Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 7237–7247
Bioorganic & Medicinal Chemistry

Identification of a series of novel derivatives as potent HCV inhibitors by a ligand-based virtual screening optimized procedure

Georgia Melagraki,[a,b,c,d] Antreas Afantitis,[a,b,c,d] Haralambos Sarimveis,[a,*] Panayiotis A. Koutentis,[d] John Markopoulos[e] and Olga Igglessi-Markopoulou[a]

[a]*School of Chemical Engineering, National Technical University of Athens, Athens, Greece*
[b]*Department of ChemoInformatics, NovaMechanics Ltd, Cyprus*
[c]*Cyano Research Corporation Ltd, PO Box 28670, 2081 Nicosia, Cyprus*
[d]*Department of Chemistry, University of Cyprus, PO Box 20537, 1678 Nicosia, Cyprus*
[e]*Department of Chemistry, University of Athens, Athens, Greece*

Received 7 June 2007; revised 2 August 2007; accepted 21 August 2007
Available online 25 August 2007

Abstract—This paper presents the results of a ligand-based virtual screening optimized procedure on 98 compounds which have been recently evaluated as inhibitors of genotype 1 HCV polymerase. First, quantitative structure–activity patterns are investigated for the selected compounds and then structural modifications are proposed to afford novel active patterns. An accurate and reliable QSAR model involving five descriptors that is able to predict successfully the HCV inhibitory potency against genotype 1 HCV polymerase is presented. Furthermore, the effects of various structural modifications on biological activity are investigated and biological activities of novel structures are estimated using the developed QSAR model. More specifically a search for optimized pharmacophore patterns by insertions, substitutions, and ring fusions of pharmacophoric substituents of the main building block scaffolds is described. The detection of the domain of applicability defines compounds whose estimations can be accepted with confidence.

1. Introduction

The hepatitis C virus (HCV) is a member of the Flaviviridae family. Chronic infection with HCV is associated with liver cirrhosis that often leads to hepatic failure and hepatocellular carcinoma. Although the number of new infections has been significantly reduced by the introduction of reliable blood testing, more than 170 million people worldwide are chronically infected with HCV, which has become a global health threat and the main cause of adult liver transplants in developed nations. There is as yet no effective therapy for HCV-associated chronic hepatitis. Hepatitis C is considered a major public health threat and current therapies still call for major improvements.[1,2]

Current treatments with interferon R (IFN-R) alone or in combination with ribavarin are effective only in limited cases and exhibit severe adverse side effects. There is thus an obvious need to develop effective therapeutic strategies to cure HCV-associated hepatitis.[3,4] HCV has become the paramount target of antiviral protease inhibitor research, particularly HCV genotype 1. This virus affects the most people worldwide and is considered the most challenging genotype to treat; indeed, for the large number of patients who fail standard therapies, there exists no alternative treatment. Protease inhibitors may be the most promising candidates to fill this unmet medical need. The most studied targets for anti-HCV therapy are the NS3 protease and the NS5b polymerase.[5,6] In the case of HCV NS5b polymerase, both nucleoside and non-nucleoside inhibitors have appeared recently in the literature.[7,8]

In this work, we have selected from the literature 98 compounds which were evaluated as inhibitors of genotype 1 HCV polymerase.[9–11] First, quantitative

Keywords: HCV; QSAR; Ligand-based design; Virtual screening.
* Corresponding author. Tel.: +30 210 772 3237; fax: +30 210 772 3138; e-mail: hsarimv@central.ntua.gr

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.08.036

structure–activity patterns were investigated for the selected compounds and then structural modifications were proposed to afford novel active patterns. The first major result is the development of an accurate and reliable QSAR model involving five descriptors that is able to predict successfully the HCV inhibitory potency against genotype 1 HCV polymerase. As a next step, the effects of various structural modifications on biological activity were investigated and biological activities of novel structures were estimated using the developed QSAR model. The detection of the domain of applicability defined the compounds whose estimations can be accepted with confidence.

2. Material and methods

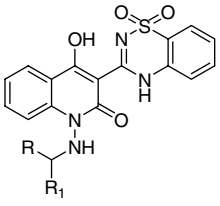
2.1. Data set

The database consists of 98 recently discovered inhibitors of genotype 1 HCV polymerase (Tables 1–6).[9–11] In order to model and predict the inhibitory activity of HCV inhibitors, 69 physicochemical constants, topological and structural descriptors (Table 7) were considered as possible input candidates to the model. Before the calculation of the descriptors, all structures were fully optimized using CS Mechanics and more specifically MM2 force fields and the Truncated-Newton–Raphson optimizer, which provide a balance between speed and accuracy (ChemOffice Manual). Before calculating the HOMO and LUMO Energies (eV) all the structures were additionally fully optimized using the semiempirical AM1 basis set. All the descriptors were calculated using ChemSar and Topix.[12,13]

2.2. Separation into a training and a validation set

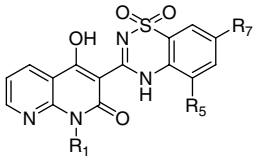
The separation of the data set into training and validation sets was performed according to the popular Kennard and Stones algorithm.[14] The algorithm starts by finding two samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, for example, the Euclidean distance. These two samples are removed from the original data set and placed into

Table 1. Biochemical potency of *N*-1-heteroalkyl-4-hydroxyquinolon-3-yl-benzothiadiazines



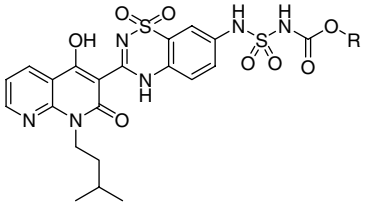
ID	R	R_1	IC_{50} (μM), observed	$\log(1/IC_{50})$ (μM), observed	Training data $\log(1/IC_{50})$ (μM), predicted	Test data $\log(1/IC_{50})$ (μM), predicted
1	Ph	H	5.09	−0.7067	−1.0280	
2	2-BrC_6H_4	H	6.72	−0.8274	−0.8012	
3[b]	3-BrC_6H_4	H	1.04	−0.0170		−0.5975
4[b]	4-BrC_6H_4	H	10.36	−1.0154		−0.5646
5[b]	2-MeC_6H_4	H	6.12	−0.7868		−0.7395
6[b]	3-MeC_6H_4	H	8.21	−0.9143		−0.5609
7	2-Thienyl	H	2.33	−0.3674	−0.6316	
8	2-Thiazolyl	H	35.8	−1.5539	−0.1901	
9[b]	2-Furyl	H	3.26	−0.5132		−0.9046
10	3-Furyl	H	1.55	−0.1903	−0.9292	
11[b]	3-Me-thien-2-yl	H	15.0	−1.1761		−0.2726
12[b]	5-Cl-thien-2-yl	H	4.78	−0.6794		−0.5892
13[b]	Pr	H	0.951	0.0218		−0.2419
14[b]	Bu	H	0.928	0.0325		−0.2823
15	*i*-Bu	H	0.629	0.2013	0.0611	
16	Neopentyl	H	0.411	0.3862	0.4836	
17	*i*-Pr	H	0.941	0.0264	−0.0380	
18[b]	Cyclopropyl	H	0.285	0.5452		−0.1014
19[b]	Cyclohexyl	H	2.528	−0.4028		−0.2092
20	Me	Me	6.35	−0.8028	−0.1390	
21[b]	Et	Et	2.46	−0.3909		−0.1014
22	Et	Pr	1.89	−0.2765	−0.0504	
23[b]	Pr	Pr	2.24	−0.3502		−0.0163
24[b]	Pr	*i*-Pr	6.40	−0.8062		−0.1141
25	Me	Ph	16.5	−1.2175	−0.5051	
26[b]	Cyclobutyl		0.278	0.5560		0.0079
27	Cyclopentyl		0.747	0.1267	−0.0674	
28	Cyclohexyl		0.356	0.4486	−0.1315	
29	Cycloheptyl		1.04	−0.0170	−0.1626	
30	4-Pyranyl		4.63	−0.6656	0.1857	

Table 2. Biochemical potency of *N*-1-benzyl and *N*-1-(3-methylbutyl)-4-hydroxy-1,8-naphthyridon-3-yl benzothiadiazine analogs containing substituents on the aromatic ring



ID	R_1	R_5	R_7	IC_{50} (μM), observed	$\log(1/IC_{50})$ (μM), observed	Training data $\log(1/IC_{50})$ (μM), predicted	Test data $\log(1/IC_{50})$ (μM), predicted
31	Benzyl	H	H	5.8	−0.7634	−1.3546	
32	Benzyl	OMe	H	18	−1.2553	−0.9298	
33	Benzyl	H	OMe	8.31	−0.9196	−0.8311	
34[b]	Benzyl	OH	H	8.37	−0.9227		−0.8887
35[b]	Benzyl	H	Me	25.7	−1.4099		−0.9581
36[b]	Benzyl	Me	H	6.7	−0.8261		−1.0038
37	Benzyl	H	Br	16.75	−1.2240	−1.0036	
38[b]	Benzyl	Br	H	6.2	−0.7924		−1.0989
39	3-Methylbutyl	H	H	0.81	0.0915	−0.3554	
40	3-Methylbutyl	OMe	H	17.58	−1.2450	−0.2026	
41[b]	3-Methylbutyl	H	OMe	1.13	−0.0531		−0.1781
42	3-Methylbutyl	H	OH	0.39	0.4089	0.1360	
43	3-Methylbutyl	H	$-OCH_2CH_2CH_3$	1.62	−0.2095	−0.0582	
44	3-Methylbutyl	H	$-OCH_2CO_2t$-Bu	5	−0.6990	0.6768	
45	3-Methylbutyl	H	$-OCH_2CO_2H$	0.367	0.4353	0.3627	
46	3-Methylbutyl	H	$-OCH_2CONMe_2$	0.934	0.0297	0.7455	
47	3-Methylbutyl	H	$-OCH_2CONHMe$	0.18	0.7447	0.4441	
48	3-Methylbutyl	H	$-OCH_2CONH_2$	0.046	1.3372	0.6624	
49[b]	3-Methylbutyl	H	$-OCH_2CH_2NH_2$	0.637	0.1959		0.2903
50	3-Methylbutyl	H	$-OCH_2CN$	0.141	0.8508	0.3864	
51	*i*-Pentyl	H	$-NH_2$	0.31	0.5086	0.4659	
52	*i*-Pentyl	H	$-NHCH_2CN$	0.12	0.9208	0.4848	
53[b]	*i*-Pentyl	H	$-NHCH_2CONH_2$	0.47	0.3279		0.6968
54	*i*-Pentyl	H	$-NHCOCF_3$	0.906	0.0429	0.1542	
55	*i*-Pentyl	H	$-NHSO_2Ph$	0.041	1.3872	0.8994	
56[b]	*i*-Pentyl	H	$-NHSO_2i$-Pr	0.008	2.0969		1.7884
57	*i*-Pentyl	H	$-NHSO_2(CH_2)_3CH_3$	0.022	1.6576	1.1332	
58	*i*-Pentyl	H	$-NHSO_2Me$	0.002	2.6990	1.5078	

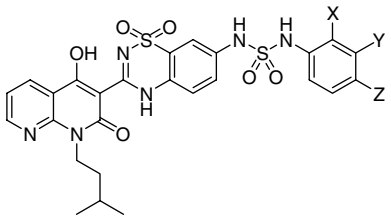
Table 3. Biochemical potency of *N*-alkyl-4-hydroxyquinolon-3-yl-benzothiadiazine



ID	R	IC_{50} (μM), observed	$\log(1/IC_{50})$ (μM), observed	Training data $\log(1/IC_{50})$ (μM), predicted	Test data $\log(1/IC_{50})$ (μM), predicted
59	CH_3	0.063	1.2007	1.3020	
60	CH_2CH_2Cl	0.189	0.7235	1.4784	
61[b]	CH_2CHCH_2	0.088	1.0555		1.5171
62	CH_2CCH	0.072	1.1427	1.2451	
63	CH_2CH_2CN	0.035	1.4559	1.4893	
64[b]	CH_2Ph	0.058	1.2366		1.1106
65[b]	$CH_2CH_2NH_2$	0.021	1.6778		1.4791
66	$CH_2CO_2CH_2CH_3$	0.061	1.2147	1.5062	
67	$CH_2CH_2OCH_3$	0.087	1.0605	1.4727	
68	$CH_2CH_2OCH_2Ph$	0.096	1.0177	1.1676	
69[b]	CH_2CO_2H	0.050	1.3010		1.5286

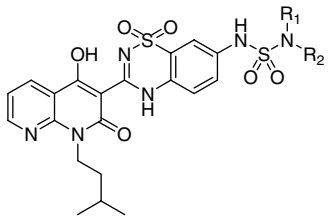
the calibration data set. This procedure is repeated until the desired number of samples has been reached in the calibration set. The advantages of this algorithm are that the calibration samples map the measured region of the

Table 4. Biochemical potency of *N*-alkyl-4-hydroxyquinolon-3-yl-benzothiadiazine



ID	X	Y	Z	IC_{50} (μM), observed	log(1/IC_{50}) (μM), observed	Training data log(1/IC_{50}) (μM), predicted	Test data log(1/IC_{50}) (μM), predicted
70	H	H	CO_2Me	0.048	1.3188	1.1582	
71[b]	H	H	CO_2H	0.020	1.6990		1.2876
72	H	H	CONHMe	0.043	1.3665	1.4642	
73	H	H	$CONH_2$	0.018	1.7447	1.5259	
74	H	CO_2Et	H	0.132	0.8794	1.3241	
75[b]	H	CO_2H	H	0.115	0.9393		1.2397
76[b]	H	$CONH_2$	H	0.043	1.3665		1.6015
77	H	$CONHCH_2CONH_2$	H	0.043	1.3665	1.9998	
78[b]	CO_2Me	H	H	0.129	0.8894		1.0767

Table 5. Biochemical potency of *N*-alkyl-4-hydroxyquinolon-3-yl-benzothiadiazine



ID	R_1	R_2	IC_{50} (μM), observed	log(1/IC_{50}) (μM), observed	Training data log(1/IC_{50}) (μM), predicted	Test data log(1/IC_{50}) (μM), predicted
79	H	CH_2CH_2Ph	0.018	1.7447	0.9244	
80[b]	H	CH_2Ph	0.0055	2.2596		0.7451
81	H	Ph	0.041	1.3872	0.8738	
82	H	CH_2CH_2OH	0.107	0.9706	1.6083	
83[b]	H	Cyclohexyl	0.067	1.1739		1.3301
84	H	Cyclopentyl	0.039	1.4089	1.4256	
85[b]	H	$CH_2CH_2NH_2$	0.010	2.0000		1.5672
86	H	4-Piperidinyl	0.032	1.4949	2.0141	
87	H	$CH_2CH_2CONH_2$	0.020	1.6990	1.4979	
88[b]	H	4-$MeOC_6H_4CH_2$	0.010	2.0000		1.0536
89[b]	H	3-$MeOC_6H_4CH_2$	0.015	1.8239		0.9987
90	H	2-$MeOC_6H_4CH_2$	0.057	1.2441	0.8640	
91		Piperidinyl	0.051	1.2924	1.3555	
92[b]		Pyrrolidinyl	0.027	1.5686		1.4121
93		Azetidinyl	0.024	1.6198	1.5494	

input variable space completely with respect to the induced metric and that the test samples all fall inside the measured region. According to Tropsha et al.[15] and Wu et al.[16] the Kennard and Stones algorithm is one of the best ways to build training and test sets.

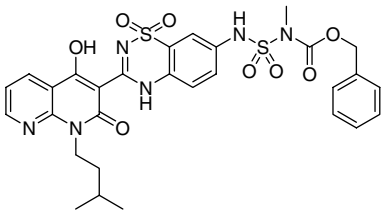
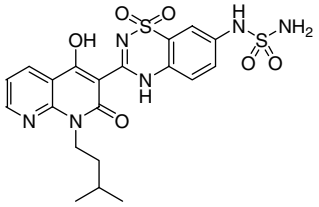
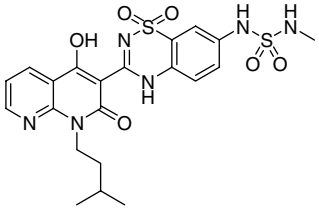
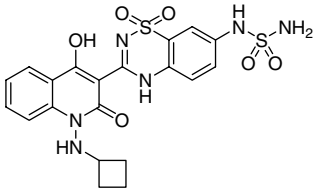
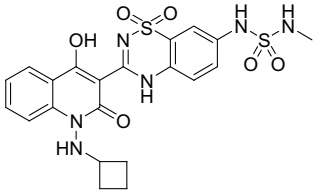
2.3. Multiple Linear Regression (MLR) model development—variable selection

Our first objective was to determinate the best variables which produce the most significant linear QSAR models linking the structure of compounds with their binding affinity. The ES-SWR algorithm was used on the training data set to select the most appropriate descriptors. Elimination Selection-Stepwise Regression (ES-SWR) is a popular stepwise technique[17] that combines Forward Selection (FS-SWR) and Backward Elimination (BE-SWR).

2.4. Model validation

The accuracy of the proposed MLR model was illustrated using the following evaluation techniques: leave-one-out (LOO) and leave-five-out (L5O) cross-validation procedures, validation through an external test set, and Y-randomization.

Table 6. Biochemical potency of *N*-alkyl-4-hydroxyquinolon-3-yl-benzothiadiazine

ID	Structure	IC_{50} (μM), observed	$\log(1/IC_{50})$ (μM), observed	Training data $\log(1/IC_{50})$ (μM), predicted	Test data $\log(1/IC_{50})$ (μM), predicted
94		0.121	0.9172	0.8928	
95		0.009	2.0458	1.3865	
96[b]		0.014	1.8539		1.1590
97		0.0052	2.2840	1.5803	
98		0.0087	2.0605	1.5371	

2.5. Cross-validation test

Cross-validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects. For each data set, an input–output model is developed, based on the utilized modeling technique. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model).[18]

2.6. Validation through the external validation set

According to Tropsha' group[15,19] a QSAR model is considered predictive, if the following conditions are satisfied:

$$R^2_{\text{pred}} > 0.6 \tag{1}$$

$$\frac{(R^2 - R^2_{\text{o}})}{R^2} \text{ or } \frac{(R^2 - R'^2_{\text{o}})}{R^2} \text{ is less than } 0.1 \tag{2}$$

$$k \text{ or } k' \text{ is close to } 1 \tag{3}$$

In Eqs. 2 and 3 R^2 is the coefficient of determination between experimental values and model prediction on the training set. Mathematical definitions of R^2_{o}, R'^2_{o}, k and k' are based on regression of the observed activities against predicted activities and the opposite (regression of the predicted activities against observed activities). The definitions are presented clearly in Ref. 20 and are not repeated here for brevity.

2.7. Y-Randomization test

This technique ensures the robustness of a QSAR model.[15,21] The dependent variable vector (biological

Table 7. Calculated descriptors

ID	Description	Notation	ID	Description	Notation
1	Molar Refractivity	MR	**2**	Diameter	Diam
3	Partition Coefficient (Octanol Water)	*C*log *P*	**4**	Molecular Topological Index	TIndx
5	Principal Moment of Inertia Z	PMIZ	**6**	Number of Rotatable Bonds	NRBo
7	Principal Moment of Inertia Y	PMIY	**8**	Polar Surface Area	PSAr
9	Principal Moment of Inertia X	PMIX	**10**	Radius	Rad
11	LUMO Energy	LUMO	**12**	Shape attribute	ShpA
13	HOMO Energy	HOMO	**14**	Shape coefficient	ShpC
15	Balaban Index	BIndx	**16**	Sum of Valence Degrees	SVDe
17	Cluster Count	ClsC	**18**	Total Connectivity	TCon
19	Wiener Index	WIndx	**20**	Total Valence Connectivity	TVCon
21	DistEqTotal	DistEqTotal	**22**	Randic 0	Chi0
23	Randic 1	Chi1	**24**	Randic 2	Chi2
25	Randic 3	Chi3	**26**	Randic 4	Chi4
27	Randic Information 0	ChiInf0	**28**	Randic Information 1	ChiInf1
29	Randic Information 2	ChiInf2	**30**	Randic Information 3	ChiInf3
31	Randic Information 4	ChiInf4	**32**	Molecular Weight	MW
33	Randic Mod	ChiMod	**34**	Xu1	Xu1
35	Xu2	Xu2	**36**	Xu3	Xu3
37	Balaban Topological	TopoJ	**38**	Number of Branches	NBranch
39	Number of Rings	NRings	**40**	Wiener Dim	Wiener Dim
41	Bertz	Bertz	**42**	AtomCompMean	AtomCompMean
43	AtomCompTot	AtomCompTot	**44**	Zagreb1	Zagreb1
45	Zagreb2	Zagreb2	**46**	Kappa1	Kappa1
47	Kappa2	Kappa2	**48**	Kappa3	Kappa3
49	Wiener Distance	WienerDistCode	**50**	Polarity	Polarity
51	DistEqMean	DistEqMean	**52**	Quadratic	Quadr
53	InfMagnitDistTot	InfMagnitDistTot	**54**	ScHultz	ScHultz
55	Gordon	Gordon	**56**	Kier-Hall 0	Ki0
57	Kier-Hall 1	Ki1	**58**	Kier-Hall 2	Ki2
59	Kier-Hall 3	Ki3	**60**	Kier-Hall 4	Ki4
61	Kier-Hall Information 0	KiInf0	**62**	Kier-Hall Information 1	KiInf1
63	Kier-Hall Information 2	KiInf2	**64**	Kier-Hall Information 3	KiInf3
65	Kier-Hall Information 4	KiInf4	**66**	Randic Cluster 3	ChiCl3
67	Randic Cluster 4	ChiCl4	**68**	Wiener Information	InfWiener
69	Wiener Index	WIndx			

action) is randomly shuffled and a new QSAR model is developed, using the given modeling algorithm. The procedure is repeated several times and the new QSAR models are expected to have low R^2 and Q^2 values. If the opposite happens then an acceptable QSAR model cannot be obtained for the specific modeling method and data.

3. Defining model applicability domain

In order for a QSAR model to be used for screening new compounds, its domain of application[15,20] must be defined and predictions for only those compounds that fall into this domain may be considered reliable. *Extent of Extrapolation*[15] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i[22] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i(X^{\mathrm{T}}X)^{-1}x_i^{\mathrm{T}} \quad (4)$$

In Eq. 4 x_i is the row vector containing the k model parameters of the query compound and X is the $n \times k$ matrix containing the k model parameters for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

4. Results and discussion

First, the data set of 98 derivatives was partitioned into a training set of 60 compounds, and a validation set of 38 compounds according to the Kennard and Stones[14] algorithm. The algorithm was applied on the complete database consisting of all 69 available descriptors. The validation examples are marked with [b] in Tables 1–6. The validation data were not involved by any means in the process of selecting the most appropriate descriptors or in the development of the QSAR model. They were considered as a completely unknown external set of data, which was used only to test the accuracy of the produced model.

Table 8. Correlation matrix of the 5 selected descriptors

	*C*log *P*	HOMO	Ki2	KiInf0	KiInf3
*C*log *P*	1				
HOMO	0.105	1			
Ki2	0.248	0.010	1		
KiInf0	0.091	0.062	0.582	1	
KiInf3	0.300	0.162	0.578	0.550	1

The MLR QSAR model was thus developed by applying the ES-SWR algorithm on the set of training data. The result was the following six-parameter (five descriptors and the intercept) equation:

$$\log(1/\mathrm{IC}_{50}) = 4.12 - 0.316 * C\log P + 0.346 * \mathrm{HOMO} + 0.434 * \mathrm{Ki2} + 2.24 * \mathrm{KiInf0} - 2.31 * \mathrm{KiInf3} \quad (5)$$

$n = 60$; $R^2 = 0.74$; $F = 29.97$; RMSE = 0.52; $Q^2 = 0.67$; $S_{\mathrm{PRESS}} = 0.58$.

From the above equation we can conclude that the most significant descriptors according to the ES-SWR algorithm are Lipophilicity ($C\log P$), HOMO energy, Kier and Hall index order 2 (Ki2), and Kier and Hall information indices order 0 and 3 (KiInf0, KiInf3). Table 8 presents the correlation matrix, where it is clear that the seven selected descriptors are not highly correlated. The chemical meaning of the seven descriptors is briefly described next.

Lipophilicity is known to be important for absorption, permeability, and in vivo distribution of organic compounds[17] and has been used as a physicochemical descriptor in QSARs with great success. Lipophilicity can be factorized in two main terms: Hydrophobicity which refers to non-polar interactions (such as dispersion forces, hydrophobic interactions) of the solute with organic and aqueous phases, and polarity which refers to polar interactions (such as ion–dipole interactions, hydrogen bond induction, orientating forces, etc.). From the derived QSAR equation we can conclude that lipophilic groups do not favor the biological action under study.

Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier Orbital Theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in determining the chemical reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction. The energies of the highest-occupied and the lowest-unoccupied molecular orbitals (HOMO/LUMO energies) are frequently used quantum chemical descriptors. As a consequence, the derived QSAR models will include information regarding the nature of the intermolecular forces involved in determining the biological activity of the compounds in question. HOMO energy in particular has been identified as being of significant value to QSAR studies.[17] Molecules with high HOMO (highest occupied molecular orbital energy) values can donate their electrons more easily compared to molecules with low HOMO energy values. The HOMO energy value is increased with the presence of electron-donating groups (EDG) such us NMe_2, NH_2, NHEt, and OMe and decreased with the presence of electron-withdrawing groups (EWG) such as halogens, cyano and nitro groups. From the derived QSAR equation we can conclude that EDGs favor the biological action under study.

In addition to the aforementioned indices, three topological indices were found to significantly influence the activity.[17,23] Topological indices give information not only about the atomic constitution of a compound but also about the presence and character of chemical bonds by which the atoms are connected to each other. Connectivity indices, such as Ki2 index, are among the most popular topological indices and can be used to characterize edges as a primitive bond order accounting for bond accessibility, that is, the accessibility of a bond to encounter another bond in intermolecular interactions. The value of Ki2 as defined by Kier and Hall is used to take into account all valence electrons of atoms and is useful for characterizing heteroatoms and carbon atoms involved in multiple bonds. Information indices such as KiInf0 and KiInf3 are graph theoretical invariants that view the molecular graph as a source of different probability distributions to which information theory definitions can be applied. They can be considered as a quantitative measure of the lack of structural homogeneity or the diversity of a graph, in this way being related to symmetry associated with structure.

Eq. 5 was used to predict the binding affinity for the validation examples. The results are presented in the last columns of Tables 1–6 and correspond to the following statistics: $R^2_{\mathrm{pred}} = 0.81$, RMSE = 0.49. The leverages for all 38 testing compounds were computed (Table 9). All 38 compounds in the test set fall inside the domain of the model (warning leverage limit 0.30). The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure are adequate to generate an efficient QSAR model for predicting the binding affinity of different compounds.

The proposed model (Eq. 5) passed all the tests related to the predictive ability (Eqs. 1–3).

$$R^2_{\mathrm{pred}} = 0.81 > 0.6$$

$$\frac{(R^2 - R^2_{\mathrm{o}})}{R^2} = -0.22 < 0.1, \quad k = 1.10$$

For a more exhaustive testing of the predictive power of the model, validation of the model was also carried out using the LOO and the L5O cross-validation techniques on the training set of compounds. The L5O method was implemented by selecting randomly groups of five compounds from the available training data. Each group was left out and that group was predicted by the model developed from the remaining observations. Three thousand random groups of five compounds were selected for the implementation of the L5O cross-validation test. It should be emphasized that the procedure for developing the QSAR models included the selection of the best descriptors. Therefore, each time one (LOO) or five (L5O) compounds were excluded from the training set, the modeling procedure selected the best descriptors and developed an MLR model based only on the remaining observations. The excluded compounds were not involved by any means in the development of the model. It was important that the model was stable to the inclusion–exclusion of compounds. The results produced by the LOO ($Q^2 = 0.67$) and the L5O

Table 9. Leverages for the test set

Compound ID	Leverage limit 0.30
3	0.0974
4	0.1003
5	0.0721
6	0.0787
9	0.1045
11	0.0483
12	0.0729
13	0.1094
14	0.1236
18	0.0597
19	0.0762
21	0.1036
23	0.1662
24	0.0857
26	0.0773
34	0.1070
35	0.0942
36	0.1003
38	0.1448
41	0.0972
49	0.0837
53	0.1157
56	0.0542
61	0.0485
64	0.0646
65	0.0898
69	0.0784
71	0.0865
75	0.0782
76	0.0845
78	0.0765
80	0.0796
83	0.0762
85	0.0663
88	0.0723
89	0.0814
92	0.0537
96	0.0403

($Q^2_{L5O} = 0.71$) cross-validation tests illustrated the quality of the obtained model.

The model was further validated by applying the Y-randomization. Several random shuffles of the Y vector were performed and the low R^2 and Q^2 values that were obtained show that the good results in our original model are not due to a chance correlation or structural dependency of the training set. It should be noted that for each random permutation of the Y vector, the complete training procedure was followed for developing the new QSAR model, including the selection of the most appropriate descriptors. The results of the Y-randomization from 10 shuffles of the Y-vector gave R^2 and Q^2 values in the ranges of 0.0–0.3 and 0.0–0.18, respectively.

While ligand flexibility is an initially useful feature in identifying compounds with good biological activity it can often result in low specificity. The introduction of rigidity to a ligand can increase specificity. In order to examine this for the above HCV inhibitors a virtual screening study was performed using the above-constructed model where the objective was to introduce rigidity and thereby provide compounds with potentially improved specificity with equal or improved predicted biological activity. Such a study could provide new synthetic targets worthy of investigation.[24–26] The representative modifications that led to novel compounds are shown in Tables 10–15. Biological activities of the compounds characterized as 'actives' were estimated using the developed MLR equation. The activity values together with the leverages are also shown in Tables 10–15.

Initially compound **95** was chosen and the C-4 substituent was screened. Replacement of the C-4 hydroxy with amino, thiol, hydroxylamino or hydrazino substituents led to compounds with predicted activities all within the domain of applicability (Table 10). The hydroxylamino had a higher activity (1.7075) although a slightly reduced domain of applicability (0.1824).

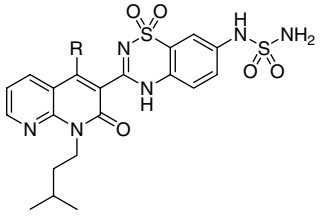
Ring fusion was introduced at the junction between the thiadiazine and the pyridopyridine for compound **95** and compound **1n** with the highest domain of applicability (0.2336). This compound was chosen since the domain of applicability indicated it would tolerate the greatest structural changes.

As shown in Table 11, only one ring fusion compound **6n** was tolerated by the model and found to be within the acceptable domain of applicability (0.0756). Furthermore compound **6n** showed increased activity (2.1518) and so this ring fusion was chosen for further screening (Table 12).

The carbonyl functionality of the pyridone was subsequently modified to that of an oxime compd **9n** and hydrazone compd **10n**. These modifications led to large increases in activity but were clearly out of the model's domain of applicability. A second ring fusion involving transformation of the diaminosulfonyl group into a benzo-1,2,5-thiadiazolidine-1,1-dioxide also gave some interesting information. The model tolerated this second ring fusion although the domain of applicability was only marginally acceptable (Table 13).

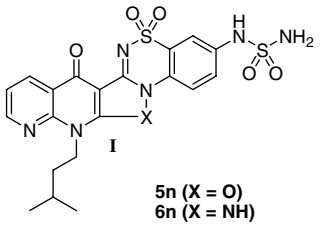
Compound **15n** was chosen for further modification owing to its excellent predicted activity (2.6828) whilst

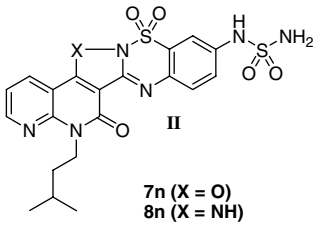
Table 10. Virtual screening, compounds **1n**–**4n**



ID	R	log(1/IC$_{50}$), predicted	Leverage-limit
1n	NH_2	1.3695	0.2336
2n	SH	1.3035	0.2266
3n	NHOH	1.7075	0.1824
4n	$NHNH_2$	1.5134	0.2129

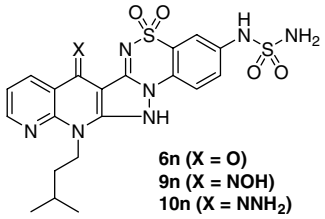
Table 11. Virtual screening, compounds **5n**–**8n**





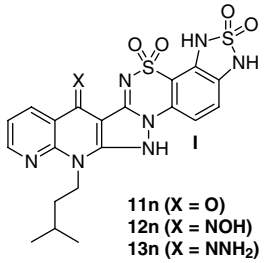
ID	X	log(1/IC_{50}), predicted	Leverage-limit
5n	O(I)	1.5777	−0.0943
6n	NH(I)	2.1518	0.0756
7n	O(II)	1.8152	−0.0723
8n	NH(II)	1.9369	−0.1293

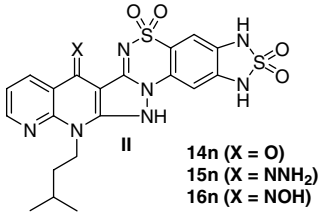
Table 12. Virtual screening, compounds **9n**–**10n**



ID	X	log(1/IC_{50}), predicted	Leverage-limit
6n	O	2.1518	0.0756
9n	NOH	2.9276	−0.1206
10n	NNH_2	2.9798	−0.2287

Table 13. Virtual screening, compounds **11n**–**16n**



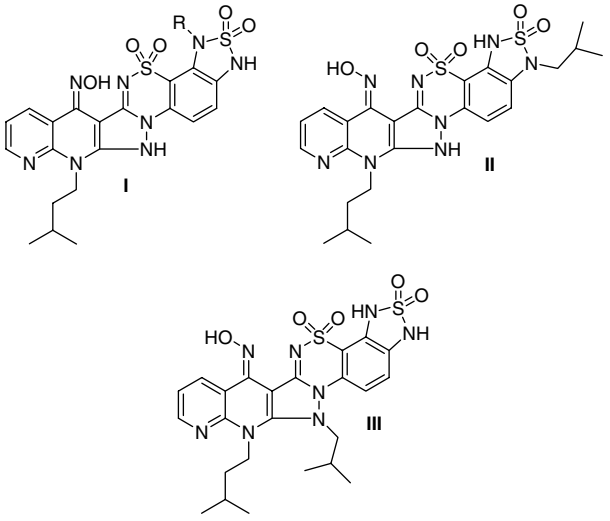


ID	X	log(1/IC_{50}), predicted	Leverage-limit
11n	O(I)	2.2774	0.0949
12n	NOH(I)	2.8525	−0.0970
13n	NNH_2(I)	2.9027	−0.1907
14n	O(II)	1.9378	0.1572
15n	NOH(II)	2.6828	−0.0482
16n	NNH_2(II)	2.7239	−0.1346

being only marginally outside of the domain of applicability (−0.0482). Alkyl groups were subsequently introduced on the thiadiazolidine ring nitrogens to observe their effect on both activity and domain of applicability (Table 14).

The introduction of *N*-alkyl groups improved the activity and the domain of applicability was superior for sec-Bu groups. Furthermore the activity and domain of applicability was not significantly dependent upon

Table 14. Virtual screening, compounds **17n**–**26n**

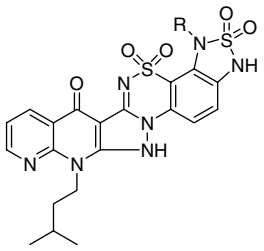


ID	R	log(1/IC_{50}), predicted	Leverage-limit
17n	Me	3.0681	−0.0619
18n	CF_3	3.1915	−0.0640
19n	Et	3.0033	−0.0192
20n	*n*-Pr	2.9792	0.0340
21n	*i*-Pr	3.0912	−0.0121
22n	Vinyl	2.7533	0.0221
23n	*n*-Bu	3.0327	0.0011
24n	s-Bu	2.9596	0.0796
25n	s-Bu(II)	2.9354	0.0567
26n	H(III)	2.4022	0.1751

which 1,2,5-thiadiazolidine ring nitrogen was alkylated. Introduction of the *N*-alkyl substituent on the fused pyrazole ring nitrogen led to a significantly improved domain of applicability but a reduced activity.

A similar improvement in the domain of applicability with concomitant reduction in activity was observed when the oxime functionality of compounds **17n**, **19n**–**21n**, **23n**, and **24n** was replaced by a simple carbonyl group (Table 15). However, even here the predicted activity is quite significantly improved compared to the original compd **95** and comfortably within the model's domain of applicability.

Table 15. Virtual screening, compounds **27n–32n**



ID	R	log(1/IC_{50}), predicted	Leverage-limit
27n	Me	2.6841	0.0689
28n	Et	2.5983	0.1090
29n	*n*-Pr	2.5952	0.1403
30n	*i*-Pr	2.6916	0.1179
31n	*n*-Bu	2.5794	0.1584
32n	s-Bu	2.4214	0.1571

The above virtual screening study has identified possible ring fusions which are tolerated by the model and maintain good predicted activities within the domain of applicability. The proposed structures promise to be rigid active HCV inhibitors derived from those chosen for the training set and could be expected to show improved specificity.

The virtual screening study has generated structures with a clear improvement in predicted biological activity within the domain of applicability. The results illustrate the utility of prediction protocols based on calculated descriptors, such as the ones found in Table 7. However, the descriptors of Table 7 only crudely reflect the receptor–ligand binding. A future extension of this work might be able to take advantage of the ready availability of receptor–ligand co-crystal structures relevant to this particular chemical series, which are available in the Protein Data Bank (PDB).[27] Additionally, a parallel 3-D QSAR model could be constructed using for example the Comparative Molecular Field Analysis (CoMFA) methodology.[28] In this way, the strategy presented in this paper will be a part of a multi-pronged approach, where several alternative computational methods will be used to identify possible leads, before investing in the synthesis of new structures.

Finally a cautionary note should be included dealing with the biological activity scales. While the data from the experimental and virtual studies have been recorded with the same units it must be noted that the predicted activities produced by the virtual model are significantly higher. It would be truly remarkable if the model was able to accurately predict such activities quantitatively but this is unlikely. The synthesis and study of these compounds would be required to truly validate the virtual model and as such is a worthy pursuit but this is outside the scope of this present paper. It must therefore be noted that the virtual screening study acts only as an aid in proposing structural modifications to assist ongoing SAR studies. The high biological activities predicted are only indicative of which structures should be targeted for synthesis on the basis that they meet or approach the optimal values for the chosen descriptors for the given model.

5. Conclusions

In the present study five descriptors [Lipophilicity (*C*log *P*), HOMO energy, Kier and Hall index order 2 (Ki2), and Kier and Hall information indices order 0 and 3 (KiInf0, KiInf3)] were found to be important for describing the inhibition activity against genotype 1 HCV polymerase. The five-descriptor set contains electronic, topological, and physicochemical information about molecules, and describes and models successfully the binding affinity of these small molecules.

The validation procedures utilized in this work (separation of data into independent training and validation sets, Y-randomization) illustrated the accuracy and robustness of the produced QSAR model not only by calculating its fitness on sets of training data, but also by testing the predictive ability of the model. The proposed method, due to the high predictive ability, offers a useful potential alternative to the costly and time-consuming experiments for determining HCV inhibition. Biological activities of novel compounds can be estimated by the produced MLR model. Furthermore, the produced QSAR model can be used to screen existing databases or virtual libraries in order to identify novel potent compounds. An attempt in this direction was carried out. Synthesis of the proposed chemistry driven small molecules using the aforementioned virtual screening procedure and experimental evaluation of their biological activity will show if the method can be used as a general rational drug discovery tool.

Acknowledgments

G. M. and A. A. would like to thank the Cyprus Research Promotion Foundation (grants no. KINHT/0505/03 and PLYPH/0506/25) for financial support.

References and notes

1. Ikegashira, K.; Oka, T.; Hirashima, S.; Noji, S.; Yamanaka, H.; Hara, Y.; Adachi, T.; Tsuruha, J.-I.; Doi, S.; Hase, Y.; Noguchi, T.; Ando, I.; Ogura, N.; Ikeda, S.; Hashimoto, H. *J. Med. Chem.* **2006**, *49*, 6950.
2. Gopalsamy, A.; Chopra, R.; Lim, K.; Ciszewski, G.; Shi, M.; Curran, K. J.; Sukits, S. F.; Svenson, K.; Bard, J.; Ellingboe, J. W.; Agarwal, A.; Krishnamurthy, G.; Howe, A. Y. M.; Orlowski, M.; Feld, B.; O'Connell, J.; Mansour, T. S. *J. Med. Chem.* **2006**, *49*, 3052.
3. Rönn, R.; Gossas, T.; Sabnis, Y. A.; Daoud, H.; Åkerblom, E.; Danielson, U. H.; Sandström, A. *Bioorg. Med. Chem.* **2007**, *15*, 4057.
4. Nittoli, T.; Curran, K.; Insaf, S.; DiGrandi, M.; Orlowski, M.; Chopra, R.; Agarwal, A.; Howe, A. Y. M.; Prashad, A.; Floyd, M. B.; Johnson, B.; Sutherland, A.; Wheless, K.; Feld, B.; O'Connell, J.; Mansour, T. S.; Bloom, J. *J. Med. Chem.* **2007**, *50*, 2108.

5. Prongay, A. J.; Guo, Z.; Yao, N.; Pichardo, J.; Fischmann, T.; Strickland, C.; Myers, J., Jr.; Weber, P. C.; Beyer, B. M.; Ingram, R.; Hong, Z.; Prosise, W. W.; Ramanathan, L.; Taremi, S. S.; Yarosh-Tomaine, T.; Zhang, R.; Senior, M.; Yang, R.-S.; Malcolm, B.; Arasappan, A.; Bennett, F.; Bogen, S. L.; Chen, K.; Jao, E.; Liu, Y.-T.; Lovey, R. G.; Saksena, A. K.; Venkatraman, S.; Girijavallabhan, V.; Njoroge, F. G.; Madison, V. *J. Med. Chem.* **2007**, *50*, 2310.
6. Rong, F.; Chow, S.; Yan, S.; Larson, G.; Hong, Z.; Wu, J. *Bioorg. Med. Chem.* **2007**, *17*, 1663.
7. Powers, J. P.; Piper, D. E.; Li, Y.; Mayorga, V.; Anzola, J.; Chen, J. M.; Jaen, J. C.; Lee, G.; Liu, J.; Peterson, M. G.; Tonn, G. R.; Ye, Q.; Walker, N. P. C.; Wang, Z. *J. Med. Chem.* **2006**, *49*, 1034.
8. Harper, S.; Pacini, B.; Avolio, S.; Di Filippo, M.; Migliaccio, G.; Laufer, R.; De Francesco, R.; Rowley, M.; Narjes, F. *J. Med. Chem.* **2005**, *48*, 1314.
9. Pratt, J. K.; Donner, P.; McDaniel, K. F.; Maring, C. J.; Kati, W. M.; Mo, H.; Middleton, T.; Liu, Y.; Ng, T.; Xie, Q.; Zhang, R.; Montgomery, D.; Molla, A.; Kempf, D. J.; Kohlbrenner, W. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1577.
10. Krueger, A. C.; Madigan, D. L.; Jiang, W. W.; Kati, W. M.; Liu, D.; Liu, Y.; Maring, C. J.; Masse, S.; McDaniel, K. F.; Middleton, T.; Mo, H.; Molla, A.; Montgomery, D.; Pratt, J. K.; Rockway, T. W.; Zhang, R.; Kempf, D. J. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3367.
11. Rockway, T. W.; Zhang, R.; Liu, D.; Betebenner, D. A.; McDaniel, K. F.; Pratt, J. K.; Beno, D.; Montgomery, D.; Jiang, W. W.; Masse, S.; Kati, W. M.; Middleton, T.; Molla, A.; Maring, C. J.; Kempf, D. J. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3833.
12. CambridgeSoft Corporation www.cambridgesoft.com.
13. ww.lohninger.com/topix.html.
14. Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137.
15. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69.
16. Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. *Chemometr. Intell. Lab. Syst.* **1996**, *33*, 35.
17. Todeschini, R.; Consonni, V.; Mannhold, R. (Series Ed.); Kubinyi, H. (Series Ed.); Timmerman, H. (Series Ed.) Handbook of Molecular Descriptors, Wiley-VCH: Weinheim, 2000.
18. (a) Efron, B. *J. Am. Stat. Assoc.* **1983**, *78*, 316; (b) Osten, D. W. *J. Chemom.* **1998**, *2*, 39.
19. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J.; Kohn, H.; Tropsha, A. *J. Med. Chem.* **2004**, *47*, 2356.
20. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Mod.* **2002**, *20*, 269.
21. Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; Van de Waterbeemd, H., Ed.; VCH Weinheim: Germany, 1995.
22. Atkinson, A. *Plots, Transformations and Regression*; Clarendon Press: Oxford (UK), 1985.
23. Kier, L. B.; Hall, L. B. *Molecular Connectivity in Structure Activity Analysis*; Wiley: Chichester, 1986.
24. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *J. Comput. Aided Design* **2006**, *20*, 83.
25. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *J. Comput. Aided Design* **2006**, *21*, 251.
26. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *Mol. Divers.* **2006**, *10*, 405.
27. Protein Data Bank, www.rcsb.org.
28. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.