Bioinformatics, 31(11), 2015, 1788–1795
doi: 10.1093/bioinformatics/btv055
Advance Access Publication Date: 31 January 2015
Original Paper

Systems biology

Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources

Zhongyang Liu[1,2,†], Feifei Guo[1,2,3,†], Jiangyong Gu[4,†], Yong Wang[5], Yang Li[1,2], Dan Wang[1,2], Liang Lu[1,2], Dong Li[1,2,*] and Fuchu He[1,2,*]

[1]State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 100850, China, [2]National Center for Protein Sciences Beijing, Beijing 102206, China, [3]Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing 100005, China, [4]Beijing National Laboratory for Molecular Sciences, State Key Lab of Rare Earth Material Chemistry and Applications, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China and [5]Preclinical School, Beijing University of Chinese Medicine, Beijing 100029, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors'.
Associate Editor: Alfonso Valencia

Received on July 8, 2014; revised on December 16, 2014; accepted on January 26, 2015

Abstract

Motivation: Anatomical Therapeutic Chemical (ATC) classification system, widely applied in almost all drug utilization studies, is currently the most widely recognized classification system for drugs. Currently, new drug entries are added into the system only on users' requests, which leads to seriously incomplete drug coverage of the system, and bioinformatics prediction is helpful during this process.

Results: Here we propose a novel prediction model of drug-ATC code associations, using logistic regression to integrate multiple heterogeneous data sources including chemical structures, target proteins, gene expression, side-effects and chemical–chemical associations. The model obtains good performance for the prediction not only on ATC codes of unclassified drugs but also on new ATC codes of classified drugs assessed by cross-validation and independent test sets, and its efficacy exceeds previous methods. Further to facilitate the use, the model is developed into a user-friendly web service *SPACE* (**S**imilarity-based **P**redictor of **A**TC **C**od**E**), which for each submitted compound, will give candidate ATC codes (ranked according to the decreasing *probability_score* predicted by the model) together with corresponding supporting evidence. This work not only contributes to knowing drugs' therapeutic, pharmacological and chemical properties, but also provides clues for drug repositioning and side-effect discovery. In addition, the construction of the prediction model also provides a general framework for similarity-based data integration which is suitable for other drug-related studies such as target, side-effect prediction etc.

Availability and implementation: The web service *SPACE* is available at http://www.bprc.ac.cn/space

Contact: hefc@nic.bmi.ac.cn or lidong.bprc@foxmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Anatomical Therapeutic Chemical (ATC) classification system developed and maintained by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOCC), is currently the most widely recognized classification system for drugs and has been widely used in almost all drug utilization studies (Dunkel *et al.*, 2008). It divides drug substances into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. The ATC classification system has five levels representing finer and finer classification for drugs, in which the first level has 14 anatomical groups, with pharmacological/therapeutic subgroups as the second level, the third and fourth levels are chemical/pharmacological/therapeutic subgroups and the fifth level is the chemical substance (http://www.whocc.no/atc/structure_and_principles/). For example, a complete classification of acetylsalicylic acid is B (blood and blood forming organs, first level), B01 (anti-thrombotic agents, second level), B01A (anti-thrombotic agents, third level), B01AC (platelet aggregation inhibitors excl. heparin, fourth level) and B01AC06 (acetylsalicylic acid, fifth level). A drug can be assigned more than one ATC code when it has multiple different therapeutic uses. For example, besides as a platelet aggregation inhibitor (B01AC06) mentioned above, acetylsalicylic acid is also used as a drug for 'local oral treatment' (A01AD05) and used as an 'analgesic and antipyretic' (N02BA01).

Currently WHOCC establishes new drug entries in the ATC classification system only on requests from the users including manufacturers, regulatory agencies and researchers, which leads to serious incompletion of the drug coverage of the system (http://www.whocc.no/atc/structure_and_principles/). Recently as the emergence and accumulation of various data sources in drug studies such as chemical structures, target proteins, side-effects and drug-induced gene expression profiles, bioinformatics prediction of ATC classification of drugs becomes feasible. The prediction of ATC classification of drugs not only contributes to the understanding of drugs' therapeutic, pharmacological and chemical properties, but also provides valuable information for drug side-effect discovery and drug repositioning. Meanwhile ATC classification prediction for chemical compounds is also helpful for new drug development, especially complementing with other virtual screening methods.

Recently several computational methods have been proposed to predict ATC classification of drugs based on different data sources. Dunkel *et al.* (2008) presented the first one to predict ATC codes based on chemical structures, the assumption of which is that structurally similar compounds tend to have similar medical indication areas, and the method was developed into a web service called SuperPred. The iSEA (indication similarity ensemble approach) method also correlated a drug and an ATC class based on the drug's chemical fingerprint similarity with all drugs belonging to this ATC class (Wu *et al.*, 2013). Expect chemical structures, Chen *et al.* (2012, 2014) successively introduced chemical-chemical associations and ChEBI ontology to predict ATC classification. Gurulingappa *et al.* (2009) assigned drugs to ATC classes based on text mining information. NetPredATC method used SVM (support vector machine) to integrate chemical structures and target proteins for ATC code prediction (Wang *et al.*, 2013). Although these methods obtained certain efficacy on ATC code prediction, there are still some limitations: (i) Most of these approaches were only based on a single or two data sources. Different data sources provide supporting evidence for ATC code prediction from different perspectives, which usually have some complementarity. The performance of the prediction method is supposed to be further improved if multiple heterogeneous data sources are efficiently integrated. (ii) All these jobs only assessed the efficacy of methods on a certain ATC classification level or a certain ATC category (e.g. Gurulingappa *et al.* (2009) only validated their method on the class of 'cardiovascular system', one of the 14 main ATC groups). And the prediction performance of ATC codes of unclassified drugs (whose ATC codes are unknown) and that of new ATC codes of classified drugs (whose ATC codes are partly known) were not distinguished. (iii) Except SuperPred mentioned above (Dunkel *et al.*, 2008), all other methods haven't provided user-friendly program package or been developed into a web service, which severely restricts the utility of these methods.

To eliminate these limitations, in this article, we propose a novel prediction model for ATC classification prediction, using logistic regression (LR) framework to integrate multiple heterogeneous data sources including chemical structures, target proteins, drug-induced gene expression profiles, side-effects and chemical–chemical associations, and comprehensively assesse its efficacy and then compare it with previous methods, aiming to further improve the performance of ATC code prediction. Further, to facilitate the use, the proposed method is developed into a user-friendly web service *SPACE* (**S**imilarity-based **P**redictor of **A**TC **C**od**E**).

2 Materials and Methods

2.1 Golden standard dataset and independent test dataset

To develop an ATC code prediction model, first the golden standard positive (GSP) and negative (GSN) datasets were determined. In this article, we developed a prediction model for each ATC classification level from 1 to 4 in a similar way, and thus GSP and GSN sets were constructed for each level. The GSP sets of four levels were composed of known drug-ATC code relationships extracted from DrugBank database (downloaded on July 7, 2013) (Law *et al.*, 2014), including 1333 small molecule drugs and their ATC codes of corresponding levels (Supplementary Table S1). For a certain level, suppose $\boldsymbol{D}$ and $\boldsymbol{C}$ are respectively drug and code spaces of the GSP set, and $\boldsymbol{n}$ is the number of drug-code pairs in the GSP set. To construct the GSN set, we removed the GSP set from $\boldsymbol{D} \times \boldsymbol{C}$ drug-code pairs, and then randomly picked $\boldsymbol{n}$ ones from the remaining pairs to construct the GSN set.

To assess the performance of the prediction model, besides cross-validation, the independent test set was also used. We removed the GSP and GSN sets from known drug-ATC code associations derived from Kyoto Encyclopedia of Genes and Genomes (KEGG) database (version: July 5, 2012) (Kanehisa *et al.* 2014), and the remaining ones constituted the independent test positive set (Supplementary Table S1). To further discriminate the efficacy of the model on the prediction of ATC codes of unclassified drugs (whose ATC codes are unknown) and that of new ATC codes of classified drugs (whose ATC codes are partly known), we divided the independent test positive set into two subsets, which were respectively 'independent test positive set-drug' and 'independent test positive set-code'. The former was composed of drugs, which don't have any overlap with those of the GSP/GSN set, and their corresponding codes, and the latter included drug-code pairs whose drug space was totally included in that of the GSP/GSN set (Supplementary Table S1). The independent test negative set for each of these three positive sets was respectively constructed in the similar way to the GSN set's construction, the drug-code pairs in which excluded not only those in the corresponding independent test positive set but also those in the GSP and GSN sets.

2.2 Similarity scores

Here we used 6 scores to measure drug-drug similarity, respectively based on chemical structures, target proteins, drug-induced gene expression profiles, side-effects and chemical–chemical associations.

FP2 fingerprint and functional group similarity scores were computed both based on chemical structures. FP2 is a hash-based binary fingerprint which is generated by indexing the molecule structure's all possible linear fragments with a length ranging from 1 to 7 atoms (O'Boyle *et al.*, 2011). Functional groups of a drug molecule are also represented as a binary vector, each dimension giving the presence (1) or absence (0) of a particular functional group in the molecule. Chemical structures represented by InChI (Heller *et al.*, 2013) were separately downloaded from DrugBank (downloaded on July 7, 2012) (Law *et al.*, 2014) for the drugs in the golden standard set and KEGG database (version: July 5, 2012) (Kanehisa *et al.*, 2014) for those in the independent test set. Based on chemical structures, FP2 fingerprints of drugs were produced by Open Babel (v2.3.2) (O'Boyle *et al.*, 2011), and functional group vectors by Checkmol program (version: April 29, 2013) which can recognize a total number of 204 functional groups (Haider *et al.*, 2010). We used the Tanimoto coefficient of FP2 fingerprints/functional group vectors of a pair of drugs as their FP2 fingerprint/functional group similarity score (Dunkel *et al.*, 2008).

Similar to the functional group vector, the target profile of a drug is also defined as a binary vector denoted by $\mathbf{x}=(x_1, x_2, \ldots, x_K)$. Each dimension of the vector represents a protein, and its value is set to 1 if the protein is targeted by the drug, and otherwise to 0. Here drug-target relationships were extracted from DrugBank database (downloaded on March 24, 2013) (Law *et al.*, 2014). In total, we obtained drug-target relationships between 4139 small molecule drugs and 1924 human genes, and thus we defined the target profile as a 1924-dimension vector (i.e. $K=1924$). The cosine correlation coefficient was used to measure the target profile similarity of two drugs **x** and **y**, which is defined as

$$s(\mathbf{x},\mathbf{y})=\frac{\sum_{i=1}^{K}x_i y_i}{\sqrt{\sum_{i=1}^{K}{x_i}^2}\sqrt{\sum_{i=1}^{K}{y_i}^2}} \quad (1)$$

The side-effect profile of a drug is defined in the similar way. Side-effect information of drugs was downloaded from SIDER database (released on October 17, 2012) (Kuhn *et al.*, 2010). The processed drug-side effect relationship dataset involved 3209 side-effects represented by MedDRA preferred terms (Brown *et al.*, 1999), and therefore here we defined side-effect profile as a 3209-dimension vector. The side-effect profile similarity score between drug **x** and **y** was calculated by weighted cosine correlation coefficient:

$$s(\mathbf{x},\mathbf{y})=\frac{\sum_{i=1}^{K}w_i x_i y_i}{\sqrt{\sum_{i=1}^{K}w_i{x_i}^2}\sqrt{\sum_{i=1}^{K}w_i{y_i}^2}} \quad (2)$$

where w_i is the weight function for the ith side-effect in the side-effect profile. w_i is defined as

$$w_i=e^{(-{d_i}^2/\sigma^2 h^2)} \quad (3)$$

where d_i is the frequency of the ith side-effect in the dataset, i.e. the number of drugs having this side-effect in the dataset, σ is the mean of d_i and h is a parameter set to 1 as Takarabe *et al.* did (Takarabe *et al.*, 2013).

Gene expression profile similarity scores between 1144 drugs were directly obtained from Cheng *et al.* (2013). These scores were calculated based on gene expression profiles in response to drug treatment downloaded from Connectivity Map (Michnick, 2006), using the Batch DMSO Control data pre-processing method and the Xtreme cosine similarity score (with 100 probes) to measure the similarity (Cheng *et al.*, 2013).

The text mining scores of chemical-chemical associations were downloaded from STITCH database (v3.1), which were computed based both on co-occurrence in the literature and on natural language processing (Kuhn *et al.*, 2008).

2.3 Minimum redundancy maximum relevance feature selection

We used minimum redundancy maximum relevance (mRMR) method for feature selection (Ding and Peng, 2005; He *et al.*, 2010; Liu *et al.*, 2013; Peng *et al.*, 2005). In mRMR, each feature is ranked based on not only its relevance to the classification variable but also its redundancy with other features. In our work, both the relevance and the redundancy were measured using mutual information (MI). MI denoted by I between two discrete random variables X and Y is computed based on their joint probabilistic distribution $p(x, y)$ and the respective marginal probability distributions $p(x)$ and $p(y)$ (Ding and Peng, 2005; Gray, 1990):

$$I(X,Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (4)$$

MI quantifies the mutual dependence between two random variables, i.e. it measures how much knowing one variable reduces uncertainty about the other (Gray, 1990; Liu *et al.*, 2013).

Suppose S and T are respectively the already-selected and to-be-selected feature sets, and $|S|$ and $|T|$ are the sizes of S and T. mRMR firstly puts the feature which is the most relevant to the classification variable into S, and then moves the remaining features within T into S one by one, requiring each time the selected feature f_j to optimize the following condition (He *et al.*, 2010; Liu *et al.*, 2013):

$$\max_{f_j\in T}\left(I(f_j,c)-\frac{1}{|S|}\sum_{f_i\in S}I(f_j,f_i)\right) \quad (5)$$

where $I(f_j, c)$ is the relevance of feature f_j to the classification variable c and $^{1}/_{|S|}\sum I_{f_i\in S}(fj,fi)$ is its redundancy with features in S.

We downloaded the mRMR program from http://penglab.janelia.org/proj/mRMR/.

2.4 LR model

Suppose c is the classification variable and $f_1, f_2, \ldots, f_n$ are n features. For a pair of drug-ATC code, when the drug is known to belong to this ATC class, $c=1$, and otherwise $c=0$. The LR model used to predict whether or not a drug belongs to an ATC code based on n features is represented as

$$P(c=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 f_1+\beta_2 f_2+\ldots\ldots+\beta_n f_n)}} \quad (6)$$

where β_0 and $\beta_1, \beta_2, \ldots, \beta_n$ are respectively constant and regression coefficients of n features, and $P(c=1)$ denoted by *probability_score* in this paper represents the possibility that a drug belongs to an ATC class (Lin *et al.*, 2004; Liu *et al.*, 2013). In this work, SPSS software was used to train the parameters of the LR model based on the golden standard dataset (SPSS, Inc., 1999).

2.5 Performance assessment

The receiver operating characteristic (ROC) curve was used to show the performance of the constructed prediction model, which plots

the model's sensitivity and specificity at varying cutoffs using 1-specificity as *x*-axis and sensitivity as *y*-axis. Sensitivity (TP/*T*) and specificity (1−FP/*F*) are respectively referred to as correctly classified fraction in the positive test set (*T*) and in the negative test set (*F*), and TP (true positive) and FP (false positive) are respectively the predicted positives in *T* and *F*. The ROC curve can show the tradeoff between sensitivity and specificity. A good dichotomous classifier has a ROC curve climbing rapidly towards the top left corner of the graph, which can be quantified by the area under the curve (AUC). The closer the AUC is to 1.0, the better the performance of the classifier is. A perfect classifier has an AUC of 1.0, and AUC = 0.5 corresponds to a non-informative prediction (Lin *et al.*, 2004; Liu *et al.*, 2013). We used SPSS software to draw ROC curves and compute corresponding AUCs (SPSS, Inc., 1999).

We used 10-fold cross-validation and the independent test set to evaluate the performance of prediction model. In the 10-fold cross-validation, drug-code pairs in the GSP and GSN sets are randomly averagely divided into 10 subsets, and 9 subsets are combined to be used as the training set to train the prediction model and the remaining a subset is used as the test set. This process is repeated in turn 10 times, and finally 10 test sets are combined to plot the ROC curve (Li *et al.*, 2008; Liu *et al.*, 2013). For the evaluation by the independent test set, we use the whole golden standard dataset as the training set to train the prediction model, and use the independent test set to plot the ROC curve.

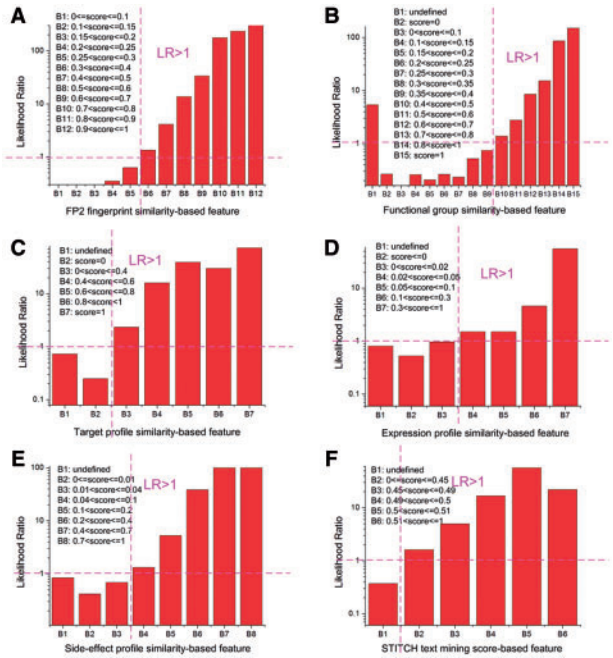

Fig. 1. LRs of six similairy-based features for the ATC code of level 4 based on 10-fold cross-validation scheme. In the 10-fold cross-validation scheme, the golden standard dataset is randomly and averagely divided into 10 subsets, and we use 9 subsets as the training set and the remaining one as the test set. Feature values of drug-ATC code pairs in the test set are identified based on the training set. This process is repeated 10 times, and finally the 10 test sets are combined to compute the LR. In each subgraph, the feature is divided into different bins. The feature value is 'undefined', because for this pair of drug-ATC code, all similarity scores between the drug and those known to belong to this ATC code cannot be computed. The reason why the similarity score cannot be computed is that the target, expression or side-effect profile of the drug is unknown, that the text mining score is not provided by STITCH (Kuhn *et al.*, 2008) or that the tanimoto coefficient cannot be calculated resulting from two '0' functional group vectors of two drugs. **(A)** FP2 fingerprint similarity-based feature. Actually the LR of B12 is infinite $\left({}^{384/1672}/_{0/1791}\right)$. **(B)** Functional group similarity-based feature. **(C)** Target profile similarity-based feature. **(D)** Expression profile similarity-based feature. **(E)** Side-effect profile similarity-based feature. Both LRs of B7 and B8 are infinite $\left({}^{89/1672}/_{0/1791}\right.$ for B7 and $\left.{}^{77/1672}/_{0/1791}\right.$ for B8$\left.\right)$. (F) STITCH text mining score-based feature

3 Results and Discussion

3.1 Features used to predict drug-ATC code associations

To develop a prediction model for ATC code assignment of drugs, first features are comprehensively collected, which can effectively discriminate true drug-ATC code associations from others.

We use the 'similarity idea' to predict drugs' ATC codes, which is commonly used in drug-related studies (Campillos *et al.*, 2008; Dunkel *et al.*, 2008; Gottlieb *et al.*, 2011). Our basic hypothesis is that if two drugs are similar in a certain aspect such as chemical structure, target protein or cellular transcriptional response etc., they might share the same therapeutic, pharmacological or chemical properties and thus belong to the same ATC class. Therefore firstly drug similarity scores are comprehensively collected. In total we collect six drug similarity scores based on five data sources and several data sources previously have been successively used to predict ATC classification (Chen *et al.*, 2012, 2014; Dunkel *et al.*, 2008; Wang *et al.*, 2013; Wu *et al.*, 2013). FP2 fingerprint and functional group similarity scores capturing structural characteristics from different perspectives are both used to measure chemical structure similarity. Another three are separately the similarity of target profiles, gene expression profiles in response to drug treatment and side-effect profiles. And the last one is the text mining score of the chemical–chemical association, whose underlying idea is that interactive compounds tend to share similar functions and thus might belong to the same ATC class (Chen *et al.*, 2014). We prove that all these six scores can be used to predict whether or not two drugs share an ATC code, and by and large the higher the score between two drugs is, the larger the possibility that they belong to the same ATC class is (Supplementary Fig. S1).

For a pair of drug-ATC code, we define the feature value of a certain similarity score-based feature as the largest one among the similarity scores between this drug and those drugs known to belong to this ATC code. We use likelihood ratio (LR) to measure the ability of the six similarity-based features to predict a drug belonging to an ATC code, which is referred to as the ratio of the probability of feature *f* observed in the GSP set to that in the GSN set. Generally LR > 1 means the feature has the prediction ability (Liu *et al.*, 2013). We find that all these six features can be effectively used to discriminate true and false drug-ATC code associations. And by and large, the larger the feature value is, the stronger its discrimination ability is (Fig. 1).

3.2 The prediction model integrating multiple features

We use LR model to integrate these effective features to construct the drug-ATC code association prediction model.

Based on LR model, we first build a single-feature prediction model for each feature. The results of ROC curves based on 10-fold cross-validation indicate that all these single-feature prediction models are effective to predict drug-ATC code associations, which is consistent with the conclusion of Figure 1. And these features have different prediction ability, among which chemical structure-based features have relatively strong ability, followed by target profile and chemical association score-based features, and the performance of

side-effect and expression profile-based features is relatively poor (Supplementary Fig. S2).

Further we use LR model to establish the prediction model integrating multiple features. All these six features have been proved to be effective for ATC code prediction, however apparently there is some redundancy between them (Supplementary Table S2). It has been recognized that integrating all individually effective features into the model does not necessarily lead to the best prediction performance. The redundancy between features often has no positive impact and even sometimes introduces negative impact on the efficiency of the prediction model, simultaneously leading to the decreased generalization of model and resource waste (Liu *et al.*, 2013; Peng *et al.*, 2005). Therefore we need feature selection before establishing the *n*-features model. Here we use mRMR feature selection method (Ding and Peng, 2005), which ranks features based on not only each feature's prediction ability but also the redundancy between pairwise features (see Section 2.3). We observe that features ranked at the bottom by mRMR are either those with low relevance (e.g. side-effect and expression) or those having relatively large redundancy with the features ranked in front (Table 1). For example, the feature of 'functional group' has large relevance (Supplementary Fig. S2), but it is ranked at the bottom because of its high redundancy with the first feature 'FP2' (Supplementary Table S2).

According to the feature order ranked by mRMR (Table 1), starting from the most effective FP2 single-feature prediction model (denoted by Model_1), features are integrated into the prediction model one by one, and ultimately the model includes all six features (Model_6). Based on the ROC AUC, first we see that the performance of the prediction model integrating multiple features is better than that of any single-feature model. Further we observe that as features are gradually added into the model, its efficacy is continuously improved. The Model_6 integrating all 6 features achieves the best performance (Fig. 2). These results indicate that it is necessary to combine multiple features for more effective ATC code prediction. Although there is some redundancy between the features we collect (Supplementary Table S2), from Figure 2, we see that the positive impact of their complementation is stronger than the potential negative impact of their redundancy on the prediction efficacy. Here we use the most powerful Model_6 integrating all six features as our ultimate prediction model for drug-ATC code association prediction, whose ROC AUC for ATC code of level4 based on 10-fold cross-validation reaches 0.9469, displaying its excellent prediction ability (Fig. 2).

Table 1. Features ranked by mRMR feature selection method based on the golden standard dataset of level 4.

Order	Feature[a]
1	FP2
2	STITCH score
3	Target
4	Expression
5	Functional group
6	Side-effect

[a]For all these features, 10-fold cross-validation scheme is used. The golden standard set is averagely and randomly divided into 10 subsets, and 9 subsets are combined as the training set and the remaining 1 subset as the test set. The six-feature values of drug-code pairs in the test set are identified based on the training set. The process is repeated 10 times, and finally the 10 test sets are combined to be used for feature selection.

Meanwhile the Model_6 also obtains good performance for ATC code prediction of other levels (Supplementary Fig. S3).

Besides 10-fold cross-validation, Model_6 also achieves good performance on the independent test set whose drug-code pairs are independent of the GSP and GSN sets (Table 2). To further discriminate Model_6's ability on the prediction of ATC codes of unclassified drugs (whose ATC codes are unknown) and that of new ATC codes of classified drugs (whose ATC codes are known in part), we divide the independent test set into two subsets (see Section 2.1). The 'independent test set-drug' is used to simulate the first condition, whose drug space is independent of that of the golden standard set, and the 'independent test set-code' whose drug space is totally included in that of the golden standard set but drug-code pairs are independent, simulating the prediction of new ATC codes of classified drugs. We observe that the performance of Model_6 is good on these two independent test subsets, indicating Model_6 can be well qualified for the prediction of not only ATC codes of unclassified drugs but also new ATC codes of classified drugs (Table 2). Supplementary Table S3 presents seven concrete examples in the independent test set. In these examples, as more available features are used for prediction, the probability scores of positive drug-code pairs gradually increase, while those of negative drug-code pairs gradually decrease, indicating that the more features we used, the larger the possibility that the drug-code relationship is correctly determined. In addition, we find that Model_6 also achieves good prediction performance for different types of drugs including hormones, natural products, synthetic molecules and salts

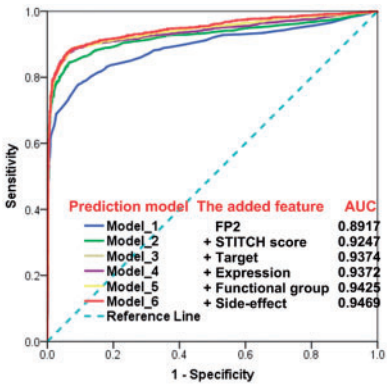

Fig. 2. ROC curves and their corresponding AUCs of prediction models integrating *n* features based on the golden standard dataset of level 4 using 10-fold cross-validation. Starting from the FP2 single-feature prediction model (Model_1), according to the order ranked by mRMR (Table 1), features are added into the prediction model one by one, and finally all six features are integrated into the model (Model_6). Different models are distinguished by different colors in the figure

Table 2. ROC AUCs of Model_6 based on the independent test set

ATC code level	ROC AUC of Model_6 based on the independent test set		
	Independent test set[a]	Independent test set-drug[a]	Independent test set-code[a]
1	0.8201	0.8223	0.7445
2	0.8642	0.8669	0.8394
3	0.8784	0.8816	0.8880
4	0.9033	0.9042	0.9148

[a]See Section 2.1 for the definitions of these three independent test sets.

(Supplementary Table S4) and different ATC classification branches of level 1 and level 2 (Supplementary Figs S4 and S5), based on corresponding subsets of the independent test set.

3.3 Comparison with previous methods

To the best of our knowledge, there are six previous methods developed specially for drug-ATC code association prediction (Chen *et al.*, 2012, 2014; Gurulingappa *et al.*, 2009; Dunkel *et al.*, 2008; Wang *et al.*, 2013; Wu *et al.*, 2013). Among these methods, NetPredATC method has been proved to be more effective than SuperPred (Dunkel *et al.*, 2008; Wang *et al.*, 2013), and the hybrid method recently proposed by Chen *et al.* also obtained better performance than their previous method (Chen *et al.*, 2012, 2014). Gurulingappa *et al.* (2009) validated their method only on ATC codes of level 3 in 'cardiovascular system' class and considered ATC

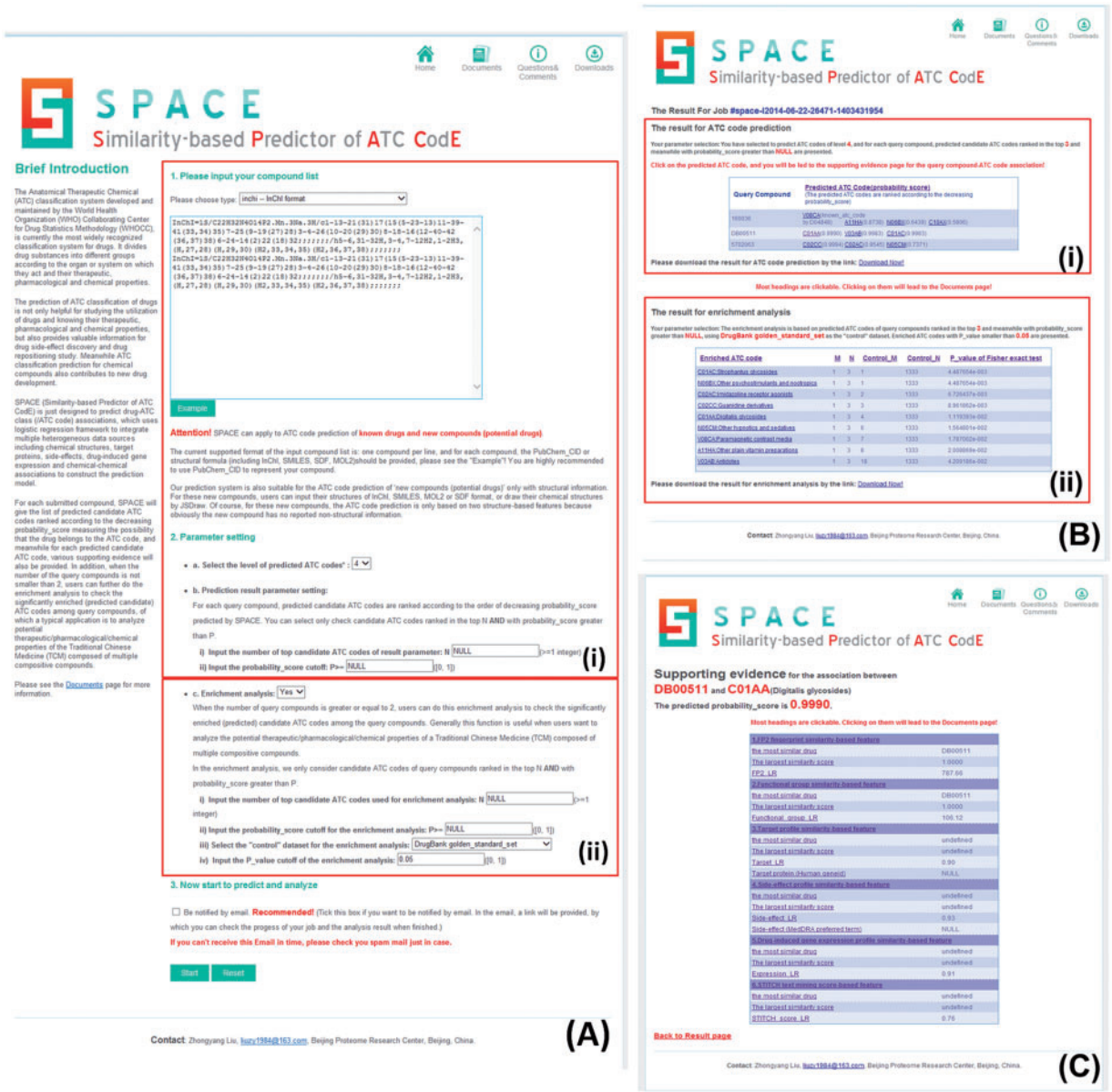

Fig. 3. Web service *SPACE* for ATC code prediction. **(A)** Homepage of *SPACE*. (i) By this page, users may submit the interested compound (list), and specify the level of ATC code prediction and result parameters. *SPACE* supports three input ways: Pubchem_CIDs of the interested compounds; Chemical structures of four commonly-used formats including SMILES, InChI, MOL2 and SDF; chemical structures drew by users through JSDraw. (ii) When the number of query compounds is not smaller than 2, users may also select the enrichment analysis to check the significantly enriched predicted candidate ATC codes among query compounds. **(B)** Result page. (i) For each submitted compound, *SPACE* will give its candidate ATC codes of user-specified level ranked according to the decreasing *probability_score*. When the query compound is a 'classified' drug (i.e. the compound belongs to the combined dataset used to train Model_6 underling *SPACE*, and thus its ATC codes are at least partly known), its known ATC codes are always listed first. For each predicted candidate ATC code, clicking on it will lead to the supporting evidence page. (ii) When the enrichment analysis is selected on the homepage, significantly enriched candidate ATC codes among query compounds are also given. The enrichment analysis is based on Fisher exact test, using the GSP set or the combined positive dataset as the 'control' group. **(C)** Supporting evidence page. This page gives supporting evidence for the predicted drug-ATC code association. For each piece of evidence, the most similar drug known to belong to the ATC code in the combined dataset to the query compound, the corresponding the largest similarity score (i.e. feature value) and LR are presented. Generally speaking, LR $>$ 1 means that this feature can act as a piece of evidence to support this drug-ATC code association

code prediction as a multi-label classification problem, which leads to difficult comparison with our method. More importantly, when an ATC class does not have sufficient drug instances, this ATC class cannot be predicted by multi-label classification model (Gurulingappa *et al.*, 2009). Meanwhile their features used for ATC code prediction were extracted from literature, which is also inconvenient in practical use of the method. Therefore, here we only compare Model_6 with three previous methods including NetPredATC (Wang *et al.*, 2013), iSEA (Wu *et al.*, 2013) and the hybrid method (Chen *et al.*, 2014). The comparison scheme is that we use the standard dataset and evaluation method previously used to assess the NetPredATC/iSEA/hybrid method to assess Model_6, and then compares the evaluation result with those of previous methods stated in their articles.

NetPredATC presented by Wang *et al.* (2013) used SVM to combine chemical structures and target proteins to infer drug-ATC code associations, using 10-fold cross-validation as evaluation scheme. Because only the evaluation result for ATC code of level 5 was given in their article, we reproduce and re-evaluate NetPredATC method on level 4 using data and programs provided by Wang *et al.*, and then compare the assessment result with that of Model_6 of level 4 evaluated based on the same dataset. On each of 4 drug datasets previously used to validate NetPredATC, ROC AUC of Model_6 is significantly larger than that of NetPredATC, indicating that Model_6 is more effective (Supplementary Table S5).

The iSEA (indication similarity ensemble approach) published in 2013 measured a pair of drug-ATC code association based on the chemical fingerprint similarity between the drug and the drug set belonging to the ATC code. Compared with the result in Table 1 of iSEA's article (Wu *et al.*, 2013), the performance of Model_6 is apparently better both on 'approved drugs' dataset and on 'external data set' (Supplementary Table S6).

Finally, we compare Model_6 with the recently published hybrid method, which integrated ChEBI ontology, STITCH chemical-chemical associations and chemical structures to predict drug-ATC code relationships (Chen *et al.*, 2014). For each drug, 14 candidate ATC codes of level 1 are ranked according to the order of decreasing scores given by the prediction model. Compared with the result in Table 2 of Chen *et al.*'s article, the prediction accuracies of first/second/third candidate ATC code predicted by Model_6 are apparently higher on the 'training dataset', 'internal validation dataset' and 'external validation dataset', also showing better performance of our Model_6 model (Supplementary Table S7).

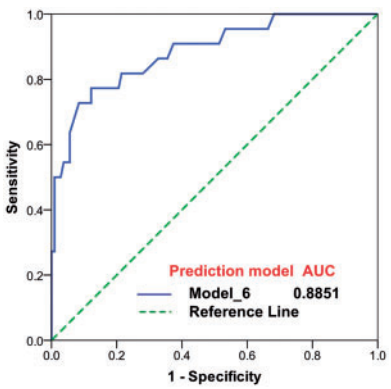

Fig. 4. The ROC curve and corresponding AUC of Model_6 underlying *SPACE* on drug-ATC code associations (denoted by P_set) composed of drugs and their newly added ATC code annotations of level 4 by latest DrugBank database (version: January 14, 2014) (relative to the combined dataset used to construct Model_6 underlying *SPACE*) and those composed of these drugs and random ATC codes (excluding the combined dataset and P_set)

In summary, compared with previous methods, our model obtains the best prediction performance on ATC code prediction, which we think mainly attributes to the effective integration of multiple heterogeneous data sources. There is more than one data source describing drug similarity which can provide evidence support for two drugs sharing therapeutic, pharmacological or chemical properties. Integrating multiple pieces of evidence which are mutually complementary to some extent can reduce the false positive and false negative of the prediction.

3.4 Web service and its application

Finally, Model_6 is developed into a user-friendly web service—*SPACE* (www.bprc.ac.cn/space). To make the prediction coverage of ATC codes of *SPACE* as large as possible, here we combine the golden standard set and independent test set to make up the combined dataset to train the Model_6 model underlying the *SPACE* (Supplementary Table S1). For each submitted compound (Fig. 3Ai), *SPACE* will give its candidate ATC code list of user-specified level ranked according to the order of decreasing *probability_score* measuring the possibility that the drug belongs to the ATC code which is predicted by Model_6 (Fig. 3Bi), and meanwhile for each predicted candidate ATC code, various supporting evidence will also be provided (Fig. 3C). In addition, when the number of the query compounds is not smaller than 2, in *SPACE* users can further do the enrichment analysis to check the significantly enriched predicted candidate ATC codes among query compounds (Figs 3Aii and 3Bii), of which a typical application is to analyze potential therapeutic/pharmacological/chemical properties of the Traditional Chinese Medicine (TCM) composed of multiple compositive compounds.

First we apply *SPACE* to the drugs whose ATC classification annotations are newly added by latest DrugBank database (version: January 14, 2014) (Law *et al.*, 2014), relative to the combined dataset used to establish *SPACE*. In result, *SPACE* obtains good performance, which further indicates that *SPACE* can effectively discriminate true drug-ATC code associations from others (Fig. 4).

Further *SPACE* is respectively applied to DrugBank drugs which still have not any ATC classification information in DrugBank ('unclassified drugs') and those whose ATC codes are (partly) known in DrugBank ('classified drugs'). We find that many newly predicted drug-code associations with high *probability_scores* are supported by known knowledge. For example, Enprofylline (DB00824) currently still lacks ATC classification annotation in DrugBank and KEGG database. Its first candidate ATC code of level 3 predicted by *SPACE* is R03D (other systemic drugs for obstructive airway diseases) (*probability_score* = 1), and its first and second candidate ATC codes of level 4 are respectively R03DA (xanthines) (*probability_score* = 1) and C03BD (xanthine derivatives) (*probability_score* = 0.9999), which is totally consistent with its description in DrugBank (Law *et al.*, 2014) and Wikipedia (http://en.wikipedia.org/wiki/Enprofylline) that Enprofylline is a xanthine derivative used in the treatment of asthma and chronic obstructive pulmonary disease. This example further shows that *SPACE* can be effectively used for ATC code assignment of unclassified drugs. Another example is Terbutaline (DB00871), which is annotated as a beta-2 adrenoreceptor agonist used for obstructive airway diseases (R03AC and R03CC) in DrugBank (Law *et al.*, 2014). Its another well-known use in clinic is to inhibit uterine contractions to delay preterm labor (Gaudet *et al.*, 2012; Nanda *et al.*, 2002), which is also successfully predicted by *SPACE* (G02C: other gynecologicals and G02CA: sympathomimetics,

labour repressants, both with *probability_score* = 1). Meanwhile its side-effect on maternal heart (Gaudet *et al.*, 2012; Nanda *et al.*, 2002) is also captured by *SPACE* (C01C: cardiac stimulants excl. cardiac glycosides with *probability_score* = 1). This example further shows that *SPACE* can not only be effectively used for ATC assignment of unclassified drugs, but also provide valuable clues for drug repositioning and side-effect discovery.

Finally, we apply *SPACE* to the TCM Qishenkeli which was previously studied in our lab (Wang *et al.*, 2012). Qishenkeli is a widely used formula in routine treatment of coronary heart disease in China, and long-term clinical practice has proved its definite effect on improving heart function (Wang *et al.*, 2012). We use *SPACE* to predict potential ATC codes of Qishenkeli's compositive compounds previously collected by us (Wang *et al.*, 2012), and further analyze the significantly enriched ATC codes (based on predicted candidate ATC codes ranked in the top 5, using the GSP set as the 'control' set). We find that many significantly enriched ATC codes among compositive compounds are consistent with Qishenkeli's known functions (Wang *et al.*, 2012), including C05: vasoprotectives (*P*-value of Fisher exact test is 4.26e-36), C01: cardiac therapy ($P = 4.38e\text{-}35$), C04: peripheral vasodilators ($P = 7.89e\text{-}04$) and C02: antihypertensives ($P = 5.43e\text{-}03$), which indicates *SPACE* is also helpful for elucidating potential therapeutic/pharmacological/chemical properties of a multicomponent drug.

Acknowledgements

We thank Yong Wang, Yongcui Wang, Pankaj Agarwal, Jie Cheng, Kuochen Chou, Lei Chen and Michael Kuhn for kindly providing related data; Qin Huang, Jiyang Zhang and Qiaosheng Xie for fruitful discussion.

Funding

This work was supported by the National Key Technology R&D Program (2012BAI29B07); Program of International S&T Cooperation (2014DFB30020); and Chinese High Technology Research and Development (2012AA020201).

Conflict of Interest: none declared.

References

Brown,E.G. *et al.* (1999) The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, **20**, 109–117.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chen,L. *et al.* (2014) A hybrid method for prediction and repositioning of drug Anatomical Therapeutic Chemical classes. *Mol. Biosyst.*, **10**, 868–877.

Chen,L. *et al.* (2012) Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, **7**, e35254.

Cheng,J. *et al.* (2013) Evaluation of analytical methods for connectivity map data. *Pac. Symp. Biocomput.*, **2013**, 5–16.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.

Dunkel,M. *et al.* (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.

Gaudet,L.M. *et al.* (2012) Effectiveness of terbutaline pump for the prevention of preterm birth. A systematic review and meta-analysis. *PLoS One*, **7**, e31679.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application topersonalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Gray,R.M. (1990) Entropy and Information. In: *Entropy and Information Theory*. Springer, New York, pp. 21–55.

Gurulingappa,H. *et al.* (2009) Concept-based semi-automatic classification of drugs. *J. Chem. Inf. Model.*, **49**, 1986–1992.

Haider,N. (2010) Functionality pattern matching as an efficient complementary structure/reactionsearch tool: an open-source approach. *Molecules*, **15**, 5079–5092.

He,Z. *et al.* (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.

Heller,S. *et al.* (2013) InChI - the worldwide chemical structure identifier standard. *J. Cheminform.*, **5**, 7.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

Kuhn,M. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

Li,D. *et al.* (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol. Cell. Proteomics*, **7**, 1043–1052.

Lin,N. *et al.* (2004) Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**, 154.

Liu,Z. *et al.* (2013) Proteome-wide prediction of self-interacting proteins based on multiple properties. *Mol. Cell. Proteomics*, **12**, 1689–1700.

Michnick,S.W. (2006) The connectivity map. *Nat. Chem. Biol.*, **2**, 663–664.

Nanda,K. *et al.* (2002) Terbutaline pump maintenance therapy after threatened preterm labor forpreventing preterm birth. *Cochrane Database Syst. Rev.*, (4), CD003933.

O'Boyle,N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.

Peng, H. *et al.* (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

SPSS, Inc. (1999) *SPSS Base 10.0, User's Guide*. SPSS, Inc., Chicago, pp. 431–434.

Takarabe,M. *et al.* (2013) Drug target prediction using adverse event report systems: a pharmacogenomics approach. *Bioinformatics*, **28**, i611–i618.

Wang,Y. *et al.* (2012) Drug target prediction based on the herbs components: the study on themultitargets pharmacological mechanism of qishenkeli acting on the coronaryheart disease. *Evid. Based Complement. Alternat. Med.*, **2012**, 698531.

Wang,Y.C. *et al.* (2013) Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics*, **29**, 1317–1324.

Wu,L. *et al.* (2013) Relating anatomical therapeutic indications by the ensemble similarity of drug sets. *J. Chem. Inf. Model.*, **53**, 2154–2160.