J Comput Aided Mol Des (2010) 24:131–141
DOI 10.1007/s10822-010-9321-0

Consensus model for identification of novel PI3K inhibitors in large chemical library

Chin Yee Liew · Xiao Hua Ma · Chun Wei Yap

Received: 25 September 2009 / Accepted: 2 February 2010 / Published online: 11 February 2010
© Springer Science+Business Media B.V. 2010

Abstract Phosphoinositide 3-kinases (PI3Ks) inhibitors have treatment potential for cancer, diabetes, cardiovascular disease, chronic inflammation and asthma. A consensus model consisting of three base classifiers (AODE, *k*NN, and SVM) trained with 1,283 positive compounds (PI3K inhibitors), 16 negative compounds (PI3K non-inhibitors) and 64,078 generated putative negatives was developed for predicting compounds with PI3K inhibitory activity of $IC_{50} \leq 10$ μM. The consensus model has an estimated false positive rate of 0.75%. Nine novel potential inhibitors were identified using the consensus model and several of these contain structural features that are consistent with those found to be important for PI3K inhibitory activities. An advantage of the current model is that it does not require knowledge of 3D structural information of the various PI3K isoforms, which is not readily available for all isoforms.

Keywords Drug discovery · Machine learning · Consensus model · Virtual screening · Phosphoinositide 3-kinases · PI3K inhibitors

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9321-0) contains supplementary material, which is available to authorized users.

C. Y. Liew · C. W. Yap (✉)
Pharmaceutical Data Exploration Laboratory, Department of Pharmacy, National University of Singapore, Singapore, Singapore
e-mail: phayapc@nus.edu.sg

X. H. Ma
Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Singapore, Singapore

Introduction

Phosphoinositide 3-kinases (PI3Ks) are a group of enzymes that can phosphorylate the 3′-hydroxyl position of phosphoinositides (PtdIns) at the inositol ring. PI3Ks are classified into three major classes on the basis of substrate specificity and sequence homology. They have a vital role in a variety of physiological processes such as metabolism regulation, cell survival, mitogenic signaling, cytoskeletal remodeling and vesicular trafficking [1, 2]. Thus, PI3Ks have been suggested to be implicated in the pathogenesis of cancer, diabetes, cardiovascular disease, chronic inflammation and asthma [3]. Consequently, the inhibitors of PI3Ks have been extensively explored as an attractive therapeutic candidates [3]. Wortmannin and LY294002 are two of the most widely used pan-PI3K inhibitors for PI3K signaling studies. Nonetheless, recent works are driven in search of isoform specific inhibitors [4, 5]. PI3K-α inhibitors (Class Ia) are being synthesized for its potential in antitumor and antidiabetic therapies [6–8]. On the other hand, inhibitors of PI3K-δ and PI3K-γ, isoforms of Class Ia and Class Ib respectively, are explored as potential anti-inflammatory agents for treatment of rheumatoid arthritis or autoimmune diseases [9].

This work will focus on the development of a computational model with large applicability domain and low false positive rate for the identification of potential PI3K inhibitors of all isoforms without the need for knowledge of 3D structural information of the protein target. The use of computational models to perform virtual screening for drug candidates is routinely conducted during the drug discovery process and has been used for drug discoveries in signal transduction [10, 11]. It is a favorable alternative to high-throughput screening (HTS) and combinatorial chemistry because virtual screening can identify drug candidates in a

fast and cheap manner. A limitation of computational virtual screening is that the predictions are predisposed to the structure-activity data in the model. However, virtual screening is still useful as it can help to overcome the limitation of HTS which may have very low hit rate or discovery of functional hits [12]. Furthermore, virtual screening is also useful because it helps to prioritize the compounds that should be biologically tested first [13]. Currently, there is a relative lack of structure-based models for PI3K inhibitors, which could be due to the limited 3D structural information. To date, PI3K-α and PI3K-γ alone or in complex with other molecules are the only isoforms with 3D-coordinates (X-ray diffraction) available in the Protein Data Bank [14, 15]. Based on these information, a study of PI3K-α selective inhibition using the approach of 3D-quantitative structure-activity relationship (QSAR) combined with homology modeling has been published [14]. Recently, the first structure-based virtual screening for PI3K inhibitors using various filtering methods like Lipinski-style rules and p110γ cavity docking was reported [16].

Ligand-based modeling is an alternative method to structure-based modeling for development of predictive models. It has the advantage of not requiring knowledge of the 3D structural information of the protein target. Thus this method was explored in this work as there is currently no 3D structural information for all PI3K isoforms. To the best of our knowledge, this work is the first ligand-based virtual screening study for PI3K inhibitors. Studies have shown that models developed using a limited number of compounds are likely to have limited applicability domain [17, 18], which may result in a large number of false positives when deployed for virtual screening of large chemical libraries [19]. Thus, a large number of compounds were used to develop the model so to expand the model's applicability domain. Consequently, 65,377 compounds from 8,423 chemical families were used to develop a consensus model for the identification of PI3K inhibitors not specific to any isoforms. First, the true negative compounds were enriched with putative inactive compounds to increase the quantity and diversity of the negative set using the method developed by Han et al. [19]. The significantly larger number of compounds in the training set will increase the applicability domain of the model and reduce the rate of false positives. Second, the consensus modeling or ensemble method was employed to improve classification accuracy by combining predictions of several base classifiers. Previous studies have shown that ensemble methods tend to perform better than single classifiers [20, 21]. Voting was chosen for the consensus method in this study and three base classifiers, *k*-nearest neighbor (*k*NN), aggregating one-dependence estimators (AODE) which is a variant of naive Bayes, and support vector machine (SVM) which has gained popularity in recent years, were used. SVM is useful for the classification of systems where there is limited knowledge on the mechanism or specific association between the activities and molecular properties [22] because it classifies compounds on the basis of the discriminative properties between active and inactive compounds rather than structural similarity to active compounds unlike most of the non-machine learning methods [23]. Hence, it is expected that a consensus model which considers the prediction results from the three base classifiers will be useful for the virtual screening of potential PI3K inhibitors from large chemical libraries.

Materials and methods

Training set

A total of 1,555 compounds and its reported IC_{50} for PI3Ks inhibition (pan-PI3K, PI3K-α, β, δ, or γ) were collected from patents and published studies within the 1994–2009 period. Information about the compounds, which includes IC_{50}(nM), structure in SMILES format and references (patents or PubMed Unique Identifier) are available in Table 1 of supplementary materials. The compounds were then categorized into positive (PI3K inhibitors) and negative (PI3K non-inhibitors) compounds using cutoff values of $IC_{50} \leq 10$ μM and $IC_{50} \geq 500$ μM respectively. Compounds with IC_{50} between these two criteria were excluded from the training set. This resulted in the selection of 1,283 positive and 16 negative compounds for the training set as shown in Fig. 1.

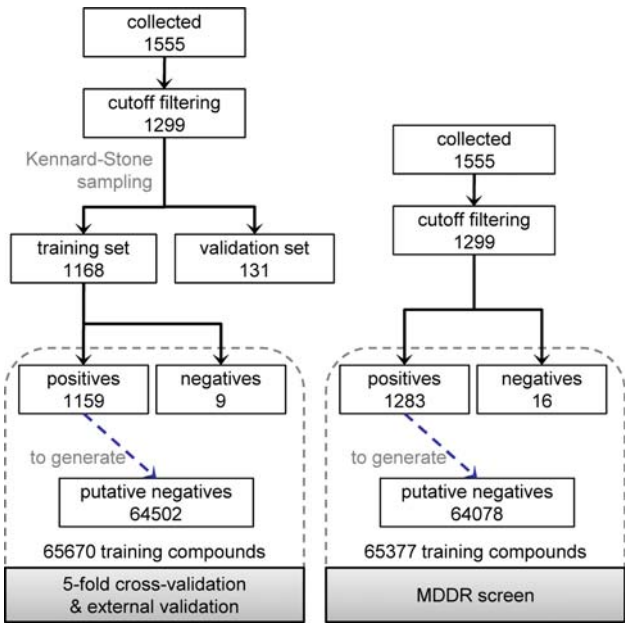

Fig. 1 Flowchart for selection of compounds for five-fold cross validation and external validation sets. The positive compounds were used as reference to generate putative negatives

A potential problem in our study is the small number of negative compounds in the training data set due to a lack of reported instances in the literature. This may cause the model to predict the negative class poorly [24] and have the problem of high false positive rate. This is undesirable when performing virtual screening of large compound libraries [25, 26]. Thus, this study has adopted the approach by Han et al. [19] to generate putative inactive compounds to augment the negative training set. This method can generate putative negatives without requiring the knowledge of actual inactive compounds and studies had shown that classification models derived from these putative negatives can perform reasonably well in virtual screening [23, 27]. Nonetheless, the effects of using a large number of putative negatives was examined to ensure that the change is not unacceptably detrimental to the identification of potential inhibitors.

The putative negative generation process was initiated by creating compound families through clustering known compounds, taken from PubChem and MDDR, on the basis of their molecular descriptors. These 13.7 million compounds with computable molecular descriptors from the program, MODEL [28], were subjected to *k*-means clustering. The clustering produced 8,423 compounds families. Given these 8,423 families, the families of the training compounds were analyzed and matched accordingly. Subsequently, an additional training data of 64,078 putative negatives were obtained by randomly selecting eight compounds from each of the families that do not contain any of the 1,283 positive compounds in the training set as illustrated in Fig. 2. For families with less than eight compounds, all their members were selected.

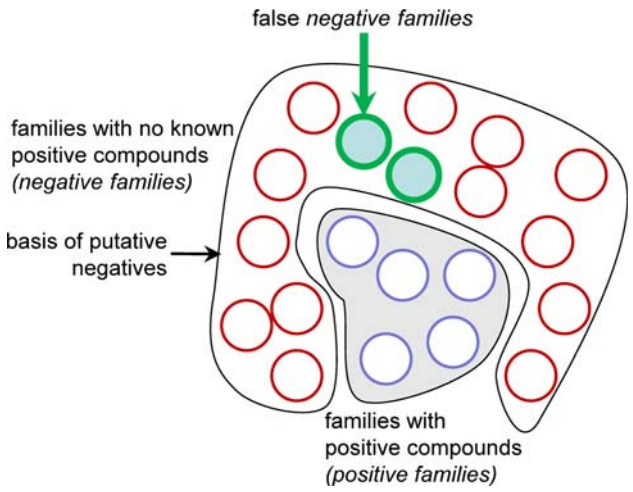

Fig. 2 Illustrating the use of negative families to obtain putative negative compounds. False negative families may arise from inclusion of undiscovered positive families

Molecular descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules. The 2D structures and 3D coordinates of the collected compounds were drawn and generated by using ChemDraw [29] and Corina [30], respectively. A total of 100 molecular descriptors, which are listed in Table 1, were computed by MODEL [28]. These include 13 simple molecular properties, 13 charge descriptors, 34 molecular connectivity and shape descriptors and 40 electrotopological state indices. The descriptors were selected from more than one thousand descriptors described in literatures by discarding those that are redundant and non-applicable to pharmaceutical agents [28]. Details of the descriptors can be found in the reference manual for MODEL [31].

Determination of structural diversity

The diversity index (DI), which is the average value of the similarity between pairs of compounds in a data set [32], was used to evaluate the structural diversity of the collected compounds:

$$\mathrm{DI} = \frac{\sum_{i,j \in D \wedge i \neq j} \mathrm{sim}(i,j)}{|D|\,(|D|-1)} \quad (1)$$

where sim(*i*, *j*)is a measure of similarity between compounds *i* and *j*, *D* is the data set and $|D|$ is set cardinality. The data set is more diverse when DI approaches 0. Tanimoto coefficient (*T*) were used to compute sim(*i*, *j*) in this study:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{k} x_{d_i} x_{d_j}}{\sum_{d=1}^{k} (x_{d_i})^2 + \sum_{d=1}^{k} (x_{d_j})^2 - \sum_{d=1}^{k} x_{d_i} x_{d_j}} \quad (2)$$

where *k* is the number of descriptors calculated for the compounds in the data set.

Modeling

All models were built and optimized using RapidMiner [33]. Three base classifiers were used: AODE, *k*NN and SVM. A compound is classified by the consensus model on the basis of the majority predictions from the three base classifiers. For example, if a compound is predicted as a non-inhibitor by the AODE and SVM model, but predicted as inhibitor by *k*-NN, the consensus model would deem the compound as a non-inhibitor based on the majority class.

k-Nearest neighbor (*k*NN)

k-Nearest neighbor is a type of lazy learner whereby it delays the learning of the training data until it is needed to

Table 1 Descriptors [31] used in this study

Descriptor class	No. of descriptors	Descriptors
Simple molecular properties	13	Molecular weight; Sanderson electronegativity sum; no. of atoms, bonds, rings; H-bond donor/acceptor; rotatable bonds; N or O heterocyclic rings; no. of C, N, O atoms
Charge descriptors	10	Relative positive/negative charge, 0th–2nd electronic-topological descriptors, electron charge density connectivity index, total absolute atomic charge, charge polarization, topological electronic index, local dipole index
Molecular connectivity and shape descriptors	37	1st–3rd order Kier shape index, Schultz/Gutman molecular topological index, total path count, 1–6 molecular path count, Kier molecular flexibility, Balaban/Pogliani/Wiener/Harary index, 0th edge connectivity, edge connectivity, extended edge connectivity, 0th–2nd valence connectivity, 0th–2nd order δ-χ index, 0th–2nd solvation connectivity, 1st–3rd order κ α shape, topological radius, centralization, graph-theoretical shape coefficient, eccentricity, gravitational topological index
Electrotopological state indices	40	Sum of E-state of atom types sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH, H-bond acceptors, all heavy/C/hetero atoms; sum of H E-state of atom types HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, H-bond donors

classify a test sample [34]. *k*NN does not produce a model and it classifies a test compound by searching for the training compounds that are similar in characteristics to the test compounds. The class of the test compound will be determined by the majority class of its neighbors. For this work, the best *k*NN model was obtained by optimizing simultaneously: (1) the number of nearest neighbor, k and (2) the distance measures, for example cosine similarity, Euclidean, or Manhattan distance; the best *k*NN model has a k of 3 when Manhattan distance was used.

Aggregating one-dependence estimators (AODE)

Naive Bayes is a simple classifier derived from the well-known Bayes' theorem. Whereas AODE is an improved implementation of the naive Bayes classifier reported to be as accurate and computationally more efficient than previous implementations [35]. In training, the classifier tries to learn the relationship between the class label and the molecular descriptors probabilistically, after which the class of an unknown compound is found by maximizing its conditional probability [34]. The details of the AODE algorithm can be found in the article by Webb et al. [35]. The data set has to be discretized first because the W-AODE module in RapidMiner is able to handle nominal data only. Hence, the MinMaxBinDiscretization module in RapidMiner was applied to the data sets prior to the learning and testing process. For this work, the best AODE model was obtained by optimizing simultaneously: (1) the number of bins for discretization and (2) the type of evaluation metrics, that is, M-estimate or Laplace correction; the best AODE model was obtained when M-estimate was used with data set in 100 bins.

Support vector machine (SVM)

Support vector machine is a machine learning method based on statistical learning theory [36]. It is a classifier that is less affected by duplicated data and has lower risk of model overfitting [34]. In linearly separable data, SVM tries to build a maximal margin hyperplane to separate positive compounds from negative compounds. The hyperplane is built on the basis of the data points called support vectors.

Nonlinear SVM is useful for classifying compounds of diverse structures which are usually not linearly separable. SVM maps the input vectors into a higher dimensional feature space by using a kernel function. The Gaussian radial basis function kernel which has been widely used and had consistently shown better performance [37, 38] were used in this study,

$$K(x_i, x_j) = \mathrm{e}^{-\left\|x_i - x_j\right\|^2 / 2\sigma^2} \quad (3)$$

For the SVM model in this study, hard margin SVM was used and σ was found to be 0.73 for the best performing model.

Applicability domain

The applicability domain (AD) defines the space where a model may be applied; the space may be of physico-chemical, structural or biological in nature [39]. A model may still be used to classify a compound that falls outside of the model's AD, however, the prediction should be deemed less reliable. The AD of the consensus model was calculated based on the range [39] of the individual descriptors; the minimum and maximum values of each

molecular descriptors in consideration of all the compounds in the training set were used. Figure 3 is a visualization of the use of ranges to define the AD for a model consisting of three descriptors. Hence, for this work, the AD is defined by a hyper-rectangle and compounds that violated one or more of the 100 molecular descriptor ranges were excluded from the prediction process.

Model validation and screening

First, the performance of the consensus model, $CM_{Tr+PutNeg}$ (subscript denotes the set of compounds used for training the model: Tr, collected training set; PutNeg, putative negative compounds; Ext, external validation set), was estimated using five-fold cross-validation which is a type of internal validation. In five-fold cross-validation, the training set was divided into five groups of approximately equal size through stratified sampling. Training of the consensus model was carried out with four subsets of data after which the performance of the model was tested with the fifth subset. This process was repeated five times resulting in five combinations so that every subset was used as the testing set once. This validation step only involved 65,670 training compounds as 131 compounds, as shown in Fig. 1, were set aside for external validation.

Second, an external validation was also conducted for the consensus model, $CM_{Tr+PutNeg}$. External validation is essential as cross-validation may not always be reflective of the predictive power of a model [40]. The external validation set in this study was selected by the Kennard–Stone sampling module in RapidMiner. A total of 131 compounds were selected from the 1,299 collected compounds and they were not used in training of the consensus model, $CM_{Tr+PutNeg}$.

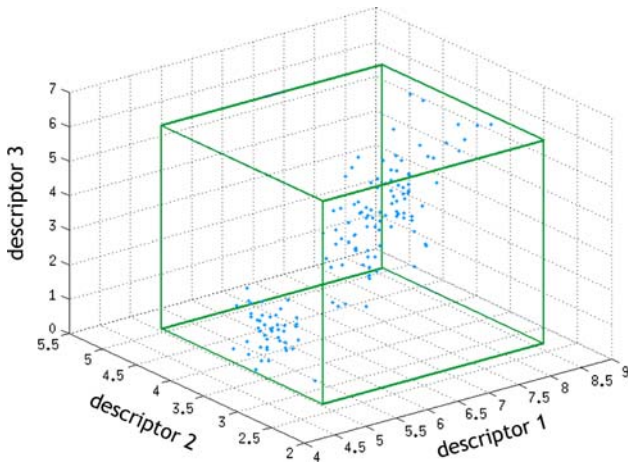

Fig. 3 The box that encloses the data is the applicability domain of a model built from a data set with three descriptors

Third, in order to evaluate the suitability of the consensus model for identifying PI3K inhibitors from large chemical libraries, compounds in MDDR were screened. As the external validation set that contains 131 compounds was subsequently found to contain a substantial number of compound families that were not represented in the original training set of 1,168 compounds, the entire external validation set was added to the training set and a set of 64,078 putative negative compounds were regenerated to match the new profile of the training set. A new consensus model ($CM_{Tr+PutNeg+Ext}$) was then developed from the new training set and used for screening MDDR compounds.

The MDDR contained eleven compounds with PI3K inhibitory activity of $IC_{50} \leq 10$ μM and these were labeled as "known inhibitors". A group of MDDR compounds were excluded from the evaluation of prediction performance of the models even though they were reported to have PI3K inhibitory activity because they did not satisfy the cutoff values or their IC_{50} values were not reported. However, this group of compounds were included in the search for novel potential inhibitors. It is to be noted that none of the MDDR compounds were present in the training set or putative negatives.

Evaluation of prediction performance

In the case of classification methods, the performance of machine learning methods can be assessed by the quantity of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) [41]. The prediction accuracy for positive compounds (PI3K inhibitors) and negative compounds (PI3K non-inhibitors) are sensitivity, $SE = \frac{TP}{TP+FN}$ and specificity, $SP = \frac{TN}{TN+FP}$, respectively. The overall prediction performance can be calculated by the overall prediction accuracy (Q):

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

and Matthew's correlation coefficient [42] (MCC):

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{5}$$

For the performance of the consensus model in virtual screening, the yield (percentage of known inhibitors that were predicted correctly), hit-rate (HR = percentage of known inhibitors in compounds predicted as positives), false positive rate (FPR = percentage of non-inhibitors predicted as positives) and enrichment factor (EF = ratio of hit-rate to the percentage of known inhibitors in MDDR) which shows the magnitude of hit-rate improvement over random selection were evaluated.

Identification of novel potential inhibitors

The selection of suitable novel candidates for biological testing of PI3K inhibitory activities was carried out by identifying those compounds that were predicted to be potential inhibitors by the consensus model ($CM_{Tr+PutNeg+Ext}$) with a prediction confidence (ranged from 0 to 1) of 1. The list of selected compounds were further refined by removing those compounds that do not meet the minimum prediction confidence (value of 1 for AODE and *k*NN, value of greater than 0.95 for SVM) in at least two of the three base classifiers. The similarity of the remaining compounds to the PI3K inhibitors in the training set were calculated and those that were sufficiently dissimilar were identified as potential candidates for biological testing. The rationale for selecting dissimilar compounds is to discover novel scaffolds for PI3K inhibitors. This is important as these novel compounds could provide new information on the mechanism of PI3K inhibition. They could also lead to a new chemical class of drugs for treatment of PI3Ks related diseases.

Results

Data set diversity and distribution

Table 2 shows that the 1,283 PI3K inhibitors have an diverse-to-intermediate DI of 0.629, which is in between that of known estrogen and benzodiazepine receptor ligands. A three dimensional visualization of the collected compounds using the first three principle components after principle component analysis (PCA) is shown in Fig. 4. The result shows that the negative compounds tend to cluster at the edge in two groups, however, there was no clear separation between the positive and negative compounds. There was no evidence to exclude any compounds as outliers, although there were a few remote compounds. Lastly, the 131 compounds isolated for external validation through Kennard–Stone sampling were well distributed in the chemical space of the collected compounds.

Table 2 Diversity index (DI) of several compounds classes in descending order of structural diversity

Chemical class	No. of compounds	DI
Satellite structures [43]	9	0.250
National Cancer Institute diversity set [43]	1,990	0.452
FDA approved drugs [43]	1,183	0.452
Estrogen receptor ligands [43]	1,009	0.511
PI3K inhibitors in training set (this study)	1,283	0.629
Benzodiazepine receptor ligands [43]	405	0.686
Dihydrofolate reductase inhibitors [43]	756	0.727
Penicillins [43]	59	0.790
Fluoroquinolones [43]	39	0.791
Cephalosporins [43]	73	0.812
Cyclooxygenase 2 inhibitors [43]	467	0.840

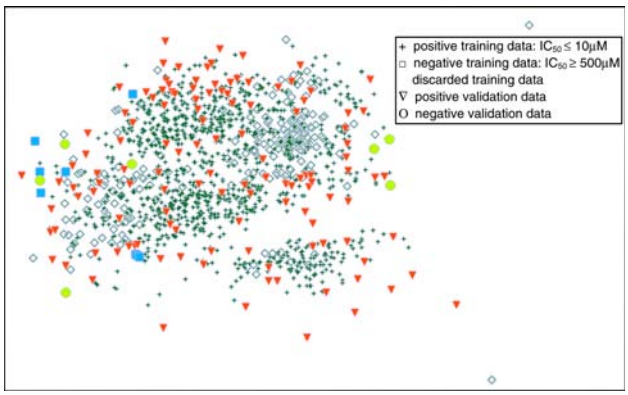

Fig. 4 Visualization of the chemical space for the collected data using the first three principle components from PCA

Figure 5 shows the distribution of PI3K inhibitors in terms of compound families. The analysis found that the 1,159 inhibitors in the training set and 124 inhibitors in the external validation set belonged to 345 and 116 families respectively. Together, they occupied 398 unique families from the total of 8,423 families. The characteristic of the external validation set was different from the positive training data set as only 54.3% of the families in the validation set were represented in the training set. These two characteristics will be useful to evaluate the model performance on familiar and unfamiliar (novel) compounds.

Applicability domain

For the consensus model trained with 65,377 compounds, all except two long chained molecules in the MDDR data set were within the applicability domain. If putative negatives were not used in model building, that is, with a

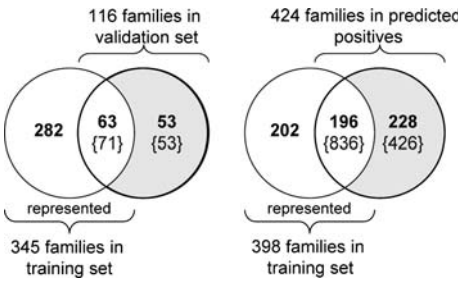

Fig. 5 Distribution of families for the 124 positive compounds in validation set and 1,262 virtual screening predicted positives (the number of compounds is given in curly brackets). Families in the shaded region are not represented in the training set

training set of 1,299 compounds, only 105,452 MDDR compounds were within the applicability domain.

Model performances

Table 3 gives the performance of the consensus model ($CM_{Tr+PutNeg}$) for predicting PI3K inhibitors and non-inhibitors by means of five-fold cross-validation and an external validation set. The consensus model in five-fold cross-validation had performed consistently well in predicting positive compounds (average SE = 96.1%) and also in predicting negative compounds (average SP = 99.7%) with an overall accuracy of 99.7% and MCC of 0.915. When tested on the external validation set, the consensus model performed with an overall sensitivity of 77.4%, specificity of 100% and accuracy of 78.6%.

168,014 compounds in MDDR were screened with the models trained with 65,377 compounds. The results are given in Table 4. The consensus model ($CM_{Tr+PutNeg+Ext}$) had predicted 1,262 compounds to have PI3K inhibitory activity with a low false positive rate of 0.75%. The consensus model was able to predict 7 out of the 11 known inhibitors correctly, giving a yield of 63.6%. Analysis of the compound families of these 1,262 compounds has shown that they belong to 424 families and 46.2% of these are represented in the training set. When Lipinski's rule of five was applied on the MDDR data set, only 85,737 MDDR compounds were eligible for screening. The results in Table 4 shows that, although the number of compounds eligible for screening was reduced to half, the consensus model ($CM_{Tr+PutNeg+Ext}$) performance in terms of yield (66.7%), hit-rate (0.7%) and false positive rate (0.67%) are comparable to those that were filtered with AD only.

Cumulative gains for the discovery of known inhibitors by the consensus model is shown in Fig. 6. The rate of known inhibitor discovery of a random model was taken as 11/168016.

A total of 26 compounds in MDDR have met the minimum prediction confidence requirements. Seven of these compounds belonged to the group that were reported to

Table 4 Performance of the consensus model in virtual screening of MDDR Compounds

Filter	Known inhibitors[a]	
	AD	AD and L5[b]
No. of MDDR compounds passed filter	168,014	85,737
Known inhibitors	11	6
Predicted positives	1,262	575
Hits[c]	7	4
Yield	63.6%	66.7%
Hit-rate	0.55%	0.70%
False positive rate	0.75%	0.67%
Enrichment factor	85	99

[a] Compounds in MDDR identified to have PI3K inhibitory activity $IC_{50} \leq 10$ μM

[b] Applicability domain and Lipinski's rule of five

[c] Predicted positive compounds that are known inhibitors in MDDR

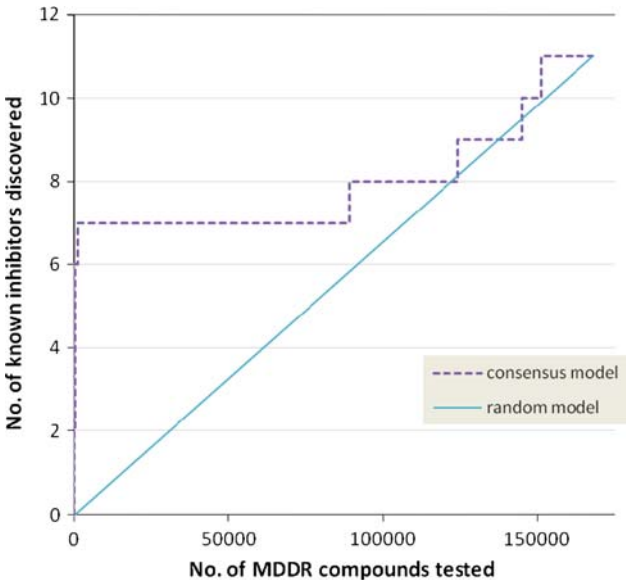

Fig. 6 Cumulative gains chart for the discovery of known inhibitors

Table 3 Classification performance of consensus model in predicting PI3K inhibitory activity

Test	No. of test compounds	TP	FN	SE (%)	TN	FP	SP (%)	*Q* (%)	MCC
Five-fold cross-validation									
Fold 1	13,135	225	7	97.0	12,876	27	99.8	99.7	0.929
Fold 2	13,134	219	13	94.4	12,875	27	99.8	99.7	0.915
Fold 3	13,134	226	6	97.4	12,855	47	99.6	99.6	0.896
Fold 4	13,134	225	7	97.0	12,866	36	99.7	99.7	0.913
Fold 5	13,133	219	12	94.8	12,877	25	99.8	99.7	0.921
Average	13,134	223	9	96.1	12,870	32	99.7	99.7	0.915
External validation	131	96	28	77.4	7	0	100	78.6	0.393

have PI3K activity but without sufficient IC_{50} information. Nine compounds, which are shown in Fig. 7, in the remaining 19 were found to be the most distant in the chemical space (average $T = 0.456$ to $T = 0.499$) from the inhibitors in the training set and thus were prioritized as suitable novel candidates for biological testing of PI3K inhibitory activity.

Molecular descriptors

An analysis of the support vectors from the SVM model was carried out to examine the differences between the 100 molecular descriptor means of the inhibitors and non-inhibitors. The difference for the means of 7 molecular descriptors were found to be statistically significant. Among the support vectors, PI3K inhibitors have higher values in terms of the number of hydrogen-bond acceptor, number of oxygen atoms, 0th valence connectivity index, and sum of E-state of atom type aaN and sSH. On the other hand, PI3K non-inhibitors has higher total path count and sum of E-state of atom type aaNH.

Discussions

The model

This work has used a few strategies that are aimed to develop a model with large applicability domain and low false positive rate so that it is suitable for virtual screening purposes even without the knowledge of 3D structural information of the protein target. The strategies include the use of two cutoff vales to divide the inhibitors from non-inhibitors, putative negatives, and consensus modeling.

First, although it is common to use a single cutoff value to separate the inhibitors from non-inhibitors in classification models, this work chose to use two cutoff values; $IC_{50} \leq 10$ μM for positive compounds and $IC_{50} \geq 500$ μM

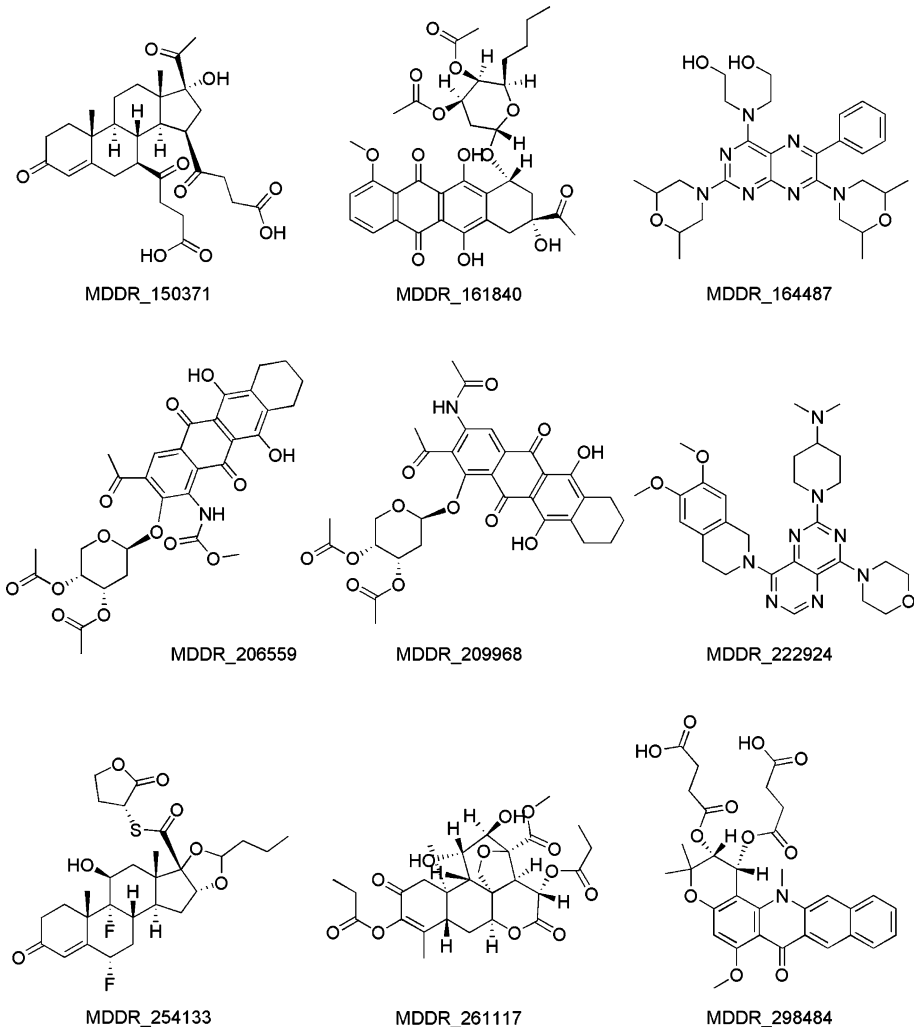

Fig. 7 A selection of MDDR compounds not reported as PI3K inhibitors that have the highest prediction confidence for the consensus model and at least two of the three base classifiers. These nine compounds are also the most dissimilar from the positive training set

for negative compounds. The use of a single cutoff value of $IC_{50} \leq 10$ μM will categorize molecules with weak activity into the negative group. Given that model learns from samples, the inaccurate information will cause the model to miss any potential drug leads when applied for screening, which is undesirable because weak inhibitors may be modified into potent drugs. Conversely, a single, large cutoff value of $IC_{50} \geq 500$ μM will include unnecessary compounds into the positive group, which will frequently result in a large number of false positive when applied for screening. This is equally undesirable as it wastes considerable resources to synthesize and biologically test these false positive compounds. Therefore, two cutoff values were chosen to minimize the risk of incorporating inhibitors into the negative group and also to reduce the false positive rate so that positive predictions will be limited to a reasonable amount for test prioritization. In addition, the wide margin between the cutoff values can, to some extent, attenuate the bias in the data set that may have arisen from inter-laboratories and equipment variations. Nonetheless, although elimination of too many in-between compounds is a potential shortcoming to this technique, it did not create much problem in this work as only 16% of the collected compounds were removed and a majority of the positive compounds have IC_{50} of ≤ 1 μM.

Second, compounds with very weak activities are rarely reported in the literature as authors typically present their most potent findings in their publications. Correspondingly, negative compounds are overwhelmed by the number of positive compounds in training which subsequently produces model with high false positive rate. To increase the amount and diversity of the negative training set in this work, true negative compounds were supplemented with putative negatives. The effect of adding putative negatives was examined to ensure that the model is not overfitted, thus becoming insensitive to potential inhibitors. Therefore, the performance of this method was evaluated by validating the models internally and externally using five-fold cross-validation and external validation respectively. The result of this study showed that, although the false positives rate was low, the consensus model ($CM_{Tr+PutNeg}$) trained with putative negative was still able to generalize well as indicated by the high sensitivity value of 96.1% and 77.4% for the five-fold cross-validation and external validation. Moreover, the applicability domain of the model was enlarged with the addition of putative negatives; 168,014 MDDR compounds were eligible for screening by $CM_{Tr+PutNeg+Ext}$ as compared to 105,452 compounds when putative negatives were excluded. Hence, the putative negatives were found to be useful, with additional benefits, to overcome the lack of negative compounds for training.

A possible disadvantage is the probable inclusion of undiscovered inhibitors into the negative set as illustrated in Fig. 2, resulting in a model that cannot identify an active compound that has similar structure to the putative negative compounds. The extent of this risk is unknown but the results of this work and two other studies [23, 27] have shown that such unwanted effect is expected to be relatively small and it was still possible for a substantial proportion of positive compounds to be classified correctly despite their membership in negative families. Nonetheless, the search for known PI3K inhibitors in this work was carried out to be as extensive as possible to minimize this risk.

Last, the consensus method was introduced to improve the prediction performance of the base classifiers. The consensus model was found to have better prediction accuracies than the base classifiers and its prediction performance in both five-fold cross-validation and external validation were consistent (results not shown). This is unlike the base classifiers which had different prediction performance ranking when different validation methods were used. The consensus model also has a much higher discovery rate for known inhibitors compared to a random model as shown in Fig. 6. Moreover, as shown in Table 4, the performance of the consensus model ($CM_{Tr+PutNeg+Ext}$) was not affected significantly by the application of Lipinski's rule of five on the chemical library before screening. Therefore, these results suggest that the consensus model is potentially useful for screening large compound libraries for PI3K inhibitors.

Application of model for novel PI3K inhibitor design

The consensus model presented in this work might be useful for novel PI3K inhibitor discovery because the model is able to predict inhibitors unrepresented in the training and also compounds that are different from the training set.

Figure 5 shows that 45.6% of the positive compound families in the external validation were not represented in the training set and they were grouped under negative families. The consensus model ($CM_{Tr+PutNeg}$) has a sensitivity of 77.4% despite the lack of positive families representation. Further analysis showed that represented compounds were predicted better than the unrepresented ones with sensitivity scores of 95.8% and 52.8% respectively. Although the sensitivity for unrepresented compounds seemed low, this result must be viewed with the perspective that the consensus model has low false positive rate, which means that the model has a high precision value. Thus, when the model predicts an unrepresented compound to be an inhibitor, it is very likely that the compound is a true inhibitor. This is in contrast to that of a random model which is only 50% certain of finding a true inhibitor. The difference in sensitivities for represented and

unrepresented compounds highlighted the importance of compound families knowledge for optimum model performance. Knowledge of more positive families will bring about the reduction of false negative families risk as illustrated in Fig. 2. Nonetheless, given that the consensus model has a reasonably good sensitivity and high precision for unrepresented compounds, it was likely that a compound classification was not decided by its membership in represented family only, but also on the basis of the differing characteristics between inhibitors and non-inhibitors. Therefore, the consensus model presented in this work have the potential to identify potential inhibitors from novel compound families.

Analysis of the three most recent publications on PI3K inhibitors synthesis showed that the calculated Tanimoto coefficient (T) of one compound to another within the same publication can range from $T_{\text{average}} = 0.703$ to $T_{\text{average}} = 0.971$. In this work, the average Tanimoto coefficient for the 1,255 predicted positive MDDR compounds (known inhibitors excluded) and the 7 hits (Table 4) calculated against the 1,283 positive training compounds ranged from 0.283 to 0.516 and 0.496 to 0.504. This suggests that the consensus model presented in this work was able to make a positive prediction even if the compound appears distant from the positive training compounds in the chemical space defined by the descriptors in this work. This is important because compounds with greater dissimilarity from currently known inhibitors may be explored as new starting points for drug design, which may have been difficult to discover through the traditional synthesis process.

Among the nine compounds in Fig. 7 that should be prioritized as suitable novel candidates for biological testing of PI3K inhibitory activity, a majority were reported as antineoplastics by MDDR and one of them is an anti-asthmatic which concurred with the potential uses of PI3K inhibitors. Some of these compounds contain structural features that were found to be essential for PI3K inhibition [14]: for PI3K inhibition, a central (hetero)aromatic scaffold carrying an hydrogen-bond acceptor is required, while a small lipophilic group on one side together with two H-bond acceptors on the other side are required for PI3K-α specific inhibitions. These features are more apparent in MDDR_164487, MDDR_222924, and MDDR_298484. Hence, these nine compounds are likely to be novel PI3K inhibitors and could serve as lead compounds for new inhibitors design.

Conclusion

A consensus model suitable for virtual screening purposes even without the knowledge of 3D structural information of the protein target was developed from a large training set of PI3K inhibitors and non-inhibitors. The consensus model was validated in a number of ways: internal validation using five-fold cross-validation, external validation with compounds not used during model development, and virtual screening of MDDR. The consensus model is capable of identifying novel PI3K inhibitors from large chemical libraries with false positive rate of 0.75%. The consensus model also a higher discovery rate for known inhibitors when compared to a random model. The use of Lipinski's filter prior to virtual screening did not affect the performance of the consensus model significantly. This suggests that the consensus model can be safely integrated into existing virtual screening pipelines without affecting its performance. Several potential drug leads were presented and they were found to contain structural features that have been reported to be associated with PI3K inhibitory activities. Hence, the consensus model presented in this work is potentially useful to complement HTS in screening large chemical libraries for novel PI3K inhibitors.

Acknowledgments Our appreciation to Professor Chen Yu Zong (Bioinformatics and Drug Design Group, National University of Singapore) for his valuable discussions.

References

1. Cantley LC (2002) Science 296:1655–1657
2. Wymann MP, Zvelebil MJ, Laffargue M (2003) Trends Pharmacol Sci 24:366–376
3. Marone R, Cmiljanovic V, Giese B, Wymann MP (2008) Biochim Biophys Acta, Proteins Proteomics 1784:159–185
4. Knight ZA, Gonzalez B, Feldman ME, Zunder ER, Goldenberg DD, Williams O, Loewith R, Stokoe D, Balla A, Toth B, Balla T, Weiss WA, Williams RL, Shokat KM (2006) Cell 125:733–747
5. Xie P, Williams DS, Atilla-Gokcumen GE, Milk L, Xiao M, Smalley KS, Herlyn M, Meggers E, Marmorstein R (2008) ACS Chem Biol 3:305–316
6. Hayakawa M, Kaizawa H, Moritomo H, Koizumi T, Ohishi T, Okada M, Ohta M, Tsukamoto S, Parker P, Workman P, Waterfield M (2006) Bioorg Med Chem 14:6847–6858
7. Kendall JD, Rewcastle GW, Frederick R, Mawson C, Denny WA, Marshall ES, Baguley BC, Chaussade C, Jackson SP, Shepherd PR (2007) Bioorg Med Chem 15:7677–7687
8. Wee S, Lengauer C, Wiederschain D (2008) Curr Opin Oncol 20:77–82
9. Pomel V, Klicic J, Covini D, Church DD, Shaw JP, Roulin K, Burgat-Charvillon F, Valognes D, Camps M, Chabert C, Gillieron C, Francon B, Perrin D, Leroy D, Gretener D, Nichols A, Vitte PA, Carboni S, Rommel C, Schwarz MK, Ruckle T (2006) J Med Chem 49:3857–3871
10. Fischer PM (2008) Biotechnol J 3:452–470
11. Seifert MH, Lang M (2008) Mini-Rev Med Chem 8:63–72
12. Chen X, Wilson LJ, Malaviya R, Argentieri RL, Yang SM (2008) J Med Chem 51:7015–7019
13. Truchon JF, Bayly CI (2007) J Chem Inf Model 47:488–508
14. Frédérick R, Denny WA (2008) J Chem Inf Model 48:629–638
15. RCSB Protein Data Bank. www.pdb.org. Accessed 6 Aug 2009

16. Frédérick R, Mawson C, Kendall JD, Chaussade C, Rewcastle GW, Shepherd PR, Denny WA (2009) Bioorg Med Chem Lett 19:5842–5847
17. Gramatica P (2007) QSAR Comb Sci 26:694–701
18. Parker CN, Bajorath J (2006) QSAR Comb Sci 25:1153–1161
19. Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, Chen YZ (2008) J Mol Graphics Model 26:1276–1286
20. Yap CW, Chen YZ (2005) J Chem Inf Model 45:982–992
21. Lau QP, Wynne H, Mong Li L, Ying M, Liang C (2007) 19th IEEE International Conference on tools with artificial intelligence. ICTAI 1:350–357
22. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ (2004) J Chem Inf Comput Sci 44:1497–1505
23. Ma XH, Wang R, Yang SY, Li ZR, Xue Y, Wei YC, Low BC, Chen YZ (2008) J Chem Inf Model 48:1227–1237
24. Schierz A (2009) Journal of Cheminformatics 1:21
25. McGovern SL, Caselli E, Grigorieff N, Shoichet BK (2002) J Med Chem 45:1712–1722
26. Shoichet BK (2004) Nature 432:862–865
27. Liew CY, Ma XH, Liu X, Yap CW (2009) J Chem Inf Model 49:877–885
28. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) J Chem Inf Comput Sci 44:1630–1638
29. CambridgeSoft Desktop Software–ChemDraw (Windows/Mac). http://www.cambridgesoft.com/. Accessed 6 Aug 2009
30. CORINA: Generation of 3D coordinates. http://www.molecular-networks.com/software/corina/index.html. Accessed 6 Aug 2009
31. MODEL reference manual. http://jing.cz3.nus.edu.sg/model/. Accessed Aug 6, 2009
32. Perez JJ (2005) Chem Soc Rev 34:143–152
33. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) KDD '06: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining 935–940
34. Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison Wesley
35. Webb GI, Boughton JR, Wang Z (2005) MLear 58:5–24
36. Vapnik V (1995) The nature of statistical learning theory. Springer, New York, London
37. Czerminski R, Yasri A, Hartsough D (2001) Quant Struct-Act Relat 20:227–240
38. Trotter M, Buxton B, Holden SB (2001) Measurement and Control 34:235–239
39. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) ATLA Altern Lab Anim 33:445–459
40. Tropsha A, Gramatica P, Gombar Vijay K (2003) QSAR Comb Sci 22:69–77
41. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Bioinformatics 16:412–424
42. Matthews BW (1975) Biochim Biophys Acta 405:442–451
43. Yap CW, Xue Y, Li H, Li ZR, Ung CY, Han LY, Zheng CJ, Cao ZW, Chen YZ (2006) Mini-Rev Med Chem 6:449–459