BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 10 2013, pages 1317–1324
doi:10.1093/bioinformatics/btt158

Systems biology

Advance Access publication April 5, 2013

Network predicting drug's anatomical therapeutic chemical code

Yong-Cui Wang[1], Shi-Long Chen[1], Nai-Yang Deng[2] and Yong Wang[3,4,*]

[1]Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810001, China, [2]College of Science, China Agricultural University, Beijing 100083, China, [3]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China and [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Discovering drug's Anatomical Therapeutic Chemical (ATC) classification rules at molecular level is of vital importance to understand a vast majority of drugs action. However, few studies attempt to annotate drug's potential ATC-codes by computational approaches.

Results: Here, we introduce drug-target network to computationally predict drug's ATC-codes and propose a novel method named NetPredATC. Starting from the assumption that drugs with similar chemical structures or target proteins share common ATC-codes, our method, NetPredATC, aims to assign drug's potential ATC-codes by integrating chemical structures and target proteins. Specifically, we first construct a gold-standard positive dataset from drugs' ATC-code annotation databases. Then we characterize ATC-code and drug by their similarity profiles and define kernel function to correlate them. Finally, we use a kernel method, support vector machine, to automatically predict drug's ATC-codes. Our method was validated on four drug datasets with various target proteins, including enzymes, ion channels, G-protein couple receptors and nuclear receptors. We found that both drug's chemical structure and target protein are predictive, and target protein information has better accuracy. Further integrating these two data sources revealed more experimentally validated ATC-codes for drugs. We extensively compared our NetPredATC with SuperPred, which is a chemical similarity-only based method. Experimental results showed that our NetPredATC outperforms SuperPred not only in predictive coverage but also in accuracy. In addition, database search and functional annotation analysis support that our novel predictions are worthy of future experimental validation.

Conclusion: In conclusion, our new method, NetPredATC, can predict drug's ATC-codes more accurately by incorporating drug-target network and integrating data, which will promote drug mechanism understanding and drug repositioning and discovery.

Availability: NetPredATC is available at http://doc.aporc.org/wiki/NetPredATC.

Contact: ycwang@nwipb.cas.cn or ywang@amss.ac.cn

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

Received on April 27, 2012; revised on February 1, 2013; accepted on March 31, 2013

*To whom correspondence should be addressed.

1 INTRODUCTION

The Anatomical Therapeutic Chemical (ATC) classification system categorizes drug substances at different levels by their therapeutic properties, chemical properties, pharmacological properties and practical applications. This classification system is recommended by the World Health Organization (WHO), and drug's ATC-codes have been widely applied in almost all drug utilization studies (WHO, 2006). Specifically, ATC classification system can be used as a basic tool for drug utilization research. It also provides the presentation and comparison of drug consumption statistics at international level and will greatly facilitate the recent drug repositioning and drug combination studies. Though useful, mapping ATC-codes to drugs is challenging.

Recently, ATC-codes for some well-characterized drugs have been deposited in databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) Biomolecular Relations in Information Transmission and Expression (BRITE) (Kanehisa *et al.*, 2006) and DrugBank (Wishart *et al.*, 2008). These databases provide high quality expert curated data. However, they are in small scale, and the coverage is far from enough to serve practical usage. Even for some well-collected drug datasets, the ATC-code assignments for drugs are far from complete. For example, the dataset in Yamanishi *et al.* (2008) contains drugs with four different type target proteins including enzymes, ion channels (ICs), G-protein couple receptors (GPCRs) and nuclear receptors (NRs). These drugs all have manually curated target proteins from KEGG BRITE (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2004), SuperTarget (Günther *et al.*, 2008) and DrugBank (Wishart *et al.*, 2008). Even in this high-quality dataset, there are 102 drugs that do not have any ATC-codes in all 445 drugs targeting enzyme, 13 drugs that do not have any ATC-codes in all 210 drugs targeting IC, 23 drugs that do not have any ATC-codes in all 223 drugs targeting GPCR, and 4 drugs that do not have any ATC-codes in all 54 drugs targeting NR. The percentage of drugs without ATC codes varies from 10 to 25%.

The bottleneck is that current data collection procedure heavily relies on human curation and is not efficient. One way out is to learn the underlying ATC classification rules from the available high quality ATC-code annotations, and further automatically assign new ATC-codes to drugs by a computational predictor. This strategy will accelerate the functional characterization of drugs under the ATC classification systems, especially those barely characterized drugs. Importantly, it will greatly speed up the mechanism understanding of a vast majority of

drugs action and narrow down the gap between the medical indications and drug effects elucidation at molecular level (Dunkel *et al*., 2008).

However, few studies attempt to address this important problem. Dunkel *et al*. tackled this challenge by proposing a computational method to classify the given compounds into ATC classification system. Their method is based on the drug similarity in chemical structures and physicochemical properties (Dunkel *et al*., 2008). They also developed a useful web-server, which allows prognoses about the medical indication of novel compounds and to find new leads for known targets (Dunkel *et al*., 2008). Nevertheless, the chemical structure only describes the static state of drugs. Cells use proteins and small molecules (drugs, metabolites, or ligands) networks to dynamically coordinate multiple biological functions. For instance, single drug may possess different biological functions by targeting different proteins. Therefore, if the drug target information is integrated into the prediction, the performance improvement can be expected. In this article, we follow this idea to design a new predictive method. That is, we map ATC-codes to a given drug based not only on its chemical structure similarity with other compounds but also on its target proteins.

The commonly accepted assumption in drug discovery is that drugs with similar pharmacological or therapeutic properties usually share common functions (Wang *et al*., 2010; Yamanishi *et al*., 2008, 2010; Zhao and Li, 2010). Existing efforts demonstrated that chemical structure similarity is useful in classifying compounds into ATC classification system (Dunkel *et al*., 2008). Here, we note that drug's pharmacological or therapeutic similarity may due to the fact that they interact with common or similar target proteins. Thus, it is reasonable to assume that drugs similar in target proteins usually share common ATC-codes. Starting with this assumption, we propose a novel computational approach called NetPredATC to predict potential ATC-codes for drugs. Specifically, we first construct the drug and ATC-code interaction network based on the known drug ATC-code annotations. Then we characterize ATC-code and drug by their similarity profiles and define kernel function to correlate drug with ATC-code. Finally, we infer drug's ATC-codes by training a machine-learning model, i.e. support vector machines (SVMs). SVMs are motivated by statistical learning theory and have been successful on many different classification problems in bioinformatics (Scholkopf *et al*., 2004). Our contributions here are not only in incorporating drug targets information for the first time into the ATC-code prediction but also in designing a novel predictive model by data integration.

The performance of our method was validated on four classes of drug target proteins, including enzymes, ICs, GPCRs and NRs. We show that both chemical structure and target protein are predictive via cross-validation experiments and statistical evaluation. Moreover, target protein information is more powerful. By combining them, our method outperforms the chemical similarity-only based method, and more experimentally observed drug ATC-code annotations can be uncovered.

2 MATERIALS AND METHODS

We propose a novel computational algorithm, NetPredATC, to infer drug's ATC-codes by using drug-target network information. Our algorithm works in three phases (Fig. 1): (i) Formulating known drug's ATC annotations as a bipartite graph. We extracted the known drug's ATC annotations from KEGG BRITE (Kanehisa *et al*., 2006) and DrugBank (Wishart *et al*., 2008) databases. (ii) Extracting drug–drug and ATC-code–ATC-code similarity metrics. Drug similarity is derived from chemical structure and target protein information. ATC-code similarity profiles are calculated by a probabilistic-based model (Lin, 1998). (iii) Feeding the similarities among drugs and similarities among ATC-codes to kernel method and applying SVM-based classifier to predict drug's unknown ATC-codes.

2.1 Constructing drug and ATC-codes interaction network

In ATC system, drugs are divided into 14 main groups (first level), with one pharmacological/theraputic subgroup (second level). The third and forth levels are chemical/pharmacological/theraputic subgroups, and the fifth level is the chemical substance. The hierarchical structure of ATC-codes makes the prediction a hierarchical multi-label classification problem. Existing models for this problem are complicated and expensive in computational cost (Cai and Hofmann, 2004; Rousu *et al*., 2006). This, thus, greatly restricts the application scope of such methods. Here, we propose a low cost computational method by treating ATC-code prediction problem as a binary classification problem. Specifically, we construct drug and ATC-code interaction network based on available drug's ATC annotations, which are extracted from KEGG BRITE (Kanehisa

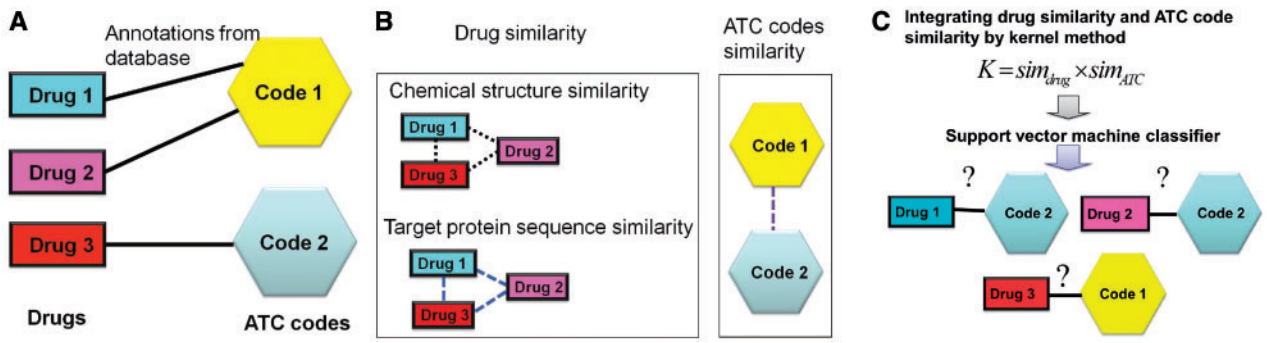

Fig. 1. The scheme of our ATC-code prediction approach for drugs. (**A**) Formulating known drug ATC-code annotations as a bipartite graph. (**B**) Extracting drug–drug and ATC-code–ATC-code similarity metrics. (**C**) Feeding the similarities among drugs and the similarities among ATC-codes to kernel method and applying SVM-based classifier to predict the unknown relationships between drugs and ATC-codes

et al., 2006) and DrugBank (Wishart *et al.*, 2008) databases. That is, by using the known ATC-codes for drugs, we construct a bipartite graph (Fig. 1A), i.e. the interactions only exist between drugs and ATC-codes. In this way drug's ATC-code prediction can be cast as a binary classification problem. We aim to determine whether a given drug and ATC-code pair interacts. The advantage is that we can use a much popular machine-learning method, SVM, to handle this high-dimensional learning problem in a relatively low cost way.

2.2 Collecting chemical structure and target protein data

Given two drug ATC-code pairs, we construct a kernel function, which correlates with their similarity. As kernel function represents the similarities among the training samples in some sense (Hofmann *et al.*, 2008), we focus on the similarity scores among drugs and similarity scores among ATC-codes. Therefore, we construct the similarity profiles to characterize drug and ATC-code in the following subsections.

2.2.1 Chemical structure data It is generally believed that drugs with similar chemical structures carry out common therapeutic function and thus likely share common ATC-codes. Therefore, each drug can be characterized by its chemical structure similarity profile with other drugs. The chemical structure similarity between two drugs d and d' is computed by SIMilar COMPound (SIMCOMP) algorithm (Hattori *et al.*, 2003), which is a graph-based method for comparing pairwise chemical structures. Suppose that we have n_c drugs in total, a matrix $S_{chem} \in R^{n_c \times n_c}$ is then constructed to represent chemical structure similarity. Each row (or column) of this matrix is chemical structure similarity profile for a single drug.

2.2.2 Target proteins Our previous drug-target prediction suggests that drugs with common protein targets often have similar therapeutic function (Wang *et al.*, 2010, 2011). Thus, drugs interacting with the same targets are likely to share common ATC-codes. Therefore, here, we fully take advantage of this feature to represent drugs by its similarity profile in target proteins with other drugs.

Given two drugs d and d', the similarity between them is calculated as follows:

$$sim(d, d') = \max_{g_i \in T(d), g_j \in T(d')} sim(g_j, g_j) \quad (1)$$

where $T(d)$ and $T(d')$ are the sets of target proteins. The sequence data are applied to measure protein similarity owing to the rapidly developed sequencing techniques. The sequence similarities among proteins are defined by a normalized version of Smith-Waterman scores (Smith and Waterman, 1981). Suppose that we have n_c drugs in total, a matrix $S_{inter} \in R^{n_c \times n_c}$ is then constructed to represent target protein similarity. Each row (or column) of this matrix is target protein similarity profile for a single drug.

2.2.3 ATC-code annotations KEGG BRITE (Kanehisa *et al.*, 2006) and DrugBank (Wishart *et al.*, 2008) databases deposited thousands of chemical compounds with detailed ATC annotations. Considering the hierarchical structure of ATC-codes, a probabilistic model (Lin, 1998) is introduced to calculate the similarity. Specifically, the similarity between two ATC-codes (t_i and t_j) is calculated as follows:

$$sim(t_i, t_j) = w(t_i)w(t_j)\exp(-\gamma d(t_i, t_j)) \quad (2)$$

where $d(t_i, t_j)$ is the shortest distance between ATC-codes t_i and t_j in the hierarchical structure of the ATC classification system, $w(t_i)$ and $w(t_j)$ represent the weights of the corresponding ATC-codes, and are defined as the inverse of ATC-code frequencies, which means that more emphasis was put on specific codes rather than the general ones (Yamanishi *et al.*, 2010). γ is a pre-defined parameter (set to be 0.25 in this study). S_{ATC} is used to denote the resulting drug therapeutic similarity matrix. Each row (or column) of this matrix is the similarity profile for a single ATC-code.

2.3 Kernel function for SVM-based predictor and data integration

With the representation of drugs and ATC-codes by their similarity profiles, the kernel function for two drug and ATC-code pairs $d_A t_A$ and $d_B t_B$ can be calculated as follows:

$$K(d_A t_A, d_B t_B) = S_{comp}(d_A, d_B) \times S_{ATC}(t_A, t_B) \quad (3)$$

where S_{comp} can be $S_{chem}(d_A, d_B)$, $S_{inter}(d_A, d_B)$ or their combination. In this article, Chem denotes the case when $S_{comp} = S_{chem}$, Inter denotes the case when $S_{comp} = S_{inter}$, and ChemInter denotes the case when $S_{comp} = \max(S_{chem}, S_{inter})$. Taken together, the rationale behind our kernel function for drug and ATC-code pairs is that two drug ATC-code pairs are similar only when the corresponding drugs and ATC-codes are simultaneously similar supported by different data sources.

2.4 Predicting drug and ATC-code interactions by the defined kernel function

With the aforementioned kernel function construction scheme, the ATC-code prediction task is formulated as a classification problem by feeding the kernel function to SVM. If we treat all drug ATC-code pairs with known interactions as the training positive samples, and others as the training negative samples, the training data imbalance problem will arise, as there are only a relatively small number of known drug and ATC-code interactions. This will make the SVM ineffective in determining the class boundary (Wu and Chang, 2003). To maintain a balance between positive and negative datasets in SVM training procedure, we randomly select a negative dataset from the unlabeled drug ATC-code pairs with almost the same size as the positive dataset.

As the kernel function (3) and training dataset are feeding to the SVM learning scheme, the predictor can be calculated by SVM algorithm.

2.5 Benchmark datasets and SVM implementation

The benchmark drug datasets, used to test the performance of our method, contain four types of target proteins, i.e. enzymes, ICs, GPCRs and NRs (Yamanishi *et al.*, 2008). The statistics for these four drug-target interaction networks are summarized in Yamanishi *et al.*, 2008.

Specifically, the numbers of drugs with known ATC-codes are 343, 197, 200 and 50, which interact with enzymes, ICs, GPCRs and NRs, respectively. The numbers of corresponding target proteins are 617, 201, 92 and 25 for enzymes, ICs, GPCRs and NRs, respectively. The numbers of drug-target interactions are 2280, 1422, 593 and 86 in four datasets, respectively. The numbers of corresponding drug–ATC-code interactions are 492, 281, 300 and 95 in four datasets (see Supplementary Table S1 for details).

We train the SVM-based predictor by using *LibSVM* (Chang and Lin, 2011). To evaluate the performance of our methods, we use 10-fold cross-validation. In our implementation, the penalty parameter C is optimized by the grid search approach with 3-fold cross-validation, and the optimal value of C is 1. The performance of our proposed method is shown by receiver operating characteristic (ROC) curve (Gribskov and Robinson, 1996), which illustrates the trade-off between the true positive (correctly predicted interactions) rate with respect to the false positive (wrongly predicted interactions) rate. Furthermore, the evaluation criteria, area under the ROC curve (AUC), accuracy (Acc) $= \frac{TP+TN}{TP+TN+FP+FN}$, sensitivity (Sn) $= \frac{TP}{TP+FN}$, specificity (Sp) $= \frac{TN}{TN+FP}$, precision (Pre) $= \frac{TP}{TP+FP}$ and F-measure $= \frac{2\times Sn\times Sp}{Sn+Sp}$, are used to assess the performance. Here, TP is the number of drug and ATC-code pairs correctly predicted to interact, FP is the number of drug and ATC-code pairs predicted to interact, but actually not. TN is the number of drug ATC code pairs that do not interact and predict correctly. FN is the number of drug and ATC-code pairs predicted not to interact but actually interact.

3 RESULTS

3.1 Proof-of-concept example for the motivation of NetPredATC

In this subsection, we explain the motivation of NetPredATC by a simple example. Drug *D*02070, annotated with 'Homatropine methylbromide', is an anticholinergic medication that inhibits muscarinic acetylcholine receptors and thus the parasympathetic nervous system. Drug *D*01297, annotated with 'Pirenzepine', is used in the treatment of peptic ulcers, as it reduces gastric acid secretion and reduces muscle spasm. It is in a class of drugs known as muscarinic receptor antagonists-acetylcholine, which serves the neurotransmitter of the parasympathetic nervous system to initiate the rest-and-digest state (as opposed to fight-or-flight). The chemical structures of *D*02070 and *D*01297 in 2D space are shown in Figure 2. The structure similarity between *D*02070 and *D*01297 based on SIMCOMP algorithm is ~0.25, which is relatively low (In the online chemical similarity search tool http://www.genome.jp/tools/simcomp/, the similar compound will be listed only when the SIMCOMP score is larger than 0.4). This indicates that two drugs are not similar in structure at all. The dissimilar structure thus fails the chemical structure-only based prediction algorithm. However, these two drugs share one common target: 'Muscarinic acetylcholine receptor M1' (HSA: 1128), which agrees with the therapeutic explanations of these two drugs. Moreover, these two drugs share a common ATC-code: *A*02*BX*03, annotated with 'pirenzepine'. This is the same as the annotation of drug *D*01297. Taken together, two drugs dissimilar in chemical structure share common ATC-code (*A*02*BX*03) owing to the fact that they interact with common target protein ('Muscarinic acetylcholine receptor M1'). This example demonstrates that drug-target interactions, introduced in our method for the first time, can help us to infer drug's ATC-codes. It is particularly useful when chemical structure information is insufficient. Furthermore, the improvement can be expected by combination of these two important information sources.

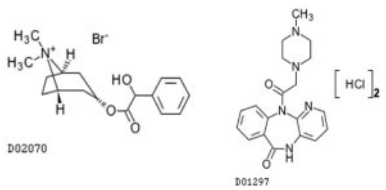

Fig. 2. The motivation of our NetPredATC is illustrated by a proof-of-concept example. Chemical structure similarity between *D*02070 and *D*01297 is 0.25, which is relatively low and indicates that two drugs are not similar in structure. However, these dissimilar drugs share the common protein target: Muscarinic acetylcholine receptor. Moreover, they share the common ATC code: *A*02*BX*03. Therefore, two drugs dissimilar in chemical structure share common ATC-code (*A*02*BX*03) owing to the evidence that they interact with common target protein ('Muscarinic acetylcholine receptor M1')

3.2 Correlation analysis shows the usefulness of chemical structure and drug-target interactions

We collect two data sources to depict drugs: chemical structure and target proteins. As the first step, we confirm each data source is predictive to ATC-codes, i.e. drugs with similar structures or target proteins tend to be annotated with similar ATC-codes. To show this, we correlate chemical structure similarity and target proteins similarity with ATC-code similarity, respectively. The ATC-codes similarity is calculated by the maximal ATC-code similarities for multiple annotations (Zhao and Li, 2010).

Pearson's correlation coefficients (PCCs) between chemical structure, drug-target network similarity and ATC-code similarity are shown in Figure 3. Besides all drug pairs, we also draw the PCCs between chemical structure similarity, target protein similarity and ATC-code similarity among drug pairs when their chemical structure similarity and target protein similarity larger than 0.5 and 0.8. Figure 3 demonstrates that, on all four datasets, PCCs are increasing when the drug similarities increase. Both PCCs from chemical structure similarity and target protein similarity reach their largest values when drug similarities are larger than 0.8. In addition, PCCs between chemical structure similarity and target protein similarity are larger than 0.4 for all the four datasets. The PCCs between target protein similarity and ATC-code similarity are consistently larger than that between-chemical structure similarity and ATC-code similarity. That is, both chemical structure similarity and target protein similarity correlate with the ATC-code similarity significantly, and, moreover, ATC-code similarity correlates better with target protein similarity.

In addition, we correlate the similarities obtained from two data sources with the topology parameters of the known drug and ATC-code interaction network. We define two drugs' distance in the network as the length of their shortest path in the drug and ATC-code interaction network. We plot the

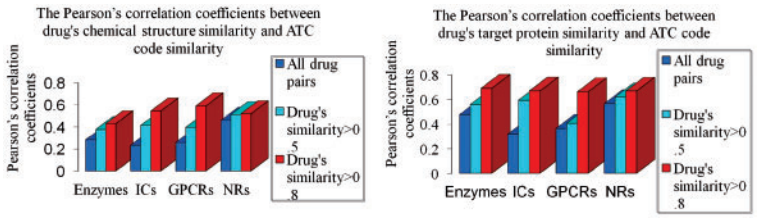

Fig. 3. Barplots of the PCCs of ATC-code similarity with the chemical structure similarity (the left subfigure) and target protein similarity (the right subfigure). The corresponding *P*-values are <1e-2 when the PCCs between chemical structure, target protein and ATC-code similarity are >0.8. Other *P*-values are near zeroes

distributions of chemical structure and target protein similarity scores with respect to drugs' distance in enzyme, ICs, GPCRs and NRs networks, respectively, in Supplementary Figure S1. As the similarity scores are close to zeroes when the corresponding distances are larger than 2 for all four datasets, we only show the similarity distributions with respect to 'distance 2' group in Supplementary Figure S1, which means that we focus on the drugs with common ATC-codes.

Supplementary Figure S1 shows that both chemical structure similarity and target protein similarity are larger than 0.8 for about ~75% drug pairs sharing common ATC-codes. That is, two drugs sharing common ATC-codes tend to have larger chemical structure and target protein similarities. It suggests that drug pairs with similar chemical structure, or target protein, tend to be annotated with the same ATC-code. Thus, both chemical structures and target proteins are predictive for ATC-code annotation. In addition, the number of drug pairs with high target protein similarity (>0.8) are more than the number of drug pairs with high chemical structure similarity (>0.8) when counting in all drug pairs with common ATC-code. This fact indicates that the drugs with common ATC-code tend to have target protein similarity than chemical structure similarity, and target protein information may have better coverage in ATC-code prediction.

3.3 Drug-target interactions are more predictable than chemical structures in ATC-code prediction

In this subsection, we compare the usefulness of chemical structures and target proteins in ATC-code prediction. The performance is evaluated and visualized by ROC curves (Gribskov and Robinson, 1996). We note that the training negative dataset is constructed only in the training process, and the validation is done with the assumption that all non-interacting drug ATC-code pairs are negative examples.

First, we show the performance of chemical structure and target protein in uncovering the experimentally observed drug and ATC-code interactions. We replace the drug similarity matrix S_{comp} in kernel function (3) with S_{chem} and S_{inter}, respectively. The ROC curves obtained by chemical structure and target proteins are drawn in Figure 4. The corresponding evaluation criteria, AUC, Acc, Sn, Sp and Pre are listed in Table 1 when the corresponding F-measure reaches its maximum.

From Figure 4 and Table 1, we can see that Chem obtains AUC score beyond 0.78 on all four datasets. That is, chemical structure is useful in ATC-code prediction. Moreover, target protein plays a more important role in predicting drug's ATC-codes. For all four datasets, Inter outperforms Chem with much higher Acc, Pre and F-measure and obtains ~3% improvement in accuracy. In addition, Inter improves Sn by >2%. This result demonstrates that more experimentally observed ATC-codes of drugs can be uncovered by applying target proteins to characterize drug similarity.

All these results suggest that each data source will do one's bit to infer drug's potential ATC-codes. Therefore, combination of these two data sources should produce a much more sophisticated picture.

3.4 Data integration improves prediction

In the previous subsection, each data source demonstrates its usefulness in uncovering the experimentally observed drug ATC-codes. In the following, we validate the effect of combining the two data sources in predicting drug's ATC-codes. The performance of data integration method, ChemInter, is evaluated and visualized by ROC curves (Fig. 4) and other criteria (Table 1).

Figure 4 shows that, except for NRs dataset, ChemInter obtains the highest true positive rate when false positive rate is <0.1. That is, ChemInter can achieve better accuracy when predicting a small fraction of drug ATC-code interactions.

Table 1 shows that ChemInter performs better than using single data source for all four datasets. For example in Enzymes dataset, Chem and Inter make the AUCs 0.805 and 0.836, respectively, whereas ChemInter obtains an AUC 0.841. For NRs dataset, Chem and Inter obtain Sns 0.715 and 0.736, respectively, whereas ChemInter obtains an Sn of 0.726, which is lower than Inter. However, Chem and Inter obtain Pres 0.701 and 0.723, respectively, whereas ChemInter obtains an Sn of 0.775, which is much higher than Inter and Chem. This fact again demonstrates that chemical structure and target protein similarity are useful in ATC-code prediction. Their combination obtains significant improvement.

3.5 Comparison with SuperPred

As we mentioned, Duntkel *et al.* proposed a computational method to assign drug's ATC-codes (Dunkel *et al.*, 2008). In their work, only drug chemical structure was used underlying the basic assumption that similar compounds belong to the same ATC-group (Dunkel *et al.*, 2008). To calculate the similarity between two drugs, their structural fingerprints, generated by Chemistry Development ToolKit (CDK) (http://almost.cubic.uni-koeln.de/cdk/), were used, and the similarity was determined by the Tanimoto coefficient. It is similar with SIMCOMP

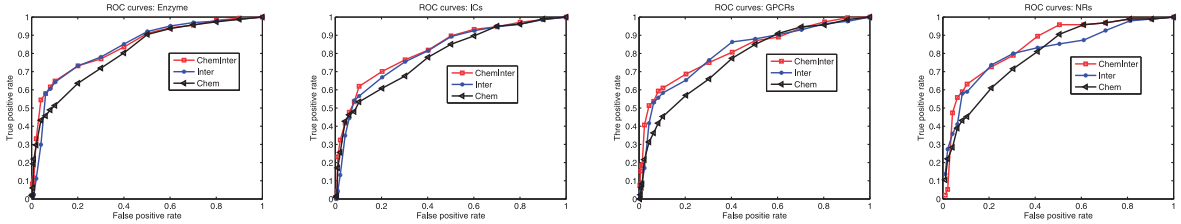

Fig. 4. ROC curves for the methods using different data sources to predict ATC-codes for drug with enzymes, ICs, GPCRs and NRs targets. Chem denotes the case when only drug chemical structure data are used, Inter denotes the case when only drug-target network data are used and ChemInter denotes the case when the two data sources are integrated

Table 1. Performance comparison of different data sources to predict drug's ATC-codes

Dataset	Methods	AUC	Acc	Sn	Sp	Pre	F-measure
Enzymes	Chem	0.805	0.709	0.719	0.699	0.705	0.709
	Inter	0.836	0.765	0.731	0.799	0.783	0.763
	ChemInter	**0.841**	**0.765**	**0.731**	**0.799**	**0.784**	**0.764**
ICs	Chem	0.783	0.702	0.608	0.797	0.750	0.690
	Inter	0.806	0.730	0.669	0.797	0.767	0.727
	ChemInter	**0.816**	**0.740**	**0.686**	**0.797**	**0.771**	**0.737**
GPCRs	Chem	0.765	0.678	0.660	0.696	0.685	0.677
	Inter	0.808	0.730	**0.763**	0.696	0.715	0.728
	ChemInter	**0.816**	**0.741**	0.686	**0.796**	**0.771**	**0.737**
NRs	Chem	0.792	0.705	0.715	0.694	0.701	0.705
	Inter	0.811	0.747	**0.736**	0.694	0.723	0.747
	ChemInter	**0.844**	**0.757**	0.726	**0.789**	**0.775**	**0.756**

The best predictions are highlighted in bold.

algorithm (Hattori *et al.*, 2003) used in our implementation, which is a graph-based method for comparing chemical structures.

Our NetPredATC is conceptually different with SuperPred by introducing drug-target network in ATC-code prediction. In addition, we performed a side-by-side comparison with SuperPred on predictive results. We found that our method NetPredATC outperforms SuperPred in terms of accuracy by considering drug target information. We took the GPCRs dataset as an example. Specifically, we submitted the drug's name in SuperPred server (http://bioinformatics.charite.de/superpred), got the corresponding ATC-codes and then calculated the predictive accuracy. As a result, 215 drug ATC-code annotations have been predicted correctly among the total 300 drug ATC-code predictions. The accuracy of SuperPred is 0.716. Our method obtains an accuracy of 0.73 by using protein target information alone. When integrating the chemical similarity with target information, our method further improves the accuracy to 0.74. The results again show that target proteins play an important role in drug's ATC-code prediction. This provides additional supports for the efficiency of NetPredATC.

Our accuracy with chemical structure information alone is 0.678 in Table 1, which is slightly smaller than SuperPred's 0.716. This difference arises from the different ways to calculate drug's chemical similarity. CDK is better than SIMCOMP owing to the fact that physiochemical properties are included.

Next, we list some examples to highlight our advantage by introducing drug target information. We specifically focus on the successful predictions by NetPredATC while failed by SuprePred. By introducing the target information, NetPredATC correctly identifies 219 drug ATC-codes, and four of them cannot be predicted by SuperPred. Specifically, drug D05740 is annotated by ATC-code N02CC04, drug D00480 is annotated by ATC-code D04AA10, drug D02884 is annotated by ATC-code R03DA10 and drug D00059 is annotated by ATC-code N04BA04. We check the data and find that these ATC annotations are borrowed from the drugs with similar target proteins but dissimilar chemical structures, similar to the example illustrated in Figure 2. These predictions demonstrate that our implementation necessitates a clear benefit in replacing SuperPred.

3.6 Novel predictions

We find that NetPredATC displays its excellent performance in discovering experimentally observed ATC-codes of drugs. To test whether it can produce biologically useful predictions, we focus on the unlabeled drug ATC-code pairs. We trained NetPredATC on the gold-standard positive dataset and randomly selected negative dataset from the unlabeled pairs and tested it on the remaining drug ATC-code pairs. Our expectation is that NetPredATC can discover the missing drugs ATC-codes. The top five predicted interactions on Enzyme, ICs, GPCRs and NRs datasets are listed in Table 2 and Supplementary Tables S2–S4, respectively. For each drug and ATC-code pair in these tables, we check out their annotation information from DrugBank (Wishart *et al.*, 2008) and World Health Organization Collaborating Centres (WHOCC) (http://www.whocc.no/atc_ddd_index/) databases. We further check the drug and its ATC-code annotations form Wikipedia (http://en.wikipedia.org/wiki/Main_Page) and finally analyzed the reliability of predicted ATC-codes.

For Enzyme dataset, *D*00969 is annotated as 'Meloxicam' and is a non-steroidal anti-inflammatory drug (NSAID) with analgesic and fever reducer effects. *M*01*AB*08, annotated as 'Etodolac', belongs to a class of drugs called NSAIDs. Moreover, the target of *D*00969 is prostaglandin G/H synthase 2 (HSA:5743), which is also the target of *D*00315, and *D*00315 has the ATC-code *M*01*AB*08. Therefore, the relationship between *D*00969 and *M*01*AB*08 may exist with high probability. The descriptions of the remaining four novel predictions are presented in the Supplementary Material.

The top five novel predictions on ICs, GPCRs and NRs datasets are listed in Supplementary Table S2–S4, the explanation of corresponding predictions is presented in the Supplementary Material. Database search, literature search and functional annotation analysis support these novel predictions. All these results

Table 2. The top five drug ATC-code predictions by our NetPredATC method for Enzymes dataset

Rank	Drug and ATC-code pair	Annotation
1	D00969	Meloxicam
	M01AB08	Etodolac
2	D01119	Temocapril hydrochloride
	C09AA13	Moexipril
3	D01582	Acemetacin
	M01AE04	Fenoprofen
4	D00968	Fenoprofen calcium
	M01AB14	Proglumetacin
5	D00463	Oxaprozin
	S01BC05	Ketorolac

suggest that NetPredATC can uncover potential ATC-codes for drugs. It can provide low-resolution predictive results for further high-resolution biological experiments.

4 DISCUSSIONS AND CONCLUSION

In this article, we propose a new computational method, NetPredATC, to infer drug's ATC-codes by integrating its chemical structure and target protein data. Our main contributions are both in characterizing the drug similarity profiles by drug-target network and in constructing data integration model by kernel method.

Specifically, we characterize the drug similarity not only form chemical structures but also from target proteins. By treating ATC-code prediction as a binary classification problem, a SVM-based predictor is used to uncover unknown ATC-codes for drugs. The improvement is achieved by incorporating target protein information on four benchmark datasets. Our method outperforms the existing chemical structure-only based method and can accurately uncover more experimentally observed drug ATC-codes. In addition, the database search and functional annotation analysis support our novel predictions. Taken together, these rigorous validations imply that our method can identify drugs' potential ATC-codes in an accurate way.

We improved the predictive performance by characterizing drugs' target protein sequence similarity. The improvement is robust to the definition of protein sequence similarity. We tried different cut-offs to measure protein sequence similarity (0.2, 0.3, 0.4 and 0.5). The results are summarized in Supplementary Table S5 and Supplementary Figure S2. We found that the AUC score is slightly lower when using a more stringent cut-off, but not too much. This is because most of sequence similarity among drug target is actually low in our datasets to avoid obvious predictions. One advantage to introduce target protein information is to fully use the indirect neighbor information in drug ATC-code annotation network. It allows us to predict drug's ATC-code when this drug has low chemical similarity and target similarity with its closest drug. We list some novel drug ATC-code predictions with low chemical similarity and target similarity in Supplementary Table S6.

We further note that there are another ways to define drug similarity by their protein targets. For example, drugs sharing common targets are often similar in their side effects (Campillos *et al*., 2008) and drugs targeting the same neighbor in network show similarity in side effects (Brouwers *et al*., 2011). Therefore, the closeness of target proteins in protein–protein interaction (PPI) network appears to be a good predictor for drug side effects. Similar to the side effects prediction, drug pairs that have similar ATC-codes may target the same neighbor in PPI network. Thus, it is necessary to validate the effect of targets neighborhood in ATC-code prediction. That is, we can characterize drugs by their targets closeness in PPIs network instead of sequence-based similarity in future.

The training negative dataset is one key problem in SVM-based predictor. There is still plenty room for the improvement on the definition and selection of the training negative dataset. This is a formidable challenge to our method and to other interaction prediction methods. Since the available ATC-code annotations for drugs is far from complete, many unknown drug and ATC-code pairs may be actually interacting in our task. To deal with this problem, a linear regression model can be introduced to uncover the new ATC-codes of drugs, which can avoid the selection of negative dataset. Specifically, the chemical structures and target proteins are used to characterize the drug similarity profiles, and then a linear regression model can be applied to correlate the ATC-code similarity with drug similarity. The similar ideas have been used to prioritize the disease genes (Jiang *et al*., 2011; Wu *et al*., 2008). In addition, our NetPredATC was validated on four relatively small benchmark datasets. The larger the dataset is, the more knowledge can be learned to construct classifier, and results are generally more reliable.

Importantly, we provide a general framework to perform data integration, and our framework is ready to be applied to another drug-related studies. For example, our method can be applied in side effects prediction, which is a critical stage in drug development. One can assess the similarity among the side effects by some text mining approaches and define the drug similarity by their chemical structures, target proteins or ATC-codes (Zhao and Li, 2010). In almost similar manner, our method can predict novel drug–disease relationships, which is also known as drug repositioning.

ACKNOWLEDGEMENTS

The authors thank the editor and reviewers for their suggestions to improve our manuscript.

Funding: National Natural Science Foundation of China (No. 11201470, No. 31270270, No. 61171007, and No. 11131009).

Conflict of Interest: none declared.

REFERENCES

Brouwers,L. *et al.* (2011) Network neighbors of drug targets contribute to drug side-effect similarity. *PLoS ONE*, **6**, e22187.

Cai,L.J. and Hofmann,T. (2004) Hierarchical document categorization with support vector machines. In: *Proceedingsof the 13th ACM International Conference on Information and Knowledge Management*. Washington, DC.

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

Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**:27:1–27:27.

Dunkel,M. *et al.* (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.
Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.
Günther,S. *et al.* (2008) Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.
Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
Hofmann,T. *et al.* (2008) Kernel methods in machine learning. *Ann. Stat.*, **36**, 1171–1220.
Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**, D354–D357.
Lin,D. (1998) An information-theoretic definition of similarity. In: Shavlik,J.W. and Shavlik,J.W. (eds) *ICML 98: Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, pp. 296–304.
Rousu,J. *et al.* (2006) Kernel-based learning of hierarchical multilabel classification models. *J. Mach. Learn. Res.*, **7**, 1601–1626.
Schomburg,I. *et al.* (2004) Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
Scholkopf,B. *et al.* (2004) *Support Vector Machine Applications in Computational Biology*. MIT Press, Cambridge, MA.
Smith,T.F. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Wang,Y.C. *et al.* (2010) Computationally probing drug-protein interactions via support vector machine. *Lett. Drug Des. Discov.*, **7**, 370–378.
Wang,Y.C. *et al.* (2011) Kernel based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.*, **35**, 353–362.
WHO Expert Committee. (2006) The selection and use of essential medicines. Report of the WHO expert committee, 2005 (including the 14th model list of essential medicines). *World Health Organ. Tech. Rep, Ser.*, 1–119, back cover.
Wishart,D.S. *et al.* (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
Wu,G. and Chang,E.Y. (2003) Class-boundary alignment for imbalanced dataset learning. In: *Workshop on Learning from Imbalanced Datasets II*. ICML, Washington DC.
Wu,X. *et al.* (2008) Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189–199.
Yamanishi,Y. *et al.* (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.
Yamanishi,Y. *et al.* (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**, i246–i254.
Zhao,S. and Li,S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS ONE*, **5**, e11764.