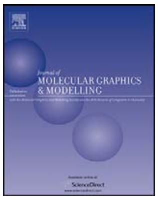
Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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



Generation and validation of the first predictive pharmacophore model for cyclin-dependent kinase 9 inhibitors

Cheng Fang, Zhiyan Xiao*, Zongru Guo

Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicine, Ministry of Education & Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, China

ARTICLE INFO

Article history:
Received 31 October 2010
Received in revised form
21 December 2010
Accepted 10 January 2011
Available online 20 January 2011

Keywords:
CDK9 inhibitors
Rational selection of training sets
Predictive pharmacophore models
Model validation
Virtual screening

ABSTRACT

A three-dimensional (3D) pharmacophore modelling approach was applied to a diverse data set of known cyclin-dependent kinase 9 (CDK9) inhibitors. Diversity sampling and principal components analysis (PCA) were employed to ensure the rational selection of representative training sets. Twelve statistically robust pharmacophore models were generated using the HypoGen algorithm. The resulting models showed high homology and indicated great convergence in ascertaining pharmacophoric features essential for CDK9 inhibitory activity. One of the best models (Hypo 6) was assessed further by external predictive capability, randomization test, as well as its performance in virtual screening. The capability of the resulting models to reliably predict the inhibitory activity of external data sets and discriminate active structures from general databases would assist the identification and optimization of novel CDK9 inhibitors.

1. Introduction

Cyclin-dependent kinases (CDKs) are serine/threonine kinases involved in cell cycle and/or transcriptional control. Currently, there are thirteen known members in the CDK family [1]. As one of the transcriptional CDKs, CDK9 complexes with cyclin T or cyclin K to form the positive transcription elongation factor b (P-TEFb), and promotes efficient mRNA elongation by phosphorylating the carboxy-terminal domain (CTD) of RNA polymerase II (RNAPII). Although P-TEFb has been recognized as a general transcription factor for the productive expression of most genes, it is revealed that P-TEFb is much more profoundly implicated in pathologic cellular processes than in normal cells. Therefore, CDK9 is expected to be a potential drug target for the treatment of cancer and cardiac hypertrophy [2]. In particular, as a kinase discovered in HIV pathogenesis, CDK9 has demonstrated a unique functional role in HIV-1 transcription and replication. The viral transcription would be aborted by negative elongation factors unless the viral protein, transactivator of transcription (Tat), binds to the nascent transactivation-responsive (TAR) RNA and recruits P-TEFb to the HIV-1 LTR promoter, which specifically activates HIV-1 transcription [1,2]. Therefore, inhibition of CDK9 activity would have potential therapeutic applications in the management of HIV infections.

Targeting cellular cofactors is an attractive strategy to develop anti-HIV drugs overriding the multidrug-resistance problem, and thereby novel CDK9 inhibitors are actively pursued in recent anti-HIV researches. Although structurally diverse CDK9 inhibitors have been reported in the literature [2,3], the pharmacophoric features essential for CDK9 inhibitory activity remain ambiguous, and further medicinal chemistry efforts to identify novel and potent CDK9 inhibitors are hampered by the deficiency in conductive information. Recently, the X-ray crystal structures of two CDK9 inhibitors, Flavopiridol and DRB, in complex with P-TEFb have been reported by Baumli et al. [4], which facilitate the understanding of the interaction mode of known CDK9 inhibitors. However, the inadequate resolution (2.8 Å) of the complex structures impairs the reliability and accuracy of structure-based design. Therefore, a ligand-based approach to formulate a generalized pharmacophore is essential and beneficial for the discovery of novel CDK9 inhibitors through virtual screening of available chemical databases.

The HypoGen algorithm is a 3D-QSAR pharmacophore modelling protocol implemented in the software packages of Catalyst and Discovery Studio, which accentuates rational selection and adequate representativeness of the training set molecules used for model generation [5]. We report herein the generation and validation of the first predictive pharmacophore model for CDK9 inhibitors with the HypoGen procedure. To ensure reliable model construction, diversity sampling and subsequent principal components analysis (PCA) were exploited to select representative training sets. The resulting models were subjected to rigorous validation with multiple approaches. These models would assist in

* Corresponding author. Tel.: +86 10 63189228; fax: +86 10 83155752.
E-mail address: xiaoz@imm.ac.cn (Z. Xiao).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.01.003

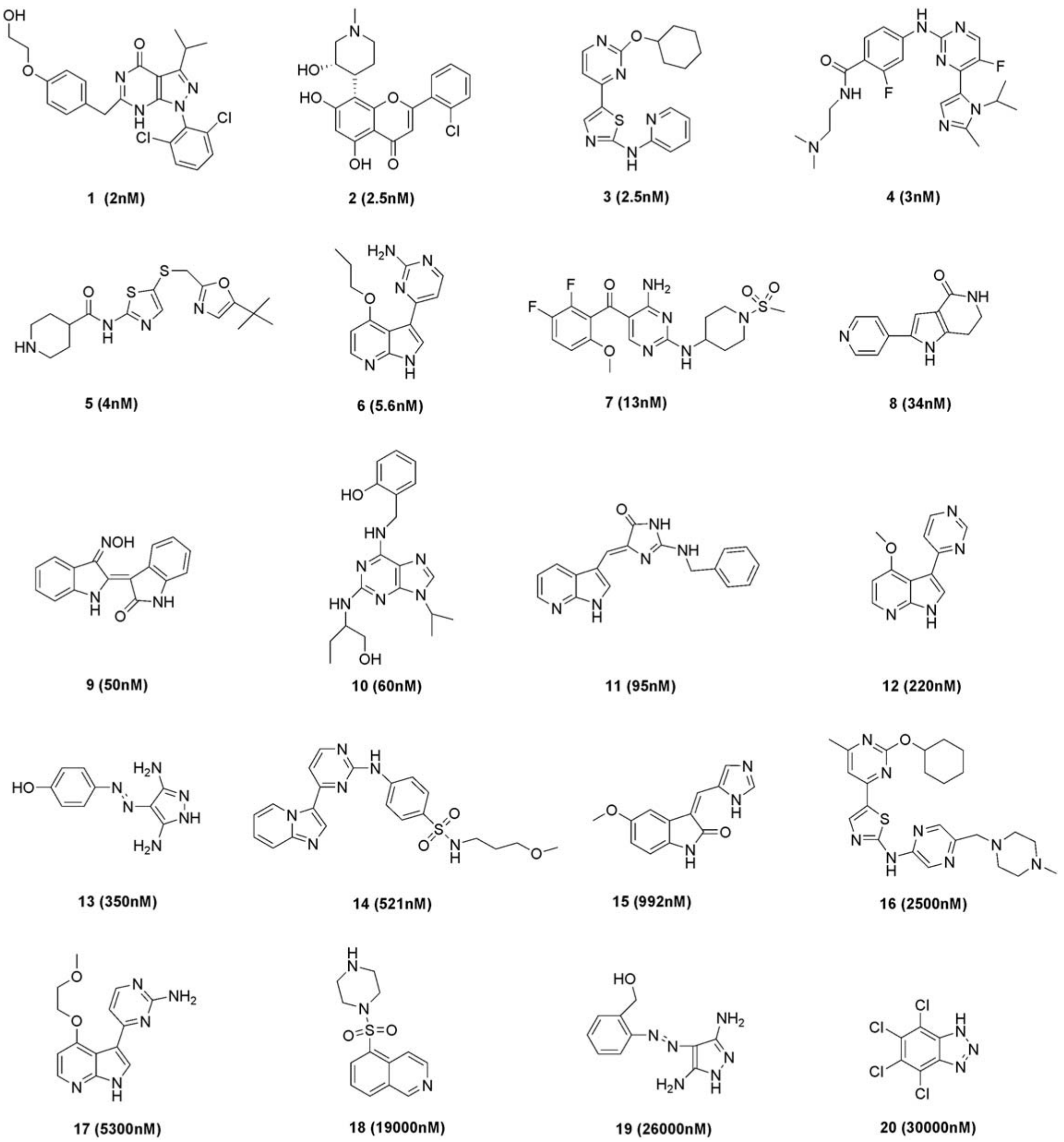


Fig. 1. Structures of representative CDK9 inhibitors.

obtaining the molecular alignments of active conformers to enable further 3D-QSAR approaches, understanding the action mode of different inhibitors and providing generalized queries for identifying novel, potentially active compounds through rapid virtual screening of chemical databases.

2. Materials and methods

2.1. General

Structures were generated, and pharmacophore modelling was performed with Discovery Studio 2.1 (DS 2.1) software package (Accelrys, San Diego, USA) [6]. The default settings were used except as otherwise noted. All calculations were performed on a DELL Precision T5500 workstation.

2.2. The data set

Seventy-five structurally diverse CDK9 inhibitors were collected from published literatures [2,3,7–26]. The activities of these inhibitors were expressed as IC_{50} values for CDK9 inhibition, which range from 2 to 30,000 nM. The structures of representative CDK9 inhibitors were shown in Fig. 1. Detailed structural and activity data of the whole data set were described in Supplementary data.

The structures of all the molecules were built using the 2D-3D sketcher in DS 2.1 program package and minimized using the CHARMM-like force field implemented in the program. The "Diverse Conformation Generation" protocol with "best" conformer generation option was chosen to generate energetically reasonable conformers with sufficient diversity for all the molecules.

2.3. "Poling" algorithm for "Diverse Conformation Generation" [5,27]

The primary objective of the conformation generation process implemented in DS 2.1 is to generate a moderate number of conformers for a given molecule, while adequately covering its conformational space within a defined energy threshold.

A "Poling" algorithm [27] is applied for this purpose, which promotes conformational variation by altering the potential hypersurface of a molecule after each new conformer is generated. Briefly, the conformer generated in the first iteration (C1) will be located at a local minimum. A “pole” will then be added to where C1 stays to change the potential surface, and the conformer generated in the second iteration (C2) is situated at another local minimum on the modified potential surface. Additional “poles” will be subsequently added to generate new conformers following the same procedure. The new conformations will be rejected if their energy exceeds the threshold value or they are too close to an existing conformation. Otherwise, they will be added to the conformation list. The algorithm stops to generate new conformations if the user defined maximal fail count is reached [27]. Thereby, the low-energy conformers produced will not be confined to the energy-valley close to the initial conformation, and broad coverage of the conformational space will be achieved.

2.4. Selection of training and test sets

The HypoGen algorithm develops predictive pharmacophore models with both activity and structural data of the training set molecules [5]. To guarantee the construction of robust models, it is crucial to generate a representative training set with sufficient coverage of both biological and chemical spaces occupied by the original data set. Therefore, diversity sampling of the original data set by considering both activity and structural information is the premise for rational selection of training sets. Furthermore, simultaneous inclusion of several similar compounds in the training sets should be avoided, since it may only provide redundant information and bias the resulting model toward those similar structures. Following these and other guidelines [5], we employed a simple diversity sampling protocol to rationally select representative training sets for pharmacophore generation. Briefly, the original data set of 75 molecules was first divided into three subgroups according to their activity data: the most active subgroup includes those with IC_{50} values lower than 10 nM, the moderately active subgroup contains compounds with IC_{50} values ranging from 10 to 3000 nM, and the rest molecules with IC_{50} values higher than 3000 nM belong to the least active subgroup. The “Find Diverse Molecules” protocol with FCFP_4 fingerprint option, which is inbuilt fingerprints based on circular substructural fragments with a maximum diameter of four bonds [28], was applied to perform diversity sampling on the three subgroups respectively, and training sets with sufficient biological and chemical diversity were thereby generated. The remaining molecules in the original data set were then taken as test sets.

2.5. Principal components analysis (PCA)

The 2D molecular properties for each compound were calculated using the “Calculate Molecular Properties” protocol implemented in the “QSAR” module. PCA method was then applied using the “Calculate Principal Components” protocol in the “Library Analysis” module to extract three principal components.

The program initially generated more than 400 descriptors for each compound and it is predictable that some of the descriptors are highly correlated. Therefore, PCA method was applied to reduce the dimensionality of the descriptor space and alleviate the correlations [29]. Basically, PCA method is a mathematical procedure that converts multiple sets of possibly correlated variables into a few orthogonal “principle components”, which are usually linear combinations of the correlated variables (Eq. (1)) and each corresponds to an axis in a multiple-dimensional space.

$$PC_i = \sum_{j=1}^{\nu} C_{i,j} X_j \tag{1}$$

where PC_i is the ith principle component, $C_{i,j}$ is the coefficient of the variable X_j, and ν is the number of variables.

Since each principle component endeavors to account for the maximum variance in the descriptor space, and in general, only a few principle components may be sufficient to explain a significant proportion of the variation in the descriptor space. For clear graphical representation of the molecular diversity, we extracted three principal components, which account for more than 70% of the variation in the descriptor space, and plotted the molecules as discrete spots in a three-dimensional coordinate system.

2.6. Pharmacophore generation

The “3D-QSAR Pharmacophore Generation” protocol with HypoGen algorithm was used for pharmacophore generation. The features of hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic or hydrophobic aromatic center (Y) and aromatic ring (R) were predefined and the parameter of “Maximum Excluded Volumes” (MEV) was set to 5 or 6, of which an MEV value of 5 is recommended by the tutorials for DS 2.1 software package [6].

2.7. Pharmacophore validation

A valid pharmacophore model should be not only statistically robust, but also predictive to internal and external data sets. The resultant pharmacophore model was assessed by cost value analysis and randomization tests, and further validated by training and test sets prediction as well as enrichment factor and hit rate in virtual screening. Its capability to reliably predict external data sets and discriminate active inhibitors from other molecules is critical criteria for high-quality models.

2.7.1. Cost value analysis [30]

The HypoGen algorithm calculates a series of statistical parameters for each possible pharmacophore model, including total cost, fixed cost, null cost, correlation coefficient and root mean square difference (RMSD) between the predicted and experimental activities of the training set molecules. The statistical significance for these parameters is elaborated and readily referable elsewhere [6,30].

All the possible pharmacophore models for each training set were ranked according to the hypothesis cost (total cost), and the top-ranked models with the lowest total cost values were selected as the best models. Furthermore, the cost difference between null and total costs, the configuration cost, the correlation coefficient and root mean square difference (RMSD) between the predicted and the experimental activities of the training set molecules were regarded as key parameters for the evaluation of model quality.

2.7.2. Internal and external prediction

Conformer assembly of the training and test set molecules was mapped onto the pharmacophore model with “Ligand Pharmacophore Mapping” protocol implemented in the program, and “Flexible” fitting option was applied. Such operations adapt to molecular flexibility by producing diverse conformations for each molecule within an energy threshold of 20 kcal/mol above the global energy minimum. The “Fit” values (i.e. how well the ligands are matched to the features of a pharmacophore model) were then correlated with the activity data to allow quantitative estimation of training and test sets activities.

2.7.3. Randomization test

Randomization test was further utilized to assess the statistical relevance of the models generated and assure that the correlation recognized was true correlation rather than chance correlation. Briefly, the experimental activity data of the training set molecules were shuffled randomly to provide "random" data sets, which were subjected to pharmacophore modelling with all the parameters for the initial model generation adopted. Thirty random pharmacophore models were obtained and their total cost values were compared with those of the original model. Five of these models were further evaluated for their predictive capability and their correlation coefficients between the predicted and the experimental activities for both training and test sets were recorded.

2.7.4. Enrichment factor and hit rate validation

In order to determine the capability of the pharmacophore models to discriminate active compounds from other molecules in virtual screening, an artificial database with 1500 compounds was generated, which merged 25 known CDK9 inhibitors with 1475 presumably inactive compounds from ACD-3D database (molecules 1–1475). The virtual screening was performed as procedures described in Section 2.7.2 for test set prediction. Enrichment factor (EF) and hit rate (HR) were calculated at a given percentage of the database, and were defined as follows [30]:

$$EF = \frac{Activ^{x\%}_{screened}}{N^{x\%}_{screened}} \times \frac{N_{total}}{Activ_{total}} \quad (2)$$

$$HR = \frac{Activ^{x\%}_{screened}}{Activ_{total}} \times 100 \quad (3)$$

where $Activ^{x\%}_{screened}$ is the number of active compounds found in x% of the database screened, $N^{x\%}_{screened}$ is the number of compounds in x% of the database screened, N_{total} is the total number of compounds in the entire database, and $Activ_{total}$ is the total number of active compounds in the entire database.

3. Results and discussion

3.1. Selection of training and test sets

As described above, a predefined diversity sampling protocol was applied to select representative training sets for model construction. With such a protocol, both biological and chemical diversities were considered in training set selection, and twelve distinct training sets were thereby generated (the constitution of the training sets are illustrated in Table A, Supplementary data).

A representative training set was utilized to exemplify the efficiency of the diversity sampling protocol, and principal component analysis (PCA) [29] was exploited. Each compound in both training and test sets was represented by a single spot in a 3D-plot with the three axes standing for three orthogonal principal components, which explained more than 70% of the variation in the descriptor space (Fig. 2). As shown in Fig. 2, the training set molecules presented a sufficient coverage of the 3D-space occupied by the whole data set, which provides a sound proof for the effectiveness of diversity sampling. It is also worth to mention that the descriptors involved in diversity sampling and principal component analysis are different, which implies sufficient diversity and representativeness of the training set molecules in various descriptor spaces.

3.2. Pharmacophore generation

For each discrete training set, ten alternative pharmacophore models were generated and the best models with the lowest total cost values were selected and illustrated in Table 1. These models could be roughly classified into three groups according to the pharmacophoric features presented. The first seven models in Group I identified four functional features, including two hydrogen bond acceptors and two hydrophobic centers. The models in Group II also recognized four functional features, with one hydrogen bond acceptor in Group I model detected as hydrogen bond donor, and one hydrophobic center perceived as either hydrophobic or aromatic center. The models in Group III characterized three pharmacophoric features of one hydrogen bond acceptor, one hydrogen bond donor and one hydrophobic center. These features also matched with those presented in models of Group I and were in consistent spatial arrangement. Furthermore, most of the best models contained 4–6 excluded volumes to demonstrate steric hindrance restriction. The high homology among the pharmacophore models derived from different training sets implied sharp convergence in establishing pharmacophoric features critical for CDK9 inhibitory activity.

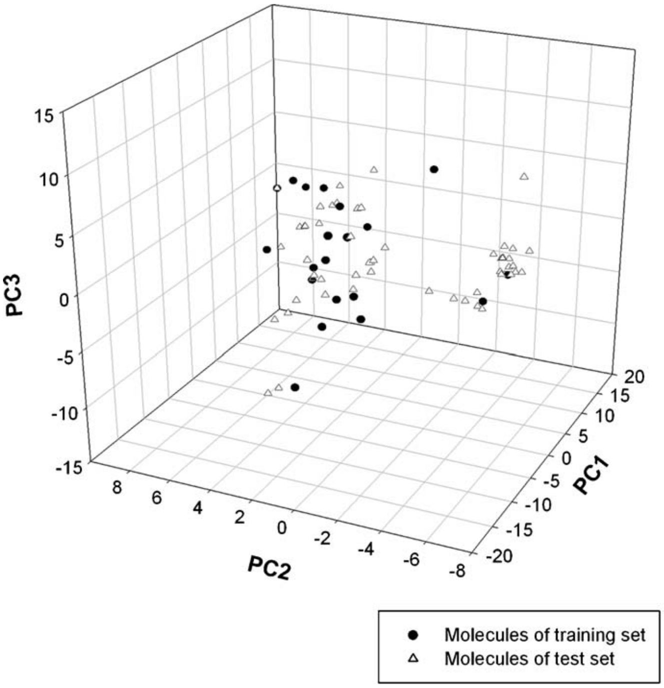


Fig. 2. Illustration of molecular diversity by principal components analysis.

A closer scrutiny of the pharmacophore models revealed subtle discrepancy among models. Despite the great similarity in pharmacophoric composition, the spatial arrangement of the pharmacophoric features varied with training sets, which pinpointed the dependence of final models on training set constitution and reinforced the importance of rational training set selection. The distances restriction between specific pharmacophoric features of the five representative models were illustrated in Table 2. As shown in Table 2, the distances between some pharmacophoric features (e.g. Distance 4) were rather constant, whereas some distances (e.g. Distance 5) fluctuated in a relatively broad range, which indicated divergent tolerance of different features to spatial variation and provided rationale for further structural modification and optimization.

Despite the significant structural diversity of known CDK9 inhibitors (cf. Fig. 1), the resulting models were readily mapped onto specific molecular areas of various potent CDK9 inhibitors in a chemically meaningful way. As exemplified in Fig. 3, a representative model, Hypo 6, was mapped with the six most active CDK9 inhibitors and identified functional features essential for CDK9 inhibitory activity. Some general structure–activity relationship (SAR) clues could be deduced from the mapping results. Basically, various scaffolds were well accommodated for the implantation of essential pharmacophoric features. Proximity between features A2 and Y1 was observed, and incorporation of A2 into heterocycles was frequently occurred. Among the two hydrophobic features, Y1 pre-

Table 1
Pharmacophore hypotheses generated with different training sets.

Class	Hypo No.[a]	Features[b]	Total cost	Fixed cost	Null cost	Cost diff. [c]	Conf. cost	Correlation square (r^2)	RMSD
I	1	AAYY	105.77	84.10	150.90	45.13	15.70	0.755	1.437
	2	AAYY	108.23	84.05	153.22	44.99	15.65	0.721	1.551
	3	AAYYE$_4$	102.34	83.42	145.06	42.72	15.02	0.759	1.372
	4	AAYYE$_4$	105.51	87.41	156.70	51.18	15.65	0.821	1.226
	5	AAYYE$_5$	103.36	83.67	154.96	51.60	15.27	0.787	1.369
	6	**AAYYE$_5$**	**100.35**	**83.67**	**152.47**	**52.12**	**15.27**	**0.805**	**1.290**
	7	AAYYE$_5$	99.70	79.90	141.24	41.54	14.87	0.752	1.290
II	8	ADYRE$_4$	86.75	75.66	131.70	44.94	13.99	0.876	1.007
	9	ADYYE$_6$	84.46	73.88	133.20	48.75	15.57	0.884	1.035
III	10	ADYE$_4$	97.20	79.02	144.18	46.98	13.99	0.808	1.289
	11	ADYE$_4$	85.50	72.29	127.54	42.04	13.99	0.837	1.174
	12	ADYE$_5$	92.09	75.66	139.34	47.25	13.99	0.841	1.211

[a] Each hypothesis was generated with a distinct training set and selected from ten possible hypotheses according to their statistical significance.
[b] A, hydrogen-bond acceptor; D, hydrogen-bond donor; Y, hydrophobic or hydrophobic aromatic center; R, aromatic center; E, excluded volume.
[c] Cost diff. = null cost − total cost. All cost units are in bits.

Table 2
Distance restrictions of representative pharmacophore models.

Hypo No.	A1-A2(D)[a] Distance 1[b]	A1-Y1 Distance 2	A1-Y2(R) Distance 3	A2(D)-Y1 Distance 4	A2(D)-Y2(R) Distance 5	Y1-Y2(R) Distance 6
4	9.909	7.045	8.905	4.592	4.937	7.468
6	8.864	5.411	10.116	4.783	11.849	7.831
8	5.450	8.460	10.456	3.345	8.641	7.378
9	7.752	6.834	11.297	3.649	3.755	5.450
12	4.878	7.803		3.052		

[a] A, hydrogen-bond acceptor; D, hydrogen-bond donor; Y, hydrophobic or hydrophobic aromatic center; R, aromatic center.
[b] The distances were depicted in Å with an error range of ±1.0 Å.

ferred to be aromatic moieties, whereas Y2 could be either aromatic or aliphatic fragments.

Hypo 6 was further examined by mapping with known CDK9 inhibitors in the ATP-binding site of CDK9. To explore whether the ligand-based pharmacophore model could reinstate key interactions between the enzyme and the inhibitors as revealed in their X-ray crystal structures, only Flavopiridol (FVP, represented in red) [4a] and DRB (represented in yellow) [4b], which had their complex

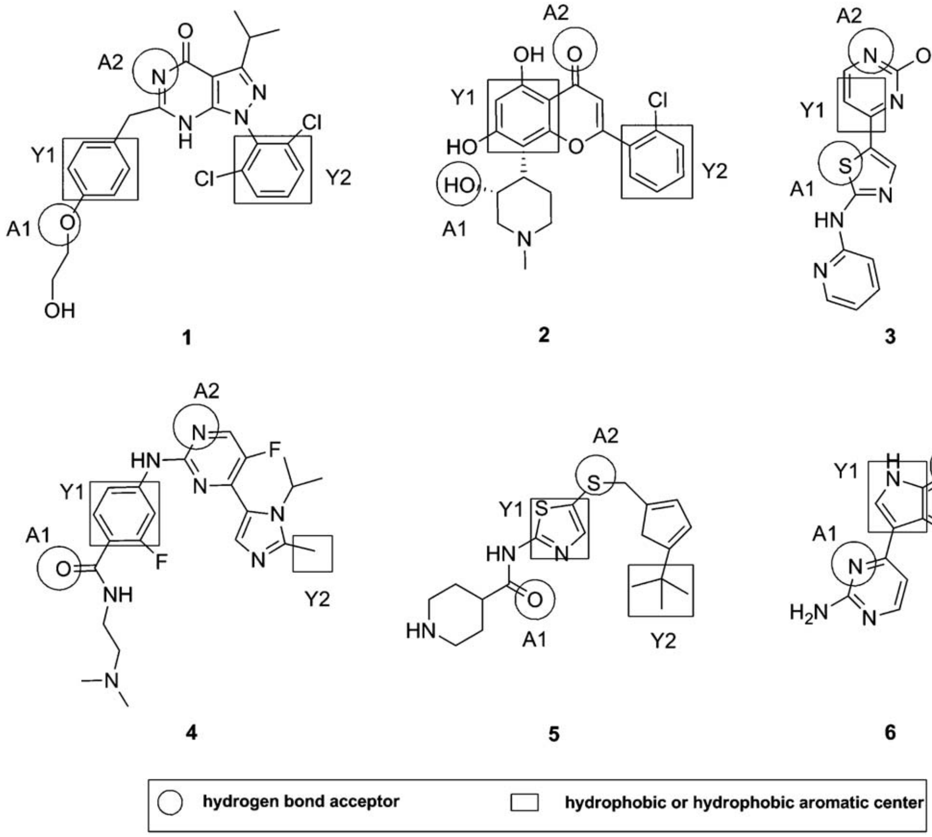


Fig. 3. Functional features identified by Hypo 6 as critical for CDK9 inhibitory activity.

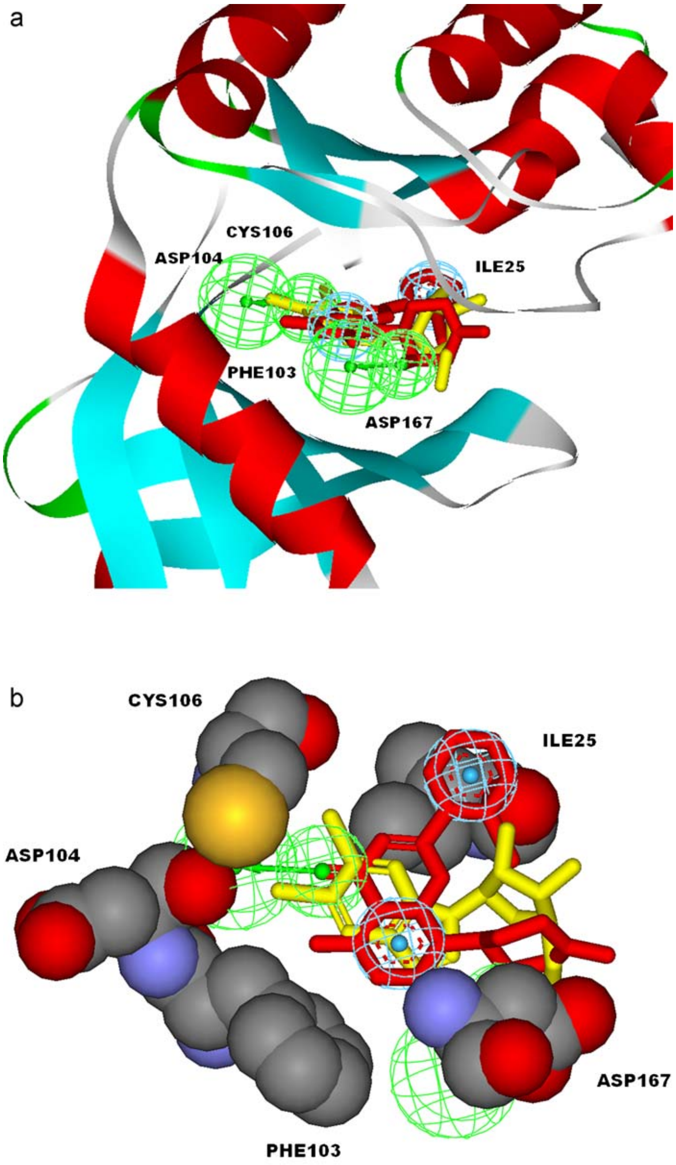


Fig. 4. Hypo 6 mapped to CDK9 inhibitors, FVP and DRB, in the ATP-binding site of CDK9 (FVP was depicted in red and DRB was depicted in yellow). (a) Overview of the interaction mode between CDK9 and FVP/DRB. Five amino acid residues involved in the interaction were displayed. (b) Key interactions between CDK9 and FVP/DRB. Five amino acid residues involved in the interaction were displayed (gray: C atoms; red: O atoms; blue: N atoms; yellow: S atoms).

structures with CDK9 resolved, were selected for molecular overlay and graphic illustration (Fig. 4).

As depicted in Fig. 4, FVP was deeply buried into the ATP-binding pocket of CDK9. The carbonyl oxygen of the benzofuran skeleton and the hydroxyl oxygen of the piperidinyl ring form hydrogen bonds with Cys106 NH and Asp167 NH respectively, which coincides with the two hydrogen-bond acceptor features presented in Hypo 6. Moreover, the chloro-phenyl ring of FVP made a hydrophobic contact to Ile25, and a π–π interaction was also observed between the benzene ring of FVP and residue Phe103, which concurs with the two hydrophobic features of Hypo 6. Thereby, all four features of Hypo 6 were mapped correctly onto the molecular areas of FVP, and revealed crucial functional features responsible for its interactions with the ATP-binding pocket of CDK9, which was consistent with those portrayed in the FVP–CDK9 complex structure [4a]. Besides the four functional groups recognized by Hypo 6, a hydrogen bonding interaction between the C5 hydroxyl of FVP and the backbone carbonyl of Asp104 was also unveiled by the crystal structure. The presence of such an extra interaction might associate with the high potency of FVP.

Similarly, DRB situated in the ATP-binding pocket of CDK9 with an orientation alike to FVP. The planar benzimidazole moiety of DRB protruded toward residue Phe103, possibly for a hydrophobic π–π interaction, which overlaid with one of the hydrophobic feature in Hypo 6. Interestingly, a feature of hydrogen bond acceptor in Hypo 6 was located near the two chlorine atoms of DRB (Fig. 4b). It was in accord with previous observation that the two chlorine atoms of DRB established halogen bonds and even hydrogen bonds with the residues of Asp104 and Cys106 [4b]. Another feature of hydrogen bond acceptor in Hypo 6 lingered in the vicinity of Asp167, which endorsed the existence of a water-mediated hydrogen bond between the benzimidazole N2 and the backbone NH group of Asp 167 [4b]. Although due to the limitation of methodology (halogen bonds and water-mediated hydrogen bonds are not adequately considered in DS 2.1), these two hydrogen bond acceptors failed to produce an exact match, they did revealed conducive clues for the action mode and structural requirements of CDK9 inhibitors. We also noticed that DRB missed a hydrophobic feature of Hypo 6, which was assumed to interact with residue Ile25. The absence of this feature at least partially explained the relatively weak potency of DRB in CDK9 inhibition and provided informative hints for further optimization of DRB analogs.

3.3. *Pharmacophore validation*

3.3.1. *Cost value analysis*

As discussed previously, the HypoGen algorithm calculates and reports a series of parameters for preliminary estimation of the models' statistical robustness. Normally, the cost difference between null and total costs, the configuration cost, the correlation coefficient and root mean square difference (RMSD) between the predicted and the experimental activities of the training set molecules were taken as critical parameters to assess statistical significance. A meaningful pharmacophore hypothesis should hold a significant cost difference to ensure the detection of true correlation rather than chance correlation. Usually, a cost difference in the range of 40–60 is required to guarantee 75–90% probability of true correlation. A reliable pharmacophore model also demands a configuration cost value less than 17, good correlation coefficient and low RMSD. The key parameters of each model were recorded in Table 1, which satisfied the aforementioned criteria and indicated statistical validity of all the twelve models generated. Among these models, Hypo 6 in Group I, which has the most significant cost difference and thus is most likely to reflect a true correlation, was selected as a representative model for further validation.

3.3.2. *Internal and external prediction*

The capability to accurately predict internal and particularly external data sets is an important attribute of a reliable pharmacophore model. As plotted in Fig. 5, the correlation coefficient between the predicted and the experimental activities of the training set molecules was 0.897, and all compounds were confined within an error range of one log unit. When Hypo 6 was applied to predict the activity of the fifty-five test set molecules, the external predictive correlation coefficient was as high as 0.846, and only four compounds slightly exceeded the one log unit error margin (Fig. 6).

To further evaluate the predictive capability of Hypo 6, the training and test sets molecules were subjectively split into three activity scales: the most active includes those with IC_{50} values lower than 20 nM, the moderately active contains compounds with IC_{50} values ranging from 20 to 1000 nM, and the least active restricts to those with IC_{50} values higher than 1000 nM. As shown in Fig. 7,

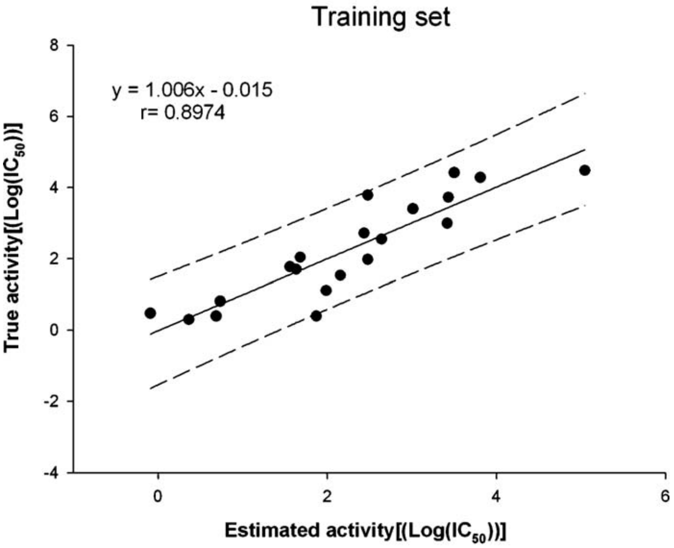


Fig. 5. Correlation between experimental versus predicted $\log(IC_{50})$ values of the training set molecules based on Hypo 6.

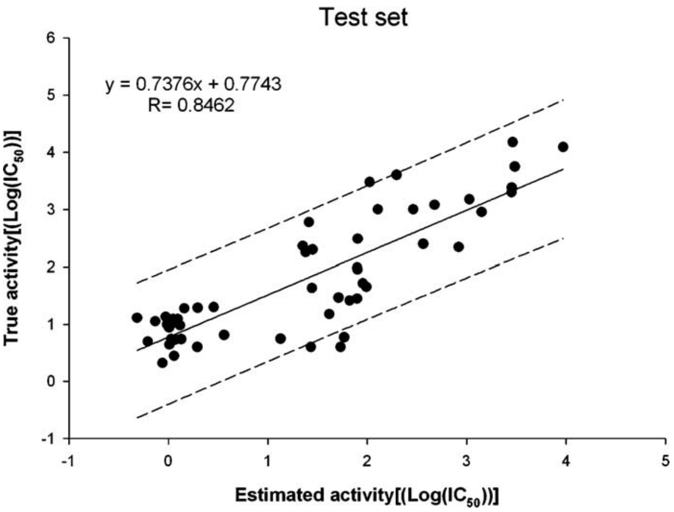


Fig. 6. Correlation between experimental versus predicted $\log(IC_{50})$ values of the test set molecules based on Hypo 6.

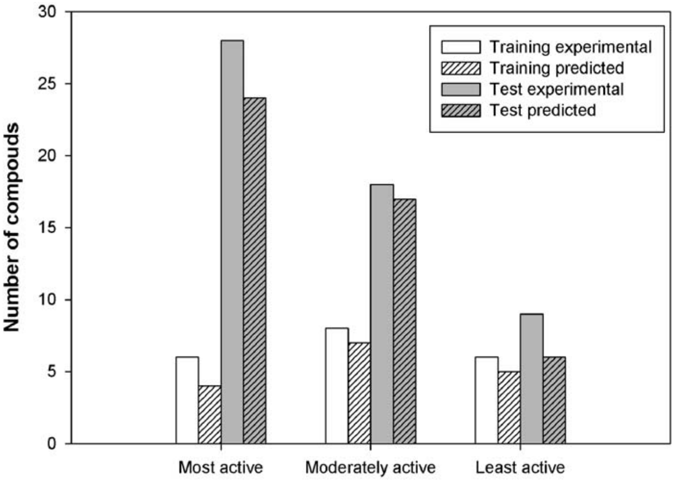


Fig. 7. Predictive capability of Hypo 6 on activity scales. The experimental columns represent the number of compounds with experimental data in respective activity scale; the predicted columns represent the number of compounds with activity data correctly predicted in the right scale.

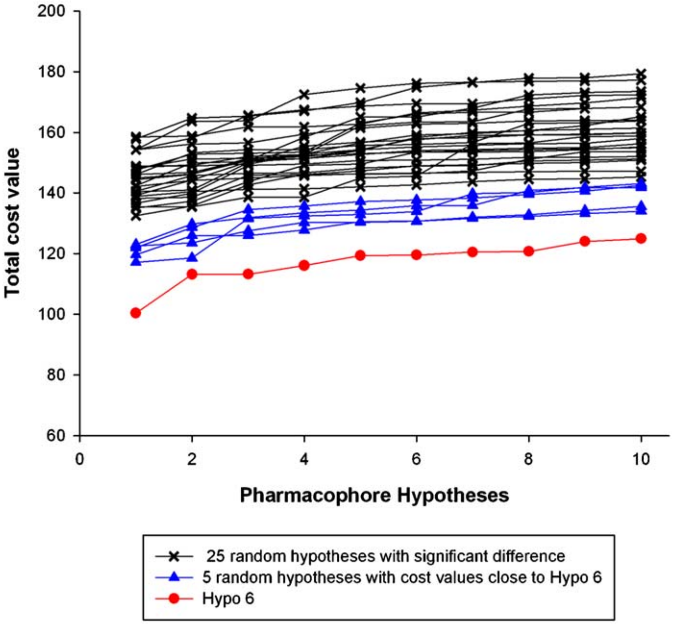


Fig. 8. Total cost values of Hypo 6 and 30 random models.

the activity scales of most compounds in both training and test sets were predicted correctly. Most notably, when considering all the compounds with IC_{50} values lower than 1000 nM as "active", 45 out of the 46 active molecules in the original data set were correctly detected by Hypo 6, which indicated its excellent potential to discriminate active compounds from the others.

3.3.3. Randomization test

Further randomization test was implemented on Hypo 6 to check whether the good correlation coefficient obtained was derived from true or chance correlation. Thirty "random" training sets were generated from the training set used for Hypo 6 and thirty "random" pharmacophore models were obtained accordingly. The total cost values of these random models were compared with those of Hypo 6 (Fig. 8). Interestingly, although most of the random models showed total cost values significantly higher than that of Hypo 6 as expected, some models did provide total cost values close to that of Hypo 6. To further assess their statistical significance, five models with relatively low total cost values (represented in blue in Fig. 8) were applied for external prediction of the 55 molecules in test set. As discussed previously [31], the external predictive power of a QSAR model is the definite criteria to estimate its reliability and eliminate the possibility of chance correlation. The parameters, including internal and external correlation coefficients (expressed in square), and the regression slopes were illustrated in Table 3. Although the internal correlation coefficients of the random models were generally acceptable, the squares of their correlation coef-

Table 3
Comparison of the five random hypotheses with Hypo 6.

Hypothesis	r^2 [a]	R^2 [b]	k [c]
Random 26	0.539	0.213	0.39
Random 27	0.521	0.100	0.39
Random 28	0.579	0.000	0.03
Random 29	0.566	0.108	0.31
Random 30	0.453	0.040	0.25
Hypo 6	**0.805**	**0.716**	**0.74**

[a] r^2 represents the square of correlation coefficient derived from training set.

[b] R^2 represents the square of correlation coefficient derived from test set.

[c] k represents the slope of the regression line for experimental versus predicted activities of test sets.

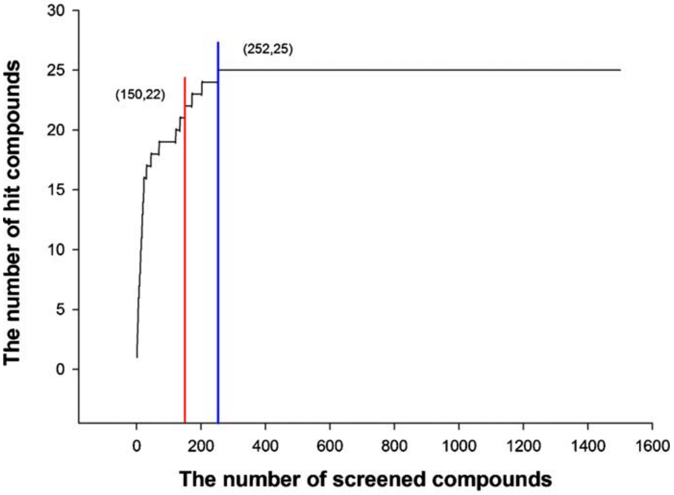


Fig. 9. Correlation between the number of hit compounds identified versus the number of compounds screened. The red line represents that 22 hits would be identified when 150 compounds screened. The blue line reflects that all of the 25 hits would be detected when 252 compounds screened.

ficients for the test set were close to zero, which implied their ominously poor external predictive power. In addition, the regression slopes of the random models were distinctly apart from the ideal value 1.0, which was in sharp contrast to the slope of 0.74 by Hypo 6. Thereby, the randomization test confirmed the statistical significance of Hypo 6.

3.3.4. *Enrichment factor and hit rate validation*

Finally, we attempted to calculate the enrichment factor (EF) and hit rate (HR) at a given percentage of the pre-defined database screened by Hypo 6, which was considered as a meaningful approach to validate the discriminative power of a pharmacophore model in virtual screening. The relationship between the number of hit compounds identified and the number of compounds screened was depicted in Fig. 9. Apparently, when *x*%, which is a given percentage of the database screened, was set to 2%, 5%, and 10%, the optimal values for EF were 50, 20 and 10, respectively. Notably, the EF values obtained from Hypo 6 at these given percentages were 32, 15.2 and 8.8 respectively, which were akin to the theoretical values. Furthermore, 22 out of the 25 pre-located active compounds were detected in the top 150 compounds of 10% database screened, and when 252 compounds were screened (16.8% of the whole database), all the 25 active compounds could be identified (Fig. 9). The hit rate of Hypo 6 at 2%, 5% and 10% of database screened was 64%, 76% and 88% respectively, which is amazingly good for a pharmacophore-based virtual screening. The discriminative power of Hypo 6 implicated that pharmacophore-based virtual screening may provide an efficient approach to find novel CDK9 inhibitors from available databases.

4. Summary

Novel and potent CDK9 inhibitors would have potential therapeutic applications in HIV infections. A diverse data set of known CDK9 inhibitors was subjected to HypoGen modelling to comprehend pharmacophoric features essential for CDK9 inhibition. The necessity to rationally select representative training sets were underscored by diversity sampling of the original data set in both biological and chemical spaces, and subsequent principal component analysis of the training set molecules in multiple dimensional spaces. The significance of adequate model validation was also accentuated by integrating statistical analysis with comprehensive evaluation of both external predictive capability and discriminative power of the models. To highlight, robust pharmacophore models were constructed from sufficiently diverse training set molecules, which demonstrated decent predictive capability to both internal and external data sets, and excellent discriminative power to active CDK9 inhibitors. These models could be used to search for novel CDK9 inhibitors through virtual screening and provide sound rationale for efficient scaffold hopping. They could also guide chemical modification and optimization of known CDK9 inhibitors to offer safe and potent novel chemical entities for HIV-1 therapy.

Acknowledgements

This investigation was supported by 863 Hi-Tech Program (2006AA020401), Beijing Natural Science Foundation (7082061) and National S&T Major Project (2009ZX09301-003-1-1).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2011.01.003.

References

[1] B.M. Klebl, A. Choidas, CDK9/cyclin T1: a host cell target for antiretroviral therapy, Fut. Virol. 1 (2006) 317–330.

[2] S. Wang, P.M. Fischer, Cyclin-dependent kinase 9: a key transcriptional regulator and potential drug target in oncology, virology and cardiology, Trends Pharmacol. Sci. 29 (2008) 302–313.

[3] V. Kryštof, I. Chamrád, R. Jorda, J. Koboutek, Pharmacological targeting of CDK9 in cardiac hypertrophy, Med. Res. Rev. 30 (2009) 1–21.

[4] (a) S. Baumli, G. Lolli, E.D. Lowe, S. Troiani, L. Rusconi, A.N. Bullock, J.É. Debreczeni, S. Knapp, L.N. Johnson, The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation, EMBO J. 27 (2008) 1907–1918;
(b) S. Baumli, J.A. Endicott, L.N. Johnson, Halogen bonds form the basis for selective P-TEFb inhibition by DRB, Chem. Biol. 17 (2010) 931–936.

[5] Y. Kurogi, O.F. Güner, Pharmacophore modeling and three-dimensional database searching for drug design using catalyst, Curr. Med. Chem. 8 (2001) 1035–1055.

[6] Information about Discovery Studio. Available from: http://accelrys.com/products/discovery-studio.

[7] H.S.Y. Mancebo, G. Lee, J. Flygare, J. Tomassini, P. Luu, Y. Zhu, J. Peng, C. Blau, D. Hazuda, D. Price, O. Flores, P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro, Genes Dev. 11 (1997) 2633–2644.

[8] A. Ali, A. Ghosh, R.S. Nathans, N. Sharova, S. O'Brien, H. Cao, M. Stevenson, T.M. Rana, Identification of flavopiridol analogues that selectively inhibit positive transcription elongation factor (P-TEFb) and block HIV-1 replication, ChemBioChem 10 (2009) 2072–2080.

[9] V. Kryštof, I.W. McNae, M.D. Walkinshaw, P.M. Fischer, P. Müller, B. Vojtěšek, M. Orság, L. Havlíček, M. Strnad, Antiproliferative activity of olomoucine II, a novel 2,6,9-trisubstituted purine cyclin-dependent kinase inhibitor, Cell. Mol. Life Sci. 62 (2005) 1763–1771.

[10] F. Popowycz, G. Fournet, C. Schneider, K. Bettayeb, Y. Ferandin, C. Lamigeon, O.M. Tirado, S. Mateo-Lozano, V. Notario, P. Colas, P. Bernard, L. Meijer, B. Joseph, Pyrazolo[1,5-a]-1,3,5-triazine as a purine bioisostere: access to potent cyclin-dependent kinase inhibitor (R)-roscovitine analogue, J. Med. Chem. 52 (2009) 655–663.

[11] K. Bettayeb, H. Sallam, Y. Ferandin, F. Popowycz, G. Fournet, M. Hassan, A. Echalier, P. Bernard, J. Endicott, B. Joseph, L. Meijer, N-&-N, a new class of cell death-inducing kinase inhibitors derived from the purine roscovitine, Mol. Cancer Ther. 7 (2008) 2713–2724.

[12] K. Bettayeb, N. Oumata, A. Echalier, Y. Ferandin, J.A. Endicott, H. Galons, L. Meijer, CR8, a potent and selective, roscovitine-derived inhibitor of cyclin-dependent kinases, Oncogene 27 (2008) 5797–5807.

[13] K. Bettayeb, O.M. Tirado, S. Marionneau-Lambot, Y. Ferandin, O. Lozach, J.C. Morris, S. Mateo-Lozano, P. Drueckes, C. Schachtele, M.H.G. Kubbutat, F. Liger, B. Marquet, B. Joseph, A. Echalier, J.A. Endicott, V. Notario, L. Meijer, Meriolins, a new class of cell death-inducing kinase inhibitors with enhanced selectivity for cyclin-dependent kinases, Cancer Res. 67 (2007) 8325–8334.

[14] A. Echalier, K. Bettayeb, Y. Ferandin, O. Lozach, M. Clément, A. Valette, F. Liger, B. Marquet, J.C. Morris, J.A. Endicott, B. Joseph, L. Meijer, Meriolins (3-(pyrimidin-4-yl)-7-azaindoles): synthesis, kinase inhibitory activity, cellular effects, and structure of a CDK2/cyclin A/meriolin complex, J. Med. Chem. 51 (2008) 737–751.

[15] M.E. Lane, B. Yu, A. Rice, K.E. Lipson, C. Liang, L. Sun, C. Tang, G. McMahon, R.G. Pestell, S. Wadler, A novel CDK2-selective inhibitor, SU9516, induces apoptosis in colon carcinoma cells, Cancer Res. 61 (2001) 6170–6177.

[16] A. Heredia, C. Davis, D. Bamba, N. Le, M.Y. Gwarzo, M. Sadowska, R.C. Gallo, R.R. Redfield, Indirubin-30-monoxime, a derivative of a Chinese antileukemia medicine, inhibits P-TEFb function and HIV-1 replication, AIDS 19 (2005) 2087–2095.

[17] A. Conroy, D.E. Stockett, D. Walker, M.R. Arkin, U. Hoch, J.A. Fox, R.E. Hawtin, SNS-032 is a potent and selective CDK 2, 7 and 9 inhibitor that drives target modulation in patient samples, Cancer Chemother. Pharmacol. 64 (2009) 723–732.

[18] S. Wang, C. Meades, G. Wood, A. Osnowski, S. Anderson, R. Yuill, M. Thomas, M. Mezna, W. McInnes, D. Zheleva, M.D. Walkinshaw, P.M. Fischer, 2-Anilino-4-(thiazol-5-yl)pyrimidine CDK inhibitors: synthesis, SAR analysis, X-ray crystallography, and biological activity, J. Med. Chem. 47 (2004) 1662–1675.

[19] T. Shimamura, J. Shibata, H. Kurihara, T. Mita, S. Otsuki, T. Sagara, H. Hirai, Y. Iwasawa, Identification of potent 5-pyrimidinyl-2-aminothiazole CDK4, 6 inhibitors with significant selectivity over CDK1, 2, 5, 7, and 9, Bioorg. Med. Chem. Lett. 16 (2006) 3751–3754.

[20] C. Zhang, K. Lundgren, Z. Yan, M.E. Arango, S. Price, A. Huber, J. Higgins, G. Troche, J. Skaptason, T. Koudriakova, J. Nonomiya, M. Yang, P. O'Connor, S. Bender, G. Los, C. Lewis, B. Jessen, Pharmacologic properties of AG-012986, a pan-cyclin dependent kinase inhibitor with antitumor efficacy, Mol. Cancer Ther. 7 (2008) 818–828.

[21] C.D. Jones, D.M. Andrews, A.J. Barker, K. Blades, P. Daunt, S. East, C. Geh, M.A. Graham, K.M. Johnson, S.A. Loddick, H.M. McFarland, A. McGregor, L. Mossa, D.A. Rudge, P.B. Simpson, M.L. Swain, K.Y. Tam, J.A. Tucker, M. Walker, The discovery of AZD5597, a potent imidazole pyrimidine amide CDK inhibitor suitable for intravenous dosing, Bioorg. Med. Chem. Lett. 18 (2008) 6369–6373.

[22] D. Cai, K.F. Byth, G.I. Shapiro, AZ703, an imidazo[1,2-a]pyridine inhibitor of cyclin-dependent kinases 1 and 2, induces E2F-1-dependent apoptosis enhanced by depletion of cyclin-dependent kinase 9, Cancer Res. 66 (2006) 435–444.

[23] V. Kryštof, P. Cankař, I. Fryšová, J. Slouka, G. Kontopidis, P. Džubák, M. Hajdúch, J. Srovnal, W.F. de Azevedo Jr., M. Orság, M. Paprskářová, J. Rolčík, A. Látr, P.M. Fischer, M. Strnad, 4-Arylazo-3,5-diamino-1H-pyrazole CDK inhibitors: SAR study, crystal structure in complex with CDK2, selectivity, and cellular effects, J. Med. Chem. 49 (2006) 6500–6509.

[24] M. Caligiuri, F. Becker, K. Murthi, F. Kaplan, S. Dedier, C. Kaufmann, A. Machl, G. Zybarth, J. Richard, N. Bockovich, A. Kluge, N. Kley, A proteome-wide CDK/CRK-specific kinase inhibitor promotes tumor cell death in the absence of cell cycle progression, Chem. Biol. 12 (2005) 1103–1115.

[25] M. Menichincheri, A. Bargiotti, J. Berthelsen, J.A. Bertrand, R. Bossi, A. Ciavolella, A. Cirla, C. Cristiani, V. Croci, R. D'Alessio, M. Fasolini, F. Fiorentini, B. Forte, A. Isacchi, K. Martina, A. Molinari, A. Montagnoli, P. Orsini, F. Orzi, E. Pesenti, D. Pezzetta, A. Pillan, I. Poggesi, F. Roletto, A. Scolaro, M. Tatò, M. Tibolla, B. Valsasina, M. Varasi, D. Volpi, C. Santocanale, E. Vanotti, First Cdc7 kinase inhibitors: pyrrolopyridinones as potent and orally active antitumor agents. 2. Lead discovery, J. Med. Chem. 52 (2009) 293–307.

[26] A. Ermoli, A. Bargiotti, M.G. Brasca, A. Ciavolella, N. Colombo, G. Fachin, A. Isacchi, M. Menichincheri, A. Molinari, A. Montagnoli, A. Pillan, S. Rainoldi, F.R. Sirtori, F. Sola, S. Thieffine, M. Tibolla, B. Valsasina, D. Volpi, C. Santocanale, E. Vanotti, Cell division cycle 7 kinase inhibitors: 1H-pyrrolo[2,3-b]pyridines, synthesis and structure–activity relationships, J. Med. Chem. 52 (2009) 4380–4390.

[27] A. Smellie, S.L. Teig, P. Towbin, Poling: promoting conformational variation, J. Comput. Chem. 16 (1995) 171–187.

[28] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures, Org. Biomol. Chem. 2 (2004) 3256–3266.

[29] A.R. Leach, Molecular Modeling: Principles and Applications, 2nd edition, Prentice Hall, 2001, pp. 497–499.

[30] J. Sutter, O. Güner, R. Hoffman, H. Li, M. Waldman, Effect of variable weights and tolerances on predictive model generation, in: O.F. Güner (Ed.), Pharmacophore Perception, Development and Use in Drug Design, International University Line, CA, 2000, pp. 501–511.

[31] A. Golbraikh, A. Tropsha, Beware of q^2!, J. Mol. Graph. Modell. 20 (2002) 269–276.