Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

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

Quantitative structure–activity relationship studies on 1-aryl-tetrahydroisoquinoline analogs as active anti-HIV agents

Ke-xian Chen, Hai-ying Xie, Zu-guang Li *, Jian-rong Gao

College of Chemical Engineering and Materials Science, Zhejiang University of Technology, 18, Chaowang Road, Hangzhou, Zhejiang 310014, China

ARTICLE INFO

Article history:
Received 27 June 2008
Revised 31 July 2008
Accepted 15 September 2008
Available online 18 September 2008

Keywords:
1-Aryl-tetrahydroisoquinoline analogs
Anti-HIV activities
Quantitative structure–activity relationship

ABSTRACT

Predictive quantitative structure–activity relationship analysis was developed for a diverse series of recently synthesized 1-aryl-tetrahydroisoquinoline analogs with anti-HIV activities in this study. The conventional 2D-QSAR models were developed by genetic function approximation (GFA) and stepwise multiple linear regression (MLR) with acceptable explanation of 94.9% and 95.5% and good predicted power of 91.7% and 91.7%, respectively. The results of the 2D-QSAR models were further compared with 3D-QSAR model generated by molecular field analysis (MFA), investigating the substitutional requirements for the favorable receptor–drug interaction and quantitatively indicating the important regions of molecules for their activities. The results obtained by combining these methodologies give insights into the key features for designing more potent analogs against HIV.

The acquired immunodeficiency syndrome (AIDS)[1] has been spreading continuously since it was first reported in 1981, and becomes one of the most hazardous diseases.[2,3] The human immunodeficiency virus type 1 (HIV-1),[4] first isolated from a patient with generalized lymphadenopathy in 1983, has been found to be the pathogenic retrovirus and causative agent of AIDS epidemic.[5,6] The number of HIV infected subjects keeps alarmingly on the rise.[7] Considerable attention has been paid to understand the viral life cycle and the functional nuances of nine genes encoded by HIV-1.[4,8] The protease (PR), reverse transcriptase (RT), and integrase[8,9] are regarded as the key enzymes in the duplication of HIV-1, thus structure-assisted design for these targets based on the knowledge of their three-dimensional structures may play a critical role in the discovery of novel anti-HIV drugs.

Intensive efforts have been madding worldwide to develop chemotherapeutic agents for the therapy of AIDS.[10] The major drugs fall into three families[11–13]: the nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), the non-nucleoside reverse transcriptase inhibitors (NNRTIs), and the protease inhibitors (PIs). Highly active antiretroviral therapy (HAART)[11,12,14] as a combination regimen, which includes two or more classes of anti-HIV drugs, can reduce HIV RNA to undetectable level, rescue $CD4^+$ cell counts and prolong survival, but cannot extinguish the infection[13,15] for the high cost, toxicity, complicated dosing schedules and considerable side effects. No thorough and successful chemotherapy has been developed so far,[16,17] because the human immunodeficiency virus (HIV) infection has remained an intractable problem and complete eradication of this virus is unrealized at present.[18,19] Therefore, it is essential and also urgent to develop new anti-HIV agents effective against the emerging drug-resistant viral strains and targeted at alternative steps of viral replication cycle.[11,13]

Recently, newly synthesized 1-aryl-tetrahydroisoquinolines[20] have been tested having the ability to protect C8166 cells against the cytopathogenicity of HIV. The previous studies also showed various biological activities including antibacterial and anticonvulsant activities for these compounds.[20] To gain further insights into the structure–activity relationships of these derivatives and understand the mechanism of their substitutional specificity influencing anti-HIV activities, the present group of authors thus developed some statistically significant QSAR models, which would aid in search for the novel anti-HIV analogs prior to synthesis.

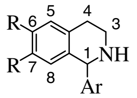
The anti-HIV activity data (EC_{50}) of 36 compounds collected from literature[20] were taken for this study. The biological activities were converted into the corresponding pEC_{50} values to get the linear relationship in equation using the following formula: $pEC_{50} = -\log EC_{50}$, where EC_{50} represents effective concentration required to protect C8166 cells against the cytopathogenicity of HIV by 50%.[20] Molecules were rationally divided into the training set (Table 1) and test set (Table 2) on the basis of suggestions by Oprea et al.,[21] which are (i) for the test set, the biological activity values should span several times but should not exceed activity values in training set by more than 10%; (ii) the test set should represent a balanced number of both active and inactive compounds for uniform sampling of the data.

All computational experiments were performed using QSAR+ module of Cerius^2 (version 4.10) running on Silicon Graphics O2 R5000 workstation.[22] The molecular geometric structures were constructed using a 3D-sketcher in the Cerius^2 Builder option.

* Corresponding author. Tel./fax: +86 571 88320306.
E-mail address: lzg@zjut.edu.cn (Z. Li).

0960-894X/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2008.09.056

Table 1
Molecular structures and anti-HIV activities of training set molecules used for QSAR study



Compound	R	Ar	EC_{50} (μM)	pEC_{50} [a](M)	Model-1		Model-2		Model-3	
					Pred. [b]	Dev.[c]	Pred. [b]	Dev. [c]	Pred. [b]	Dev. [c]
01	–OMe	Phenyl	119.6	3.922	3.960	−0.038	3.975	−0.053	4.054	−0.062
02	–OH	Phenyl	67.3	4.172	4.207	−0.035	4.150	0.022	4.184	−0.012
03	–OMe	2-Methylphenyl	42.5	4.372	4.258	0.114	4.273	0.099	4.275	0.097
04	–OH	2-Methylphenyl	61.9	4.210	4.258	−0.048	4.256	−0.046	4.382	−0.172
05	–OMe	4-Methylphenyl	67.3	4.172	4.155	0.017	4.216	−0.044	4.181	−0.009
08	–OH	2-Methoxyphenyl	42.2	4.375	4.420	−0.045	4.408	−0.033	4.375	0.000
09	–OMe	4-Methoxyphenyl	127.2	3.896	3.941	−0.045	3.973	−0.077	4.045	−0.149
10	–OH	4-Methoxyphenyl	34.1	4.467	4.399	0.068	4.408	0.059	4.253	0.214
11	–OMe	2,4-Dimethoxyphenyl	43.8	4.359	4.388	−0.029	4.363	−0.004	4.267	0.092
12	–OH	2,4-Dimethoxyphenyl	46.5	4.333	4.345	−0.012	4.311	0.022	4.374	−0.041
13	–OMe	4-Trifluoromethylphenyl	27.8	4.556	4.513	0.043	4.525	0.031	4.560	−0.004
14	–OH	4-Trifluoromethylphenyl	27.6	4.559	4.564	−0.005	4.537	0.022	4.595	−0.036
16	–OH	4-Bromophenyl	36.2	4.441	4.495	−0.054	4.536	−0.095	4.413	0.028
17	–OMe	3-Bromophenyl	43.5	4.362	4.513	−0.151	4.502	−0.140	4.413	−0.051
19	–OMe	2-Bromophenyl	31.3	4.504	4.453	0.051	4.458	0.046	4.431	0.073
20	–OH	2-Bromophenyl	25.7	4.625	4.458	0.167	4.506	0.119	4.521	0.104
21	–OMe	2-Fluorophenyl	69.0	4.161	4.170	−0.009	4.160	0.001	4.116	0.045
22	–OH	2-Fluorophenyl	58.4	4.234	4.175	0.059	4.255	−0.021	4.257	−0.023
23	–OMe	3-Chlorophenyl	31.8	4.498	4.404	0.094	4.350	0.148	4.419	0.079
25	–OMe	2-Chlorophenyl	58.4	4.234	4.342	−0.108	4.306	−0.072	4.260	−0.026
26	–OH	2-Chlorophenyl	51.8	4.286	4.343	−0.056	4.238	0.048	4.369	−0.083
27	–OMe	4-Nitrilephenyl	60.1	4.221	4.127	0.094	4.214	0.007	4.183	0.038
28	–OH	4-Nitrilephenyl	53.7	4.270	4.367	−0.097	4.386	−0.116	4.302	−0.032
29	–OMe	2-Furyl	152.4	3.817	3.858	−0.041	3.795	0.022	3.957	−0.140
30	–OH	2-Furyl	61.7	4.210	4.112	0.098	4.136	0.074	4.088	0.122
31	–OMe	Pyridin-2-yl	171.2	3.766	3.796	−0.030	3.782	−0.016	3.760	0.006
35	–OMe	2-Naphthyl	6.9	5.161	5.109	0.052	5.100	0.061	5.184	−0.023
36	–OH	2-Naphthyl	5.3	5.276	5.330	−0.054	5.339	−0.063	5.308	−0.032

[a] $pEC_{50} = -\log EC_{50}$. EC_{50} represents effective concentration required to protect C8166 cells against the cytopathogenicity of HIV by 50%.
[b] Pred. is the pEC_{50} predicted by the corresponding best model.
[c] Dev. is calculated by the formula: Dev. = Actual pEC_{50} − Predicted pEC_{50}.

Table 2
Structures and anti-HIV activities of test set molecules used for QSAR study

Compound	R	Ar	EC_{50} (μM)	pEC_{50} [a] (M)	Model-1		Model-2		Model-3	
					Pred. [b]	Dev. [c]	Pred. [b]	Dev. [c]	Pred. [b]	Dev. [c]
06	–OH	4-Methylphenyl	8.2	5.086	4.279	0.807	4.286	0.800	4.301	0.785
07	–OMe	2-Methoxyphenyl	45.9	4.338	3.901	0.437	3.931	0.407	4.237	0.101
15	–OMe	4-Bromophenyl	16.9	4.772	4.377	0.395	4.417	0.355	4.297	0.475
18	–OH	3-Bromophenyl	53.5	4.272	4.649	−0.377	4.690	−0.418	4.579	−0.518
24	–OH	3-Chlorophenyl	4.6	5.337	4.535	0.802	4.393	0.944	4.543	0.794
32	–OH	Pyridin-2-yl	48.4	4.315	4.036	0.279	3.948	0.367	3.880	0.435
33	–OMe	1-Naphthyl	22.6	4.646	5.101	−0.455	5.083	−0.437	5.333	−0.687
34	–OH	1-Naphthyl	33.2	4.479	5.316	−0.837	5.292	−0.813	5.421	−0.942

[a] $pEC_{50} = -\log EC_{50}$. EC_{50} represents effective concentration required to protect C8166 cells against the cytopathogenicity of HIV by 50%.
[b] Pred. is the pEC_{50} predicted by the corresponding best model.
[c] Dev. is calculated by the formula: Dev. = Actual pEC_{50} − Predicted pEC_{50}.

Then, an energy minimization procedure named UFF-VALBOND1.1[23] was employed to generate the lower energy conformation for each molecule. Partial atomic charges were assigned using the Gasteiger method.[24] All the structures were subsequently energy minimized until a root mean square derivation 0.001 kcal/mol was achieved and used in this study.

2D-QSAR analysis. Initially, 2D-QSAR analysis was performed by genetic function approximation (GFA). Different types of physicochemical descriptors for each molecule were generated in the study table using default setting. Before generating models, the inter-correlation of 147 descriptors with nonzero values was taken into account and highly correlated descriptors were removed.[25] Description of remaining descriptors used for generating 2D-QSAR models is given in Table 3.

Genetic function approximation (GFA),[26–28] genetically involved in the combination of Fried machs multivariate adaptive regression splines (MARS) and Holland's genetic algorithm (GA), is a useful statistical analysis tool to correlate biological activity or property with characteristic parameters of molecules, and also greatly improves the ease of successful model interpretation. The length of equation was initially fixed to five terms including a constant, the population size was established as 100, the equation term was set to linear polynomial and the mutation probability was specified as 50%. After some preliminary runs for observations, GFA crossover of 5000 and

Table 3
Descriptors used for building 2D-QSAR models (model-1 and model-2)

Type	Descriptors
E-state-indices	Electrotopological-state indices
Spacial	Jurs descriptors, radius of gyration, principal moment of inertia, shadow indices molecular surface area, density, molecular volume
Electronic	Sum of atomic polarizabilities, dipole moment, energy of highest occupied orbital (HOMO), energy of lowest unoccupied orbital (LUMO), superdelocalizability
Thermodynamic	Ghose and Crippen molar refractivity, heat of formation, log of the partition coefficient, log of the partition coefficient atom type value, desolvation free energy for water, desolvation free energy for octanol
Structural	Number of chiral centers, number of rotatable bonds, number of hydrogen-bond donors, number of hydrogen-bond acceptors, molecular weight
Conformation	The energy of the currently selected conformation
Information_content	Multigraph information content indices, information of atomic composition index

smoothing parameter "*d*" value of 2.0 were set to give reasonable convergence. Other default settings were maintained. Cross-validated r^2 (r^2_{CV})[29] was calculated using cross-validated test option in the statistical tool in Cerius^2 software.

A brute force approach[30] was first employed to investigate the number of descriptors necessary and adequate for the QSAR equation. As shown in Table 4, adding the number of descriptors in the equation does increase the r^2_{CV} value of the best model but r^2_{CV} and r^2 increase a little when the number of descriptors in equations ranges from 5 to 6. Based upon our experiments, the values of LOF (Friedman's Lack of Fit)[30] and *F*-test begin to decrease after the number of descriptors is set to 5 including a constant. The $r^2_{\text{-pred}}$ of the best model with four descriptors is 0.197, which is unsatisfactory for prediction.[25] The number of descriptors in equations thus was restricted to 5 for the final model. The selection of the best model[29] was based on the values of r^2 (square of the correlation coefficient for the training set molecules); LOF; *F*-test; r^2_{CV} (cross-validated r^2); r^2_{BS} (bootstrap correlation coefficient)[25]; PRESS (predicted sum of deviation squares). The statistically significant QSAR model is shown below:

Model-1

$$\begin{aligned}\text{pEC}_{50} = {} & 0.158002 + 0.302009A\log P98 + 0.225407\text{S\_aaaC} \\ & - 0.040666\text{Fh2o} + 0.871537\text{IC} - 0.032735\text{Jurs-PPSA-3}\end{aligned}$$

$$N = 28, \text{LOF} = 0.025, r^2 = 0.949, r^2_{\text{-adj}} = 0.938,$$

$$F\text{-test} = 82.555, \text{LSE} = 0.005,$$

$$r = 0.974, r^2_{CV} = 0.917, r^2_{\text{-BS}} = 0.949 \pm 0.001, \text{PRESS} = 0.244;$$

$$N' = 8, r^2_{\text{-pred}} = 0.961, \text{LSE} = 0.005, F\text{-test} = 9.770$$

where *N* is the number of compounds in training set; *N*′ is the number of compounds in test set; r^2 is an indicators of the model data fit; r^2_{CV} is an indication of the predictive capability of the model.[31,32]A high bootstrap r^2 with a low standard deviation indicates the robustness of the model.[25] $r^2_{\text{-pred}}$ is the predicted correlation coefficient which indicates that the model can be predicted well the activity of molecules in test set. The inter-correlation of the descriptors appeared in the above best model was taken into account and the descriptors were found to be reasonably orthogonal (Table 5). Descriptor values appeared in the above model of training set and test set molecules are shown in Table 6.

To determine reliability and significance of these generated models, the leave-one-out (LOO) test and randomization tests were employed. From the cross-validation test, r^2_{CV} of 0.917 indicated that the results obtained for the best QSAR equation was not by chance correlation.[29] The randomization tests[33] were performed at 90% (9 trials), 95% (19 trials), 98% (49 trials), and 99% (99 trials) confidence levels and carried out by repeatedly permuting the dependent variable set. The results of randomization tests (Table 7) showed that none of the permuted data sets produced the random *r* comparable to nonrandom *r* of 0.974, suggesting that the value obtained for the original GFA model was significant.[29,33] The robust and highly predictive ability of the models was reflected insufficiently only by the cross-validation test, thus the external predictive power of the model was evaluated with the test set molecules. The predictive power of the model is calculated by $r^2_{\text{-pred}} = (\text{SD-PRESS})/\text{SD}$,[34] where SD is the sum of squared deviations between the pEC_{50} of each molecule and the mean pEC_{50} of the molecules in the training set and PRESS is the sum of squared deviations between the predicted and calculated pEC_{50} values for

Table 4
Statistical evaluation of 2D-QSAR models with varying number of descriptors by genetic function approximation (GFA)

Descriptor	Equation	LOF	r^2	$r^2_{\text{-adj}}$	*F*-test	LSE	*r*	$r^2_{\text{-BS}}$	r^2_{CV}
1	$\text{pEC}_{50} = 4.27021 + 0.381656\text{S\_aaaC}$	0.058	0.565	0.548	33.705	0.046	0.751	0.562	0.517
2	$\text{pEC}_{50} = 3.72795 + 0.345058\text{S\_aaaC} + 0.007726\text{Jurs-WNSA-1}$	0.038	0.779	0.762	44.131	0.023	0.883	0.778	0.738
3	$\text{pEC}_{50} = -2.6753 + 0.42542A\log P98 - 0.042445\text{Fh2o} + 1.32558\text{IC}$	0.031	0.864	0.847	50.666	0.014	0.929	0.864	0.799
4	$\text{pEC}_{50} = -0.645278 + 0.364041A\log P98 + 0.183415\text{S\_aaaC} - 0.034992\text{Fh2o} + 0.844282\text{IC}$	0.019	0.942	0.932	93.227	0.006	0.971	0.942	0.913
5	$\text{pEC}_{50} = 0.158002 + 0.302009A\log P98 + 0.225407\text{S\_aaaC} - 0.040666\text{Fh2o} + 0.871537\text{IC} - 0.032735\text{Jurs-PPSA-3}$	0.025	0.949	0.938	82.555	0.005	0.974	0.949	0.917
6	$\text{pEC}_{50} = 0.294587 + 0.725444\text{IC} + 0.330201A\log P98 + 0.208728\text{S\_aaaC} - 0.026158\text{Fh2o} - 0.041854\text{Sr} - 0.000628\text{Jurs-DPSA-1}$	0.035	0.958	0.946	79.739	0.004	0.979	0.958	0.920

Table 5
Correlation matrix of the descriptors appeared in model-1 and model-2

	pEC_{50}	$A\log P98$	S_aaaC	Fh2o	IC	Jurs-PPSA-3	$A\log P$	Foct	Density	Sr
pEC_{50}	1									
$A\log P98$	0.533	1								
S_aaaC	0.751	0.263	1							
Fh2o	−0.233	0.456	−0.047	1						
IC	0.501	0.231	0.526	0.323	1					
Jurs-PPSA-3	0.127	−0.635	0.315	−0.735	−0.013	1				
$A\log P$	0.679	0.924	0.325	0.136	0.140	−0.443	1			
Foct	−0.319	0.352	−0.099	0.991	0.298	−0.710	0.032	1		
Density	0.258	0.186	−0.142	−0.293	−0.161	−0.271	0.302	−0.290	1	
Sr	−0.126	−0.022	−0.091	−0.190	−0.337	−0.025	0.075	−0.187	0.196	1

Table 6
Descriptors appeared in model-1 and model-2

Compound	Descriptors								
	S_aaaC	IC	*A*log*P*98	Fh2o	Jurs-PPSA-3	*A*log*P*	Foct	Density	Sr
01	0.000	3.446	3.02	−17.167	24.817	2.90	−21.790	1.04	1.9210
02	0.000	3.461	2.57	−28.216	27.233	2.83	−32.130	1.08	3.4983
03	0.000	3.594	3.51	−17.344	24.357	3.36	−22.390	1.03	0.4989
04	0.000	3.326	3.06	−28.394	26.795	3.30	−32.730	1.06	1.6477
05	0.000	3.499	3.51	−17.085	24.654	3.36	−22.390	1.03	0.1959
06	0.000	3.366	3.06	−28.134	26.895	3.30	−32.730	1.06	1.6705
07	0.000	3.300	3.01	−20.832	27.123	2.64	−25.550	1.05	0.9825
08	0.000	3.622	2.55	−31.881	29.406	2.58	−35.890	1.08	0.6909
09	0.000	3.356	3.01	−20.832	27.407	2.64	−25.550	1.05	0.9853
10	0.000	3.622	2.55	−31.881	30.056	2.58	−35.890	1.08	0.6976
11	0.000	3.573	3.99	−17.254	24.191	3.83	−22.900	1.02	1.0198
12	0.000	3.246	3.54	−28.304	26.391	3.77	−33.240	1.04	1.9339
13	0.000	3.605	3.96	−18.556	22.591	3.78	−24.720	1.16	0.1932
14	0.000	3.386	3.51	−29.605	24.763	3.72	−35.060	1.21	1.6647
15	0.000	3.499	3.77	−19.042	22.712	3.69	−23.760	1.25	1.6647
16	0.000	3.366	3.32	30.091	25.138	3.62	−34.100	1.32	1.6647
17	0.000	3.654	3.77	−19.042	22.688	3.69	−23.760	1.26	1.6647
18	0.000	3.537	3.32	−30.091	24.984	3.62	−34.100	1.32	1.6647
19	0.000	3.594	3.77	−19.355	23.314	3.69	−23.760	1.25	1.6647
20	0.000	3.326	3.32	−30.404	25.603	3.62	−34.100	1.32	1.6647
21	0.000	3.594	3.23	−17.299	24.400	3.04	−21.990	1.09	1.6780
22	0.000	3.326	2.78	−28.348	26.662	2.97	−32.330	1.13	0.3556
23	0.000	3.654	3.69	−17.155	22.881	3.41	−22.760	1.11	1.2747
24	0.000	3.537	3.24	−28.204	25.338	3.35	−33.100	1.16	3.7488
25	0.000	3.594	3.69	−17.401	23.510	3.41	−22.760	1.11	1.2745
26	0.000	3.326	3.24	−28.450	25.929	3.35	−33.100	1.15	3.7357
27	0.000	3.607	2.90	−18.994	25.147	2.93	−22.880	1.07	0.1958
28	0.000	3.662	2.45	−30.043	27.764	2.87	−33.220	1.10	1.8253
29	0.000	3.471	2.31	−20.654	26.366	1.90	−24.500	1.08	2.3500
30	0.000	3.499	1.86	−31.704	28.908	1.84	−34.840	1.13	0.3561
31	0.000	3.446	2.19	−19.651	25.260	1.86	−22.460	1.06	0.7163
32	0.000	3.461	1.74	−30.700	27.887	1.80	−32.800	1.10	2.5208
33	2.560	3.741	3.93	−19.316	26.457	3.90	−25.180	1.06	1.2806
34	2.417	3.754	3.48	−30.365	28.853	3.83	−35.520	1.09	1.2849
35	2.536	3.772	3.93	−19.330	26.913	3.90	−25.180	1.06	1.2775
36	3.424	3.789	3.48	−30.379	29.395	3.83	−35.520	1.09	0.8170

Table 7
Results of randomization tests for QSAR models generated by genetic function approximation (GFA)

Randomization test:				
Confidence level	90%	95%	98%	99%
Total trials	9	19	49	99
r from nonrandom	0.974	0.974	0.974	0.974
Random *r*'s < nonrandom	9	19	49	99
Random *r*'s > nonrandom	0	0	0	0
Mean of *r* from random trial	0.691	0.732	0.752	0.748
Standard deviation of random trials	0.297	0.138	0.050	0.105
Standard deviation from nonrandom *r* to mean	0.954	1.539	3.616	1.880

Table 8
Descriptor usage during generating 2D-QSAR models by genetic function approximation (GFA) with the number of descriptors in equations ranging from 3 to 6

No.	3		4		5		6	
	Descriptor	Usage	Descriptor	Usage	Descriptor	Usage	Descriptor	Usage
	S_aaaC	87	S_aaaC	92	IC	100	S_aaaC	100
	*A*log*P*	45	IC	86	S_aaaC	100	Sr	96
	*A*log*P*98	25	*A*log*P*	56	*A*log*P*98	87	IC	94
	Jurs-WNSA-1	19	Foct	36	Fh2o	69	*A*log*P*98	91
	IC	15	*A*log*P*98	33	Foct	35	Fh2o	72
	Jurs-FPSA-1	14	Fh2o	16	*A*log*P*	15	Foct	43

each molecule in the test set. The high $r^{2}_{\text{-pred}}$ value of 0.961 for the test set account for good predictive ability. The developed 2D-QSAR model-1 thus was robust and was found satisfactory for predicting the activities of the test set (Table 2).

Model-1 with four descriptors could explain 94.9% of the variance and predict 91.7% of the variance. The significance of descriptors in model-1 can be seen in (Table 8). The presence of three positive descriptors indicates that the large value of *A*log*P*98 (log of the partition coefficient, atom-type value), S_aaaC (atomic type of =C<in aromatic ring) and IC (multigraph information content indices) increase the anti-HIV activities. The terms of the atomic charge weighted surface areas (Jurs-PPSA-3) and the desolvation free energy for water (Fh2o) with negative coefficients are the important descriptors in favor for the bioactivity. Descriptor usage during generating 2D-QSAR models by GFA with the number of descriptors ranging from 3 to 6 is shown in Table 8.

The stepwise multiple linear regression (MLR) procedure was also employed for the model selection because of many descriptors used in this study. The multiple linear regression method with stepwise selection[22,31] calculates QSAR equations by adding one variable at a time and testing each addition for significance. Only variables found to be significant are used in the QSAR equation. This regression method is especially useful when the number of variables is large and the key descriptors are not known. The forward regression calculation mode was selected because backward regression calculation can lead to overfitting.[22] The parameters were specified to gain appropriate results. The maximum number of steps to be run in the calculation was set at 100, which can be specified to avoid hysteresis. *F* value of 4.000 was to evaluate the significance of a variable when a variable is added to or deleted from the equation. If the *F* value of a variable falls below a specified value, the variable is removed. The anti-HIV activity (pEC_{50}) was expressed with acceptable statistical significance in model-2:

Model-2

$$\begin{aligned}\text{pEC}_{50} = {} & -0.249275 + 0.262154A\log P - 0.022114\text{Foct} \\ & + 0.522538\text{Density} + 0.73178\text{IC} - 0.042674\text{Sr} \\ & + 0.202565\text{S\_aaaC}\end{aligned}$$

$$N = 28, r^2 = 0.955, F\text{-test} = 74.848, r^2_{\text{CV}} = 0.917,$$
$$r^2_{\text{-BS}} = 0.956 \pm 0.001, \text{PRESS} = 0.246;$$
$$N' = 8, r^2_{\text{-pred}} = 1.000, \text{LSE} = 0.000, F\text{-test} = 496.432$$

Model-2 contains much more descriptors than model-1, and also gains more significant results. According to model-2, it can explain and predict 95.5% and 91.7% of descriptors, respectively, which can be proved in predicting the test set (Table 2). The residuals of model-2 also are much smaller than that of model-1 (Table 1). Thus the anti-HIV activity (pEC_{50}) should be considered in terms of various descriptors in each molecule. It has been found to be positively significant correlated with the descriptors of S_aaaC and IC, which is equal to model-1. The positive slope of Ghose and Crippen log$P$ ($A$log$P$) and molecular density in model-2 represents that activity increases with addition of hydrophobic property in molecules. Desolvation free energy for octanol (Foct) shows a negative contribution to biological activity, suggesting that the anti-HIV effect can be enhanced by decreasing these values. Superdelocalizability (S_r)[22] is an index of reactivity in aromatic hydrocarbons (AH), proposed by Fukui as following formula:

$$S_r = 2\sum_{j=1}^{m}\left(\frac{c_{jr}^2}{e_j}\right)$$

In the above formula, S_r = superdelocalizability at position r; e_j = bonding energy coefficient in jth MO (eigenvalue); c = molecular orbital coefficient at position r in the HOMO; m = index of the HOMO. This index for all atomic positions of a molecule gives a metric of electrophilicity, which may be used to predict relative reactivity in a series of molecules.

3D-QSAR analysis. Molecular field analysis (MFA) was employed to derive 3D-QSAR model of 1-aryl-tetrahydroisoquinoline analogs in this study. MFA[34,35] is an effective method for evaluating the interaction energy between a probe molecule and a set of aligned target molecules at a series of points defined by a rectangular grid, especially for the analysis of data sets with available activity data but unknown 3D receptor site structure. The interaction energy values measured for each point of a 3D-grid were computed using atomic coordinates of binding molecules and can be used in subsequent QSAR study.

In this study, the core substructure search (CSS) method[36] coupled with root mean square (RMS) alignment method[22] was employed to rigidly align all of the structures in the analogous series based on a defined substructure of 10 atoms in active molecule 36 based on a defined substructure of 1,2,3,4-tetrahydroisoquinoline. So that common features are discernable from any other random arrangement of orientations and the sum of squares of the distances between all atoms to be superimposed are functionally minimized. The consensus RMS is 0.0128. Stereo-view of aligned molecules in training set and test set is shown in Figure 1.

The molecular field descriptors of the molecules were calculated using three different probes named CH_3, H^+, and HO^- once they had been aligned. CH_3, H^+, and HO^- were used to create the fields and simulate van der Waals, electrostatic and hydrogen bonding interactions, respectively.[37] The fields were computed at each point of a regularly spaced grid of 2 Å and the energy cutoff of ±30.00 kcal/mol was truncated. The total grid points generated were 343 and 10% of all new significant descriptors with highest variance of probes were automatically set as independent X variables for the subsequent 3D-QSAR modeling.

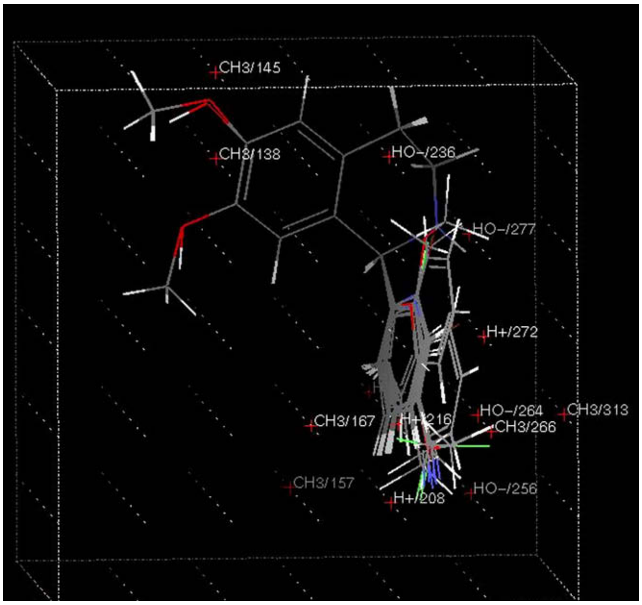


Figure 1. Stereo-view of aligned molecules in training set and test set within the 3D point grid of the model-3 is shown. CH_3, H^+, and HO^- represent steric interaction electrostatic interaction, and hydrogen bonding interaction, respectively.

Regression analysis was carried out using genetic partial least squares (G/PLS) method consisting of over 5000 generations with a population size of 100. The length of equation was fixed at 15 terms containing constant. The optimum number of components was set to 5. Cross-validation was performed with the leave-one-out (LOO) procedure by clicking the appropriate buttons in the validate control panel. The statistically significant 3D-QSAR model in terms of the most relevant descriptors is expressed as:

Model-3

$$\begin{aligned}\text{pEC}_{50} = {} & 3.87508 + 0.025199(\text{CH}_3/157) + 0.006084(\text{H}^+/208) \\ & + 0.004381(\text{H}^+/216) - 0.018174(\text{CH}_3/313) \\ & - 0.005032(\text{HO}^-/256) + 0.004254(\text{CH}_3/145) \\ & + 0.019511(\text{CH}_3/266) + 0.013327(\text{H}^+/272) \\ & + 0.024657(\text{CH}_3/167) + 0.014692(\text{HO}^-/277) \\ & - 0.006924(\text{HO}^-/264) + 0.002739(\text{HO}^-/236) \\ & - 0.007715(\text{CH}_3/138) + 0.005976(\text{H}^+/212)\end{aligned}$$

$$N = 28, r^2 = 0.932, \text{LSE} = 0.007, r = 0.965, r^2_{\text{CV}} = 0.618,$$
$$r^2_{\text{-BS}} = 0.744 \pm 0.441, \text{PRESS} = 1.105;$$
$$N' = 8, r^2_{\text{-pred}} = 1.000, \text{LSE} = 0.000$$

The descriptors H^+/a, H^+/b, and H^+/c[38] in model-3 are the energies between a proton probe and the molecule at the rectangular points a, b, and c, respectively. The descriptors CH_3/x and HO^-/x are the corresponding energies of interaction for the methyl probe and donor/acceptor probe, respectively.

Though the predicted value r^2_{CV} of model-3 is much worse than that obtained by GFA and MLR, the results predicted using model-3 was found satisfactory for residuals compared well with that of the above both methods. Model-3 consists of less methyl probes (CH_3) than proton and HO^- probes. The numbers associated with the variations specify their location in the 3D-grid around the aligned molecules is shown in Figure 1. The presence of CH_3/145 with positive coefficient indicates that the steric group is favorable at these positions, which can account for the higher activities with the group of OMe at C6 and C7 than that of OH (Tables 1 and 2) in compounds, but the negative slope of CH_3/138 with activity weaken this effect. As shown in Figure 1, the favorable presence of HO^-/

277 and HO^-/236 near to C4 and N atoms in core structure indicates the importance of the hydrogen bonding interactions at these positions. It can be seen that most of 3D-descriptors are distributing in the Ar regions of side chains distant from core structure, indicating the small transformation in these regions may change the anti-HIV activities of 1-aryl-tetrahydroisoquinoline analogs. Appearance of H^+/208, H^+/216, H^+/272, and H^+/212 with positive coefficients at Ar region can increasing the activity of molecules with electron withdrawing group.[34] Besides, three positive steric parameters (CH_3/157, CH_3/167, and CH_3/266) mean the groups with bulky substituents are favored to activity.[39]

In conclusion, diverse approaches of GFA, MLR, and MFA-G/PLS for the variable selection showed good predictive results of test set molecules and yielded some statistical information in the form of descriptors, which are significantly correlated with bioactivities. 2D-QSAR model-1 and model-2 show that the thermodynamic descriptors (*A*log*P*98, *A*log*P*, Fh2o, and Foct), the atomic type of =C<in aromatic ring (S_aaaC) and multigraph information content indices (IC) are strongly correlated with the activities. 3D-QSAR model-3 quantitatively explains the important regions of molecules influencing their activities. The results of various QSAR studies validate and supplement each other, and may provide a preliminary valuable guidance for improving the biological activity of the analogs and continuing search for potent anti-HIV agents prior to synthesis.

Acknowledgments

This work is financially supported by the National Natural Science Foundation of China (No. 30500339), the Natural Science Foundation Program of Zhejiang Province (No. Y407308) and the Sprout Talented Project Program of Zhejiang Province (No. 2008R40G2020019). The authors are also indebted to regional editor Masakatsu Shibasaki for valuable advice and reviewers for their important technical suggestion.

References and notes

1. Menéndez-Arias, L.; Tözsér, J. *Trends Pharmacol.* **2007**, *29*, 42.
2. Thakur, A.; Thakur, M.; Bharadwaj, A.; Thakur, S. *Eur. J. Med. Chem.* **2008**, *43*, 471.
3. Walgate, R.; Degett, J.; Eagles, N. *Biotechnology* **2008**, *25*, 29.
4. Andersen, J. L.; Le Rouzic, E.; Planelles, V. *Exp. Mol. Pathol.* **2008**. doi:10.1016/j.yexmp.2008.03.015.
5. Ravichandran, V.; Agrawal, R. K. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2197.
6. Vangelista, L.; Secchi, M.; Lusso, P. *Vaccine* **2008**, *26*, 3008.
7. Dessalew, N. *QSAR Comb. Sci.* **2008**, *27*, 901.
8. Yuan, H.; Parrill, A. L. *THEOCHEM* **2000**, *529*, 273.
9. Chen, X.; Tsiang, M.; Yu, F.; Hung, M.; Jones, G. S.; Zeynalzadegan, A.; Qi, X.; Jin, H.; Kim, C. U.; Swaminathan, S.; Chen, J. M. *J. Mol. Biol.* **2008**, *380*, 504.
10. Bonina, F.; Puglia, C.; Rimoli, M. G.; Avallone, L.; Abignente, E.; Boatto, G.; Nieddu, M.; Meli, R.; Amorena, M.; de Caprariis, P. *Eur. J. Pharm. Sci.* **2002**, *16*, 167.
11. Vicini, P.; Incerti, M.; La Cola, P.; Loddo, R. *Eur. J. Med. Chem.* **2008**. doi: 10.1016/j.ejmech.2008.05.030.
12. Jochmans, D. *Virus Res.* **2008**, *134*, 171.
13. Dubey, S.; Satyanarayana, Y. D.; Lavania, H. *Eur. J. Med. Chem.* **2007**, *42*, 1159.
14. Martinez-Picado, J.; Martínez, M. A. *Virus Res.* **2008**, *134*, 104.
15. Alterman, M.; Sjöbom, H.; Säfsten, P.; Markgren, P. O.; Danielson, U. H.; Hämäläinen, M.; Löfås, S.; Hultén, J.; Classon, B.; Samuelsson, B.; Hallberg, A. *Eur. J. Pharm. Sci.* **2001**, *13*, 203.
16. Leonard, J. T.; Roy, K. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4467.
17. Leonard, J. T.; Roy, K. *Bioorg. Med. Chem.* **2006**, *14*, 1039.
18. Dureja, H.; Gupta, S.; Madan, A. K. *J. Mol. Graphics Modell.* **2008**, *26*, 1020.
19. Musah, R. A. *Curr. Med. Chem.* **2004**, *4*, 1605.
20. Cheng, P.; Huang, N.; Jiang, Z.-Y.; Zhang, Q.; Zheng, Y.-T.; Chen, J.-J.; Zhang, X.-M.; Ma, Y.-B. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2475.
21. Oprea, T. J.; Waller, G. L.; Marshall, G. R. *J. Med. Chem.* **1994**, *37*, 2206.
22. Cerius^2, version 4.10, A. San Diego, Inc.: CA, USA, 2005.
23. Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. *J. Am. Chem. Soc.* **1992**, *114*, 10024.
24. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3291.
25. Kadam, R. U.; Roy, N. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5136.
26. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854.
27. Shi, L. M.; Yi, F.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189.
28. Leonard, J. T.; Roy, K. *Eur. J. Med. Chem.* **2008**, *43*, 81.
29. Ramar, S.; Bag, S.; Tawari, N. R.; Degani, M. S. *QSAR Comb. Sci.* **2007**, *26*, 608.
30. Nair, P. C.; Sobhia, M. E. *Eur. J. Med. Chem.* **2008**, *43*, 293.
31. Jung, M.; Tak, J.; Lee, Y.; Jung, Y. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1082.
32. Sivakumar, P. M.; Seenivasan, S. P.; Kumar, V.; Doble, M. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1695.
33. Deswal, S.; Roy, N. *Eur. J. Med. Chem.* **2006**, *41*, 1339.
34. Equbal, T.; Silakari, O.; Ravikumar, M. *Eur. J. Med. Chem.* **2008**, *43*, 204.
35. Hirashima, A.; Eiraku, T.; Kuwano, E.; Eto, M. *Internet Electron. J. Mol. Des.* **2003**, *2*, 511.
36. Vijayan, R. S. K.; Ghoshal, N. *J. Mol. Graphics Modell.* **2008**. doi:10.1016/j.jmgm.2008.05.003.
37. Yuan, H.; Parrill, A. *J. Mol. Graphics Modell.* **2005**, *23*, 317.
38. Hirashima, A.; Morimoto, M.; Kuwanoa, E.; Eto, M. *Bioorg. Med. Chem.* **2003**, *11*, 3753.
39. Shaikh, A. R.; Ismael, M.; Del Carpio, C. A.; Tsuboi, H.; Koyama, M.; Endou, A.; Kubo, M.; Broclawik, E.; Miyamoto, A. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5917.