ORIGINAL PAPER

Three-dimensional quantitative structure-activity relationships of pyrrolopyridinone as cell division cycle kinase inhibitors by CoMFA and CoMSIA

**Junxia Zheng · Gaokeng Xiao · Jialiang Guo ·
Longyi Rao · Wei Chao · Kun Zhang · Pinghua Sun**

Abstract Seventy-five 1,5,6,7-tetrahydro-pyrrolo[3,2-C] pyridinone derivatives displaying potent activities against Cdc7 kinase were selected to establish 3D-QSAR models using CoMFA and CoMSIA methods. Internal and external cross-validation techniques were investigated as well as some measures including region focusing, progressive scrambling, bootstraping and leave-group-out. The satisfactory CoMFA model predicted a q^2 value of 0.836 and an r^2 value of 0.950, indicating that electrostatic and steric properties play a significant role in potency. The best CoMSIA model, based on a combination of steric, electrostatic and H-bond acceptor effects, predicted a q^2 value of 0.636 and an r^2 value of 0.907. The models were graphically interpreted using contour plots which provided insight into the structural requirements for increasing the activity of a compound. The final 3D-QSAR results could be used for rational design of potent inhibitors against Cdc7 kinase.

Keywords Cdc7 inhibitors · CoMFA · CoMSIA · Pyrrolopyridinone · QSAR

J. Zheng · K. Zhang
Faculty of Chemical Engineering and Light Industry, Guangdong University of Technology,
Guangzhou 510006, People's Republic of China

G. Xiao · J. Guo · L. Rao · W. Chao · P. Sun (✉)
Guangdong Province Key Laboratory of Pharmacodynamic Constituents of TCM and New Drugs Research, College of Pharmacy, Jinan University,
Guangzhou 510632, People's Republic of China
e-mail: biochemdoctor@sina.com

Introduction

The inhibition of DNA synthesis has proven a useful strategy in the treatment of hyper-proliferative disorders, but there is a great need for novel agents that can tackle the toxicity and resistance of current DNA replication inhibitors such as gemcitabine, active metabolites of 5-fluorouracil and hydroxyurea possibly by a different mechanism [1, 2]. The cell division cycle 7 (Cdc7), an evolutionary conserved serine-threonine kinase, is responsible for origin activation and the establishment of active replication forks by phosphorylating the "pre-replicative" complex of Cdc6 and MCM in the S-phase [3–6]. Cdc7 inhibition through siRNA or prototype small molecule inhibitors can cause tumor-cell death in a p53-independent manner while cell cycle progression is arrested in normal cells without loss of viability, surviving Cdc7 inhibition for long periods of time [7, 8]. Cdc7 kinase has been considered a potential target for anticancer therapy and a small-molecule inhibitor of Cdc7 kinase could be a useful drug for the treatment of cancers.

As the crystal structure of Cdc7 kinase remained unclear, ligand-based drug design techniques were employed to guide the synthesis of future generations of Cdc7 inhibitors. The three-dimensional quantitative structure-activity relationship (3D-QSAR) techniques are useful methods of ligand-based drug design used to correlate physicochemical descriptors from a set of related compounds to their known molecular activity or molecular property values [9]. The present 75 pyrrolopyridinones as first Cdc7 kinase inhibitors [10, 11] were evaluated for structure-activity correlation using comparative molecular field analysis (CoMFA) [12] and comparative similarity indices analysis (CoMSIA) [13] in order to derive the essential structural requirements

for antitumor activity and to discover more potent derivatives without toxicity and resistance in biological test systems.

Materials and methods

Molecular structures and optimization

Seventy-five molecules selected for the present study were taken from the published works of Oliver Werz and co-workers [10, 11]. The structures of the compounds and their biological data are given in Tables 1 and 2. For convenience, the IC_{50} values against Cdc7 kinase were often converted to their negative logarithm (pIC_{50}) values. Sixteen structures were randomly selected as the test set based on the structural diversity using "Selector", a program which is available in the Sybyl 8.1 [14] molecular modeling suite, while the remaining 55 structures were taken as the training set. Atom pairs and 2D fingerprints are used to form the diversity clusters by hierarchical clustering. To maximize the spread of activity data, one compound was randomly taken out from each cluster. In general, the spread of activity should cover at least 3 log units for a reliable 3D-QSAR model [15]. The pIC_{50} values of these compounds have a span of 3.4 log units from 5.30 to 8.70, providing a broad and homogenous data set for 3D-QSAR study (see Table 2).

One method of 3D-QSAR optimization is known as region focusing. As an iterative procedure, it involves giving additional weight to the lattice points in a given CoMFA region to enhance or attenuate the contribution of those points in a further analysis. Technically, region focusing can maximize q^2 value by rotating the extracted principal components [16].

Molecular alignment

A good alignment is the most essential element for CoMFA and CoMSIA analysis although a number of other factors such as lattice shifting step size, probe atom type and over all orientation of the aligned compounds may have bearing on their results [17]. The quality and the predictive ability of the model are directly dependent on the alignment rules. Once the active conformation was determined by energy minimization using Powell method and Tripos standard force field with a distance-dependent dielectric function, common substructure or pharmacophore alignment was performed on some rules [18]. In the present study, the pyrrole-3-formamide ring with structural rigidity which comprises the common core of atoms for all molecules in the data set was selected as the common substructure to overlap and to align all of the molecules. The most active compound 68 was used as the alignment template to produce a more robust CoMFA and CoMSIA models with a good cross-validated r^2 value. Alignment of all compounds was shown in Fig. 1. It can be seen that all the compounds studied have similar active conformations.

CoMFA and CoMSIA studies

Three-dimensional grid spacing was set at 2 Å in the x, y, and z directions and its grid region was automatically generated to be a 3D cubic lattice that extended at least 4 Å beyond van der Waals volume of all aligned molecules in all directions. Steric energy (Lennard-Jones potential) and electrostatic energy (Coulomb potential) were calculated using the Tripos force field [19] for each molecule in which the sp^3-hybridized carbon atom with a +1 charge was taken as the probe atom to determine the magnitude of the field values. To reduce domination by large steric and electrostatic energies to a minimum, all energies that exceeded the cutoff value of 30 kcal mol^{-1} were passed over. With standard options for scaling of variables, the regression analysis was carried out using partial least squares (PLS) method [20]. The column filtering was set to 2.0 kcal mol^{-1} to improve the signal to noise ratio by omitting those lattice points whose energy variation was below this threshold [21]. The final model was developed with the optimum number of components to yield a non cross-validated r^2 value. Despite being unable to describe all of the binding forces, CoMFA is still a widely useful tool for QSAR analysis at 3D level.

CoMSIA is similar to CoMFA in that a group of structurally aligned molecules are represented in terms of fields around the molecule, based on the same assumption that changes in binding affinities of ligands are related to changes in molecular properties represented by fields. Moreover, in CoMSIA, besides steric and electrostatic fields, three other different fields are calculated: hydrophobic, hydrogen bond donor, and hydrogen bond acceptor [22]. These fields were selected to cover the major contributions to ligand binding. A Gaussian function was also introduced to determine the distance between the probe atom and the molecule atoms. Similarity indices can be calculated at all grid points, inside and outside different molecular surfaces due to the different shape of the Gaussian function, while the calculations are not done inside the molecular surface in CoMFA. Equation used to calculate the similarity indices is as follows:

$$A_{F,K(j)}^q = \sum_i W_{probe,k} W_{ik} e^{-\alpha r_{iq}^2} \tag{1}$$

Where, A is the similarity index at grid point q, summed over all atoms i of the molecule j under investigation.

Table 1 Structures and inhibitory activities of Pyrrolopyridinone derivatives

1-11,13,15-33, 42-75   12   14

34-41

Compd.	R₁	R₂	R₃	R₄	IC50 (μM)
1		H	H	H	0.010
2[a]		H	H	H	5.0
3		H	H	H	0.009
4		H	H	H	0.020
5		H	H	H	0.007
6[a]		H	H	H	0.014

Table 1 (continued)

7[a]		H	H	H	0.178
8		H	H	H	5.0
9[a]		H	H	H	0.147
10		H	H	H	0.028
11[a]		H	H	H	0.004
12		H	H	H	0.045
13[a]		Me	H	H	0.404
14[a]		H	H	H	0.863
15		Et	H	H	0.067
16		HO—CH2CH2—	H	H	3.30
17[a]		CF3CH2—	H	H	0.160
18		Me	H	H	0.088

Table 1 (continued)
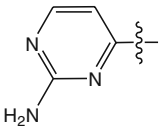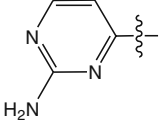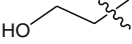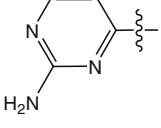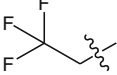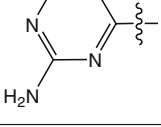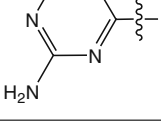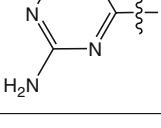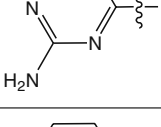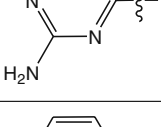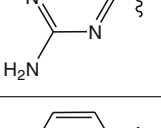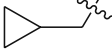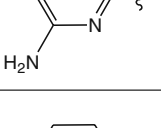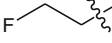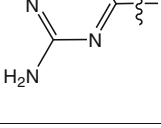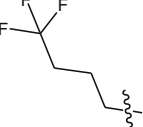
19		Et	H	H	0.008
20			H	H	0.138
21			H	H	0.005
22			H	H	0.012
23			H	H	0.042
24			H	H	0.014
25			H	H	0.023
26			H	H	0.013
27			H	H	0.004
28			H	H	0.003
29			H	H	0.047

Table 1 (continued)
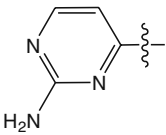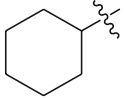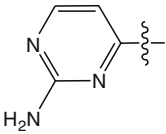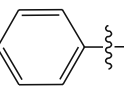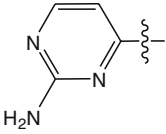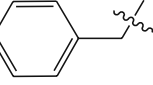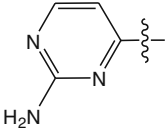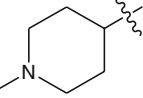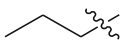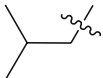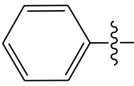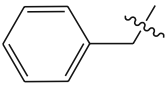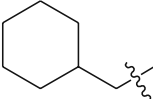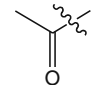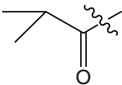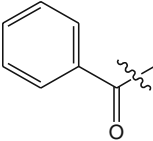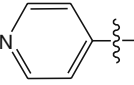
30			H	H	5.0
31			H	H	5.0
32			H	H	0.290
33			H	H	0.120
34		H	H	H	0.259
35		H	H	H	0.636
36[a]		H	H	H	0.007
37		H	H	H	0.220
38		H	H	H	0.655
39		H	H	H	0.323
40		H	H	H	1.397
41		H	H	H	0.262
42		H	H		0.013

Table 1 (continued)
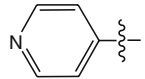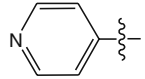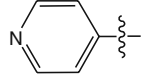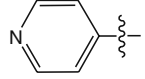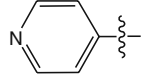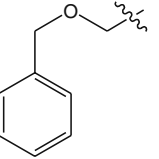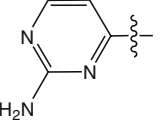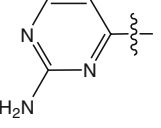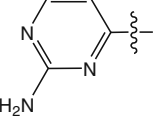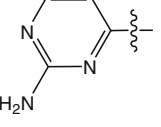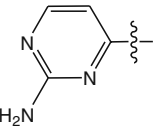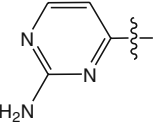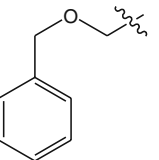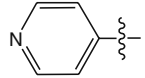
43	(pyridin-4-yl)	H	H	(isopropyl)	0.006
44	(pyridin-4-yl)	H	H	(cyclopropyl)	0.005
45[a]	(pyridin-4-yl)	H	H	(isobutyl)	0.003
46	(pyridin-4-yl)	H	H	(tert-butyl)	0.032
47	(pyridin-4-yl)	H	H	(benzyloxymethyl)	0.044
48	(2-aminopyrimidin-4-yl)	H	H	Me	0.026
49[a]	(2-aminopyrimidin-4-yl)	H	H	(isopropyl)	0.012
50	(2-aminopyrimidin-4-yl)	H	H	(cyclopropyl)	0.011
51[a]	(2-aminopyrimidin-4-yl)	H	H	(isobutyl)	0.008
52	(2-aminopyrimidin-4-yl)	H	H	(tert-butyl)	0.064
53	(2-aminopyrimidin-4-yl)	H	H	(benzyloxymethyl)	0.052
54	(pyridin-4-yl)	H	Me	H	0.010

Table 1 (continued)
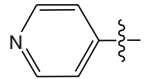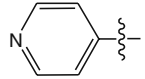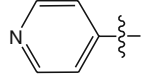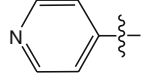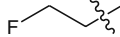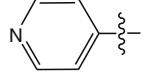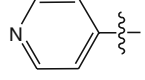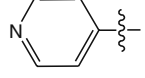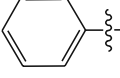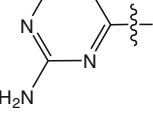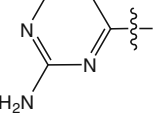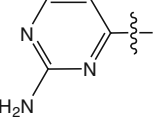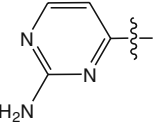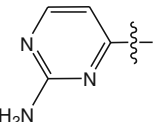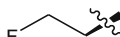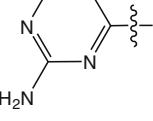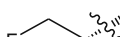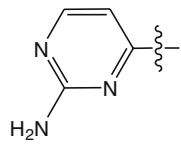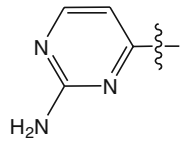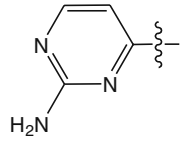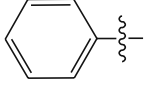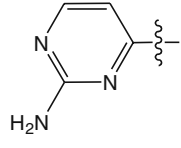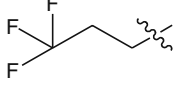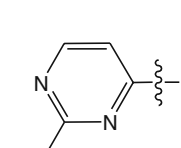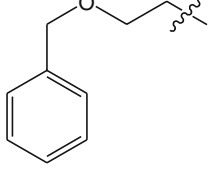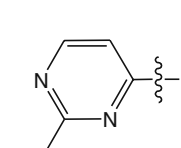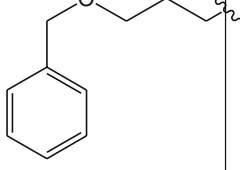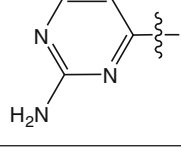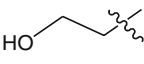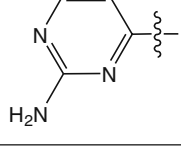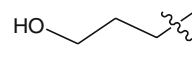
55	(pyridin-4-yl)	H	Et	H	0.002
56	(pyridin-4-yl)	H	isopropyl	H	0.010
57	(pyridin-4-yl)	H	cyclobutyl	H	0.003
58	(pyridin-4-yl)	H	F–CH₂CH₂–	H	0.002
59	(pyridin-4-yl)	H	(methyl-vinyl)	H	0.003
60	(pyridin-4-yl)	H	(allyl-ethyl)	H	0.006
61	(pyridin-4-yl)	H	phenyl	H	0.020
62	(2-amino-pyrimidin-4-yl)	H	Me	H	0.006
63	(2-amino-pyrimidin-4-yl)	H	Et	H	0.005
64	(2-amino-pyrimidin-4-yl)	H	isopropyl	H	0.004
65	(2-amino-pyrimidin-4-yl)	H	cyclobutyl	H	0.005
66	(2-amino-pyrimidin-4-yl)	H	F–CH₂CH₂–	H	0.005
67[a]	(2-amino-pyrimidin-4-yl)	H	F–CH₂CH₂–	H	0.002

Table 1 (continued)

68		H		H	0.002
69[a]		H		H	0.005
70		H		H	0.019
71		H		H	0.010
72[a]		H		H	0.007
73		H		H	0.024
74		H		H	0.004
75[a]		H		H	0.012

[a] Test set molecules

$W_{probe, k}$ is the probe atom with radius 1 Å, charge +1, hydrophobicity +1, hydrogen bond donating +1 and hydrogen bond accepting +1. W_{ik} is the actual value of the physicochemical property k of atom i. r_{iq} is the mutual distance between the probe atom at grid point q and atom i of the test molecule. α is the attenuation factor whose optimal value is normally between 0.2 and 0.4, with a default value of 0.3 [23, 24].

Table 2 The experimental pIC50, predicted pIC_{50} (Pred.) and their residuals (Res.) of the training and test set molecules

Compd.	Experimental	CoMFA		Focus CoMFA		CoMSIA	
		Pred.	Res.	Pred.	Res.	Pred.	Res.
1	8.000	7.866	0.134	7.893	0.107	7.851	0.149
2[a]	5.301	5.598	−0.297	5.449	−0.148	6.053	−0.752
3	8.046	7.929	0.117	8.027	0.019	7.941	0.105
4	7.699	7.282	0.417	7.141	0.558	7.262	0.437
5	8.155	7.765	0.390	7.792	0.363	7.728	0.427
6[a]	7.854	7.557	0.297	7.783	0.071	7.476	0.378
7[a]	6.750	6.604	0.146	6.779	−0.029	7.093	−0.343
8	5.301	4.710	0.591	4.984	0.317	4.407	0.894
9[a]	6.833	7.872	−1.039	7.701	−0.868	7.769	−0.936
10	7.553	7.658	−0.105	7.605	−0.052	7.652	−0.099
11[a]	8.398	7.616	0.782	7.859	0.539	7.581	0.817
12	7.347	7.442	−0.095	7.375	−0.028	7.575	−0.228
13[a]	6.394	6.762	−0.368	6.603	−0.209	6.855	−0.461
14[a]	6.064	6.880	−0.816	6.515	−0.451	6.712	−0.648
15	7.174	7.092	0.082	7.106	0.068	7.249	−0.075
16	5.481	5.388	0.093	5.450	0.031	5.255	0.226
17[a]	6.796	6.011	0.785	6.310	0.486	6.007	0.789
18	7.055	7.533	−0.478	7.409	−0.354	7.757	−0.702
19	8.097	8.352	−0.255	8.320	−0.223	8.031	0.066
20	6.860	6.933	−0.073	6.908	−0.048	6.897	−0.037
21	8.301	8.312	−0.011	8.225	0.076	8.541	−0.240
22	7.921	7.845	0.076	7.930	−0.009	7.772	0.149
23	7.377	7.146	0.231	7.218	0.159	7.168	0.209
24	7.854	7.980	−0.126	7.951	−0.097	7.980	−0.126
25	7.638	7.497	0.141	7.518	0.120	7.740	−0.102
26	7.886	7.718	0.168	7.720	0.166	7.617	0.269
27	8.398	8.108	0.290	8.087	0.311	7.852	0.546
28	8.523	8.506	0.017	8.489	0.034	8.343	0.180
29	7.328	7.415	−0.087	7.364	−0.036	7.670	−0.342
30	5.301	5.091	0.210	5.118	0.183	5.435	−0.134
31	5.301	5.553	−0.252	5.542	−0.241	5.594	−0.293
32	6.538	6.895	−0.357	6.728	−0.190	7.300	−0.762
33	6.921	6.932	−0.011	6.899	0.022	7.032	−0.111
34	6.587	6.483	0.104	6.564	0.023	6.139	0.448
35	6.196	6.196	0.000	6.196	0.000	6.463	−0.267
36[a]	8.155	7.828	0.327	7.923	0.232	8.003	0.152
37	6.658	6.467	0.191	6.590	0.068	6.493	0.165
38	6.184	6.166	0.018	6.173	0.011	6.307	−0.123
39	6.491	6.648	−0.157	6.641	−0.15	6.300	0.191
40	5.855	5.945	−0.090	5.899	−0.044	5.871	−0.016
41	6.582	6.437	0.145	6.179	0.403	6.556	0.026
42	7.886	7.677	0.209	7.801	0.085	7.884	0.002
43	8.222	8.291	−0.069	8.200	0.022	8.031	0.191
44	8.301	8.065	0.236	8.178	0.123	7.955	0.346
45[a]	8.523	8.055	0.468	8.085	0.438	8.057	0.466
46	7.495	7.498	−0.003	7.496	−0.001	7.768	−0.273
47	7.356	7.441	−0.085	7.297	0.059	7.264	0.092

Table 2 (continued)

Compd.	Experimental	CoMFA		Focus CoMFA		CoMSIA	
		Pred.	Res.	Pred.	Res.	Pred.	Res.
48	7.585	7.557	0.028	7.689	−0.104	7.763	−0.178
49[a]	7.921	7.953	−0.032	7.958	−0.037	7.852	0.069
50	7.959	7.969	−0.010	7.953	0.006	7.841	0.118
51[a]	8.097	8.008	0.089	8.048	0.049	7.960	0.137
52	7.194	7.385	−0.191	7.343	−0.149	7.673	−0.479
53	7.284	7.475	−0.191	7.366	−0.082	7.282	0.002
54	8.000	8.366	−0.366	8.380	−0.380	8.113	−0.113
55	8.699	8.431	0.268	8.479	0.220	8.078	0.621
56	8.000	8.255	−0.255	8.229	−0.229	8.375	−0.375
57	8.523	8.404	0.119	8.384	0.139	8.251	0.272
58	8.699	8.592	0.107	8.616	0.083	8.448	0.251
59	8.523	8.578	−0.055	8.508	0.015	8.580	−0.057
60	8.222	8.377	−0.155	8.238	−0.016	8.557	−0.335
61	7.699	7.720	−0.021	7.716	−0.017	7.758	−0.059
62	8.222	8.175	0.047	8.212	0.010	7.995	0.227
63	8.301	8.207	0.094	8.272	0.029	8.204	0.097
64	8.398	8.209	0.189	8.315	0.083	8.292	0.106
65	8.301	8.317	−0.016	8.315	−0.014	8.272	0.029
66	8.301	8.380	−0.079	8.338	−0.037	8.351	−0.050
67[a]	8.699	8.403	0.296	8.559	0.140	8.414	0.285
68	8.699	8.417	0.282	8.473	0.226	8.490	0.209
69[a]	8.301	8.054	0.247	8.341	−0.040	8.630	−0.329
70	7.721	7.556	0.165	7.683	0.038	7.675	0.046
71	8.000	8.053	−0.053	7.964	0.036	8.089	−0.089
72[a]	8.155	7.697	0.458	7.771	0.384	8.421	−0.266
73	7.619	7.639	−0.020	7.591	0.028	7.731	−0.112
74	8.398	8.331	0.067	8.337	0.061	8.647	−0.249
75[a]	7.921	8.102	−0.181	7.945	−0.024	8.598	−0.677

[a] Test set molecules

Partial least squares (PLS) analysis

PLS analysis [20], used to linearly correlate the CoMFA and CoMSIA fields to biological activity values, was first carried out by the leave-one-out (LOO) and leave-group-out (10 compound groups) cross-validation methods, respectively, to determine cross-validated r^2 (q^2) values and the optimal number of components. In this particular method, one compound was removed from the data set and its activity was predicted using the model from the rest of the data set. The optimal number of components, usually corresponding to the highest cross-validated squared coefficient (q^2), was selected on the basis of the lowest standard error of prediction (SEP). In order to avoid over-fitting the models, a higher component was accepted and used only

when the q^2 differences between two components was larger than 10%. q^2 is calculated as follows:

$$q^2 = 1 - \frac{\sum \left(Y_{obs} - Y_{pre}\right)^2}{\sum \left(Y_{obs} - Y_{mean}\right)^2} \quad (2)$$

Where, Y_{obs} = experimental activity, Y_{pre} = predicted activity, Y_{mean} = mean activity.

Non-cross-validation was performed to establish the final 3D-QSAR model using the optimal number of components, in which conventional correlation coefficient (r^2), standard errors of estimate (SEE), and F ratio between the variances of calculated and observed activities were given. To further assess the robustness of the derived models, bootstraping analysis (10 runs) was also utilized to
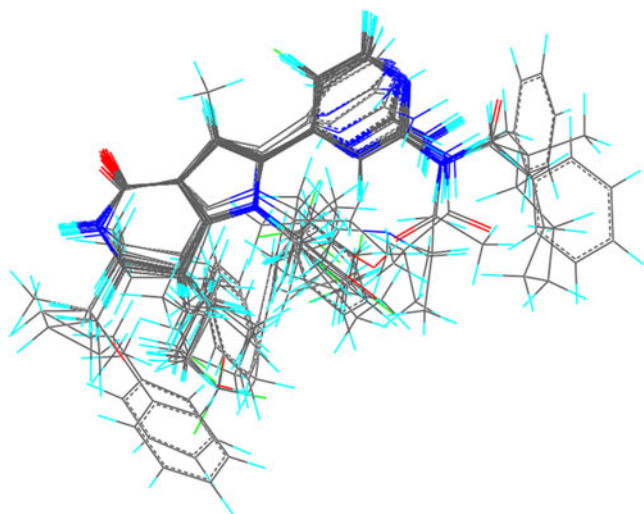
Fig. 1 Molecular alignment of pyrrolopyrindinone derivatives

calculate confidence intervals for the r^2 and SEE [20, 25]. The equation for SEE is given below.

$$SEE = \sqrt{\frac{PRESS}{n - c - 1}} \qquad (3)$$

Where n means number of compounds, c means number of components, and PRESS (predicted sum of squares) means $\sum (Y_{obs} - Y_{pre})^2$.

The q^2 has been a good indicator of the accuracy of actual predictions. In general, q^2 values can be separated into three categories: $q^2 > 0.6$ means a fairly good model, $q^2 = 0.4 - 0.6$ means a questionable model, and $q^2 < 0.4$ a poor model [26].

Sensitivity of a PLS model

Cross-validation including leave-one-out and leave-group-out generally confound redundancy with predictivity when applied to large data sets. Because most members of the data set may have "twins," which makes a near twin of each left-out molecule likely remain in the training data and usually obtain good predictions, but the q^2 statistic obtained from cross-validation may give you a false sense of confidence [27]. To check the model's stability, progressive scrambling is used to determine the sensitivity of a QSAR model to small systematic perturbations of the response variable [28, 29]. In the SYBYL code for progressive scrambling, a third order polynomial was used for application over a wider range of bin sizes. The values of Q^2, cSDEP, and $dq^2/dr_{yy'}^2$ are returned and can aid in interpreting the predictivity of the model without the potentially confusing redundancy.

$$cSDEP[r_{yy'}^2] = a_0 + a_1(r_{yy'}^2) + a_2(r_{yy'}^2)^2 + a_3(r_{yy'}^2)^3 \qquad (4)$$

$$q^2(r_{yy'}^2) = b_0 + b_1(r_{yy'}^2) + b_2(r_{yy'}^2)^2 + b_3(r_{yy'}^2)^3 \qquad (5)$$

$$\frac{dq^2}{dr_{yy'}^2} = b_1 + 2b_2(r_{yy'}^2) + 3b_3(r_{yy'}^2)^2 \qquad (6)$$

Here, y' indicates the perturbed (scrambled) responses and the statistic $r_{yy'}^2$ indicates the degree of correlation between the perturbed responses and the original ones. The cSDEP statistic is an estimated cross-validated standard error at a specific critical point for $r_{yy'}^2$ and calculated using a third-order polynomial equation, where scramble critical-point is default value of 0.85. The Q^2 statistic is an estimate of the predictivity of the model after removing the effects of redundancy. The Q^2 statistic obtained in this way is very conservative, in that it is necessarily reduced by the level of noise introduced to remove redundancy, so a value as low as 0.35 will signify that the original, unperturbed model is

Table 3 Statistical results of CoMFA and best CoMSIA models

	CoMFA (before region focusing)	CoMFA (after region focusing)	CoMSIA (Model 4)
PLS statistics			
LOO cross q^2/SEP	0.682/0.516	0.836/0.376	0.636/0.552
Group cross q^2/SEP	0.720/0.482	0.818/0.391	0.588/0.588
Non-validated r^2/SEE	0.951/0.202	0.950/0.205	0.907/0.280
F	157.391	160.925	80.820
$r_{bootstrap}^2$	0.967±0.010	0.971±0.008	0.933±0.013
$S_{bootstrap}$	0.163±0.079	0.141±0.065	0.228±0.120
Optimal components	6	6	6
Field distribution%			
steric	76.9	69.5	47.6
electrostatic	23.1	30.5	36.8
H-bond acceptor			15.6
r_{pred}^2	0.735	0.871	0.698

Fig. 2 Region focusing. The CoMFA field calculations are shown for 68 before (left) and after (right) region focusing. Electrostatic fields (upper): Blue fields indicate electropositive groups favored, red fields indicate electronegative groups favored. Steric fields (lower): Green fields indicate steric bulk favored, yellow fields indicate steric bulk disfavored
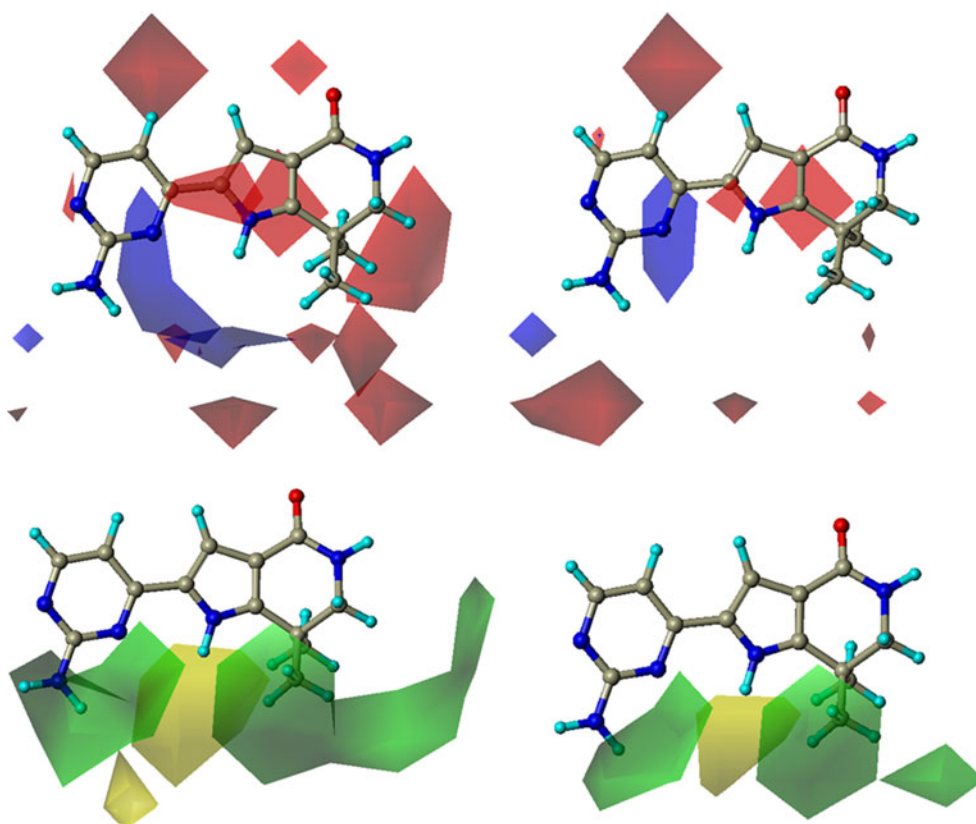
robust. [28] The slope of q^2 evaluated at the specified critical point with respect to $r_{yy'}^2$ is reported as $dq^2/dr_{yy'}^2$, indicating to what extent the model changes with small changes to the dependent variable. In an unstable model, changing greatly with small changes in underlying response values, the effective slope is generally greater than 1.2, while stable models which change proportionally with small changes in underlying data have slopes near unity. This method was employed to verify the number of components used to build the final model and to check

the cross-validation against the possibility of such a redundancy in our training set [30].

Predictive correlation coefficient

An external test sets (r_{pred}^2) [31] was recommended for the estimation of predictive ability because q^2 is a useful but not sufficient criterion for model validation. Equation of predictive values r_{pred}^2 is as follows:

$$r_{pred}^2 = 1 - (PRESS/SD) \tag{7}$$

Therein, SD means the sum of squared differences between the measured activities of the test set and the average measured activity of the training set.

Fig. 3 Graph of experimental versus predicted pIC50 of the training set and the test set using the CoMFA model after region focusing

Table 4 Progressive scrambling results of the focus CoMFA model

components	Q^2	cSDEP	$dq^2/dr_{yy'}^2$
2	0.477	0.637	0.281
3	0.540	0.603	0.420
4	0.574	0.585	0.618
5	0.602	0.571	0.746
6	0.597	0.578	0.963
7	0.589	0.589	1.152
8	0.567	0.608	1.277
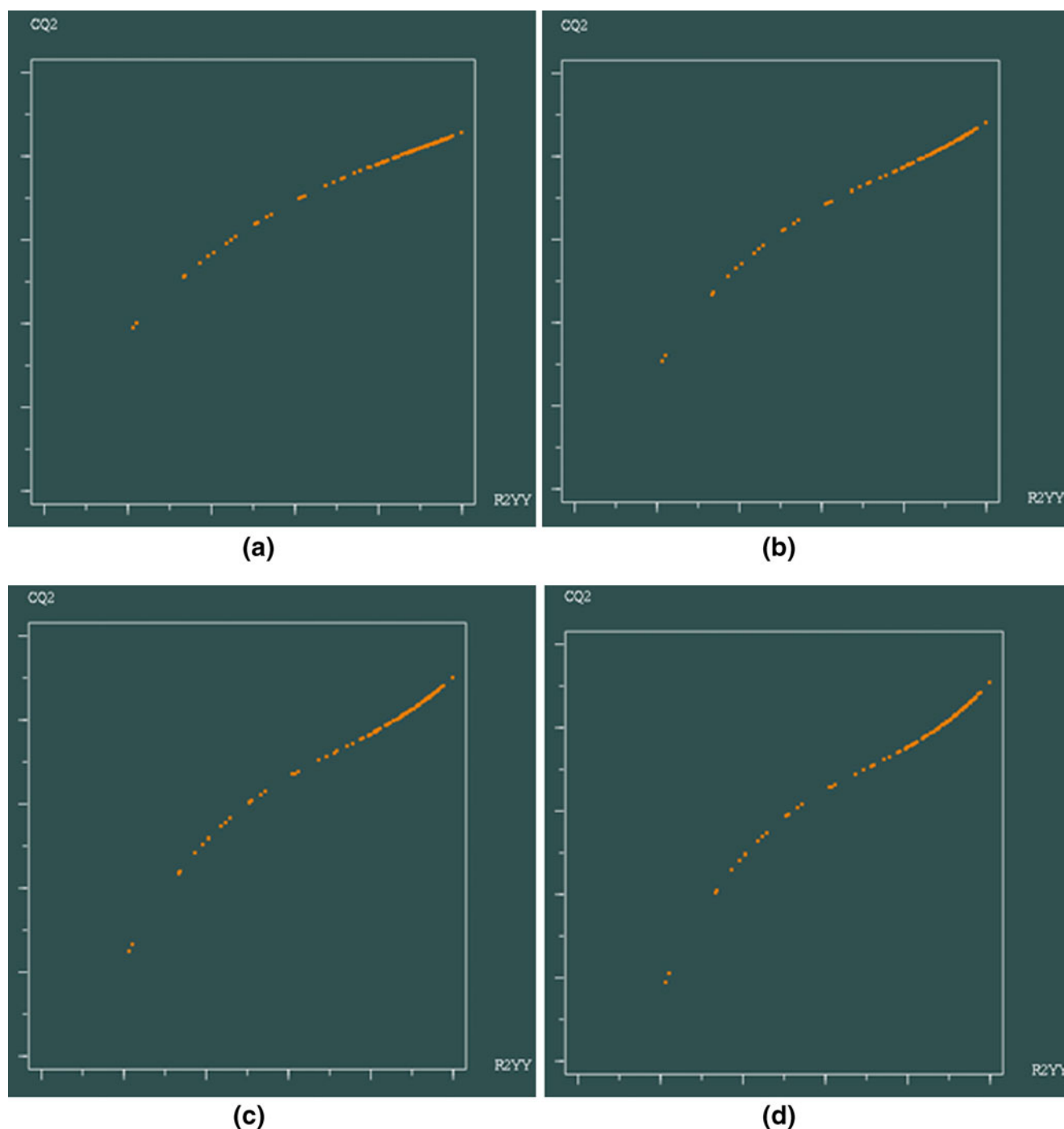
Fig. 4 Variation fitted curves for progressive scrambling analyses with random number seed: (**a**) five components; (**b**) six components; (**c**) seven components; (**d**) eight components

Results and discussion

CoMFA and CoMSIA analysis

The most active compound **68** (Table 1) was selected as the template for alignment as shown in Fig. 1. The **CoMFA** model provided a cross-validation q^2 value of 0.682 with six components and an r^2 value of 0.951 (Table 3).

The new model was generated through region focusing with increased predictive power (q^2), enhanced resolution, tighter grid spacing, and greater stability at a higher number of components. Figure 2 shows the CoMFA fields for molecule **68** before and after region focusing. Application of region focusing resulted in a significant increase from 0.682 to 0.836 for the internal validity, from 0.720 to 0.818 for group cross-validation, and from 0.735 to 0.871 for test

Table 5 Results of CoMSIA models using combinations of the five field descriptors

Model	Descriptors	LOO cross q^2/SEP	Group cross q^2/SEP	Bootstrap r^2	Bootstrapped SEE	Non-validated r^2/SEE
1	S and E	.718/.486	.703/.499	.949±.013	.204±.113	.925/.251
2	D and A	.002/.880	.056/.905	.639±.076	.552±.240	.499/.647
3	S, E and H	.589/.587	.564/.604	.956±.010	.186±.094	.948/.208
4	S, E and A	.636/.552	.588/.588	.933±.013	.228±.120	.907/.280
5	S, E and D	.619/.564	.574/.597	.923±.021	.256±.124	.883/.313
6	E, H and A	.255/.970	.284/.767	.963±.010	.180±.093	.938/.228
7	E, H and D	.358/.733	.300/.765	.950±.021	.195±.083	.916/.264
8	D, A and H	.269/.767	.253/.761	.942±.018	.217±.098	.889/.305
9	D, A and S	.316/.728	.308/.733	.683±.077	.495±.232	.590/.564
10	D, A and E	.112/.830	.091/.839	.575±.076	.591±.259	.474/.638
11	S, D and H	.463/.658	.422/.682	.910±.018	.281±.128	.835/.364
12	S, A and H	.483/.658	.518/.635	.951±.020	.196±.101	.921/.257
13	S, E, D and A	.526/.630	.535/.624	.913±.017	.263±.127	.874/.324
14	S, E, D and H	.529/.628	.531/.627	.956±.016	.193±.101	.933/.236
15	S, E, A and H	.515/.637	.486/.656	.964±.010	.181±.096	.939/.225
16	D, A, H and S	.433/.676	.460/.659	.894±.033	.282±.150	.832/.368
17	D, A, H and E	.273/.751	.237/.769	.742±.043	.456±.182	.710/.474
18	S, E, D, A and H	.451/.678	.482/.658	.902±.026	.274±.130	.854/.343

set activity predictions, with small to negligible effect to the non-validated r^2 (Table 3). The inhibitory activity values predicted for the tested compounds are in good agreement with the experimental values (Fig. 3). The r^2_{pred} value of 0.871 further confirms that the model is reliable and accurate with higher predictive capacity. Therefore, this model will be used to predict the activity and guide future synthetic efforts on novel Cdc7 inhibitors.

The progressive scrambling method was employed against the final focus CoMFA model to verify the number of components used to build the model and to check the cross-validation. Table 4 lists the results of the
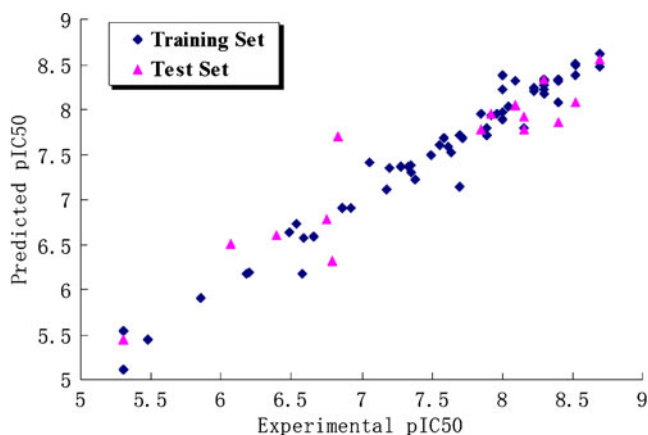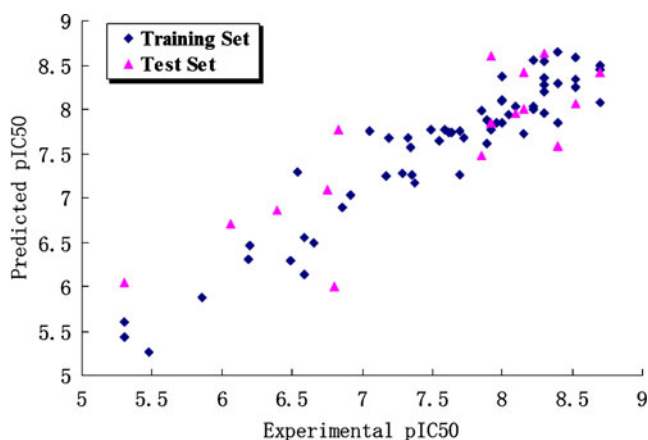

Fig. 5 Graph of experimental versus predicted pIC50 of the training set and the test set using the CoMSIA model 4

progressive scrambling of the final model, from which the values for the slope in the six and seven component models are acceptable (also see Fig. 4), and the optimum statistics are seen for six components: *cSDEP* is at a minimum and Q^2 is at a maximum. The eight component model has a less acceptable slope (1.28), but this model is also rejected as being over predicted on the basis of the other two statistics. Still, several models are good and well behaved, with the best found for six components.

In CoMSIA, 18 CoMSIA models were generated using combinations of two, three, four, and all five descriptors as shown in Table 5. Model 4 based on steric, electrostatic and hydrogen-bond acceptor fields was found to be the most accurate yielding a q^2 value of 0.636 and an r^2 value of 0.907. The Group cross q^2 value of 0.588, bootstrapped value of 0.933±0.013 and test set r^2 value of 0.698 conformed that model 4 was the best CoMSIA model. The predicted values are consistent with the experimental data (Fig. 5). This model was subsequently selected to generate the final CoMSIA model.

CoMFA and CoMSIA contour maps

The results of 3D-QSAR models are presented in the contour coefficient maps as shown in Figs. 2 and 6. The CoMFA steric contour map of the most active molecule **68** displays a large green polyhedron at position 7 of the six-membered lactam ring and around position 2' and 3' of aromatic ring substituted at position 2 of the pyrrole ring,

Fig. 6 CoMSIA fields. The CoMSIA fields from model 4 are shown with active compound 68. (**a**) Steric fields, green indicates steric bulk favored, yellow indicates bulk disfavored. (**b**) Electrostatic fields, blue indicates electropositive groups favored, red fields indicate electronegative groups favored. (**c**) H-bond acceptor fields, cyan indicates acceptor favored, magenta indicates disfavored
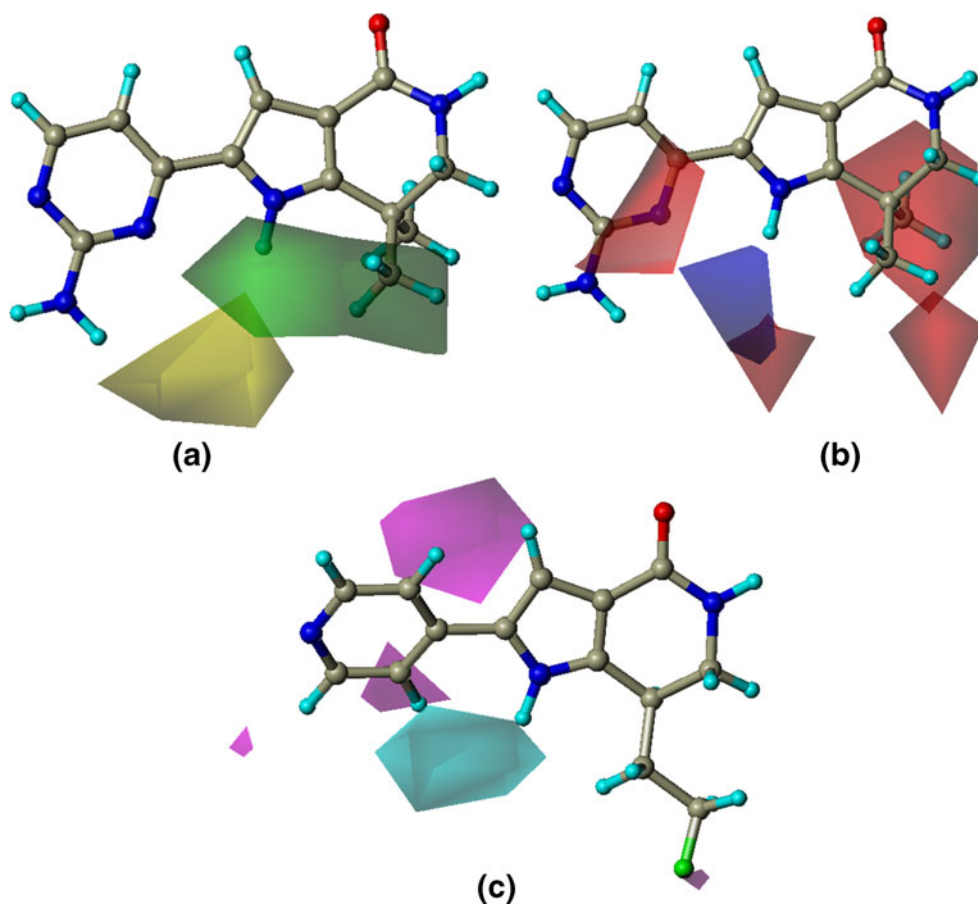
respectively, indicating, in both cases, that binding affinity should be increased, and one yellow polyhedra at *N*-position of the pyrrole ring indicating that bulk is disfavored. This may be the reason why compound **58** with a 2-fluoroethyl substitution at position 7 and no substitution at *N*-position showed the most active while compound **30** with only cyclohexanyl substitution at *N*-position shows the lowest activity (Fig. 7). The large green polyhedron at position 7 also interpreted that compounds **54–60**, **62–69**, **71**, **72** and **74** have IC50 values in low nanomolar range, and that compounds **42**, **46**, **48**, **49** and

52 with some substituents at position 6 and no substitution at position 7 showed lower potency although compounds **54** and **56** have similar activity as compounds **42** and **43**, respectively. The free NH with no other substitutions is necessary for keeping higher potency, so compounds **13**, **15**–**18** are all less active than parent compound **1** with a free NH at pyrrole ring, and compounds **18**, **20**, **22–26** and **29–33** with some substituents at *N*-position of pyrrole ring as less active than their parent compound **5**, in which compounds **27** and **28** have similar activity as compound **5** for other factors such as electrostatic fields and lipophilic interaction.

Fig. 7 CoMFA steric contour maps for most highly active compound 58 (left) and less potent compound 30 (right)
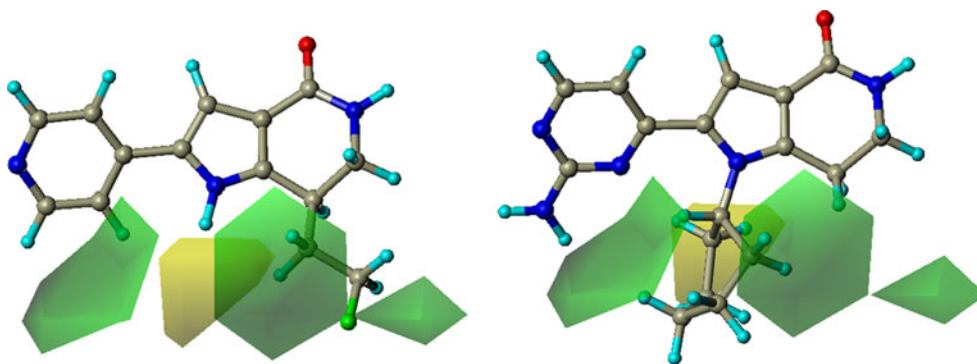
For the electrostatic contour plots, a large blue polyhedron located near position 2′ of compound **68**, indicating that negatively charged groups are disfavored at this position, and /or a green polyhedron at position 3' in the steric contour maps (Fig. 2) may be the reason why compounds **48**–**53** are less potent than corresponding compounds **42**–**47** with no electron-deficient substituent at position 2' and no amino group at position 3', respectively. The presence of negative charge favoring red contours around the lactam ring indicates that the addition of electron donating groups may increase the binding affinity. The steric and electrostatic field contributions to the final model were 69.5% and 30.5%, respectively.

The best CoMSIA model contour maps of the most active analog (compound **68** in Table 1) are shown in Fig. 6. Its steric contour plots (Fig. 6a) correlate well with the CoMFA contour maps described above except one green polyhedron at position 3' However, its electrostatic contour plots (Fig. 6b) don't go well with the CoMFA contour maps. Consideration of the values of q^2 and r^2, the CoMFA electrostatic contour map (Fig. 2) was chosen to analyze the model. Figure 6c shows the CoMSIA hydrogen-bond acceptor fields denoted by magenta and cyan contours. Cyan contours represent regions where hydrogen-bond acceptor substituents are preferred and magenta contours indicate unfavorable regions. One large cyan contour around position 1 of pyrrole ring indicates once again that the free NH is necessary for potencies. The CoMSIA steric field explains 47.6% of the variance, the electrostatic field explains 36.8% and the hydrogen-bond acceptor explains 15.6% of the variance.

Conclusions

Based on a series of pyrrolopyridinones with known Cdc7 inhibitory activity, some predictive 3D-QSAR models have been developed. Both CoMFA and CoMSIA models provided good statistical results in terms of q^2 and r^2 values for pyrrolo[3,2-C]pyridine derivatives, suggesting the significant correlations of biological activities with molecular structures. Compared with CoMSIA, CoMFA provided a slightly better statistical model. The final model was optimized by region focusing and progressive scrambling, and validated by a variety of methods including crossvalidation, non-crossvalidation and test set predictions. The model developed has high internal validity (q^2 above 0.5) and high predictive ability (test set r^2 above 0.7). The 3D-QSAR results also revealed some important sites, where steric, electrostatic and hydrogen-bond acceptor modifications should significantly affect the bioactivities of these compounds. The 3D-QSAR models are being used to

predict the Cdc7 inhibitory activity of proposed new compounds and quite reliable to efficiently guide further modification in the molecules for obtaining better Cdc7 inhibitors.

References

1. Swords R, Mahalingam D, O'Dwyer M, Santocanale C, Kelly K, Carew J, Giles F (2010) Eur J Cancer 46:33–40
2. Longley DB, Johnston PG (2005) J Pathol 205:275–292
3. Labib K (2010) Genes Dev 24:1208–1219
4. Ermoli A, Bargiotti A, Brasca MG, Ciavolella A, Colombo N, Fachin G, Isacchi A, Menichincheri M, Molinari A, Montagnoli A, Pillan A, Rainoldi S, Sirtori FR, Sola F, Thieffine S, Tibolla M, Valsasina B, Volpi D, Santocanale C, Vanotti E (2009) J Med Chem 52:4380–4390
5. Zhao C, Tovar C, Yin X, Xu Q, Todorov IT, Vassilev LT, Li C (2009) Bioorg Med Chem Lett 19:319–323
6. Montagnoli A, Valsasina B, Brotherton D, Troiani S, Rainoldi S, Tenca P, Molinari A, Santocanale C (2006) J Biol Chem 281:10281–10290
7. Montagnoli A, Tenca P, Sola F, Carpani D, Brotherton D, Albanese C, Santocanale C (2004) Cancer Res 64:7110–7116
8. Im J, Lee J (2008) J Biol Chem 283:25171–25177
9. Hansch C, Smith N, Engle R, Wood H (1972) Cancer Chemother Rep 56:443–456
10. Menichincheri M, Bargiotti A, Berthelsen J, Bertrand JA, Bossi R, Ciavolella A, Cirla A, Cristiani C, Croci V, D'Alessio R, Fasolini M, Fiorentini F, Forte B, Isacchi A, Martina K, Molinari A, Montagnoli A, Orsini P, Orzi F, Pesenti E, Pezzetta D, Pillan A, Poggesi I, Roletto F, Scolaro A, Tato M, Tibolla M, Valsasina B, Varasi M, Volpi D, Santocanale C, Vanotti E (2009) J Med Chem 52:293–307
11. Vanotti E, Amici R, Bargiotti A, Berthelsen J, Bosotti R, Ciavolella A, Cirla A, Cristiani C, D'Alessio R, Forte B, Isacchi A, Martina K, Menichincheri M, Molinari A, Montagnoli A, Orsini P, Pillan A, Roletto F, Scolaro A, Tibolla M, Valsasina B (2008) J Med Chem 51:487–501
12. Cramer RD III, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959–5967
13. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130–4146
14. SYBYL 8.1, Tripos International. St. Louis, MO, USA
15. Cramer RD III, Patterson DE, Bunce JD (1989) Prog Clin Biol Res 291:161–165
16. Lindgren F, Geladi P, Rännar S, Wold S (1994) J Chemom 8:349–363
17. Cho SJ, Tropsha A (1995) J Med Chem 38:1060–1066
18. Cramer RD III, Clark M, Simeroth P, Patterson DE (1991) Pharmacochem Libr 16:239–242
19. Clark MC, Cramer RD III, Bosch NVO (1989) J Comput Chem 10:982–1012
20. Bush BL (1993) J Comput Aided Mol Des 7:587–619
21. Srivastava V, Gupta SP, Siddiqi MI, Mishra BN (2010) Eur J Med Chem 45:1560–1571
22. Klebe G, Abraham U (1999) J Comput Aided Mol Des 13:1–10
23. Böhm M, Stürzebecher J, Klebe G (1999) J Med Chem 42:458–477
24. Woodard B, Saleh MA (2008) J Environ Sci Heal B 43:281–287

25. Clark M, Cramer RD III, Jones DM, Patterson DE, Simeroth PE (1990) Tetrahedron Comput Methodol 3:47–59
26. Clark M, Cramer RD III (1993) Quant Struct Act Relat 12:137–145
27. Clark RD, Sprous DG, Leonard JM (2001) In: Holtje HD, Sippl W (eds) Rational Approaches to Drug Design. Prous Science SA, pp 475–485
28. Clark RD, Fox PC (2004) J Comput Aided Mol Des 18:563–576
29. Luco JM, Ferretti FH (1997) J Chem Inf Comput Sci 37:392–401
30. Hevener KE, Ball DM, Buolamwini JK, Lee RE (2008) Bioorg Med Chem 16:8042–8053
31. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269–276