Journal of Molecular Graphics and Modelling 27 (2008) 430–438

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2

Jun Zou, Huan-Zhang Xie, Sheng-Yong Yang *, Jin-Juan Chen, Ji-Xia Ren, Yu-Quan Wei

State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, and School of Life Sciences, Sichuan University, Chengdu 610041, China

ARTICLE INFO

Article history:
Received 25 May 2008
Received in revised form 23 July 2008
Accepted 28 July 2008
Available online 7 August 2008

Keywords:
Pharmacophore
CDK2 inhibitors
Protein kinase
Virtual screening
Structure-based method

ABSTRACT

Pharmacophore modeling, including ligand- and structure-based approaches, has become an important tool in drug discovery. However, the ligand-based method often strongly depends on the training set selection, and the structure-based pharmacophore model is usually created based on *apo* structures or a single protein–ligand complex, which might miss some important information. In this study, multicomplex-based method has been suggested to generate a comprehensive pharmacophore map of cyclin-dependent kinase 2 (CDK2) based on a collection of 124 crystal structures of human CDK2–inhibitor complex. Our multicomplex-based comprehensive pharmacophore map contains almost all the chemical features important for CDK2–inhibitor interactions. A comparison with previously reported ligand-based pharmacophores has revealed that the ligand-based models are just a subset of our comprehensive map. Furthermore, one most-frequent-feature pharmacophore model consisting of the most frequent pharmacophore features was constructed based on the statistical frequency information provided by the comprehensive map. Validations to the most-frequent-feature model show that it can not only successfully discriminate between known CDK2 inhibitors and the molecules of focused inactive dataset, but also is capable of correctly predicting the activities of a wide variety of CDK2 inhibitors in an external active dataset. Obviously, this investigation provides some new ideas about how to develop a multicomplex-based pharmacophore model that can be used in virtual screening to discover novel potential lead compounds.

1. Introduction

Pharmacophore modeling method, as a key tool of computer-aided drug design, has been widely used in the lead discovery and optimization [1–3]. The pharmacophore modeling method involves two aspects: pharmacophore hypothesis generation and 3D structural database search. The latter is very mature right now and many advanced 3D database searching algorithms have been implemented in commercial programs. However, the pharmacophore hypothesis generation method is still under development although there are several commercial programs available presently [4]. The generation method of pharmacophore model can be classified into two categories: direct method and indirect method [5–7]. Direct method uses *apo* structure or receptor–ligand complex information (usually called receptor-based or structure-based method), whereas indirect method uses a set of ligands that have been experimentally observed to interact with a specific biological target (called ligand-based method).

The ligand-based method has been very popular since for a long time just a limited number of protein structures or protein–ligand complex structures are available. It has been shown that the quality of the pharmacophore model generated by ligand-based method can be affected significantly by two factors, namely, the conformation generation method and the selection of the training set molecules [8–10]. For the conformation generation method, systematic researches and large-scale survey works have been carried out by Langer, Gillet and others. It has been demonstrated that most of the conformation model generators are competent for the pharmacophore construction [11–14]. However, the selection of training set compounds used for the

* Corresponding author at: State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, No. 1, Keyuan Road 4, High Technology Development Zone, Chengdu 610041, China.
Tel.: +86 28 85164063; fax: +86 28 85164060.
E-mail address: yangsy@scu.edu.cn (S.-Y. Yang).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.07.004

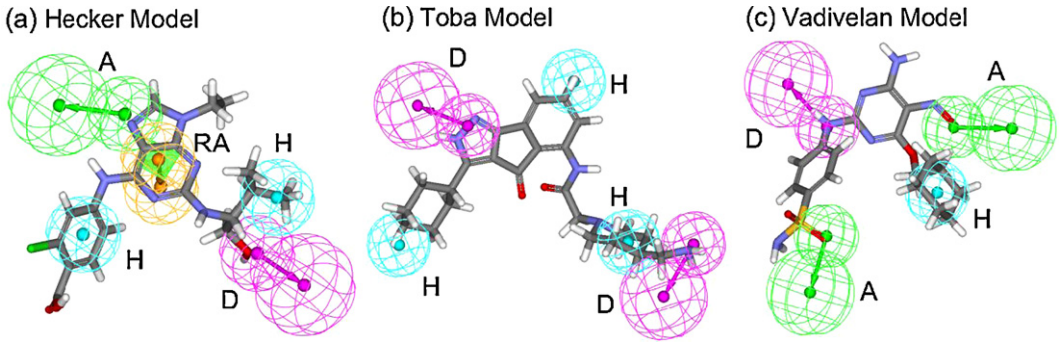

Fig. 1. The ligand-based pharmacophore models of CDK2 inhibitors reported previously by (a) Hecker, (b) Toba and (c) Vadivelan. (A, hydrogen bond acceptor; B, hydrogen bond donor; H, hydrophobic feature; RA, ring aromatic feature.)

model generation is quite complicated. In order to establish a good pharmacophore model, specific guidelines are required to choose appropriate training set molecules, such as the activity range and the structural diversity of the selected compounds. The difference in the training set selection has a large influence on the final pharmacophore model. A possible case is that completely different pharmacophore models of ligands interacting with the same protein target could be generated with the use of the same algorithm and program but different training set. One important case in point is that of cyclin-dependent kinase 2 (CDK2) inhibitors. Hecker et al. [15], Toba et al. [16] and Vadivelan et al. [17] have independently reported three ligand-based pharmacophore models for CDK2 inhibitors. However, the three pharmacophore models are found to be different from each other in terms of the feature categories as well as the location constraint of features (Fig. 1). These are mainly due to that different training set molecules were used in the three studies. A possible way to overcome this type of shortcomings is to adopt the structure-based pharmacophore modeling method, which just use the information of receptor or receptor–ligand complex.

Usually structure-based pharmacophore models are generated based on *apo* structures or a single protein–ligand complex. However, in the case of the identical binding modes of different ligands, multiple complexes based pharmacophore would be the best since it allows for detecting all the protein–ligand interaction patterns and for evaluating the importance of each protein–ligand interaction. The resulting comprehensive pharmacophore map should contain much more information and be more accurate over others.

As a part of our recent attempts to explore how to generate more accurate and reasonable pharmacophore models, in this account, we shall develop a comprehensive pharmacophore map of CDK2 inhibitors by utilizing all available CDK2–inhibitor complex structures. We chose CDK2 here since it has been identified as an important target for anti-proliferative drug design [18–22] and a wealth of CDK2–inhibitor complexes have been publicly reported. At the time we prepared this manuscript, a total of 124 crystal structures of CDK2–inhibitor complexes were available from the protein data bank (PDB) [23], which are summarized in Table 1. The constructed comprehensive pharmacophore map will then be used to compare previously reported ligand-based pharmacophore models, which purpose is to find out their relationship and to check their validity. Finally, one most-frequent-feature pharmacophore model will be generated based on the most frequent features of the comprehensive pharmacophore map. Validation of the obtained pharmacophore model will also be carried out. It is hoped that the present study can provide valuable information for analyzing the pharmaco-

Table 1
List of 124 CDK2 protein–ligand complexes used in this study

No.	PDB	Resolution (Å)	Ligand	Release
1	1AQ1	2.00	STU	12 November 1997
2	1CKP	2.05	PVB	13 January 1999
3	1DI8	2.20	DTQ	29 November 2000
4	1DM2	2.10	HMD	31 May 2000
5	1E1V	1.95	CMG	10 May 2001
6	1E1X	1.85	NW1	10 May 2001
7	1E9H	2.50	INR	11 October 2001
8	1FVT	2.20	106	17 January 2001
9	1FVV	2.80	107	17 January 2001
10	1G5S	2.61	I17	2 November 2001
11	1GIH	2.80	1PU	6 February 2002
12	1GII	2.00	1PU	6 February 2002
13	1GIJ	2.20	2PU	6 February 2002
14	1GZ8	1.30	MBP	12 June 2003
15	1H00	1.60	FAP	11 July 2003
16	1H01	1.79	FBL	11 July 2003
17	1H07	1.85	MFP	11 July 2003
18	1H08	1.80	BWP	11 July 2003
19	1H0V	1.90	UN4	27 June 2003
20	1H0W	2.10	207	27 June 2003
21	1H1P	2.10	CMG	19 September 2002
22	1H1Q	2.50	2A6	19 September 2002
23	1H1R	2.00	6CP	19 September 2002
24	1H1S	2.00	4SP	19 September 2002
25	1JSV	1.96	U55	29 August 2001
26	1JVP	1.53	LIG	21 December 2001
27	1KE5	2.20	LS1	14 May 2002
28	1KE6	2.00	LS2	14 May 2002
29	1KE7	2.00	LS3	14 May 2002
30	1KE8	2.00	LS4	14 May 2002
31	1KE9	2.00	LS5	14 May 2002
32	1OGU	2.60	ST8	2 September 2003
33	1OI9	2.10	N20	13 July 2004
34	1OIQ	2.31	HDU	4 September 2003
35	1OIR	1.91	HDY	4 September 2003
36	1OIT	1.60	HDT	4 September 2003
37	1OIU	2.00	N76	13 July 2004
38	1OIY	2.40	N41	13 July 2004
39	1P2A	2.50	5BN	15 July 2003
40	1P5E	2.22	TBS	1 July 2003
41	1PF8	2.51	SU9	23 December 2003
42	1PKD	2.30	UCN	24 June 2003
43	1PXI	1.95	CK1	9 December 2003
44	1PXJ	2.30	CK2	9 December 2003
45	1PXK	2.80	CK3	9 December 2003
46	1PXL	2.50	CK4	9 December 2003
47	1PXM	2.53	CK5	13 April 2004
48	1PXN	2.50	CK6	13 April 2004
49	1PXO	1.96	CK7	13 April 2004
50	1PXP	2.30	CK8	13 April 2004
51	1PYE	2.00	PM1	13 July 2004
52	1R78	2.00	FMD	20 January 2004
53	1URW	1.60	I1P	23 April 2004
54	1V1K	2.31	3FP	4 May 2004
55	1VYW	2.30	292	10 June 2004

Table 1 (*Continued*)

No.	PDB	Resolution (Å)	Ligand	Release
56	1VYZ	2.21	N5B	17 June 2004
57	1W0X	2.20	OLO	14 January 2005
58	1W8C	2.05	N69	30 August 2006
59	1WCC	2.20	CIG	27 January 2005
60	1Y8Y	2.00	CT7	8 February 2005
61	1Y91	2.15	CT9	8 February 2005
62	1YKR	1.80	628	24 January 2006
63	2A0C	1.95	CK9	24 January 2006
64	2A4L	2.40	RRC	3 October 2006
65	2B52	1.88	D42	11 October 2005
66	2B53	2.00	D23	11 October 2005
67	2B54	1.85	D05	11 October 2005
68	2B55	1.85	D31	11 October 2005
69	2BHE	1.90	BRY	9 March 2005
70	2BHH	2.60	RYU	9 March 2005
71	2BKZ	2.60	SBC	8 March 2006
72	2BPM	2.40	529	8 December 2005
73	2BTR	1.85	U73	9 November 2005
74	2BTS	1.99	U32	9 November 2005
75	2C4G	2.70	514	23 November 2005
76	2C5N	2.10	CK8	1 March 2006
77	2C5O	2.10	CK2	1 March 2006
78	2C5P	2.30	CK7	1 March 2006
79	2C5V	2.90	CK4	1 March 2006
80	2C5X	2.90	MTW	1 March 2006
81	2C5Y	2.25	MTW	1 March 2006
82	2C68	1.95	CT6	7 December 2005
83	2C69	2.10	CT8	7 December 2005
84	2C6I	1.80	DT1	7 December 2005
85	2C6K	1.90	DT2	7 December 2005
86	2C6L	2.30	DT4	7 December 2005
87	2C6M	1.90	DT5	7 December 2005
88	2C6O	2.10	4SP	7 December 2005
89	2C6T	2.61	DT5	7 December 2005
90	2CLX	1.80	F18	1 November 2006
91	2DS1	2.00	1CD	19 June 2007
92	2DUV	2.20	371	27 January 2007
93	2EXM	1.80	ZIP	27 December 2005
94	2FVD	1.85	LIA	10 October 2006
95	2G9X	2.50	NU5	23 May 2006
96	2I40	2.80	BLZ	10 October 2006
97	2IW6	2.30	QQ2	6 September 2006
98	2IW8	2.30	4SP	6 September 2006
99	2IW9	2.00	4SP	6 September 2006
100	2J9M	2.50	PY8	6 November 2007
101	2R3F	1.50	SC8	22 January 2008
102	2R3G	1.55	SC9	22 January 2008
103	2R3H	1.50	SCE	22 January 2008
104	2R3I	1.28	SCF	22 January 2008
105	2R3J	1.65	SCJ	22 January 2008
106	2R3K	1.70	SCQ	22 January 2008
107	2R3L	1.65	SCW	22 January 2008
108	2R3M	1.70	SCX	22 January 2008
109	2R3N	1.63	SCZ	22 January 2008
110	2R3O	1.80	2SC	22 January 2008
111	2R3P	1.66	3SC	22 January 2008
112	2R3Q	1.35	5SC	22 January 2008
113	2R3R	1.47	6SC	22 January 2008
114	2UUE	2.06	MTZ	27 March 2007
115	2UZB	2.70	C75	26 June 2007
116	2UZD	2.72	C85	26 June 2007
117	2UZE	2.40	C95	26 June 2007
118	2UZL	2.40	C94	26 June 2007
119	2UZN	2.30	C96	26 June 2007
120	2UZO	2.30	C62	26 June 2007
121	2V0D	2.20	C53	26 June 2007
122	3BHV	2.10	VAR	12 February 2008
123	3BHU	2.30	MHR	12 February 2008
124	3BHT	2.00	MFR	12 February 2008

phoric interactions of a protein and different inhibitors, and for developing a multicomplex-based pharmacophore model that can be used for virtual screening to discover novel potential lead compounds.

2. Materials and methods

2.1. Protein and ligand preparation

The aim of this study was to use the structural data to generate a structure-based pharmacophore for the ATP-competitive inhibitors of CDK2; hence only *holo* structures with CDK2 inhibitors targeted against the ATP binding pocket were used. The crystal structures with ATP, the natural ligand of CDK2, were not used in the analyses in order to avoid the unnecessary noise likely to be introduced into the pharmacophore model. 124 X-ray crystallography structures of CDK2 in complex with chemical inhibitors were obtained from the protein data bank (PDB) [23]. Due to the high mobility of water molecules and ions, and their less conservative location in the active pocket [24], it is difficult and complicated to locate the pharmacophore features accurately for the cases where they exist. Thus they were all removed from the structures. Any structural issue of the protein and the ligand was carefully examined upon visual inspection.

An important part of the present analysis involved transforming the coordinates of each receptors and ligands to a common reference frame for further analysis. Three multiple protein structure alignment tools, including Align3D in Modeller [25], MUSTANG [26], and 3DMA in Accelrys Discovery Studio [27], were used in this study to examine whether different algorithms influence our structural alignment or not. The crystal structure with PDB code "2C5O", the one with the high crystallographic resolution and no missing residues, was arbitrarily taken as the reference structure.

2.2. Generation of multicomplex-based comprehensive pharmacophore map

The whole process of generating the multicomplex-based comprehensive pharmacophore map is illustrated in Fig. 2 and detailed as follows. The software LigandScout 1.03 [28], which uses algorithm that allows the automatic construction of 3D pharmacophore from the structural data of protein–ligand complex, was used to generate 124 individual complex-based pharmacophore models based on the previously aligned structures. For the purposes of creating multicomplex-based comprehensive map, all the pharmacophore features identified by LigandScout were clustered according to their interaction pattern with the receptor. Hydrogen bond donor and acceptor features were clustered in term of the protein atom with which they were formed. For hydrophobic, positive and negative ionizable, and ring aromatic features, density-based clustering methods was used. Excluded volume features were clustered based on the protein atom that they represent. In-house software was developed with the aim of speeding up and automating the above process. The cluster centers were identified using the Accelrys Discovery Studio [27]. The model obtained was further refined by the modification of the constraint tolerance of the spheres in accordance with the default values of Catalyst software [29].

2.3. Pharmacophore model validation

A total of 194 unique CDK2 inhibitors (named *external active dataset*) with diverse sizes and chemical characteristics were obtained from the published literatures [15–17,30–33] and from the MDL Drug Data Report database (MDDR, version 2007.2) [34] (see Table s2 in Supporting Information).

Verdonk et al. had proposed a method to eliminate biases in lower dimensions in the validation of structure-based virtual screening approaches [35]. In this study, we have taken the similar

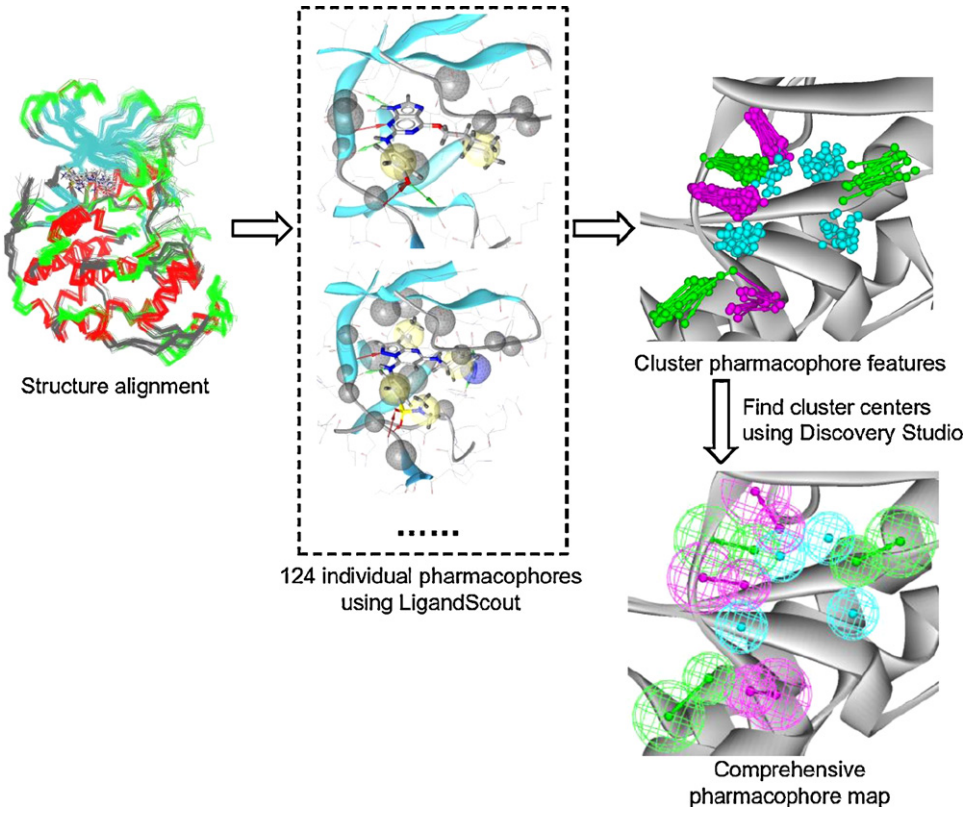

Fig. 2. Flowchart of the generation of the multicomplex-based comprehensive pharmacophore map. The complex-based pharmacophore models were generated using LigandScout based on the previously aligned protein structures. All the pharmacophore features were then clustered, and the cluster centers were identified by Discovery Studio.

protocol to prevent obtaining a high (but meaningless) enrichment factor. All validations presented here are carried out against a *focused inactive dataset* that contains 300 compounds with experimentally confirmed no CDK2 inhibition activities [30]. Here, we used three simple 1D physicochemical properties: (1) number of hydrogen-bond acceptors, (2) number of hydrogen-bond donors, and (3) molecular weight, to ensure that for each selected inactive compound the 1D properties are similar to those of the known active compounds (see Table s3 in Supporting Information).

The molecular structures were prepared using SciTegic Pipeline Pilot 6.1.5 [36]. Multiple conformations of each compound were generated using the FAST conformational search protocol implemented in Catalyst [29] with an energy threshold of 20 kcal/mol and a maximum of 250 conformers. The FAST method uses the Poling algorithm [37] that generates energetically feasible diverse conformations. It is demonstrated that the FAST generation algorithm with the energy cut off is appropriate for finding the protein-bound ligand conformation within reasonable calculation time [11–13]. All compounds were fit to the pharmacophore models using the Best Fit Compare algorithm in Catalyst [29], which considers the conformational flexibility by modifying the molecular conformation within a given energy threshold during the computation [4,8]. The best-fit value of the molecule to the respective pharmacophore was calculated. Molecules are ranked based on their fit value computed, and the receiver operating characteristic (ROC) curve, in which the percentage of known CDK2 inhibitors (true positives) identified by the model is plotted against the percentage of false positives found, was generated to evaluate the performance of the pharmacophore model. The ideal model would be at the point (0, 100) with all of the true positives identified with no false positives found, whereas a random predictor would lie on a line of slope equal to one.

Furthermore, a relationship between the molecule activities and the corresponding fit values was computed using the "Regress Hypothesis" method in Catalyst [29] in order to make the pharmacophore model can be used to quantitatively estimate the activities of the molecules screened from dataset. The predictive capacity of the model is validated with the external active dataset.

3. Results and discussion

3.1. Multicomplex-based comprehensive pharmacophore map

124 X-ray crystallography structures of CDK2 in complex with small molecular inhibitors were used to construct pharmacophore. It is illuminated that the resulting superposition from the three alignment programs are not significantly different (see Table s1 in Supporting Information), and only these based on Modeller [25] were reported below.

The detected pharmacophore features as well as their statistical frequency, which measures how many complexes a given pharmacophore feature can be found in, are listed in Table 2. One can see that there are 24 pharmacophore features, including 6 hydrogen bond acceptors (A1–A6), 8 hydrogen bond donors (D1–D8), 5 hydrophobic features (H1–H5), 3 ring aromatic features (RA1–RA3), and 2 positive ionizable features (PI1 and PI2).

Of the 24 detected pharmacophore features, 7 features (A1, A2, D1, D2, H1, H2, and H3) were found to present in the 124 complexes with more than 25% probability. It is believed that the pharmacophore features, which present in the complexes with a high probability, are likely to be more important than features that exhibiting a low probability. The most frequently occurred feature A1 (91.94%) was mapped to the hydrogen bond acceptor of the

Table 2
Spreading of comprehensive pharmacophore map features

No.	Feature Name	Id	Count	Statistical frequency (%)	Interaction
1	HBA-F 1	A1	114	91.94	Leu83:N
2	HBA-F 2	A2	32	25.81	Lys89:NZ
3	HBA-F 3	A3	29	23.39	Asp86:N
4	HBA-F 4	A4	28	22.58	Lys33:NZ
5	HBA-F 5	A5	7	5.65	Asp145:N
6	HBA-F 6	A6	1	0.81	Gln131:NE
7	HBD 1	D1	93	75.00	Leu83:O
8	HBD 2	D2	48	38.71	Glu81:O
9	HBD 3	D3	13	10.48	Asp86:OD
10	HBD 4	D4	12	9.68	Asp145:OD
11	HBD 5	D5	5	4.03	Gln131:O
12	HBD 6	D6	3	2.42	Glu51:OE
13	HBD 7	D7	2	1.61	Ile10:O
14	HBD 8	D8	2	1.61	Asn132:OD
15	Hydrophobic 1	H1	85	68.55	Ile10, Phe82, Gln85
16	Hydrophobic 2	H2	62	50.00	Val64, Phe80, Ala144
17	Hydrophobic 3	H3	58	46.77	Val18
18	Hydrophobic 4	H4	28	22.58	Ala31, Leu134
19	Hydrophobic 5	H5	15	12.10	Ala144
20	Ring Aromatic 1	RA1	11	8.87	Phe80
21	Ring Aromatic 2	RA2	10	8.06	Phe82, His84
22	Ring Aromatic 3	RA3	10	8.06	Lys33
23	Positive Ionizable 1	PI1	5	4.03	Asp145
24	Positive Ionizable 2	PI2	3	2.42	Asp86

ligand, whose partner donor is the backbone amino nitrogen of Leu83 locating in the hinge region that links the two lobes of the kinase (see Fig. 3a and c). The second remarkable feature is D1 (75.00%) that represents the hydrogen bond donor whose partner acceptor is the backbone carbonyl oxygen of Leu83 (Fig. 3a and c). The third feature H1 (68.55%, Fig. 3a) corresponds to the hydrophobic region formed by residues Ile10, Phe82, and Gln85 at the solvent-accessible region. The fourth one H2 (50.00%) represents the hydrophobic region formed by residues Val64, Phe80, and Ala144 at the small-buried region. The fifth one H3 was mapped to the hydrophobic region formed by residue Val18 in the ribose/phosphate binding site. The sixth feature D2 stands for the hydrogen bond donor interaction with the backbone carbonyl oxygen of Glu81. The seventh feature A2 represents the hydrogen bond acceptor whose partner donor is the side chain nitrogen of Lys89, locating at the solvent-accessible region that cannot be utilized by ATP (Fig. 3a and c).

For a full pharmacophore map, it is also important to include excluded volume features, which reflect potential steric restriction and correspond to the positions that are inaccessible to any

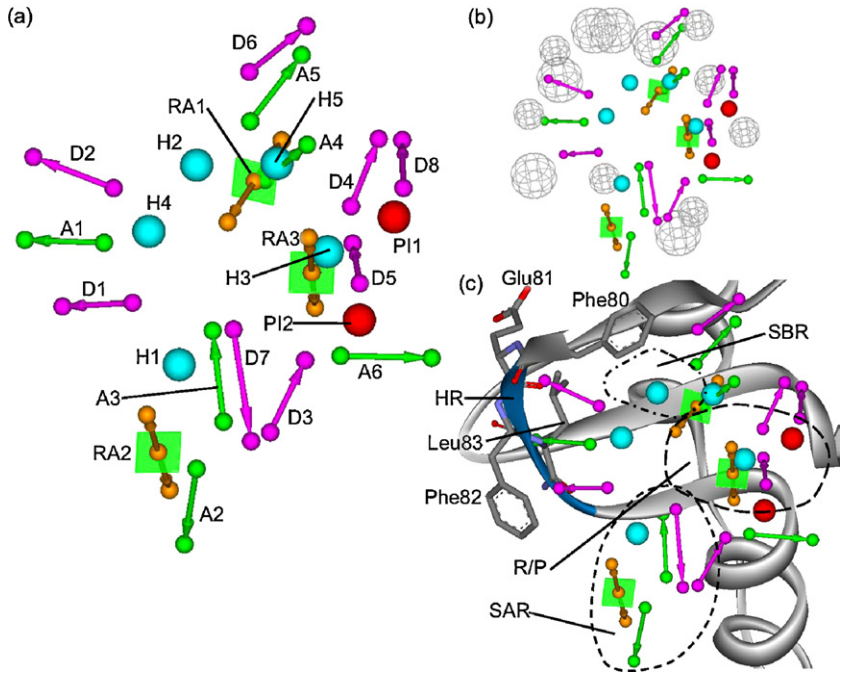

Fig. 3. (a) All the chemical features of the multicomplex-based comprehensive pharmacophore map. (b) The comprehensive map with excluded volume features. For clarity, some excluded volume features were not drawn. (c) View looking into the ATP-binding site of CDK2 with comprehensive map from the N-terminal domain. The Cα ribbon representation of the hinge region (HR) is colored in blue. (SBR, small buried region; R/P, ribose/phosphate binding site; SAR, solvent-accessible region.)

potential ligand. 21 excluded volume features were found in the ATP-binding site, which spaces were occupied by residues Ile10, Gly11, Glu12, Gly13, Val18, Ala31, Lys33, Val64, Phe80, Phe82, Leu83, His84, Gln85, Asp86, Lys88, Lys89, Gln131, Asn132, Leu134, Ala144, Asp145. Comprehensive pharmacophore map involving excluded volume spheres has been shown in Fig. 3b.

3.2. Comparisons with ligand-based pharmacophore models

As stated in the Introduction section, Hecker et al. [15], Toba et al. [16] and Vadivelan et al. [17] independently reported three ligand-based pharmacophore models for CDK2 inhibitors. However, it was found that they were different from each other (Fig. 1). In this section, a comparison will be made between these models and our comprehensive pharmacophore map. Firstly, the three ligand-based pharmacophore models were regenerated according to the methods described in the literatures [15–17].

Hecker et al. [15] generated a 4-feature-shape hypothesis and a 5-feature hypothesis firstly, and then the 2 were merged into a new 5-feature model. The final Hecker model consists of a hydrogen bond acceptor, a hydrogen bond donor, two hydrophobic groups and a ring aromatic feature (Fig. 1a). An alignment of the Hecker model and ours shows that the hydrogen bond acceptor feature of Hecker model corresponds to the feature A1 of our comprehensive map that reflects the interaction between ligand and hinge region. The hydrogen bond donor feature in Hecker model was mapped to the feature D5 that represents the interaction with the backbone oxygen of Gln131. One of the hydrophobic features was mapped to the H1 feature at the solvent-accessible region, and the other one corresponds to the feature H3 in the ribose/phosphate binding site. The ring aromatic feature in Hecker model is stacked directly against the residues Ala31 and Leu134, and was mapped close to the feature H4 (Table 3). The difference of the chemical feature in this position between the ligand-based pharmacophore model and multicomplex-based pharmacophore map (i.e., one is ring aromatic feature and one is hydrophobic feature) is mainly due to the distinct methodologies that have been employed. In LigandScout, the pharmacophore feature is added to the model only if a reasonable interaction pattern between the ligand and the receptor is found. In contrast, the pharmacophore hypothesis generated in Catalyst merely includes ligand information.

The pharmacophore hypothesis produced by Toba et al. [16] contains two hydrogen bond donor features and three hydrophobic features (Toba model, Fig. 1b). In comparison with our multicomplex-based comprehensive pharmacophore map, one of the hydrogen bond donor features of Toba model corresponds to the feature D1. The other hydrogen bond donor feature was mapped to the feature D5, which is positioned to interact with the backbone oxygen of Gln131. The three hydrophobic features correspond to the features H1, H2, and H3, respectively (Table 3).

The pharmacophore model generated by Vadivelan et al. [17] consists of two hydrogen bond acceptors, one hydrogen bond donor, and one hydrophobic feature (Vadivelan model, Fig. 1c). Compared with our comprehensive pharmacophore map, one of the hydrogen bond acceptor features corresponds to the feature A3 that is positioned to interact with the backbone nitrogen of Asp86. The other one corresponds to the feature A4 that reflects the interaction with the amino group of Lys33. The hydrogen bond donor feature corresponds to the feature D1. The hydrophobic feature was mapped to the feature H3 (Table 3).

In summary, for each pharmacophore feature in the ligand-based pharmacophore models, one can find a corresponding feature, indicating that our comprehensive pharmacophore map contains much more information over all the three ligand-based models. It also implied that each of the ligand-based pharmacophore models is just a reduced version of our comprehensive map. In addition, it should be noticed that some conserved features with high statistical frequency were missed in these ligand-based pharmacophore models, such as the feature D2 (statistical frequency: 38.71%) and the feature A2 (statistical frequency: 25.81%). This might limit the predictive ability of these ligand-based models.

3.3. Most-frequent-feature pharmacophore model

The comprehensive pharmacophore map obtained initially is too restrictive and not suitable for the virtual screening since it contains a large number of chemical features and the fit of a

Table 3
Comparison of pharmacophore model features

No.	Pharmacophore features	Comprehensive pharmacophore map	Hecker model	Toba model	Vadivelan model	Most-frequent-feature pharmacophore model
1	A1	✓	✓			✓
2	A2	✓				✓
3	A3	✓			✓	
4	A4	✓			✓	
5	A5	✓				
6	A6	✓				
7	D1	✓		✓	✓	✓
8	D2	✓				✓
9	D3	✓				
10	D4	✓				
11	D5	✓	✓	✓		
12	D6	✓				
13	D7	✓				
14	D8	✓				
15	H1	✓	✓	✓		✓
16	H2	✓		✓		✓
17	H3	✓	✓	✓	✓	✓
18	H4	✓	✓			
19	H5	✓				
20	RA1	✓				
21	RA2	✓				
22	RA3	✓				
23	PI1	✓				
24	PI2	✓				

molecule to such a pharmacophore is still out of reach for today's state-of-the-art of the computational tools. A correctly reduced pharmacophore model would be much more preferred in terms of practical application. However, it is still a hard task to determine which feature should be involved in the reduced model. A feasible solution is to select the most frequent features that were recognized as the features important to the activity of the CDK2 inhibitors. According to our experience, the top ranked seven features (A1, A2, D1, D2, H1, H2, and H3), which were found to present in the 124 complexes with more than 25% probability, would be more appropriate in practice, and consequently they were selected from the comprehensive pharmacophore map and were merged to generate a most-frequent-feature pharmacophore model (Fig. 4a).

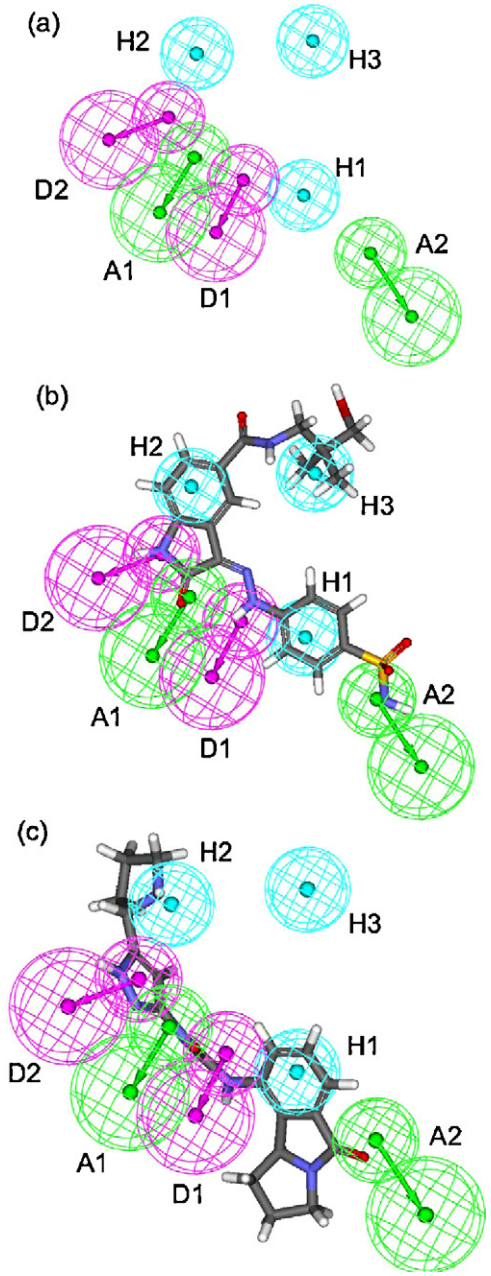

Fig. 4. (a) The most-frequent-feature pharmacophore model. Example of (b) high-active molecule Compound-079 and (c) low-active molecule Compound-092 aligned with the most-frequent-feature pharmacophore model.

3.4. Pharmacophore model validation

Subsequently, the most-frequent-feature pharmacophore model was validated by using it to screen against an external test dataset that includes known active and inactive compounds, and by using it to predict the activities of CDK2 inhibitors in the external active dataset.

The most-frequent-feature pharmacophore model was firstly screened against external test dataset, which includes the external active dataset containing 194 known CDK2 inhibitors (see Table s2 in Supporting Information) and the focused inactive dataset consisting of 300 noninhibitors (see Table s3 in Supporting Information). The receiver operating characteristic (ROC) curve, which was introduced very recently by Triballeau et al. [38], was used to estimate the performance of the pharmacophore model. The best model identifies the greatest number of true positives and the least number of false positives. The performance of our pharmacophore model at discriminating known inhibitors versus a focused dataset of noninhibitors is shown in Fig. 5a. It is obvious that the most-frequent-feature pharmacophore model was very successful at differentiating between these two populations. It identifies 75.3% of the true positives and only 5.0% of the false positives. The ROC curve of our model has a very steep beginning, which means that the number of true positives can be sacrificed to reduce the amount of false positives when screening large databases of compounds. Considering that the false positive rate is reduced to 1.0%, the most-frequent-feature pharmacophore model can retrieve 57.7% of the true active CDK2 inhibitors. Furthermore, in order to evaluate the efficiency of the most-frequent-feature pharmacophore in virtual screening of a large

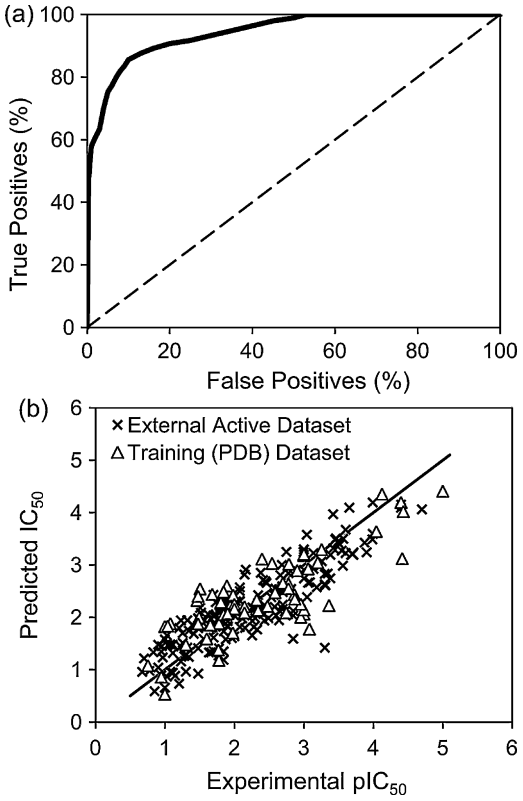

Fig. 5. (a) Receiver operator characteristic (ROC) curves generated from screening the external test dataset, which includes 194 known CDK2 inhibitors and 300 inactive molecules. (b) The correlation between the experimental and predicted activities (pIC_{50} values) obtained from the most-frequent-feature pharmacophore model.

database containing diverse compounds, the same validation protocol described above was applied to screen against the MDL Comprehensive Medicinal Chemistry database (CMC, version 2007.1) consisting of 8896 compounds. Our pharmacophore model displays a superior performance with 76.3% retrieval efficiency when the false-positive tolerance is only 1.0% (see Figure s1 in Supporting Information). The enrichment factor (E) calculated according to Gopalakrishnan et al. [39] was found to be 12.2, indicating that it is 12 times more probable to pick an active compound from the database than an inactive one. The validation has shown that our model is capable of being used in virtual screening of large databases.

Fig. 4b and c represents the most-frequent-feature pharmacophore model aligned with the high- and low-active molecules (Compound-079 and Compound-092, see Table s2 in Supporting Information) with IC_{50} values of 6.8 nM and 25 μM, respectively. It is shown that the high-active Compound-079 fits all the pharmacophore features in the given relative position in space with high fit value. On the other hand, the molecule Compound-092 with low activity does not fulfill these criteria and fits the most-frequent-feature pharmacophore poorly and partially. It is implied that the most-frequent-feature model provided important information on the structural and chemical properties of the CDK2 inhibitors.

At last, the most-frequent-feature pharmacophore model is used to calculate activities for both the training (PDB) dataset and the external active dataset molecules. The plot showing the correlation between the actual and predicted activity for the molecules is given in Fig. 5b. The experimental and predicted activity values for the molecules are detailed in Supporting Information Tables s2 and s4. The results gave a Pearson's correlation coefficient with the value of 0.879 for the external active dataset molecules. This indicates that the most-frequent-feature pharmacophore model generated is capable of predicting the activity of external compounds with reasonable accuracy.

4. Conclusion

In this study, we have utilized a collection of 124 crystal structures of human CDK2 bound to a variety of inhibitors to generate a comprehensive pharmacophore map by using multi-complex-based method. The comprehensive pharmacophore map was used to compare previously reported three ligand-based pharmacophore models. It was clearly demonstrated that each one of the ligand-based pharmacophore models is just a subset of the comprehensive map. Our multicomplex-based comprehensive map has managed to involve almost all the pharmacophore features for CDK2–inhibitor interactions. Based on the statistical frequency value, it also provides us insight into how important each pharmacophore feature is. Thus the most-frequent-feature pharmacophore model can be constructed based on the most frequent chemical features of the comprehensive pharmacophore map. The receiver operating characteristic (ROC) curve has indicated that the most-frequent-feature pharmacophore model is able to differentiate between known CDK2 inhibitors and the compounds in the focused inactive dataset. In addition, it has also been validated that the most-frequent-feature pharmacophore model is capable of predicting the activities of a wide variety of CDK2 inhibitors in the external active dataset.

In conclusion, the work conducted here has provided an approach to generate a comprehensive pharmacophore map and a most-frequent-feature pharmacophore model based on a set of crystal structures of protein–ligand complex. The pharmacophore model obtained can be used to retrieve potential inhibitors from a database in virtual screening and to evaluate the newly engineered compound in *de novo* design. It is expected that the information provided here is helpful for the study toward more accurate pharmacophore modeling.

Acknowledgements

We do apologize to the many research groups whose work could not be cited here due to space limitations. We gratefully acknowledge the Inte:Ligand GmbH for providing the academic license of the LigandScout 1.03. And we also thank the National Natural Sciences Foundation of China (Grant number:30772651) for financial support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.07.004.

References

[1] C.G. Wermuth, Pharmacophores: historical perspective and viewpoint from a medicinal chemist, in: T. Langer, R.D. Hoffmann (Eds.), Pharmacophores and Pharmacophore Searches, Wiley–VCH, 2006, pp. 3–13.
[2] S. Ekins, J. Mestres, B. Testa, In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling, Br. J. Pharmacol. 152 (2007) 9–20.
[3] S. Ekins, J. Mestres, B. Testa, In silico pharmacology for drug discovery: applications to targets and beyond, Br. J. Pharmacol. 152 (2007) 21–37.
[4] K. Poptodorov, T. Luu, R.D. Hoffmann, Pharmacophore model generation software tools, in: T. Langer, R.D. Hoffmann (Eds.), Pharmacophores and Pharmacophore Searches, Wiley–VCH, 2006, pp. 17–47.
[5] O. Dror, A. Shulman-Peleg, R. Nussinov, H.J. Wolfson, Predicting molecular interactions in silico. I. A guide to pharmacophore identification and its applications to drug design, Curr. Med. Chem. 11 (2004) 71–90.
[6] S.A. Khedkar, A.K. Malde, E.C. Coutinho, S. Srivastava, Pharmacophore modeling in drug discovery and development: an overview, Med. Chem. 3 (2007) 187–197.
[7] G. Wolber, T. Seidel, F. Bendix, T. Langer, Molecule-pharmacophore superpositioning and pattern matching in computational drug design, Drug Discov. Today 13 (2008) 23–29.
[8] Y. Kurogi, O.F. Güner, Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, Curr. Med. Chem. 8 (2001) 1035–1055.
[9] O.F. Güner, History and evolution of the pharmacophore concept in computer-aided drug design, Curr. Topics Med. Chem. 2 (2002) 1321–1332.
[10] Z. Xiao, S. Varma, Y.-D. Xiao, A. Tropsha, Modeling of p38 mitogen-activated protein kinase inhibitors using the Catalyst HypoGen and k-nearest neighbor QSAR methods, J. Mol. Graphics Modell. 23 (2004) 129–138.
[11] J. Kirchmair, C. Laggner, G. Wolber, T. Langer, Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms, J. Chem. Inf. Model. 45 (2005) 422–430.
[12] J. Kirchmair, G. Wolber, C. Laggner, T. Langer, Comparative performance assessment of the conformational model generators omega and catalyst: a large-scale survey on the retrieval of protein-bound ligand conformations, J. Chem. Inf. Model. 46 (2006) 1848–1861.
[13] J. Kirchmair, S. Ristic, K. Eder, P. Markt, G. Wolber, C. Laggner, T. Langer, Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches, J. Chem. Inf. Model. 47 (2007) 2182–2196.
[14] R. Kristam, V.J. Gillet, R.A. Lewis, D. Thorner, Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst, J. Chem. Inf. Model. 45 (2005) 461–476.
[15] E.A. Hecker, C. Duraiswami, T.A. Andrea, D.J. Diller, Use of catalyst pharmacophore models for screening of large combinatorial libraries, J. Chem. Inf. Comput. Sci. 42 (2002) 1204–1211.
[16] S. Toba, J. Srinivasan, A.J. Maynard, J. Sutter, Using pharmacophore models to gain insight into structural binding and virtual screening: an application study with CDK2 and human DHFR, J. Chem. Inf. Model. 46 (2006) 728–735.
[17] S. Vadivelan, B.N. Sinha, S.J. Irudayam, S.A.R.P. Jagarlapudi, Virtual screening studies to design potent CDK2-cyclin A inhibitors, J. Chem. Inf. Model. 47 (2007) 1526–1535.
[18] A.M. Senderowicz, Small molecule modulators of cyclin-dependent kinases for cancer therapy, Oncogene 19 (2000) 6600–6606.
[19] T.M. Sielecki, J.F. Boylan, P.A. Benfield, G.L. Trainor, Cyclin-dependent kinase inhibitors: useful targets in cell cycle regulation, J. Med. Chem. 43 (2000) 1–18.
[20] P.M. Fischer, D.P. Lane, Inhibitors of cyclin-dependent kinases as anti-cancer therapeutics, Curr. Med. Chem. 7 (2000) 1213–1245.
[21] P.L. Toogood, Cyclin-dependent kinase inhibitors fortreating cancer, Med. Res. Rev. 21 (2001) 487–498.
[22] A. Huwe, R. Mazitschek, A. Giannis, Small molecules as inhibitors of cyclin-dependent kinases, Angew. Chem., Int. Ed. 42 (2003) 2122–2138.

[23] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucl. Acids Res. 28 (2000) 235–242.
[24] J.E. Ladbury, Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design, Chem. Biol. 3 (1996) 973–980.
[25] M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali, Comparative protein structure modeling of genes and genomes, Annu. Rev. Biophys. Biomol. Struct. 29 (2000) 291–325.
[26] A.S. Konagurthu, J.C. Whisstock, P.J. Stuckey, A.M. Lesk, MUSTANG: a multiple structural alignment algorithm, Proteins 64 (2006) 559–574.
[27] Discovery Studio, Version 1.7, Accelrys, Inc., San Diego, CA, 2006.
[28] G. Wolber, T. Langer, LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters, J. Chem. Inf. Model. 45 (2005) 160–169.
[29] Catalyst, Version 4.11, Accelrys, Inc., San Diego, CA, 2005.
[30] E.K. Bradley, J.L. Miller, E. Saiah, P.D.J. Grootenhuis, Informative library design as an efficient strategy to identify and optimize leads: application to cyclin-dependent kinase 2 antagonists, J. Med. Chem. 46 (2003) 4360–4364.
[31] D.A. Nugiel, A.-M. Etzkorn, A. Vidwans, P.A. Benfield, M. Boisclair, C.R. Burton, S. Cox, P.M. Czerniak, D. Doleniak, S.P. Seitz, Indenopyrazoles as novel cyclin dependent kinase (CDK) inhibitors, J. Med. Chem. 44 (2001) 1334–1336.
[32] R. Lin, P.J. Connolly, S. Huang, S.K. Wetter, Y. Lu, W.V. Murray, S.L. Emanuel, R.H. Gruninger, A.R. Fuentes-Pesquera, C.A. Rugg, S.A. Middleton, L.K. Jolliffe, 1-acyl-1H-[1,2,4]triazole-3,5-diamine analogues as novel and potent anticancer cyclin-dependent kinase inhibitors: synthesis and evaluation of biological activities, J. Med. Chem. 48 (2005) 4208–4211.
[33] X.-J. Chu, W. DePinto, D. Bartkovitz, S.-S. So, B.T. Vu, K. Packman, C. Lukacs, Q. Ding, N. Jiang, K. Wang, P. Goelzer, X. Yin, M.A. Smith, B.X. Higgins, Y. Chen, Q. Xiang, J. Moliterni, G. Kaplan, B. Graves, A. Lovey, N. Fotouhi, Discovery of [4-amino-2-(1-methanesulfonylpiperidin-4-ylamino)pyrimidin-5-yl](2,3-difluoro-6-methoxyphenyl)methanone (R547), a potent and selective cyclin-dependent kinase inhibitor with significant in vivo antitumor activity, J. Med. Chem. 49 (2006) 6549–6560.
[34] MDDR Drug Data Report, MDL Information Systems, Inc., San Leandro, CA, 2007.
[35] M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T.M. Mooij, C.W. Murray, R.D. Taylor, P. Watson, Virtual screening using protein-ligand docking: avoiding artificial enrichment, J. Chem. Inf. Comput. Sci. 44 (2004) 793–806.
[36] SciTegic Pipeline Pilot, version 6.1.5, Accelrys, Inc., San Diego, CA, 2007.
[37] A. Smellie, S.L. Teig, P. Towbin, Poling: promoting conformational variation, J. Comput. Chem. 16 (1995) 171–187.
[38] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4, J. Med. Chem. 48 (2005) 2534–2547.
[39] B. Gopalakrishnan, V. Aparna, J. Jeevan, M. Ravi, G.R. Desiraju, A virtual screening approach for thymidine monophosphate kinase inhibitors as antitubercular agents based on docking and pharmacophore models, J. Chem. Inf. Model. 45 (2005) 1101–1108.