Mol Divers
DOI 10.1007/s11030-014-9561-3

FULL-LENGTH PAPER

Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery

Jiansong Fang · Ranyao Yang · Li Gao · Shengqian Yang · Xiaocong Pang · Chao Li · Yangyang He · Ai-Lin Liu · Guan-Hua Du

Received: 19 July 2014 / Accepted: 25 November 2014

Abstract Cyclin-dependent kinase 5 (CDK5) has emerged as a principal therapeutic target for Alzheimer's disease. It is highly desirable to develop computational models that can predict the inhibitory effects of a compound towards CDK5 activity. In this study, two machine learning tools (naive Bayesian and recursive partitioning) were used to generate four single classifiers from a large dataset containing 462 CDK5 inhibitors and 1,500 non-inhibitors. Then, two types of consensus models [combined classifier-artificial neural networks (CC-ANNs) and consensus prediction] were applied to combine four single classifiers to obtain superior performance. The results showed that both consensus models outperformed four single classifiers, and (MCC = 0.806) was superior to consensus prediction (MCC = 0.711) for an external test set. To illustrate the practical applications of the CC-ANN model in virtual screening, an in-house dataset containing 29,170 compounds was screened, and 40 compounds were selected for further bioactivity assays. The assay results showed that 13 out of 40 compounds exerted CDK5/p35 inhibitory activities with IC_{50} values ranging from 9.23 to 229.76 μM. Interestingly, three new scaffolds that had not been previously reported as CDK5 inhibitors were found in this study. These studies prove that our protocol is an effective approach to predict small-molecule CDK5 affinity and identify novel lead compounds.

Keywords Cyclin-dependent kinase 5 (CDK5) · Consensus model · Artificial neural networks · Virtual screening

Introduction

Alzheimer's disease (AD) is a progressively degenerative disease along with memory decline, cognitive impairment, and visual–spatial disorientation [1]. It is reported that more than 35 million people worldwide suffer from AD [2]. Several acetylcholinesterase inhibitors and one NMDA receptor antagonist are clinically used to temporarily restrain the effects of AD; however, currently there are no effective treatments to prevent or cure this pathology [3].

The proline-directed serine/threonine kinase cyclin-dependent kinase 5 (CDK5) belongs to the large family of

Electronic supplementary material The online version of this article (doi:10.1007/s11030-014-9561-3) contains supplementary material, which is available to authorized users.

Jiansong Fang and Ranyao Yang have contributed equally to this work.

J. Fang · R. Yang · L. Gao · S. Yang · X. Pang · C. Li · Y. He · A.-L. Liu (✉) · G.-H. Du (✉)
Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, 1 Xian Nong Tan Street, Beijing 100050, People's Republic of China
e-mail: liuailin@imm.ac.cn

G.-H. Du
e-mail: dugh@imm.ac.cn

A.-L. Liu · G.-H. Du
Beijing Key Laboratory of Drug Target and Screening Research, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, People's Republic of China

A.-L. Liu · G.-H. Du
State Key Laboratory of Bioactive Substance and Function of Natural Medicines, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, People's Republic of China

Published online: 16 December 2014

cyclin-dependent kinases (CDKs) [4,5]. CDK5 and its cofact tors p25 and p35 are thought to hyperphosphorylate the tau protein [6,7] which leads to the formation of paired helical filaments and the deposition of cytotoxic neurofibrillary tangles (NFTs) [8–11]. CDK5 can also phosphorylate the dopamine and cyclic AMP-regulated phosphoprotein (DARPP-32) at threonine 75, indicating that it plays a role in dopaminergic neurotransmission.

Furthermore, it was revealed that activation of CDK5 in vivo models recapitulated lots of AD features, such as tau hyperphosphorylation, the formation of NFTs, neurodegeneration, cognitive impairment, and increase in amyloid-β (Aβ) levels [12–14]. Thus, considerable efforts have been invested to explore CDK5 inhibitors as new drugs for Alzheimer's disease treatment. Because the discovery of inhibitors by experimental methods is time-consuming, costly, and labour intensive, it is necessary to develop in silico methods to facilitate the screening and identification of CDK5 inhibitors.

Structure-based drug discovery is a general approach to find potential inhibitors towards AD. Due to the limitation of available crystallographic structures of biological targets, scientists have pursued 3D structures of important targets towards AD using *in silico* methods [15–19]. Such 3D structures are important for drug design, particularly for conducting mutagenesis studies [20]. Chou et al. constructed a *in silico* model of CDK5-p25-ATP to provide structural basis for the study of the mechanisms of CDK5 activation [15], and the model was experimentally validated by identifying functional domains stimulating relevant truncation experiments [21]. Pielak and Chou also developed *in silico* models to define the binding pockets of many other receptor–ligand interactions important for drug design [22,23].

Quantitative structure–activity relationship (QSAR) studies are also effective computational methods that can speed up drug discovery. In recent years, considerable progress has been made in QSAR studies [24], such as multiple field QSAR (MF-QSAR) [25], fragment-based QSAR (FB-QSAR) [26], and multi-tasking QSAR (mt-QSAR) [27]. QSAR studies have been widely applied to virtual screening [28–30], and ligand biological activity prediction studies [31]. To date, numerous QSAR models have been developed to predict CDK inhibitor activity [32–40]. For example, in 2005, Fernandez et al. used artificial neural network ensembles to model the CDK inhibitory activity of 1H-pyrazolo[3,4-*d*]pyrimidine derivatives. The model explained approximately 87 % of the data variance with a q^2 of 0.648 of by LOO-cross-validation experiments [32]. In 2007, Li et al. applied least-squares support vector machines (LS-SVMs) and linear discriminant analysis (LDA) to develop models to classify oxindole-based inhibitors of CDKs, achieving best performance with LS-SVMs models with a 100 % prediction accuracy on the test set and 90.91 % for CDK1 and CDK2, respectively, as well as that of LDA models 95.45 % for CDK1 and 86.36 % for CDK2 [33].

Overall, the studies using in silico CDK5 inhibitors predictions are not enough for identifying hit compounds. For example, most QSAR studies examining CDK inhibitors are not directed towards CDK5 [34–38]. Furthermore, the datasets used for previous QSAR studies of CDK5 inhibitors did not contain a large number of compounds with chemical scaffold diversity, thus limiting their application.

In this study, a workflow for classification models, model validations, and their application to the virtual screening of CDK5 inhibitors is shown in Fig. 1. First, we present a large dataset containing 1,962 compounds (including CDK5 inhibitors and non-inhibitors) that were divided into a training set and a test set. Then, we used two machine learning methods (naive Bayesian and recursive partitioning) to generate four single classifiers using the training set. Additionally, we applied two types of consensus models to combine different single machine learning classifiers: (a) four single classifiers were integrated by a combined classifier-artificial neural network (CC-ANN) model fused with a back-propagation (BP) artificial neural network algorithm, and (b) a consensus prediction protocol was applied to enrich datasets by merging predictive results from the four single classifiers previously identified. A test set and an external test set were used to mea-

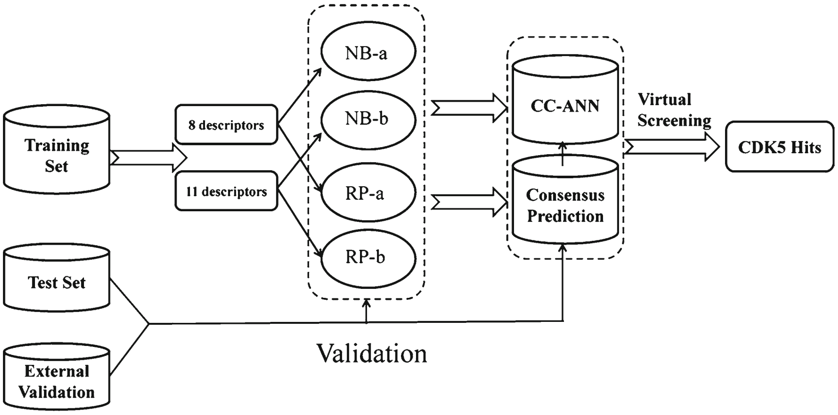


Fig. 1 Workflow for classification model building, validation, and virtual screening (VS) as applied to CDK5 inhibitors and non-inhibitors. *NB* naive Bayesian, *RP* recursive partitioning, *CC-ANN* combined classifier-artificial neural network, *CDK5* cyclin-dependent kinase 5

sure the performance of both models. The results showed that the overall performance of both consensus models (CC-ANN and consensus prediction) was superior to that of the four single classifiers, and that the CC-ANN model performed better than the consensus prediction protocol. Consequently, the better-performing consensus model CC-ANN was used as a new classifier of CDK5 inhibitors and non-inhibitors to screen an in-house database. Finally, several selected compounds were subjected to a CDK5 enzymatic inhibitory assay to assess their actual CDK5 inhibitory profile.

Materials and methods

Preparation of the dataset

510 CDK5 reported inhibitors were obtained from the BindingDB database [41] and related references [42–47]. After removing 48 duplicate compounds, 462 inhibitors with $IC_{50} < 10\ \mu M$ were considered as active compounds. Additionally, a dataset of 1,500 decoys were randomly obtained from the CDK5 decoy database, which was generated using the DUD online database [48] with known CDK5 inhibitors. For the four single classifiers and the BP-ANN classifiers, both the inhibitor and decoy datasets were randomly separated into two parts. The training set contained 360 CDK5 inhibitors and 1,080 decoy compounds, and the test set 102 CDK5 inhibitors and 420 decoys (detailed data can be found in the Supporting Information, see Tables S1, S2).

All the compounds were processed as follows: salts were converted into the corresponding acids or bases; water molecules were removed from hydrates. For each compound, hydrogen atoms were added; strong acids were deprotonated, and strong bases were protonated. Three-dimensional conformation was generated through the minimization of energy module using the Molecular Operating Environment (MOE) [49] software. All inhibitors were considered active compounds and labelled "1", while decoys exhibiting no activity towards CDK5 were labelled "0".

Molecular descriptors calculation and selection

Two combinations of molecular descriptors (Table 1) were used to construct the classification models using the Discovery Studio 3.1 software [50]. The first combination utilized eight default descriptors (LogP, Molecular_Weight, Num_AromaticRings, Num_H_Acceptors, Num_H_Donors, Num_Rings, Num_RotatableBonds, and Molecular_ Fractional Polar Surface Area) available in the software. The other combination was optimized by applying Pearson correlation analysis and stepwise regression, which is described as follows.

Discovery Studio 3.1 was employed to calculate two-dimensional molecular descriptors, which consisted of six subtype descriptors including a total of 301 descriptors for the training set (1,440 compounds). Pearson correlation analysis [51] can select the descriptors that are significantly correlated with activity while avoiding any high correlation with each other. In this study, descriptors exhibiting a Pearson correlation coefficient lower than 0.1 ($P < 0.1$) regarding activity were removed. If a pairwise correlation coefficient ($P > 0.9$) between any two descriptors was found, the descriptor with the lower P value of descriptor–activity was then removed. After that, a stepwise regression was carried out to further optimize the descriptors. At the end of this process, 11 descriptors were selected for model building.

Table 1 Molecular descriptors used in this work

Descriptor class	Number of descriptors	Descriptors
DS_2D_default	8	ALogP, Molecular_Weight, Num_AromaticRings, Num_H_Acceptors, Num_H_Donors, Num_Rings, Num_RotatableBonds, Molecular_FractionalPolarSurfaceArea
DS_2D_stepwise	11	Molecular_Solubility, Num_AromaticRings, Num_H_Donors, Num_Rings, Num_Rings5, Molecular_FractionalPolarSASA, Molecular_PolarSASA, BIC, CHI_3_C, CHI_V_3_C, Kappa_1_AM

Model methods

Three different machine learning tools [naive Bayesian (NB), recursive partitioning (RP), and back-propagation artificial neural networks (BP-ANNs)] were employed in this study using Discovery Studio 3.1 [50] for NB and RP, and Matlab R2007b for BP-ANN [52].

Naive Bayesian (NB) classifiers

The Bayesian algorithm is robust for classification, and it can distinguish between active and inactive compounds [53]. In general, a Bayesian method depends on the frequency of occurrence of diverse descriptors which are found in two or more sets of compounds that can discriminate best between these sets. Bayesian classification can process large amounts of data. For naive Bayesian classifiers, Bayesian can generate the posterior probabilities based on the core of the function, which are given by Eq. 1.

$$P(+|\mathrm{A}_1, \ldots, \mathrm{A}_n) = \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n |+) P(+)}{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}. \qquad (1)$$

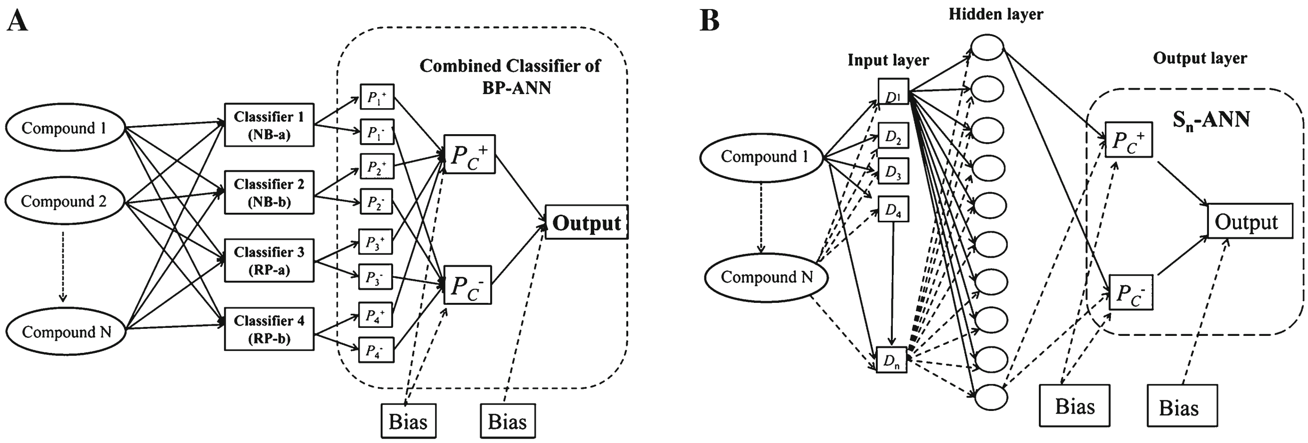


Fig. 2 The architecture of the combined classifier of BP-ANN (**a**) and single classifiers of BP-ANN (**b**). Combined classifiers take a set of probability output for inhibitors (+1) or non-inhibitors (0) by a single classifier (NB or RP) and produce a combination probability output for each class. *NB* naive Bayesian, *RP* recursive partitioning, *BP-ANN* back-propagation artificial neural network

Here, $P(\mathrm{A}_1, \ldots, \mathrm{A}_n|+)$ represents the conditional probability of one compound being classified as CDK5 active; $P(+)$ stands for the prior probability, a probability calculated from various compounds in the training set; and $P(\mathrm{A}_1, \ldots, \mathrm{A}_n)$ represents the marginal probability of the given descriptors which will occur in the training set.

Recursive partitioning classifiers

Recursive partitioning (RP) in Discovery Studio 3.1 was employed to build decision trees to categorize molecules into CDK5 inhibitors and non-inhibitors. RP is a classification approach for interpreting the relationship between a dependent property Y (activity class: 1 or 0) to a set of independent molecular descriptors X. Models are developed by dividing the training set into progressively homogeneous subsets until it is infeasible to proceed according to some "stopping rules" [54].

The result of an RP model can be depicted by a "decision tree" or "graph", which can be used to make predictions about data. When making a prediction with an RP model, a sample is assigned to a certain node of the tree. The class predicted for that sample is the class which the majority of the training samples in the node belong to. In this work, a fivefold cross-validation was adopted to identify the degree of pruning for the best predictive accuracy. The specific parameters for our RP model were set as follows: minimum number of samples at each node, the maximum tree depth, and the maximum tree depth were 10, 20, and 20, respectively. The 1,440 compounds present in the training set were used to develop the decision trees, whereas 522 compounds from the test set and 496 compounds from the external test set were used to evaluate the predictive performance of the models.

Consensus models

Consensus models were also constructed to integrate the four single classifiers (two from NB classifiers and two from RP classifiers). Consensus scoring or data fusion can enhance the performance of a single classifier by improving prediction reliability [55,56]. In general, a single model contains various kinds of noise, which may be reduced by consensus modelling. Various strategies can be adopted in consensus analysis, such as ranking, artificial neural network fusion, or the use of conditional probability. In this study, two different consensus approaches were used: (a) four single classifiers were combined by fusing with a back-propagation artificial neural network algorithm, and (b) a consensus prediction protocol was applied to enrich datasets by merging predictive results obtained from four single classifiers.

Back-propagation artificial neural networks (BP-ANNs)

BP-ANN is a generalization of the delta rule for training multi-layer feed-forward neural networks with non-linear units [57]. Previous work has shown that an ANN combiner compared favourably to other combination methods in a pattern recognition problem [58]. In this study, as shown in Fig. 2a, the probability output (P_i^{+1} and P_i^{-1} $\;i = 1, 2, 3, 4$) for each compound was predicted with four single classifiers; then, all of these probability outputs were selected as new descriptors to develop a combined classifier BP-ANN (CC-ANN) model that would generate the final combination decision probability (P_C^{+1} and P_C^{-1}). To compare the performance of the CC-ANN model with the single BP-ANN models, two single BP-ANN models (S8-ANN and S11-ANN) were also developed (Fig. 2b).

Each network contains three layers: an input layer, a hidden layer, and an output layer. For each BP-ANN model, the amount of input layer neurons is set to the number of the descriptors. For example, both the CC-ANN and S8-ANN models have eight descriptors; therefore, their amount of input neurons is set to eight. In the same manner, there are 11 input neurons for S11-ANN. For all models, the amount of hidden layer neurons was set at 10. The scaled conjugate gradient back propagation (trainscg) was used to update weights and biases and the tangent sigmoid transfer functions (tansig) for the hidden layer and the output layer. The training set (1,440 compounds) was used for training, and the network was adjusted according to its error. Training was automatically stopped when the resulting generalization stopped improving. The test set (522 compounds) was used to provide an independent measure of network error during and after training.

Consensus prediction

Based on the prediction results of the four single classifiers, four new consensus models were generated using a "consensus prediction" procedure [56]. First, we screened the training set, test set, and external test set with four single classifiers, and then the compounds, which were predicted as "+1" by one of the four single classifiers, were considered as CDK5 inhibitors. This procedure is called consensus prediction C1. Similarly, we performed consensus prediction C2 (predicted as "+1" by two of the four single classifiers), C3 (predicted as "+1" by three of the four single classifiers), and C4 (predicted as "+1" by all of the four single classifiers).

Model validation

The performance of the developed models was evaluated by the quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are the number of CDK5 inhibitors predicted as CDK5 inhibitors, non-inhibitors predicted as non-inhibitors, non-inhibitors predicted as CDK5 inhibitors, and CDK5 inhibitors predicted as non-inhibitors, respectively. Several accuracy functions were used to measure prediction performance: sensitivity (SE), specificity (SP), prediction accuracy of CDK5 inhibitors (Q^+), prediction accuracy of non-inhibitors (Q^-), and Matthews correlation coefficient (MCC), which are given by Eqs. 2–6. A more detailed interpretation for these equations can be found in the following references [59,60]. This set of metrics is valid only for single-label systems. For multi-label systems in system biology [61] and system medicine [62], a completely different set of metrics as defined in the reference [63] is needed.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{2}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{3}$$

$$Q^+ = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4}$$

$$Q^- = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}, \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}}. \tag{6}$$

The high performance of the test set does not signify that the model can also exhibit good predictive performance for an external test set. To avoid this complication, a dataset including 96 CDK5 inhibitors ($\mathrm{IC}_{50} < 10\ \mu\mathrm{M}$) collected from recent literature [64–67] and 400 decoys not contained in the training and test set was used as an external test set to further evaluate the performance of the classification models (see Table S3).

In vitro CDK5/p35 test

The CDK5/p35 inhibitory activity of the selected compounds was tested using the Kinase-Glo luminescent technique, which is a safe non-radioactive assay [68]. (R)-Roscovitine (Sigma-Aldrich catalog number R7772) was used as reference. The substrate Histone H1 peptide (PKTPKKAKKL) was synthesized by Apeptide Co., Ltd (Shanghai, China), and adenosine triphosphate (ATP) was purchased from Roche. Recombinant full-length human CDK5/p35 (catalog number V3271) and Kinase-Glo luminescent kinase assays (catalog number V6713) were supplied by Promega Corporation (Madison, WI). CDK5/p35 (10 μg, 0.1 μg/μL) was diluted 10 times in reaction Buffer A (4×) (160 mM Tris-HCl, pH 7.5, 100 mM MgCl_2, and 0.4 μg/μLBSA).

Kinase-Glo assays were carried out in opaque white-walled 384-well microtiter plates. The reaction system consisted of 2.5 μL of test compounds, 5 μL of CDK5/p35 kinase (50 ng), and 5 μL of a complex of Histone H_1 peptide and adenosine triphosphate (ATP). After 1 h of incubation at 37 °C, the enzymatic reaction was stopped by the addition of 12.5 μL of Kinase-Glo Reagent. Then the plate was mixed, incubated for 1 h at 37 °C, and read with a SpectraMax M5 (Molecular Devices, Sunnyvale, CA, USA) after 10 min. The mean of three independent experiments was calculated to express the data.

Results and discussion

Chemical space analysis

The chemical space of the training set (1,440 compounds), test set (522 compounds), and external test set (496 compounds) was investigated using principal component analy-

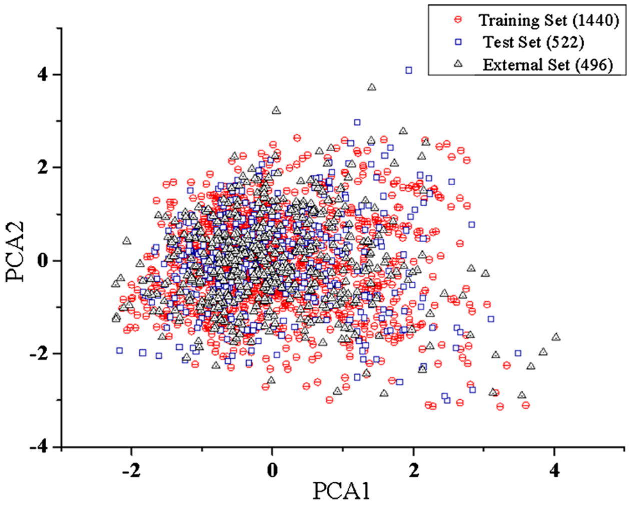


Fig. 3 Diversity distribution of the training set ($n = 1,440$ compounds), test set ($n = 522$ compounds), and external test set ($n = 496$ compounds) as described by principal component analysis (PCA)

sis (PCA) [69]. The input variables were the 11 molecular descriptors selected by the Pearson correlation analysis and stepwise regression. According to the chemical space defined by PCA (Fig. 3), there are diverse chemical space distributions for all compounds, and most of the compounds in the test set and the external test set are well within the chemical space of the training set.

Performance of binary classification models by single classifier

The classification performance of the four single classifiers was collected, and the results are given in Table 2. In Table 2, the statistical results for the training sets were achieved using fivefold cross-validation. In general, the performance of the recursive partitioning (RP) models was superior to the naive Bayesian (NB) models, which was in accordance with the cross-validation results for training set. The best values of sensitivity (SE), specificity (SP), prediction accuracy of CDK5 inhibitor activity (Q^+), prediction accuracy of non-inhibitors (Q^-), and Matthew's correlation coefficient (MCC) on the test set were 0.853, 0.850, 0.580, 0.960, and 0.616, respectively. All of these values were obtained for the RP-b model using the recursive partitioning algorithm. Similarly, the performance of SE (0.885), SP (0.848), Q^+(0.582), Q^-(0.969), and MCC (0.635) on the external test set was obtained for the RP-b model. When compared to the recursive partitioning performance, the Bayesian models performed worse. For example, with the same descriptor combinations, the MCC of the Bayesian-a model on the test set was 0.445, while that of the RP-a model was 0.539. Similarly, the MCC of the Bayesian-a model on the external test set was 0.450, while that of the RP-a model was 0.661. The same explanation accounts for the MCC differences (0.524 and 0.616) between the Bayesian-b and RP-b models on the test set and the MCC differences (0.462 and 0.635) between the Bayesian-b and RP-b models on the external test set.

Additionally, the Bayesian-b and RP-b models, generated by Pearson correlation analysis and stepwise regression, performed better than those built with the default eight descriptors given in Discovery Studio. For example, when using the same algorithm, the MCC of the Bayesian-a model on the test set was 0.445, which was lower than the one (MCC = 0.524) obtained for the Bayesian-b model. The same explanation accounts for the MCC differences (0.539 and 0.616) between the RP-a and RP-b models on the test set.

Performance of consensus models

Back-propagation artificial neural networks (BP-ANNs)

Three BP-ANN models were constructed to compare the performance of the combined classifier-artificial neural network (CC-ANN) with two single BP-ANN models (S8-ANN and S11-ANN). Here, S8-ANN represented an artificial neural network model built by the eight DS_2D_default descriptors, and S11-ANN represented a network built by the 11 molecular descriptors obtained by descriptor selection. The performance of these models is given in Table 3. All the statistical results for the training sets were obtained using fivefold cross-validation. All of the models were trained by the training set (1,440 compounds) and evaluated by the test

Table 2 Performance of four single classifier models

Model	Descriptors	Training set (1,440)					Test set (522)					External test set (496)				
		SE	SP	Q^+	Q^-	MCC	SE	SP	Q^+	Q^-	MCC	SE	SP	Q^+	Q^-	MCC
Bayesian-a	8	0.850	0.714	0.498	0.935	0.494	0.843	0.707	0.411	0.949	0.445	0.813	0.738	0.426	0.942	0.450
Bayesian-b	11	0.867	0.778	0.565	0.946	0.574	0.853	0.776	0.481	0.956	0.524	0.760	0.785	0.459	0.932	0.462
RP-a	8	0.925	0.887	0.732	0.973	0.756	0.794	0.826	0.526	0.943	0.539	0.813	0.900	0.661	0.952	0.661
RP-b	11	0.950	0.877	0.720	0.981	0.762	0.853	0.850	0.580	0.960	0.616	0.885	0.848	0.582	0.969	0.635

Table 3 Performance comparison of combined BP-ANN to single BP-ANN

Model	Training set (1,440)					Test set (522)					External test set (496)				
	SE	SP	Q^+	Q^-	MCC	SE	SP	Q^+	Q^-	MCC	SE	SP	Q^+	Q^-	MCC
CC-ANN	0.863	0.956	0.856	0.958	0.816	0.836	0.961	0.871	0.949	0.808	0.750	0.988	0.938	0.943	0.806
S11-ANN	0.683	0.928	0.745	0.905	0.630	0.682	0.933	0.762	0.904	0.640	0.450	0.934	0.616	0.876	0.434
S8-ANN	0.541	0.933	0.706	0.872	0.523	0.533	0.940	0.770	0.855	0.541	0.646	0.939	0.717	0.917	0.608

CC-ANN combined classifier-artificial neural network model
S11-ANN artificial neural networks model built based on 11 molecular descriptors
S8-ANN artificial neural networks model built based on 8 DS_2D_default descriptors

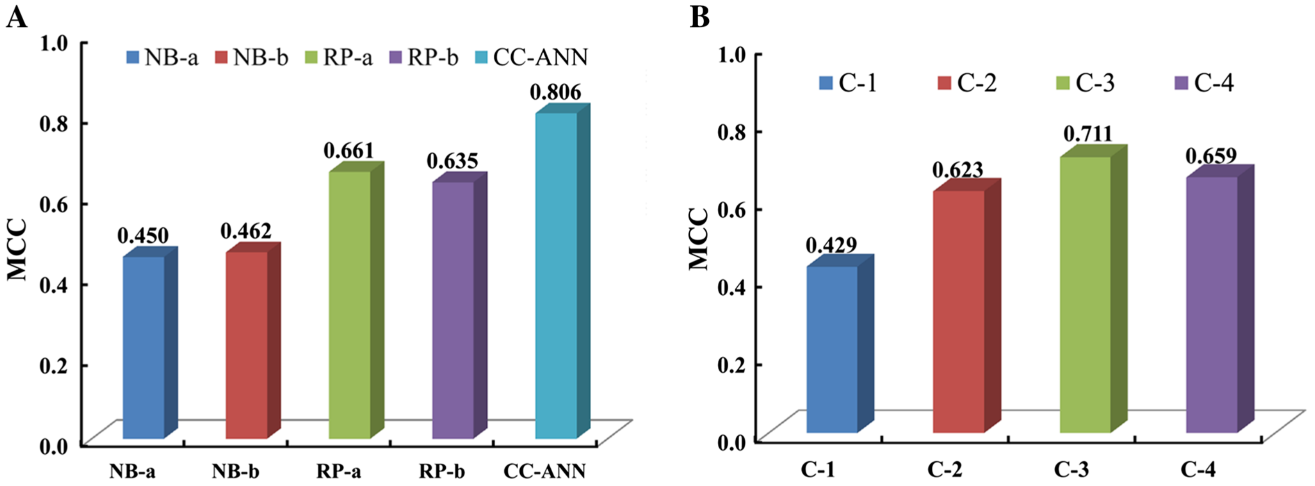


Fig. 4 Comparison of the MCC values for four single classifiers and the CC-ANN model (**a**) and the four consensus prediction models (**b**) for the external test set (496 compounds)

set (522 compounds), and the highest performance among the three models was obtained for the CC-ANN model.

As shown in Table 3, the values of SE, SP, Q^+, Q^-, and MCC for the CC-ANN model on the training set were 0.863, 0.956, 0.856, 0.958, and 0.816, respectively, which were significantly higher than those obtained by the S11-ANN (0.683, 0.928 0.745, 0.905, and 0.630) and S8-ANN (0.541, 0.933, 0.706, 0.872, and 0.523) models. In addition, the CC-ANN model also exhibited significantly higher values of SE (0.836), SP (0.961), Q^+(0.871), Q^-(0.949), and MCC (0.808) on the test set than those obtained for the S11-ANN (0.682, 0.933, 0.762, 0.904, and 0.640) and S8-ANN (0.533, 0.940, 0.770, 0.855, and 0.541) models.

A similar phenomenon occurred when the external test set (496 compounds) was used. As shown in Table 3, the MCC value of the CC-ANN model was 0.806, which was much higher than the values obtained for the S11-ANN (MCC = 0.434) and S8-ANN (MCC = 0.608) models. In addition, the CC-ANN model also exhibited higher values for SE (0.750), SP (0.988), Q^+(0.938), and Q^-(0.943) on the external test set than observed for the S11-ANN and S8-ANN models.

Moreover, the MCC value was also compared among the four single classifiers (NB-a, NB-b, RP-a, and RP-b) and the CC-ANN model using the external test set (496 compounds). As shown in Fig. 4a, the MCC value for the CC-ANN model on the external test set was 0.806, which was much higher than that obtained for any single classifier (ranging from 0.450 to 0.661).

As discussed above, the overall performance of the combined classifier fused with BP-ANN (CC-ANN) was better than two single BP-ANN models (S11-ANN and S8-ANN) and four single classifiers. Why did the CC-ANN model perform better than the single model? The key step in building the CC-ANN model was to obtain probability output for every single classifier. The single classifier here could be considered as a mechanism that assisted in the conversion of the initial descriptors into new descriptors with a more reasonable orientation. Subsequently, all of the spatial positions for each chemical instance were rearranged, and BP-ANNs were then applied to minimize the total error of the output (P_C^{+1} and P_C^{-1}) computed by the network. Therefore, the CC-ANN model performed better than any single model.

Table 4 Performance of four consensus prediction models

Model	Training set (1,440)					Test set (522)					External test set (496)				
	SE	SP	SE	Q^-	MCC	SE	SP	Q^+	Q^-	MCC	SE	SP	Q^+	Q^-	MCC
C-1	0.994	0.619	0.465	0.997	0.532	0.990	0.560	0.353	0.996	0.438	1.000	0.538	0.342	1.000	0.429
C-2	0.969	0.774	0.589	0.987	0.654	0.922	0.752	0.475	0.975	0.551	0.938	0.808	0.539	0.982	0.623
C-3	0.897	0.896	0.743	0.963	0.748	0.794	0.890	0.638	0.947	0.633	0.813	0.928	0.729	0.954	0.711
C-4	0.750	0.975	0.909	0.921	0.776	0.637	0.957	0.783	0.916	0.645	0.500	0.998	0.980	0.893	0.659

Consensus prediction

The performance of four consensus prediction models is presented in Table 4. Among the four models, the highest performance was obtained for the C-3 model, which maintained a reasonable balance between specificity and sensitivity on both the training set (SE = 0.897, SP = 0.896) and the test set (SE = 0.794, SP = 0.890). The MCC values of the C-3 model on the training set and test set were 0.748 and 0.633, respectively, which were higher than those of the C-1 (0.532 and 0.438) and C-2 (0.654 and 0.551) models but less than those of the C-4 (0.776 and 0.645) model. However, the overall performance of the C-3 model on the external test set was superior to that of the C-4 model. For the C-3 model, the values of sensitivity, specificity, Q^+, Q^-, and MCC were 0.813, 0.928, 0.729, 0.954, and 0.711, respectively, whereas the corresponding values for the C-4 model on the external validation set were 0.500, 0.998, 0.980, 0.893, and 0.659, respectively.

The performance was compared among the four single classifiers and the four consensus prediction models on the external test set. As shown in Fig. 4a and b, the consensus prediction for the C-1 model (predicted as "+1" by at least one of the four single classifiers) had an MCC value of 0.429, which was lower than that of any model derived from the four single classifiers. This result was likely due to the loose criterion for the compounds predicted as "+1", which made more noise compounds got involved in. For the C-1 and C-3 models, the MCC value gradually increased from 0.429 to 0.711 as the selection criterion became progressively stricter. However, the C-4 model (predicted as "+1" by all four single classifiers) only had an MCC value of 0.659. This result suggested that the selection conditions of the compounds were so strict that fewer true positives (48 out of 96 compounds) were detected. Thus, the C-3 model exhibited the best balance and it performed better than the best single classifier RP-a (MCC = 0.661).

Comparing the CC-ANN model to the C-3 consensus prediction model

To exploit the differences in performance between the CC-ANN and C-3 consensus prediction models, the performance comparisons of the two models on the external test set were plotted. As given in Fig. 5, the performance of the CC-ANN model was better than that of the C-3 model. The values of SE, SP, Q^+, and Q^- for the external test set in the CC-ANN model were 0.75, 0.988, 0.938, and 0.943, respectively, while those for the C-3 model were 0.813, 0.928, 0.729, and 0.954. Both of the MCC values were much larger than 0.5. This result suggested that both of the combined models exhibited excellent accuracy prediction capabilities. At the same time, the MCC value (0.806) of the CC-ANN model was obviously higher than that (0.711) of the C-3 model, which implied that the combined method fused in the ANN model possessed a competitive advantage over the consensus prediction method in this study.

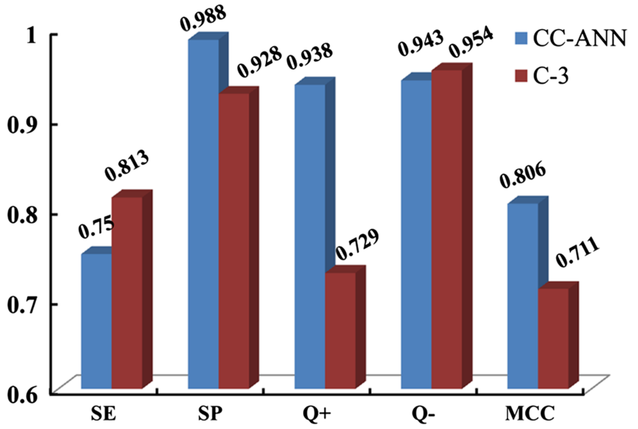


Fig. 5 Performance comparisons between the CC-ANN and C-3 models on the external test set (496 compounds)

Virtual screening of an in-house database for CDK5 inhibitors with CC-ANN

To illustrate the application of our best CC-ANN model and search for CDK5 inhibitors, we implemented a virtual screen of our in-house database, which belongs to the National Center for Pharmaceutical Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences. This database included 29,170 compounds not reported to be CDK5 inhibitors. First, the structure of each compound was converted into 11 two-dimensional descriptors selected using a

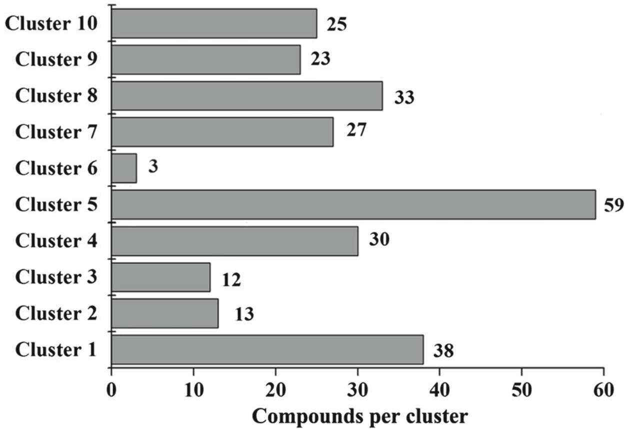


Fig. 6 Structure clustering of 263 compounds by FCFP_6 fingerprint

Table 5 IC_{50} values (μM) of 13 CDK5/p35 hits

No.	MW	P_C^{+1} (CC-ANN) %	IC50 (μM/L) ± SD
(R)-Roscovitine	354.45	–	0.36 ± 0.19
J18911	363.45	98.13	9.23 ± 2.33
J18842	396.45	98.32	22.45 ± 5.14
J14683	258.23	98.26	29.80 ± 4.28
J18854	412.91	98.27	36.70 ± 2.42
J18836	392.49	98.37	41.35 ± 0.90
J18803	446.46	97.99	51.97 ± 4.84
J18848	396.45	98.28	57.05 ± 4.29
J18879	379.40	98.18	68.26 ± 11.65
J18811	464.45	97.98	95.57 ± 4.89
J18809	446.46	97.98	139.19 ± 10.64
J18723	353.38	98.19	154.23 ± 38.04
J33146	441.58	98.14	187.30 ± 9.70
J18999	357.41	98.25	229.76 ± 26.35

(R)-Roscovitine served as reference compound for the bioassay

Pearson correlation analysis and stepwise regression, which were calculated by Discovery Studio software, and then eight probability outputs (P_i^{+1} and P_i^{-1} $i = 1, 2, 3, 4$) were predicted for each compound using four single classifiers. Additionally, each compound had the eight new descriptors (the eight probability outputs), and then two final combination decision probabilities (P_C^{+1} and P_C^{-1}) were predicted using the CC-ANN model. Out of the 29,710 compounds screened, 263 compounds received both P_C^{+1} values higher than 0.95 and P_C^{-1}values lower than 0.05 (Tables S4, S5), and were chosen for further studies.

In addition, as shown in Fig. 6, these 263 compounds were clustered into 10 groups based on their FCFP_6 fingerprint using the Cluster ligands module in Discovery Studio. Clustering is based on the root-mean-square (RMS) difference of the Tanimoto distance for fingerprinting. For each cluster, compounds with the top 50 % CC-ANN scores were selected. After that, 132 compounds were obtained. Then, 60 compounds were chosen from the remaining 132 compounds by visual evaluation. Finally, 40 of the remaining 60 compounds were pulled from our in-house compound library for in vitro CDK5/p35 inhibitory assay analysis.

In vitro CDK5/p35 inhibitory assays

For these studies, (R)-Roscovitine (a known CDK5 inhibitor, $IC_{50} = 0.36$ μM) served as the reference compound. Among the 40 tested compounds, nine hits exhibited moderate ($IC_{50} < 100$ μM) inhibitory activity towards CDK5/p35, and four hits exhibited low activity (100 μM $< IC_{50} <$ 250 μM) towards CDK5/p35. J18911, as the best hit found in this study, has an IC_{50} value of 9.23 μM (Table 5). Because we had a success rate of 33 % in this virtual screening protocol, these results are encouraging.

Inhibitory curves of the five best hits (J18911, J18842, J14683, J18854, and J18836) and the reference compound (R)-Roscovitine towards CDK5/p35 are given in Fig. 7. The chemical structures of all of the 13 CDK5/p35 hits are shown in Fig. 8. Among these 13 hit compounds, J18911 ($IC_{50} = 9.23$ μM) is the most active compound and can be further modified structurally to increase its inhibitory potency towards CDK5.

The structures of 13 CDK5/p35 hits can be classified into four scaffolds (Fig. 9), including scaffold A (J18911, J18842, J18854, J18836, J18803, J18848, J18879, J18811, J18809, and J18723), scaffold B (J14683), scaffold C (J33146), and scaffold D (J18999). With the exception of scaffold B, scaffolds A, C, and D are new compound classes not previously reported as CDK5 inhibitors.

Conclusions

In this study, we developed reliable consensus models (CC-ANN and consensus prediction) to discriminate between CDK5 inhibitors and non-inhibitors by means of combining different single classifiers. The consensus model CC-ANN exhibited the best performance and was successfully applied to the discovery of CDK5 inhibitors. Based on the virtual screening results of the CC-ANN model, 33.3 % of compounds displayed moderate or low CDK5/p35 inhibitory activity. Among the 13 hits, a new compound J18911 was discovered exhibiting an IC_{50} of 9.23 μM towards CDK5/p35, which suggested that the virtual screening tool utilizing the CC-ANN model could be used to discover new scaffold inhibitors.

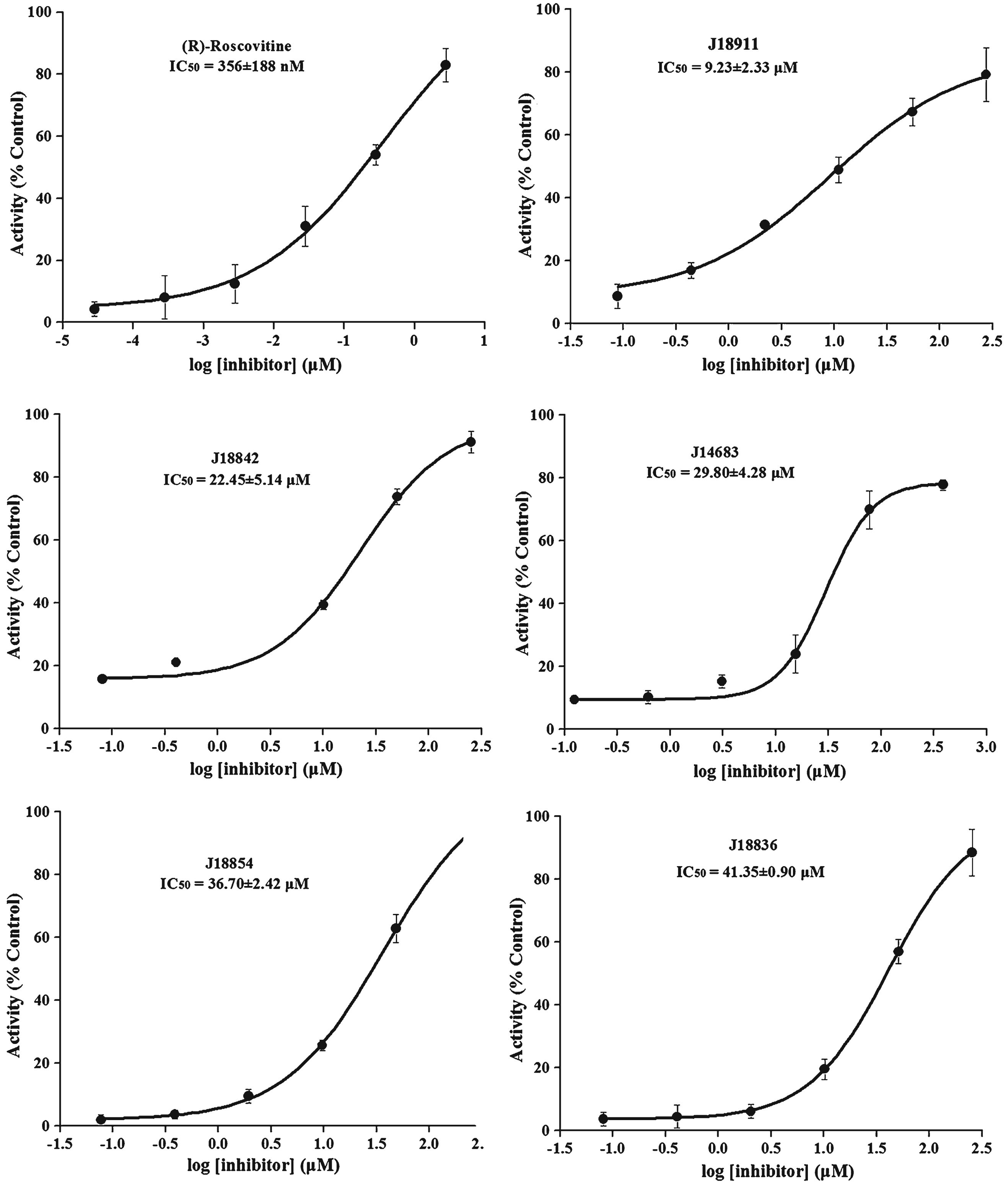


Fig. 7 Inhibitory curves of the five best hits and the reference compound (R)-Roscovitine towards CDK5/p35 activity

Fig. 8 Chemical structures and IC_{50} values of 13 hit compounds against CDK5/p35

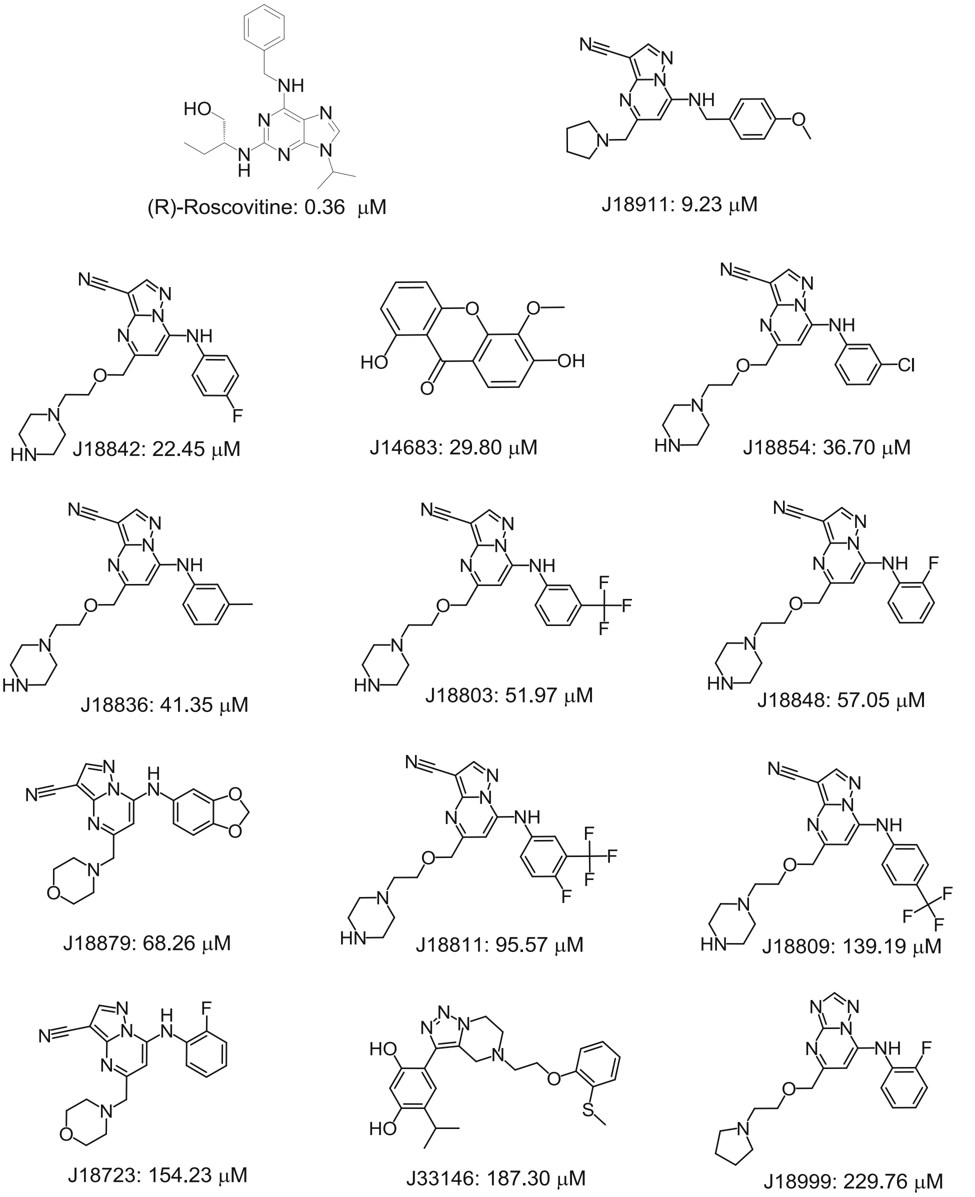
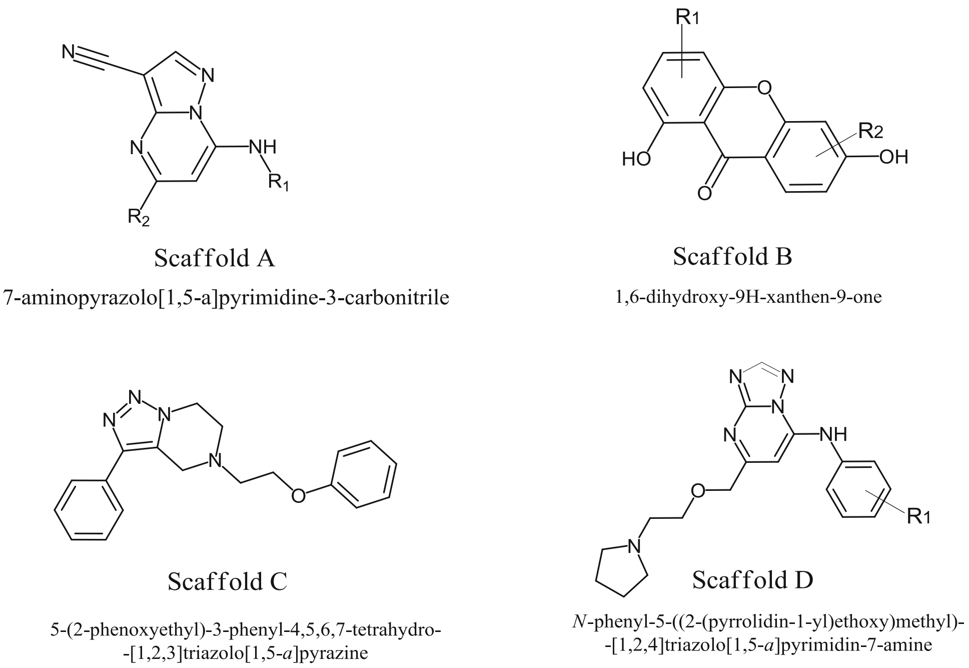


Fig. 9 Scaffold classification of 13 CDK5/p35 hits found in this study



Scaffold A

7-aminopyrazolo[1,5-a]pyrimidine-3-carbonitrile

Scaffold B

1,6-dihydroxy-9H-xanthen-9-one

Scaffold C

5-(2-phenoxyethyl)-3-phenyl-4,5,6,7-tetrahydro- -[1,2,3]triazolo[1,5-*a*]pyrazine

Scaffold D

N-phenyl-5-((2-(pyrrolidin-1-yl)ethoxy)methyl)- -[1,2,4]triazolo[1,5-*a*]pyrimidin-7-amine

Supporting information

The structures (in SMILE format) of the 1,440 compounds of the training set, 522 compounds of the test set, 496 compounds of the external test set with the corresponding activities (Tables S1–S3), and the structures (in SMILE format) of 263 compounds with the value of P_C^{+1} higher than 0.95 and P_C^{-1} lower than 0.05 are located in Tables S4 and S5.

Acknowledgments This work was funded in part by the Research Special Fund for Public Welfare Industry of Health (No. 200802041), the National Great Science and Technology Projects (2012ZX09301002, 2014ZX09507003-002), the International Collaboration Project (2011DFR31240), and Peking Union Medical College graduate student innovation fund (2013-1007-18)

Conflict of interest The authors declare no competing financial interest.

References

1. Melnikova I (2007) Therapies for Alzheimer's disease. Nat Rev Drug Discov 6:341–342. doi:10.1038/nrd2314
2. Querfurth HW, LaFerla FM (2010) Alzheimer's disease. N Engl J Med 362:329–344. doi:10.1056/NEJMra0909142
3. Dudash K (2011) Alzheimer's disease: new therapies and the role of biomarkers. Biotechnol Healthc 8:22–23
4. Dhavan R, Tsai LH (2001) A decade of cdk5. Nat Rev Mol Cell Biol 2:749–759. doi:10.1038/35096019
5. Cruz JC, Tsai LH (2004) Cdk5 deregulation in the pathogenesis of Alzheimer's disease. Trends Mol Med 10:452–458. doi:10.1016/j.molmed.2004.07.001
6. Goedert M (1993) Tau protein and the neurofibrillary pathology of Alzheimer's disease. Trends Neurosci 16:460–465. doi:10.1111/j.1749-6632.1996.tb34410.x
7. Tolnay M, Probst A (1999) REVIEW: tau protein pathology in Alzheimer's disease and related disorders. Neuropathol Appl Neurobiol 25:171–187. doi:10.1046/j.1365-2990.00182.x
8. Patrick GN, Zukerberg L, Nikolic M, Monte S, Dikkes P, Tsai LH (1999) Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. Nature 402:615–622. doi:10.1038/45159
9. Lau LF, Patricia AS, Mark AS, Schachter JB (2002) Cdk5 as a drug target for the treatment of Alzheimer's disease. J Mol Neurosci 19:267–270. doi:10.1385/JMN:19:3:267
10. Hosoi T, Uchiyama M, Okumura E, Saito T, Ishiguro K, Uchida T, Okuyama A, Kishimoto T, Hisanaga S (1995) Evidence for cdk5 as a major activity phosphorylating tau protein in porcine brain extract. J Biochem 117:741–749
11. Cardone A, Hassan SA, Albers RW, Sriram RD, Pant HC (2010) Structural and dynamic determinants of ligand binding and regulation of cyclin-dependent kinase 5 by pathological activator p25 and inhibitory peptide CIP. J Mol Biol 401:478–492. doi:10.1016/j.jmb.2010.06.040
12. Ahlijanian MK, Barrezueta NX, Williams RD, Jakowski A, Kowsz KP, McCarthy S, Coskran T, Carlo A, Seymour PA, Burkhardt JE, Nelson RB, McNeish JD (2000) Hyperphosphorylated tau and neurofilament and cytoskeletal disruptions in mice overexpressing human p25, an activator of cdk5. Proc Natl Acad Sci USA 97:2910–2915. doi:10.1073/pnas.040577797
13. Fischer A, Sananbenesi F, Pang PT, Lu B, Tsai LH (2005) Opposing roles of transient and prolonged expression of p25 in synaptic plasticity and hippocampus-dependent memory. Neuron 48:825–838. doi:10.1016/j.neuron.2005.10.033
14. Cruz JC, Tseng HC, Goldman JA, Shih H, Tsai LH (2003) Aberrant Cdk5 activation by p25 triggers pathological events leading to neurodegeneration and neurofibrillary tangles. Neuron 40:471–483. doi:10.1016/S0896-6273(03)00627-5
15. Chou KC, Watenpaugh KD, Heinrikson RL (1999) A model of the complex between cyclin-dependent kinase 5 (Cdk5) and the activation domain of neuronal Cdk5 activator. Biochem Biophys Res Commun 259:420–428. doi:10.1006/bbrc.1999.0792
16. Chou KC (2004) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor.

Biochem Biophys Res Commun 319:433–438. doi:10.1016/j.bbrc.2004.05.016

17. Chou KC (2004) Insights from modeling the tertiary structure of human BACE2. J Proteome Res 3:1069–1072. doi:10.1021/pr049905s
18. Chou KC (2005) Modeling the tertiary structure of human cathepsin-E. Biochem Biophys Res Commun 331:56–60. doi:10.1016/j.bbrc.2005.03.123
19. Chou KC, Howe WJ (2002) Prediction of the tertiary structure of the beta-secretase zymogen. Biochem Biophys Res Commun 292:702–708. doi:10.1006/bbrc.2002.6686
20. Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134. doi:10.2174/0929867043364667
21. Chou KC, Luan CH, Chou KC, Johnson GV (2002) Identification of the N-terminal functional domains of Cdk5 by molecular truncation and computer modeling. Proteins 48:447–453. doi:10.1002/prot.10173
22. Pielak RM, Schnell JR, Chou JJ (2009) Mechanism of drug inhibition and drug resistance of influenza A M2 channel. Proc Natl Acad Sci USA 106:7379–7384. doi:10.1073/pnas.0902548106
23. Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem Biophys Res Commun 308:148–151. doi:10.1016/S0006-291X(03)01342-1
24. Du QS, Huang RB, Chou KC (2008) Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design. Curr Protein Pept Sci 9:248–260. doi:10.2174/138920308784534005
25. Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008) Multiple field three dimensional quantitative structure–activity relationship (MF-3D-QSAR). J Comput Chem 29:211–219. doi:10.1002/jcc.20776
26. Du QS, Huang RB, Wei YT, Pang ZW, Du LQ, Chou KC (2009) Fragment-based quantitative structure–activity relationship (FB-QSAR) for fragment-based drug design. J Comput Chem 30:295–304. doi:10.1002/jcc.21056
27. Prado-Prado FJ, Martinez de la Vega O (2009) Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug–drug complex networks. Bioorg Med Chem 17:569–575. doi:10.1016/j.bmc.2008.11.075
28. Fang J, Huang D, Zhao W, Ge H, Luo HB, Xu J (2011) A new protocol for predicting novel GSK-3β ATP competitive inhibitors. J Chem Inf Model 51:1431–1438. doi:10.1021/ci2001154
29. Fang J, Yang R, Gao L, Zhou D, Yang S, Liu AL, Du GH (2013) Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. J Chem Inf Model 53:3009–3020. doi:10.1021/ci400331p
30. Singh N, Chaudhury S, Liu R, AbdulHameed MD, Tawa G, Wallqvist A (2012) QSAR classification model for antibacterial compounds and its use in virtual screening. J Chem Inf Model 52:2559–2569. doi:10.1021/ci300336v
31. Myint KZ, Wang L, Tong Q, Xie XQ (2012) Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm 9:2912–2923. doi:10.1021/mp300237z
32. Fernandez M, Tundidor-Camba A, Caballero J (2005) Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo [3,4-d]pyrimidine derivatives using artificial neural network ensembles. J Chem Inf Model 45:1884–1895. doi:10.1021/ci050263i
33. Li J, Liu H, Yao X, Liu M, Hu Z, Fan B (2007) Structure–activity relationship study of oxindole-based inhibitors of cyclin-dependent kinases based on least-squares support vector machines. Anal Chim Acta 581:333–342. doi:10.1016/j.aca.2006.08.031
34. Ducrot P, Legraverend M, Grierson DS (2000) 3D-QSAR CoMFA on cyclin-dependent kinase inhibitors. J Med Chem 43:4098–4108. doi:10.1021/jm000965t
35. Kunick C, Lauenroth K, Wieking K, Xie X, Schultz C, Gussio R, Zaharevitz D, Leost M, Meijer L, Weber A, Jorgensen FS, Lemcke T (2004) Evaluation and comparison of 3D-QSAR CoMSIA models for CDK1, CDK5, and GSK-3 inhibition by paullones. J Med Chem 47:22–36. doi:10.1021/jm0308904
36. Singh SK, Dessalew N, Bharatam PV (2006) 3D-QSAR CoMFA study on indenopyrazole derivatives as cyclin dependent kinase 4 (CDK4) and cyclin dependent kinase 2 (CDK2) inhibitors. Eur J Med Chem 41:1310–1319. doi:10.1016/j.ejmech.2006.06.010
37. Singh SK, Dessalew N, Bharatam PV (2007) 3D-QSAR CoMFA study on oxindole derivatives as cyclin dependent kinase 1 (CDK1) and cyclin dependent kinase 2 (CDK2) inhibitors. Med Chem 3:75–84. doi:10.2174/157340607779317517
38. Caballero J, Fernandez M, Gonzalez-Nilo FD (2008) Structural requirements of pyrido[2,3-d]pyrimidin-7-one as CDK4/D inhibitors: 2D autocorrelation, CoMFA and CoMSIA analyses. Bioorg Med Chem 16:6103–6115. doi:10.1016/j.bmc.2008.04.048
39. Babu PA, Smiles DJ, Narasu ML, Srinivas K (2008) Identification of novel CDK2 inhibitors by QSAR and virtual screening procedures. QSAR Comb Sci 27:1362–1373. doi:10.1002/qsar.200860041
40. Dessalew N, Singh SK (2008) 3D-QSAR CoMFA and CoMSIA study on benzodipyrazoles as cyclin dependent kinase 2 inhibitors. Med Chem 4:313–321. doi:10.2174/157340608784872244
41. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res 35:D198–201. doi:10.1093/nar/gkl999
42. Chioua M, Samadi A, Soriano E, Lozach O, Meijer L, Marco-Contelles J (2009) Synthesis and biological evaluation of 3,6-diamino-1H-pyrazolo[3,4-b]pyridine derivatives as protein kinase inhibitors. Bioorg Med Chem Lett 19:4566–4569. doi:10.1016/j.bmcl.2009.06.099
43. Helal CJ, Kang Z, Lucas JC, Gant T, Ahlijanian MK, Schachter JB, Richter KE, Cook JM, Menniti FS, Kelly K, Mente S, Pandit J, Hosea N (2009) Potent and cellularly active 4-aminoimidazole inhibitors of cyclin-dependent kinase 5/p25 for the treatment of Alzheimer's disease. Bioorg Med Chem Lett 19:5703–5707. doi:10.1016/j.bmcl.2009.08.019
44. Jain P, Flaherty PT, Yi S, Chopra I, Bleasdell G, Lipay J, Ferandin Y, Meijer L, Madura JD (2011) Design, synthesis, and testing of an 6-O-linked series of benzimidazole based inhibitors of CDK5/p25. Bioorg Med Chem 19:359–373. doi:10.1016/j.bmc.2010.11.022
45. Kassis P, Brzeszcz J, Beneteau V, Lozach O, Meijer L, Le Guevel R, Guillouzo C, Lewinski K, Bourg S, Colliandre L, Routier S, Merour JY (2011) Synthesis and biological evaluation of new 3-(6-hydroxyindol-2-yl)-5-(Phenyl) pyridine or pyrazine V-Shaped molecules as kinase inhibitors and cytotoxic agents. Eur J Med Chem 46:5416–5434. doi:10.1016/j.ejmech.2011.08.048
46. Laha JK, Zhang X, Qiao L, Liu M, Chatterjee S, Robinson S, Kosik KS, Cuny GD (2011) Structure–activity relationship study of 2,4-diaminothiazoles as Cdk5/p25 kinase inhibitors. Bioorg Med Chem Lett 21:2098–2101. doi:10.1016/j.bmcl.2011.01.140
47. Shiradkar M, Thomas J, Kanase V, Dighe R (2011) Studying synergism of methyl linked cyclohexyl thiophenes with triazole: synthesis and their cdk5/p25 inhibition activity. Eur J Med Chem 46:2066–2074. doi:10.1016/j.ejmech.2011.02.059
48. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55:6582–6594. doi:10.1021/jm300687e

49. Chemical Computing Group Inc. (2010) Molecular Operating Environment (MOE), version 2010.10. Chemical Computing Group Inc., Montreal, Canada
50. Accelrys Inc. (2012) Discovery Studio 3.1. Accelrys Inc., San Diego. http://www.accelrys.com
51. Wang L, Wang M, Yan A, Dai B (2013) Using self-organizing map (SOM) and support vector machine (SVM) for classification of selectivity of ACAT inhibitors. Mol Divers 17:85–96. doi:10.1007/s11030-012-9404-z
52. MathWorks Inc. Matlab Version 7.5.0.342 (R2007b). MathWorks Inc., Natick
53. Xia X, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. J Med Chem 47:4463–4470. doi:10.1021/jm0303195
54. Lamanna C, Bellini M, Padova A, Westerberg G, Maccari L (2008) Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process. J Med Chem 51:2891–2897. doi:10.1021/jm701407x
55. Cheng F, Yu Y, Shen J, Yang L, Li W, Liu G, Lee PW, Tang Y (2011) Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. J Chem Inf Model 51:996–1011. doi:10.1021/ci200028n
56. Mansouri K, Ringsted T, Ballabio D, Todeschini R, Consonni V (2013) Quantitative structure–activity relationship models for ready biodegradability of chemicals. J Chem Inf Model 53:867–878. doi:10.1021/ci4000213
57. Andrea TA, Kalayeh H (1991) Applications of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors. J Med Chem 34:2824–2836. doi:10.1021/jm00113a022
58. Huang YS, Liu K, Suen CY (1994) A neural network approach for multiclassifier recognition systems. In: Proceedings of the fourth international workshop on frontiers in handwriting recognition, Taiwan, Dec, pp 235–244
59. Fan YN, Xiao X, Min JL, Chou KC (2014) iNR-Drug: predicting the interaction of drugs with nuclear receptors in cellular networking. Int J Mol Sci 15:4915–4937. doi:10.3390/ijms15034915
60. Xu Y, Wen X, Wen LS, Wu LY, Deng NY, Chou KC (2014) iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9:e105018. doi:10.1371/journal.pone.0105018
61. Xiao X, Wu ZC, Chou KC (2014) iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J Theor Biol 284:42–51. doi:10.1016/j.jtbi.2011.06.005
62. Chen L, Zeng WM, Cai YD, Feng KY, Chou KC (2012) Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. PLoS One 7:e35254. doi:10.1371/journal.pone.0035254
63. Chou KC (2013) Some remarks on predicting multi-label attributes in molecular biosystems. Mol Biosyst 9:1092–1100. doi:10.1039/c3mb25555g
64. Zatloukal M, Jorda R, Gucky T, Reznickova E, Voller J, Pospisil T, Malinkova V, Adamcova H, Krystof V, Strnad M (2013) Synthesis and in vitro biological evaluation of 2,6,9-trisubstituted purines targeting multiple cyclin-dependent kinases. Eur J Med Chem 61:61–72. doi:10.1016/j.ejmech.2012.06.036
65. Demange L, Abdellah FN, Lozach O, Ferandin Y, Gresh N, Meijer L, Galons H (2013) Potent inhibitors of CDK5 derived from roscovitine: synthesis, biological evaluation and molecular modelling. Bioorg Med Chem Lett 23:125–131. doi:10.1016/j.bmcl.2012.10.141
66. Malmstrom J, Viklund J, Slivo C, Costa A, Maudet M, Sandelin C, Hiller G, Olsson LL, Aagaard A, Geschwindner S, Xue Y, Vasange M (2012) Synthesis and structure–activity relationship of 4-(1,3-benzothiazol-2-yl)-thiophene-2-sulfonamides as cyclin-dependent kinase 5 (cdk5)/p25 inhibitors. Bioorg Med Chem Lett 22:5919–5923. doi:10.1016/j.bmcl.2012.07.068
67. Demange L, Lozach O, Ferandin Y, Hoang NT, Meijer L, Galons H (2012) Synthesis and evaluation of new potent inhibitors of CK1 and CDK5, two kinases involved in Alzheimer's disease. Med Chem Res 22:3247–3258. doi:10.1007/s00044-012-0334-1
68. Baki A, Bielik A, Molnar L (2007) A high throughput luminescent assay for glycogen synthase kinase-3 inhibitors. Assay Drug Dev Technol 5:75–83. doi:10.1089/adt.2006.029
69. Ma S, Dai Y (2011) Principal component analysis based methods in bioinformatics studies. Brief Bioinform 12:714–722. doi:10.1093/bib/bbq090