Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 44 (2009) 25–34
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

Prediction of CCR5 receptor binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas based on the heuristic method, support vector machine and projection pursuit regression

Yongna Yuan [a], Ruisheng Zhang [a,b,*], Rongjing Hu [a,c], Xiaofang Ruan [a]

[a] Department of Chemistry, Lanzhou University, Lanzhou 730000, China
[b] Department of Computer Science, Lanzhou University, Lanzhou 730000, China
[c] Université Paris 7-Denis Diderot, ITODYS, 1 rue Guy de la Brosse, 75005 Paris, France

Received 26 July 2007; received in revised form 1 March 2008; accepted 6 March 2008
Available online 20 March 2008

Abstract

Quantitative structure–activity relationship (QSAR) models were developed to predict for CCR5 binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas using linear free energy relationship (LFER). Eight molecular descriptors selected by the heuristic method (HM) in CODESSA were used as inputs to perform multiple linear regression (MLR), support vector machine (SVM) and projection pursuit regression (PPR) studies. Compared with MLR model, the SVM and PPR models give better results with the predicted correlation coefficient (R^2) of 0.867 and 0.834 and the squared standard error (s^2) of 0.095 and 0.119 for the training set and R^2 of 0.732 and 0.726 and s^2 of 0.210 and 0.207 for the test set, respectively. It indicates that the SVM and PPR approaches are more adapted to the set of molecules we studied. In addition, methods used in this paper are simple, practical and effective for chemists to predict the human CCR5 chemokine receptor.

Keywords: QSAR; LFER; CCR5; HM; SVM; PPR

1. Introduction

Among infectious diseases, acquired immunodeficiency syndrome (AIDS) is the most fatal disorder for which no curative chemotherapy has been developed so far [1]. This infection targets cells of the immune system expressing the $CD4^+$ receptors and lead to defects in cell-mediated immunity. The replicative cycle of HIV-1, which is the causative agent of AIDS, can be divided into many steps [2,3]. Gp120-co-receptor (CCR5) interaction, which results in the exposure of a co-receptor-binding domain in gp120 on the cell surface, is one of the very important step. The chemokine receptor CCR5 is expressed on T-lymphocytes, monocytes, macrophages, dendritic cells, microglia and other cell types. After severe depletion of immuno-competent cells, the ultimate phase is the appearance of opportunistic infections, neurologic and neoplastic diseases, and ultimately death [4]. Over the last few years many groups have successfully engaged in the search for small molecules that interfere with the interaction between chemokines and their receptors [5]. In particular CCR5 has generated much interest as a drug target and a recent clinical trial has provided proof of concept for this approach in the treatment of HIV-1 infection [6]. Therefore, the inhibition of this key biochemical event in the viral life cycle provides the most attractive target for anti-HIV drug development. And in general, IC_{50} (the molar concentration of the drug required to inhibit 50% of the contraction of guinea pig ileum induced by methyl-furmethide) is used to evaluate the efficiency of a drug.

Among various anti-HIV activity screening, some important methods are cytoprotection assay, integration enzyme assay, RT inhibition assay, HIV attachment assay, fusion assay, etc. [7–10]. However, the experimental determination of such IC_{50} values is difficult, costly and time-consuming, and there are many uncertainties in chamber conditions [11]. For the

* Corresponding author. Department of Chemistry, Lanzhou University, Lanzhou 730000, China.
E-mail address: zhangrs@lzu.edu.cn (R. Zhang).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.03.004

reasons above, reliable theoretical models to estimate CCR5 binding affinity of the substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas are strongly required. Among them, quantitative structure property–activity relationship (QSPR/QSAR) represents an attempt to correlate structural descriptors of compounds with their physicochemical properties and biological activities. It is now widely used for the prediction of physicochemical properties and biological activities in chemical, environmental, and pharmaceutical areas [12,13]. The advantage of this approach over other methods lies in the fact that it requires only the knowledge of chemical structure and is not dependent on the experimental properties. This approach involves modeling a continuous activity for quantitative prediction of the activity of previously unseen compounds. The advances in QSAR studies have widened the scope of rationalizing drug design and the search for the mechanisms of drug actions [14–16]. In addition, they are useful in areas like design of virtual compound libraries, computational–chemical optimization of compounds, and design of combinatorial libraries with appropriate ADME (absorption, distribution, metabolism and excretion) properties.

In recent years, several binding affinity QSAR models have been published. Relatively few works concern the substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas for CCR5 binding affinity using linear free energy relationship (LFER). Leonard et al. developed several linear models to explore the structural and physicochemical requirements of 79 chemicals: stepwise regression, stepwise regression, multiple linear regression with factor analysis as the data preprocessing step (FA-MLR), partial least squares with factor analysis as the preprocessing step (FA-PLS), principal component regression analysis (PCRA), multiple linear regression with genetic function approximation (GFA-MLR), and genetic partial least squares (G/PLS). The equations obtained from stepwise regression, FA-MLR, FA-PLS, and PCRA were of acceptable statistical range (explained variance ranging from 71.9% to 80.4%, while predicted variance ranging from 67.4% to 77.0%) and were statistically validated using leave-one-out technique [17].

In addition to linear methods, a number of non-linear modeling methods, such as genetic algorithm, support vector machine and projection pursuit regression have been developed in the field of statistics to handle non-linearity exhibited in a given dataset. The support vector machine (SVM) has attracted attention and gained extensive applications in recent QSPR/QSAR analysis owing to its remarkable generalization performance. The projection pursuit regression (PPR) is another important nonparametric statistical technique, which seeks the ''interesting'' projections of data from high dimensional to lower dimensional space to try to find the intrinsic structure information hidden in the high dimensional data [18]. With the obtained interesting projections' direction, it can be used for further study of visual pattern recognition and regression (projection pursuit regression, PPR) [19,20]. Similar to SVM, it can also effectively overcome the curse of dimensionality. In previous work, Ren et al. have successfully applied PPR in the QSAR study of ion mobility [21]. In the present work, SVM and PPR are used to establish the quantitative relationship between molecular structure and CCR5 binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas using linear free energy relationship (LFER) for the same data used by Leonard et al. [17]. We use heuristic method to reduce the number of descriptors' space and to select the structural features of the molecules relevant to the CCR5 binding affinity. Finally, using the selected variables as inputs, QSAR models were constructed by MLR, SVM and PPR. The ultimate objective is to establish reliable QSAR models for CCR5 binding affinity prediction and to obtain knowledge of the binding affinity of substances not yet tested or for which reliable experimental data are not available as well.

2. Materials and methods

2.1. Datasets

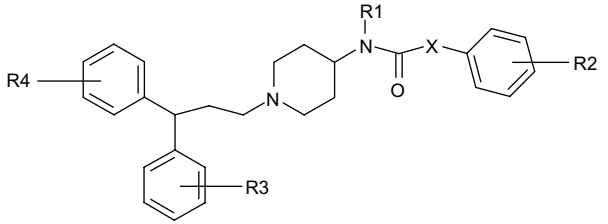
The CCR5 binding affinity data of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas reported by Thomas et al. [17] have been used as the model dataset for the present QSAR study: the affinity (50% inhibitory concentration) data [IC_{50} (lM) and IC_{50} (nM)] of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas (Table 1) for I-labeled RANTES (regulated on activation normal T-cell expressed and secreted) to Chinese hamster ovary (CHO) cells expressing human CCR5 have been converted to the logarithmic scale [pIC_{50} (mM)] and then used for subsequent QSAR analyses as the response variable. In order to build a general model, the rest of the 79 compounds were studied. The data set was randomly separated into a training set of 59 compounds and a test set of 20 compounds.

2.2. Descriptor calculation and feature selection

All structures were drawn and preoptimized using the MM+ molecular mechanics method within the framework of the Hyperchem program [22]. The preoptimized structures were submitted to the MOPAC 6.0 program [23] for further geometry refinement and for the calculation of molecular orbital parameters. The AM1 parametrizations were used to calculate the quantum-chemical molecular descriptors. The output files from MOPAC were transferred to the program CODESSA to calculate various descriptors. In CODESSA there are implemented procedures for the calculation of a large selection of molecular descriptors including a variety of constitutional, topological, geometric, and electrostatic descriptors [24].

Once molecular descriptors were generated, the heuristic method [24] in CODESSA was used to accomplish the preselection of the descriptors and build the linear model. Its advantages are the high speed and lack of software restrictions on the size of the data set. The heuristic method can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, descriptors with missing or constant values

Table 1
Structural features, experimental, and calculated values of LFER



Compound	X	R1	R2	R3	R4	pIC_{50}										
						Obsd[a]	Calcd[b]	Calcd[b]	Calcd[b]	Calcd[b]	HM[c]	Res.[d]	PPR[e]	Res.[d]	SVM[f]	Res.[d]
1[t]	–	CH_3	4-F	–	–	2.143	3.174	3.081	3.172	2.867	2.275	0.132	2.858	0.715	2.018	−0.125
2	–	CH_3	3-NO_2	–	–	2.292	2.488	2.815	2.158	2.309	2.68	0.388	2.662	0.33	2.203	−0.089
3	$CH(CH_3)$	CH_3	–	–	–	2.167	1.822	1.766	1.933	3.068	2.814	0.647	2.82	0.653	2.285	0.118
4	CH_2	CH_3	–	–	–	3.114	3.101	2.854	3.17	2.836	2.926	−0.188	2.952	−0.162	3.225	0.111
5[t]	CH_2	CH_3	2-Cl	–	–	2.444	3.101	2.743	2.665	2.379	2.900	0.456	2.758	0.314	2.842	0.398
6	CH_2	CH_3	3-Cl	–	–	2.658	2.781	2.739	2.894	2.655	2.796	0.138	2.657	0.001	2.687	0.001
7	CH_2	CH_3	4-Cl	–	–	3.097	3.383	3.409	3.174	3.194	2.975	−0.122	2.928	−0.169	3.149	0.052
8	CH_2	CH_3	3,4-Cl	–	–	3.108	3.064	3.023	2.899	3.142	2.992	−0.116	2.853	−0.255	2.997	−0.111
9[t]	CH_2	CH_3	2,4-Cl	–	–	2.585	3.383	2.975	2.670	2.776	2.932	0.347	2.709	0.124	2.798	0.213
10	CH_2	CH_3	2-F	–	–	2.721	3.101	2.853	2.665	2.488	2.761	0.04	2.869	0.148	2.832	0.111
11	CH_2	CH_3	3-F	–	–	2.854	2.807	2.854	2.937	2.602	3.008	0.154	2.883	0.029	3.103	0.249
12	CH_2	CH_3	4-F	–	–	3.18	3.174	3.119	3.172	3.002	2.760	−0.420	2.915	−0.265	2.878	−0.302
13[t]	CH_2	CH_3	3,4-F	–	–	3.161	2.881	3.105	2.939	2.823	2.811	−0.350	2.749	−0.412	2.795	−0.366
14	CH_2	CH_3	3-OMe	–	–	3.167	2.997	2.856	3.141	2.876	2.282	0.115	3.495	0.328	3.131	−0.036
15	CH_2	CH_3	4-OCH_3	_	_	3.327	2.769	3.110	3.106	3.008	3.203	−0.034	3.321	0.084	3.294	0.057
16	CH_2	CH_3	3,4-OMe	–	–	3.187	2.665	3.110	3.078	3.023	2.914	−0.273	2.802	−0.385	3.093	−0.094
17[t]	CH_2	CH_3	3,5-OMe	–	–	2.569	2.894	2.854	3.053	2.803	3.325	0.756	2.893	0.324	3.412	0.843
18	CH_2	CH_3	2,4,5-OMe	–	–	2.959	2.665	3.113	2.574	2.721	3.269	0.310	2.868	−0.091	3.070	0.741
19	CH_2	CH_3	4-Br	–	–	3.237	3.383	3.477	3.144	3.294	3.129	−0.108	2.888	−0.349	3.141	−0.096
20	CH_2	CH_3	4-Benzyloxy	–	–	2.456	2.818	2.351	2.771	2.547	2.578	0.122	2.662	0.206	2.567	0.111
21t	CH_2	CH_3	4-Phenyl	–	–	2.638	3.088	2.853	2.771	2.778	2.532	−0.106	2.568	−0.070	3.303	0.392
22	CH_2	CH_3	4-CF_3	–	–	3.432	3.765	3.599	3.471	3.585	3.916	0.484	3.597	0.165	3.764	0.443
23	CH_2	CH_3	4-OCF_3	–	–	3.538	3.531	3.424	3.218	3.231	3.720	0.182	3.733	0.195	3.509	−0.029
24	CH_2	CH_3	4-NHCOMe	–	–	3.167	3.101	3.500	3.365	3.093	3.151	−0.016	3.140	−0.027	3.056	0.111
25[t]	CH_2	CH_3	4-CN	–	–	4.222	3.912	4.051	3.931	3.770	3.949	−0.273	4.294	0.072	4.010	−0.212
26	CH_2	CH_3	4-SO_2NH_2	–	–	4.041	3.838	3.547	4.095	4.077	4.074	0.033	4.091	0.050	3.930	−0.111
27	CH_2	CH_3	4-SO_2N(Me)2	–	–	4.337	3.900	3.889	3.959	4.036	4.143	−0.194	3.970	−0.367	4.448	0.111
28	CH_2	CH_3	4-SMe	–	–	3.252	3.101	3.426	3.047	3.154	4.321	0.069	0.324	0.072	0.286	0.034
29[t]	CH_2	CH_3	4-CO_2Me	–	–	3.201	3.654	3.939	3.539	3.361	4.180	0.320	3.795	0.594	3.281	0.080
30	CH_2	CH_3	4-OH	–	–	3.328	2.646	2.698	3.241	3.053	3.051	−0.277	3.367	0.039	3.217	−0.111
31	CH_2	CH_3	4-NO_2	–	–	3.824	4.060	4.050	4.063	3.821	3.486	−0.338	3.419	−0.405	3.731	−0.111
32	CH_2	Ethyl	4-OCF_3	–	–	3.509	3.531	3.318	3.218	3.430	3.978	0.469	3.845	0.336	3.620	0.111
33[t]	CH_2	Ethyl	4-CN	–	–	4.180	3.912	4.097	3.931	3.954	4.334	0.154	4.788	0.608	4.132	−0.048
34	CH_2	Ethyl	4-SO_2NH_2	–	–	4.420	3.838	3.983	4.095	4.354	4.354	−0.066	4.508	0.088	4.309	−0.111
35	CH_2	Ethyl	4-SO_2Me			4.119	3.986	4.138	4.236	4.326	4.504	0.385	4.679	0.560	4.513	0.394
36	CH_2	Ethyl	4-NO_2	–	–	3.959	4.06	4.055	4.063	4.054	4.137	0.178	3.657	−0.302	4.070	0.111
37[t]	CH_2	c-propyl	4-SO_2NH_2	–	–	4.481	3.838	3.903	4.095	4.163	4.161	−0.320	4.508	0.027	3.792	−0.689

(*continued on next page*)

Table 1 (*continued*)

Compound	X	R1	R2	R3	R4	pIC_{50}										
						Obsd[a]	Calcd[b]	Calcd[b]	Calcd[b]	Calcd[b]	HM[c]	Res.[d]	PPR[e]	Res.[d]	SVM[f]	Res.[d]
38	CH_2	c-propyl	4-SO_2Me	–	–	4.292	3.986	4.106	4.236	4.164	3.723	−0.569	4.216	−0.076	4.181	−0.111
39	CH_2	c-propyl	4-NO_2	–	–	3.509	4.060	3.969	4.063	3.689	3.488	−0.021	3.687	0.178	3.620	0.111
40	CH_2	Allyl	4-OCF_3	–	–	3.456	3.531	3.188	3.218	3.247	3.794	0.338	3.706	0.250	3.510	0.054
41[t]	CH_2	Allyl	4-SO_2Me	–	–	4.432	3.986	4.138	4.236	4.248	4.350	−0.082	4.108	−0.324	4.468	0.036
42	CH_2	Allyl	4-NO_2	–	–	3.745	4.060	3.939	4.063	3.813	3.456	−0.289	3.239	−0.506	3.634	−0.111
43	NH	CH_3	3-CN	–	–	2.310	2.617	2.757	2.539	2.517	2.250	−0.061	2.534	0.224	2.421	0.111
44	NH	CH_3	3-CH_3	–	–	2.229	3.161	2.850	3.159	2.952	3.080	0.850	2.825	0.596	2.977	0.748
45[t]	NH	Ally	–	–	–	2.678	3.101	2.826	3.170	2.980	2.839	0.161	2.861	0.183	2.793	0.115
46	NH	CH_3	3,4-Cl	–	–	3.432	3.064	2.83	2.899	3.449	2.512	−0.920	2.703	−0.729	3.321	−0.111
47	NH	CH_3	4-F	–	–	3.721	3.174	3.056	3.172	3.359	3.452	−0.269	3.64	−0.081	3.610	−0.111
48	NH	Ethyl	4-CH_3	–	–	3.495	2.891	2.865	3.001	3.640	3.292	−0.203	3.329	0.166	2.998	−0.497
50	NH−CH_2	Ethyl	–	–	–	4.208	3.892	3.712	3.836	3.759	3.995	−0.213	4.008	−0.200	4.01	−0.198
51	NH-CH(CH_3)	Ethyl	–	–	–	2.268	2.613	2.669	2.599	3.577	2.430	0.162	2.916	0.648	2.379	0.111
52	NH-CH_2	Allyl	3-CH_3	–	–	3.620	3.952	3.711	3.826	3.533	3.685	0.065	3.811	0.191	3.594	−0.026
53[t]	NH−CH_2	Allyl	4-OCH_3	–	–	3.959	3.56	3.983	3.773	3.656	3.860	−0.099	3.564	−0.395	3.936	−0.023
54	NH−CH_2	Ethyl	3-CH_3	–	–	4.456	3.952	3.734	3.826	3.958	4.273	−0.183	4.447	−0.009	4.345	−0.111
55	NH−CH_2	Ethyl	4-OCH_3	–	–	3.377	3.56	3.977	3.773	3.67	4.234	0.857	3.652	0.275	3.487	0.110
56	NH−CH_2	Ethyl	4-SO_2CH_3	–	–	4.310	4.777	4.836	4.903	4.808	4.143	−0.167	4.355	0.045	4.199	−0.111
57[t]	CH_2	Ethyl	–	4-F	4-F	3.108	2.813	2.876	2.905	2.827	4.272	1.164	4.052	0.944	4.19	1.082
58	CH_2	Ethyl	–	4-F	–	3.509	4.019	4.101	4.353	4.342	3.742	0.233	3.956	0.447	3.619	0.11
59	CH_2	Ethyl	–	4-Cl	–	5.071	4.621	4.819	4.558	5.007	4.06	−1.011	4.184	−0.887	4.214	−0.857
60	CH_2	Ethyl	–	4-Cl	4-Cl	3.119	3.414	3.351	3.11	3.044	4.638	1.52	3.818	0.699	4.542	1.423
61[t]	CH_2	Ethyl	–	3-Cl	–	4.237	3.986	4.066	4.236	4.019	3.97	−0.267	3.94	−0.297	3.844	−0.393
62	CH_2	Ethyl	–	3,4-Cl_2	–	4.108	4.621	4.587	4.558	4.401	4.32	0.212	4.337	0.229	4.219	0.111
63	CH_2	Ethyl	–	4-CH_3	–	4.553	4.295	4.506	4.304	4.517	3.98	−0.574	4.852	0.299	4.267	0.286
64	CH_2	Ethyl	–	4-CF3	–	5.638	4.77	5.036	5.098	5.522	5.192	−0.446	5.573	−0.065	5.527	−0.111
65t	CH_2	Ethyl	–	4-CO_2CH_3	–	5.149	5.132	5.39	5.097	5.101	4.483	−0.666	4.377	−0.772	5.518	0.369
66	CH_2	Ethyl	–	4-$CONH_2$	–	3.585	4.951	4.158	5.004	4.584	4.134	0.549	3.578	−0.007	3.696	0.111
67	CH_2	Ethyl	–	4-OCH_3	–	5.201	4.382	4.629	4.472	4.543	4.436	−0.765	4.704	−0.497	5.09	−0.111
68	CH_2	Ethyl	–	4-Ph	–	4.071	4.562	4.277	4.129	4.2	3.948	−0.123	3.691	−0.38	4.014	−0.057
69t	CH_2	Ethyl	–	4-SCH_3	–	4.921	4.821	5.141	4.671	4.774	4.959	0.326	4.509	−0.412	5.215	0.294
70	CH_2	Ethyl	–	4-SO_2CH_3	–	5.77	5.345	5.3	5.296	5.471	5.292	−0.478	5.592	−0.178	5.66	−0.11
71	CH_2	Ethyl	–	4-NH_2	–	3.699	3.918	3.964	3.853	3.679	4.162	0.462	4.082	0.383	4.328	0.539
72	CH_2	Ethyl	–	4-$NHCOCH_3$	–	4.585	4.838	4.776	4.727	4.477	4.61	0.026	4.694	0.109	4.696	0.111
73[t]	CH_2	Ethyl	–	4-NHCOPh	–	3.886	3.548	3.61	3.679	3.681	3.979	0.093	3.663	−0.223	3.477	−0.409
74	CH_2	Ethyl	–	4-$NHSO_2CH_3$	–	4.745	4.861	4.814	4.916	4.552	5.052	0.307	5.003	0.258	4.856	0.111
75	CH_2	Ethyl	–	4-$NHSO_2Ph$	–	3.131	3.233	3.295	3.604	3.414	3.341	0.21	2.671	−0.46	3.02	−0.111
76	CH_2	Ethyl		4-Cl	3-F	4.721	4.621	4.89	4.558	5.205	4.434	−0.287	4.003	−0.718	4.535	−0.186
77[t]	CH_2	Ethyl	–	4-NH_2	3-F	3.979	3.918	3.756	3.853	4.05	4.475	0.496	4.311	0.332	4.35	0.371
78	CH_2	Ethyl	–	4-$NHCOCH_3$	3-F	5.102	4.838	4.816	4.727	4.894	4.836	−0.266	40840	−0.262	4.965	−0.137
79	CH_2	Ethyl	–	4-$NHSO_2CH_3$	3-F	5.041	4.861	4.854	4.916	4.908	4.533	−0.508	4.922	−0.049	4.93	−0.111

t – Test set.

[a] Experiment pIC_{50}.

[b] Ref. [13].

[c] Predicted by HM.

[d] Relative error of (calculated – experimental).

[e] Predicted by PPR.

[f] Predicted by SVM.

were discarded. This information will be helpful in reducing the number of descriptors involved in the search for the best QSAR/QSPR model.

The heuristic method usually produces correlations 2–5 times faster than other methods with comparable quality [25]. The rapidity of calculations from the heuristic method renders it the first method of choice in practical research. Thus, in the present investigation, we used this method to select structural descriptors and build the linear model.

2.3. Support vector machine (SVM)

A detailed description about the theory of SVM can be found in a tutorial [26]. Here, we give only a brief description about it. In support vector machine, the input data is first mapped into high dimensional feature space by the use of kernel function and then linear regression is performed in the feature space. The non-linear feature mapping will allow the treatment of non-linear problems in a linear space. After training on the training set data, SVM can be used to predict the objects whose values are unknown. The prediction or approximation function used by SVM is:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{1}$$

where α_i is some real value, x_i is a feature vector corresponding to a training object. The components of vector α and the constant b represent the hypothesis and are optimized during training. $K(x, x_i)$ is a kernel function. Training points with nonzero weight α_i are called the support vectors. The elegance of using a kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly and it may be useful to think of the kernel, $K(x, x_i)$ as comparing patterns or as evaluating the proximity of objects in their feature space. Thus, a test point is evaluated by comparing it with all training points. In the function estimation problems, the Gaussian radial basis function kernel is most commonly used because of its effectiveness and speed in training process. The form of the Gaussian function in R is

$$K(u, v) = \exp\left(-\gamma^* |u - v|^2\right) \tag{2}$$

where γ is the parameter of the kernel, u and v are two independent variables.

2.4. Projection pursuit regression

Projection pursuit (PP) is a powerful tool for seeking interesting projections of high dimensional data in one, two or three dimensions. It can overcome the curse of dimensionality. At present, projection pursuit has been applied in density estimation, classification and regression problems [27–29]. In this investigation, we mainly apply this technique to solve a regression problem. So, we introduce only the principle of projection pursuit regression (PPR).

For many practical problems, the data is usually high dimensional. It has been a common practice to use lower dimensional linear projections of the data for visual inspection. The lower dimension is usually 1 or 2 (or may be 3). More precisely, $X_1, \ldots, X_n$, $X \in IR^p$ are p-dimensional data, then a k-dimensional $(k < p)$ linear projections is $Z_1, \ldots, Z_n$, $Z \in IR^k$ where $Z_i = \alpha^T X_i$ for some $p \times k$ matrix α such that $\alpha^T \alpha = I_k$, the k-dimensional identity matrix. Such a matrix α is often called orthonormal. Since there are infinitely many projections from a higher dimension to a lower dimension, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. Friedman and Stuetzle successfully implemented the idea combining both projection and pursuit, which is called projection pursuit (PP) [27].

The two basic elements of projection pursuit are: a PP index and a PP algorithm. A PP index, $I(\alpha)$, is a measure of how interesting a projection by α is, where $I(\alpha) = I(\alpha|X)$ implicitly depends on the data $X = (X_1, \ldots, X_n)$. The larger the index value, the more interesting the projection is. A PP algorithm is a numerical optimization algorithm which maximizes the index over all possible α. In the implementation of PP by Friedman and Turkey [27], the first several maxima found by a PP algorithm provide the most interesting projections.

In a typical regression problem, (X, Y) is an observable pair of random variables from a distribution F, where $X \in IR^p$ is a p-dimensional variable (called predictor) and $Y \in IR$ is a response; and the goal is to estimate the regression function.

$$f(x) = E(Y|X = x)$$

i.e. the conditional expectation of Y given $X = x$, using a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from F PPR approximates the regression function $f(x)$ by a finite sum of ridge functions

$$g^{(m)}(x) = \sum_{i=1}^{m} g_i\left(\alpha_i^T x\right) \tag{3}$$

where α_i are $p \times k$ orthonormal matrices, m is the number of ridge functions. PPR model can be used to approximate a large class of function by suitable choices of α_i and g_i.

In this investigation, the PPR algorithm proposed by Friedman was used to construct a PPR model. In this algorithm, g_i are found by smoothing operation that entails a backfitting. Specially, given $g^{(0)} = 0$, for $i \geq 1$, it iteratively estimates α_i by maximum of an index and g_i by a low dimensional nonparametric regression estimate based on the projected data (z_j, r_j), where $r_j = Y_j - g^{(i-1)}(X_j)$ are the residuals at the ith step and $z_j = \alpha_i^T X_j, j = 1, \ldots, n$. The procedure is repeated forward (and perhaps a backward fitting is allowed to adjust for the previous fitted pair) until the residual sum of squares $\sum r_j^2$ is less than a predetermined values. A different smoother for g_i, or index, or fitting order may be used and hence yields a different PPR algorithm.

In the present work, we used Friedman's super smoother as a smoother. Friedman's super smoother is a running lines smoother which choose among three spans for the

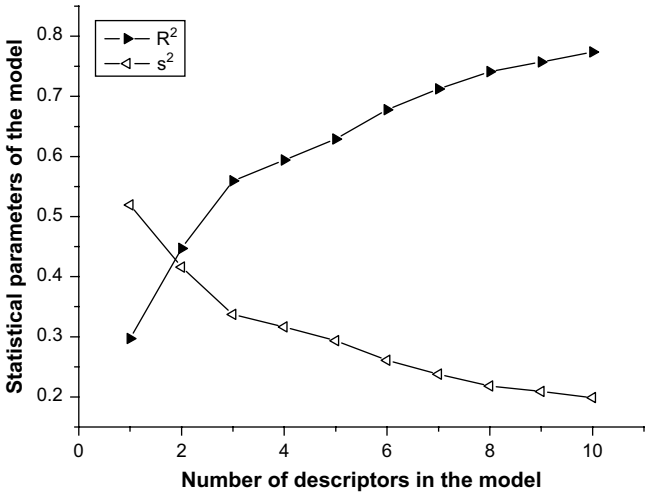


Fig. 1. Influence of the number of descriptors on the correlation coefficient (R^2) and the squared standard error (s^2) of the regression models.

lines: 0.5*n, 0.2*n and 0.05*n, where n is the number of data points and ''span'' defines the fraction of the observations in the span of the running lines smoother. If ''span'' is specified, a single smoother with span ''span*n'' is used. PPR algorithm was performed using R script.

3. Results and discussion

3.1. The HM and MLR models

A total of 632 descriptors were calculated by the CODESSA program for each compound. After the heuristic reduction, the pool of descriptors was reduced to 183. To select the set of descriptors that are most relevant to the mobility of compounds, the linear models with the number of variables from 1 to 10 were built. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size was achieved. The influences of the number of the descriptors on the correlation coefficient (R^2) and the squared standard error (s^2) are shown in Fig. 1. From Fig. 1, it can be seen that eight descriptors appear to be sufficient for a successful regression model. Thus, the 8-parameter model was chosen as the best linear model, which is given in Table 2.

Using these parameters, an MLR model was built as shown in Table 2. Multi-collinearity between the above four descriptors was detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$\mathrm{VIF} = \frac{1}{1 - r^2} \tag{4}$$

where r is the correlation coefficient of multiple regression between one variable and the others in the model. If VIF equals to 1.0, no intercorrelation exists for each variable; if VIF falls into the range of 1.0–5.0, the related model is acceptable; and if VIF is larger than 10.0, the related model is unstable and recheck is necessary [30]. The corresponding VIF values of the seven descriptors are shown in Table 2. As can be seen from this table, most variables have VIF values less than 5, indicating that the obtained model has obvious statistic significance.

By interpreting the descriptors selected by HM in the regression model, it is possible to gain some insight into factors that are likely to govern the LFER of the compounds and understand which factors play an important role for the human CCR5 chemokine receptor. Of the selected eight descriptors, one constitutional (relative number of benzene rings), one geometrical (XY Shadow/XY Rectangle), one topological (Balaban index), and the rest five are quantum-chemicals (Principal moment of inertia A/# of atoms, Avg bond order of a C atom, Avg bond order of a N atom, Min e–n attraction for a C–H bond and Min atomic orbital electronic population). These descriptors encode different aspects of the molecular structure information and are highly informative of different aspects of the studied CCR5 binding affinity.

The constitutional descriptors include the number of benzene rings (RNBR), it accounts for the chain stiffness and steric hindrance effect of compounds. The number of benzene rings encode the size of the compounds, thus, an increase in this descriptor strengthens the hydrodynamic friction of the molecule between the solute and the solvent, and then disfavoring the CCR5 binding affinity. RNBR receives a negative coefficient in the regression, and this indicates that increasing the value of this descriptor leads to a low linear free energy relationship (LFER).

The geometrical descriptors describe the size and shape of the molecules, the only descriptor contained in the model that belongs to this group is the XY Shadow/XY Rectangle

Table 2
The linear model based on the eight parameters selected by heuristic method

Chemical meaning	Descriptor	Coefficient	Standard error	t value	VIF
Intercept	Constant	−0.316	13.833	−0.023	
Principal moment of inertia A/# of atoms	PMIAA	−12,735.000	3255.346	−3.917	1.743
Balaban index	BI	4.899	1.435	3.413	1.436
Avg bond order of a C atom	BOAvg(C)	−38.62	9.568	−4.04	1.223
Relative number of benzene rings	RNBR	−71.851	30.325	−2.368	1.81
XY Shadow/XY Rectangle	XYS/XYR	−6.314	2.110	−2.995	1.591
Avg bond order of a N atom	BOAvg(N)	7.267	1.810	4.02	1.707
Min e–n attraction for a C–H bond	$E_{\min}$(C–H)	0.6717	0.175	3.852	1.674
Min atomic orbital electronic population	MAOEP	−10.704	3.171	−3.167	2.043

$n = 59$, $R^2 = 0.741$, $F = 17.88$, $s^2 = 0.218$; 95% confidence interval. VIF: variation inflation factors.

(XYS/XYR). It receives a negative coefficient in the regression; this indicates that the CCR5 binding affinity decreases with the increasing of the XY Shadow.

The Balaban index (BI), a topological descriptor, describes the atomic connectivity and branching information in the molecule and has some correlation with the hydrophobic interaction of the molecules. Because of its positive coefficient in the linear model, increasing this descriptor also increases the calculated LFER values, indicating that the large degree of branching for molecules is in favor of the CCR5 binding affinity. This echoes the importance of hydrophobicity in binding.

Quantum-chemical descriptors include the principal moment of inertia A/# of atoms (PMIAA), Avg bond order of a C atom (BOAvg(C)), Avg bond order of a N atom (BOAvg(N)), Min e−n attraction for a C−H bond (E_{min} (C−H)) and the Min atomic orbital electronic population (MAOEP). IA, the principal moment of inertia of the molecules around the *x*-axis, is particularly useful for distinguishing between the isomers [31]. Here, the descriptor principal moment of inertia A (PMIA) belongs to quantum-chemical descriptors, which is a measure of the mass distribution in the molecule. BOAvg(C) and BOAvg(N) are related to the strength of intramolecular bonding interactions and characterize the stability of the molecules, their conformational flexibility, and other valency-related properties [32]. MAOEP for a given atomic species in the molecule is a simplified index to describe the electrophilic ability of the molecule. E_{min}(C−H) is the minimum nuclear attraction energy for a C−H bond. Because the main weight atoms are C atoms, this energy describes the nuclear attraction driven processes in the molecule and may be related to the conformational changes or atomic reactivity in the molecule. These five descriptors indicate the importance of the intramolecular electronic effects and reactivity of a molecule in determining the binding ability between the molecule and receptor. The influence of individual descriptors on the CCR5 binding affinity is different from each other and can be reflected by the standardized regression coefficients. As can be seen from Table 2, these five descriptors play a major role in the binding process. The negative coefficient of any descriptor in the regression function means increasing this descriptor will decrease the pIC_{50} values and therefore increase the CCR5 binding ability. On the contrary, an increase of the descriptor with positive coefficient would lead to a decrease of the binding ability to CCR5 chemokine receptor.

From the above discussion, it can be seen that the above descriptors can account for the factors influencing the linear free energy relationship (LFER), i.e., the molecular size, structure, inertia of atoms and the bond order of C/N atom. According to the *t*-test in Table 2, BOAvg(C) is the most relevant descriptor, and this indicates that the bond order of a C atom plays a prevailing influence on the linear free energy relationship.

In QSAR study, generally, the quality of a model is expressed by its fitting ability and prediction ability, and the prediction ability is more important. Table 2 gave the obtained linear model for the training set based on the selected parameters by using multiplier linear regression (MLR) method. With the test set, the prediction results were obtained,

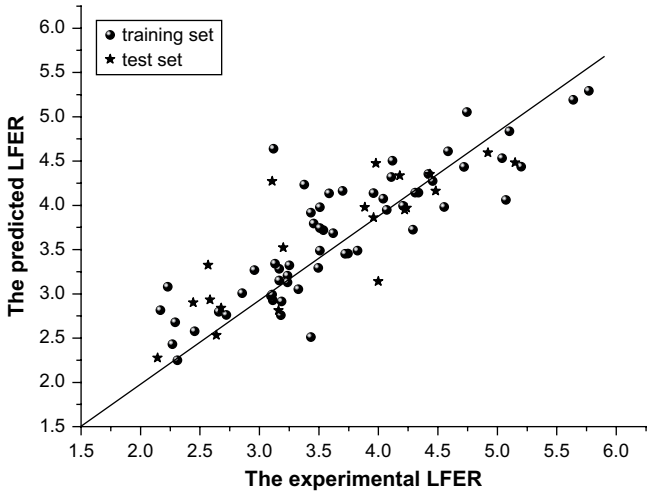


Fig. 2. Predicted vs experimental linear free energy relationship by HM.

confirming the predictive capability of the model. The statistical parameters for the training set were $R^2 = 0.741$; $F = 17.88$; $n = 59$; $s^2 = 0.218$. Fig. 2 shows a plot of the calculated vs experimental LFER for the training set and the test set.

3.2. The SVM model

The quality of SVM for regression depends on several parameters: kernel type *k*, which determines the sample distribution in the mapping space, and its corresponding parameters γ, capacity parameter *C*, ε, and ε-insensitive loss function. The three parameters were optimized in a systematic grid search way and the final optimal model was determined as $C = 100$, $\gamma = 0.02$, $\varepsilon = 0.13$, and $SVs = 47$.

Using this model, very satisfactory results were obtained. The results are shown in Table 3 and Fig. 3. With this model, R^2 for the training set was increased to 0.867 and s^2 was reduced to 0.095, respectively. For the test set, R^2 was increased to 0.732, while s^2 was reduced to 0.210, respectively, showing

Table 3
Comparison of the statistical parameters by different QSAR models for the prediction of the linear free energy relationship (LFER)

Methods	Number of descriptors	Data set	R^2	s^2
Our work				
HM	8-Descriptor	Training	0.741	0.218
		Test	0.715	0.238
PPR	8-Descriptor	Training	0.837	0.119
		Test	0.726	0.207
SVM	8-Descriptor	Training	0.867	0.095
		Test	0.732	0.210
Ref. [17]				
Stepwise regression	8-Descriptor	Training	0.747	0.207
		Test	–	–
FA-MLR	11-Descriptor	Training	0.810	0.162
		Test	0.725	–
FA-PLS	13-Descriptor	Training	0.797	0.191
		Test	–	–
PCRA	12-Descriptor	Training	0.834	0.145
		Test	–	–

the good generalization ability of the SVM model. The smaller scatter of data points in Fig. 3 demonstrates that the SVM model is clearly superior both in fitness and in prediction performance.

3.3. The PPR model

With the selected eight descriptors as input, PPR was also applied to build the non-linear model. In this investigation, all calculation programs implementing projection pursuit regression were written in R-file based on R script. For projection pursuit regression, there are several parameters ''nterms'', ''optlevel'' and ''span'' need to be determined. The parameter ''nterms'' controls the number of terms to be included in the final model. The levels of optimization (argument 'optlevel') differ in how thoroughly the models are refitted during this process. At level 0 the existing ridge terms are not refitted. At level 1 the projection directions are not refitted, but the ridge functions and the regression coefficients are. Levels 2 and 3 refit all the terms and are equivalent for one response; level 3 is more careful to re-balance the contributions from each regressor at each step and so is a little less likely to converge to a saddle point of the sum of squares criterion. Span defined the fraction of the observations in the span of the running lines smoother. In this investigation, the three parameters ''nterms'', ''optlevel'' and ''span'' were determined as 8, 3 and 0.17, respectively. The predicted results of the optimal PPR model are shown in Fig. 4. The model gave a squared standard error (s^2) of 0.119 for the training set, 0.207 for the prediction set and the corresponding correlation coefficients (R^2) were 0.837, 0.726, respectively.

Compared with the SVM and PPR models developed above, the improvement of the above goodness-of-fit parameters on those of MLR model indicates that the training set is described more accurately by these descriptors and the SVM and PPR models are expected to be better predictors for CCR5 binding ability than MLR model. It indicated the good generalization capability of the SVM and PPR models.

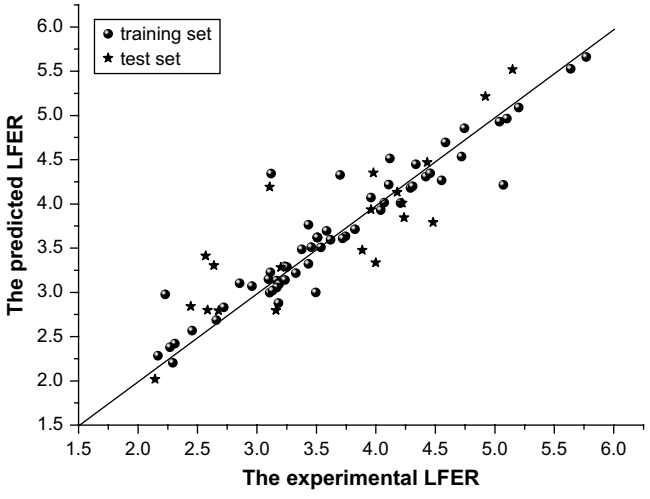


Fig. 3. Predicted vs experimental linear free energy relationship by SVM.

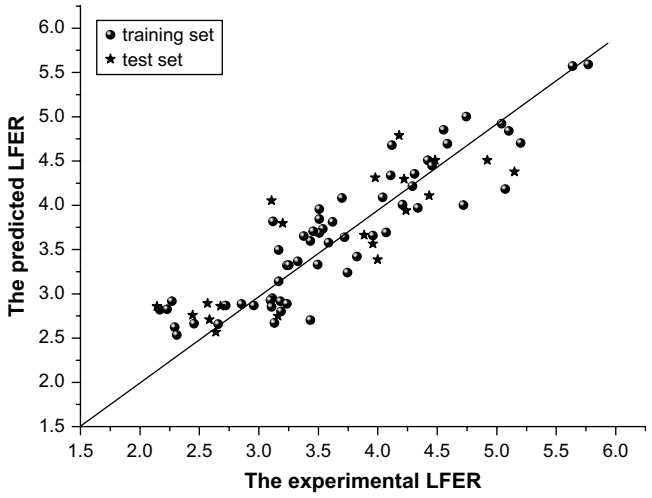


Fig. 4. Predicted vs experimental linear free energy relationship by PPR.

This is also demonstrated in Fig. 5, in which we compared the predicted results for the test set alone by three models, HM, SVM and PPR. Clearly, the results by the non-linear models, SVM and PPR, show relatively smaller bias than those by the MLR model.

Fig. 6 is a further comparison of the results by HM, SVM and PPR models, which plots the number of compounds as a function of the absolute deviation of the linear free energy relationship (LFER). The plot shows that, for MLR model, it can correctly predict rate constants for 72.88% and 70% compounds in the training and test sets, respectively, within an absolute error of 0.4 unit; for SVM model, the corresponding proportions were 77.97% and 60%, respectively; for PPR model, the respective proportions were 88.14% and 75%. Within an absolute error of 0.8 unit, the SVM model can give corresponding respective accuracy as 98.30% and 95%; the PPR model can correctly predict 96.61% and 90% of the compounds for the training and test sets, respectively; whereas the MLR model can only predict 93.22% and 90% of compounds in the respective data set.

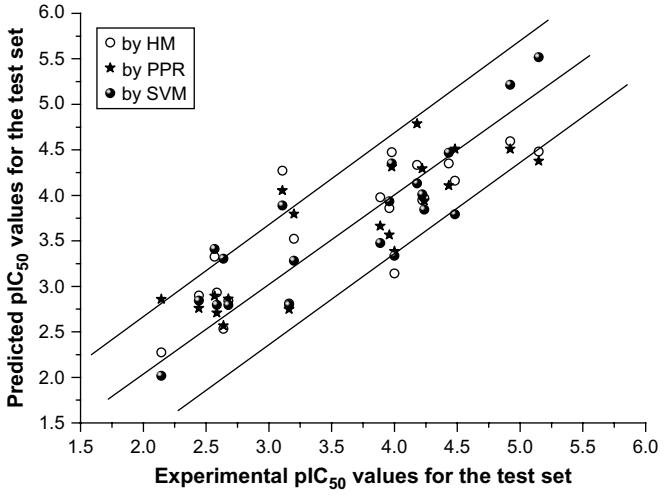


Fig. 5. Comparison of the predicted retention time for the test set by HM, SVM and PPR.

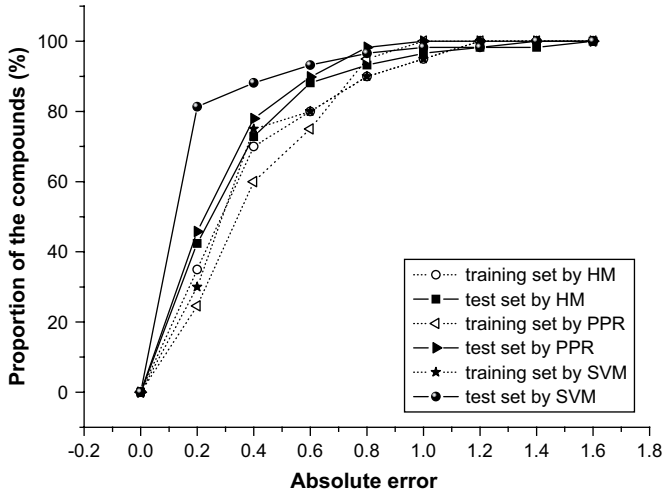


Fig. 6. Proportion of compounds within a given deviation from the experimental linear free energy relationship by HM, SVM and PPR.

3.4. Comparison of the results obtained by different approaches

From Table 3, we can see that our results are much better than the previous results even with a number of descriptors less than the previous models. The method of Ref. [17] is Hansch analysis with descriptors of electronic (Hammett σ), hydrophobicity (π), and steric (molar refractivity and STERIMOL L, B1, and B5) of phenyl ring substituents. The descriptors calculated by CODESSA are various of parameters which can describe the structures. In addition, to test the suitability of non-linear the models, SVM and PPR, we have compared the obtained predicted results by SVM and PPR with MLR and those obtained in Ref. [17]. Table 3 shows the statistical parameters of the results obtained from the three studies for the same set of compounds. Thus, from Table 3, it can be seen that the SVM and PPR models give better results.

4. Conclusions

In the present work, we built linear and non-linear models to predict the CCR5 binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas using linear free energy relationship (LFER) bases on MLR, SVM and PPR methods. HM was used to select descriptors responsible for bioactivity of inhibitors and develop the linear model. We can conclude that the non-linear SVM and PPR models produced more satisfactory results than the MLR model with good predictive ability. When comparing with previous research, our non-linear models with descriptors from CODESSA are much better than the previous model with descriptors of Hansch analysis. We can conclude that (1) better relationship between the structures and their activity can be obtained using descriptors from CODESSA; (2) the proposed models might identify and provide some insight into what structural features are related to the CCR5 binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas; (3) non-linear models using SVM and PPR produced better models with good predictive ability than heuristic method. Non-linear relationships can describe accurately the relationship between the structural parameter and the LFER of the studied compounds. Additionally, this paper provided three simple, practical and effective methods for analytical chemists to predict the human CCR5 chemokine receptor.

Acknowledgements

The authors thank the R Development Core Team for affording the free R1.7.1 software. The authors thank the National Natural Science Foundation of China (NSFC) Fund (No. 90612016) for supporting this project.

References

[1] L. Douali, D. Villemin, A. Zyad, D. Cherqaoui, Mol. Div. 8 (2004) 1–8.
[2] S. Jiang, Q. Zhao, A.K. Debanth, Curr. Pharm. Des. 8 (2002) 563–580.
[3] R.W. Sanders, M.M. Dankers, E. Busser, M. Caffrey, J.P. Moore, B. Berkhout, Retrovirology 1 (2004) 3–13.
[4] G. Campiani, A. Ramunno, G. Maga, V. Nacci, C. Fattorusso, B. Catalanotti, E. Morelli, E. Novellino, Curr. Pharm. Des. 8 (2002) 615–657.
[5] For reviews, see: (a) Z. Gao, W.A. Metz, Chem. Rev. 103 (2003) 3733–3752;
(b) J.J. Onuffer, R. Horuk, Trends Pharmacol. Sci. 23 (2002) 459–467.
[6] A. Palani, S. Shapiro, J.W. Clader, W.J. Greenlee, D. Blythin, K. Cox, N.E. Wagner, J. Strizki, B.M. Baroudy, N. Dan, Bioorg. Med. Chem. Lett. 13 (2003) 705–708.
[7] G. Xu, A. Kannan, T.L. Hartman, H. Wargo, K. Watson, J.A. Turpin, R.W. Buckheit, A.A. Johnson, Y. Pommier, M. Cushman, Bioorg. Med. Chem. 10 (2002) 2807–2816.
[8] M. Stevens, C. Pannecouque, E. DeClercq, J. Balzarini, Antimicrob. Agents Chemother. 47 (2003) 3109–3116.
[9] J.N. Burrows, J.G. Cumming, S.M. Fillery, G.A. Hamlin, J.A. Hudson, R.J. Jackson, S. McLaughlin, J.S. Shaw, Bioorg. Med. Chem. Lett. 15 (2005) 25–28.
[10] J.G. Cumming, A.E. Cooper, K. Grime, C.J. Logan, S. McLaughlin, J. Oldfield, J.S. Shaw, H. Tucker, J. Winter, D. Whittaker, Bioorg. Med. Chem. Lett. 15 (2005) 5012–5015.
[11] L.H. Wang, J.B. Milford, W.P.L. Carter, Atmos. Environ. 34 (2000) 4337–4348.
[12] A.R. Katritzky, D.C. Fara, R.O. Petrukhin, D.B. Tatham, U. Maran, A. Lomaka, M. Karelson, Curr. Top. Med. Chem. 2 (2002) 1333–1356.
[13] A.R. Katritzky, U. Maran, V.S. Lobanov, M.J. Karelson, Chem. Inf. Comput. Sci. 4 (2000) 1–8.
[14] A. Borghini, D. Pietra, P. Domenichelli, A.M. Bianucci, Bioorg. Med. Chem. 13 (2005) 5330–5337.
[15] A.I. Khlebnikov, I.A. Schepetkin, M.T. Quinn, Bioorg. Med. Chem. 14 (2006) 352–365.
[16] W. Li, Y. Tang, Y.L. Zheng, Z.B. Qiu, Bioorg. Med. Chem. 14 (2006) 601–610.
[17] J. Thomas Leonard, K. Roy, Bioorg. Med. Chem. Lett. 16 (2006) 4467–4474.
[18] P.J. Huber, Ann. Stat. 13 (1985) 435–475.
[19] J.H. Friedman, J.W. Tukey, IEEE Trans. Comput. C-23 (1974) 881–889.
[20] J.H. Friedman, J. Am. Stat. Assoc. 82 (1987) 249–266.
[21] Y.Y. Ren, H.X. Liu, S.Y. Li, X.J. Yao, M.C. Liu, Bioorg. Med. Chem. Lett. 17 (2007) 2474–2482.
[22] HyperChem. 4.0, hypercube, 1994.
[23] J.P.P. Stewart, MOPAC 6.0: Quantum Chemistry Program Exchange QCPE, No. 455, Indiana University, Bloomington, IN, 1989.
[24] A.R. Katritzky, V.S. Lobanov, M. Karelson, Comprehensive Descriptors for Structural and Statistical Analysis, Reference Manual, Version 2.0, University of Florida, Gainesville, FL, 1994.

[25] A.R. Katritzky, R. Petrukhin, R. Jain, M. Karelson, J. Chem. Inf. Comput. Sci. 1 (2001) 1521–1530.
[26] A.J. Smola, B. Schölkopf, NeuroCoLT2 Technical Report Series, NC2-TR-1998-030, 1998.
[27] J.H. Friedman, W. Stuetzle, J. Am. Atat. Assoc. 76 (1981) 817–823.
[28] Y. Du, Y. Liang, D. Yun, J. Chem. Inf. Comput. Sci. 42 (2002) 1283–1299.
[29] Q.Z. Hu, Y.Z. Liang, X.L. Peng, H. Yin, K.T. Fang, J. Chem. Inf. Comput. Sci. 44 (2004) 437–446.
[30] G.R. Famini, C.A. Penski, L.Y. Wilson, J. Phys. Org. Chem. 5 (1992) 395–408.
[31] Handbook of Chemistry and Physics, CRC Press, Cleveland, OH, 1974, pp. 112–135.
[32] A.B. Sannigrahi, Adv. Quantum Chem. 23 (1992) 301–351.