Mining Drug–Disease Relationships as a Complement to Medical Genetics-Based Drug Repositioning: Where a Recommendation System Meets Genome-Wide Association Studies

H Wang[1,2], Q Gu[3], J Wei[4], Z Cao[1] and Q Liu[1]

A novel recommendation-based drug repositioning strategy is presented to simultaneously determine novel drug indications and side effects in one integrated framework. This strategy provides a complementary method to medical genetics-based drug repositioning, which reduces the occurrence of false positives in medical genetics-based drug repositioning, resulting in a ranked list of new candidate indications and/or side effects with different confidence levels. Several new drug indications and side effects are reported with high prediction confidences.

Finding new uses for existing drugs, also known as drug repositioning, is a promising strategy in translational medicine for obtaining more therapeutic drugs.[1] *In silico* drug repositioning often focuses on inferring novel drug–disease associations. There are two types of widely used drug repositioning methodologies: (1) Systematic models are used to infer novel associations based on the mining of known drug–disease relationships. Such models often follow the rule of "guilt by association," which assume that similar drugs tend to have similar indications.[1,2] (2) Direct inferences of drug–disease relationships are established by mining novel drug–target and gene–disease relationships. Such methods, often referred to as the medical genetics-based drug repositioning, are implemented based on the assumption that disease genes are highly druggable and the drug indications frequently match the genetic disease traits. Typical published works in this area include the pioneering strategy to use genome-wide association studies (GWAS) for drug repositioning[3] followed by a validation phase which considers the loss of function or gain of function of mutated genes as well as the mode of action of the ligands (agonists or antagonists) in the repositioning.[4]

The first type of drug repositioning, like PREDICT,[2] is designed to infer novel drug indications based on the physiochemical properties of the compounds, the pathophysiological characteristics of the diseases as well as the known chemical–protein interactome. These methods made novel indication inferences based on explicitly suggested positive and negative data to build the prediction model.[1,2] In this scenario, the true associations are easy to obtain while the negative associations are difficult to identify. Therefore, traditional methods have not used the ground-truth negative data in identifying drug-indication associations because they were generally considered unprofitable.[1,2] Extended network-based models[1] improved the prediction results but they only used positive-sample–based inferences. We noticed that substantial drug side effect profiles exist in addition to drug indication information that can be explored.[5] So far, this information has not been exploited or properly considered as negative data. By harnessing the side effect profiles, well known drug–disease relationships can be used to mine novel relationships analogous to collaborative filtering (CF) in the recommendation system, as initially presented in the social network community (see Supplementary Texts, which are available online). In a recommendation system, the *Users* (drugs) may have their historical preferences to *Commodities* (disease; known drug–disease relationships including indications and side effects), where such information can be used to predict potential user-commodity preferences as recommendations. Borrowing this idea from the social network sphere, we mined the known drug indications and

[1]Department of Central Laboratory, Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China; [2]Department of Computer Science, Hefei University of Technology, Hefei, China; [3]Department of Systems and Information Engineering, University of Virginia, U.S.; [4]R&D Information, AstraZeneca, Shanghai, China. Correspondence: Q Liu (qiliu@tongji.edu.cn)

Received 18 September 2014; accepted 19 January 2015; advance online publication 23 January 2015. doi:10.1002/cpt.82

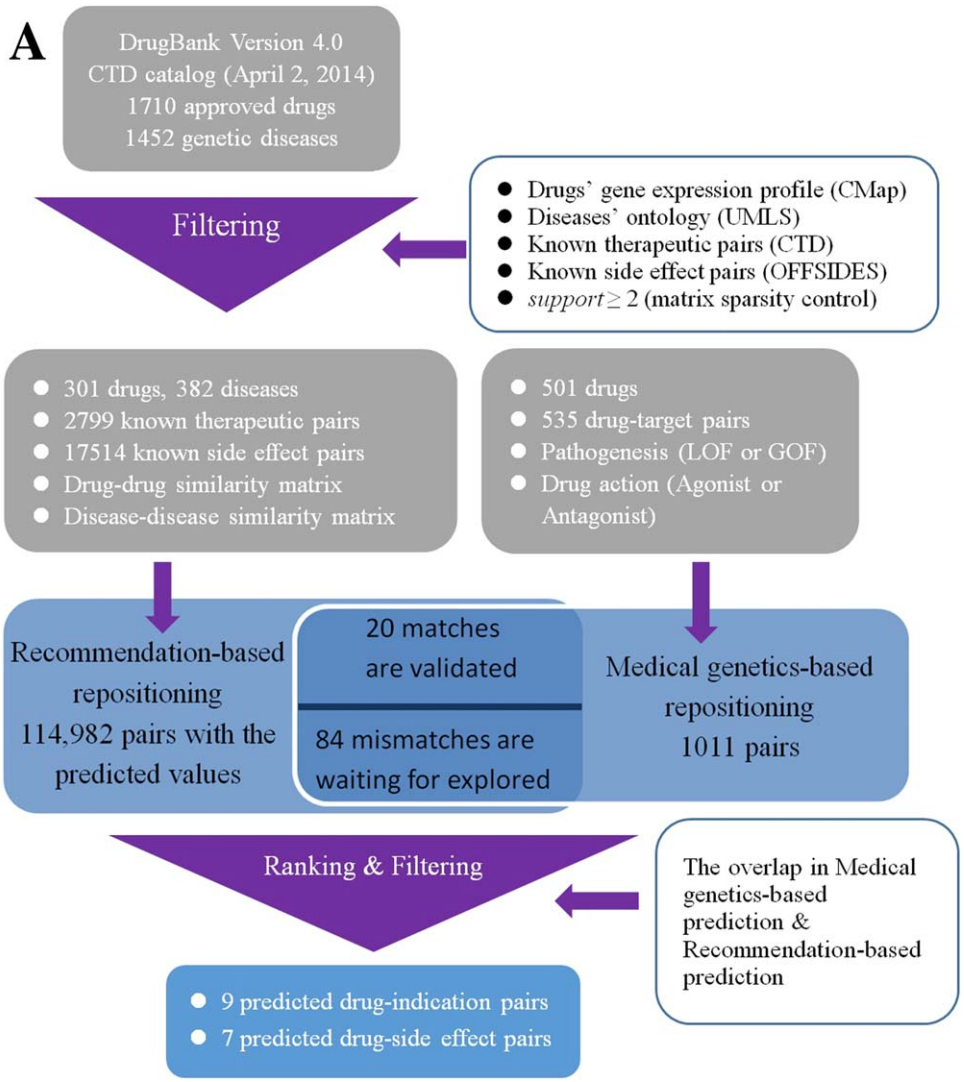
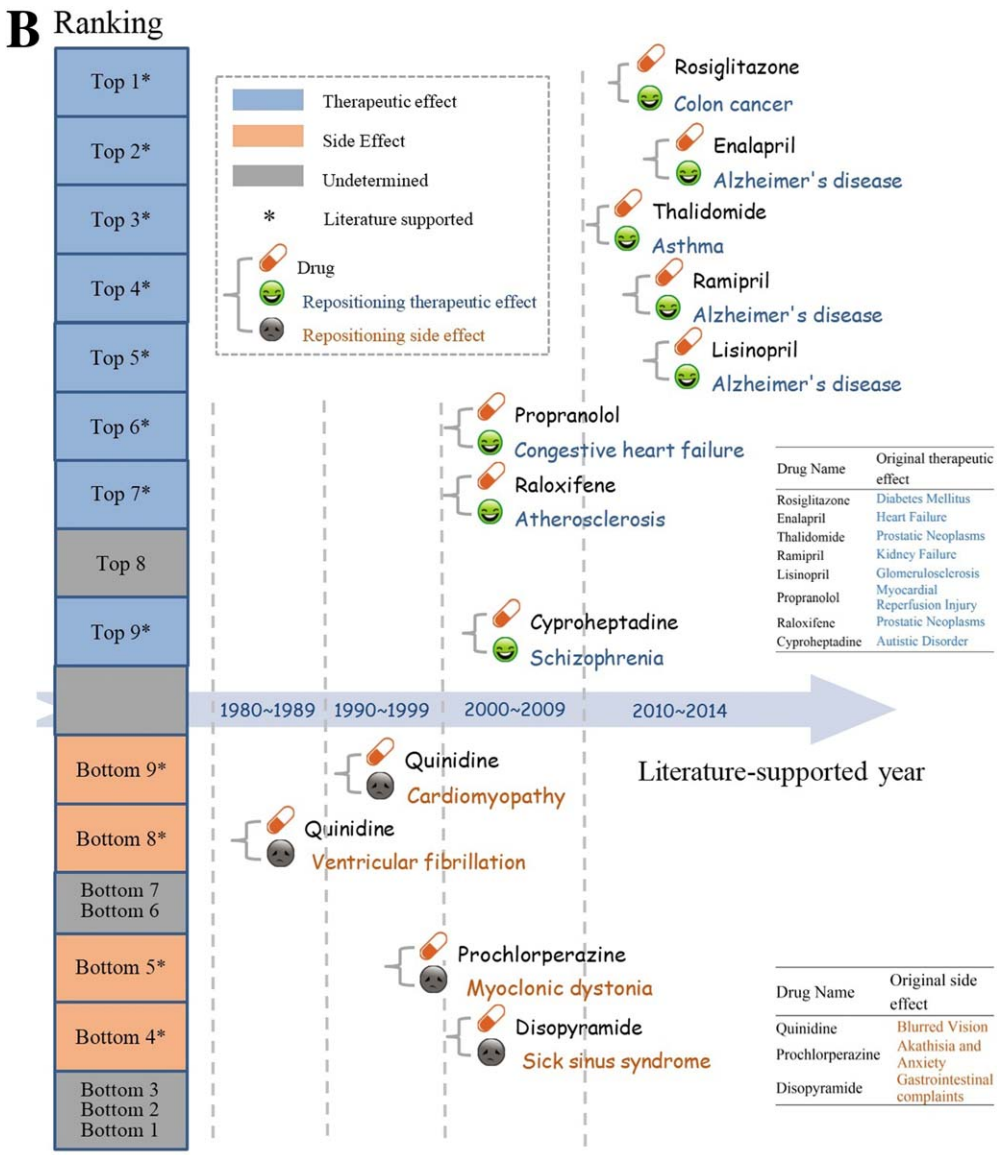


Figure 1 (**A**) Data processing pipeline. A total of 114,982 drug–disease pairs were presented in our recommendation system curated from DrugBank, the Comparative Toxicogenomics Database and OFFSIDES database after going through a filtering step (Supplementary Texts). Among them, 104 pairs overlapped with 1,011 medical genetics-based repositioning mismatches. A complementary study resulted in 16 potential pairs (9 novel drug-indication pairs, 7 novel drug-side effect pairs) based on the consistent results of recommendation-based and medical genetics-based repositioning strategies. The existing gene expression profiles for the drugs were obtained from the Connectivity Map (https://www.broadinstitute.org/cmap/). The ontology information for diseases was derived from the UMLS (http://www.nlm.nih.gov/research/umls/). (**B**) The ranking list obtained by the complementary repositioning strategy is presented, including the nine novel drug-indication pairs and seven novel drug-side effect pairs. Twelve of the 16 pairs are supported by literature published between 1980 and 2014.

drug side effects in an integrative way, and here we present a novel recommendation-based drug repositioning strategy, which simultaneously predicts novel drug indications and side effects in one computational framework.

The second type of drug repositioning methods generally involves mining the GWAS catalog or a genetic disease database such as Mendelian Inheritance in Man (OMIM). Nevertheless, we found that inferences based on the pathogenesis of genetic diseases and drug modes of action can also be substantially improved. Incorrect inferences exist in such repositioning (see Supplementary Texts), likely due to the fact that: (1) Our knowledge regarding drug modes of action and the pathogenesis of genetic diseases remains insufficient, which may lead to less confident inferences of drug–disease relationships; and (2) most complex diseases are often related to multiple gene malfunctions rather than a single malfunction, thus inferences based on only a single gene mutation related to a genetic disease may be of limited value. Furthermore, medical genetics-based drug repositioning only provides the candidate new indications or side effects with equal confidence levels, while a candidate ranking list with different confidence levels is preferred clinically.

We propose that these two types of drug repositioning methods can be integrated together. Applying a recommendation-based strategy should provide a complementary method to



Drug Name	Original therapeutic effect
Rosiglitazone	Diabetes Mellitus
Enalapril	Heart Failure
Thalidomide	Prostatic Neoplasms
Ramipril	Kidney Failure
Lisinopril	Glomerulosclerosis
Propranolol	Myocardial Reperfusion Injury
Raloxifene	Prostatic Neoplasms
Cyproheptadine	Autistic Disorder

Drug Name	Original side effect
Quinidine	Blurred Vision
Prochlorperazine	Akathisia and Anxiety
Disopyramide	Gastrointestinal complaints

Figure 1 (Continued)

medical genetics-based drug repositioning, resulting in a ranked list of new candidate indications or side effects with different confidence levels.

Our experiment was focused on the collected 1,710 FDA-approved drugs from DrugBank (records dated through April 2014). After manual curation, 963 drugs with their 1,263 disease UMLS terms representing indications or side effects were selected based on The Comparative Toxicogenomics Database and OFF-SIDES database (Supplementary Table S1). We presented a novel matrix factorization-based CF method to predict unknown drug–disease relationships based on the known drug–disease relationships by integrating drug–drug similarity measurements and disease–disease similarity measurements (see Supplementary Texts). In our study, we measured drug–drug similarity based on the drug gene expression profiles using Connectivity Map data, and disease–disease similarities were measured by determining the disease ontology similarity (see Supplementary Texts). Based on the existing gene expression profile information for drugs in the Connectivity Map as well as the matrix sparsity control for known drug–disease relationships, we eventually obtained 301 FDA-approved drugs associated with UMLS information from 382 diseases (**Figure 1A**). Our recommendation system contained 20,313 known drug–disease pairs in total, and the remaining 94,669 unknown pairs were ranked based on the CF method. Based on a specific cutoff, our ranking resulted in 6,022 pairs of novel drug indications, 20,341 pairs of novel side effects, and the remaining pairs were marked as "undetermined" (Supplementary Texts, Supplementary Table S2).

As a complementary study, the "mismatched pairs" identified by a former medical genetics-based repositioning study were used

for comparison[3] (Supplementary Table S3). Among them, 104 pairs overlapped with the pairs in our recommendation system, 20 pairs actually already have known drug–disease relationships based on the literature, and the remaining 84 pairs remain to be explored. Using these 104 pairs as a comparison, excluding the undetermined pairs in the recommendation-based repositioning, we found that the Pearson correlation coefficient between the two repositioning strategies was 0.925 ($P < 1e^{-10}$; Supplementary Texts), indicating that the two prediction results were highly consistent. Furthermore, among the 20 validated pairs in medical genetics-based drug repositioning, 14 were correctly predicted and 6 inferences were incorrect, while for the recommendation-based drug repositioning, 8 among the 20 pairs were correctly predicted, and the others, including the 6 incorrectly predicted pairs in the medical genetics-based model, were denoted as "undetermined" (Supplementary Table S4). The result indicates that recommendation-based repositioning can greatly reduce the error rate in medical genetics-based drug repositioning. This method only recommended the most confident drug repositioning candidates based on the ranking results, while reducing false positives by marking them as "undetermined."

Based on this method, from the former 84 candidate pairs among the 104 pairs waiting to be explored, we selected consistent inference results between medical genetics-based drug repositioning and recommendation-based drug repositioning, resulting in 9 novel drug-indication pairs and 7 novel drug-side effect pairs (Supplementary Table S6). Surprisingly, we found that 8 novel drug-indications and 4 novel drug-side effects are supported by published studies (**Figure 1B**), using either preclinical or clinical data. For example, based on the medical genetics-based prediction, perindopril as an angiotensin-converting enzyme (ACE) inhibitor can be repositioned to treat Alzheimer's disease (AD).[2] In our study, several other angiotensin-converting enzyme inhibitors, i.e., enalapril, ramipril, and lisinopril, which have similar gene expression profiles, were repositioned to Alzheimer's disease. The ACE inhibitors were previously reported to be widely used as anti-hypertensive agents. Of interest, direct experimental evidence in 2013 indicated the beneficial effects of ACE-I on endothelial dysfunction and endothelial apoptosis, which is regarded as a basis of AD pathogenesis (see Supplementary Table S6).

In summary, we presented a novel recommendation-based drug repositioning strategy and integrated the two distinct perspectives of rational drug repositioning to obtain more reliable and ranked drug repositioning results. We speculated that these two strategies may be related in that if two drugs have a similar gene expression profile (as in our drug similarity study), they tend to have similar target profiles and thus tend to correlate with similar genetic diseases. From this point of view, medical genetics-based drug repositioning, which assumes that only one gene acts as the "bridge" between the drug and disease, may be a simplified method of examining drug–drug similarities as in the recommendation system. We also examined other methods for measuring drug–drug similarities, such as the use of drug fingerprint information, and arrived at similar drug repositioning results (Supplementary Table S5).

Additional Supporting Information may be found in the online version of this article.

ACKNOWLEDGMENTS

This work was supported by the National High Technology Research and Development Program ("863" Program) of China (grant 2012AA020405), National Natural Science Foundation of China (grants 31100956 and 61173117), the Fundamental Research Funds for the Central Universities of China (grant 2000219084), and the key project of Scientific Research Innovation, Shanghai (grant 20002360059).

CONFLICT OF INTEREST

The authors declared no conflict of interest.

1. Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P. & Agarwal. P. Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* **93**, 335–341 (2013).
2. Gottlieb, A., Stein, G.Y. & Ruppin, E. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **7**, 496 (2011).
3. Sanseau, P. *et al*. Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30**, 317–320 (2012).
4. Wang, Z.Y. & Zhang, H.Y. Rational drug repositioning by medical genetics. *Nat. Biotechnol.* **31**, 1080–1082 (2013).
5. Campillos, M. *et al*. Drug target identification using side-effect similarity. *Science* **11**, 263–266 (2008).