Scientific paper

QSAR Analysis of 1,1-Dioxoisothiazole and Benzo[*b*]thiophene-1,1-dioxide Derivatives as Novel Inhibitors of Hepatitis C Virus NS5B Polymerase

Ke-Xian Chen, Hai-Ying Xie and Zu-Guang Li*

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

** Corresponding author: E-mail: lzg@zjut.edu.cn*
Phone/Fax: +86-571-88320306

Received: 26-08-2008

Abstract

Quantitative structure-activity relationship studies were carried out on some novel HCV NS5B polymerase inhibitors comprising 1,1-dioxoisothiazoles and benzo [b] thiophene-1,1-dioxides using genetic function algorithm (GFA) and molecular field analysis (MFA) techniques. The statistically significant 2D/3D-QSAR models ($r^2 > 0.975$) showed the indispensable structural requirements to improve the activity of this class. High r^2_{CV} values of 0.961 and 0.945 and r^2_{pred} values of 0.856 and 0.992 respectively for 2D/3D-QSAR models indicated the significant predictive ability of derived models. The validation of the models was done by full cross validation tests and external test set prediction. The results obtained can be exploited for modifications of the anti-HCV NS5B polymerase activity of this class of analogs.

Keywords: Hepatitis C virus NS5B polymerase inhibitors, 1,1-dioxoisothiazole, benzo[b]thiophene-1,1-dioxide, quantitative structure-activity relationship (QSAR), genetic function algorithm (GFA), molecular field analysis (MFA)

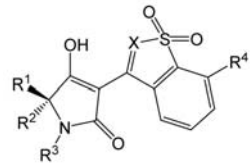
1. Introduction

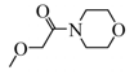
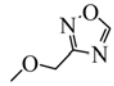
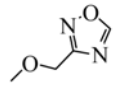
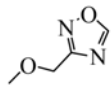
Hepatitis C virus (HCV), a member of the Flaviviridae family of viruses, has caused a global health crisis since it was first identified in 1989.[1–3] There is an estimated 3% or nearly 200 million of the world's population at risk of this disease,[4,5] which is about five times as prevalent as the acquired immunodeficiency syndrome (AIDS).[1] Many researchers have found that HCV is a major causative agent of liver cirrhosis, liver fibrosis, hepatocellular carcinoma and other forms of liver dysfunction.[2,6] Besides, HCV is also the leading indication for liver transplantation.[5] Unfortunately, the detailed mechanism of the action of HCV remains largely obscure at current level of knowledge.[1,7] It is also reported that HCV will be a serious global health threat for many years to come because this disease has the chronic nature of the infection, and high prevalence and the significant morbidity.[8] Hence, there is an increasing demand for new anti-HCV therapies to fulfill this unmet medical need.[2, 9]

The interferon (IFN) monotherapy has been the mainstream treatment for chronic HCV infection since the early 1990s.[9,10] The introduction of ribavirin as combination therapy with IFN has improved the treatment of HCV infection. However, current standard therapy has demonstrated limited success because of the development of drug resistance and severe adverse side effects.[9,11] Many infected peoples with genotype 1 virus have poor response to standard therapy. There is no effective vaccine to prevent the disease and no specific antiviral drugs directed against HCV infection.[12–14] No effective therapy for HCV associated chronic hepatitis C has been developed so far.[11] Thus there is an urgent need to further understand the mechanism of the action of hepatitis C virus and to discover improved therapeutic agents that can effectively combat chronic HCV infection.[13,15,16]

The urgent demand for novel anti-HCV agents has provided an impetus for understanding the structural requirements of NS5B polymerase inhibitors at molecular level.[17] NS5B polymerase, one of the virus-encoded pro-

Table 1. Structures and anti-HCV NS5B polymerase activity of molecules used for QSAR study.



Molecule	R^1	R^2	R^3	R^4	X (μM)	IC_{50} (M)[b]	$_pIC_{50}$	Model-1[c] predicted	Model-2[d] predicted
1	$CH(CH_3)_2$	H	CH_2Ph	H	N	1.4	5.854	5.785	5.966
2	$CH(CH_3)_2$	H	CH_2Ph	OEt	N	3.6	5.444	5.309	5.404
3	$CH(CH_3)_2$	H	CH_2Ph	OH	N	0.47	6.328	6.324	6.509
4	$CH(CH_3)_2$	H	CH_2Ph	OCH_2CONH_2	N	0.52	6.284	6.443	6.320
5	$CH(CH_3)_2$	H	CH_2Ph	$OCH_2CH_2CONH_2$	N	1.2	5.921	6.110	6.011
6[a]	$CH(CH_3)_2$	H	CH_2Ph	$OCH_2COC(CH_3)_3$	N	21	4.678	6.195	5.170
7	$CH(CH_3)_2$	H	CH_2Ph	OCH_2CN	N	1.6	5.796	5.760	5.680
8	$CH(CH_3)_2$	H	CH_2Ph		N	3.2	5.495	5.477	5.259
9	$CH(CH_3)_2$	H	CH_2Ph		N	0.35	6.456	6.363	6.463
10	$CH(CH_3)_2$	H	$CH_2CH_2C(CH_3)_3$	OEt	N	9.0	5.046	5.200	5.102
11[a]	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	OEt	N	1.5	5.824	5.136	5.298
12	$CH_2CH(CH_3)_2$	H	$CH_2CH_2C(CH_3)_3$	OEt	N	5.1	5.292	5.124	5.381
13	C_6H_{11}	H	$CH_2CH_2C(CH_3)_3$	OEt	N	4.0	5.398	5.384	5.335
14	2-Thiophene	H	$CH_2CH_2C(CH_3)_3$	OEt	N	2.9	5.538	5.589	5.505
15	$C(CH_3)_3$	H	CH_2(3-Cl-Ph)	OEt	N	1.6	5.796	5.844	6.039
16	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	OEt	N	0.76	6.119	6.313	5.935
17	$C(CH_3)_3$	H	$CH_2CH_2CH(CH_3)_2$	OEt	N	5.1	5.292	5.154	5.336
18[a]	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	H	N	0.42	6.377	5.824	6.286
19	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	OCH_2CONH_2	N	0.033	7.481	7.320	7.393
20	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	OCH_2CN	N	0.07	7.155	7.259	7.257
21[a]	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)		N	0.18	6.745	7.389	5.097
22[a]	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	$CH_2NHSO_2CH_3$	N	<0.01	8.000	7.622	7.397
23	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	$CH_2N(CH_3)SO_2CH_3$	N	0.023	7.638	7.438	7.666
24[a]	$C(CH_3)_3$	H	CH_2(3-Cl-4-F-Ph)	$NHSO_2CH_3$	N	0.094	7.027	8.362	6.657
25	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	OCH_2CONH_2	N	0.2	6.699	6.621	6.587
26	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$		N	0.65	6.187	6.339	6.142
27	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	$CH_2NHSO_2CH_3$	N	0.018	7.745	7.857	7.696
28	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	$CH_2CH_2SO_2CH_3$	N	0.045	7.347	7.324	7.286
29	$-CH_2CH_2-$	–	CH_2(3-Cl-4-F-Ph)	$CH_2NHSO_2CH_3$	N	0.13	6.886	6.735	6.763
30	$-CH_2CH_2-$	–	CH_2(3-Cl-4-F-Ph)	$CH_2N(CH_3)SO_2CH_3$	N	0.62	6.208	6.363	6.346
31	CH_2CH_3	CH_3	CH_2(3-Cl-4-F-Ph)	$CH_2N(CH_3)SO_2CH_3$	N	0.41	6.387	6.266	6.459
32	CH_3	CH_3	CH_2(3-Cl-4-F-Ph)	$CH_2N(CH_3)SO_2CH_3$	N	0.46	6.337	6.361	6.277
33	$CH(CH_3)_2$	CH_3	$CH_2CH_2C(CH_3)_3$	$CH_2N(CH_3)SO_2CH_3$	CH	13	4.886	4.979	4.960
34	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	$CH_2N(CH_3)SO_2CH_3$	CH	0.92	6.036	6.010	5.972
35[a]	$C(CH_3)_3$	H	$CH_2CH_2C(CH_3)_3$	$CH_2NHSO_2CH_3$	CH	0.24	6.620	6.222	4.130

[a] Molecules in the test set;
[b] IC_{50} is concentration of compounds required to achieve 50% inhibition against HCV NS5B polymerase, which is converted into corresponding negative logarithm (pIC_{50});
[c] Model-1 is the best 2D-QSAR model derived by GFA method;
[d] Model-2 is the best 3D-QSAR model derived by MFA-G/PLS method

teins in the HCV genome, is initially recognized as an RNA-dependent RNA polymerase (RdRp) due to its critical function in the replication cycle of the virus, which has been confirmed by extensive in vivo studies having essential roles for viral replication in cell cultures and in chimpanzees.[2,9] Therefore, NS5B polymerase has been one of the ideal targets for anti-HCV therapy.[11] Medicinal chemistry pursuits of this polymerase for discovery of anti-HCV drugs have led to the identification of many structurally diversified inhibitors.[2] Recently, Kim et al. have reported some novel HCV NS5B polymerase inhibitors comprising 1,1-dioxoisothiazoles and benzo[b]thiophene-1,1-dioxides.[18] Elucidation of the structural requirements for receptor binding of these inhibitors is important in the development of novel therapeutic and diagnostic agents. Computational chemistry including QSAR study was developed as an important contributor to rational drug design.[19] The aim of this study is to derive some statistically significant 2D/3D-QSAR models to indicate the indispensable structural requirements for improving the activity of this class of potent anti-HCV agents by combining genetic function algorithm (GFA) and molecular field analysis (MFA) methodologies. The results obtained may contribute to design novel anti-HCV agents and to further understanding of the mechanism of action of hepatitis C virus (HCV).

2. Experimental

2.1. Data Set

A data set of 35 novel HCV NS5B polymerase inhibitors comprising 1,1-dioxoisothiazoles and benzo[b]thiophene-1,1-dioxides were taken from the literature (Table 1).[18] The biological activity was expressed as IC_{50} values (IC_{50} values are concentration of compounds required to achieve 50% inhibition against HCV NS5B polymerase). The biological data were converted to -log molar concentration (pIC_{50}) to reduce the skewness of the data set.[20] Molecules were rationally divided into the training set (**28**) and test set (**7**) (Table 1) 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. The test set molecules captured structural features of the training set molecules, thus their activity could be well predicted.[22]

2. 2. Molecular Modeling

All the molecular modeling and statistical analysis were performed using Cerius^2 (version 4.10) running on Silicon Graphics O2 R5000 workstation.[23] The molecular geometric structures were constructed using a 3D-sketcher in the Cerius^2 Builder option. Partial atomic charges were assigned using the Gasteiger method.[24] Multiple conformations of each molecule were generated using the Boltzmann jump as a conformational search method. The upper limit of the number of conformations per molecule was 150. Each conformer was subsequently subjected to an energy minimization procedure of UFF-VALBOND1.1 to generate the lowest energy conformation for each structure.[25] The minimum energy difference of 0.001kcal/mol was set as a convergence criterion.

2. 3. 2D-QSAR Analysis

Different types of physicochemical descriptors for each molecule were generated in the study table using default setting within QSAR+ module of Cerius^2. Before generating models, 163 nonzero descriptors of E-state-indices, conformational, structural, thermodynamic, electronic and spatial were considered for their inter-correlation and the highly correlated descriptors, and the descriptors that are difficult to interpret were removed.[26] Descriptors used for generating 2D-QSAR models are listed and described in Table 2.

2D-QSAR analysis was performed using genetic function algorithm (GFA). GFA was developed by Rogers and Hopfinger,[27] which was genetically involved in the combination of Fried machs multivariate adaptive regession splines (MARS) and Holland's genetic algorithm (GA),[28,29] It is a useful statistical analysis tool to correlate

Table 2. Descriptors used for building 2D-QSAR models.

Type	Descriptors
E-state-indices	Electrotopological-state indices
Spatial	Jurs descriptors, radius of gyration, principal moment of inertia, 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
Conformational	The energy of the currently selected conformation

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 smoothing parameter "d" value of 1.0 were set to give reasonable convergence. Other default settings were maintained. Cross-validated r^2 (r^2_{CV}) was calculated using cross-validated test option in the statistical tool in Cerius[2].

2. 4. 3D-QSAR Analysis

Molecular field analysis (MFA) was employed to derive 3D-QSAR models for HCV NS5B polymerase inhibitors in this study. MFA attempts to postulate and represent the essential features of a receptor site from the aligned common features of the molecules that bind to it.[30–32] This approach can effectively evaluate the interaction energy between a probe and the set of aligned molecules at a series of points defined by a rectangular grid, especially for the analysis of data sets with available activity data but unknown receptor site structure.[31,33,34] The probe interaction energy on a rectangular grid were computed using atomic coordinates of binding molecules, which can be used for the subsequent 3D-QSAR study.

Having a proper alignment of the structures is critical for obtaining reliable 3D-QSAR models. It is also vital that all compounds are aligned in a pharmacological active orientation since the 3D-QSAR model assumes that each structure exhibits activity at the same binding site of the receptor. The method used for performing the alignment was the maximum common subgroup (MCSG) method.[23,31,33,35] This method looks at molecules as points and lines, and uses the techniques of graph theory to identify patterns. It finds the largest subset of atoms in the shape reference compound that is shared by all the structures in the study table and uses this subset for alignment. A rigid fit of atom pairings was performed to superimpose each structure so that it overlays the shape reference compound. A conformer of the most active molecule **22** was selected as the shape reference compound to which all the structures in the study compounds were aligned through pair–wise superposition (Figure 1).

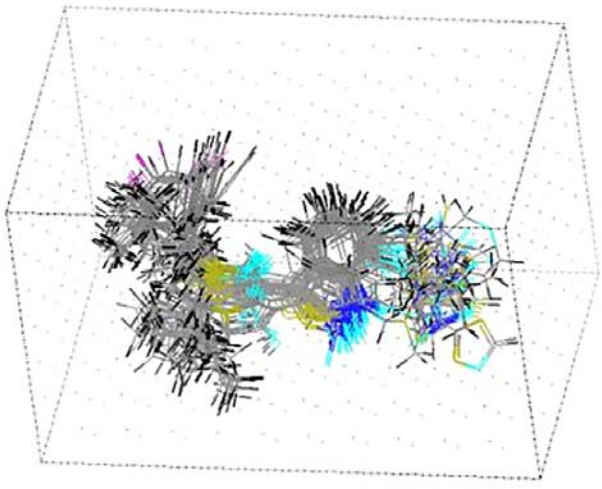


Figure 1. Stereo-view of aligned molecules in the training set and test set by the maximum common subgroup (MCSG) alignment method.

The MFA fields were calculated using proton probe (H^+) and methyl probe (CH_3) at each lattice interaction of a regularly spaced grid of 2.0 Å within defined three-dimensional region and an energy cutoff of –30 to +30 kcal/mol was truncated. CH_3 and H^+ represent the steric and electrostatic interaction, respectively. The total grid points generated were 768 and only ten percent of total columns of H^+ and CH_3 probes with highest variance were automatically selected as independent X variables, which were directly used as input for 3D-QSAR analysis. Regression analysis was carried out using the genetic partial least squares (G/PLS) method consisting of 10000 generations with a population size of 100.[30,31,33] The optimum number of component was set to 4 and the length of equation was fixed to 15 containing a constant. Cross-validation was performed with the leave-one-out (LOO) procedure.

3. Results and Discussion

3. 1. 2D-QSAR Model

Different sets of 2D-QSAR equations with several descriptors were generated using the genetic function algorithm (GFA) in Cerius[2]. A brute force approach was first employed to investigate the number of descriptors necessary and adequate in the QSAR equations.[34,36] As the square correlation coefficient (r^2) can be easily increased by number of terms in the equation, the cross-validated r^2 (r^2_{CV}) was selected as the limiting factor for number of descriptors in the equation. As shown in Table 3, the r^2_{CV} value increases till the number of descriptors in the equation reaches up to 5 and then starts decreasing as the number of descriptors increases further. Thus, the number of descriptors was restricted to 5 in the equation for the final model. The selection of the best model was based on the values of $\mathbf{r^2}$ (square of the correlation coefficient for the training set molecules), **LOF** (Friedman's lack of fit score), **F-Test**; **LSE** (least square error), $\mathbf{r^2_{CV}}$ (cross-validated r^2), $\mathbf{r^2_{BS}}$ (bootstrap correlation coefficient), and **PRESS** (predicted sum of deviation squares).[34] The statistically significant 2D-QSAR model with no outliners is shown below.

Table 3. Statistical evaluation of 2D-QSAR models with varying number of descriptors by genetic function algorithm (GFA).

Descriptor	Equation	LOF	r^2	r^2_{adj}	F-test	LSE	r	r^2_{BS}	r^2_{CV}
1	pIC_{50} = 2.23347 – 0.046664(Jurs-PNSA-3)	0.237	0.655	0.641	49.286	0.205	0.809	0.655	0.602
2	pIC_{50} = 0.007649 – 0.062216(Jurs-PNSA-3) + 1.10834(Atype_C_8)	0.092	0.886	0.877	97.202	0.068	0.941	0.886	0.862
3	pIC_{50} = 1.4776 – 1.15221(Atype_C_11) – 0.063625(Jurs-PNSA-3) – 0.228339(Atype_C_6)	0.056	0.942	0.934	129.042	0.035	0.970	0.942	0.919
4	pIC_{50} = 0.066656 – 0.230102(Atype_C_6) + 1.01908(Atype_C_8) + 0.174428(Sr) – 0.065147(Jurs-PNSA-3)	0.036	0.969	0.963	178.578	0.018	0.984	0.955	0.969
5	pIC_{50} = 1.26999 – 1.01674(Atype_C_11) – 0.240401(Atype_C_6) – 0.140126 (S_dssC)-0.06192(Jurs-PNSA-3) +0.152684(Sr)	0.034	0.976	0.971	178.944	0.014	0.988	0.976	0.961
6	pIC_{50} = 0.118409 – 0.062861(Jurs-PNSA-3) + 0.150551(Sr) – 0.175949(S_dssC) – 0.235841(Atype_C_6) + 0.318926 (S_sssCH) + 1.24354(Atype_C_8)	0.039	0.979	0.972	159.781	0.013	0.989	0.979	0.959

Model-1

pIC_{50} = 1.26999 – 1.01674(Atype_C_11) – 0.240401(Atype_C_6) – 0.140126(S_dssC) – 0.06192(Jurs-PNSA-3) + 0.152684(Sr)

N = 28, LOF = 0.034, r^2 = 0.976, r^2_{adj} = 0.971, F-Test = 178.944, LSE = 0.014, r = 0.988, r^2_{CV} = 0.961, r^2_{BS} = 0.976+/-0.000, PRESS = 0.640, N' = 7, r^2_{pred} = 0.856

N is the number of compounds in training set and LOF (Friedman's lack of fit score) is used to assess the goodness of each progeny equation using the following formula:[36,37] LOF = LSE / $\{1 – (c + dp) / m\}^2$, where LSE is least square error, c is the number of basis function in the model, d is smoothing parameter, p is the number of descriptors and m is the number of observations in the training set; r^2 is an indicator of the model data fit; r^2_{CV} is an indication of the predictive capability of the model;[38] N' is the number of compounds in test set; r^2_{-pred} is the predictive squared correlation coefficient of test set,[39] which is calculated by r^2_{-pred} = (SD-PRESS)/SD,[33,34] where SD is the sum of squared deviations between the pIC_{50} of each molecule and the mean pIC_{50} of the molecules in the training set and PRESS is the sum of squared deviations between the predicted and calculated pIC_{50} values for each molecule in the test set. The high r^2_{-pred} value of 0.856 for the test set accounted for the good predictive ability of model-1. The inter-correlation of the descriptors appeared in the above model-1 was taken into account and the descriptors were found to be reasonably orthogonal (Table 4). Summary of best QSAR equations with five descriptors are shown in Table 5.

The randomization tests and full cross-validation tests were employed to determine reliability and significance of these generated models. The randomization tests 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.[34,36,37] The results of randomization tests in Table 6 showed that none of the permuted data sets produced the random r comparable to nonrandom r of 0.988, suggesting that the value obtained for the original model-1 was significant. Cross-validation is a practical and reliable method for testing significance.[33] The full cross-validation tests encompass the entire algorithm including both the choice of descriptors and the optimization of regression coefficients.[40] The cross-validated r^2 (r^2_{CV}) was computed using the predicted values of the missing molecules by the models obtained from the remaining compounds in the data set. The results based on the rules of "leave-1-out", "leave-2-out", "leave-3-out", "leave-4-out", "leave-5-out", "leave-7-

Table 4. Correlation matrix of the descriptors appeared in 2D-QSAR model-1.

	pIC_{50}	Atype_C_11	Jurs-PNSA-3	Sr	S_dssC	Atype_C_6
pIC_{50}	1					
Atype_C_11	–0.024	1				
Jurs-PNSA-3	–0.809	–0.489	1			
Sr 0.039	–0.430	0.318	1			
S_dssC	–0.513	–0.157	0.482	–0.058	1	
Atype_C_6	–0.149	–0.042	–0.058	0.017	–0.142	1

Table 5. Summary of best QSAR equations generated by genetic function algorithm (GFA) method with five descriptors.

Descriptor	Equation	LOF	r^2	r^2-adj	F-test	LSE	r
1	pIC_{50} = 1.26999 – 1.01674(Atype_C_11)– 0.240401 (Atype_C_6) – 0.140126(S_dssC) – 0.06192 (Jurs-PNSA-3) + 0.152684(Sr)	0.034	0.976	0.971	178.944	0.014	0.988
2	pIC_{50} = 1.37196 – 0.057496(Jurs-PNSA-3) – 1.10783 (Atype_C_11) – 0.249135(Atype_C_6) + 0.160746(Sr) – 0.000367(Jurs-PNSA-2)	0.038	0.973	0.967	159.950	0.016	0.987
3	pIC_{50} = 1.271 – 0.234089(Atype_C_6) + 0.157638(Sr – 0.2603(Jurs-FNSA-2) – 1.11158 (Atype_C_11) – 0.058589(Jurs-PNSA-3)	0.039	0.973	0.967	158.491	0.016	0.986
4	pIC_{50} = 0.966113 – 0.064248(Jurs-PNSA-3) – 0.991227(Atype_C_11) – 0.229949 (Atype_C_6) + 0.204867(Atype_C_39) + 0.153011(Sr)	0.039	0.973	0.967	156.910	0.016	0.986
5	pIC_{50} = 1.31634 – 0.223683(Atype_C_6) – 0.062435 (Jurs-PNSA-3) – 1.09332(Atype_C_11) + 0.010869(S_sF) + 0.143171(Sr)	0.040	0.972	0.966	152.856	0.017	0.986
6	pIC_{50} = 1.31602 – 0.223698(Atype_C_6) + 0.143166(Sr) – 0.06244(Jurs-PNSA-3) – 1.09365(Atype_C_11) + 0.148092 (Atype_F_84)	0.040	0.972	0.966	152.767	0.017	0.986
7	pIC_{50} = 1.06779 – 0.066316(Jurs-PNSA-3) – 1.08581 (Atype_C_11) – 0.238002(Atype_C_6) + 0.171716(Sr) – 0.065183(Atype_C_4)	0.040	0.972	0.965	151.998	0.017	0.986
8	pIC_{50} = 1.10075 – 0.230753(Atype_C_6) + 0.144789 (Sr) – 0.064054(Jurs-PNSA-3) – 1.0514 (Atype_C_11) + 0.055958(Atype_C_26)	0.041	0.972	0.965	151.232	0.017	0.986
9	pIC_{50} = 2.68512 – 0.060797(Jurs-PNSA-3) – 1.01493 (Atype_C_11) – 0.217129(Atype_C_6) + 0.171065(Sr) – 1.717(Jurs-RASA)	0.041	0.972	0.965	151.199	0.017	0.986
10	pIC_{50} = 1.36288 – 0.062509(Jurs-PNSA-3) – 1.03628 (Atype_C_11)) – 0.225279(Atype_C_60) – 162164(Sr) – 0.091528(S_aaaC)	0.041	0.972	0.965	150.520	0.017	0.986

Table 6. Results of randomization tests for 2D-QSAR models.

Randomization tests				
Confidence leve	**l90%**	**95%**	**98%**	**99%**
Total trials	9	19	49	99
r from non-random	0.988	0.988	0.988	0.988
Random rs< non-random	9	19	49	99
Random rs> non-random	0	0	0	0
Mean of r from random trial	0.666	0.659	0.673	0.634
Standard deviation of random trials	0.050	0.056	0.066	0.061
Standard deviation from non-random r to mean	4.429	4.068	3.257	4.135

Table 7. Results of full cross-validation tests for 2D-QSAR models.

Rule	PRESS	Sum of sq dev.	r^2_{cv}
Leave-1-out	0.415	16.590	0.975
Leave-2-out	0.345	16.590	0.979
Leave-3-out	0.475	16.590	0.971
Leave-4-out	0.435	16.590	0.974
Leave-5-out	0.449	16.590	0.973
Leave-7-out	0.485	16.590	0.971
Leave-10-out	0.410	16.590	0.974

out" and "leave-10-out" are shown in Table 7, indicating that the QSAR models obtained were not by chance correlation. The developed 2D-QSAR model-1 thus was robust and was found satisfactory for predicting the activities of the test set (Table 1).

An important observation during generating QSAR models was the occurrence of Atype_C_11, Atype_C_6, Sr and Jurs-PNSA-3 as the frequent descriptors (Figure 2). The model-1 with five descriptors could explain

97.6% of the variance and predict 96.1% of the variance. Descriptors of Atype_C_6 and Atype_C_11 are the atom type AlogP descriptors used to characterize the hydrophobicity (logP) of molecules. The atomic contribution of individual atom types was proposed by Ghose and Crippen[41] toward the overall hydrophobicity of molecules where carbon, hydrogen, oxygen, nitrogen, sulfur and halogens were classified into 120 atom types.[36] Hydrogen and halogens are classified by the hybridization and oxidation state of the carbon they are bonded to, and carbon atoms are classified by their hybridization state and the chemical nature of their neighboring atoms. A total of 44 carbon types alone attest the complexity of the classification procedure. The negative slope of Atype_C_6 and Atype_C_11 in model-1 represents that activity decreases with an increase in lipophilicity related to C_6 and C_11 atom types for these inhibitors. The atom type C_6 is C in CH_2RX and C_11 is C in CR_3X where X represents any heteroatom (O, N, S, and halogens).[41] The E-state indices encode information about both the topological environment and the electronic interaction of an atom due to all other atoms in the molecule.[42] S_dssC is one descriptor of the E-state indices, representing the atomic type of =C< in the cyclic ring. S stands for the sum of the E-state values for a given atom type in a molecule. The set of bonds to a skeletal atom is given by a string of lower case letters: s (single), d (double), t (triple) and a (aromatic).[43] Jurs-PNSA-3 is the sum of the product of solvent-accessible surface area X partial charge for all negatively charged atoms.[44] It is negatively correlated with the activity, indicating that increasing this value in molecules could decrease biological activity. That may explain why molecules **2**, **6**, **10**, **12** and **17** with high values of Jurs-PNSA-3 were less active than molecules **22**, **23**, **24**, **27** and **28** with low values of Jurs-PNSA-3 (Table 1). Superdelocalizability (Sr) is an index of reactivity of occupied and unoccupied orbitals in aromatic hydrocarbons (AH),[23,45] which is proposed by Fukui using following formula:[46]

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

where S_r = superdelocalizability at position r; e_j = bonding energy coefficient in j^{th} MO (eigenvalue); c = molecular orbital coefficient at position r in the HOMO, and m = index of the HOMO. This index is based on the idea that early interaction of the molecular orbitals of two reactants may be regarded as a mutual perturbation, so that the relative energies of the two orbitals change together and maintain a similar degree of overlap as the reactants approach one another. This parameter for all atomic positions of a molecule gives a metric of electrophilicity, which is frequently employed to characterize molecular interactions.

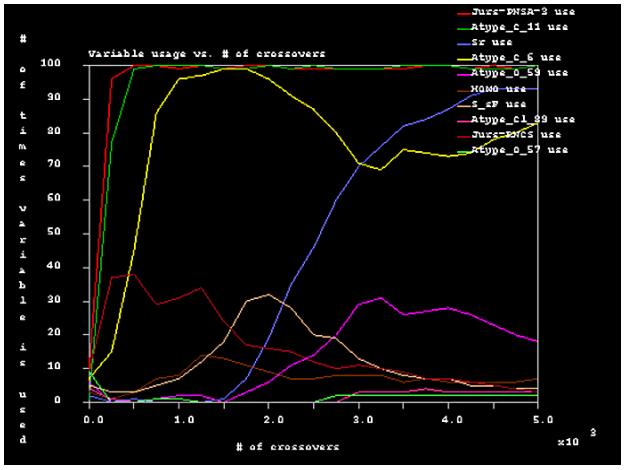


Figure 2. Descriptor usage graph during generating 2D-QSAR models by genetic function algorithm (GFA).

3. 2. 3D-QSAR Model

MFA samples the steric and electrostatic fields surrounding a set of ligands and constructs 3D-QSAR models by correlating these 3D field interaction energies computed using atomic coordinates of binding molecules with the corresponding bioactivity. The statistically significant 3D-QSAR model below was selected based on r^2 and r^2_{CV}. In the model-2, the descriptors H^+/x, H^+/y and H^+/z are the interaction energies between a proton probe and the molecule at the rectangular points x, y and z, respectively.[31,34] The descriptor CH_3/x is the corresponding interaction energy for the methyl probe. [30,34]

Model-2

pIC_{50} = 0.004785(H^+/239) – 0.01012(H^+/540) + 0.008195(H^+/491)
– 0.022389 (H^+/619) – 0.013445(CH_3/621) + 0.00728(CH_3/636)
+ 0.02768 (CH_3/404) + 0.023923(H^+/546) + 0.025499(CH_3/110)
– 0.015767(H^+/430) – 0.019974(CH_3/347) + 0.022202(CH_3/91)
– 0.014428(CH_3/307) + 0.006893(H^+/214) + 5.59793

N= 28, r^2 = 0.980, r^2_{CV} = 0.945, r^2_{BS} = 0.963+/–0.063, LSE = 0.012, r = 0.990, PRESS = 2.957, N' = 7, r^2_{pred} = 0.992

N is the number of compounds used in training set, and N' is the number of compounds in test set. The reasonable values of correlation coefficient r^2 of 0.980 and cross-validated r^2 of 0.945 indicate that this model could explain satisfactorily the variances in the activity, which can be used to design novel HCV NS5B polymerase inhibitors. The robust and highly predictive ability of the mo-

dels was reflected insufficiently only by the cross-validation test, thus the external predictive power of the model was evaluated with the test set molecules.[33] High r^2_{pred} value of 0.992 for the test set accounts for good predictive ability of model-2, which was also proved satisfactory for predicting the activity of the test set (Table 1).

Model-2 consists of the same number of methyl (CH_3) probes as proton (H^+) probes, indicating both of steric and electrostatic interactions are important to the biological activity. In order to visualize the models effectively, the most active molecule **22** (IC_{50} < 0.01 μM) and the least active molecule **6** (IC_{50} = 21 μM) specifying their location in the 3D-grid with selected field energies of model-2 are highlighted in Figure 3 and Figure 4, respectively. Through the comparison of these field energies of active and inactive compounds in Figure 3 and Figure 4, there are several obvious differences. Appearance of descriptor of CH_3/91 with positive coefficient between the R^2 group and R^3 group indicates that moderate steric substituents are favorable.[34,35] Electrostatic descriptor of H^+/214 with positive coefficient around the R^3 position suggests that electron donating group can increase the activity of molecules.[33] The negative coefficient of descriptor (H^+/540) near to the region of group R^4 shows that the subtle balance of electrostatic parameter is required at this position.[34,35] The unfavorable presence of CH_3/621 indicates that groups with big steric parameters at region of group R^4 lead to drop in activity.

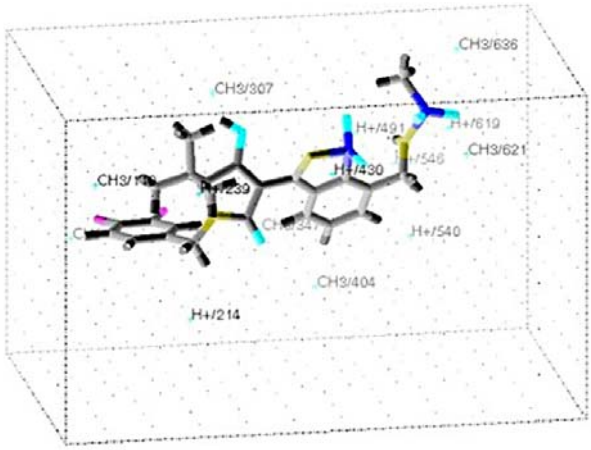


Figure 3. The most active molecule **22** (IC_{50} < 0.01 μM) within the 3D point grid of the model-2 is shown. CH_3 and H^+ represent steric interaction and electrostatic interaction, respectively.

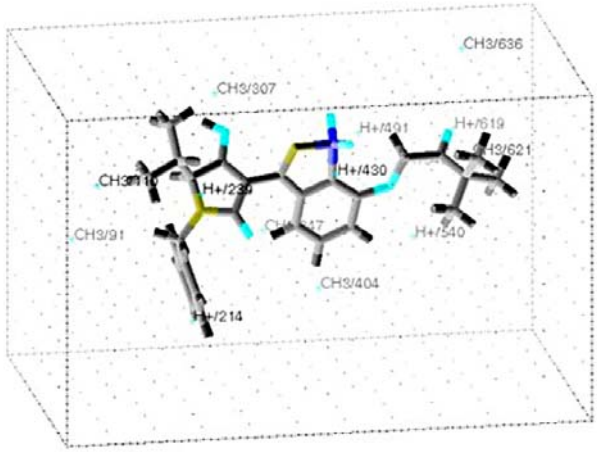


Figure 4. The least active molecule **6** (IC_{50} = 21 μM) within the 3D point grid of the model-2 is shown. CH_3 and H^+ represent steric interaction and electrostatic interaction, respectively.

4. Conclusions

Statistically significant QSAR models were generated with the purpose of deriving indispensable structural requirements for a series of novel HCV NS5B polymerase inhibitors comprising 1,1-dioxoisothiazoles and benzo[b]thiophene-1,1-dioxides. The 2D-QSAR models constructed by GFA methodology were validated by full cross-validation tests, randomization tests and external test set prediction, indicating that the bioactivities were principally governed by the atom type AlogP descriptors (Atype_C_6 and Atype_C_11), atomic type of =C in the cyclic ring (S_dssC), the sum of the product of solvent-accessible surface area X partial charge for all negatively charged atoms (Jurs-PNSA-3) and superdelocalizability (Sr). 3D-QSAR models developed based on steric and electrostatic descriptors by MFA-G/PLS method were to investigate the substitutional requirements for the favorable receptor-drug interaction. Well predicted activities compared with actual activities of test set molecules supports the significant predictive ability of the derived 2D/3D-QSAR models, thus, the models can be used to predict the anti-HCV activities of new analogs. The results in this research can provide a preliminary valuable guidance for continuing search of potent HCV NS5B polymerase inhibitors prior to synthesis.

5. Acknowledgements

The authors gratefully acknowledge the financial supports for this research 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 reviewers for their important technical suggestion.

6. References

1. T. A. Jennings, Y. Chen, D. Sikora, M. K. Harrison, B. Sikora, L. Huang, E. Jankowsky, M. E. Fairman, C. E. Cameron, K. D. Raney, *Biochemistry* **2008**, *47*, 1126–1135.

2. S. Yan, T. Appleby, E. Gunic, J. H. Shim, T. Tasu, H. Kim, F. Rong, H. Chen, R. Hamatake, J. Z. Wu, Z. Hong, N. Yao, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 28–33.
3. Q. L. Choo, G. Kuo, A. J. Weiner, L. R. Overby, D. W. Bradley, M. Houghton, *Science* **1989**, *244*, 359–362.
4. N. M. Dixit, J. E. Layden-Almer, T. J. Layden, A. S. Perelson, *Nature* **2004**, *432*, 922–924.
5. N. J. Liverton, M. K. Holloway, J. A. McCauley, M. T. Rudd, J. W. Butcher, S. S. Carroll, J. DiMuzio, C. Fandozzi, K. F. Gilbert, S. S. Mao, C. J. McIntyre, K. T. Nguyen, J. J. Romano, M. Stahlhut, B. L. Wan, D. B. Olsen, J. P. Vacca, *J. Am. Chem. Soc.* **2008**, *130*, 4607–4609.
6. C. R. Corbeil, P. Englebienne, C. G. Yannopoulos, L. Chan, S. K. Das, D. Bilimoria, L. L'Heureux, N. Moitessier, *J. Chem. Inf. Model.* **2008,** *48*, 902–909.
7. A. Schulze-Krebs, D. Preimel, Y. Popov, R. Bartenschlager, V. Lohmann, M. Pinzani, D. Schuppan, *Gastroenterology* **2005**, *129*, 246–258.
8. R. De Francesco, G.. Migliaccio, *Nature* **2005,** *436*, 953–960.
9. Z. Huang, M. G. Murray, J. A. Secrist, *Antiviral Res.* **2006**, *71*, 351–362.
10. M. Ikeda, N. Kato, *Adv. Drug Delivery Rev.* **2007**, *59*, 1277–1289.
11. G. Melagraki, A .Afantitis, H. Sarimveis, P. A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, *Bioorg. Med. Chem.* **2007**, *15*, 7237–7247.
12. H. Li, A. Linton, J. Tatlock, J. Gonzalez, A. Borchardt, M. Abreo, T. Jewell, L. Patel, M. Drowns, S. Ludlum, M. Goble, M. Yang, J. Blazel, R. Rahavendran, H. Skor, S. Shi, C. Lewis, S. Fuhrman, *J. Med. Chem.* **2007**, *50*, 3969–3972.
13. J. L. Clark, L. Hollecker, J. C. Mason, L. J. Stuyver, P. M. Tharnish, S. Lostia, T. R. McBrayer, R. F. Schinazi, K. A. Watanabe, M. J. Otto, P. A. Furman, W. J. Stec, S. E. Patterson, K. W. Pankiewicz, *J. Med. Chem.* **2005,** *48*, 5504–5508.
14. G. Maga, S. Gemma, C. Fattorusso, G. A. Locatelli, S. Butini, M. Persico, ? G. Kukreja, M. P. Romano, L. Chiasserini, L. Savini, E. Novellino, V. Nacci, S. Spadari, G. Campiani, *Biochemistry* **2005,** *44*, 9637–9644.
15. J. A. Lee, H. O. Kim, D. K. Tosh, H. R. Moon, S. Kim, L. S. Jeong, *Org. Lett.* **2006**, *8*, 5081–5083.
16. X. Liao, G. Butora, D. B. Olsen, S. S. Carroll, D. R. McMasters, J. F. Leone, M. Stahlhut, G. A. Doss, L. Yang, M. MacCoss, *Tetrahedron Lett.* **2008**, *49*, 4149– 4152.
17. P. D. Patel, M. R. Patel, N. Kaushik-Basu, T. T. Talele, *J. Chem. Inf. Model.* **2008,** *48*, 42–55.
18. S. H. Kim, M. T. Tran, F. Ruebsam, A. X. Xiang, B. Ayida, H. McGuire, D. Ellis, J. Blazel, C. V. Tran, D. E. Murphy, S. E. Webber, Y. Zhou, A. M. Shah, M. Tsan, R. E. Showalter, R. Patel, A. Gobbi, L. A. LeBrun, D. M. Bartkowski, T. G. Nolan, D. A. Norris, M. V. Sergeeva, L. Kirkovsky, Q. Zhao, Q. Han, C. R. Kissinger, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4181–4185.
19. K. K. Sahu, V. Ravichandran, P. K. Jain, S. Sharma, V. K. Mourya, R. K. Agrawal, *Acta Chim. Slov.* **2008**, *55*, 138–145.
20. P. Silakari, S. D. Shrivastava, G. Silakari, D. V. Kohli, G. Rambabu, S. Srivastava, S. K. Shrivastava, O. Silakari, *Eur. J. Med. Chem.* **2008**, *43*, 1559–1569.
21. T. J. Oprea, G. L. Waller, G. R. Marshall, *J. Med. Chem.* **1994**, *37*, 2206– 2215.
22. J. T. Leonard, K. Roy, *Eur. J. Med. Chem.* **2008**, *43*, 81–92.
23. Cerius^2, Version 4.10, A. San Diego, Inc.CA, USA, **2005**.
24. J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.
25. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III, W. M. Skiff, *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
26. Q. Shen, Q. Z. Lü, J. H. Jiang, G. L. Shen, R. Q. Yu, *Eur. J. Pharm. Sci.* **2003**, *20*, 63–71.
27. D. Rogers, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
28. J. H. Friedman, *Ann. Statistics* **1991**, *19*, 1–141.
29. J. Holland, *Adaptation in Artificial and Natural System*. University of Michigan Press, Ann Arbor, **1975**, pp.1–80.
30. A. Hirashima, M. Morimoto, E. Kuwano, M. Eto, *Bioorg. Med. Chem.* **2003**, *11*, 3753–3760.
31. A. Hirashima, T. Eiraku, E. Kuwano, M. Eto, *Internet Electron. J. Mol. Des.* **2003**, *2*, 511–526.
32. P. Bhattacharya, J. T. Leonard, K. Roy, *J. Mol. Model.* **2005**, *11*, 516–524.
33. T. Equbal, O. Silakari, M. Ravikumar, *Eur. J. Med. Chem.* **2008**, *43*, 204–209.
34. K.-X. Chen, H.-Y. Xie, Z-G. Li, J-R. Gao, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5381–5386
35. A. R. Shaikh, M. Ismael, C. A. Del Carpio, H. Tsuboi, M. Koyama, A. Endou, M. Kubo, E. Broclawik, A. Miyamoto, *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5917– 5925.
36. P. C. Nair, M. E. Sobhia, *Eur. J. Med. Chem.* **2008**, *43*, 293–299.
37. S. Deswal, N. Roy, *Eur. J. Med. Chem.* **2006**, *41*, 1339–1346.
38. P. M. Sivakumar, S. P. Seenivasan, V. Kumar, M. Doble, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1695–1700.
39. S. V. Sambasivarao, L. K. Soni, A. K. Gupta, S.G. Kaskhedikar, *Acta Chim. Slov.* **2008**, *55*, 338–342.
40. Y. Fan, L. M. Shi, K. W. Kohn, Y. Pommier, J. N. Weinstein, *J. Med. Chem.* **2001**, *44*, 3254–3263.
41. A. K. Ghose, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.
42. L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
43. P. Crivori, A. Morelli, D. Pezzetta, M. Rocchetti, I. Poggesi**,** *Eur. J. Pharm. Sci.* **2007**, *32*, 169–181.
44. R. H. Rohrbaugh, P. C. Jurs, *Anal. Chim. Acta.* **1987**, *199*, 99–109.
45. M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1043.
46. K. Fukui, *Theory of Orientation and Stereoselection*, Springer-Verlag: New York, **1975**, pp. 34–39.

Povzetek

V prispevku je predstavljena kvantitativna študija povezave med strukturo in reaktivnostjo nekaterih novih inhibitorjev HCV NS5B polimeraze, ki obsegajo 1,1-dioksoizotiazole in benzo[b]tiofen-1,1-diokside, z uporabo tehnik algoritma genetske funkcije (GFA) in analize molekulskega polja (MFA). Statistično pomembna 2D/3D-QSAR modela (r^2 > 0.975) sta pokazala na neobhodno potrebne strukturne zahteve za izboljšanje aktivnosti tega tipa spojin. Visoke vrednosti r^2_{CV} (0.961 in 0.945) in r^2_{pred} (0.856 in 0.992) za 2D/3D-QSAR modela kažejo na pomembno napovedovalno sposobnost uporabljenih modelov. Validacija uporabljenih modelov je bila narejena z navzkrižnim validacijskim testom in napovedjo neodvisnega testnega niza. Rezultate raziskave bi lahko uporabili za spreminjanje anti-HCV NS5B polimerazne aktivnosti pri študiranemu razredu analogov.