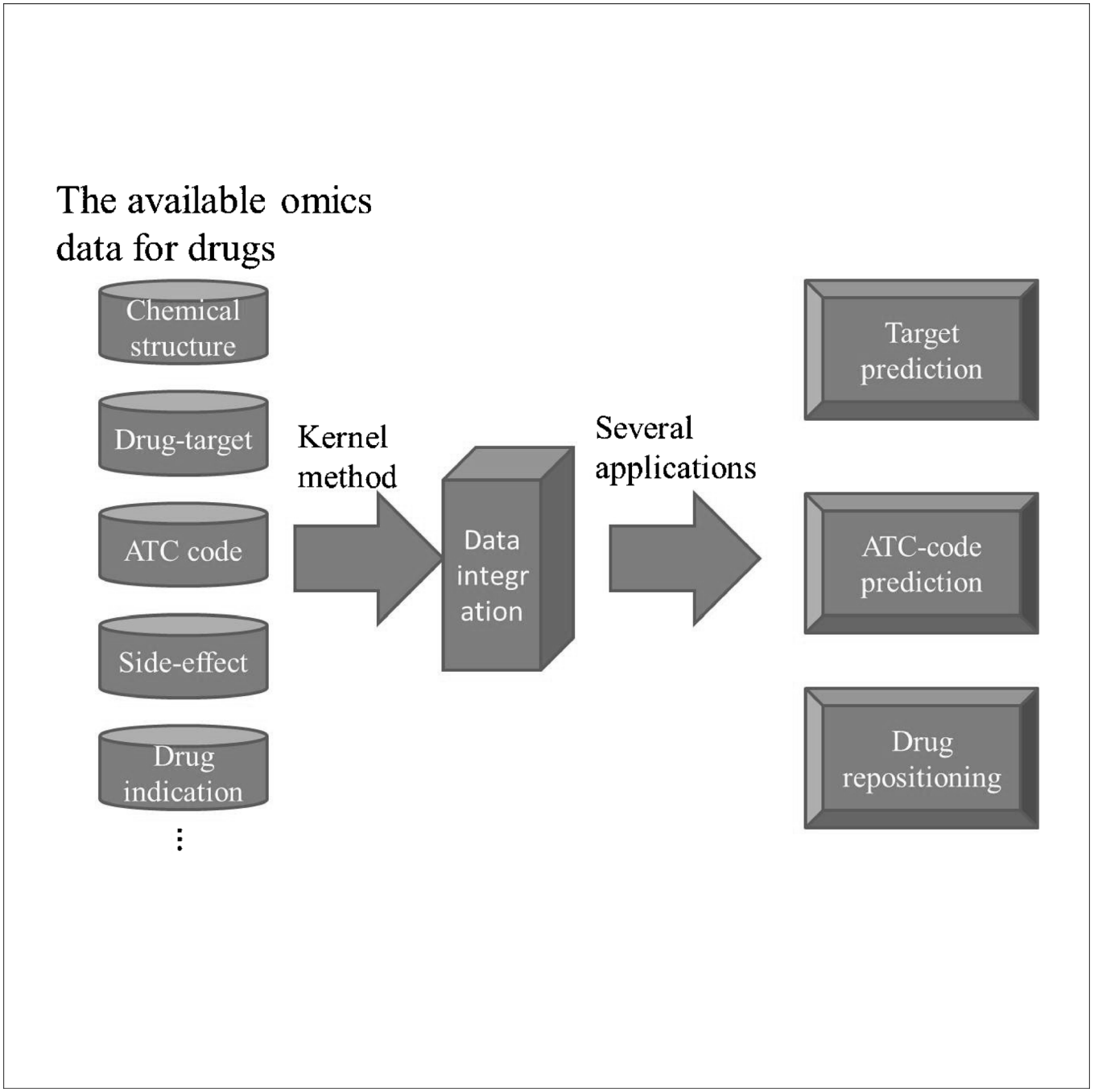
Review

DOI: 10.1002/minf.201300090

Computational Study of Drugs by Integrating Omics Data with Kernel Methods

Yongcui C. Wang,[a] Naiyang Deng,[b] Shilong Chen,*[a] and Yong Wang*[c, d]



molecular informatics

Abstract: With the rapid development of genomic and chemogenomic techniques, many omics data sources for drugs have been publicly available. These data sources illustrate drug's biological function in the living cell from different levels and different aspects. One straightforward idea is to learn understandable rules via computational models and algorithms to mine and integrate these data sources. Here, we review our recent efforts on developing kernel-based methods to integrate drug related omics data sources. Three promising applications of our framework are shown to predict drug targets, assign drug's ATC-code annotation, and reveal drug repositioning. We demonstrate that data integration does provide more information and improve the accuracy by recovering more experimentally observed target proteins, ATC-codes, and drug repositioning. Importantly, data integration can indicate novel predictions which are supported by database search and functional annotation analysis and worthy of further experimental validation. In conclusion, kernel methods can efficiently integrate heterogeneous data sources to computationally study drugs, and will promote the further research in drug discovery in a low-cost way.

Keywords: Omics data · Kernel methods · Data integration · Drug-targets · ATC-codes of drugs · Drug repositioning

1 Introduction

With the rapid development of genomic and chemogenomic projects, many omics data sources for drugs are publicly available. For example, PubChem database at NCBI deposits millions of chemical compounds with structure information;[1] Japan Pharmaceutical Information Center (JAPIC) database curates more than ten hundreds of keywords to describe the pharmacological information of compounds;[2] The Anatomical Therapeutic Chemical (ATC) classification system categorizes drug substances by their therapeutic and chemical characteristics;[3] The KEGG BRITE,[4] DrugBank,[5] BRENDA,[6]and SuperTarget[7] deposit high-quality drug-target interactions; Until Dec. 2012, there are a total of 4,192 side-effects in the SIDER database from the chemical structures of 996 approved drugs;[8] From DrugBank,[5] and Online Mendelian Inheritance in Man (OMIM),[9] the abundant drug–disease associations can be obtained.

Above data sources illustrate drug's biological function in the living cell from different levels and different aspects. For example, chemical structure provides information by the 'structure determines function' paradigm; Compound's JAPIC annotation and ATC-code describe drug effect from molecular function level; Target protein provides the direct effect at molecular activity level; Drugs' side-effects and their indications hint the unwanted and desired effects at phenotype level.

Computational drug study aims to learn understandable rules for drug effects from these data sources. Each data source is important in some ways and will contribute in different ways in understanding the mechanisms of action of drugs. Therefore, fusion of multiple data sources from different levels should produce a much more sophisticated picture of the effects of drugs. One popular strategy is to explore this fused representation by kernel-based statistical learning methods,[10] which have the capacity of evade the sample representation problem in the high-dimensional space. The trick is to apply a kernel function to replace the inner-products of samples in the high-dimensional space and to facilitate the construction and analysis of leaning algorithm.[11] Kernel-based statistical learning methods have been proven as very useful tools in bioinformatics.[12] One big advantage is that kernel-based methods provide a principled framework in which many types of data sources can be integrated. For example, Noble et al. applied kernel-based method to integrate heterogeneous data sources, such as protein domain, protein–protein interactions (PPIs), and gene expression data, to infer protein functions.[13] They also proposed kernel-based methods to predict PPIs by incorporating protein sequences and GO annotations information.[14] Those observations motivate us that applying kernel-based method to elaborate drug effects by integrating multiple omics data sources.

In the process of drug development, elucidating drug's targets, potential ATC-codes, and possible disease connections are fundamental challenges. Identification of drug–target interactions is a key area in genomic drug discovery. Kuhn et al. reviewed some attempts to apply large-scale computational analyses to predict novel interactions between drugs and targets from molecular and cellular features.[15] Meanwhile, ATC classification system provides the presentation and comparison of drug consumption statistics at international level (see Report of the WHO Expert

[a] *Y. C. Wang, S. Chen*
Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences
No. 23, Xinning Road, Xining, Qinghai Province, P. R. China
**e-mail: slchen@nwipb.cas.cn*

[b] *N. Deng*
College of Science, China Agriculture University
No. 17. Qinghua East Road, Beijing, P. R. China

[c] *Y. Wang*
National Centre for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
N0.55, Zhongguancun East Road, Beijing, P. R. China
**e-mail: ywang@amss.ac.cn*

[d] *Y. Wang*
Molecular Profiling Research Center for Drug Discovery, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

Committee, 2005), and some efforts have been made to study the ATC-classification system by computational methods.[16,17] What's more, inferring new disease treatments for existing drugs (drug repositioning) offers the possibility of faster, safer, low-risk, and low-cost drug development. Dudley et al. classified recent computational predictions for drug repositioning in two axes: drug based, where discovery initiates from the chemical perspective, or disease based, where discovery initiates from the clinical perspective of disease or its pathology.[18]

The key step to address above challenges is to develop methods that can correlate and integrate the information across drugs, proteins, ATC-codes, and diseases from multiple omics data sources. In this review, we survey our recent efforts in a heterogeneous data integration framework by developing kernel-based methods to uncover drug-targets, drug ATC-codes, and drug-disease connections. Roughly, our framework consists of three steps. First we characterize drug, protein, ATC-code, and disease by their similarity-based profiles, and define the kernel function to correlate drug with target protein, ATC-code, and disease, respectively. Secondly, we train a machine learning model, support vector machine (SVM), to automatically predict novel drug-target, drug ATC-code, drug-disease interactions. Finally, we validate our method by cross-validation on well-established datasets. We will make sure that each single data source is predictive in drug-target, drug ATC-code, and drug-disease interactions prediction. Moreover, by combination of multiple properties, more experimentally observed interactions can be uncovered. Database search and functional annotation analysis indicate that our new predictions are worthy of future experimental validation.

Yong Wang is Associate Professor at the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS). He received his Ph.D. degree in Operations Research and Control Theory from AMSS of CAS in 2005, his Master's Degree in Operations Research and Control Theory from the Dalian University of Technology in 2002 and his Bachelor's Degree in Mathematics and Physics from the Inner Mongolia University in 1999. His current interest is Bioinformatics and Systems Biology.

Nai-Yang Deng received his B.Sc. and M.Sc. degrees from the Department of Mathematics and Mechanics of the Peking University, China, in 1962 and 1966, respectively. He joined the Department of Science College of China Agriculture University as a professor in 1990. He is a Part-time Professor of Shanghai University since 1994. He has wide research interests, mainly including computational methods for optimization, operation research, support vector machine in data mining and bioinformatics.

Shilong Chen is Deputy Director of the Northwest Plateau Institute of Biology, Professor of the Chinese Academy of Science, and member of the Flora of China Editor Committee. In July 1997 he got his Ph.D. degree from the Institute of Botany, Chinese Academy of Sciences. He received the Suquan Fang Award and the Diao Award from the Chinese Academy of Science in 1995. He is second-prize winner of the National Natural Science Awards in 2004 and received the Youth Scientific and Technological Reward of the China Society on Tibetan Plateau in 2005.

Yongcui Wang is Associate Professor at the Northwest Institute of Plateau Biology, Chinese Academy of Science (CAS). She received her B.S. from the Academy of Mathematics and Systems Science at Shandong University in 2005 and her Ph.D. from the College of Science, China Agriculture University in 2010.

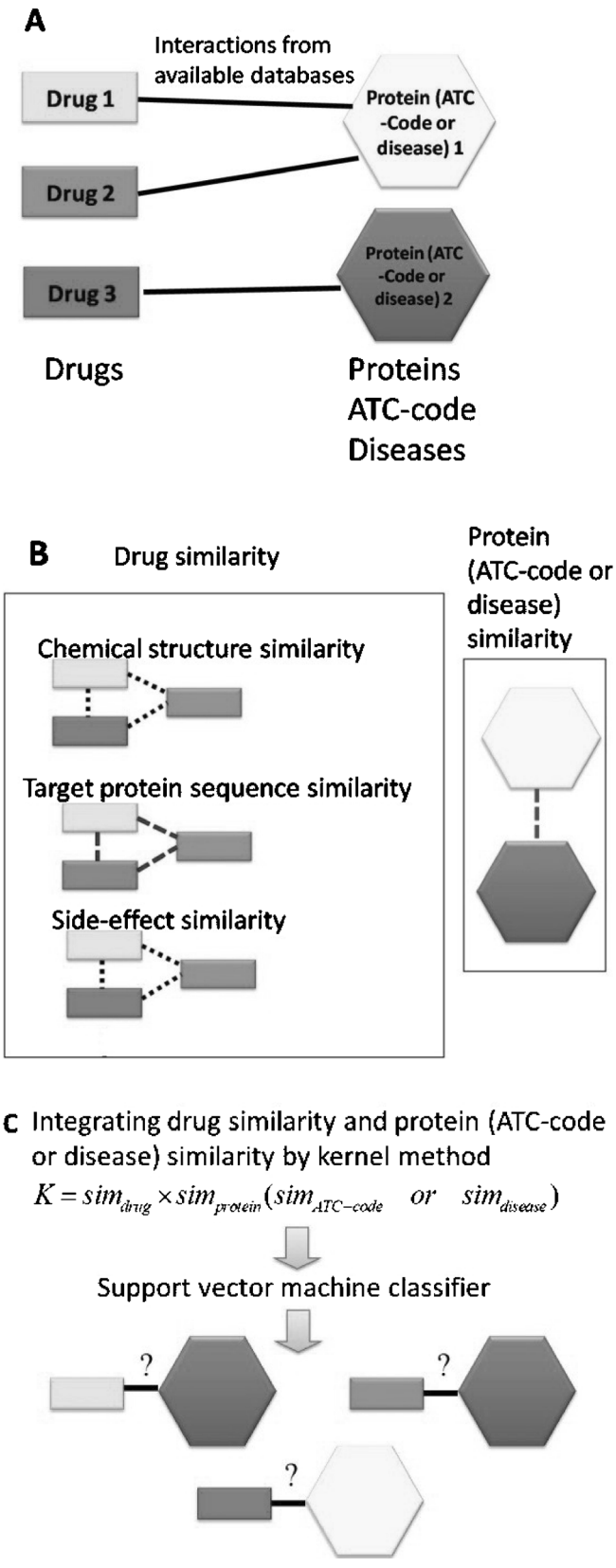
2 Methods Overview

2.1 Data Representation

Inferring potential target proteins, ATC-codes, and disease associations for drugs can be all treated as binary classification problem, i.e., to predict whether a given pair of drug-protein, drug ATC-code, or drug-disease interacts or not. We apply kernel method to integrate multiple omics data sources and introduce SVM-based algorithm to cope with these prediction tasks. The methodology works in three phases (Figure 1): (A) Collecting all known drug-target interactions (drug ATC-code or drug-disease interactions) as gold-standard positives in a bipartite graph. (B) Creating drug-drug and protein-protein (ATC-code and ATC-code or disease-disease) similarity metrics based on multiple data sources. (C) Relating the similarity among drugs and similarity among proteins (similarity scores among ATC-codes or similarity scores among diseases) by kernel methods, and apply SVM-based algorithm to predict unknown relationships between drugs and proteins (relationships between drug and ATC-codes or relationships between drug and diseases).

To implement the SVM-based algorithm, the kernel function and standard training dataset should be prepared. The kernel function represents the similarities among the training samples in some sense.[11] Given two drug-protein pairs (drug ATC-code or drug-disease pairs), we consider to construct a kernel function to reflect their similarity. Naturally

molecular informatics



we consider the similarity scores among drugs, the similarity scores among proteins (similarity scores among ATC-codes or diseases), and then integrate the two scores together. Therefore, we will introduce the ideas to extract the similarity profile from the available omics data in the following subsections.

2.1.1 Drug Similarity from Molecular Structure, Molecular Function, Molecular Activity, and Phenotype Data

Molecular structure Data

Under the umbrella that molecular structure determines function, it is generally believed that drugs with similar chemical structures carry out common therapeutic function, thus likely share common target proteins (ATC-codes or diseases). So here, each drug is first characterized by its chemical structure similarity profile with other drugs.

The chemical structure similarity between two drugs d and d′ is calculated in the following two ways. First, it is calculated by SIMCOMP algorithm,[19] which is a graph-based method for comparing pairwise chemical structures. Second, it is calculated by weighted cosine correlation of drugs' substructure profiles. Specifically, a given drug is firstly represented by a binary vector x, each element of x is encoded as 1 or 0, means the presence or absence of corresponding PubChem substructure,[20] and then the similarity between drug d and d′ is computed by their weighted cosine correlation coefficient.[21] There are a total of 881 chemical substructures in PubChem database,[1] thus each drugs' substructure profile has 881 elements.

Suppose that we have n_c drugs in total, a $n_c \times n_c$ matrix $\boldsymbol{S}_{chem}$ is then constructed to represent chemical structure similarity. The i-th row of this matrix is the chemical structure similarity profile for the i-th drug.

Molecular Function Data with Pharmacological Information as Representative

One abundant information source for drug is the chemical pharmacological properties and annotations. For example, JAPIC database[2] contains more than ten hundreds of unique keywords to describe the compounds pharmacological information. If there are totally K unique keywords to annotate the compounds, each compound can be then

Figure 1. The flowchart of our kernel-based prediction algorithm. (A) Collecting known interactions between drugs and proteins (drugs and ATC-codes or drugs and diseases) as gold-standard positives in a bipartite graph. (B) Creating drug-drug and protein-protein (ATC-code and ATC-code or disease-disease) similarity metrics. (C) Relating the similarities among drugs and similarities among proteins (similarities among ATC-codes or similarities among diseases) by kernel function, and apply SVM-based algorithm to predict the unknown relationships between drugs and proteins (relationships between drugs and ATC-codes or relationships between drugs and diseases).

encoded as a K-dimensional binary vector y, with its element denoting whether the corresponding pharmacological keyword is used to annotate the given compound or not. The pharmacological similarity between drugs d and d′ is evaluated by their weighted cosine correlation coefficient.[21]

$\boldsymbol{S}_{\text{phar}}$ is then constructed to represent pharmacological similarity. The *i*-th row of this matrix is pharmacological similarity profile for the *i*-th drug.

Molecular Function Data with Therapeutic Information (ATC-Code) as Representative

Ontology is a systematic method to describe the properties of molecules. For example, gene ontology is important to depict the biological processes, molecular functions, and subcellular localization for genes. For drugs, the useful data source for annotations is ATC classification system. In the ATC system, each drug is annotated by some ATC-codes. According to the ATC codes, drugs are divided into fourteen main groups (1st level), and further into one pharmacological/therapeutic subgroup (2nd level). The 3rd and 4th levels are chemical/pharmacological/therapeutic subgroups and the 5th level is the chemical substance. The drug similarity in its therapeutic sense (ATC-code metric) can be then calculated as follows:

$$sim_{ther}(d_i, d') = \max_{t_i \in A(d), t_j \in A(d')} sim(t_i, t_j)$$

where $A(d)$ and $A(d')$ are the sets of ATC-codes annotating the corresponding drugs, $sim(t_i,t_j)$ is the similarity between ATC code t_i and t_j, which is calculated by a probabilistic model:[22]

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

where $sim(t_i,t_j)$ 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. This means that more emphasis is put on specific ATC-codes rather than the common ones.[21] γ is a predefined parameter (set to be 0.25 in this study).

$\boldsymbol{S}_{\text{ther}}$ is used to denote the resulting drug therapeutic similarity matrix. The *i*-th row of this matrix is therapeutic similarity profile for the *i*-th drug.

Phenotype Data with Side-Effect as Representative

Drug side-effect is high level phenotype data for drugs to indicate the malfunction by off-targets. These effects have been used to infer whether two drugs share a common target protein.[23] Furthermore, side-effects had been utilized to relate with drug repositioning.[24,25] Similarly, drug side-effects information can be applied to characterize the drug by the similar profile concept. Until 2010, in total there are 1,450 unique side-effects in the SIDER database[8] for 888 approved drugs. That is, each drug can be represented by a 1,450 dimensional binary vector, whose element is encoded as 1 or 0, means the presence or absence of corresponding side-effect, respectively. We then define the drug similarity under their side-effects metric as their weighted cosine correlation coefficient.[21]

The matrix $\boldsymbol{S}_{\text{side-effect}}$ is then constructed to represent the drug similarity matrix by their side-effects. The *i*-th row of this matrix is the side-effect similarity profile for the *i*-th drug.

Molecular Activity Data with Target Protein as Representative

Drugs perform their biological functions inside cell via their target proteins. High-quality drug-target interactions can be manually constructed from the KEGG BRITE,[4] DrugBank,[5] BRENDA,[6] and SuperTarget.[7] In addition, the interactions among therapeutic drugs and their targets are well-studied in the previous studies.[21,26–29] Therefore we introduce target proteins information deposited in DrugBank[5] and define the drug similarity by their targets. Given two drugs *d* and *d*′, their similarity can be calculated as follows.

$$sim_{inter}(d_i, d') = \max_{g_i \in T(d), g_j \in T(d')} sim(g_i, g_j)$$

where $T(d)$ and $T(d')$ are the sets of target proteins for drug *d* and *d*′, $sim(g_i,g_j)$ is the sequence similarities among the protein g_i and g_j defined by a normalized version of Smith–Waterman scores.[30]

The matrix $\boldsymbol{S}_{\text{inter}}$ is then constructed to represent drug similarity matrix in target protein sense. The *i*-th row of this matrix is the target protein similarity profile for the *i*-th drug.

2.1.2 Protein Similarity by Genomic Data

Due to the rapidly developed sequencing techniques to accumulate large-scale data, we use the amino acid sequence data to measure protein similarity. The sequence similarities among the proteins are defined by a normalized version of Smith-Waterman scores.[30] Suppose that we have n_g proteins in total, matrix $\boldsymbol{S}_{\text{geno}} \in R^{ng \times ng}$ represents the protein sequence similarity matrix. Each row (or column) of this matrix is the similarity profile for a single protein.

2.1.3 ATC-Code Similarity by Ontology Structure

As we mentioned above, a probabilistic model is introduced to calculate the pairwise similarity $sim(t_i,t_j)$ between two ATC-codes (t_i and t_j) by considering their weighted distance in the hierarchical structure of the ATC classification system.[22] As a result, $\boldsymbol{S}_{\text{ATC}}$ is used to denote the resulting

drug therapeutic similarity matrix. Suppose that we have n_A ATC codes, a matrix $\boldsymbol{S}_{ATC} \in R^{nA \times nA}$ is constructed. Each row (or column) of this matrix is the similarity profile for a single ATC-code.

2.1.4 Disease Similarity by Phenotypes

The disease-disease similarity measures are based on semantic similarity of disease phenotypes according to the text mining scheme in van Driel et al.,[31] where over 5000 human phenotypes in OMIM database are collected and classified to describe diseases. The phenotype similarity data are accessible through a web interface.[32] If there are totally P phenotypes to annotate the diseases in OMIM, each disease can be then encoded as a P-dimensional binary vector, with its element denoting whether the corresponding phenotype is proper to annotate the given disease or not. The similarity between two diseases can be evaluated by defining vector distance. Suppose that we have n_s diseases, a matrix $\boldsymbol{S}_{disease} \in R^{ns \times ns}$ is constructed. The matrix $\boldsymbol{S}_{disease}$, is then applied to represent the disease similarity matrix. Each row (or column) of this matrix is the phenotype similarity profile for a single disease.

2.2 The Pairwise Kernel to Integrated Data

With the representation of drugs, proteins, ATC-codes, and diseases by their similarity profiles, the kernel function with two drug-protein pair: $d_A g_A$ and $d_B g_B$, two drug ATC-code pair: $d_A t_A$ and $d_B t_B$, and two drug-disease pairs: $d_A D_A$ and $d_B D_B$ can be calculated as Kronecker product kernel:[14,33–35]

$$\boldsymbol{K}(d_A g_A,\, d_B g_B) = \boldsymbol{S}_{comb}(d_A, d_B) \times \boldsymbol{S}_{geno}(g_A, g_B)$$

$$\boldsymbol{K}(d_A t_A,\, d_B t_B) = \boldsymbol{S}_{comb}(d_A, d_B) \times \boldsymbol{S}_{ATC}(t_A, t_B)$$

$$\boldsymbol{K}(d_A D_A,\, d_B D_B) = \boldsymbol{S}_{comb}(d_A, d_B) \times \boldsymbol{S}_{disease}(D_A, D_B)$$

where $\boldsymbol{S}_{comb}$ can be any one of $\boldsymbol{S}_{chem}$, $\boldsymbol{S}_{phar}$, $\boldsymbol{S}_{ther}$, $\boldsymbol{S}_{inter}$, and $\boldsymbol{S}_{side\text{-}effect}$ or their combination. In this paper, we use "Chem" to denote the case when $\boldsymbol{S}_{comb} = \boldsymbol{S}_{chem}$, "Phar" denotes the case when $\boldsymbol{S}_{comb} = \boldsymbol{S}_{ther}$, "Ther" denotes the case when $\boldsymbol{S}_{comb} = \boldsymbol{S}_{ther}$, "Inter" denotes the case when $\boldsymbol{S}_{comb} = \boldsymbol{S}_{inter}$, "Side-effect" denotes the case when $\boldsymbol{S}_{comb} = \boldsymbol{S}_{side\text{-}effect}$ and "Comb" denotes the case when

$$\boldsymbol{S}_{comb} = \max\,(\boldsymbol{S}_{chem}, \boldsymbol{S}_{phar}, \boldsymbol{S}_{ther})$$

$$\boldsymbol{S}_{comb} = \max\,(\boldsymbol{S}_{chem}, \boldsymbol{S}_{inter})$$

$$\boldsymbol{S}_{comb} = \max\,(\boldsymbol{S}_{chem}, \boldsymbol{S}_{inter}, \boldsymbol{S}_{side\,effect})$$

in drug-target, drug ATC-code, and drug-disease prediction task, respectively, which requires drug similarity supported by one or more than one metrics.

Taken together, the rationale behind our kernel function construction scheme for drug-protein pairs (drug ATC-code or drug-disease pairs) is that two drug- protein pairs (drug ATC-code or drug-disease pairs) are similar only when the corresponding compound and protein (ATC-code or disease) are simultaneously similar supported by different kinds of data source. We note that not all the above matrices are positive semi-definite. To remedy this issue, we need to normalize them into kernel matrix as a pre-process step.[36]

2.3 SVM-Based Predictors for Drugs

With the above defined pairwise kernel construction scheme, the drug-protein interactions (drug ATC-code or drug-disease interactions) prediction task is ready to be formalized as a classification problem. We collected publicly available and known interacting drug-protein pairs (drug ATC-code or drug-disease pairs) as the positives and the others as the negatives. Then we simply feed the kernel function to SVM. One possible problem is the training data imbalance. Because only a relatively small number of drug-protein pairs (drug ATC-code or drug-disease pairs) is known to be interacted, compared to the large amount of unknown pairs. This situation will make SVM ineffective in determining the class boundary.[37] To maintain a balance between training positives and negatives in SVM training procedure, we usually randomly select a set of negatives from the unlabelled data (unknown drug-protein pairs (drug ATC-code or drug-disease pairs)) to make sure that it has the similar size with the training positives. With the well-defined kernel function and training dataset, we take SVM learning scheme to cope with prediction task and a score can be calculated by SVM algorithm for every drug-protein pair (drug ATC-code or drug-disease pair). Ranking all the drug-protein pairs (drug ATC-code or drug-disease pairs) by their scores, we can assess the predictive accuracy. Importantly, the well-trained model is ready to predict novel interactions.

2.4 Benchmark Datasets and SVM Implementation

Before the large-scale predictions, validation in a small and well-annotated dataset is necessary. The benchmark dataset to test the performance of predicting drug-target and drug ATC-code was summarized by Yamanishi et al.[27] This dataset is widely used as a community standard and contains four kinds of target proteins, i.e., enzymes, ion channels (ICs), G-protein couple receptors (GPCRs), and nuclear receptors (NRs).[27] The statistics for these drug-target interaction data and more information were summarized by Yamanishi et al.[27] The gold standard dataset used to test the performance of the prediction algorithm for drug repositioning was summarized by Gottlieb et al.[38] It spans 1,933 associations between 593 drugs taken from DrugBank[5] and 313 diseases listed in OMIM database.[9]

We trained the SVM-based predictor by using LibSVM.[39] In our implementation, the penalty parameter C was optimized by a grid search approach with 3-fold cross-validation, and the optimal value of *C* is 10, 1, 1 for drug-target prediction, drug ATC-code prediction, and drug-disease prediction, respectively. To evaluate the performance of our methods, the 10-fold cross-validation was applied. The performance was assessed by receiver operating characteristic (ROC) curve,[40] which shows the trade-off between the true positive (correctly predicted interactions) rate (TPR) with respect to the false positive (wrongly predicted interactions) rate (FPR). Furthermore, the evaluation criterions AUC (area under the ROC curve),

$$\text{Accuracy } (Acc) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

$$\text{Sensitivity } (Sn) = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity } (Sp) = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

$$\text{Precision } (Pre) = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

$$F\text{-measure} = \frac{2 \times \text{Sn} \times \text{Sp}}{\text{Sn} + \text{Sp}},$$

and are used to further assess the performance of the proposed predictive methods. Here TP is the number of drug-protein pairs (drug ATC-code or drug-disease pairs) correctly predicted to interact, while FP is the number of drug-protein pairs (drug ATC-code or drug-disease pairs) predicted to interact but actually not. And TN is the number of drug-protein pairs (drug ATC-code or drug-disease pairs) do not interact and predicted correctly, while FN is the number of drug-protein pairs (drug ATC-code or drug-disease pairs) predicted not to interact but actually interact.

3 Results

3.1 Correlation of the Training Data with each Data Source

We collected three data sources to provide descriptions for drugs in drug-target prediction: chemical structures, pharmacological and therapeutic information; two data sources in drug ATC-code prediction: chemical structure and target proteins; three data sources in drug repositioning prediction: chemical structure, target proteins and side-effects. As the first step, we want to make sure each data source is indeed predictive by simple correlation analysis, that is, drugs with similar structures, pharmacological, or therapeutic effect tend to interact with similar proteins; drugs with similar structures or target proteins tend to be annotated with similar ATC-codes; drugs with similar structures, target proteins or side-effects will cure similar diseases. To show these, we correlated the similarity obtained from different data sources with the topology of the known interaction network, respectively.

We defined the distance of two compounds in the network as the length of their shortest path in network. For concise, we just plotted the distributions of chemical structure, pharmacological, and therapeutic similarity scores with respect to network distance for drugs, respectively, in Enzyme dataset in Figure 2A (see the distribution in other three kinds of datasets in Figure 1 in Wang et al.[41]), and drew the distributions of chemical structure, and target protein similarity scores with respect to network distance in Enzyme dataset in Figure 2B (see the distribution in other three kinds of datasets in Figure S1 in Wang et al.[42]).

Figure 2A showed that, two drugs with higher chemical, pharmacological, or therapeutic similarities tend to have shorter network distance. It suggests that drug pairs with similar chemical structure, pharmacological, or therapeutic profile may interact with the same target protein. Figure 2B showed that 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 common ATC-codes. Both chemical structures and target proteins are predictive for ATC-codes annotation.

In drug repositioning study,[36] we also observed that all chemical structures, target proteins, and side-effects similarities are larger than 0.6 for about 75% drug pairs with common diseases. That is, two drugs with larger similarity scores in all three kinds of metrics tend to share common diseases. All these results confirm that each data source is predictive. In addition it shows that correlation analysis provides some insights and is necessary before data integration.

3.2 Drug-Target, Drug ATC-Code and Repositioning by Single Data Source

With the rough picture that each omics data source is useful in our prediction, we next quantitatively assess the predictive power for each data source in different prediction tasks: chemical structures, pharmacological and therapeutic information in drug-target prediction; chemical structures and target proteins in drug ATC-code prediction; chemical structures, target proteins and side-effects in drug repositioning prediction. The performance were evaluated and visualized by ROC curves[40] and some evaluation criteria.

Firstly, we showed the effect of each data source in uncovering the experimentally observed interactions by replacing the drug similarity matrix $\boldsymbol{S}_{\text{comb}}$ in pairwise kernel function with the similarity matrix defined by corresponding data source. For example, when revealing experimentally observed drug-disease interactions, we replaced drug

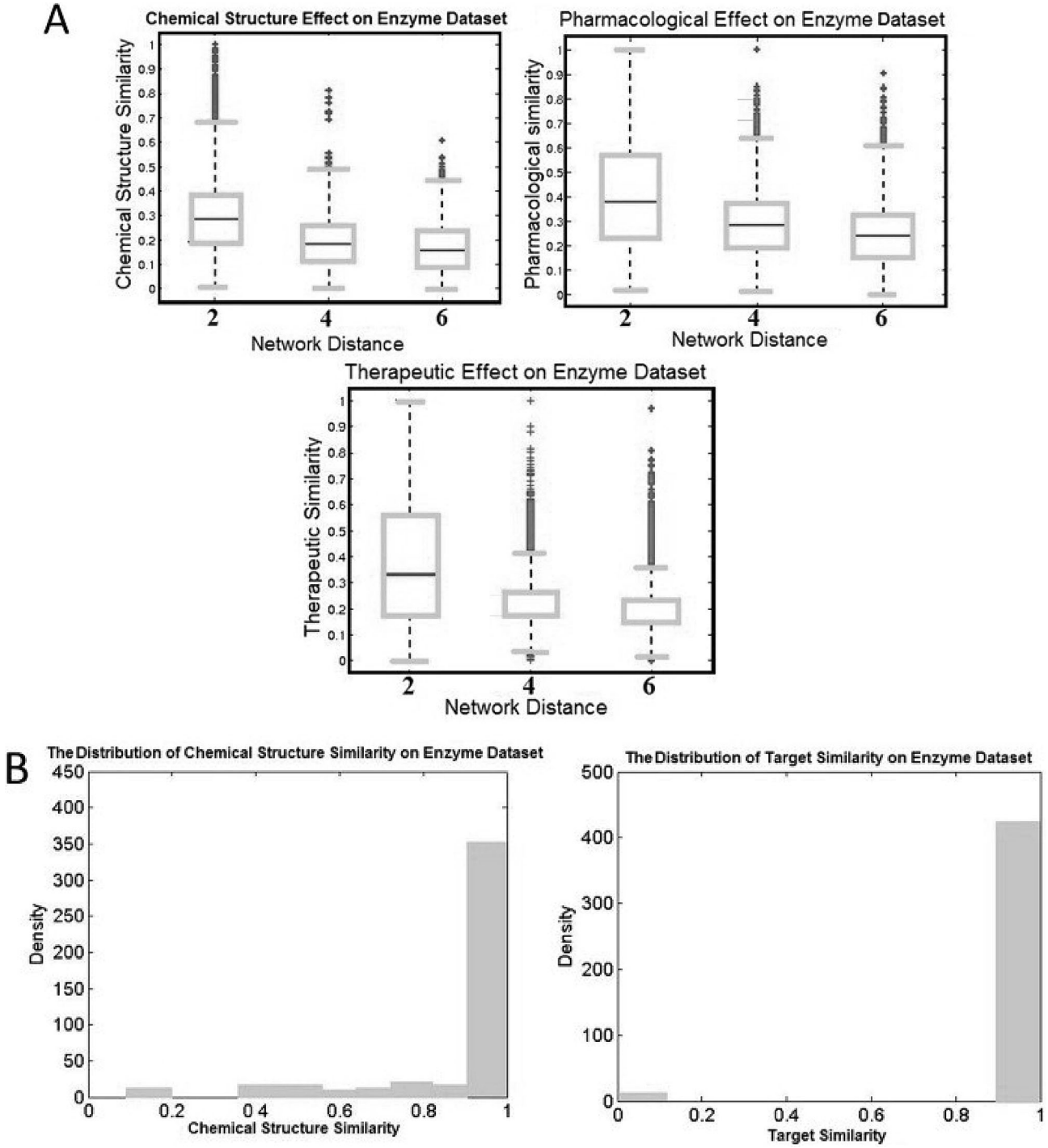


Figure 2. The correlation of various drug similarity scores with the distance in network on Enzyme dataset. A) The correlation of the chemical structure, pharmacological effect and therapeutic similarity scores with network distance for drugs targeting enzymes. It shows that drugs tend to have larger similarities when they are much closer in network, that is, drug pairs with similar chemical structure, pharmacological or therapeutic profile may interact with the same target proteins. B) The distribution of drug similarity scores among the drugs sharing common ATC-codes for enzyme dataset. It shows that two drugs sharing common ATC-codes tend to have larger similarities, that is, drug pairs with similar chemical structure, or target protein tend to be annotated with the common ATC-codes.

similarity matrix $\boldsymbol{S}_{comb}$ with $\boldsymbol{S}_{chem}$, $\boldsymbol{S}_{inter}$ and $\boldsymbol{S}_{side\text{-}effect}$ respectively.

For concise, we just plotted ROCs on NRs dataset in drug-target prediction in Figure 3A (ROCs on other three kinds of datasets can be seen in Figure 3 in Wang et al.[41]). We drew ROCs on NRs dataset in drug ATC-code prediction in Figure 3B (ROCs on other three kinds of datasets can be seen in Figure 4 in Wang et al.[42]); Figure 3A showed that,

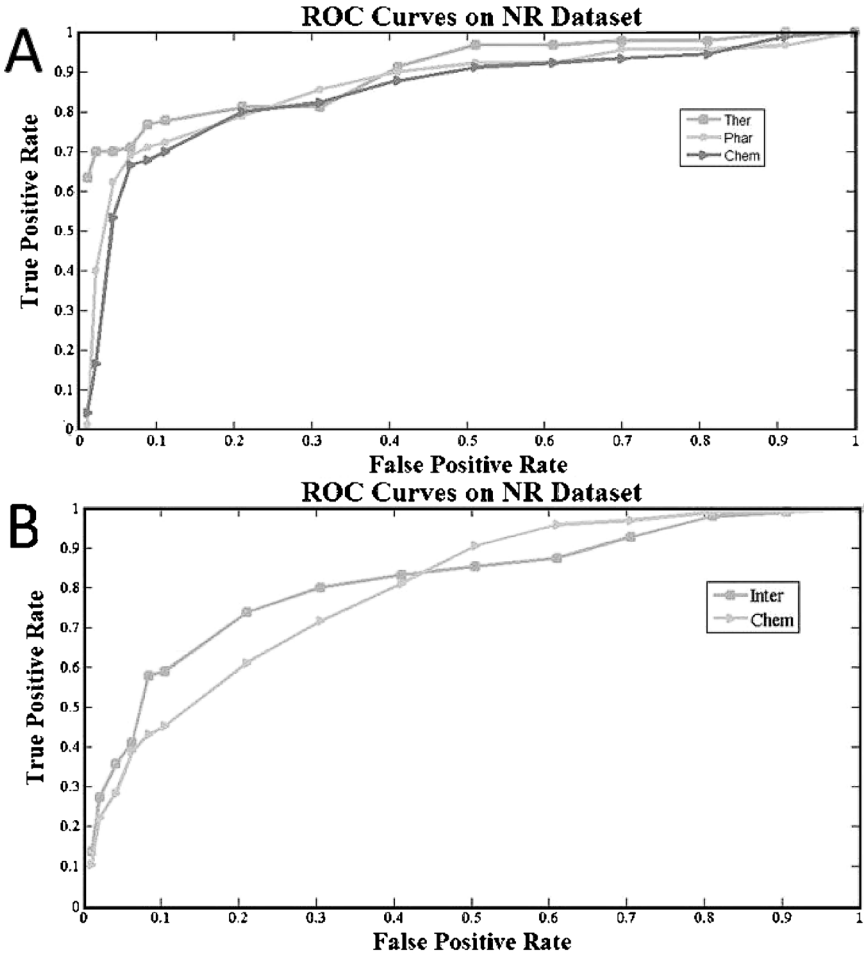


Figure 3. ROCs obtained by various data sources on NR dataset. A) The ROC curves derived from chemical structure (Chem), pharmacological (Phar) and therapeutic (Ther) data source on NR dataset. It shows that "Ther" outperforms "Chem" and "Phar". B) ROC curves for our methods using chemical structure (Chem) and target protein (Inter) data sources to predict ATC-codes for drugs interecting with NRs. It shows that target protein plays a more important role when false positive rate is small.

"Chem" and "Phar" perform almost the same. "Ther" outperforms "Chem" and "Phar" when FPR is very small. Figure 3B showed that, chemical structure is useful in ATC-codes prediction. Moreover, target protein plays a more important role in predicting drug's ATC-codes when FPR is very small. In drug repositioning task,[36] we observed that, "Chem" obtains the highest TPR when FPR is very small, and with the number of known interactions increasing, "Side-effect" reveals more experimentally observed drug-disease interactions.

The corresponding evaluation criteria when the corresponding F-measure reaches its maximum in discovering experimentally observed drug-target, drug ATC-code, and drug repositioning, were also calculated by Wang et al.[36,41,42] The evaluation criteria showed that each omics data source will do one's bit about inferring the potential rules from the existing interactions. Therefore, combination of these three data sources should produce a much more sophisticated picture of the interactions.

3.3 Data Fusion Improves Prediction

In the previous subsection, the usefulness of each data source was validated in uncovering the experimentally observed properties of drugs. In the following, we checked the performance for the combination of multiple data sources in prediction of exist drug-target, drug ATC-code, drug-disease interactions. The combination method, "Comb" is evaluated and visualized by ROC curve and evaluation criteria. The results in Wang et al.[36,41,42] showed that "Comb" could achieve better performance when predicting a small fraction of known interactions. What's more, "Comb" outperformed other methods with the highest AUC, sensitivity, specificity, and precision, upon the maximal *F*-measure. All these results suggest that the predictive performance can be further improved when multiple data sources are further incorporated into a single predictive model.

3.4 Novel Predictions for Further Validation

On cross-validation, "Comb" displayed its excellent performance in predicting experimentally observed interactions. To test whether it can produce biologically useful predictions, we tested our method on the unknown (non-interacting) drug-protein, drug ATC-code, and drug-disease pairs. We trained "Comb" on the gold standard positives and randomly selected gold standard negatives from the unknown pairs, and tested it on the remaining drug-target, drug ATC-code, and drug-disease pairs. Our expectation is that "Comb" can discover the novel interactions besides the gold standard positives.

For drug-target prediction, we took GPCRs network as an example for concise. The top ten predicted interactions on GPCRs dataset were listed in Table 4 in Wang et al.[41] For each novel prediction, we searched the corresponding evidences in KEGG[4] and DrugBank[5] and found evidence for eight of the top ten predictions. Furthermore, we noted that the annotations of two remaining predictions may indicate supporting evidences in biology. Took novel prediction: drug-protein pair of 'Clozapine' and 'dopamine receptor D3', as an example, the targets of 'Clozapine' in DrugBank and KEGG are 'dopamine receptor D1', 'dopamine receptor D2', and 'dopamine receptor D4'. They are in the same pathway of 'Neuroactive ligand-receptor interaction'. While 'dopamine receptor D3', also participates in the pathway of 'Neuroactive ligand-receptor interaction', that is, 'dopamine receptor D3' may relate with 'Clozapine' with a high probability. The analysis for another drug-protein pairs of 'Albuterolcan' and 'Chemokine None (C–X–C motif) receptor 1' can be found in Wang et al.[41] All these results suggested that, "comb" could uncover potential drug-protein interactions, and at least could provide low-resolution predictive results for further high-resolution experiments such as docking in drug discovery.

For drug ATC-code prediction, the top five predicted interactions on Enzyme, ICs, GPCRs, and NRs datasets were listed in Table 2 and Table S2–4 in Wang et al.,[42] respectively. For each drug and ATC-code pair in these tables, we checked their annotations from DrugBank[5] and WHOCC databases.[3] We further checked the explanation of drug and ATC-codes annotations form Wikipedia,[43] and finally analysed the reliability of predicted ATC-codes. Database search, literature search, and functional annotation analysis support these novel predictions. All these results suggest that "Comb" could uncover potential ATC-codes of drugs.

For drug repositioning prediction, the novel predicted drug-disease network was presented in.[36] We specifically took a close look at the top 100 newly predicted drug-disease associations. For each novel prediction, we checked the target proteins form DrugBank,[5] the disease genes from OMIM,[9] and the corresponding pathway information from KEGG BRITE.[4] We also checked whether novel predictions appear in current clinical trials.[44] Took the most confident prediction as an example, target protein 'Endothelin-1 receptor' (EDNRA) for 'Bosentan', and the disease gene 'KCNMB1' (Kca) for 'Hypertension, Diastolic, Resistance To' belong to the same pathway 'Arachidonic Acid metabolism'. Furthermore, we found that this drug-disease pair appeared in current clinical trials, the 'ClinicalTrials.gov Identifier' is NCT00820352. That is, this novel drug-disease pair may interact in vivo with high probability. Again, database search, literature search, and functional annotation analysis support these novel predictions.

4 Discussions and Conclusions

In this review, we described the data integration framework to computationally study drugs. We first surveyed the available omics data for drugs from different levels and different aspects, such as, compound chemical structure, drug pharmacology and therapeutic annotations, target proteins, side-effect, drug cured diseases and so on. Then we proposed kernel methods to integrate these data sources. Finally, we applied the kernel method to infer novel properties for drugs, including drug-target, drug ATC-code, and drug repositioning.

These three predictive tasks are similar in procedure. Specifically, we characterize drug, protein, ATC-code, and disease by their multiple similarity-based profiles, and define the kernel function to correlate drug with target protein, ATC-code, and disease. Then we train SVM-based classifier to predict novel drug-target, drug ATC-code, and drug-disease interactions. By cross-validation on well-established datasets, we found that each single data source is predictive. Moreover, more experimentally observed interactions can be uncovered by combination of multiple properties. In addition, database search and functional annotation analysis indicate that our new predictions are worthy of future experimental validation. We observed that the same framework works well in all three applications and can efficiently predict drug properties in different levels, i.e., drug target in molecular level, ATC-code in annotation level, and disease association in phenotype level. Thus we demonstrate that our kernel-based integration strategy serves as a useful tool to study drugs and will promote further research in drug discovery.

In addition to the common procedure to integrate data for the three applications, we also find each predictive task has its own feature. Thus different biological insights can be learned for each task and indicate further improvements. In drug-target and ATC-code prediction tasks, drug with four kinds of target network are used to validate the performance of our method. One possible concern is that the good predictive results are due to the homology proteins and similar compounds in our datasets. Our results showed that the average sequence similarities among the proteins and chemical similarities among the drugs are less than 0.2 in our dataset, that is, a redundancy cut-off has

been applied to reduce the homology bias when these datasets were constructed by Yamanishi et al.[27]

For drug ATC-codes prediction task, improved predictive performance was obtained by characterizing target protein sequence similarity. We noted that the improvement was robust to the definition of protein sequence similarity under different cut-offs to measure protein sequence similarity. We found that, the AUC score was slightly lower when using a more stringent cut-off, but not too much. This was because that most of sequence similarity among drug target was actually low in our dataset to avoid obvious results. One advantage to introduce target protein information is to fully utilize the indirect neighbor information in drug-target network. This allowed us to predict drug ATC-code interactions when this drug has low chemical similarity and target similarity with its closest drug. We listed some particularly interesting drug ATC-code predictions with low chemical similarity and target similarity in Table S6 in Wang et al.[42]

High performance in drug repositioning prediction may be due to those 'trivial' predictions, which are predictions that are obvious to anyone working on drug development. For example, those drugs sharing common target are easily to be predicted to treat the same disease. To address this issue, we filtered out the potential 'trivial' predictions and the performance was still far better than random classification. This result suggested that data integration can reveal 'non-trivial' predictions.

When predicting drug ATC-code and drug repositioning, we integrated target proteins to improve the prediction performance. The experimental results showed that, comparing with chemical structures, the performance was indeed improved by characterizing drugs in target sequence-based similarity. In fact, there are other ways to define the drug similarity based on their targets. For example, the targets closeness in PPI network can be considered to measure the target protein similarity.[38] In the future, we will utilize more target protein information by mining the PPI networks collected from multiple curated databases, including HPRD,[45] OPHID[46] and BIND[47] databases.

In drug repositioning prediction task, we only applied the phenotypic similarity to characterize diseases. However, with the development of systems biology, studies have shown that phenotype similar diseases are often caused by functionally related genes.[48] In addition, many large-scale studies support the idea that genes sharing similar diseases are tightly linked in the network.[49,50] Therefore, PPI network is useful to correlate disease with candidate genes.[48] Apart from gene closeness, function linkages among genes can be explored in different aspects.[51–53] Thus, it is promising to incorporate PPI network and other information to characterize disease.

In our experience, high quality training data is a key to our three predictive tasks. The training negative dataset is a formidable challenge to SVM-based algorithm as well as to other methods. Since the limitation of the available drug-target, drug ATC-code, drug-disease interactions, many unknown drug-protein, drug ATC-code, drug-disease pairs may be actually interacting in prediction task. To address this issue, a linear regression model can be introduced to uncover the potential interactions, which can avoid the bias in selection of negative dataset. The similar ideas have been used to prioritize the disease genes.[48] There is still much room to improve the predictive accuracy along this direction.

Acknowledgements

This work is supported by the *National Natural Science Foundation of China* (No. 11201470, No. 31270270, No. 11131009, No. 61171007, No. 11371365, and No. 11071252).

References

[1] http://pubchem.ncbi.nlm.nih.gov.
[2] http://www.japic.or.jp.
[3] http://www.whocc.no/atc_ddd_index/.
[4] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, *Nucleic Acids Res.* **2006**, *34*, D354 – D357.
[5] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, *Nucleic Acids Res.* **2008**, *36*, D901 – D906.
[6] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, D. Schomburg, *Nucleic Acids Res.* **2004**, *32*, D431 – D433.
[7] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, R. Preissner, *Nucleic Acids Res.* **2008**, *36*, D919 – D922.
[8] http://sideeffects.embl.de.
[9] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, V. A. McKusick, *Nucleic Acids Res.* **2002**, *30*, 52 – 55.
[10] J. Shawe-Taylor, N. Cristianini, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, **2000**.
[11] T. Hofmann, B. Schölkopf, A. J. Smola, *Ann. Statist.* **2008**, *36(3)*, 1171 – 1220
[12] B. Schölkopf, K. Tsuda, J. P. Vert, *Kernel Methods in Computational Biology*, MIT Press, Cambridge, MA, **2004**.
[13] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, W. S. Noble, *Pacific Symp. Biocomp.* **2004**, 300 – 311.
[14] A. Ben-Hur, W. S. Noble, *Bioinformatics* **2005**, *21 (Suppl. 1)*, i38 – i46 (Proc. Intelligent Syst. Mol. Biol. Conf.).
[15] M. Kuhn, M. Campillos, P. González, L. J. Jensen, P. Bork, *FEBS Lett.* **2008**, *582(8)*, 1283 – 1290.
[16] M. Dunkel, S. Günther, J. Ahmed, B. Wittig, R. Preissner, *Nucleic Acids Res.* **2008**, *36*, W55 – W59.
[17] L. Chen, W. M. Zeng, Y. D. Cai, K. Y. Feng, K. C. Chou, *PLoS One* **2012**, *7(4)*, e35254. doi:10.1371.
[18] J. T. Dudley, T. Deshpande, A. J. Butte, *Brief Bioinform.* **2011**, *12(4)*, 303 – 311.
[19] M. Hattori, Y. Okuno, S. Goto, M. Kanehisa, *J. Am. Chem. Soc.* **2003**, *125*, 11853 – 11865.
[20] E. Pauwels, V. Stoven, Y. Yamanishi, *BMC Bioinform.* **2011**, *12*, 169.

[21] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, *Bioinformatics* **2010**, *26*, i246 – i254.
[22] D. Lin, *Proc. 15th Internat. Conf. Machine Learning*, Morgan Kaufmann, San Francisco, CA, **1998**, pp. 296 – 304.
[23] M. Campillos, M. Kuhn, A. C. Gavin, L. Jensen, P. Bork, *Science* **2008**, *321*, 263 – 266.
[24] L Yang, P. Agarwal, *PLoS One* **2011**, *6(12)*, e28025.
[25] M. Duran-Frigola, P. Aloy, *Genome Medicine* **2012**, *4*, 3.
[26] Z. Xia, L. Y. Wu, X. B. Zhou, S. T. C. Wong, *BMC Syst. Biol.* **2010**, *4(Suppl 2)*, S6.
[27] Y. Wang, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, *Bioinformatics* **2008**, *24*, i232 – i240.
[28] S. W. Zhao, S. Li, *PLOS One* **2010**, *5(7),* e11764.
[29] Y. C. Wang, Z. X. Yang, Y. Wang, N. Y. Deng, *Lett. Drug Des. Discov.* **2010**, *7*, 370 – 378.
[30] T. F. Smith, M. Waterman, *J. Mol. Biol.* **1981**, *147*, 195 – 197.
[31] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, J. A. Leunissen, *Europ. J. Human Genet.* **2006**, *14*, 535 – 542.
[32] http://www.cmbi.ru.nl/MimMiner
[33] J. Basilico, T. Hofmann, in *Proc. 27th Ann. Int. ACM SIGRI Conf. Res. Develop. in Information Retrieval*, **2004**, 550—551.
[34] S. Oyama, C. D. Manning, in *Eur. Conf. Machine Learning*, **2004**, pp. 322–333.
[35] M. Hue, J. P. Vert, *Int. Conf. Machine Learning*, **2010**, pp. 463–470.
[36] Y. C. Wang, S. L. Chen, N. Y. Deng, Y. Wang, *PLOS One* **2013**, *8(11)*, e78518.
[37] G. Wu, E. Y. Chang, in *ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, *Washington, DC*, **2003**, 49–56.
[38] A. Gottlieb, G. Y. Stein, E. Ruppin, R. Sharan, *Mol. Syst. Biol.* **2011**, *7*, 496.
[39] C. C. Chang, C. J. Lin, *ACM Transacti. Intell. Syst. Technol.* **2011**, *2(27)*, 1 – 27.
[40] M. Gribskov, N. L. Robinson, *Comp. Chem.* **1996**, *20*, 25 – 33.
[41] Y. C. Wang, C. H. Zhang, N. Y. Deng, Y. Wang, *Comp. Biol. Chem.* **2011**, *35(6),* 353 – 362.
[42] Y. C. Wang, S. L. Chen, N. Y. Deng, Y. Wang, *Bioinformatics* **2013**, *29(10)*, 1317 – 1324.
[43] http://en.wikipedia.org/wiki/Main_Page.
[44] http://clinicaltrials.gov.
[45] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, A. Pandey, *Nucleic Acids Res.* **2009**, *37*, D767 – 772.
[46] K. R. Brown, I. Jurisica, *Bioinformatics* **2005**, *21*, 2076 – 2082.
[47] G. D. Bader, C. Wolting, B. F. Ouellette, T. Pawson, C. W. Hogue, *Nucleic Acids Res.* **2003**, *31*, 248 – 250.
[48] X. B. Wu, R. Jiang, M. Q. Zhang, S. Li, *Mol. Syst. Biol.* **2008**, *4*, 189 – 199.
[49] H. B. Fraser, J. B. Plotkin, *Genome Biol.* **2007**, *8*, R252.
[50] K. L. McGary, I. Lee, E. M. Marcotte, *Genome Biol.* **2007**, *8*, R258.
[51] J. C. Whisstock, A. M. Lesk, *Quart. Rev. Biophys.* **2003**, *36*, 307 – 340.
[52] P. D. Dobson, Y. D. Cai, B. J. Stapley, A. J. Doig, *Curr.Med. Chem.* **2004**, *11*, 2135 – 2142.
[53] Z. Wu, Y. Wang, L. Chen, *Mol. BioSyst.* **2013**, *9(6)*, 1268 – 1281.

Received: May 12, 2013
Accepted: November 13, 2013
Published online: December 11, 2013