JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Systematic Drug Repositioning for a Wide Range of Diseases with Integrative Analyses of Phenotypic and Molecular Data

Hiroaki Iwata,[†] Ryusuke Sawada,[†] Sayaka Mizutani,[‡] and Yoshihiro Yamanishi*[,†,§]

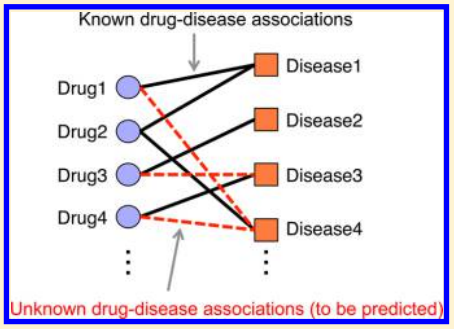
[†]Division of System Cohort, Multi-Scale Research Center for Medical Science, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan

[‡]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

[§]Institute for Advanced Study, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka, Fukuoka 812-8581, Japan

Supporting Information

ABSTRACT: Drug repositioning, or the application of known drugs to new indications, is a challenging issue in pharmaceutical science. In this study, we developed a new computational method to predict unknown drug indications for systematic drug repositioning in a framework of supervised network inference. We defined a descriptor for each drug−disease pair based on the phenotypic features of drugs (e.g., medicinal effects and side effects) and various molecular features of diseases (e.g., disease-causing genes, diagnostic markers, disease-related pathways, and environmental factors) and constructed a statistical model to predict new drug−disease associations for a wide range of diseases in the International Classification of Diseases. Our results show that the proposed method outperforms previous methods in terms of accuracy and applicability, and its performance does not depend on drug chemical structure similarity. Finally, we performed a comprehensive prediction of a drug−disease association network consisting of 2349 drugs and 858 diseases and described biologically meaningful examples of newly predicted drug indications for several types of cancers and nonhereditary diseases.



INTRODUCTION

In recent years, it has become very difficult to develop new drugs using conventional drug discovery procedures,[1] as reflected by the fact that the number of newly approved drugs has tended to decrease worldwide. Drug repositioning, or the application of known drugs to new indications (i.e., new diseases), is an emerging strategy for drug development, and it has received remarkable attention in pharmaceutical science.[2,3] In fact, many drugs have enormous potential for new drug indications in terms of polypharmacology.[4−6] Various sources of information on known drugs (e.g., pharmacokinetics, safety for humans, and manufacturing processes) can be reused; thus, the drug repositioning approach is expected to increase the success rate of drug development and to reduce the cost in terms of time, risk, and expenditure.[7]

From the perspective of systems biology, drugs can be regarded as molecules that induce perturbations to biological systems, leading to various phenotypic effects on the human body. Phenotypic effects include the expected medicinal effects caused by interactions with the primary target as well as side effects caused by interactions with off-targets.[8,9] Drug side effects are usually undesired, but they can sometimes be beneficial and lead to new therapeutic indications. For example, sildenafil (Viagra) was developed to treat angina, but it is now used for the treatment of erectile dysfunction owing to the discovery of erectile response as a side effect. However, such unexpected drug activities have been discovered largely by chance from observed phenotypes, and the previous drug repositioning approach has been dependent on serendipity. There is therefore a strong incentive to develop computational methods for systematic drug repositioning on a large scale.[10]

Recently, a variety of computational methods for systematic drug repositioning have been proposed. Examples include the similarity-based approach[11−18] and the machine learning approach.[19−22] In both these approaches, prediction was based on drug features (e.g., chemical structures, side effects, target protein sequences, and biological pathways) and disease features (e.g., symptomatic state and gene expression). However, previous methods have several limitations in terms of applicability and prediction accuracy. Previous methods depend on disease information stored in the Online Mendelian Inheritance in Man (OMIM) database;[23] thus, they tend to make predictions only for Mendelian inheritance disorders. In addition, prediction accuracy depends heavily on drug chemical structure similarity and the availability of different data sources for the same drugs or the same diseases; thus, there remains some potential for improvement regarding prediction accuracy and applicability.

In this study, we developed a new computational method to predict unknown drug indications for systematic drug repositioning in a framework of supervised network inference. We defined a descriptor for each drug−disease pair based on the phenotypic features of drugs (e.g., medicinal effects and side

Received: November 6, 2014
Published: January 20, 2015

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

effects) and various molecular features of diseases (e.g., disease-causing genes, diagnostic markers, disease-related pathways, and environmental factors). On the basis of the descriptors, we constructed a statistical model to predict new drug−disease associations for a wide range of diseases defined in the International Classification of Diseases (ICD-10)[24] using the KEGG DISEASE database,[25] which stores various molecular features of ICD-defined diseases including common diseases, compared with OMIM. Our results show that the proposed method outperforms previous methods in terms of accuracy and applicability, and its performance does not depend on drug chemical structure similarity. Finally, we performed a comprehensive prediction of a drug−disease association network consisting of 2349 drugs and 858 diseases and described biologically meaningful examples of newly predicted drug indications for several types of cancers (e.g., breast cancer, cervical cancer, and Hodgkin lymphoma) and nonhereditary diseases (e.g., allograft rejection, Whipple's disease, and viral myocarditis).

MATERIALS AND METHODS

Data Sets. Information on drug phenotypic effects was downloaded from FDA Adverse Event Reporting System (FAERS), which has been collecting drug adverse event reports submitted to the FDA by healthcare professionals from the first quarter of 2004 to the third quarter of 2012. We used adverse drug reactions (ADRs) and side effects provided in the FDA database as the phenotypic effects of drugs in this study. FAERS drug names were converted into KEGG D numbers by complete and partial name matching followed by manual inspection. Partial name matching used the Anatomical Therapeutic Chemical (ATC) classification system, in which FAERS drug entities and KEGG drug entities were matched if they were classified into the same ATC category at the fifth level (chemical structure level). Drugs tagged as primary and secondary causes of the ADRs were used. As a result, 2349 drugs were associated with a total of 15,188 phenotypic effects. For each of 2349 drugs, we constructed a 15,188-dimensional binary vector whose elements encode for the presence or absence of the phenotypic effects by 1 or 0, respectively.

The chemical substructures of drugs were encoded by a chemical fingerprint with 881 chemical substructures defined in the PubChem database.[26] Chemically identical drugs with the same structures (duplicates) were removed, and thus, the chemical structures of all drugs were unique. Each drug was represented by an 881-dimensional binary vector whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively.

Various features of diseases were obtained from the 10th revision of the International Classification of Diseases (ICD-10) (http://apps.who.int/classifications/icd10/browse/2010/en)[24] and the KEGG DISEASE database.[25,27] As a result, 858 diseases were represented by 2649 features (e.g., diagnostic markers, disease-causing genes, biological pathways, and environmental factors). Each of 858 diseases was represented by a 2649-dimensional binary vector whose elements encode for the presence or absence of each disease feature by 1 or 0, respectively.

Drug−disease associations were obtained from the KEGG DISEASE database,[27] where the information about drug indication was based on medical books and drug package inserts.[28] The resulting data set consisted of 1432 drug−disease associations involving 694 drugs (identified by D numbers) and 323 diseases (identified by H numbers). This data set was used as gold standard data in the cross-validation (CV) experiment and as training data in the new prediction model.

Proposed Methods. In this study, we considered solving the drug−disease association network prediction problem in the framework of supervised classification. We represented a pair of drug X and disease Z by (X, Z). Suppose that we were given a learning set of drug−disease pairs (X_i, Z_j) ($i = 1, 2,..., n_X$; $j = 1, 2,..., n_Z$), where the pairs were known to be or not to be associating pairs, n_X is the number of drugs, and n_Z is the number of diseases in the learning set.

We represented a drug−disease pair (X, Z) by a feature vector $\Phi(X, Z)$ and then estimated a function $f(X, Z) = w^{\mathrm{T}}\Phi(X, Z)$ that would predict whether the drug−disease pair (X, Z) was an associating pair or not. We optimized the weight vector w based on the learning set with label information.

The profile of drug X was represented by an M-dimensional binary vector as follows: $\Phi(X) = (x_1, x_2,..., x_M)^{\mathrm{T}}$, where $x_k \in \{0, 1\}$, $k = 1,..., M$. $\Phi(X)$ is a phenotypic profile of drugs or chemical substructure profile of drugs in this study. In the same manner, the molecular profile of disease Z was represented by an N-dimensional binary vector as follows: $\Phi(Z) = (z_1, z_2,..., z_N)^{\mathrm{T}}$, where $z_l \in \{0, 1\}$, $l = 1,..., N$.

We represented a feature vector for each drug−disease pair using the tensor product between $\Phi(X)$ and $\Phi(Z)$ as follows:

$$\Phi(X, Z) = \Phi(X) \otimes \Phi(Z) = (x_1z_1, ..., x_1z_N, ..., x_Mz_1, ..., x_Mz_N)^{\mathrm{T}}$$

where $\Phi(X, Z)$ is an $(M \times N)$-dimensional feature vector. This tensor product descriptor is similar to that in a previous work on a different biological problem.[29]

In this study, we used logistic regression as a binary classifier to predict whether a drug X was associated with a disease Z. The predictive model was learned by minimizing the loss function with L_1-regularization to avoid overfitting. Suppose that we have a learning set of drug−disease pairs and association labels $(\Phi(X_i, Z_j), y_{ij})$, $y_{ij} \in \{+1, -1\}$ ($i = 1, 2,..., n_X$; $j = 1,2,..., n_Z$), where n_X is the number of drugs and n_Z is the number of diseases in the learning set. The weight vector w of the linear logistic regression was learned with L_1-regularization as follows:

$$\min_w \| w \|_1 + C \sum_{i=1}^{n_X} \sum_{j=1}^{n_Z} \log(1 + \exp(-y_{ij} w^{\mathrm{T}} \Phi(X_i, Z_j)))$$

where $\|\cdot\|_1$ is L_1 norm (the sum of absolute values), and C is a hyper parameter to control the sparsity. We examined various values (0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000) for the hyper parameter C and selected the value that gave the highest the area under the receiver operating characteristic curve (AUC) score in the CV experiment. We used the LIBLINEAR algorithm[30] to train L_1-regularized logistic regression model. LIBLINEAR is known to be the most efficient and suitable for learning linear models on binary feature vectors, and the software is freely available at http://www.csie.ntu.edu.tw/~cjlin/liblinear/. The input of LIBLINEAR is a list of existing features for each object (drug−disease pair in this study), which enables to reduce the consumed memory in practice.

Performance Evaluation Protocol. The pairwise CV was performed as follows. First, we split all drug−disease pairs in the gold standard set into five subsets in an independent manner. Second, we regarded one subset of drug−disease pairs

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

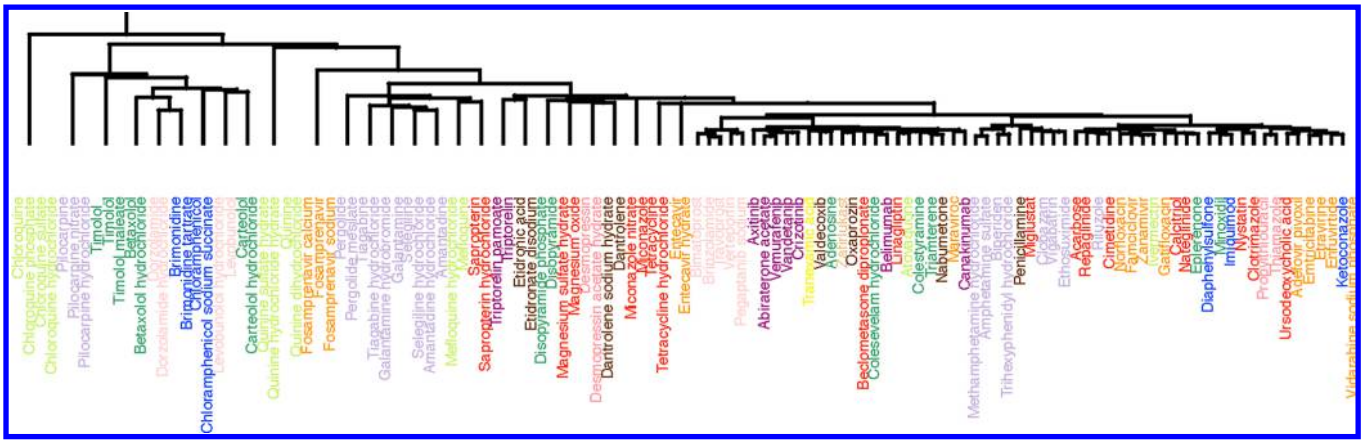


Figure 1. Clustering of drugs using the phenotypic similarity scores. Drug names are colored according to the first level Anatomical Therapeutic Chemical classification system as follows: (1) red, alimentary tract and metabolism, (2) yellow, blood and blood-forming organs, (3) green, cardiovascular system, (4) blue, dermatologicals, (5) light blue, genitourinary system and sex hormones, (6) pink, systemic hormonal preparations, excluding sex hormones and insulin, (7) orange, anti-infectives for systemic use, (8) purple, antineoplastic and immunomodulating agents, (9) brown, musculoskeletal system, (10) light purple, nervous system, (11) beige, antiparasitic products, insecticides, and repellents, (12) light yellowish green, respiratory system, (13) light pink, sensory organs, and (14) light gray, various.

as a test set and regarded the remaining four subsets of drug−-disease pairs as a training set. Third, we optimized a predictive model based on drug−disease pairs in the training set. Finally, we applied the predictive model to drug−disease pairs in the test set. Note that drug−disease pairs were considered independent of each other, and thus, the drugs and diseases in test pairs overlapped with those in the training set to some extent.

The blockwise CV was performed as follows. First, we split drugs and diseases in the gold standard set into five subsets of drugs and five subsets of diseases. Second, we regarded one subset of drugs (respectively diseases) as test drugs (respectively test diseases), and the remaining four subsets of drugs were used as training drugs (respectively training diseases). Third, we optimized a predictive model based on drug−disease pairs consisting of training drugs and training diseases. Finally, we computed the prediction scores for three types of drug−disease pairs: test drugs vs training diseases (referred to as "drug-based CV"), training drugs vs test diseases (referred to as "disease-based CV"), and test drugs vs test diseases (referred to as "drug&disease-based CV"). Note that drugs and diseases in test pairs were not completely different from those in the training set.

Gold standard data contained many drugs that were chemically and structurally almost identical because they were derived from the same lead compound. If these identical drugs were divided into a training set and a test set, prediction in the CV experiment would be trivial.[31,32] To avoid overestimation of prediction accuracy, we performed the filtering of similar drugs based on their chemical structures and used only drugs that were chemically and structurally different to some extent (Figure S1, Supporting Information). First, we calculated the Tanimoto coefficient (Jaccard coefficient)[33] of chemical fingerprints for all drugs. Second, we detected drugs sharing high Tanimoto coefficients and selected one representative drug based on a threshold. Third, we constructed a set of representative drugs sharing low Tanimoto coefficients. Finally, we prepared five sets of benchmark data by varying the threshold (e.g., from 0.6 to 1.0 via an increment of 0.1). When the thresholds were 0.1−0.5, the numbers of drug clusters were less than 20, making model learning difficult and unstable. Thus, we did not test the thresholds of 0.1−0.5 in the CV.

We compared the proposed method with two previous methods: PREDICT[19] and PreDR.[21] In PREDICT, chemical fingerprints, side effects in package inserts, amino acid sequences of target proteins, protein−protein interactions of target proteins, functional terms of target proteins were used as features of drugs, and phenotypes were used for diseases. In PreDR, chemical fingerprints, side effects in package inserts, and target proteins were used for drugs, and semantic phenotype similarities were used as features of diseases. Note that the number of drugs/diseases with all data sources is very limited. We implemented the algorithm of existing methods (PREDICT and PreDR) and optimized parameters (hyper parameters or regularization parameters) with AUC scores for each existing method as well by an internal cross-validation using only a training set for each method.

RESULTS

Clustering of Drugs and Diseases with Phenotypic and Molecular Profiles. As a preliminary analysis, we performed the clustering of 2349 drugs using their phenotypic similarities, in which drug phenotypic similarity was based on the drug phenotypic profiles of 15,188 phenotypic effects (see Materials and Methods for more details). Figure 1 shows the partial result of drug clusters after applying hierarchical clustering, in which drug names are colored according to the first level Anatomical Therapeutic Chemical (ATC) classification system (Figure S2, Supporting Information for the high-resolution version of Figure 1 for all drugs). We used the Ward method in order to divide drugs into many compact clusters with distinct functions because it is known that Ward method tends to generate many compact clusters. It was observed that drugs with different ATC codes were clustered into the same group, and drugs with the same ATC code were separately clustered. This result implies that drug phenotypic effects are independent of the original drug indications to some extent, and drugs with similar phenotypic patterns are likely to have similar effects regardless of the predefined drug codes.

Next, we performed the clustering of 858 diseases using their molecular similarities, in which disease molecular similarity was

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

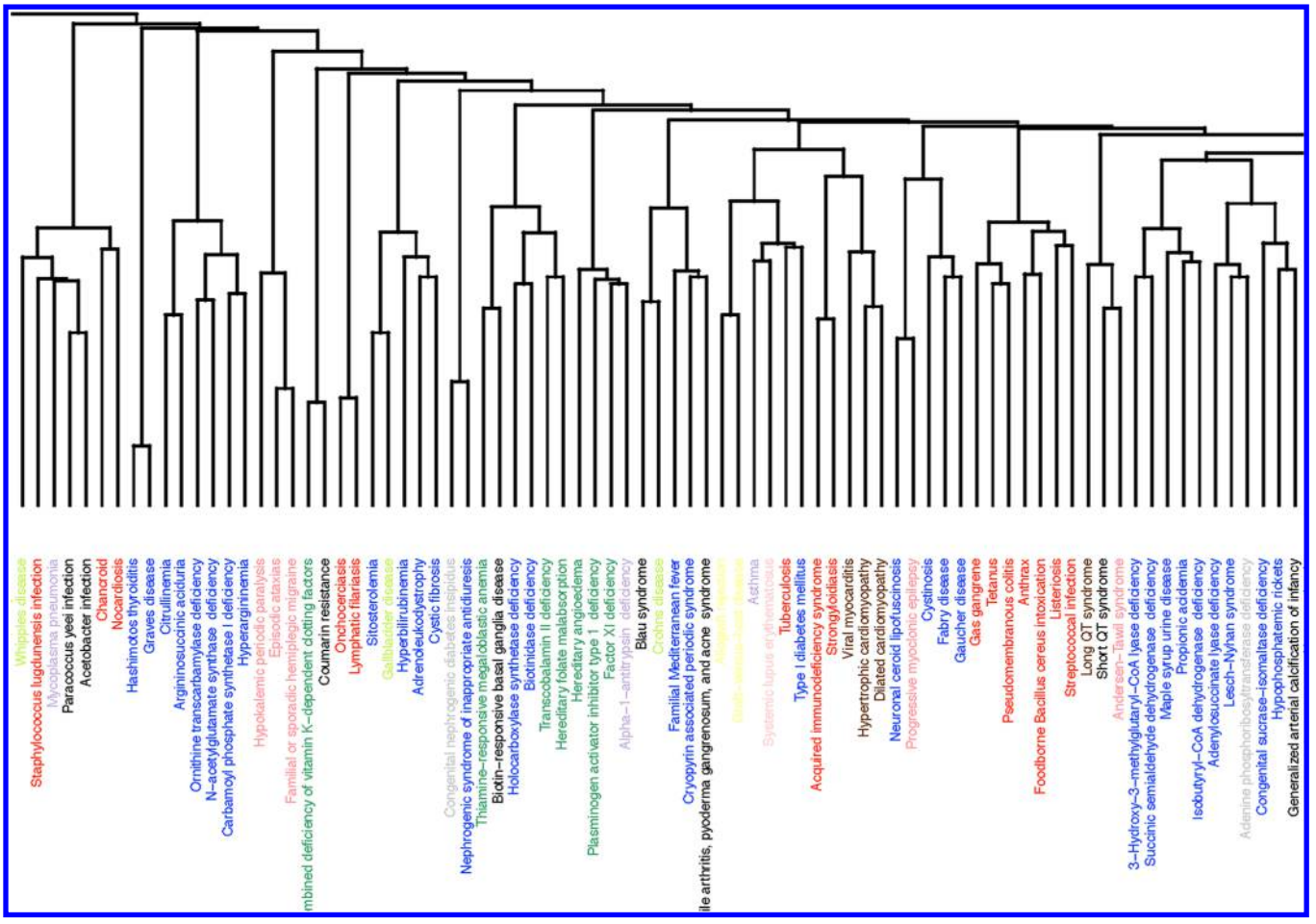


Figure 2. Clustering of diseases using the molecular similarity scores. Disease names are colored according to the 10th version of the International Classification of Disease as follows: red, chapter I (certain infectious and parasitic diseases [A00−B99]); yellow, chapter II (neoplasms [C00−D48]); green, chapter III (diseases of blood, blood-forming organs, and certain disorders involving the immune mechanism [D50−D89]); blue, chapter IV (endocrine, nutritional, and metabolic diseases [E00−E90]); light blue, chapter V (mental and behavioral disorders [F00−F99]); pink, chapter VI (diseases of the nervous system [G00−G99]); orange, chapter VII (diseases of the eye and adnexa [H00−H59]), brown, chapter IX (diseases of the circulatory system [I00−I99]); light purple, chapter X (diseases of the respiratory system [J00−J99]); beige, chapter XI (diseases of the digestive system [K00−K93]); light pink, chapter XIII (diseases of the musculoskeletal system and connective tissue [M00−M99]); light gray, chapter XIV (diseases of the genitourinary system [N00−N99]); purple, chapter XVII (congenital malformations, deformations, and chromosomal abnormalities [Q00−Q99]; light yellowish green, chapter XVIII (symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified [R00−R99]); cream, chapter XIX (injury, poisoning, and certain other consequences of external causes [S00−T98]); and black, chapter XXII (codes for special purposes [U00−U99]).

based on the molecular profiles of 2649 disease features (e.g., disease-causing genes, diagnostic markers, disease-related pathways, and environmental factors) (see Materials and Methods for more details). Figure 2 shows the partial result of disease clusters after applying hierarchical clustering using the Ward method, in which disease names are colored according to the 10th revision of the International Classification of Diseases (ICD-10) (http://apps.who.int/classifications/icd10/browse/2010/en)[24] (see Figure S3, Supporting Information, for the high-resolution version of Figure 2 for all diseases). It was observed that diseases with different ICD chapters were clustered into the same group, and diseases with the same ICD chapter were separately clustered. This result implies that disease features do not always correspond to the original disease classification, and diseases with similar features may be treated by the same drug regardless of the predefined disease classification.

Proposed Algorithm for Systematic Drug Repositioning. Drug indication predictions can be formulated as the problem of inferring an unknown part of the drug−disease association network from data on drugs and diseases. We proposed a statistical method for predicting the unknown part of the drug−disease association network in the framework of supervised network inference using all known drug−disease associations as a learning data set. Figure 3 shows an illustration of our proposed algorithm.

First, we obtained a set of known drug−disease associations from the KEGG database,[25] where the information was based on drug package inserts and medical books.[28] Second, we constructed a descriptor of each drug−disease pair by computing a tensor product of the drug phenotypic profile and the disease molecular profile. Third, we learned a predictive model using logistic regression based on training data. Finally, we applied the predictive model to all possible drug−disease pairs and assigned prediction scores to all drug−disease pairs. More detailed explanations regarding the algorithm can be found in the Materials and Methods.

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

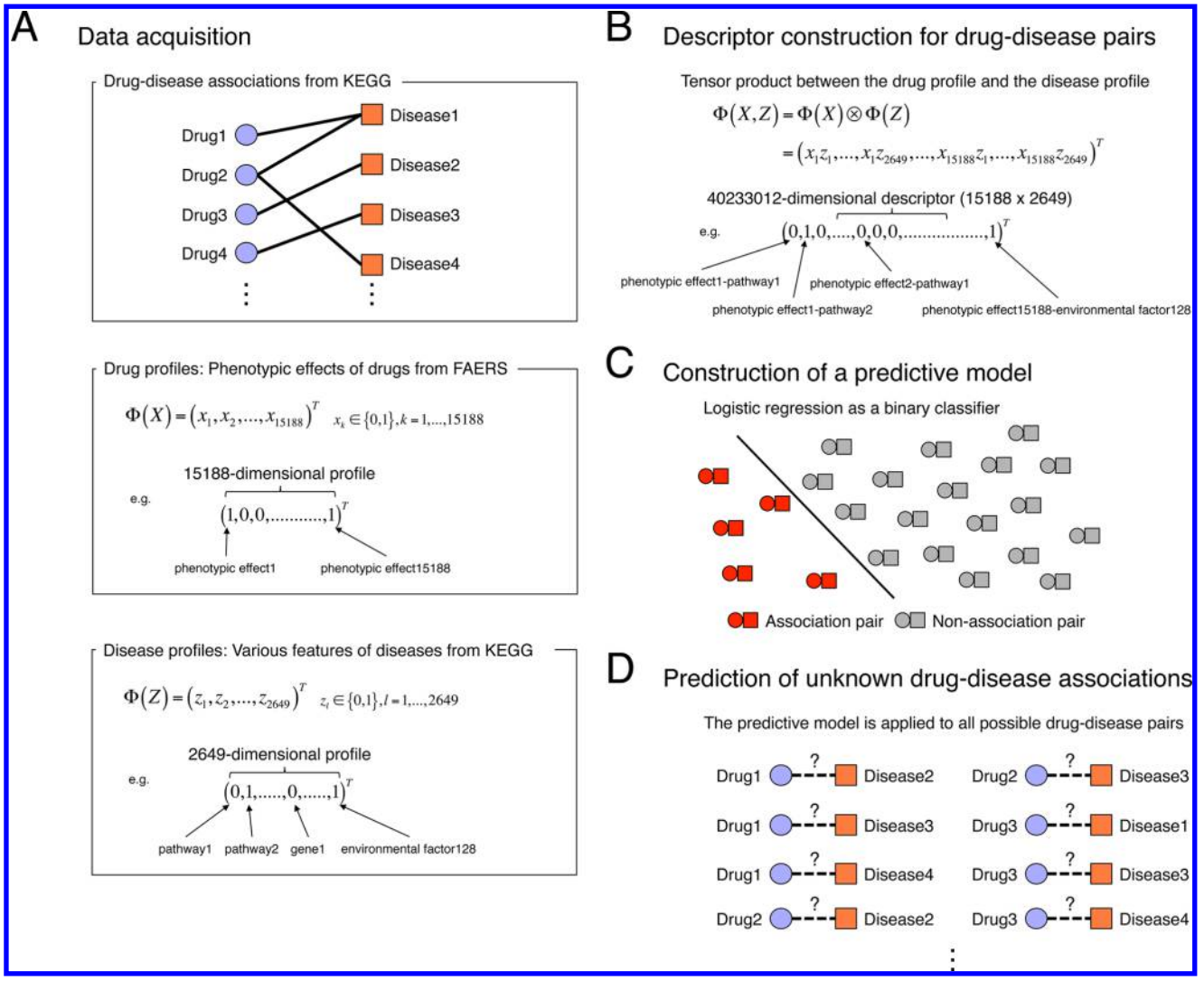


Figure 3. Workflow of the proposed method. First, we obtained known drug–disease associations from the KEGG database and the medical literature. Second, we constructed a descriptor of each drug–disease pair by computing a tensor product of the drug phenotypic profile and the disease molecular profile. Third, we learned a predictive model using logistic regression based on training data. Finally, we applied the predictive model to all possible drug–disease pairs and assigned prediction scores to all drug–disease pairs.

Performance Evaluation of the Reconstruction of Known Drug–Disease Associations. We tested the proposed method for its ability to reconstruct 1432 known drug–disease associations involving 694 drugs and 323 diseases. We regarded the 1432 known drug–disease associations as positive instances, and regarded the other 222,730 drug–disease pairs as negative instances. In addition, we tested the use of drug chemical structures as a baseline method with the same predictive model of the proposed method because chemical structures have often been used in many chemoinformatics problems such as the quantitative structure–activity relationship (QSAR), and they contribute to high prediction accuracy in the drug indication problem. We evaluated the performance of the methods by two types of cross-validations (CVs): pairwise CV and blockwise CV (see Materials and Methods for more details).

We evaluated prediction accuracy by the receiver operating characteristic (ROC) and precision-recall (PR) curves. The ROC curve is a plot of true positive rates (correctly predicted pairs) against false positive rates (incorrectly predicted pairs), and the area under the ROC curve (AUC) returns 1 for perfect inference and 0.5 for random inference. The PR curve is a plot of precision (positive predictive value) against recall (sensitivity), and the area under the PR curve (AUPR) returns 1 for perfect inference and returns the ratio of positive samples to all samples for random inference. We performed the CV experiments on seven benchmark data sets that were constructed with various thresholds for drug chemical structure similarities (see Materials and Methods for more details and Figure S1, Supporting Information).

Figure 4 shows the AUC and AUPR scores of the proposed (phenotype-based method) and baseline methods (chemical structure-based method) in the pairwise and blockwise CVs on seven benchmark data sets, in which the horizontal axis in each panel indicates clustering thresholds for the chemical structure similarity of drugs. A higher clustering threshold implied that the benchmark data set consisted of many structurally similar drugs, whereas a lower clustering threshold implied that the benchmark data set consisted of structurally diverse drugs. The proposed method worked better than the baseline method for most benchmark data sets in terms of both AUC and AUPR. It was observed that the scores of the blockwise CV were lower

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446–459

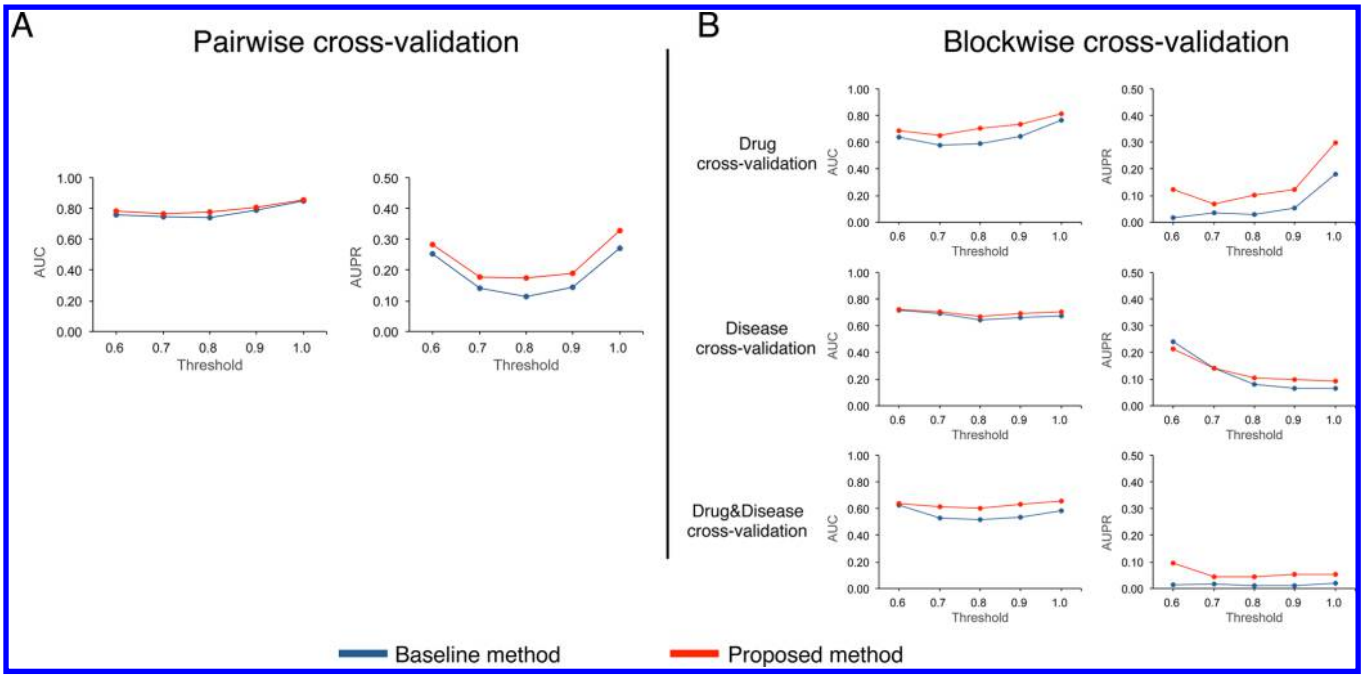


Figure 4. AUC and AUPR scores in the cross-validation experiments. The results of the pairwise CV are shown in panel (A), and the results of the blockwise CV are shown in panel (B). The horizontal axis in each panel denote the threshold for the chemical structure similarity of drugs in constructing benchmark data sets.

than those of the pairwise CV. This implies that it is more difficult to predict new indications for orphan drugs/diseases (without any known drug−disease association) than to predict new indications for drugs/diseases of known associations (with at least one known drug−disease association).

We investigated differences in three types of blockwise CVs: drug-based (test drugs vs training diseases), disease-based (training drugs vs test diseases), and drug&disease-based CVs (test drugs vs test diseases). In the drug-based CV, the proposed method worked much better than the baseline method when the benchmark data had many structurally diverse drugs, suggesting an advantage of using phenotypic effects rather than chemical structures. The same tendency was also observed in the disease-based and drug&disease-based CVs, but the tendency was smaller than that in the drug-based CV. One explanation is that the drug similarity is crucial in the drug-based CV because drugs differed between the training set and test set in the CV simulation. These results suggest that the proposed method is capable of predicting drug indications without depending on the chemical structures of drugs in practical situations.

Performance Comparison with Other Methods. We compared the performance of the proposed method with that of the existing methods for drug indication prediction. The state-of-the-art methods are PREDICT[19] and PreDR,[21] both of which uses machine learning algorithms and both of which are based on disease information in the OMIM database. It was demonstrated that PREDICT and PreDR worked much better than the other drug repositioning methods reported previously.[19,21] Therefore, we compared the performance of the proposed method with that of PREDICT and PreDR in this study.

First, we compared prediction accuracy among the three methods. We used a set of drug−disease pairs involving common drugs and common diseases across the three methods as gold standard data with the correspondence of IDs of drugs and diseases. We used only drugs and diseases for which all features suggested in previous works could be obtained. Figure S4 of the Supporting Information presents the ROC and PR curves of the proposed method, PREDICT, and PreDR. In the pairwise CV, the AUC and AUPR scores of the proposed method were 0.80 and 0.30, respectively, those of PREDICT were 0.77 and 0.30, respectively, and those of PreDR were 0.66 and 0.076, respectively. The same tendency was also observed in the blockwise CV (Figure S4, Supporting Information). These results suggest the superiority of the proposed method over the other methods in terms of prediction accuracy.

We used all known drug−disease associations in KEGG as gold standard data. However, some features (e.g., drug side effects, drug targets, disease pathways, and disease phenotypes) were not always available for all drugs and all diseases in each method, and thus, missing data in the corresponding feature profiles were imputed by zeros in each method. Figure S5 of the Supporting Information shows the ROC and PR curves of the proposed method, PREDICT, and PreDR. In the pairwise CV, the AUC and AUPR scores of the proposed method were 0.79 and 0.31, respectively, those of PREDICT were 0.61 and 0.018, respectively, and those of PreDR were 0.60 and 0.015, respectively. The same tendency was also observed in the blockwise CV (Figure S5, Supporting Information). These results suggest the superiority of the proposed method over the other methods in terms of applicability.

Large-Scale Prediction of Novel Drug Indications. After confirming the usefulness of our proposed method, we conducted a comprehensive prediction of unknown associations between all possible drugs and all possible diseases. We learned a predictive model based on all gold standard data as training data and applied the model to 2349 drugs with phenotypic profiles and 858 diseases with molecular profiles. Note that there remained numerous marketed drugs whose indications were not fully characterized. As a result, we were able to predict 117,880 new drug−disease associations

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

Table 1. Distribution of Drugs Repositioned from Original Disease Class to Other Disease Classes by the Proposed Method

ICD chapter[a] (# of drugs)	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX	XXI	XXII
chapter I (190)	185	128	139	171	131	139	111	6	143	146	133	82	166	128	0	138	179	6	5	0	1	0
chapter II (149)	140	145	128	144	115	127	111	1	119	75	115	54	140	124	0	117	149	14	19	0	7	0
chapter III (34)	32	27	28	31	25	25	22	2	28	26	25	17	32	25	0	27	33	1	5	0	0	0
chapter IV (125)	115	100	94	124	90	101	91	19	95	85	95	86	113	73	0	95	123	21	9	0	0	0
chapter V (2)	2	1	1	2	2	2	1	0	2	2	1	1	1	2	0	1	2	1	0	0	0	0
chapter VI (86)	83	71	75	84	71	79	57	5	72	73	72	64	80	70	0	70	86	16	1	0	0	0
chapter VII (31)	29	28	29	24	23	23	12	2	26	22	27	18	24	13	0	23	31	5	0	0	0	0
chapter VIII (0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chapter IX (29)	29	23	28	29	24	24	21	0	29	24	25	17	26	28	0	26	29	3	3	0	0	0
chapter X (58)	53	47	43	43	39	42	36	4	46	31	40	33	54	45	0	40	57	4	3	0	0	0
chapter XI (15)	14	11	13	15	11	14	13	2	13	14	12	8	14	10	0	14	14	3	2	0	0	0
chapter XII (3)	3	2	3	3	1	2	2	0	1	3	3	2	3	1	0	1	3	0	1	0	0	0
chapter XIII (49)	46	43	44	44	39	39	46	15	39	39	42	32	49	37	0	39	49	2	11	0	0	0
chapter XIV (9)	9	9	7	9	7	7	4	0	7	9	7	7	9	7	0	7	9	1	0	0	0	0
chapter XV (0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chapter XVI (13)	13	7	13	12	9	11	9	0	9	7	10	4	13	11	0	9	13	0	1	0	0	0
chapter XVII (2)	2	1	2	2	1	1	1	0	1	2	1	0	2	2	0	1	2	0	0	0	0	0
chapter XVIII (6)	6	6	6	6	6	6	6	0	6	5	6	6	6	6	0	6	6	0	1	0	0	0
chapter XIX (14)	14	14	14	14	14	14	14	0	14	10	14	12	14	14	0	14	14	2	7	0	0	0
chapter XX (0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chapter XXI (0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
chapter XXII (1)	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0

[a]Each element represents the number of drugs repositioned from the original disease class to new disease classes by the proposed method. The rows indicate the original ICD disease chapters, and the columns indicate the newly predicted ICD disease chapters. Chapter I: certain infectious and parasitic diseases (A00-B99); chapter II: neoplasms (C00-D48); chapter III: diseases of the blood, blood-forming organs, and certain disorders involving the immune mechanism (D50−D89); chapter IV: endocrine, nutritional, and metabolic diseases (E00−E90); chapter V: mental and behavioral disorders (F00−F99); chapter VI: diseases of the nervous system (G00−G99); chapter VII: diseases of the eye and adnexa (H00−H59); chapter VIII: diseases of the ear and mastoid process (H60−H95); chapter IX: diseases of the circulatory system (I00−I99); chapter X: diseases of the respiratory system (J00−J99); chapter XI: diseases of the digestive system (K00−K93); chapter XII: diseases of the skin and subcutaneous tissue (L00−L99); chapter XIII: diseases of the musculoskeletal system and connective tissue (M00−M99); chapter XIV: diseases of the genitourinary system (N00−N99); chapter XV: pregnancy, childbirth, and the puerperium (O00−O99); chapter XVI: certain conditions originating in the perinatal period (P00−P96); chapter XVII: congenital malformations, deformations; and chromosomal abnormalities (Q00−Q99); chapter XVIII: symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified (R00−R99); chapter XIX: injury, poisoning, and certain other consequences of external causes (S00−T98); chapter XX: external causes of morbidity and mortality (V01−Y98); chapter XXI: factors influencing health status and contact with health services (Z00−Z99); and chapter XXII: codes for special purposes (U00−U99).

J. Chem. Inf. Model. 2015, 55, 446−459 DOI: 10.1021/ci500670q

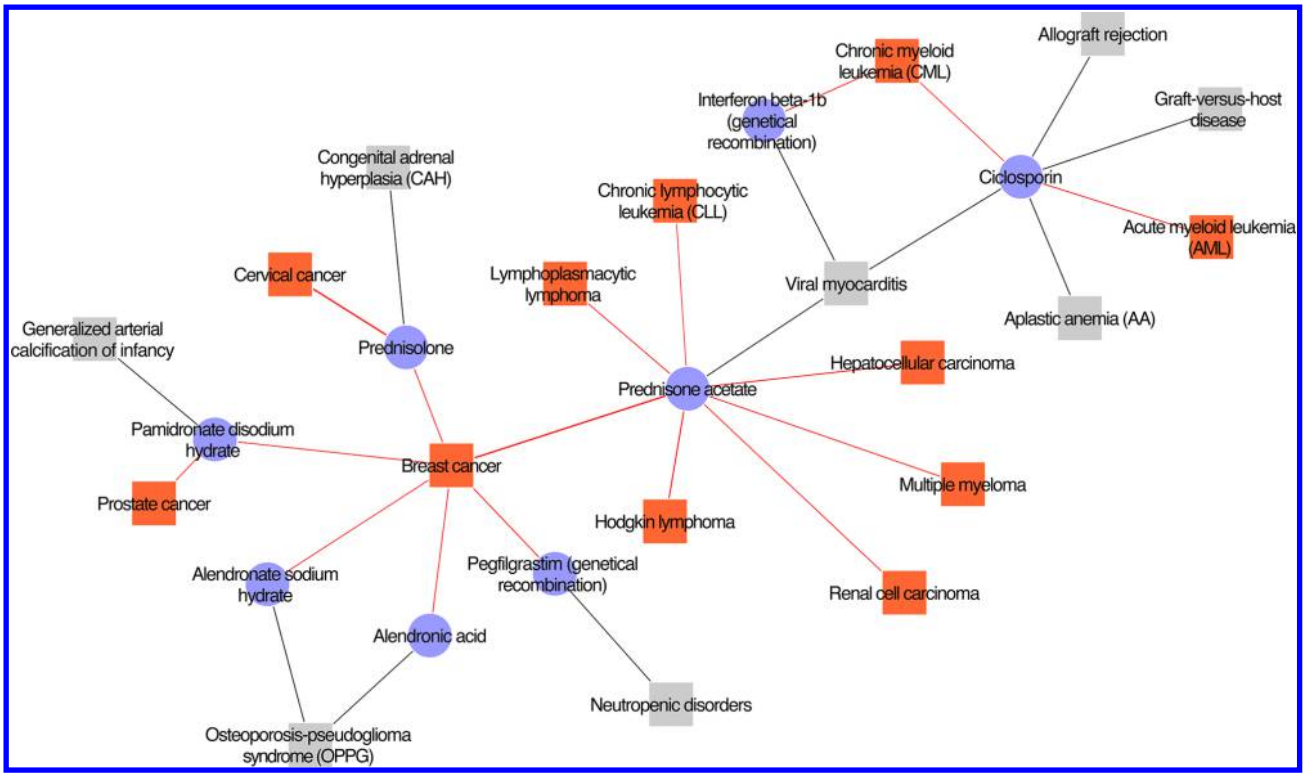


Figure 5. Drug−disease association network predicted for cancers. The network consists of known associations and predicted associations, in which edges of known drug−disease associations are colored in gray, and edges of newly predicted drug−disease associations are colored in red.

(involving 2300 drugs and 545 diseases) as potential drug indications. It took 36 min for training and 43 min for new prediction, where one core of two eight-core Xeon(R) CPUs E5-4617 (2.90 GHz) machine with 512GB memory was used, and the consumed memory was about 30GB.

We investigated the distribution of the repositioned drugs across different ICD-10 disease classes called "ICD chapters", and there are 22 ICD chapters in ICD-10. Table 1 shows the numbers of drugs repositioned from the original disease class to the other disease classes by the proposed method, in which the rows indicate the original ICD chapters, and the columns indicate the newly predicted ICD chapters. For example, the number of drugs that were repositioned from a chapter I disease to a chapter II disease was 128. The proposed method could predict drug indications for most ICD chapters (19 of 22 ICD chapters), suggesting that the proposed method can be applicable to most diseases defined in ICD-10. However, the method could not predict drug indications for chapters XV (pregnancy, childbirth, and puerperium [O00−O99]), XX (external causes of morbidity and mortality [V01−Y98]), and XXII (codes for special purposes [U00−U99]). For example, recurrent hydatidiform moles (H00289) of chapter XV and severe acute respiratory syndrome (H00402) of chapter XXII were not predicted because they did not have any molecular features in the database. The other chapter XX diseases had some disease features, but they were not predicted.

We compared the prediction results of our method with those of PREDICT and PreDR by obtaining the prediction results from the Supporting Information of previous papers.[19,21] Figure S6 of the Supporting Information shows a Venn diagram of the predicted drug−disease pairs for the proposed method, PREDICT, and PreDR. Most drug−disease associations predicted by our method were not predicted by the other methods. Our predicted associations overlap with 386 associations in PREDICT and 563 associations in PreDR, but the others (117,007 associations) were unique. This result suggests that the proposed method could predict many diseases that previous methods failed to predict. One explanation about the little overlap is that disease IDs are different between our proposed method and the previous methods. OMIM disease IDs (used in the previous methods) did not correspond to KEGG disease IDs (used in our proposed method) in the manner of one-to-one relationship, and the number of overlapped diseases with the same disease names was 83. This might have led to small overlap of the predicted drug−disease associations.

We also investigated the distribution of the repositioned drugs across different ICD-10 disease classes in the prediction results of PREDICT and PreDR (Tables S1 and S2, Supporting Information). Similar to the proposed method, chapter XX and XXII diseases were not predicted by PREDICT and PreDR. Because chapter XX and XXII diseases were not registered in the OMIM database, it was impossible to make predictions for these diseases using PREDICT and PreDR. Diseases of chapters X (diseases of the respiratory system [J00−J99]), XI (diseases of the digestive system [K00−K93]), XVI (certain conditions originating in the perinatal period [P00−P96]), XVIII (symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified [R00−R99]), and XXI (factors influencing health status and contact with health services [Z00−Z99]) were not predicted by PREDICT, but they were predicted by our proposed method. Diseases of chapters X, XI, XII (diseases of the skin and subcutaneous tissue [L00−L99]), XVI, and XVIII were not predicted by PreDR, but they were predicted by our proposed method.

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

Table 2. List of Drug−Disease Pairs for Cancers

rank	drug name	original disease name	predicted disease name	prediction score	reference
4	prednisone acetate	viral myocarditis	breast cancer	1.64	44
5	prednisolone	congenital adrenal hyperplasia	cervical cancer	1.59	45
10	prednisone acetate	viral myocarditis	Hodgkin lymphoma	1.24	46
13	pegfilgrastim	neutropenic disorders	breast cancer	1.08	47
18	prednisone acetate	viral myocarditis	hepatocellular carcinoma	1.02	48
19	prednisone acetate	viral myocarditis	lymphoplasmacytic lymphoma	0.99	49
20	prednisone acetate	viral myocarditis	renal cell carcinoma	0.98	50
21	interferon beta-1b	viral myocarditis	chronic myeloid leukemia	0.97	51
26	prednisone acetate	viral myocarditis	multiple myeloma	0.93	52
27	ciclosporin	allograft rejection, etc.	acute myeloid leukemia	0.93	53
28	pamidronate disodium hydrate	generalized arterial calcification of infancy	breast cancer	0.91	54
30	alendronate sodium hydrate	osteoporosis−pseudoglioma syndrome	breast cancer	0.88	55
31	alendronic acid	osteoporosis−pseudoglioma syndrome	breast cancer	0.88	55
32	pamidronate disodium hydrate	generalized arterial calcification of infancy	prostate cancer	0.86	56
33	prednisolone	congenital adrenal hyperplasia	breast cancer	0.86	44
38	ciclosporin	allograft rejection, etc.	chronic myeloid leukemia	0.77	57
40	prednisone acetate	viral myocarditis	chronic lymphocytic leukemia	0.76	58

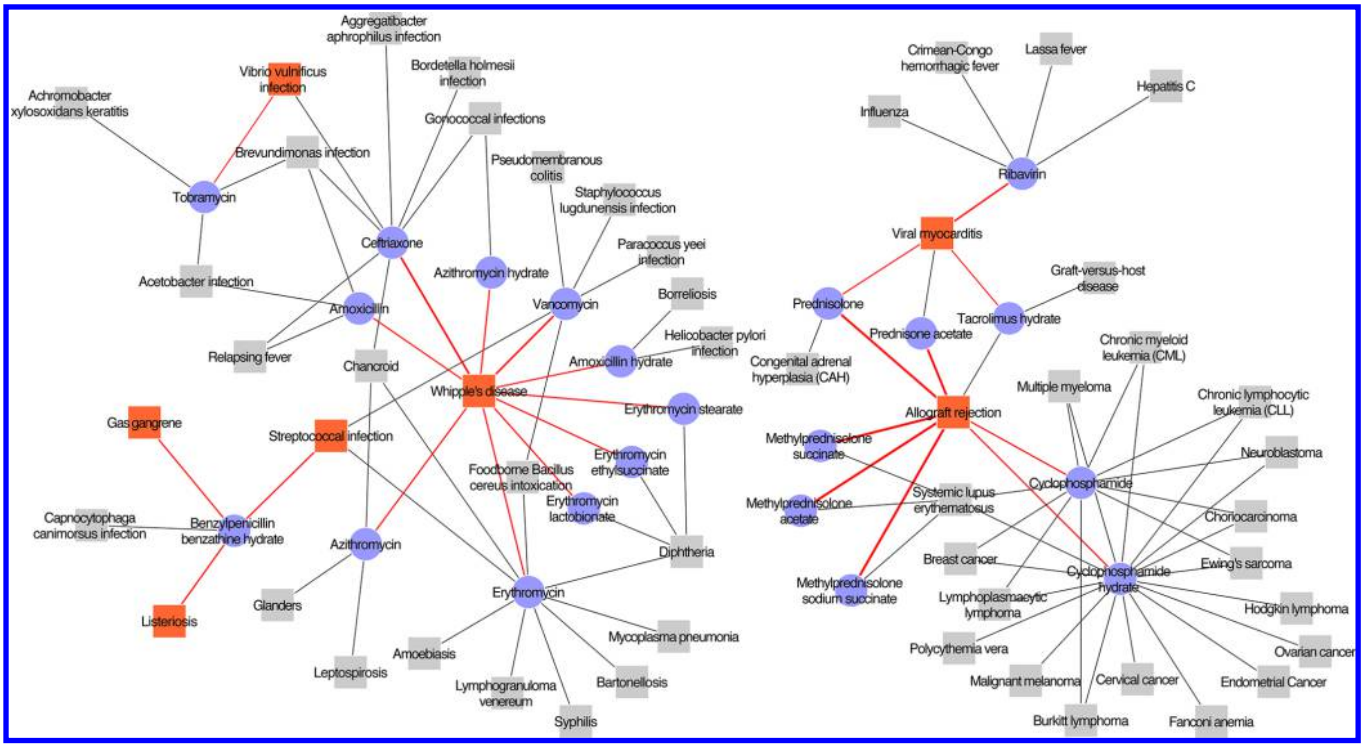


Figure 6. Drug−disease association network predicted for nongenetic disorders. The network consists of known associations and predicted associations, in which edges of known drug−disease associations are colored in gray, and edges of newly predicted drug−disease associations are colored in red.

These results suggest that the proposed method can predict drug indications for a wider range of diseases than other previous methods. Because of space limitations, we focused on the prediction results obtained only by our proposed method in the following sections.

Examples of Drugs Repositioned to Cancers. Drug therapy has not been established for many cancers, and thus, the discovery of new anticancer drugs will be of great benefit to patients with cancer. Most conventional anticancer drugs have cellular toxicity; therefore, they kill both cancerous and normal cells. Thus, the resulting adverse drug reactions (ADRs) represent a severe problem and a considerable burden for patients with cancer. The drug repositioning approach is expected to identify safer drugs that have potential anticancer activity from the list of drugs approved for other indications (excluding cancers). In this study, we analyzed drugs that were repositioned from the original indications to cancers (they were not anticancer agents but were predicted to work for cancers by our method). Among 117,880 predicted drug−disease associations, we selected 239 drug−disease pairs (consisting of 100 drugs and 46 diseases) involving drugs that were predicted to work for cancers in addition to their original indications (noncancers).

Figure 5 shows a predicted drug−disease association network, in which edges of known drug−disease associations are colored in gray, and edges of newly predicted drug−disease

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

Table 3. List of Drug−Disease Pairs for Nonhereditary Diseases

rank	drug name	original disease name	predicted disease name	prediction score	reference
1	methylprednisolone acetate	systemic lupus erythematosus	allograft rejection	3.45	59
2	methylprednisolone sodium succinate	systemic lupus erythematosus	allograft rejection	3.45	59
3	methylprednisolone succinate	systemic lupus erythematosus	allograft rejection	3.45	59
4	prednisone acetate	viral myocarditis	allograft rejection	3.27	60
5	prednisolone	congenital adrenal hyperplasia	allograft rejection	3.23	59
6	ceftriaxone	chancroid, etc.	Whipple's disease	2.79	61
12	vancomycin	foodborne *Bacillus cereus* intoxication, etc.	Whipple's disease	2.54	62
13	ribavirin	Crimean−Congo hemorrhagic fever, etc.	viral myocarditis	2.49	63
16	benzylpenicillin benzathine hydrate	*Capnocytophaga canimorsus* infection	Listeriosis	2.09	64
17	benzylpenicillin benzathine hydrate	*Capnocytophaga canimorsus* infection	streptococcal infection	2.09	65
18	erythromycin ethylsuccinate	diphtheria	Whipple's disease	2.05	66
19	erythromycin stearate	diphtheria	Whipple's disease	2.05	66
20	erythromycin lactobionate	diphtheria	Whipple's disease	2.05	66
21	erythromycin	chancroid, etc.	Whipple's disease	2.05	66
22	benzylpenicillin benzathine hydrate	*Capnocytophaga canimorsus* infection	gas gangrene	2.00	67
23	benzylpenicillin benzathine hydrate	*Capnocytophaga canimorsus* infection	tetanus	2.00	68
24	azithromycin	chancroid, etc.	Whipple's disease	1.99	69
25	azithromycin hydrate	gonococcal infections	Whipple's disease	1.99	69
26	cyclophosphamide	chronic myeloid leukemia, etc.	allograft rejection	1.99	70
27	cyclophosphamide hydrate	chronic myeloid leukemia, etc.	allograft rejection	1.99	70
31	amoxicillin hydrate	*Helicobacter pylori* infection	Whipple's disease	1.89	71
32	amoxicillin	relapsing fever, etc.	Whipple's disease	1.89	71
43	prednisolone	congenital adrenal hyperplasia	viral myocarditis	1.75	72
44	tacrolimus hydrate	allograft rejection, etc.	viral myocarditis	1.70	73
48	tobramycin	*Achromobacter xylosoxidans* keratitis, etc.	*Vibrio vulnificus* infection	1.61	74

associations are colored in red. We attempted to confirm the validity of the prediction results using independent resources. Of the top 50 predicted pairs, we were able to confirm the validity of 17 pairs (34%) in the literature.

Table 2 presents the list of drug−disease pairs that were predicted to have anticancer activity and were confirmed in the literature. Note that the confirmed drug−disease pairs were absent from the training set used for learning the predictive model. For example, prednisone acetate (D08416) is a drug originally developed for viral myocarditis (H00295), but it was predicted to work for breast cancer (H00031). Prednisolone (D00472) is a drug originally developed for congenital adrenal hyperplasia (H00216), but it was predicted to work for cervical cancer (H00030). Other examples of successful predictions are shown in Table 2. This result implies that our proposed method successfully reproduced recently reported drug effects.

Examples of Drugs Repositioned to Nonhereditary Diseases. In this study, we analyzed drugs that were repositioned to nonhereditary diseases (e.g., allograft rejection, Whipple's disease, and viral myocarditis) from the original indications (predicted for diseases absent from the OMIM database). Note that all previous methods on computational drug repositioning are based on disease information in the OMIM database, and therefore, previous methods tend to make predictions only for Mendelian hereditary disorders. In contrast, our proposed method is based on disease information in the KEGG database, which covers most diseases defined in ICD-10, and thus, the proposed method can make predictions for a wide range of diseases, including nonhereditary diseases. In this study nonhereditary diseases are defined as diseases absent from the OMIM database. Among 117,880 newly predicted drug−disease pairs, we selected 2004 drug−disease pairs (consisting of 512 drugs and 42 diseases) involving drugs that were predicted to work for nonhereditary diseases in addition to their original indications.

Figure 6 shows a predicted drug−disease association network, in which edges of known drug−disease associations are colored in gray, and edges of newly predicted drug−disease associations are colored in red. We attempted to confirm the validity of the prediction results using independent resources. Of the top 50 predicted pairs, we were able to confirm the validity of 25 pairs (50%) in the literature.

For example, methylprednisolone acetate (D00979) and methylprednisolone sodium succinate (D00751) are originally developed for systemic lupus erythematosus (H00080), but they were predicted to work for allograft rejection (H00083). This is a reasonable prediction because these drugs are the derivatives of methylprednisolone, a corticosteroid used to treat inflammation, also known to be used in allograft rejection. Ceftriaxone (D07659), a cephem antibiotic used for Gram-negative bacterial infections, and vancomycin (D00212), another antibacterial obtained from particular bacterial species, were repositioned to Whippel's disease (H00352), which is caused by another type of Gram-negative bacterium *Tropheryma whippeli*. Although a few other antibiotics have been used to treat this disease, our predicted results can be a good candidate for alternative drugs, especially in the case where resistance occurred in the known drugs. Other examples of successful predictions are shown in Table 3. Remarkably, most of the diseases predicted for repositioned drugs were infectious diseases, which would not be predicted by PREDICT and PreDR, because these methods used the OMIM database, which is limited to hereditary diseases. This implies that our proposed method provides a versatile system in predicting new drug uses in diseases including nonhereditary diseases.

Prediction Does Not Depend on Drug Chemical Structures. As an example of predicted drug indications, we

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

took the top prediction pair of prednisone acetate (D08416) and breast cancer (H00031) with the highest score for cancers in Table 2. Figure 7(A) shows the chemical structure of

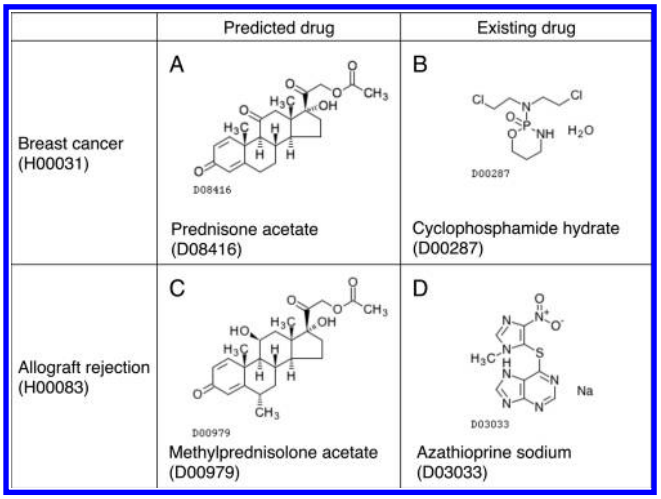


Figure 7. Chemical structure comparison between newly predicted drugs and existing drugs. Panel (A) shows the chemical structure of prednisone acetate (D08416), and panel (B) shows the chemical structure of cyclophosphamide hydrate (D00287), which is a known drug for breast cancer. Panel (C) shows the chemical structure of methylprednisolone acetate (D00979), and panel (D) shows the chemical structure of azathioprine sodium (D03033), which is a known drug for allograft rejection.

prednisone acetate (D08416), and Figure 7(B) shows the chemical structure of cyclophosphamide hydrate (D00287), which is a known drug for breast cancer (H00031). It appears that the chemical structure of prednisone acetate (D08416) is completely different from that of cyclophosphamide hydrate (D00287), and the corresponding Tanimoto coefficient is 0.090 (see Materials and Methods for more details).

As another example of a predicted drug indication, we took the top prediction pair of methylprednisolone acetate (D00979) and allograft rejection (H00083) with the highest score for nonhereditary diseases in Table 3. Figure 7(C) shows the chemical structure of methylprednisolone acetate (D00979), and Figure 7(D) shows the chemical structure of azathioprine sodium (D03033), which is a known drug for allograft rejection (H00083). It appears that the chemical structure of methylprednisolone acetate (D00979) is completely different from that of azathioprine sodium (D03033), and the corresponding Tanimoto coefficient is 0.059.

These predictions were not obtained by a simple chemical structure similarity search (a commonly used technique in chemoinformatics). These results suggest that the use of phenotypic profiles is useful for systematic drug repositioning, and it enables predictions for drug indications that were not expected from the chemical structures of drugs.

DISCUSSION

In this article, we presented a computational method for systematic drug repositioning based on drug phenotypic profiles and disease molecular profiles. The originality of the proposed method lies in applicability to a wide range of diseases, in the lack of dependence on drug chemical structures, and in the high-performance prediction with state-of-the-art machine learning methods. In fact, we demonstrated the usefulness of the proposed method via various CV experiments and described examples of successfully predicted drug indications for cancers and nonhereditary diseases.

A work related to this study is Yang's work,[20] which is also based on phenotypic effects of drugs. They used data on drug side effects in SIDER[34,35] and data on drug−disease associations in PharmGKB.[36] However, the number of drugs with side effect information in SIDER is limited (approximately 800 drugs). In addition, the confidence level of the learning set of drug−disease associations was 3 (the reproducibility of most drug indications was not fully confirmed), and the number of diseases was very small (only 102 diseases in their analysis), making it difficult for large-scale applications. Conversely, an advantage of our proposed method is the inclusion of a wide range of diseases for analysis.

Most previous works on computational drug repositioning are based on disease information in the OMIM database.[23] OMIM is the most well-known disease database, but there are two limitations in OMIM. First, it basically includes information only on Mendelian inheritance diseases. Most common diseases are non-Mendelian disorders, and their information is not stored in the OMIM database. Second, one disease can be assigned multiple disease IDs. For example, type 1 diabetes mellitus is assigned four disease IDs (OMIM: 601942, OMIM: 601388, OMIM: 125852, and OMIM: 612520), and they were regarded as different diseases in previous works. If multiple disease IDs corresponding to the same disease are split into a training set and a test set in the CV, then the resulting prediction accuracy does not reflect the difficulty in practical applications and may overestimate prediction accuracy. To overcome these problems, we used disease information in the KEGG database[25] in this study because the definition of diseases in KEGG is based on the definition of diseases in ICD-10.[24] Each disease is assigned a unique disease ID, and information on most diseases defined in ICD-10 is stored in KEGG, which enables analysis of non-Mendelian disorders in the drug repositioning framework.

The basic strategy of computational drug discovery is based on the assumption that similar drugs are likely to have similar targets and indications.[6] For example, a global network analysis of drug-therapy interactions revealed that drugs with similar ATC codes tended to have similar indications.[37] In this study, we used phenotypic similarity of drugs and molecular similarity of diseases, but there are many possibilities for data sources. Recently, many public databases have been established to store information on drugs and diseases, including DrugBank,[38] STITCH,[39] BindingDB,[40] MATADOR,[41] KEGG,[25] and ChEMBL.[42] These databases can be useful sources for drug repositioning, and most previous works are based on drug data and disease data stored in such public databases. Examples of drug data include chemical structures,[22] drug side effects,[18,20,43] target proteins or genes,[17,22] gene expressions,[11,12,14,15,22] and biological pathways.[16] Examples of disease data are phenotypes, target proteins or genes,[17] and gene expressions.[12,15] The state-of-the-art method features the integration of multiple different data sources for the same drugs or diseases,[19,21] and our proposed method was compared with this method in this study. However, all data (e.g., drug side effects, drug targets, disease pathways, and disease phenotypes) were not always available for the same drugs and diseases, and thus, the number of drugs/diseases with all data sources tends to be very small, limiting scalable applications. An important future work would be to

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

develop a useful method for handling missing data in data integration.

■ ASSOCIATED CONTENT

Supporting Information

Figure S1 provides the procedure of reconstructing benchmark data. Figure S2 provides clustering of drugs using the phenotypic similarity scores. Figure S3 provides clustering of diseases using the molecular similarity scores. Figures S4 and S5 provide the ROC and PR curves of the proposed method and previous methods. Figure S6 provides a Venn diagram of the predicted drug−disease pairs across the proposed method and previous methods. Tables S1 and S2 contain the distributions of drugs repositioned from the original disease class to other disease classes by the previous methods. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yamanishi@bioreg.kyushu-u.ac.jp. Phone: +81 92 642 6699. Fax: +81 92 642 6692.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Program to Disseminate Tenure Tracking System, MEXT, Japan, and Kyushu University Interdisciplinary Programs in Education and Projects in Research Development.

■ REFERENCES

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: New estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151−185.

(2) Ashburn, T. T.; Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673−683.

(3) Chong, C. R.; Sullivan, D. J. New uses for old drugs. *Nature* **2007**, *448*, 645−646.

(4) Hopkins, A. L.; Mason, J. S.; Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127−136.

(5) Nacher, J. C.; Schwartz, J. M. Modularity in protein complex and drug interactions reveals new polypharmacological properties. *PLoS One* **2012**, *7*, e30028.

(6) Csermely, P.; Korcsmáros, T.; Kiss, H. J.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery. A comprehensive review. *Pharmacol. Ther.* **2013**, *138*, 333−408.

(7) Novac, N. Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci.* **2013**, *34*, 267−272.

(8) Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. Keynote review: *In vitro* safety pharmacology profiling: an essential tool for successful drug development. *Drug discovery today* **2005**, *10*, 1421−1433.

(9) Blagg, J. Structure−activity relationships for in vitro and in vivo toxicity. *Annu. Rep. Med. Chem.* **2006**, *41*, 353.

(10) Terstappen, G. C.; Reggiani, A. *In silico* research in drug discovery. *Trends Pharmacol. Sci.* **2001**, *22*, 23−26.

(11) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929−1935.

(12) Hu, G.; Agarwal, P. Human disease-drug network based on genomic expression profiles. *PLoS One* **2009**, *4*, e6536.

(13) Chiang, A. P.; Butte, A. J. Systematic evaluation of drug−disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* **2009**, *86*, 507−510.

(14) Iorio, F.; Bosotti, R.; Scacheri, E.; Belcastro, V.; Mithbaokar, P.; Ferriero, R.; Murino, L.; Tagliaferri, R.; Brunetti-Pierri, N.; Isacchi, A.; Bernardo, D. d. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 14621−14626.

(15) Sirota, M.; Dudley, J. T.; Kim, J.; Chiang, A. P.; Morgan, A. A.; Sweet-Cordero, A.; Sage, J.; Butte, A. J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **2011**, *3*, 96ra77−96ra77.

(16) Ye, H.; Tang, K.; Yang, L.; Cao, Z.; Li, Y. Study of drug function based on similarity of pathway fingerprint. *Protein & cell* **2012**, *3*, 132−139.

(17) Zhao, S.; Li, S. A co-module approach for elucidating drug−disease associations and revealing their molecular basis. *Bioinformatics* **2012**, *28*, 955−961.

(18) Ye, H.; Liu, Q.; Wei, J. Construction of drug network based on side effects and its application for drug repositioning. *PLoS One* **2014**, *9*, e87864.

(19) Gottlieb, A.; Stein, G. Y.; Ruppin, E.; Sharan, R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 496.

(20) Yang, L.; Agarwal, P. Systematic drug repositioning based on clinical side-effects. *PLoS One* **2011**, *6*, e28025.

(21) Wang, Y.; Chen, S.; Deng, N.; Wang, Y. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* **2013**, *8*, e78518.

(22) Napolitano, F.; Zhao, Y.; Moreira, V. M.; Tagliaferri, R.; Kere, J.; D'Amato, M.; Greco, D. Drug repositioning: A machine-learning approach through data integration. *J. Cheminformatics* **2013**, *5*, 30.

(23) Hamosh, A.; Scott, A. F.; Amberger, J.; Bocchini, C.; Valle, D.; McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **2002**, *30*, 52−55.

(24) *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*; World Health Organization: Geneva, 1992.

(25) Kanehisa, M.; Goto, S.; Furumichi, M.; Tanabe, M.; Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **2010**, *38*, D355−D360.

(26) Chen, B.; Wild, D.; Guha, R. PubChem as a source of polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044−2055.

(27) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, *36*, D480−D484.

(28) Papadakis, M. A.; McPhee, S. J.; Rabow, M. W. *CURRENT Medical Diagnosis and Treatment 2014*; McGraw-Hill Medical: New York, 2013.

(29) Tabei, Y.; Pauwels, E.; Stoven, V.; Takemoto, K.; Yamanishi, Y. Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. *Bioinformatics* **2012**, *28*, i487−i494.

(30) Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; Lin, C.-J. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871−1874.

(31) Takarabe, M.; Kotera, M.; Nishimura, Y.; Goto, S.; Yamanishi, Y. Drug target prediction using adverse event report systems: A pharmacogenomic approach. *Bioinformatics* **2012**, *28*, i611−i618..

(32) Iwata, H.; Mizutani, S.; Tabei, Y.; Kotera, M.; Goto, S.; Yamanishi, Y. Inferring protein domains associated with drug side effects based on drug-target interaction network. *BMC Syst. Biol.* **2013**, *7*, 1−11.

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

(33) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corporation: New York, 1958.
(34) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263−266.
(35) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **2010**, *6*, 343.
(36) Whirl-Carrillo, M.; McDonagh, E.; Hebert, J.; Gong, L.; Sangkuhl, K.; Thorn, C.; Altman, R.; Klein, T. E. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **2012**, *92*, 414−417.
(37) Nacher, J. C.; Schwartz, J.-M. A global view of drug-therapy interactions. *BMC pharmacology* **2008**, *8*, 5.
(38) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.
(39) Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T. H.; von Mering, C.; Jensen, L. J.; Bork, P. STITCH 4: Integration of protein−chemical interactions with user data. *Nucleic Acids Res.* **2013**, *42*, D401−D407.
(40) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(41) Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E. G.; Gewiess, A.; Jensen, L. J.; Schneider, R.; Skoblo, R.; Russell, R. B.; Bourne, P. E.; Bork, P.; Preissner, R. SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **2008**, *36*, D919−D922.
(42) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(43) Lin, S. F.; Xiao, K. T.; Huang, Y. T.; Chiu, C. C.; Soo, V. W. Analysis of adverse drug reactions using drug and drug target interactions and graph-based methods. *Artif. Intell. Med.* **2010**, *48*, 161−166.
(44) Wong, N. S.; Buckman, R. A.; Clemons, M.; Verma, S.; Dent, S.; Trudeau, M. E.; Roche, K.; Ebos, J.; Kerbel, R.; DeBoer, G. E.; Sutherland, D. J.; Emmenegger, U.; Slingerland, J.; Gardner, S.; Pritchard, K. I. Phase I/II trial of metronomic chemotherapy with daily dalteparin and cyclophosphamide, twice-weekly methotrexate, and daily prednisone as therapy for metastatic breast cancer using vascular endothelial growth factor and soluble vascular endothelial growth factor receptor levels as markers of response. *J. Clin. Oncol.* **2010**, *28*, 723−730.
(45) Nozawa, S.; Engvall, E.; Kano, S.; Kurihara, S.; Fishman, W. H. Sodium butyrate produces concordant expression of "early placental" alkaline phosphatase, pregnancy-specific beta1-glycoprotein and human chorionic gonadotropin beta-subunit in a newly established uterine cervical cancer cell line (SKG-IIIa). *Int. J. Cancer* **1983**, *32*, 267−272.
(46) Böll, B.; Bredenfeld, H.; Görgen, H.; Halbsguth, T.; Eich, H. T.; Soekler, M.; Markova, J.; Keller, U.; Graeven, U.; Kremers, S.; Geissler, M.; Trenn, G.; Fuchs, M.; Tresckow, B. v.; Eichenauer, D. A.; Borchmann, P.; Engert, A. Phase 2 study of PVAG (prednisone, vinblastine, doxorubicin, gemcitabine) in elderly patients with early unfavorable or advanced stage Hodgkin lymphoma. *Blood* **2011**, *118*, 6292−6298.
(47) Yardley, D. A.; Zubkus, J.; Daniel, B.; Inhorn, R.; Lane, C. M.; Vazquez, E. R.; Naot, Y.; Burris, H. A., III; Hainsworth, J. D. A phase II trial of dose-dense neoadjuvant gemcitabine, epirubicin, and albumin-bound paclitaxel with pegfilgrastim in the treatment of patients with locally advanced breast cancer. *Clin. Breast Cancer* **2010**, *10*, 367−372.
(48) Moore, C. M.; Lamzabi, I.; Bartels, A. K.; Jakate, S.; Van Thiel, D. H. Colonic diffuse large B-cell lymphoma in a liver transplant patient with historically very low tacrolimus levels. *Case Rep. Transplant.* **2012**, *2012*, 952359.
(49) Buske, C.; Hoster, E.; Dreyling, M.; Eimermacher, H.; Wandt, H.; Metzner, B.; Fuchs, R.; Bittenbring, J.; Woermann, B.; Hohloch, K.; Hess, G.; Ludwig, W. D.; Schimke, J.; Schmitz, S.; Kneba, M.; Reiser, M.; Graeven, U.; Klapper, W.; Unterhalt, M.; Hiddemann, W. The addition of rituximab to front-line therapy with CHOP (R-CHOP) results in a higher response rate and longer time to treatment failure in patients with lymphoplasmacytic lymphoma: results of a randomized trial of the German Low-Grade Lymphoma Study Group (GLSG). *Leukemia* **2008**, *23*, 153−161.
(50) Haarstad, H.; Jacobsen, A. B.; Schjølseth, S. A.; Risberg, T.; Fosså, S. D. Interferon-α, 5-FU and prednisone in metastatic renal cell carcinoma: A phase II study. *Ann. Oncol.* **1994**, *5*, 245−248.
(51) Legros, L.; Guilhot, J.; Huault, S.; Mahon, F.; Preudhomme, C.; Guilhot, F.; Hueber, A. Interferon decreases VEGF levels in patients with chronic myeloid leukemia treated with imatinib. *Leuk. Res.* **2014**, *38*, 662−665.
(52) Kim, M. Y.; Sposto, R.; Swaika, A.; Asano, H.; Alamgir, A.; Chanan Khan, A.; Ailawadhi, S. Pharmacoeconomic implications of lenalidomide maintenance therapy in multiple myeloma. *Oncology* **2014**, *87*, 224−231.
(53) Chauncey, T. R.; Gundacker, H.; Shadman, M.; List, A. F.; Dakhil, S. R.; Erba, H. P.; Slovak, M. L.; Chen, I.; Willman, C. L.; Kopecky, K. J.; Appelbaum, F. R. Sequential phase II Southwest Oncology Group studies (S0112 and S0301) of daunorubicin and cytarabine by continuous infusion, without and with ciclosporin, in older patients with previously untreated acute myeloid leukaemia. *Br. J. Hamaetol.* **2010**, *148*, 48−58.
(54) Newcomb, P.; Trentham-Dietz, A.; Hampton, J. Bisphosphonates for osteoporosis treatment are associated with reduced breast cancer risk. *Br. J. Cancer* **2010**, *102*, 799−802.
(55) Hue, T. F.; Cummings, S. R.; Cauley, J. A.; Bauer, D. C.; Ensrud, K. E.; Barrett-Connor, E.; Black, D. M. Effect of bisphosphonate use on risk of postmenopausal breast cancer: Results from the randomized clinical trials of alendronate and zoledronic acid. *JAMA Intern. Med.* **2014**, *174*, 1550−1557.
(56) Lee, M. V.; Fong, E. M.; Singer, F. R.; Guenette, R. S. Bisphosphonate treatment inhibits the growth of prostate cancer cells. *Cancer Res.* **2001**, *61*, 2602−2608.
(57) Blake, S. J.; Hughes, T. P.; Lyons, A. B. Drug-interaction studies evaluating T-cell proliferation reveal distinct activity of dasatinib and imatinib in combination with cyclosporine A. *Exp. Hematol.* **2012**, *40*, 612−621.
(58) Schmidt, C.; Fetscher, S.; Görg, C.; Kornek, P.; Nusch, A.; Kegel, T.; Kellermann, L.; Hiddemann, W.; Fingerle-Rowson, G.; Dreyling, M. Treatment of indolent lymphoma in Germany: Results of a representative population-based survey. *Clin. Lymphoma, Myeloma Leuk.* **2011**, *11*, 204−211.
(59) Costa, D.; De Castro, R.; Kara-Jose, N. Case-control study of subconjunctival triamcinolone acetonide injection vs intravenous methylprednisolone pulse in the treatment of endothelial corneal allograft rejection. *Eye* **2008**, *23*, 708−714.
(60) Roberti, I.; Geffner, S.; Vyas, S. Successful rescue of refractory acute antibody-mediated renal allograft rejection with splenectomy: A case report. *Pediatr. Transplant.* **2012**, *16*, E49−E52.
(61) Feurle, G. E.; Moos, V.; Bläker, H.; Loddenkemper, C.; Moter, A.; Stroux, A.; Marth, T.; Schneider, T. Intravenous ceftriaxone, followed by 12 or three months of oral treatment with trimethoprim-sulfamethoxazole in Whipple's disease. *J. Infect.* **2013**, *66*, 263−270.
(62) Boulos, A.; Rolain, J. M.; Raoult, D. Antibiotic susceptibility of Tropheryma whipplei in MRC5 cells. *Antimicrob. Agents Chemother.* **2004**, *48*, 747−752.
(63) Kalimuddin, S.; Sessions, O. M.; Hou, Y.; Ooi, E. E.; Sim, D.; Cumaraswamy, S.; Tan, T. E.; Lai, S. H.; Low, C. Y. Successful clearance of human parainfluenza virus type 2 viraemia with intravenous ribavirin and immunoglobulin in a patient with acute myocarditis. *J. Clin. Virol.* **2013**, *56*, 37−40.

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459

(64) Penduka, D.; Mosa, R.; Simelane, M.; Basson, A.; Okoh, A.; Opoku, A. Evaluation of the anti-Listeria potentials of some plant-derived triterpenes. *Ann. Clin. Microbiol. Antimicrob.* **2014**, *13*, 37.
(65) Chhetri, U.; Shrestha, S.; Pradhan, R.; Shrestha, A.; Adhikari, N.; Thorson, S.; Pollard, A.; Murdoch, D.; Kelly, D. Clinical profile of invasive pneumococcal disease in Patan Hospital, Nepal. *Kathmandu Univ. Med. J.* **2012**, *9*, 45−49.
(66) Hawkins, C.; Farr, M.; Morris, C.; Hoare, A.; Williamson, N. Detection by electron microscope of rod-shaped organisms in synovial membrane from a patient with the arthritis of Whipple's disease. *Ann. Rheum. Dis.* **1976**, *35*, 502−509.
(67) Takahira, N.; Shindo, M.; Tanaka, K.; Soma, K.; Ohwada, T.; Itoman, M. Treatment outcome of nonclostridial gas gangrene at a level 1 trauma center. *J. Orthop. Trauma* **2002**, *16*, 12−17.
(68) Ganesh Kumar, A.; Kothari, V.; Krishnan, A.; Karnad, D. Benzathine penicillin, metronidazole and benzyl penicillin in the treatment of tetanus: A randomized, controlled trial. *Ann. Trop. Med. Parasitol.* **2004**, *98*, 59−63.
(69) Dearment, M. C.; Woodward, T. A.; Menke, D. M.; Brazis, P. W.; Bancroft, L. W.; Persellin, S. T. Whipple's disease with destructive arthritis, abdominal lymphadenopathy, and central nervous system involvement. *J. Rheumatol.* **2003**, *30*, 1347−1350.
(70) Kanakry, C. G.; Ganguly, S.; Zahurak, M.; Bolaños-Meade, J.; Thoburn, C.; Perkins, B.; Fuchs, E. J.; Jones, R. J.; Hess, A. D.; Luznik, L. Aldehyde dehydrogenase expression drives human regulatory T cell resistance to posttransplantation cyclophosphamide. *Sci. Transl. Med.* **2013**, *5*, 211ra157−211ra157.
(71) Bureš, J.; Kopáčová, M.; Douda, T.; Bártová, J.; Tomš, J.; Rejchrt, S.; Tachecí, I. Whipple's disease: Our own experience and review of the literature. *Gastroenterol. Res. Pract.* **2013**, *2013*.
(72) Aziz, K. U.; Patel, N.; Sadullah, T.; Tasneem, H.; Thawerani, H.; Talpur, S. Acute viral myocarditis: Role of immunosuppression: A prospective randomised study. *Cardiol. Young* **2010**, *20*, 509−515.
(73) Kanda, T.; Nagaoka, H.; Kaneko, K.; Wilson, J. E.; McManus, B. M.; Imai, S.; Suzuki, T.; Murata, K.; Kobayashi, I. Synergistic effects of tacrolimus and human interferon-alpha A/D in murine viral myocarditis. *J. Pharmacol. Exp. Ther.* **1995**, *274*, 487−493.
(74) Chuang, Y. G.; Young, C.; Chen, C. W. *Vibrio vulnificus* infection. *Scand. J. Infect. Dis.* **1989**, *21*, 721−726.

DOI: 10.1021/ci500670q
J. Chem. Inf. Model. 2015, 55, 446−459