JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Drug−Disease Association and Drug-Repositioning Predictions in Complex Diseases Using Causal Inference−Probabilistic Matrix Factorization

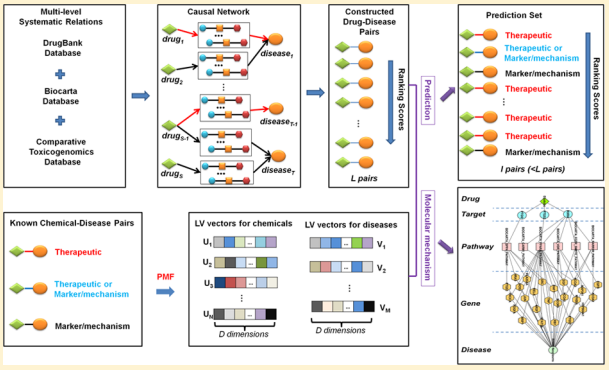
Jihong Yang,[†,§] Zheng Li,*[,‡,§] Xiaohui Fan,[†] and Yiyu Cheng*[,†]

[†]Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[‡]State Key Laboratory of Modern Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 300193, China

Supporting Information

ABSTRACT: The high incidence of complex diseases has become a worldwide threat to human health. Multiple targets and pathways are perturbed during the pathological process of complex diseases. Systematic investigation of complex relationship between drugs and diseases is necessary for new association discovery and drug repurposing. For this purpose, three causal networks were constructed herein for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively. A causal inference-probabilistic matrix factorization (CI-PMF) approach was proposed to predict and classify drug−disease associations, and further used for drug-repositioning predictions. First, multilevel systematic relations between drugs and diseases were integrated from heterogeneous databases to construct causal networks connecting drug−target−pathway−gene−disease. Then, the association scores between drugs and diseases were assessed by evaluating a drug's effects on multiple targets and pathways. Furthermore, PMF models were learned based on known interactions, and associations were then classified into three types by trained models. Finally, therapeutic associations were predicted based upon the ranking of association scores and predicted association types. In terms of drug−disease association prediction, modified causal inference included in CI-PMF outperformed existing causal inference with a higher AUC (area under receiver operating characteristic curve) score and greater precision. Moreover, CI-PMF performed better than single modified causal inference in predicting therapeutic drug−disease associations. In the top 30% of predicted associations, 58.6% (136/232), 50.8% (31/61), and 39.8% (140/352) hit known therapeutic associations, while precisions obtained by the latter were only 10.2% (231/2264), 8.8% (36/411), and 9.7% (189/1948). Clinical verifications were further conducted for the top 100 newly predicted therapeutic associations. As a result, 21, 12, and 32 associations have been studied and many treatment effects of drugs on diseases were investigated for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively. Related chains in causal networks were extracted for these 65 clinical-verified associations, and we further illustrated the therapeutic role of etodolac in breast cancer by inferred chains. Overall, CI-PMF is a useful approach for associating drugs with complex diseases and provides potential values for drug repositioning.



1. INTRODUCTION

Complex diseases, such as cancers, diabetes mellitus, and cardiovascular diseases, are caused by a combination of genetic and environmental factors.[1] These diseases have become continuing global threats for their increasing prevalence and high mortality rate. According to Heron's report,[2] cardiovascular diseases, diabetes mellitus, and malignant neoplasms are all within the top 10 leading causes of death in the United States. There is a pressing need to develop effective drugs for complex diseases.

Drug repositioning has been used as an alternative strategy for drug development[3] and investigation of drug−disease association for new indication discovery. However, with respect to complex diseases induced by genes' collective abnormalities[4] and polypharmacological profiles of drugs,[5,6] it is difficult to investigate the complicated relationships. Computational methods provide a promising avenue for systematic investigation of these relationships. Two main strategies—namely, drug-based and disease-based computational methods—have been developed to exploit drug−disease associations for drug repositioning.[7] Chemical similarity, molecular activity similarity, and molecular docking are used in the former strategy, while the latter makes use of associative indication transfer, shared molecular pathology, and side effect similarity to measure drug-disease association. Moreover, some other approaches, such as predictive toxicogenomic-derived model,[8] network propagation,[9] pathway-based Bayesian inference,[10] and network-based

Received: June 9, 2014
Published: August 13, 2014

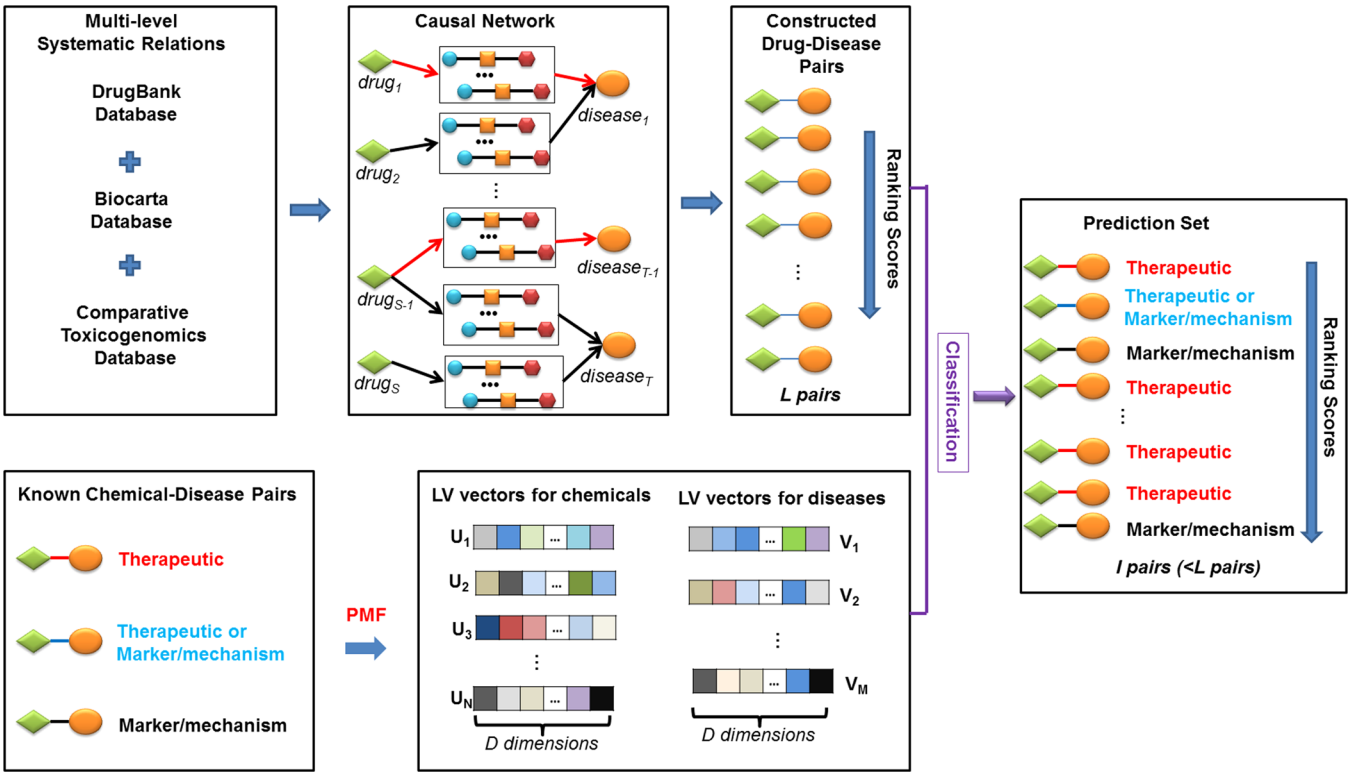


Figure 1. Scheme of CI-PMF approach. The drug-target-pathway-gene-disease causal network is first constructed for calculating the association score (S_a), and observed chemical-disease associations form CTD are used for learning PMF models. Then, constructed PMF model is used to classify constructed associations into different groups, namely, "therapeutic", "marker/mechanism", and "therapeuticlmarker/mechanism". Finally, therapeutic drug−disease association is predicted based on ranking of association score and predicted type. [Symbol legend: diamond, drug; circle, target; foursquare, pathway; hexagon, disease-associated gene; and ellipse, disease. Red arrow is only for chains concluding known chemical−disease pairs from CTD.]

inference,[11] have been proposed to infer drug-disease associations.

However, few methods unveil the underlying mechanisms for predicted associations.[10,12] It is of critical importance to bridge the molecular effect of drug and disease phenotypes.[4] Therefore, we argue that it is likely to predict more accurate associations by taking all relevant biological processes into consideration. In addition, drugs are not always therapeutic, as they may only correlate with or play roles in the etiology of the disease. Thus, identification of therapeutic associations is critical for drug repositioning. Based on this, we developed a new approach, causal inference-probabilistic matrix factorization (CI-PMF), for three purposes: (i) predicting drug-disease associations, (ii) inferring new drug repositioning, and (iii) uncovering molecular mechanisms underlying drug-disease associations. The relatedness between drugs and diseases was measured by association score (Figure 1), which was determined by investigating all possible chains for each drug−disease association in the causal network. On the other hand, PMF was applied to summarize known interactions and train models for type classification. Therapeutic drug−disease pairs were then predicted on the basis of the ranking of association score and predicted type. Moreover, molecular pathways connecting drugs and diseases were conveniently used to interpret the underlying mechanisms.

2. MATERIALS AND METHODS

2.1. Constructing Weighted Causal Network. As shown in Figure 1, a causal network was constructed by evaluating all chains from drugs to diseases. Five layers of nodes were included in the network, namely, drugs, targets, pathways, genes, and diseases. Causal links between every two layers from left to right represent (i) drug's act on targets; (ii) influence of drug targets on pathways; (3) influence of pathways on genes; (4) genes' association with diseases. Heterogeneous resources were collected for network construction. Approved drugs and their targets were collected from DrugBank database (V3.0),[13] with species of target restricted to "Homo sapiens". Target-involved pathways and pathway-related genes were from Biocarta of MsigDB (v4.0) Pathway ontology. Terms of neoplasms, diabetes, and cardiovascular diseases were extracted from the MeSH database and their associated genes from Comparative Toxicogenomics Database (CTD)[14] with direct evidence to be "marker/mechanism" or "therapeutic", as of December 6, 2013. The initial weight for each connection was 1, and additional weight was added for connections involved in enriched chains (C*), where the primal drug and final disease form a known association assembled from CTD. Chemical−disease interaction types without "therapeutic" or "marker/mechanism" were removed in this study. The final transition weight was calculated as follows:

$$W(i \to j) = \begin{cases} 1 + \dfrac{P^*(i \to j)}{P(i \to j)} & \text{if } i \to j \text{ exists in C}^* \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

dx.doi.org/10.1021/ci500340n | J. Chem. Inf. Model. 2014, 54, 2562−2569

where $P(i \rightarrow j)$ and $P^*(i \rightarrow j)$ represent the transition probability of connection from node i to node j $(i \rightarrow j)$ in complete chains and enriched chains, respectively.

Both directed graphs containing all chains or enriched chains can be regarded as two Markov models, and maximum likelihood estimation was used to compute the transition probability as follows:

$$P(i \rightarrow j) = \frac{N(i \rightarrow j)}{N(i \rightarrow \bullet)} \tag{2}$$

Here, $N(i \rightarrow j)$ is the number of paths from i to j in a given graph, and $N(i \rightarrow \bullet)$ is the total number of connections originated from i.

Each connection of each chain was assigned a transition weight. Weighted causal networks of cardiovascular diseases, diabetes mellitus, and neoplasms were constructed.

2.2. Computing Association Score. Based on the constructed weighted causal network, multiple chains might relate to one drug−disease association and the likelihood (L_c) of each chain was computed using the following equation:

$$L_c = W_{d\rightarrow t} \times W_{t\rightarrow p} \times W_{p\rightarrow g} \times W_{g\rightarrow dis} \tag{3}$$

$W_{d\rightarrow t}$, $W_{t\rightarrow p}$, $W_{p\rightarrow g}$, $W_{g\rightarrow dis}$ are precomputed transition weights for drug to target, target to pathway, pathway to disease associated gene, and gene to disease, respectively.

Different from previous work,[15] association scores S_a between drugs and diseases were calculated using the mean likelihood of all possible chains in this study:

$$S_a(x, y) = \text{mean}(L_{C(x\rightarrow y)}) \tag{4}$$

x represents drug x, and y represents disease y, $L_{C(x\rightarrow y)}$ are likelihoods of all chains starting from drug x and ending with disease y.

2.3. Learning Probabilistic Matrix Factorization Model. Constructed causal networks predict systematic drug−disease associations. However, therapeutic associations could not be distinguished from the large number of predicted associations. To overcome this limit, probabilistic matrix factorization (PMF)[16,17] was used for association-type classification through the analysis of known information on chemical disease association type. PMF is a factor-based model for collaborative filtering; it does not require complete information on the drug−target interactions, pathway−gene interactions, and disease−gene interactions used in the causal network to make accurate predictions.

Chemical−disease association-type information was extracted from CTD; integer rating values of 3, 2, and 1 represents "therapeutic", "therapeuticlmarker/mechanism", and "marker/mechanism" associations, respectively. Then, the bipartite graph of chemical−disease interaction was converted to a target matrix, $R_{N\times M}$, for N chemicals and M diseases. R_{ij} is 0 when drug i is not associated with disease j; otherwise, the integer rating value is set as 1. Two latent variable (LV) matrixes, $U^T_{N\times D}$ and $V_{D\times M}$, should be trained for the matrix. Thus, each chemical/disease can be expressed by a D-dimensional LV.

A probabilistic linear model with Gaussian observation noise was taken for modeling the association. The conditional distribution over the observed associations was presented as

$$p(R|U, V, \sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}[f(R_{ij}|U_i^T V_j, \sigma^2)]^{I_{ij}} \tag{5}$$

where the probability density function $f(x|\mu,\sigma^2)$ represents for Gaussian distribution with mean μ and variance σ^2. The indicator function, I_{ij}, is equal to 1 if the association is known and 0 otherwise. U_i^T is the transpose of U_i.

In addition, zero-mean spherical Gaussian priors was placed on LVs:

$$p(U|{\sigma_U}^2) = \prod_{i=1}^{N} f(U_i|0, {\sigma_U}^2 I) \tag{6}$$

$$p(V|{\sigma_V}^2) = \prod_{j=1}^{M} f(V_j|0, {\sigma_V}^2 I) \tag{7}$$

Thus leading to a log-likelihood plot of chemical and disease features, given by

$$\ln p(U, V|R, \sigma^2, {\sigma_U}^2, {\sigma_V}^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - U_i^T V_j)^2 - \frac{1}{2{\sigma_U}^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2{\sigma_V}^2}\sum_{j=1}^{M} V_j^T V_j + C \tag{8}$$

Here, C is a constant. Training optimal models must find D-dimensional vectors, U_i and V_j, to maximize the function with fixed observation noise variance and prior variances. Thus, it is equivalent to minimizing the following objective function:

$$E = \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{1}{2{\sigma_U}^2}\sum_{i=1}^{N} U_i^T U_i + \frac{1}{2{\sigma_V}^2}\sum_{j=1}^{M} V_j^T V_j \tag{9}$$

The first term is the squared error, and the two following terms are trained for regularization, which forces the model to make default no-interaction predictions. Finally, optimal LVs are trained for each chemical and each disease.

2.4. Association Type Classification. When chemical i−disease j is an association in the causal network, a score (S_p) of the association will be calculated by

$$S_{p_{(i,j)}} = U_i^T V_j + \mu' \tag{10}$$

Here, μ' represents the mean value of rating for all known associations. Association type was then predicted by

$$\text{type}_{pre} = \begin{cases} \text{marker/mechanism} & \text{if } S_p < 1 \\ \text{marker/mechanism} & \text{if round}(S_p) = 1 \\ \text{therapeuticlmarker/mechanism} & \text{if round}(S_p) = 2 \\ \text{therapeutic} & \text{if round}(S_p) = 3 \\ \text{therapeutic} & \text{if } S_p > 3 \end{cases} \tag{11}$$

2.5. Criteria for New Therapeutic Drug−Disease Association. A new therapeutic association between a drug and a disease is predicted based on both association type and ranking of association score. Associations with high association scores ranking and predicted "therapeutic" type are more likely to be used for drug repositioning predictions.

dx.doi.org/10.1021/ci500340n | J. Chem. Inf. Model. 2014, 54, 2562−2569

Table 1. Detailed Information of Constructed Causal Networks

	Cardiovascular Diseases		Diabetes Mellitus		Neoplasms	
parameter[a]	complete chain	enriched chain	complete chain	enriched chain	complete chain	enriched chain
chains	114840	8616	19680	428	133022	4849
No. of drugs	726	209	566	25	721	90
No. of targets	280	92	245	29	274	72
No. of pathways	207	168	146	69	200	145
No. of genes	286	223	80	53	307	220
No. of diseases	97	62	12	10	72	43
|drug-target|	1472	318	1007	37	1461	148
|target−pathway|	1348	539	1043	145	1319	372
|pathway−gene|	1665	1112	417	146	1662	1089
|gene−disease|	736	560	141	88	835	552
drug−disease pairs	19103	1058	3290	43	15650	291

[a] |*| denotes the number of associations.

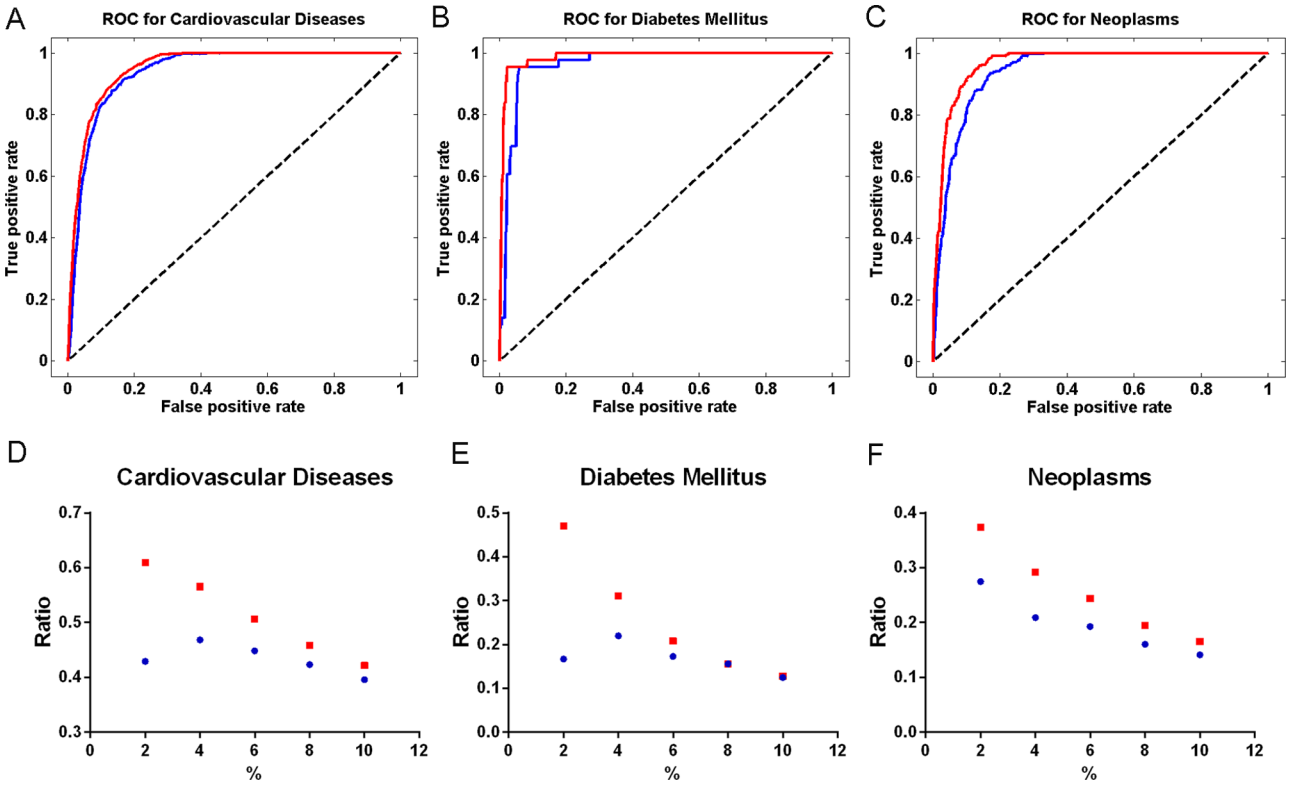


Figure 2. Comparison between existing causal inference and modified causal inference: ROC curves of drug-disease association predictions (A, B, C), and ratios of known drug-disease associations in top K genes (D, E, F). Red traces represent results obtained by modified causal inference, and blue traces represent results obtained by existing causal inference.

3. RESULTS

3.1. Weighted Causal Network. Weighted causal networks were constructed for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively. Detailed statistics are described in Table 1. Take cardiovascular diseases for an example; 114 840 chains are constructed, 19 103 unique drug−disease pairs could be investigated by above chains, in which 726 drugs and 97 disease terms are included. Furthermore, the transition weight of each connection in the network was calculated to obtain a likelihood score for each chain. Scores corresponding to cardiovascular diseases, diabetes mellitus, and neoplasms ranged from 0 to 4.8655, 0 to 4.6138, and 0 to 5.2309, respectively.

3.2. Comparison between Existing Causal Inference and Modified Causal Inference. In the present study, only known associations were treated as positive instances. The true positive rate and the false positive rate were calculated at different ranking scores. AUC (area under receiver operating characteristic curve) scores were obtained from these ROC (receiver operating characteristic) curves. We compared modified causal inference with the existing method in predicting drug−disease associations. AUC scores obtained with mean likelihood and maximal likelihood were 0.9507 vs 0.9380 for cardiovascular diseases, 0.9872 vs 0.9642 for diabetes mellitus, and 0.9660 vs 0.9413 for neoplasms, respectively (Figure 2). In addition, precision comparison was also conducted for these two methods. The top 2%, 4%, 6%, 8%, and 10% of predicted associations were collected, and ratios of

dx.doi.org/10.1021/ci500340n | J. Chem. Inf. Model. 2014, 54, 2562−2569

Table 2. 65 Clinical-Verified Associations

condition	intervention	condition	intervention	condition	intervention
heart failure	riboflavin	diabetic nephropathies	fosinopril	breast neoplasms	lidocaine
heart failure	clenbuterol	diabetes mellitus	vitamin E	breast neoplasms	naltrexone
hypertension	allopurinol	diabetes mellitus, type 2	fenofibrate	ovarian neoplasms	simvastatin
hypertension	fenofibrate	diabetes mellitus, type 2	gemfibrozil	breast neoplasms	melatonin
heart failure	allopurinol	diabetic neuropathies	acetaminophen	breast neoplasms	theophylline
hypertension	minocycline	diabetes mellitus, type 2	nifedipine	ovarian neoplasms	lovastatin
heart failure	perindopril	diabetes mellitus	fenofibrate	breast neoplasms	sirolimus
hypertension	iloprost	diabetes mellitus	gemfibrozil	breast neoplasms	etodolac
myocardial infarction	fenofibrate	diabetic angiopathies	vitamin E	melanoma	indomethacin
arrhythmias, cardiac	vitamin E	diabetes mellitus, type 2	cyclosporine	leukemia	epirubicin
cardiomegaly	ramipril	diabetes mellitus, type 2	losartan	leukemia	etoposide
myocardial ischemia	losartan	mouth neoplasms	lidocaine	breast neoplasms	lovastatin
heart diseases	nicardipine	ovarian neoplasms	allopurinol	melanoma	dexamethasone
myocardial infarction	iloprost	ovarian neoplasms	carmustine	breast neoplasms	simvastatin
thrombosis	lidocaine	breast neoplasms	estramustine	precancerous conditions	progesterone
myocardial ischemia	bisoprolol	breast neoplasms	mifepristone	ovarian neoplasms	prednisolone
myocardial ischemia	pindolol	neoplasms, experimental	acetaminophen	ovarian neoplasms	prednisone
myocardial ischemia	ramipril	neoplasms, experimental	hydrocortisone	melanoma	lovastatin
brain ischemia	losartan	neoplasms, experimental	cyclosporine	neoplasm metastasis	dexamethasone
cardiomegaly	felodipine	breast neoplasms	flurbiprofen	liver neoplasms	vincristine
Behcet syndrome	methylprednisolone	breast neoplasms	thalidomide	ovarian neoplasms	auranofin
diabetes mellitus	melatonin	breast neoplasms	ibuprofen		

known associations in collected sets were calculated. As shown in Figure 2, causal inference with mean likelihood outperformed existing causal inference for better precisions.

3.3. PMF Models Construction and Association-Type Prediction. As we can see from above results, modified causal inference can be used to predict drug−disease associations. However, the specific type of these associations cannot be distinguished. On the other hand, much association-type information on chemical−disease has been curated in databases such as CTD. Therefore, the dataset of chemical−disease association type was extracted from CTD with a total of 80 537 records, as of December 6, 2013. 8235 chemicals and 3033 diseases were included, and association types consist of three types, namely, "therapeutic" (26 381 records), "marker/mechanism|therapeutic" (3228 records), and "marker/mechanism" (50 928 records). If a chemical exerts a potential or known therapeutic effect on a disease, the association is curated in the first group. The last group consists of associations that a chemical is associated with a disease, or may influence the etiology of a disease, while the second term is assigned for associations that cannot be clearly distinguished from the last two terms. Then, the dataset was used to train optimal models for each chemical and disease. In the process of training, the dataset was divided into five subsets, with four as training sets and one subset as a validation set (detailed parameters of PMF is shown in Table S1 in the Supporting Information). Optimal models were obtained and used to calculate the prediction score. The process was repeated 30 times to overcome the limit of randomness involved in subset selection. Mean value of the 30 prediction scores was calculated and then used to predict association type. PMF models correctly classified 86.22% of the association types in the entire dataset. In this case, accuracies for "marker/mechanism" and "therapeutic" type were 92.35% and 75.16%, and ratio of incorrect prediction between "therapeutic" type and "maker/mechanism" type was only 0.02%. Most false predictions happened because some known therapeutic associations were classified into "marker/mechanism|therapeutic" group.

3.4. Association Type Predictions for Constructed Associations. Given the good performance of trained models in association classification, they were further used to predict association type for constructed associations in causal networks of cardiovascular diseases, diabetes mellitus, and neoplasms. Types were predicted for 7545, 1369, and 6492 drug−disease pairs for these three types of diseases (details shown in Tables S2, S3, and S4, respectively, in the Supporting Information). Also, 772, 202, and 1172 associations were predicted as therapeutic associations, in which 72.2% (236/327), 86.1% (31/36), and 89.0% (170/191) associations were recovered. Thus, it can be seen that trained models were suitable for association type classification in cardiovascular diseases, diabetes mellitus, and neoplasms.

3.5. Comparison of Our Approach with Modified Causal Inference. In order to compare the performance of our approach with modified causal inference in predicting therapeutic drug−disease associations. Among all associations with predicted type, we collected the top 30% of associations with the highest association scores by single modified causal inference and CI-PMF. In the case of the former, only 10.2% (231/2264), 8.8% (36/411), and 9.7% (189/1948) associations were known as therapeutic associations for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively, while corresponding accuracies of our approach were 58.6% (136/232), 50.8% (31/61), and 39.8% (140/352). A hypergeometric test was then taken for the comparisons, and p-values were 7.65×10^{-91}, 1.42×10^{-24}, and 1.95×10^{-73}; thus, it is clear that our approach displayed better performance in predicting therapeutic associations. We took a closer look at the top 10 predictions by modified causal inference for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively, 3, 2, and 4 associations were known as "marker/mechanism", indicating that the approach is insufficient to distinguish different association types. In contrast, no known "marker/mechanism"

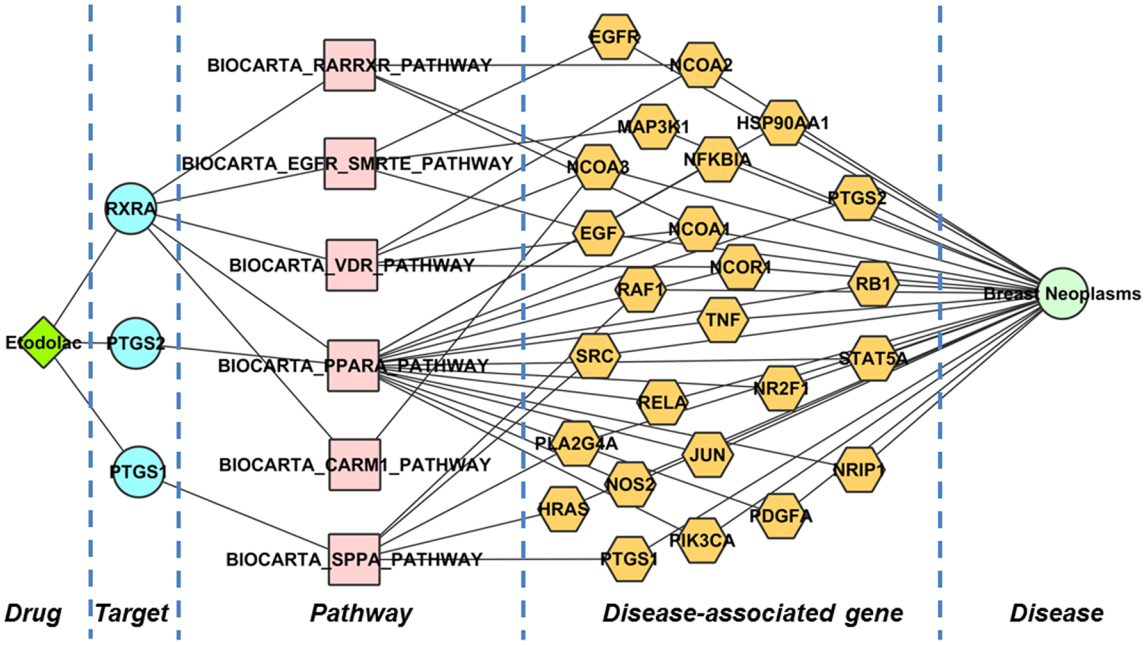


Figure 3. Causal network between etodolac and breast neoplasms.

associations were predicted in the top 10 associations by our approach.

3.6. Drug-Repositioning Predictions and Clinical Evaluation. Based on the above results, both association score and predicted type should be considered to predict new therapeutic associations. High association score suggests high possibility to be an association, and therapeutic associations are distinguished by the predicted association type. In the current research, each top 100 associations with highest association scores and "therapeutic" type were predicted as new therapeutic associations for cardiovascular diseases, diabetes mellitus, and neoplasms, respectively (details shown in Tables S5, S6 and S7, respectively, in the Supporting Information). Furthermore, we cross-checked the ClinicalTrials.gov website for verification of our drug-repositioning predictions. ClinicalTrial is an authoritative public database that collects information about clinical trials conducted around the world. A total of 33 238 clinical studies of human participants have been included as of the day we conducted the search (May 29, 2014). Drug names were matched to interventions, and disease names were matched to conditions. In our results, 21, 12, and 32 predicted drug−disease associations have been studied by clinical studies (see Table 2), and lots of drugs' therapeutic effects were evaluated. For instance, the drug clenbuterol has been investigated for whether it can improve the left ventricular function of chronic heart failure patients (ClinicalTrials.gov Identifier No. NCT00585546). A clinical study has been conducted for investigating effects of drug allopurinol treatment on an essential hypertension in adolescents (ClinicalTrials.gov Identifier No. NCT00288184). The effect of melatonin treatment on glucose tolerance was investigated in a phase 3 clinical study (ClinicalTrials.gov Identifier No. NCT01705639). The combination of indomethacin and biological therapy was investigated for whether they can treat advanced melanoma patients (ClinicalTrials.gov Identifier No. NCT00002535). Overall, predicted therapeutic drug−disease associations could provide valuable information for drug repositioning.

3.7. Revealing Mechanisms for Newly Predicted Associations. The constructed causal network provides a convenient way for investigating detailed information connecting drugs and diseases to reveal underlying mechanisms of the associations. There were 237, 87, and 883 chains extracted for the aforementioned 21, 12, and 32 clinical-verified associations, respectively (detailed information is shown in Tables S8, S9, and S10, respectively, in the Supporting Information).

In the current study, we took the use of related chains to illustrate etodolac's potential treatment effect on breast neoplasms. Six molecular pathways were inferred connecting etodolac and breast neoplasms. (Figure 3, plotted in Cytoscape[18]). Many biological processes are included, such as regulation of platelet activation, gene regulation by PPARα (peroxisome proliferator-activated receptor alpha), and gene expression control by VDR (vitamin D receptor). In detail, RXRα (retinoid X receptor alpha) is proposed to be of critical importance in the association with breast cancer. As an isotype of RXR, RXRα is able to form heterodimers with PPARs, RAR (retinoic acid receptor), and VDR (vitamin D receptor) through interacting proper ligands.[19,20] The PPAR and VDR signaling pathways have been shown to modulate invasion, metastasis, and proliferation of cancer, including breast cancer.[21−23] Moreover, CARM1 (coactivator-associated arginine methyl-transferase I) was reported to methylate nucleosomes and potentiate the transcriptional activation of RAR/RXR,[24] which has been treated as an therapeutic strategy in cancer.[25] In addition, CARM1 was suggested as a putative epigenetic target in ER-positive breast cancer and played roles in the regulation of breast cancer cells differentiation and proliferation.[26] On the other hand, the platelet is of great importance in cancer progression; cancer growth and dissemination are facilitated by coagulation via platelet activation,[27] and etodolac is proposed to influence the breast cancer progression through inhibition of the platelet activation. Therefore, these related chains suggested that etodolac may exert a treatment effect on breast cancer through regulation of cell differentiation, proliferation, and metastasis. Furthermore,

dx.doi.org/10.1021/ci500340n | J. Chem. Inf. Model. 2014, 54, 2562−2569

etodolac has been studied for its treatment effect on decreasing metastatic potential and cancer recurrence (ClinicalTrials.gov Identifier No. NCT00502684). Hence, inferred chains by our approach could benefit the understanding of drug−disease associations, thus gaining insight into mechanisms of drug treatment on the disease.

4. DISCUSSION

Integration of different types of interactions from heterogeneous databases, which enables construction of causal networks for cardiovascular diseases, diabetes mellitus, and neoplasms, facilitate the investigation of complex relationships between drugs and diseases. CI-PMF combined modified causal inference and PMF to predict possible drug-disease associations based on causal networks. PMF approach further classified associations into different types to predict therapeutic associations with possible underlying mechanisms revealed.

4.1. Weighted Causal Network Construction: The Fundamental Step. Drug−complex disease association is a many-to-many relationship. First, a complex disease encompassing a collection of related conditions may be caused by one mutation with a strong biological effect, or interactions of such mutations under a specific condition.[28,29] Second, numerous drugs have shown multiple-targeting activities, e.g., aspirin shows both analgesic and antipyretic effects, coined as polypharmacology.[5,6] Such a complex relationship can hardly be expressed by simple multiple regression equations. In the present work, we integrated drug−target, pathway−gene, and disease−gene interactions into the causal network, which intuitively exhibited the complex relationship. The drug's possible effect on diseases in each chain was encompassed in our network with the pathways serving as bridges to connect drug targets and disease-associated genes. Moreover, known drug−disease associations were extracted for weight calculation of each connection. Each connection constructed in final weight causal networks was quantitatively evaluated, providing the basis for new association predictions.

4.2. Relevant Biological Processes: An Indispensable Factor for Association Prediction in Complex Diseases. As shown in constructed networks, one drug always exerts effects on one disease through many chains, representing biological processes of a certain drug on one disease. The relatedness can be measured based on the weighted causal network, although specific roles of the process remains elusive. The many-to-many relationship is evaluated by mean likelihood of all relevant chains. Analysis of all possible chains would reveal complex influence of drugs on diseases with overall relevancy quantitatively assessed. The comparison results between modified and existing causal inference revealed by our study proved the necessity to take all relevant biological processes into consideration for more-accurate predictions.

4.3. Conducting Causal Inference before PMF. PMF dissects the existing knowledge to mine patterns of known associations and guides classification of new association. It is common that the negative results were ignored, while only those positive results have the opportunity to be reported and curated. Hence, the trained models in the current study could not distinguish the existence of an association between a drug and a disease. Modified causal inference complemented PMF, with its excellent ability to predict associations to overcome this limit. On the other hand, PMF makes up for the defect of causal inference in the classification of association types. Therefore, combining the two methods gives complementary advantages for drug-repositioning predictions.

4.4. Methods Comparisons in Predicting Drug−Disease Associations. As mentioned above, many relevant methods have been reported to predict drug−disease associations. Several drawbacks exist among these approaches. Methods based on chemical similarity cannot predict many physiological effects, and molecular activity similarity-based approaches are always restricted by the quality of data. Other approaches, based on shared molecular pathology and side effect similarity, are limited by precise definition and measurement of concerning issues. Moreover, the lack of a three-dimensional (3D) structure of protein targets makes it impossible to apply molecular docking methods. Compared to aforementioned methods, data used for our approach can be easily curated from expert-knowledge databases, and there is no need to clearly define the molecular activity profile and complex pathology. Also, 3D structures are no longer necessary. In addition, CI-PMF completes multiple missions at the same time: both drug-disease associations and underlying molecular mechanisms are inferred. The types of associations are also predicted, thus contributing to further investigation of these associations.

4.5. Data Requirement for the Methodology. Complete datasets regarding drug−target, pathway−target, disease−gene, and known drug−disease associations should be curated for causal network construction and PMF model training. To note, it is necessary to control the quality of collected data. In order to reduce the noise or error of PMF model and causal network, computationally inferred data should not be included. In addition, CI-PMF was designed for complex diseases; thus, datasets related to monogenic diseases still need to be tested.

4.6. Limits and Perspective. The application scope of CI-PMF is limited by the existing data used for network construction and PMF model training. The results in association type classification showed that many constructed associations could not be classified, because of the absence of corresponding drugs/diseases in training data. In future studies, relevant data should be collected from more databases to expand the application scope of our approach. It is also important to collect more-detailed drug pharmacological profiles, such as inhibitor, agonist, and antagonist to enrich the information on biological mechanisms. Aside from the diseases mentioned above, CI-PMF can also be employed to predict drug repositioning of other complex diseases (i.e., Alzheimer's disease).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information includes the following: (1) Detail parameters in PMF model construction (Table S1); (2) 7545 predicted associations for cardiovascular diseases (Table S2); (3) 1369 predicted associations for diabetes mellitus (Table S3); (4) 6492 predicted associations for neoplasms (Table S4); (5) 100 newly predicted therapeutic association for cardiovascular diseases (Table S5); (6) 100 newly predicted therapeutic association for diabetes mellitus (Table S6); (7) 100 newly predicted therapeutic association for neoplasms (Table S7); (8) detail chains of 21 clinical-verified associations for cardiovascular diseases; (9) detail chains of 12 clinical-verified associations for cardiovascular diseases; (10) detail chains of 32 clinical-verified associations for neoplasms (EXCEL). This

material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: lizheng1@gmail.com (Dr. Zheng Li).
*E-mail: chengyy@zju.edu.cn (Dr. Yiyu Cheng).

Author Contributions

§These authors contributed equally to this work.

Funding

This work was financially supported by the National Science & Technology Major Project (No. 2012ZX09503001−001).

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

CI-PMF, causal inference-probabilistic matrix factorization; CTD, Comparative Toxicogenomics Database; AUC, area under receiver operating characteristic curve; LV, latent variable; ROC, receiver operating characteristic; PPARα, peroxisome proliferator-activated receptor alpha; RAR, retinoic acid receptor; VDR, vitamin D receptor; CARM1, coactivator-associated arginine methyl-transferase I

REFERENCES

(1) Schork, N. J. Genetics of complex disease: Approaches, problems, and solutions. *Am. J. Respir. Crit. Care Med.* **1997**, *156*, S103−S109.

(2) Heron, M. Deaths: Leading causes for 2010. In *National Vital Statistics Reports*, Vol. *62*; U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System: Hyattsville, MD, 2013; pp 1−97.

(3) Hurle, M. R.; Yang, L.; Xie, Q.; Rajpal, D. K.; Sanseau, P.; Agarwal, P. Computational drug repositioning: From data to therapeutics. *Clin. Pharmacol. Ther.* **2013**, *93*, 335−341.

(4) Wu, Z.; Wang, Y.; Chen, L. Network-based drug repositioning. *Mol. BioSyst.* **2013**, *9*, 1268−1281.

(5) Apsel, B.; Blair, J. A.; Gonzalez, B.; Nazif, T. M.; Feldman, M. E.; Aizenstein, B.; Hoffman, R.; Williams, R. L.; Shokat, K. M.; Knight, Z. A. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat. Chem. Biol.* **2008**, *4*, 691−699.

(6) Reddy, A. S.; Zhang, S. Polypharmacology: Drug discovery for the future. *Expert Rev. Clin. Pharmacol.* **2013**, *6*, 41−47.

(7) Dudley, J. T.; Deshpande, T.; Butte, A. J. Exploiting drug−disease relationships for computational drug repositioning. *Briefings Bioinf.* **2011**, *12*, 303−311.

(8) Cheng, F.; Li, W.; Zhou, Y.; Li, J.; Shen, J.; Lee, P. W.; Tang, Y. Prediction of human genes and diseases targeted by xenobiotics using predictive toxicogenomic-derived models (PTDMs). *Mol. BioSyst.* **2013**, *9*, 1316−1325.

(9) Huang, Y. F.; Yeh, H. Y.; Soo, V. W. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med. Genomics* **2013**, *6* (Suppl. 3), S4.

(10) Pratanwanich, N.; Lio, P. Pathway-based Bayesian inference of drug−disease interactions. *Mol. BioSyst.* **2014**, *10*, 1538−1548.

(11) Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503.

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

(13) 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.

(14) Davis, A. P.; Murphy, C. G.; Johnson, R.; Lay, J. M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; King, B. L.; Rosenstein, M. C.; Wiegers, T. C.; Mattingly, C. J. The Comparative Toxicogenomics Database: Update 2013. *Nucleic Acids Res.* **2013**, *41*, D1104−D1114.

(15) Li, J.; Lu, Z. Pathway-based drug repositioning using causal inference. *BMC Bioinf.* **2013**, *14*.

(16) Salakhutdinov, R.; Mnih, A. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems; NIPS*, 2007; Vol. *20*, pp 1257−1264.

(17) Cobanoglu, M. C.; Liu, C.; Hu, F.; Oltvai, Z. N.; Bahar, I. Predicting drug−target interactions using probabilistic matrix factorization. *J. Chem. Inf. Model.* **2013**, *53*, 3399−3409.

(18) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498−2504.

(19) Hu, X.; Funder, J. W. The evolution of mineralocorticoid receptors. *Mol. Endocrinol.* **2006**, *20*, 1471−1478.

(20) Perez, E.; Bourguet, W.; Gronemeyer, H.; de Lera, A. R. Modulation of RXR function through ligand design. *Biochim. Biophys. Acta* **2012**, *1821*, 57−69.

(21) Lopes, N.; Sousa, B.; Martins, D.; Gomes, M.; Vieira, D.; Veronese, L. A.; Milanezi, F.; Paredes, J.; Costa, J. L.; Schmitt, F. Alterations in Vitamin D signalling and metabolic pathways in breast cancer progression: A study of VDR, CYP27B1 and CYP24A1 expression in benign and malignant breast lesions. *BMC Cancer* **2010**, *10*, 483.

(22) Suchanek, K. M.; May, F. J.; Robinson, J. A.; Lee, W. J.; Holman, N. A.; Monteith, G. R.; Roberts-Thomson, S. J. Peroxisome proliferator-activated receptor alpha in the human breast cancer cell lines MCF-7 and MDA-MB-231. *Mol. Carcinog.* **2002**, *34*, 165−171.

(23) Matsuda, S.; Kitagishi, Y. Peroxisome proliferator-activated receptor and vitamin D receptor signaling pathways in cancer cells. *Cancers* **2013**, *5*, 1261−1270.

(24) Xu, W.; Chen, H.; Du, K.; Asahara, H.; Tini, M.; Emerson, B. M.; Montminy, M.; Evans, R. M. A transcriptional switch mediated by cofactor methylation. *Science* **2001**, *294*, 2507−2511.

(25) Altucci, L.; Leibowitz, M. D.; Ogilvie, K. M.; de Lera, A. R.; Gronemeyer, H. RAR and RXR modulation in cancer and metabolic disease. *Nat. Rev. Drug Discovery* **2007**, *6*, 793−810.

(26) Al-Dhaheri, M.; Wu, J.; Skliris, G. P.; Li, J.; Higashimato, K.; Wang, Y.; White, K. P.; Lambert, P.; Zhu, Y.; Murphy, L.; Xu, W. CARM1 is an important determinant of ERα-dependent breast cancer cell differentiation and proliferation in breast cancer cells. *Cancer Res.* **2011**, *71*, 2118−2128.

(27) Bambace, N. M.; Holmes, C. E. The platelet contribution to cancer progression. *J. Thromb. Haemostasis* **2011**, *9*, 237−249.

(28) Mitchell, K. J. What is complex about complex disorders? *Genome Biol.* **2012**, *13*, 237.

(29) Buchanan, A. V.; Weiss, K. M.; Fullerton, S. M. Dissecting complex disease: The quest for the Philosopher's Stone? *Int. J. Epidemiol.* **2006**, *35*, 562−571.