Journal of Molecular Graphics and Modelling 28 (2010) 775–787

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

PhDD: A new pharmacophore-based *de novo* design method of drug-like molecules combined with assessment of synthetic accessibility

Qi Huang, Lin-Li Li, Sheng-Yong Yang *

State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, #1 Keyuan Road 4, Chengdu, Sichuan 610041, China

ARTICLE INFO

Article history:
Received 2 August 2009
Received in revised form 25 January 2010
Accepted 7 February 2010
Available online 11 February 2010

Keywords:
Pharmacophore model
De novo design
Assessment of synthetic accessibility
Drug-like molecule

ABSTRACT

This account describes a new pharmacophore-based *de novo* design method of drug-like molecules (PhDD). The method PhDD first generates a set of new molecules that completely conform to the requirements of a given pharmacophore model, followed by a series of assessments to the generated molecules, including assessments of drug-likeness, bioactivity, and synthetic accessibility. PhDD is tested on three typical examples, namely, pharmacophore hypotheses of histone deacetylase (HDAC), cyclin-dependent kinase 2 (CDK2) and HIV-1 integrase (IN) inhibitors. The test results demonstrate that PhDD is able to generate molecules with novel structures but having similar biological functions with existing inhibitors. The validity of PhDD together with its ability of assessing synthetic accessibility makes it a useful tool in rational drug design.

1. Introduction

De novo design methods of drug molecules have increasingly attracted much attention in recent years since they can produce completely novel chemical structures with desired pharmacological properties. So far a considerable number of *de novo* design methods have been developed, including HSITE/2D Skeletons [1], 3D Skeletons [2], LEGEND [3], LUDI [4], NEWLEAD [5], CONCEPTS [6], SPROUT [7], MCSS&HOOK [8], SMoG [9], CONCERTS [10], LEA [11], LigBuilder [12], TOPAS [13], F-DycoBlock [14], ADAPT [15], SYNOPSIS [16], CoG [17], BREED [18], etc. (for a review, see Ref. [19]). These methods have already been used in drug discovery and some of them have shown a good performance. However, all of these methods except NEWLEAD, which will be mentioned later, adopt receptor-based strategy. This strategy is difficult to process the cases in which the receptors or their structures are unknown. On the other hand, very few methods consider the synthetic accessibility of the designed compounds, which has been thought as the most challenging problem in *de novo* molecule design. Other problems which still have no optimal solution in the known methods include (1) how to sample the molecular search space effectively, and (2) how to evaluate the potency of designed molecules [19].

NEWLEAD, proposed by Tschinke and Cohen [5], is the first pharmacophore-based *de novo* design method, which has ever played an important role in promoting the development of *de novo* design methods. It uses as input a set of disconnected molecule fragments that are consistent with a pharmacophore model. Then, it tries to link the disconnected fragments with spacers (such as atoms, chains, or ring moieties). Actually the NEWLEAD can only process the cases that the pharmacophore features are concrete functional groups (not abstract chemical features, like hydrogen bond acceptor, hydrogen bond donor, hydrophobic feature). Additionally, sterically forbidden regions of the receptor binding site are not considered in NEWLEAD.

In order to overcome problems mentioned above, we developed a new pharmacophore-based *de novo* design method of drug-like molecules, called PhDD (a *Ph*armacophore-based *De Novo D*esign *M*ethod). The method PhDD has the following characteristics that are distinct from the commonly used receptor-based *de novo* design methods and NEWLEAD: (1) PhDD works with abstract pharmacophore models. The pharmacophore models might be established by using receptor-based methods, such as LigandScout [20], Pocket [21], or ligand based methods, such as Catalyst (Accelrys Inc., USA), Galahad [22], GASP [23], and DISCO [24]. Further, PhDD also works with pharmacophore models with excluded volumes involved. (2) PhDD incorporates with the assessment of synthetic accessibility of the designed molecules. (3) Fragments as well as linkers which are used to link the different fragments were obtained by splitting drug molecules that are clinically used or in clinical trials. This would help to reduce the molecular search space effectively and make the generated molecules more drug-like. (4) The bioactivity of designed molecules is estimated by using a fit value, which describes how well a ligand is aligned with a pharmacophore model.

The rest of this paper is organized as follows: the second part presents a detailed description of the algorithms used in PhDD. In

* Corresponding author. Tel.: +86 28 85164063; fax: +86 28 85164060.
E-mail address: yangsy@scu.edu.cn (S.-Y. Yang).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.02.002

the third part, PhDD is tested on three typical examples. The fourth part is a short discussion about the performance of PhDD. Conclusions will be offered in the final part.

2. Methodology

2.1. An overview of PhDD

The purpose of PhDD is to generate drug-like molecules which completely conform to the requirements of a given pharmacophore model. An overall flow chart for PhDD is presented in Fig. 1. The working process can be briefly described as follows. Given a pharmacophore hypothesis, in the first step, PhDD chooses proper fragments from different fragment databases that match chemical features of the pharmacophore hypothesis, followed by installing them to the 3D framework of the pharmacophore model. In the second step, PhDD tries to link all the disconnected fragments together by suitable linkers to form a complete molecule. In the third step, PhDD performs a series of assessments to the generated molecules, including assessments of drug-likeness, bioactivity, and synthetic accessibility. Details for all the algorithms used are given as follows.

2.2. Fragment and linker databases

Eight types of fragment databases and one linker database were established in advance. The eight types of fragment databases correspond to eight popular pharmacophore features respectively, including hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), positive ionizable (PI), negative ionizable (NI), ring aromatic (RA), hydrophobic (H), hydrophobic aromatic (HAR), and hydrophobic aliphatic (HAL) features. All the fragments and linkers were obtained by splitting the molecules in the databases of MDDR (MDL Drug Data Report) and CMC (Comprehensive Medicinal Chemistry), which might be helpful to make the created molecules more drug-like. The splitting operation was accomplished by using a module named Generate Fragments in Pipeline Pilot (Accelrys Inc., USA). The pharmacophore feature category for each fragment was identified by Catalyst (Accelrys Inc., USA), and confirmed visually. The molecular weight of each fragment and linker was restricted to be less than 250 and 200 Da, respectively. All the fragments and linkers were minimized by using the CHARMm force field [25] and stored in MOL2 format. The number of fragments is 385, 749, 43, 52, 530, 609, 436, and 173 for HBD, HBA, PI, NI, RA, H, HAR, and HAL, respectively. The linker database contains 1974 fragments. Users are allowed to modify each library to meet their specific purposes, including adding or deleting fragments/linkers from corresponding libraries.

2.3. Installation of fragments in the 3D framework of pharmacophore model

PhDD starts its work from a given pharmacophore model, which is the only input. The default format of pharmacophore model input file used here is the CHM format since it has been widely used and has become a *de facto* standard of pharmacophore model. Firstly, PhDD reads the pharmacophore feature information from the input, such as the number of pharmacophore features, coordinates of centers of pharmacophore features, tolerance of each pharmacophore feature, etc. For each pharmacophore feature, PhDD randomly chooses a fragment from its corresponding fragment database. Subsequently the chosen fragments are properly positioned in the 3D framework defined by the pharmacophore model. The placement of fragment should satisfy the following conditions: (1) the center of fragment should be superposed with that of its corresponding pharmacophore feature. For example, given a fragment 1-methyl-1H-indole and a ring aromatic feature (see Fig. 2a), PhDD places the center of the benzene ring over that of the ring aromatic feature (since 1-methyl-1H-indole has two ring aromatic centers, one benzene ring and one pyrrole ring, the larger one is automatically chosen); (2) if the pharmacophore feature is HBA, HBD or RA, a proper orientational adjustment of the fragment is needed to make it

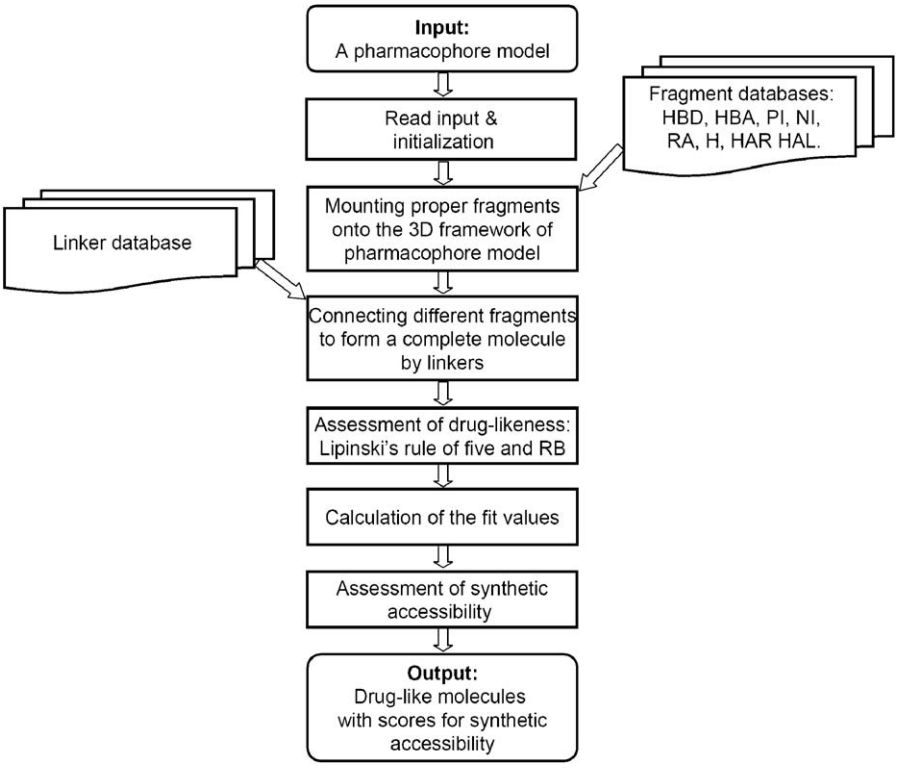

Fig. 1. Overall flow chart of PhDD.

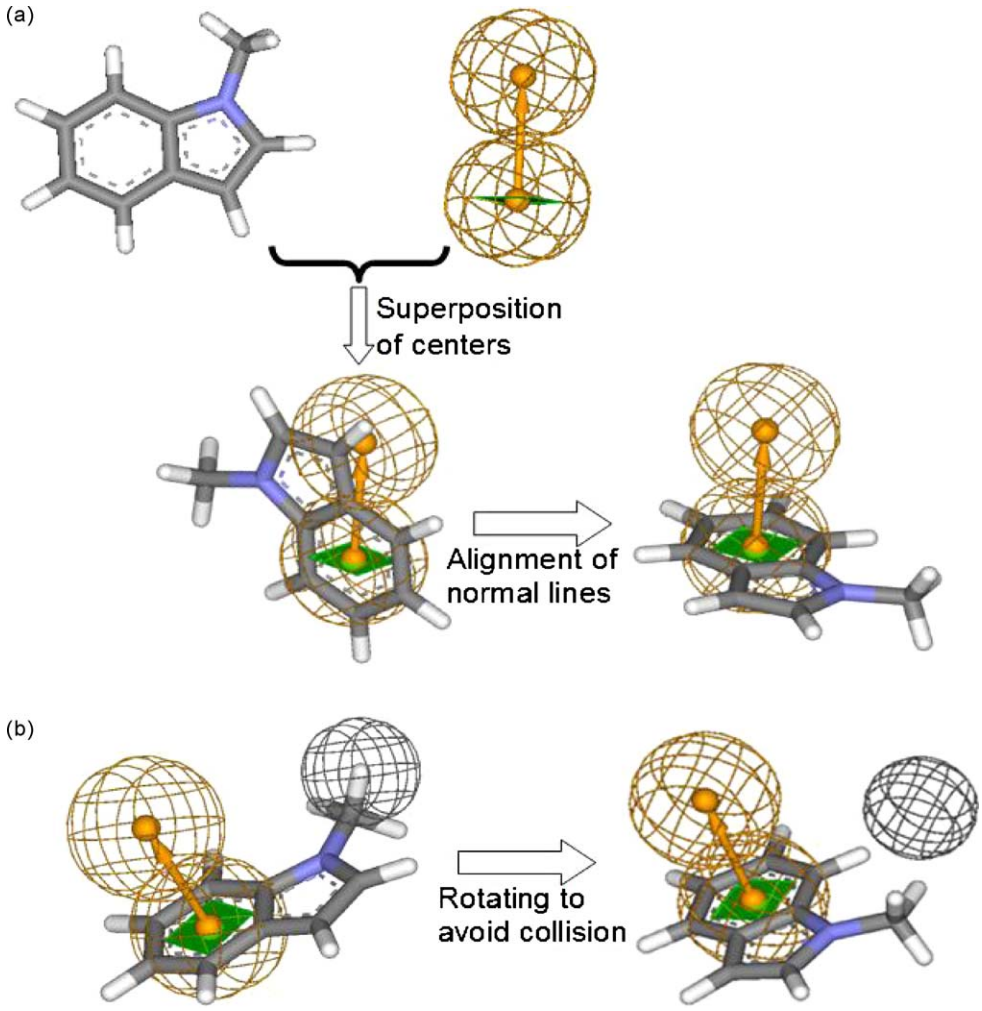

Fig. 2. Schematic illustration of the fragment installation onto the 3D framework of a pharmacophore model. (a) Superposition of centers and alignment of normal lines. (b) Rotation around the normal direction to avoid collisions. Features are color coded with orange, aromatic ring; gray, excluded volume.

to conform to the requirement of the pharmacophore feature. For example, one needs to adjust the orientation of 1-methyl-1H-indole to make the normal direction of the benzene ring consistent with that of the RA pharmacophore feature (see Fig. 2a); (3) a proper rotation might also be needed if a collision happens between fragments or fragments and excluded volumes (see Fig. 2b).

2.4. Connecting fragments by using linkers

Through the preceding step, all the locations where pharmacophore features reside have been occupied by appropriate fragments, which means that a set of disconnected fragments exist in the 3D space now. In this step, PhDD chooses a pair of disconnected fragments arbitrarily and links them together to form a new intermediate fragment. This procedure is repeated until all the fragments are connected. The procedure for attaching a linker between two fragments is shown in Fig. 3, and a detailed description is given as follows.

Supposing that two fragments F1 and F2 are needed to link, PhDD first searches all the heavy atom pairs (**A1**, **B1**) as the linking point pairs; in each pair, **A1** is from F1, and **B1** from F2. The qualification of a heavy atom to be a linking point is that it must bond with at least one hydrogen atom. For each linking point pair, a vector pair is defined, in which the heavy atom and its bonding H-atom correspond to the vector tail and head, respectively (see Fig. 3a). Four parameters are used to describe the vector pair: the distance (d) between the two heavy atoms, the two angles α_1 and α_2, and one dihedral angle ϕ, namely (d, α_1, α_2, ϕ) (see Fig. 3a). Subsequently, the program starts to systematically search for all the matching linkers from the linker database; each of the selected linkers should have a vector pair (d', α_1', α_2', ϕ'), in which $d' = d$, $\alpha_1' = \pi - \alpha_1$, $\alpha_2' = \pi - \alpha_2$, and $\phi = \phi'$. Tolerance values can be defined for these parameters; here the default tolerance value for d is 0.25 Å, for angle is 15° and 20° for dihedral angle. At this moment, all the possible linking point pairs for the two fragments together with their matching linkers are obtained. The next thing is to select a combination of linking point pair and linker. Here the roulette-wheel selection algorithm [26] is used. In the roulette-wheel selection algorithm, the individuals (corresponding to the combinations of linking point pair and linker) are mapped to contiguous segments of a line, such that each individual's segment is equal in size to its fitness. A random number is generated and the individual whose segment spans the random number is selected. In PhDD, the fitness is defined by the following formula:

$$Fit_{roulette\text{-}wheel} = \frac{1}{d \times linker\_weight} \quad (1)$$

where d is the distance component in the vector pair (d, α_1, α_2, ϕ) (its definition see Fig. 3a), *linker_weight* represents the molecular weight of linker. Clearly, the cases in which the linking point pair has a smaller distance d and/or the linker has a smaller molecular

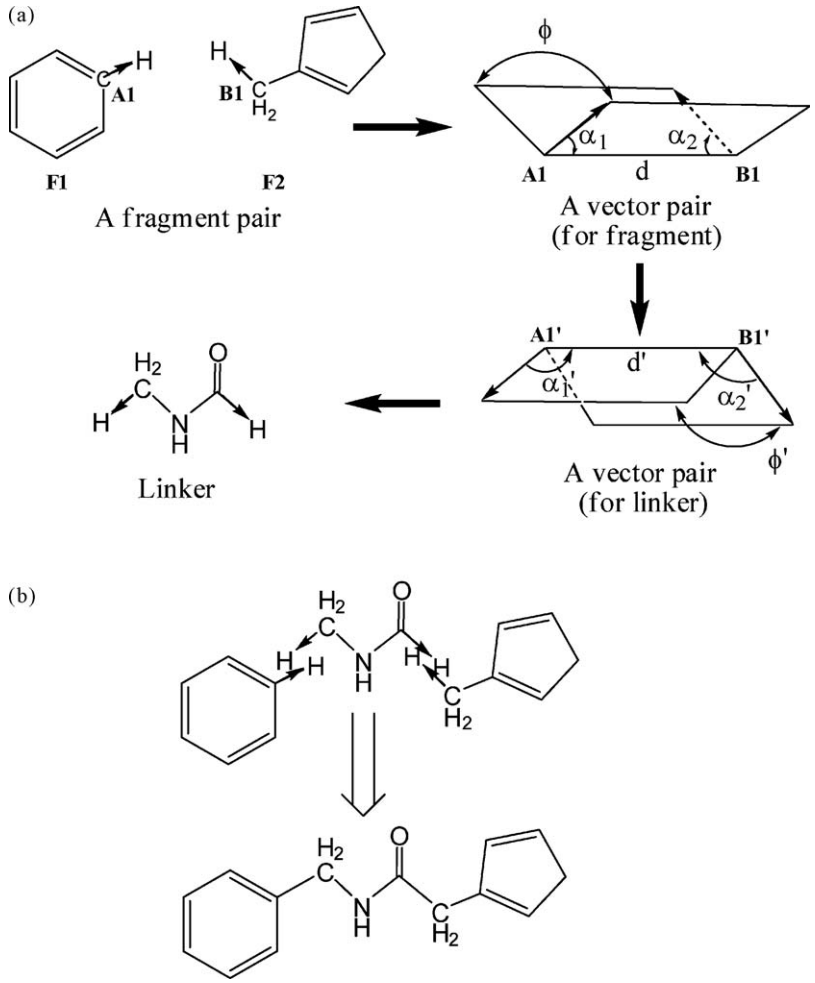

Fig. 3. Schematic illustration of the connection of fragments by using a linker. (a) The process of locating a proper linker from a fragment pair. (b) Attaching a linker between two fragments.

weight have a larger probability to be selected. Finally PhDD directly connects the two heavy atoms by removing the hydrogen atoms (see Fig. 3b).

The above procedures are repeated until only one fragment left, which is the final molecule. Since the molecule just generated is usually heavily distorted, a geometry cleaning operation is then carried out. Here a simplified MMFF94 molecular force field [27] is used for the energy minimization. Molecules with their geometry cleaned will be subjected to the subsequent assessments.

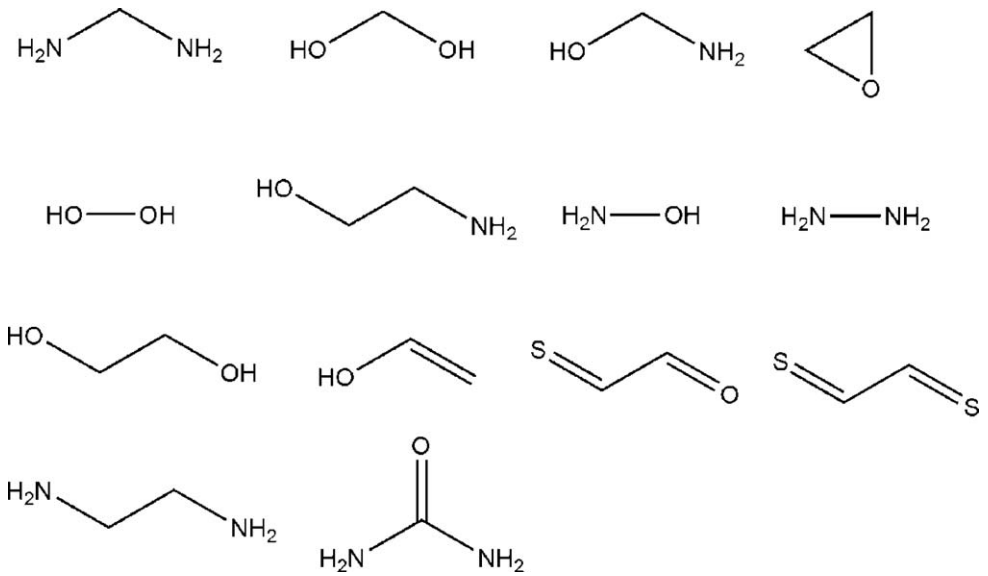

Fig. 4. Not-allowed atomic connections in PhDD.

Here we have to mention that some connections are chemically improper. For examples, the same hetero-atoms are connected, such as, O–O, N–N. The situation that two or more hetero-atoms bond to a same carbon atom, such as N–C–N, O–C–O, is also not allowed. All the not-allowed situations in PhDD are shown in Fig. 4. If a connection results in any one of the not-allowed situations, it will be rejected.

2.5. Assessments to the generated molecules

PhDD performs a series of assessments to the generated molecules, including assessments of drug-likeness, bioactivity, and synthetic accessibility. Drug-unlike molecules will be skipped, in other words, all of the output molecules should satisfy the criteria of drug-likeness. Estimations of bioactivity and synthetic accessibility will help users choose compounds with higher potency and easier synthetic accessibility for further experimental studies.

2.6. The assessment of drug-likeness

The classical Lipinski's rule of five [28] is used to assess the drug-likeness of the generated molecules. The Lipinski's rule requires the following criteria to be satisfied for an active drug molecule: (1) not more than 5 hydrogen bond donors (nitrogen or oxygen atoms with one or more hydrogen atoms); (2) not more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms); (3) a molecular weight under 500 Da; and (4) a partition coefficient log *P* less than 5. In addition, a criterion related to rotatable bonds, namely, the number of rotatable bonds is not more than 10, is also used since this has recently been thought as another important descriptor of drug-likeness [29–31]. Here a rotatable bond is defined as any single non-ring bond, bounded to nonterminal heavy (i.e., non-hydrogen) atom. Amide C–N bonds are not considered because of their high rotational energy barrier [32]. If a molecule violates any of these rules, it will be rejected.

2.7. The estimation of bioactivity

Fit values, which are usually used to define how well molecules are mapped with a given pharmacophore model, are adopted to estimate the biological activity of the generated molecules. A fit value is calculated according to the following formula:

$$Fit = \sum_{all\_features} weight \times \sqrt{1 - \left(\frac{d}{t}\right)^2} \tag{2}$$

where d represents the distance between the center of fragment and that of pharmacophore feature, t is the tolerance of pharmacophore feature, and weight refers to the weight of feature.

Here one may argue that the estimation of bioactivity through the fit value is not necessary since in the beginning the fragments have been exactly placed onto the centers of pharmacophore features. However, in the actual running process, almost all of the fragments have been found to deviate a little from their original positions. This situation is due to that it is difficult to find a vector pair of linker which matches perfectly to that of the fragments (that is the reason why we introduce tolerances as mentioned before).

2.8. The assessment of synthetic accessibility

In general, most of the molecules generated by *de novo* design methods have novel chemical structures, which means that they cannot be purchased from the market directly and have to synthesize. Thus their synthetic accessibility becomes a critical problem. In order to assess the synthetic accessibility of the newly generated molecules, PhDD adopts a modified complexity-based method [33]. Here the complexity-based method other than more complicated methods, such as starting material-based and retro-synthesis-based methods, is used, since the complexity-based method is simple and fast, and in many cases it can give considerably good results [34].

In PhDD, a value of scoring function defined by Eq. (3) is assigned to each molecule.

$$F = F_{ring} + F_{connect} + F_{type} + F_{chirality} \tag{3}$$

where F_{ring} represents the contribution of rings, which is calculated by the following formula:

$$F_{ring} = \sum_{i=1}^{i=nring} cons \times size(i) \tag{4}$$

where *nring* is the number of rings in the molecule, $size(i)$ is the number of atoms in the ring i, *cons* is a constant coefficient (which is set to 6 here).

$F_{connect}$ in Eq. (3) refers to the contribution of inter-atomic connections. The calculation formula is:

$$F_{connect} = \sum_{i=1}^{natom} f(atom_i) \tag{5}$$

$$f(atom_i) = \begin{cases} 24 & if\ \ con\_degree = 4 \\ 12 & if\ \ con\_degree = 3 \\ 6 & if\ \ con\_degree = 2 \\ 3 & if\ \ con\_degree = 1 \end{cases} \tag{6}$$

where *natom* is the number of atoms in the molecule, *con_degree* is the connection degree of atom, i.e. the number of atoms bonded to the specified atom.

F_{type} in Eq. (3) refers to the contribution of atom type. If it is a carbon atom, the value is set to 3. Otherwise, the value is 6.

$$F_{type} = \sum_{i=1}^{natom} f(at_i) \tag{7}$$

$$f(at_i) = \begin{cases} 3 & if\ atomic\_type = C \\ 6 & if\ atomic\_type \neq C \end{cases} \tag{8}$$

The last term in Eq. (3), namely $F_{chirality}$, represents the contribution of chiral centers. This calculation formula is:

$$F_{chirality} = diff \times nchiral \tag{9}$$

where *diff* refers to a "difficulty" coefficient for each chiral center and it is set to 20 here. *nchiral* is the number of chiral centers in the molecule.

To sum up, an empirically derived rule based scoring function is used to estimate the synthetic accessibility of a chemical compound. The scoring function includes the contributions of rings, connections, atomic types and chiral centers. A larger value of scoring function means that the molecule is more difficult to synthesize. On the contrary, a smaller value means that the compound is easier for chemical synthesis. By the way, all the parameters used here were taken from Ref. [33], except *diff* in Eq. (9), which was set based on a statistical analysis to more than 100 chiral drug molecules and their synthetic reactions.

2.9. Implementation of PhDD

The overall flow chart of PhDD is presented in Fig. 1. The source code for the implementation of PhDD was written in C++ programming language (gcc) under the UNIX/LINUX operating system. The PhDD program is available free of charge to not-for-profit institution upon request from the corresponding author.

3. Results

Here three test examples, namely pharmacophore models of Histone deacetylase (HDAC) [35,36] inhibitors, cyclin-dependent kinase 2 (CDK2) [37,38] inhibitors, and HIV-1 integrase (IN) [39,40] inhibitors, were adopted for the evaluation of the performance of PhDD. The three examples were chosen, which are mainly due to the following reasons: (1) they are all targets of pharmaceutical interest and belong to different kinds of targets; (2) they have been extensively studied and a considerable number of small molecular inhibitors have been publicly reported; (3) pharmacophore hypotheses in these examples cover two different pharmacophore model types, namely pharmacophore models with and without excluded volumes.

3.1. HDAC inhibitors

Histone deacetylases are a class of enzymes that catalyze the removal of acetyl groups from a ε-N-acetyl lysine amino acid on a histone [35,36]. They play critical roles in epigenetic regulation of gene expression. HDAC inhibition causes hyperacetylation of histones, which leads to a more open chromatin structure, thereby facilitating gene transcription and ultimately leading to cell-growth arrest, differentiation, and apoptosis. Thus HDAC inhibitors have been seen as potential anti-cancer drugs and increasingly attracted much attention in recent years.

Here a pharmacophore model of HDAC inhibitors developed by Vadivelan et al. [41] was chosen, which contains one hydrogen bond acceptor, one hydrophobic aliphatic, and two ring aromatic properties (see Fig. 5a). This model was used as input of PhDD. The maximum number of output molecules was set to 100. This job was run on a desktop PC (3.20 GHz Intel Pentium CPU, Red Hat Linux 9) and 100 molecules were generated in about 20 min.

A structural analysis of the whole generated molecular ensemble shows that the generated molecules are totally different in chemical structures. All of them satisfy the criteria of drug-likeness, including Lipinski's rule of five and the restriction of the number of rotatable bonds. Each of them has been assigned a fit value, which reflects how well the molecule is mapped with the given pharmacophore model. Fig. 5b presents four of the generated molecules (**A1**, **A2**, **A3**, and **A4**). Their alignments with the pharmacophore model of HDAC inhibitors are also shown in Fig. 5b. The calculated fit values are 8.325, 5.652, 9.027, and 8.762 for **A1**, **A2**, **A3**, and **A4**, respectively. A larger fit value indicates better match between the molecule and pharmacophore model, hence presumably having a higher bioactivity.

Further, we examined the similarity of the generated molecules with known HDAC inhibitors in chemical structure and biological function. In order to do this, a set of 40 known HDAC inhibitors with higher potency were first collected from literature [35,36,41]. A comparison of chemical structures between the generated molecules and the known HDAC inhibitors indicates that the generated molecules have novel chemical structures. Here one may wonder whether the generated molecules with novel chemical structures have some similarity in biological function with known HDAC inhibitors. For addressing this question, we calculated Tanimoto similarity coefficients [42–44] of the generated molecules with the known HDAC inhibitors. The Tanimoto similarity coefficient has been thought as a good indicator of similarity in biological function for molecules with different structures, which is defined as follows:

$$sim(i,j) = \frac{\sum_{d=1}^{p} x_{di}x_{dj}}{\sum_{d=1}^{p}(x_{di})^2 + \sum_{d=1}^{p}(x_{dj})^2 - \sum_{d=1}^{p} x_{di}x_{dj}} \quad (11)$$

where $sim(i, j)$ is the Tanimoto similarity coefficient of molecule i and j. p is the number of molecular descriptors used. x_{di} is the value of descriptor d of molecule i. Five descriptors were used here, including AlogP (a parameter of lipophilicity), numbers of hydrogen bond donors and acceptors, and number of rotatable bonds. These descriptors were chosen since their corresponding properties have been shown to be important for an enzyme inhibitor [45]. The calculated Tanimoto similarity coefficients of the generated molecules with the known inhibitors span a range of values between 0.738 and 0.969, implying that the generated molecules should have considerably good similarity in biological function with the known HDAC inhibitors. Fig. 6 shows superposition maps of **A1**–**A4** with one HDAC inhibitor that has the highest values of Tanimoto coefficient; here the superposition of molecular structures was carried out by Discovery Studio (Accelrys Inc., USA), in which molecular field based strategy was used. Obviously, the generated molecules **A1**–**A4** are also very similar in 3D molecular shape with the known HDAC inhibitors, which are usually necessary for a potential ligand due to the spatial constraint of the active site of receptor.

Finally, synthetic accessibility of the generated molecules was assessed with the use of a scoring function defined by Eq. (3). Table 1 presents calculated scores for **A1**, **A2**, **A3**, and **A4**. For comparison, scores calculated by CAESA [46] and SYLVIA [47], which are two widely used commercial programs for the assessment of synthetic accessibility, are also given in Table 1. CAESA, developed by Myatt [46], can automatically rank sets of molecules according to their ease of synthesis. A larger score means that the molecule is easier to synthesize. SYLVIA, programmed by Gasteiger's group [47] is another program that can estimate the synthetic accessibility of given compounds. Contrary to CAESA, a smaller score in SYLVIA indicates that the molecule is easier to synthesize. From Table 1, we can see that the ranking order of **A1**, **A2**, **A3**, and **A4** in terms of the difficulty of synthetic accessibility calculated by PhDD is consistent with that calculated by SYLVIA. But it does not completely match that calculated by CAESA. In the overall trend, however, they are still coincident: the synthesis of compound **A1** and **A2** is predicted by CAESA to be easier than that of **A3** and **A4**, coincident with the result predicted by PhDD. Although the ranking order of **A1** and **A2** and that of **A3** and **A4** are not consistent, the scores given by CAESA are very close. All of these demonstrate that PhDD has a comparable performance in estimation of synthetic accessibility with currently used commercial programs.

3.2. CDK2 inhibitors

Cyclin-dependent kinase 2 is a member of the cyclin-dependent kinase family of Ser/Thr protein kinases [37,38]. Its activity is restricted to the G1–S phase of the cell cycle, and is essential for the G1/S transition. The effects of CDK2 inhibitors on the cell cycle and their potential value for the treatment of cancer have been widely studied. The main reasons why CDK2 inhibitors are attractive potential antitumor agents could be (1) they are potent anti-proliferative agents, arresting cells in G1/S phase, (2) they trigger apoptosis, alone or in combination with other treatments, and (3) in some instances, inhibitors of CDK2 contribute to cell differentiation [37,38].

A pharmacophore model of CDK2 inhibitors, which was taken from Ref. [48], was used as another test example. This model, shown in Fig. 7a, contains two HBA, one HBD, and one hydrophobic feature, as well as one excluded volume. Again the model was used as input of PhDD. And the maximum number of molecules was set to 100. PhDD ran about 30 min and produced 100 molecules.

A structural analysis shows that the generated molecules are structurally different. They all satisfy the Lipinski's rule of five and the criterion of the number of rotatable bonds (not more than 10). As examples, four of the generated molecules (**B1**, **B2**, **B3**, **and B4**)

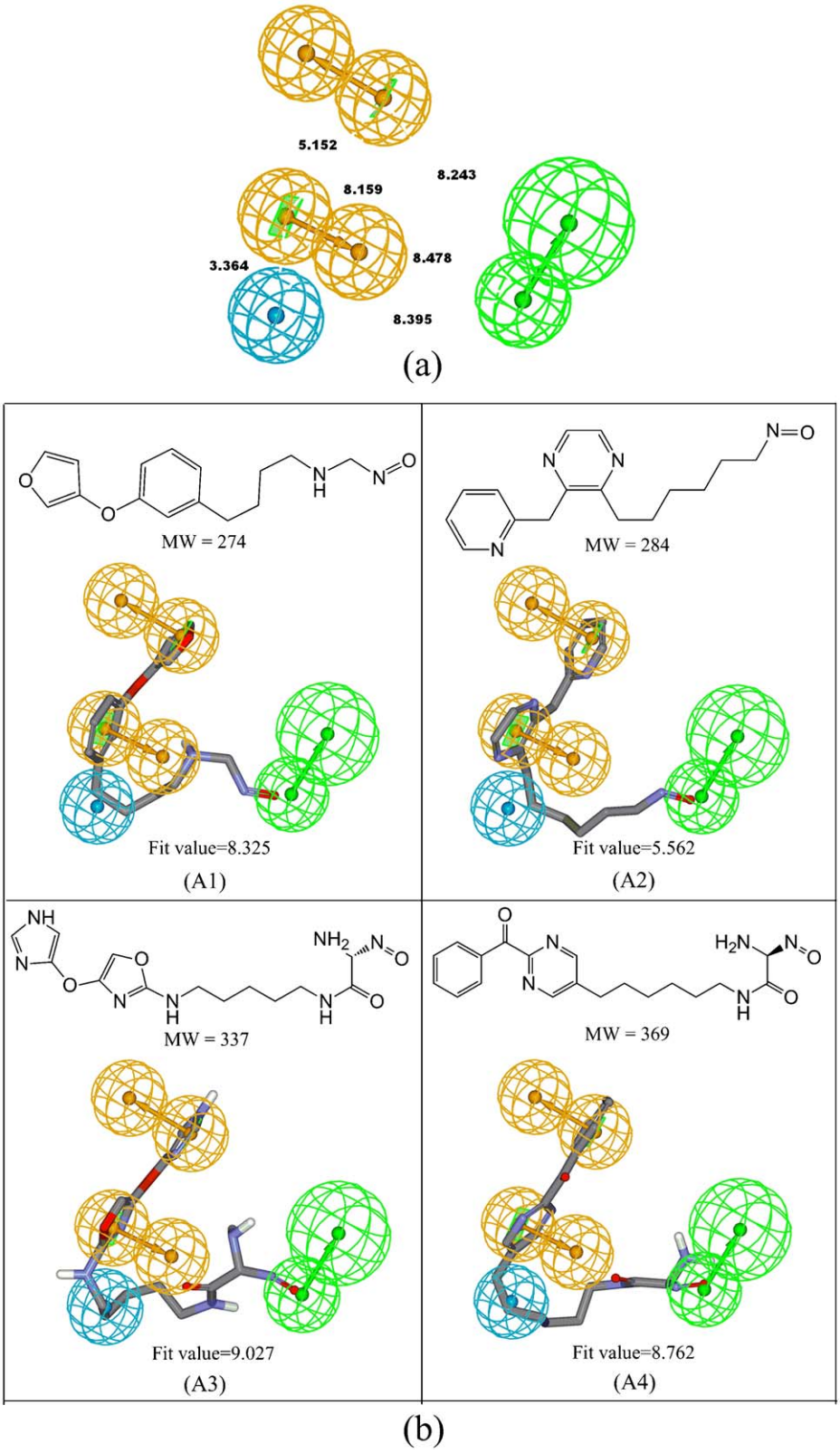

Fig. 5. (a) A pharmacophore model of HDAC inhibitors. (b) Four molecules (**A1**, **A2**, **A3**, and **A4**) generated by PhDD together with their alignments with the pharmacophore model of HDAC inhibitors. The features are color coded with green, hydrogen bond acceptor; light blue, hydrophobic aliphatic feature; orange, ring aromatic property.

together with their alignments with the pharmacophore model of CDK2 inhibitors are shown in Fig. 7b. The calculated fit values are 8.327, 8.193, 8.563, and 8.797 for **B1**, **B2**, **B3**, and **B4**, respectively. Obviously, these molecules are mapped very well with the pharmacophore model (see Fig. 7b).

In order to examine the similarity of the generated molecules with known CDK2 inhibitors, we collected 104 inhibitors of CDK2, which were publicly reported and have higher potency [37,38,48]. Again, we calculated Tanimoto similarity coefficients of the four molecules (**B1**–**B4**) with the known inhibitors. The results show

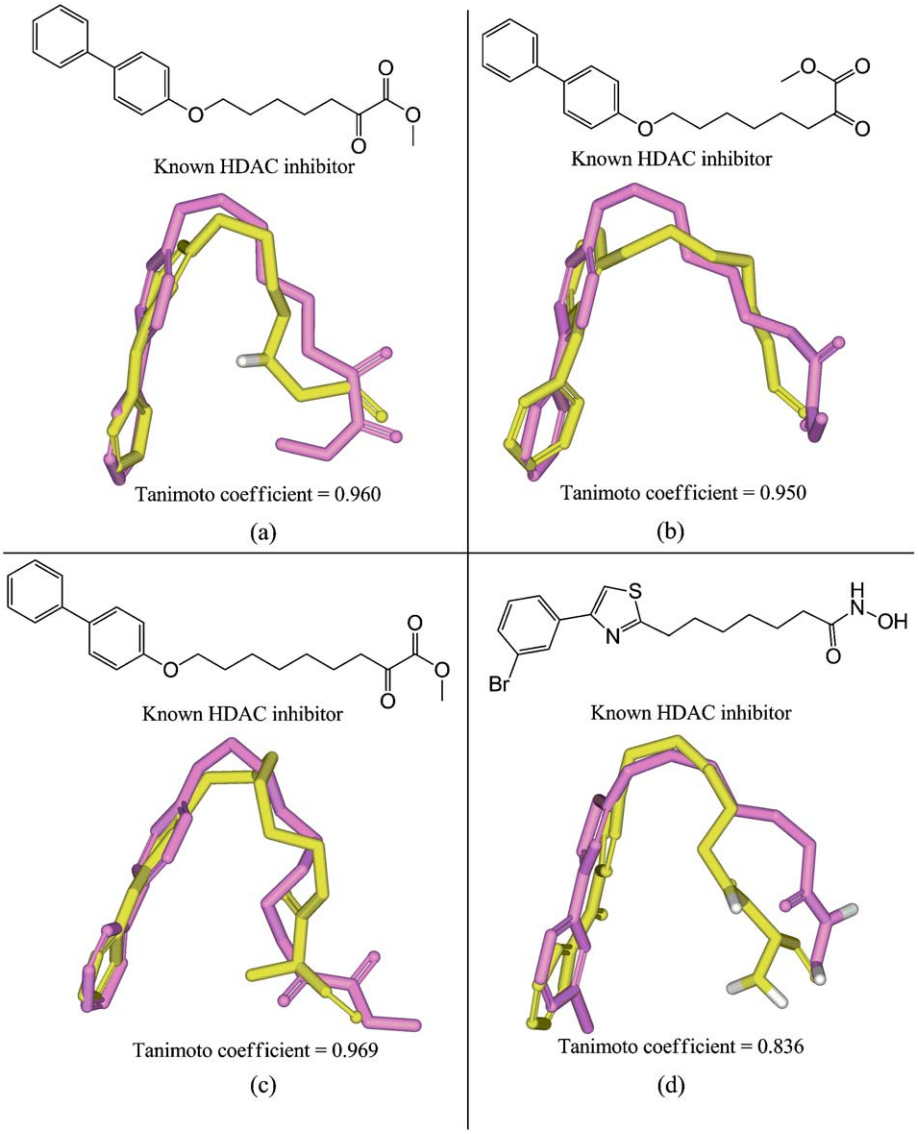

Fig. 6. Superposition maps of **A1**–**A4** (in yellow) with known HDAC inhibitors (violet) that have the highest value of Tanimoto coefficient. (a) is for **A1**, (b) for **A2**, (c) for **A3**, and (d) for **A4**.

that the Tanimoto similarity coefficients are between 0.691 and 0.785, indicating that these molecules should have a good similarity in biological function with the known CDK2 inhibitors. Superpositions of **B1**–**B4** with known CDK2 inhibitors that have the highest values of Tanimoto coefficient are shown in Fig. 8. Clearly, the generated molecules have very similar shapes with these known CDK2 inhibitors.

Again, the assessment of synthetic accessibility of the generated molecules was carried out with the use of a scoring function defined by Eq. (3). Table 2 shows the calculated scores for **B1**, **B2**, **B3**, and **B4**. For comparison, the corresponding scores calculated by CAESA and SYLVIA are also given in Table 2. From Table 2, we can see that the ranking order of the scoring function values calculated by PhDD is very similar with that calculated by SYLVIA. However the ranking order does not match that obtained by CAESA, which might be due to the different algorithms used in PhDD and CAESA.

Table 1
Calculated scoring function values for synthetic accessibility of the designed molecules **A1**, **A2**, **A3**, and **A4** by PhDD together with those by SYLVIA and CAESA.

Molecules	Calculated scores for synthetic accessibility		
	PhDD	SYLVIA	CAESA
A1	363	3.51	63.2
A2	390	3.70	63.4
A3	440	4.70	53.8
A4	512	4.90	59.8

3.3. IN inhibitors

HIV-1 integrase is a DNA nucleotidyltransferase encoded by the pol gene. It is an enzyme of HIV that is required to integrate viral DNA into cellular DNA in the nucleus of a host cell [39,40]. It represents an attractive and validated target for the development of antiretroviral agents. Drugs that selectively inhibit this enzyme, when used alone and in combination regimens, have shown potent anti-HIV activity and a good safety profile in clinical trials.

Fig. 9a shows a chosen pharmacophore model of HIV-1 integrase inhibitors, which was taken from Ref. [49]. This model involves two HBD, one HBA, and one hydrophobic aromatic. Again this model was used as input of PhDD. The maximum number of molecules was set to 100. PhDD ran about 20 min and output 100 drug-like molecules. Fig. 9b depicts four of them (**C1**, **C2**, **C3**, and **C4**). Alignments of the four molecules with the pharmacophore

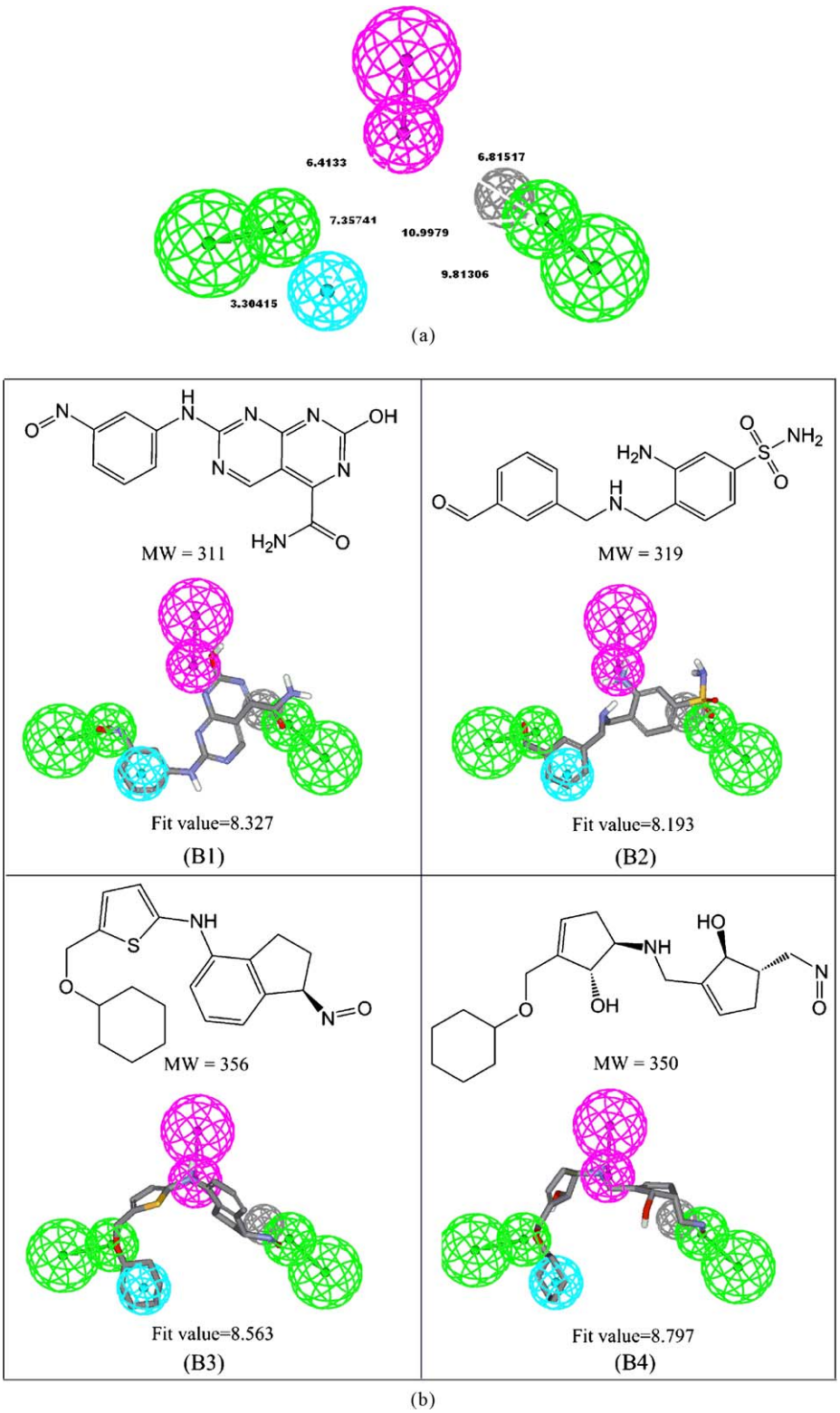

Fig. 7. (a) A pharmacophore model of CDK2 inhibitors. (b) Four molecules (**B1**, **B2**, **B3**, and **B4**) generated by PhDD together with their alignments with the pharmacophore model of CDK2 inhibitors. The features are color coded with green, hydrogen bond acceptor; blue, hydrophobic feature; magenta, hydrogen bond donor.

model are also shown in Fig. 9b. The calculated fit values are 3.73, 3.752, 3.388, and 3.341 for **C1**, **C2**, **C3**, and **C4**, respectively. Obviously, these molecules are mapped very well with the pharmacophore model (see Fig. 9b).

The similarity of the generated molecules with known IN inhibitors was analyzed by calculating the Tanimoto similarity coefficients. A set of 66 known IN inhibitors was collected from different literature resources [39,40,49]. The calculated Tanimoto similarity coefficients of the four molecules with the known inhibitors span a range of values from 0.642 to 0.785, indicating that the four molecules should have similar biological activity with the known IN inhibitors. Alignments of the four molecules with IN

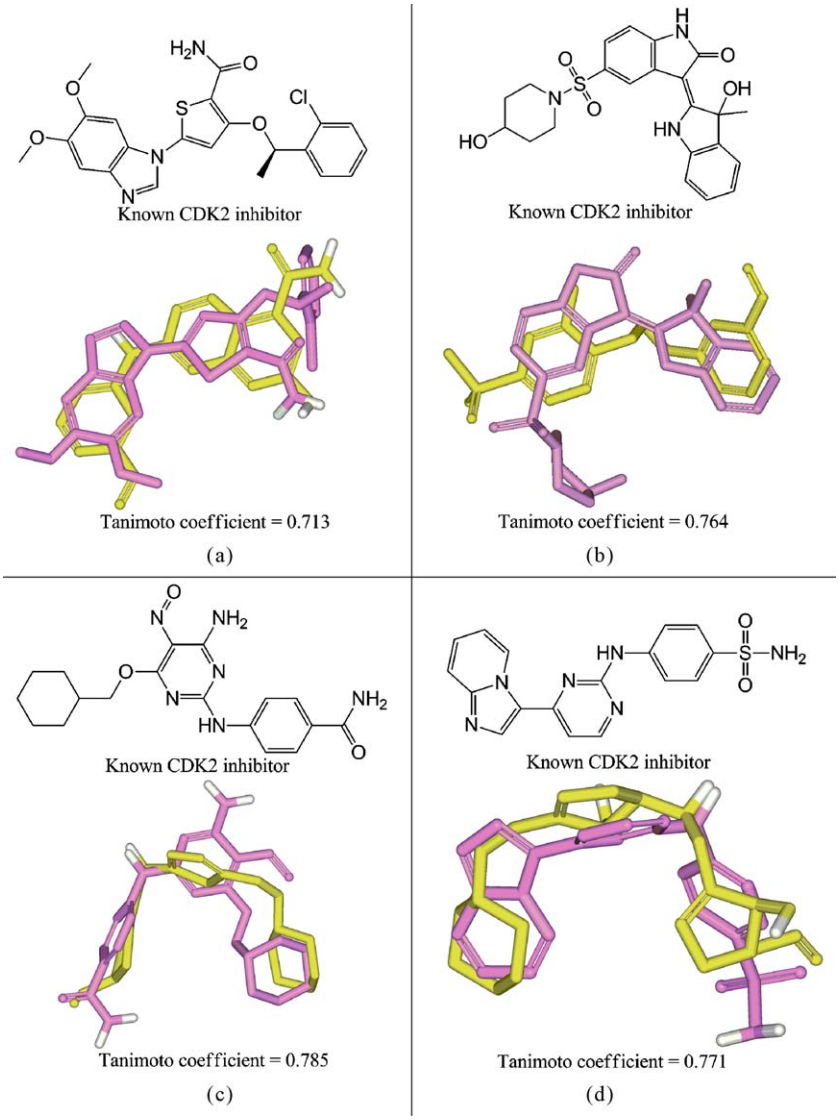

Fig. 8. Superposition maps of **B1**–**B4** (in yellow) with known CDK2 inhibitors (violet) that have the highest value of Tanimoto coefficient. (a) is for **B1**, (b) for **B2**, (c) for **B3**, and (d) for **B4**.

inhibitors that have the highest values of Tanimoto coefficient are shown in Fig. 10. Clearly, the four molecules are very similar in 3D structure with these IN inhibitors.

Finally, scores of synthetic accessibility of the generated molecules were calculated again according to Eq. (3). Table 3 shows the calculated scores for **C1**, **C2**, **C3**, and **C4**. For comparison, the corresponding scores calculated by CAESA and SYLVIA are also given in Table 3. From Table 3, we can see that the ranking order of the scores calculated by PhDD is very similar with that calculated by SYLVIA. However the ranking order does not match that obtained by CAESA. Again this should be ascribed to the different algorithms used in PhDD and CAESA.

Table 2
Calculated scoring function values for synthetic accessibility of the designed molecules **B1**, **B2**, **B3**, and **B4** by PhDD together with those by SYLVIA and CAESA.

Molecules	Calculated scores for synthetic accessibility		
	PhDD	SYLVIA	CAESA
B1	603	4.84	49.8
B2	738	5.46	37.7
B3	778	4.97	45.3
B4	781	6.05	53.4

4. Discussion

The three test examples clearly demonstrate that PhDD is able to generate molecules with completely novel structure. Higher values of Tanimoto similarity coefficients of the generated molecules with known inhibitors indicate that the newly generated molecules should have similar biological properties with the existing inhibitors. Moreover, all the fragments and linkers were obtained by splitting molecules in the databases of CMC and MDDR, which are composed of drug or drug-like compounds. This together with the application of Lipinski's rule of five and the restriction of the number of rotatable bonds ensures the drug-likeness of the generated molecules. On the other hand, new fragments and linkers can be easily added into the corresponding databases. The use of large databases of fragments and linkers is helpful for increasing the diversity of newly generated molecules. The application of roulette-wheel algorithm for the selection of linking point pairs and linkers further benefits the topological diversity of the generated molecules.

Since molecules generated by *de novo* design program usually do not exist, a pre-evaluation of synthetic accessibility to the designed molecules appears extremely necessary. And several

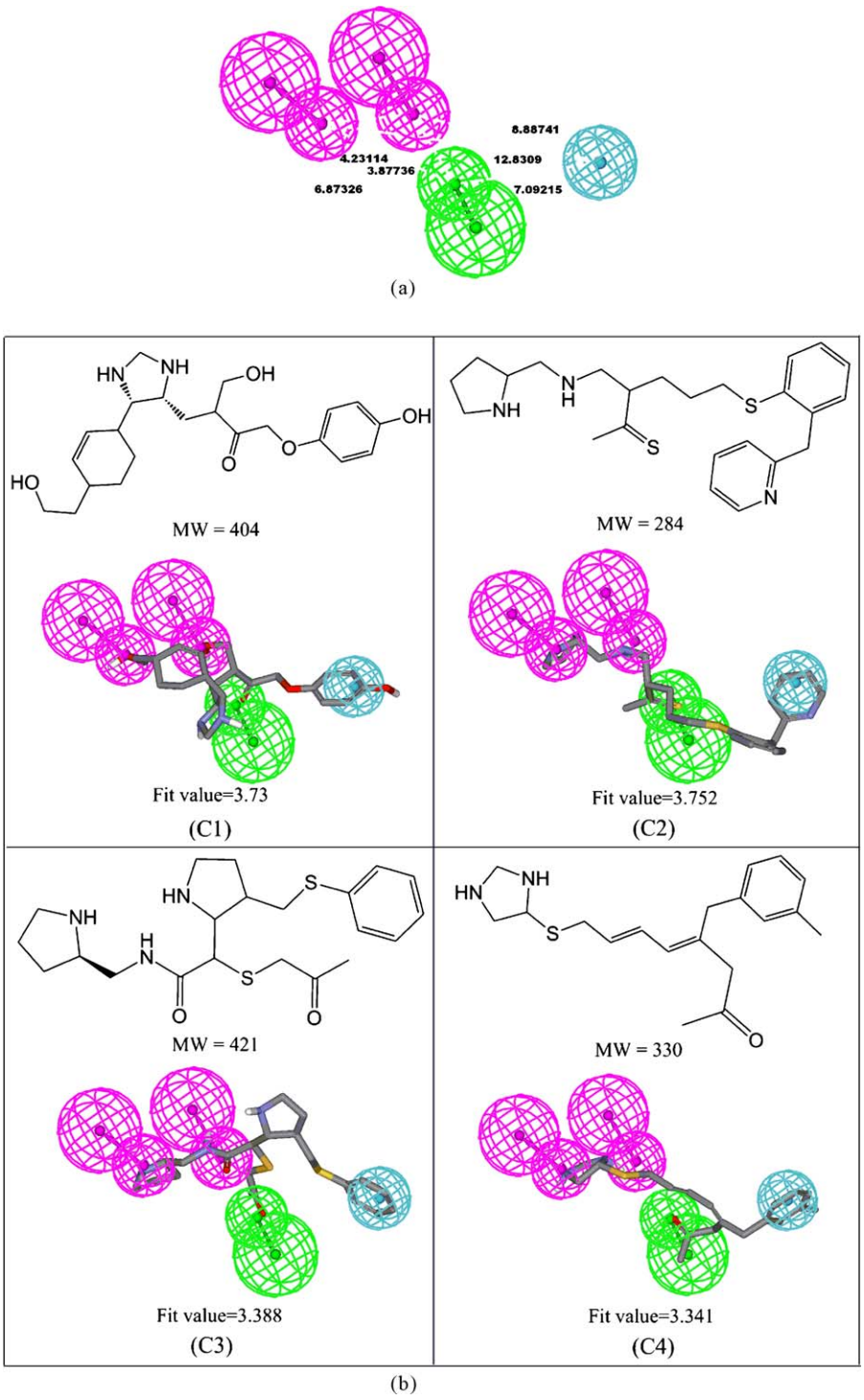

Fig. 9. (a) A pharmacophore model of IN inhibitors. (b) Four molecules (**C1**, **C2**, **C3**, and **C4**) generated by PhDD together with their alignments with the pharmacophore model of IN inhibitors. The features are color coded with green, hydrogen bond acceptor; blue, hydrophobic feature; magenta, hydrogen bond donor.

methods are already available in literature for predicting synthetic accessibility [34,46,47]. These can be classified into three categories: complexity-based, starting material-based or retro-synthetic-based method. All of these methods have their own strong points and weak points. In PhDD, a modified complexity-based method was incorporated, in which the contribution of chirality to the synthetic accessibility was also involved in addition to the contributions of rings, inter-atomic connections and atom types. Here we adopted the complexity-based method other than other methods, since this method is simple and fast, and in many cases can give reasonable results. Our test examples also demonstrate that the modified complexity-based method has a comparable performance with currently used commercial programs, such as CAESA and SYLVIA.

In summary, the algorithms of PhDD ensure that the generated molecules completely conform to the requirements of a given pharmacophore hypothesis. The applications of Lipinski's rule of five and the restriction of the number of rotatable bonds lead to the output molecules more drug-like. The use of roulette-wheel algorithm as well as large databases of fragments and linkers

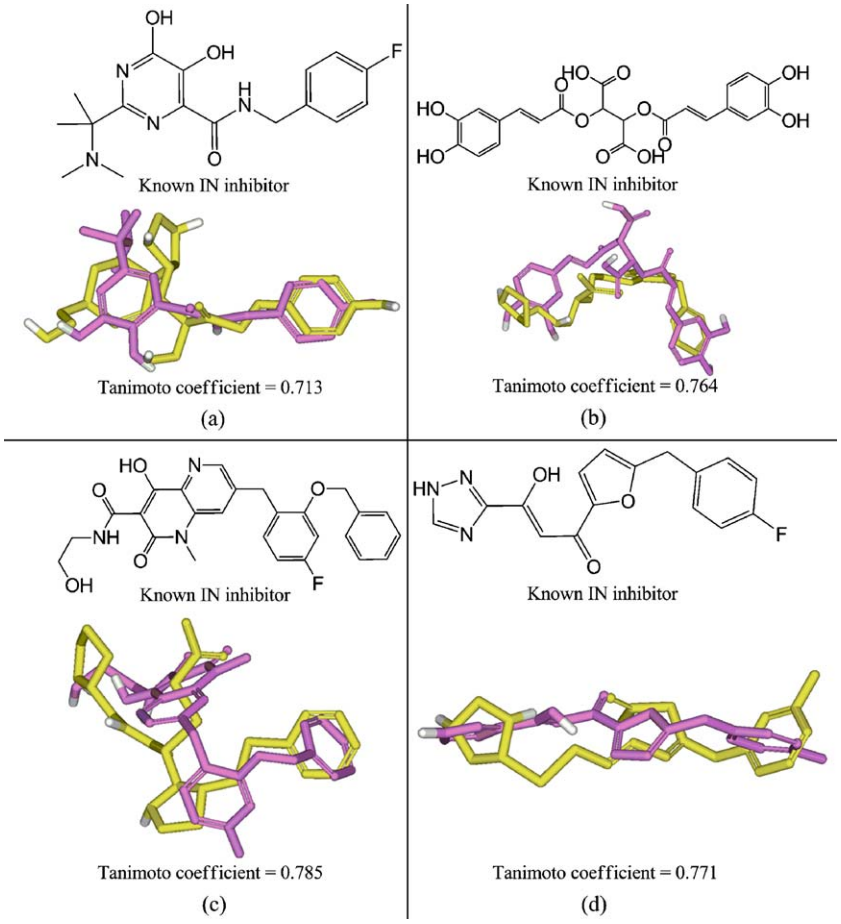

Fig. 10. Superposition maps of **C1**–**C4** (in yellow) with known IN inhibitors (violet) that have the highest value of Tanimoto coefficient. (a) is for **C1**, (b) for **C2**, (c) for **C3**, and (d) for **C4**.

Table 3
Calculated scoring function values for synthetic accessibility of the designed molecules **C1**, **C2**, **C3**, and **C4** by PhDD together with those by SYLVIA and CAESA.

Molecules	Calculated scores for synthetic accessibility		
	PhDD	SYLVIA	CAESA
C1	595	5.43	33
C2	553	5.15	50
C3	604	5.74	51
C4	419	4.48	55

guarantees the structural diversity of generated molecules. Incorporation of assessments of bioactivity and synthetic accessibility can further facilitate the researchers to select such compounds with higher potency and easier synthetic accessibility for subsequent experimental studies.

5. Conclusions

A new pharmacophore-based *de novo* design method of drug-like molecules, called PhDD, is described here. PhDD not only can generate new molecules which conform to the requirements of the given pharmacophore model, but also performs comprehensive assessments to the generated molecules, including assessments of drug-likeness, bioactivity, and synthetic accessibility. Three test cases, namely, pharmacophore models of HDAC inhibitors, CDK2 inhibitors, IN inhibitor, were used to evaluate the method PhDD. The results show that PhDD is able to generate molecules with completely novel structures. Further Tanimoto similarity coefficients of the generated molecules with the existing inhibitors were calculated, which gave higher Tanimoto coefficients, indicating that the generated molecules should have similar biological functions with the existing inhibitors. The validity of PhDD as well as its ability of assessing synthetic accessibility of generated molecules makes it a useful tool in rational drug design.

Acknowledgements

This work was supported by the 863 Hi-Tech Program (2006AA020400) and the National Natural Science Foundation of China (30772651), and partly by the NCET program of the ministry of education of China.

References

[1] D. Danziger, P. Dean, Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces, Proceedings of the Royal Society of London. Series B, Biological Sciences (1934–1990) 236 (1989) 101–113.
[2] V. Gillet, P. Johnson, S. Sike, Automated structure design in 3D, Tetrahedron 3 (1990) 681–696.
[3] Y. Nishibata, A. Itai, Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation, Tetrahedron 47 (1991) 8985–8990.
[4] H. Bohm, The computer program LUDI: a new method for the de novo design of enzyme inhibitors, Journal of Computer-Aided Molecular Design 6 (1992) 61–78.
[5] V. Tschinke, N. Cohen, The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses, Journal of Medicinal Chemistry 36 (1993) 3863–3870.
[6] D. Pearlman, M. Murcko, CONCEPTS: new dynamic algorithm for de novo drug suggestion, Journal of Computational Chemistry 14 (1993) 1184–1193.

[7] V. Gillet, A. Johnson, P. Mata, S. Sike, P. Williams, SPROUT: a program for structure generation, Journal of Computer-Aided Molecular Design 7 (1993) 127–153.
[8] M. Eisen, D. Wiley, M. Karplus, R. Hubbard, HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site, Proteins 19 (1994) 199–221.
[9] R. DeWitte, E. Shakhnovich, SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence, Journal of American Chemical Society 118 (1996) 11733–11744.
[10] D. Pearlman, M. Murcko, CONCERTS: dynamic connection of fragments as an approach to de novo ligand design, Journal of Medicinal Chemistry 39 (1996) 1651–1663.
[11] D. Douguet, E. Thoreau, G. Grassy, A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm, Journal of Computer-Aided Molecular Design 14 (2000) 449–466.
[12] R.-X. Wang, Y. Gao, L.-L. Lai, LigBuilder: a multi-purpose program for structure-based drug design, Journal of Molecular Modeling 6 (2000) 498–516.
[13] G. Schneider, M. Lee, M. Stahl, P. Schneider, De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks, Journal of Computer-Aided Molecular Design 14 (2000) 487–494.
[14] J. Zhu, H. Fan, H. Liu, Y. Shi, Structure-based ligand design for flexible proteins: application of new F-DycoBlock, Journal of Computer-Aided Molecular Design 15 (2001) 979–996.
[15] S. Pegg, J. Haresco, I. Kuntz, A genetic algorithm for structure-based de novo design, Journal of Computer-Aided Molecular Design 15 (2001) 911–933.
[16] H. Vinkers, M. de Jonge, F. Daeyaert, J. Heeres, L. Koymans, J. van Lenthe, P. Lewi, H. Timmerman, K. Van Aken, P. Janssen, SYNOPSIS: synthesize and optimize system in silico, Journal of Medicinal Chemistry 46 (2003) 2765–2773.
[17] N. Brown, B. McKay, F. Gilardoni, J. Gasteiger, A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules, Journal of Chemical Information and Computer Sciences 44 (2004) 1079–1087.
[18] A. Pierce, G. Rao, G. Bemis, BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease, Journal of Medicinal Chemistry 47 (2004) 2768–2775.
[19] G Schneider, U. Fechner, Computer-based *de novo* design of drug-like molecules, Nature Reviews Drug Discovery 4 (2005) 649–663.
[20] G. Wolber, T. Langer, LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters, Journal of Chemical Information and Modeling 45 (2005) 160–169.
[21] J. Chen, L.-L. Lai, Pocket v. 2: further developments on receptor-based pharmacophore modeling, Journal of Chemical Information and Modeling 46 (2006) 2684–2691.
[22] N. Richmond, C. Abrams, P. Wolohan, E. Abrahamian, P. Willett, R. Clark, GALAHAD. 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D, Journal of Computer-Aided Molecular Design 20 (2006) 567–587.
[23] G. Jones, P. Willett, R. Glen, GASP: genetic algorithm superposition program, in: Pharmacophore Perception, Development and Use in Drug Design, International University Line, CA, 2000 , pp. 85–106.
[24] Y. Martin, M. Bures, E. Danaher, J. DeLazzer, I. Lico, P. Pavlik, A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists, Journal of Computer-Aided Molecular Design 7 (1993) 83–102.
[25] B. Brooks, R. Bruccceri, B. Olafson, D. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, Journal of Computational Chemistry 4 (1983) 187–217.
[26] D Fogel, An introduction to simulated evolutionary optimization, IEEE Transactions on Neural Networks 5 (1994) 3–14.
[27] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, Journal of Computational Chemistry 17 (1996) 490–519.
[28] C. Lipinski, F. Lombardo, B. Dominy, P. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery, Advanced Drug Delivery Reviews 23 (1997) 3.
[29] C. Lipinski, Lead- and drug-like compounds: the rule-of-five revolution, Drug Discovery Today: Technologies 1 (2004) 337–341.
[30] D. Veber, S. Johnson, H.-Y. Cheng, B. Smith, K. Ward, K. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, Journal of Medicinal Chemistry 45 (2002) 2615–2623.
[31] H. Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise? Nature Reviews Drug Discovery 2 (2003) 192–204.
[32] R. Bruce Martin, Free energies and equilibria of peptide bond hydrolysis and formation, Biopolymers 45 (1998) 351–353.
[33] R. Barone, M. Chanon, A new and simple approach to chemical complexity. Application to the synthesis of natural products, Journal of Chemical Information and Computer Sciences 41 (2001) 269–272.
[34] J.C. Baber, M. Feher, Predicting synthetic accessibility: application in drug discovery and development, Mini-Reviews in Medicinal Chemistry 4 (2004) 681–692.
[35] W. Douglas Cress, E. Seto, Histone deacetylases, transcriptional control, and cancer, Journal of Cellular Physiology 184 (2000) 1–16.
[36] P. Marks, R. Rifkind, V. Richon, R. Breslow, T. Miller, W. Kelly, Histone deacetylases and cancer: causes and therapies, Nature Reviews Cancer 1 (2001) 194–202.
[37] D. Morgan, Cyclin-dependent kinases: engines, clocks, and microprocessors, Annual Review of Cell and Developmental Biology 13 (1997) 261–291.
[38] X. Graña, E. Reddy, Cell cycle control in mammalian cells: role of cyclins, cyclin dependent kinases (CDKs), growth suppressor genes and cyclin-dependent kinase inhibitors (CKIs), Oncogene 11 (1995) 211–219.
[39] O. Jegede, J. Babu, R. Santo, D. McColl, J. Weber, M. Quiñones-Mateu, HIV Type 1 integrase inhibitors: from basic research to clinical implications, AIDS Reviews 10 (2008) 172–189.
[40] F. Dyda, A. Hickman, T. Jenkins, A. Engelman, R. Craigie, D. Davies, Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases, Science 266 (1994) 1981–1986.
[41] S. Vadivelan, B. Sinha, G. Rambabu, K. Boppana, S. Jagarlapudi, Pharmacophore modeling and virtual screening studies to design some potential histone deacetylase inhibitors as new leads, Journal of Molecular Graphics and Modelling 26 (2008) 935–946.
[42] P. Willett, J. Barnard, G. Downs, Chemical similarity searching, Journal of Chemical Information and Computer Sciences 38 (1998) 983–996.
[43] L. Molnár, G. Keseru, A neural network based virtual screening of cytochrome P450 3A4 inhibitors, Bioorganic & Medicinal Chemistry Letters 12 (2002) 419–421.
[44] T. Poetter, H. Matter, Random or rational design? Evaluation of diverse compound subsets from chemical structure databases, Journal of Medicinal Chemistry 41 (1998) 478–488.
[45] X. Xia, E. Maliski, P. Gallant, D. Rogers, Classification of kinase inhibitors using a Bayesian model, Journal of Medicinal Chemistry 47 (2004) 4463–4470.
[46] V.J. Gillet, G. Myatt, Z. Zsoldos, A.P. Johnson, SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility, Perspectives in Drug Discovery and Design 3 (1995) 34–50.
[47] K. Boda, T. Seidel, J. Gasteiger, Structure and reaction based evaluation of synthetic accessibility, Journal of Computer-Aided Molecular Design 21 (2007) 311–325.
[48] S. Vadivelan, B. Sinha, S. Irudayam, S. Jagarlapudi, Virtual screening studies to design potent CDK2-cyclin a inhibitors, Journal of Chemical Information and Modeling 47 (2007) 1526–1535.
[49] R. Dayam, T. Sanchez, O. Clement, R. Shoemaker, S. Sei, N. Neamati, β-Diketo acid pharmacophore hypothesis. 1. Discovery of a novel class of HIV-1 integrase inhibitors, Journal of Medicinal Chemistry 48 (2005) 111–120.