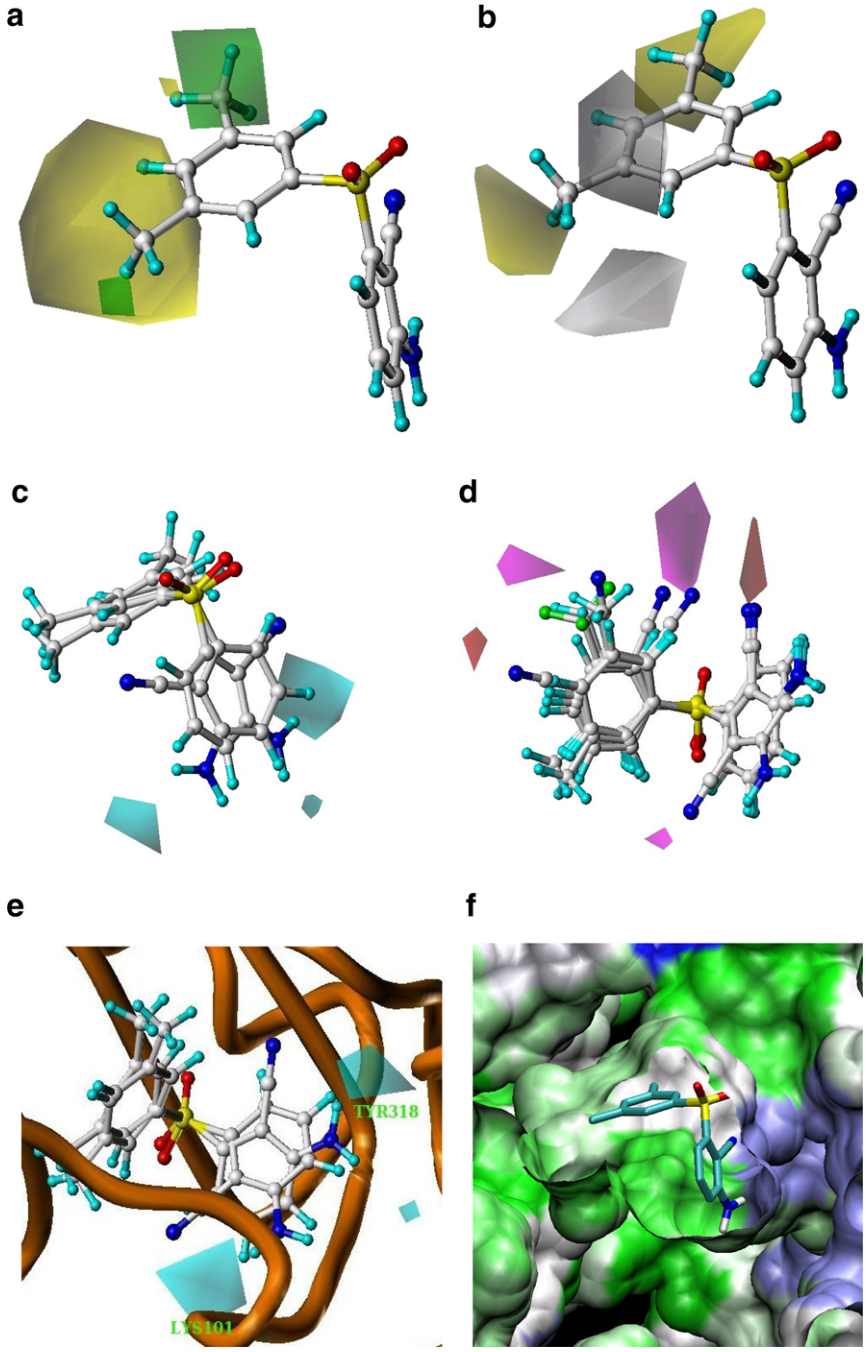


Figure 7. 3D CoMSIA contour maps (standard deviation × coefficient). (a) Steric field for compound **60** (green: sterically favorable; yellow: sterically unfavorable); (b) hydrophobic field for compound **60** (yellow: hydrophobic favorable; white: hydrophobic unfavorable); (c) H-bond donor field for compounds **35** and **60** (cyan: H-bond donor favorable; purple: H-bond donor unfavorable); (d) H-bond acceptor field for compounds **32**, **34**, **35**, **54**, **55**, **56** and **60** (magenta: H-bond acceptor favorable; red: H-bond acceptor unfavorable); (e) H-bond donor field with the RT protein and compounds **35** and **60** (magenta: H-bond acceptor favorable; red: H-bond acceptor unfavorable); (f) compound **60** and distribution of hydrophobic and hydrophilic residues at the surface of the protein according to the hydrophobic scale of Eisenberg[25,26] (green: hydrophobic part; blue: hydrophilic part; white: intermediate).

by donating a lone pair. H-bond acceptors contour field can be seen in Figure 7d. Compounds **32**, **34**, **35**, **54**–56 and **60** are overlaid. Surrounding the B ring, two large magenta-colored areas located in the 2,3-position, are favorable to H-bond acceptors. One small red-colored area located in the 4-position is not favorable to H-bond acceptors. This is confirmed by the activities of compound **54** with a 2-CN group, of compound **55** with a 3-CN group and of compound **56** with a 4-CN group which are 2.268, 2.620 and 1.097, respectively.

Around the A ring, there are a large red area and a small magenta area. For a better visual effect, the protein is overlapping the acceptor field. The cyano substituents on ring A of molecules from the sulfide family and the sulfone family are located near the magenta area and the cyano groups substituents on ring A of molecules from the sulfoxide family are located in the red area. From Table 1, it may be remarked that the mean activity of inhibitors in the sulfoxide family is lower than the mean activity of the sulfide and especially the sulfone family. This might be due to the fact

that the H bond in the sulfone family is stronger that in the sulfoxide family (see Section 3.1).

3.6. Prediction of activity values of compounds without experimental data

This work allows not only prediction of activities for molecules with no experimental activity available, but also rationalization of the predictions on the basis of our COMFA/COMSIA results. We give a few examples in what follows.

Compounds **9**, **13**, 18 and **20** have no known experimental activity values. Compound **13,** 9 has a -Cl substituent in the 4-position of the B ring. According to the CoMFA and CoMSIA steric field, its activity should be the lowest compared with compounds **7** (2.387) and **13**. According to the electrostatic field in CoMFA, its activity should be the higher than that of compound **15** (1.359) with a more electronegative –CN group in the same position. The predicted value 1.530 from CoMFA model is absolutely consistent with this.

We can infer that compound **13** is not very active (but more than compound **15** with 4-CN) considering the steric, electrostatic and H-bond acceptor fields, because it has a –CN in 2-position. This is consistent with the fact that its predicted activity values are 1.860 from CoMFA and 1.636 from CoMSIA.

For compound **20**, the activity is expected to be less than that of compounds **19** and **21** and more than that of compound **23** from hydrophobic field. The predicted value of 2.577 from CoMSIA supports this.

Compound **18** should have an activity close to that of compound **20** from the CoMSIA fields. However, from the steric field, the activity should in fact be lower than that of compound **20**. This ties in with the predicted values of 2.577 from CoMFA and 2.979 from CoMSIA.

3.7. Comparison with previous work

Roy and Leonard[9,10] built several 2D-QSAR models for this series of molecules with various parameters. With eight different molecular connectivity Hall and Kier E-state parameters, and for the whole data set (64 compounds with anti-HIV activity values),[9] the best model yielded $q^2 = 0.761$ and $R^2 = 0.799$. The best of 2D-QSAR models[10] built using different physicochemical parameters like hydrophobicity, electronic and steric descriptors, for the whole data set and seven descriptors, gave $q^2 = 0.758$ and $R^2 = 0.809$. Freitas performed a MIA-QSAR[11] study on the same compounds and obtained $q^2 = 0.712$ and $R^2 = 0.814$. The present work offers good statistical results and relates activity to steric, electrostatic, hydrophobic, donor and acceptor fields surrounding the structures. It is not only a predictive model of anti-HIV activity, but also a direct tool to help design new inhibitor candidates.

4. Conclusion

A study of 68 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners was carried out using docking and 3D-QSAR (CoMFA and CoMSIA) techniques. The conformations were generated from docking. A complex alignment method (receptor-based and ligand-based) was chosen to align the molecules after a comparison with a single docking alignment. Bootstrapping analysis showed that a good internal consistency exists in the data set.

Quite satisfactory cross validated and non-validated correlations were obtained and showed the superiority of the complex alignment. The models were also validated using a predictive correlation coefficient. The good agreement between observed and predicted activity values for the test set indicated the reliability of the QSAR models.

The bioactivity of the inhibitors can be rationalized to some extent owing to our results. The CoMFA steric and CoMSIA hydrophobic fields show that bulkier and hydrophobic groups are favorable to bioactivity in the 3- and 5-positions of the B benzene ring. On the contrary, they are unfavourable in the 4-position. The CoMSIA H-bond donor and acceptor fields suggest that the sulfide and sulfone inhibitors are more active than the sulfoxide ones due to H-bonding with protein residues.

In conclusion, our 3D-QSAR models based on docking, CoMFA and CoMSIA studies show a satisfactory predictive power in the three families of molecules and will be useful to evaluate novel potential drugs.

Acknowledgment

We are indebted to the National Natural Science Foundation of China for support through grants 90612016 and 60773108, and to the Ministry of Science and Technology of China for support through grant 2005DKA64001. We are also indebted to the China Research Council and the Conseil Régional d'Ile de France for financial support to Miss HU.

References and notes

1. Marzo Veronese, F.; Copeland, T. D.; DeVico, A. L.; Rahman, R.; Oroszlan, S.; Gallo, R. C.; Sarngadharan, M. G. *Science* **1986**, *231*, 1289.
2. Merluzzi, V. J.; Hargrave, K. D.; Labadia, M.; Grozinger, K.; Skoog, M.; Wu, J. C.; Shih, C. K.; Eckner, K.; Hattox, S.; Adams, J., et al *Science* **1990**, *250*, 1411.
3. Kopp, E. B.; Miglietta, J. J.; Shrutkowski, A. G.; Shih, C. K.; Grob, P. M.; Skoog, M. T. *Nucleic Acids Res.* **1991**, *19*, 3035.
4. Spence, R. A.; Kati, W. M.; Anderson, K. S.; Johnson, K. A. *Science* **1995**, *267*, 988.
5. Grozinger, K.; Proudfoot, J.; Hargrave, K.. In *Drug Discovery and Development*; Chorghade, M. S., Ed.; Wiley-Interscience, 2006; Vol. 1, pp 353–363.
6. Hargrave, K. D.; Proudfoot, J. R.; Grozinger, K. G.; Cullen, E.; Kapadia, S. R.; Patel, U. R.; Fuchs, V. U.; Mauldin, S. C.; Vitous, J.; Behnke, M. L. *J. Med. Chem.* **1991**, *34*, 2231.
7. Proudfoot, J. R.; Hargrave, K. D.; Kapadia, S. R.; Patel, U. R.; Grozinger, K. G.; McNeil, D. W.; Cullen, E.; Cardozo, M.; Tong, L.; Kelly, T. A. *J. Med. Chem.* **1995**, *38*, 4830.
8. Chan, J. H.; Hong, J. S.; Hunter, R. N., III; Orr, G. F.; Cowan, J. R.; Sherman, D. B.; Sparks, S. M.; Reitter, B. E.; Andrews, C. W., III; Hazen, R. J.; Clair, M. St.; Boone, L. R.; Ferris, R. G.; Creech, K. L.; Roberts, G. B.; Short, S. A.; Weaver, K.; Ott, R. J.; Ren, J.; Stuart, A. H. D.; Stammers, D. K. *J. Med. Chem.* **2001**, *44*, 1866.
9. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
10. Leonard, J. T.; Roy, K. *QSAR Comb. Sci.* **2004**, *23*, 23.
11. Freitas, M. P. *Org. Biomol. Chem.* **2006**, *4*, 1154.
12. Kubinyi, H.. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons Ltd, 1998; Vol. 1, pp 448–460.
13. Klebe, G.; Abraham, U. *J. Comput. Aided Mol. Des.* **1999**, *13*, 1.
14. www.rcsb.org.
15. Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. *J. Comput. Chem.* **1989**, *10*, 982.
16. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
17. 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.
18. Raichurkar, A. V.; Kulkarni, V. M. *J. Med. Chem.* **2003**, *46*, 4419.
19. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
20. Wehrens, R.; van der Linden, W. E. *J. Chemom.* **1997**, *11*, 157.
21. Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. *QSAR Comb. Sci.* **2006**, *7*, 18.
22. Agarwal, A.; Pearson, J. P. P.; Taylor, E. W.; Li, H. B.; Dahlgren, T.; Herslof, M.; Yang, Y.; Lambert, G.; Nelson, D. L.; Regan, J. W.; Martin, A. R. *J. Med. Chem.* **1993**, *36*, 4006.
23. Böhm, M.; Stürzebecher, J.; Klebe, G. *J. Med. Chem.* **1999**, *42*, 458.
24. Norinder, U. *Perspect. Drug Discovery Des.* **1998**, *12–14*, 25.
25. Barbault, F.; Landon, C.; Guenneugues, M.; Meyer, J. P.; Schott, V.; Dimarcq, J. L.; Vovelle, F. *Biochemistry* **2003**, *42*, 14434.
26. Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. *J. Mol. Biol.* **1984**, *179*, 125.