DrPOCS: Drug Repositioning Based on Projection Onto Convex Sets

Yin-Ying Wang, Chunfeng Cui, Liqun Qi, Hong Yan, and Xing-Ming Zhao

Abstract—Drug repositioning, i.e., identifying new indications for known drugs, has attracted a lot of attentions recently and is becoming an effective strategy in drug development. In literature, several computational approaches have been proposed to identify potential indications of old drugs based on various types of data sources. In this paper, by formulating the drug-disease associations as a low-rank matrix, we propose a novel method, namely DrPOCS, to identify candidate indications of old drugs based on projection onto convex sets (POCS). With the integration of drug structure and disease phenotype information, DrPOCS predicts potential associations between drugs and diseases with matrix completion. Benchmarking results demonstrate that our proposed approach outperforms popular existing approaches with high accuracy. In addition, a number of novel predicted indications are validated with various types of evidences, indicating the predictive power of our proposed approach.

Index Terms—Drug repositioning, projection onto convex sets (POCS), matrix completion, singular value decomposition (SVD)

1 Introduction

Drug discovery is a long and expensive procedure, where it typically takes about fourteen years and two billion dollars to bring a new drug to market [1], [2]. Due to the costly and time-consuming procedure, drug repositioning is becoming a popular strategy in drug discovery. A notable example of drug repositioning is Thalidomide which was originally developed as sleeping pill and was soon found useful for pregnant women with morning sickness [3]. More recently, ongoing research implies its possible indications for erythema nodosum leprosum and type II diabetes [4]. However, it is a big challenge to identify novel indications of existing drugs since new indications may not have any relations with the original use of old drugs.

With the explosive growth of large-scale genomic and phenotypic data, as well as the chemical and bioactivity data of drugs, much effort has been developed for the purpose of repositioning drugs. By assuming that novel indications can be identified for a drug if the drug can target a new protein related to a certain disease, some computational approaches have been proposed to predict new targets of known drugs [5], [6], [7], [8], [9]. For example, Wang et al. predicted drug targets based on the derived interactions between drugs and protein domains [5], [10], and Zhang et al. constructed a post-translational regulatory network to explore network motifs as potential drug targets which can help design multi-component or combinatorial drugs [11]. Furthermore, based on the idea of 'anti-correlation between drug and disease signatures', several methods have been proposed to find new indications of drugs by exploiting the disease related signatures [12], [13], [14], [15]. For instance, Iorio et al. built a drug network based on the transcriptional profiles induced by drugs and predicted new indications for old drugs [15], while Iskar et al. defined the drug signatures in another way by removing batch effects inherited in the transcription profiles [16].

Furthermore, with the biosimilar principle that assumes biological entities with similar behaviors to have same functions, computational approaches have been proposed for drug repositioning. For instance, with multiple drug–drug and disease–disease similarity measures, Gottlieb et al. proposed a computational approach named as PREDICT to predict novel indications of drugs [17]. Wu et al. constructed a weighted drug-disease heterogeneous network based on known disease–gene and drug–target relationships, and identified potential drug-disease associations with clustering [18]. In recent years, a number of methods have been proposed for drug repositioning based on matrix decompositions, where the drug-disease associations were described as a matrix. Generally, the drug-disease association matrix was decomposed into multiple low-rank matrices consisting of latent features that are assumed to govern the associations [19], [20], [21]. For example, Mehmet et al.

- *Y.-Y. Wang is with the Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China and the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, China. E-mail: yy_wang@tongji.edu.cn.*
- *C. Cui and H. Yan are with the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong, China. E-mail: chunfengcui89@gmail.com, h.yan@cityu.edu.hk.*
- *L. Qi is with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China. E-mail: liqun.qi@polyu.edu.hk.*
- *X.-M. Zhao is with the Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Shanghai 200433, China. E-mail: xmzhao@fudan.edu.cn.*

Manuscript received 11 Feb. 2018; accepted 13 Apr. 2018. Date of publication 26 Apr. 2018; date of current version 4 Feb. 2019.
(Corresponding author: Xing-Ming Zhao.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2830384

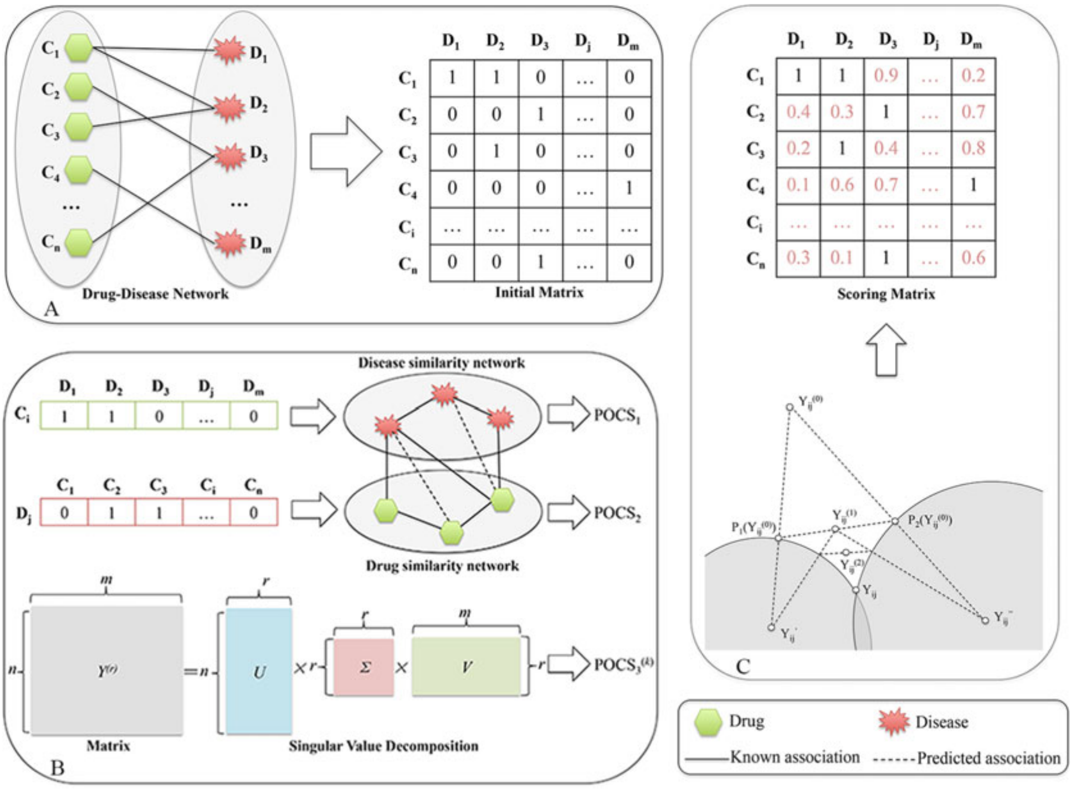

Fig. 1. The pipeline of repositioning drugs based on DrPOCS. A. Problem formulation: formulated the drug-disease associations into an initial binary matrix with missing values. B. Construction of convex sets and projections based on different constraints: 1) Projections based on local pair-wise correlations using disease phenotype information. 2) Projections based on local pair-wise correlations using drug chemical structure. 3) Projections based on global structure of matrix using singular vector decomposition. C. Drug repositioning with DrPOCS: identifying candidate indications of old drugs based on the iterative projections onto different convex sets.

proposed a novel Bayesian approach named as KBMF2K to predict new targets for known drugs which can be further used for drug repositioning [22]. Taguchi et al. repositioned drugs for posttraumatic stress disorder-mediated heart disease with unsupervised feature extraction based on principal com-ponent analysis (PCA) [23].

In this paper, by formulating the drug-disease associations as a low-rank matrix, we propose a novel method, namely DrPOCS, to identify candidate indications of old drugs based on projection onto convex sets (POCS, also known as the alternating projection) which has been successfully used in microarray missing data imputation [24] and graph matching [25]. The main idea of DrPOCS is to formulate the prior knowledge into a corresponding convex set and then use a convergence-guaranteed iterative procedure to obtain a solution in the intersection of all these sets. With the integration of drug chemical structure and disease phenotype information as well as the global structure of the drug-disease association matrix, DrPOCS predicts potential associations between drugs and diseases with matrix completion. Benchmarking results demonstrate that our proposed approach outperforms popular existing approaches with high accuracy. In addition, several novel indications predicted are validated with various types of evidences, indicating the predictive power of our proposed approach.

2 Materials and Method

To identify potential indications of old drugs, we hereby proposed an approach namely DrPOCS based on projection onto convex sets (POCS). Fig. 1 illustrates the pipeline of the DrPOCS approach. First, an initial binary matrix with missing values was constructed with known drug-disease associations. Second, different types of convex sets and corresponding projections were constructed based on drug chemical structure and disease phenotype information. Finally, the potential associations between drugs and diseases were predicted with matrix completion by formulating the drug-disease association as a low-rank matrix. The details can be found below.

2.1 Data Resources

2.1.1 *Gold Standard Drug-Disease Associations*

In this paper, the human FDA-approved drugs were obtained from the DrugBank database (Version 5.0), which is a blended cheminformatics resource with detailed drug data and comprehensive target information [26]. The known drug-disease associations used in this study were obtained from the Comparative Toxicogenomics Database (CTD) [27], where only those marked with direct evidences (therapeutic or maker/mechanism) were collected to make sure the associations more confident. We further required each drug or disease has no less than two associations. As a result, 80972 drug-disease associations involving 656 human drugs approved in DrugBank and 1921 diseases were used as the gold standard dataset.

2.1.2 *Drug and Disease Similarity*

The disease similarity was calculated by MinMiner based on disease phenotypes as described in [28], which was calculated analogous to the term frequency-inverse document frequency technique widely used in information retrieval. Briefly, each disease was described as a feature vector by using the anatomy (A) and the disease (C) sections of medical subject heading vocabulary (MeSH) to automatically extract MeSH terms from its OMIM records [29], where every entry in the feature vectors represents an MeSH concept relevance to the phenotype. For each concept, its relevance was calculated by the actual count of the concept in a document plus the relevance sum of the concept hyponyms. The similarity between a pair of diseases $\{d_i, d_j\}$ was calculated as the cosine similarity between two Mesh concept vectors $t_i = \{t_{i1}, t_{i2}, \cdots t_{ik}\}$ and $t_j = \{t_{j1}, t_{j2}, \cdots t_{jk}\}$. To calculate the drug chemical similarity, each compound was described as an 880 dimensions binary vector based on the PubChem fingerprints, where the element in the vector is 1 if the corresponding fingerprint is contained in the drug and 0 otherwise. Then the 2D similarity between two compounds was calculated as the Tanimoto coefficient which is defined as the ratio of the number of common fingerprints to the total number of fingerprints [30].

2.2 Problem Formulation

In this paper, the drug-disease associations were formulated as a matrix with drug and disease sets respectively denoted as $C = \{C_1, C_2, \cdots C_n\}$ and $D = \{D_1, D_2, \cdots D_m\}$, where n is the number of drugs and m is the number of diseases considered here. Specifically, let X be an $n \times m$ binary matrix describing the drug-disease associations. Each element (i, j) in the matrix is the index of the association between drug i and disease j, where the known association set was denoted as K and the rest unobserved set was denoted as Ω. For all elements $(i, j) \in K$, $x_{ij} = 1$ means drug i and disease j are known to be associated, while $x_{ij} = 0$ for $(i, j) \in \Omega$ means that the association between drug i and disease j is uncertain as shown in Fig. 1A. That is, X is defined as follows.

$$x_{ij} = \begin{cases} 1, & \text{if } (i, j) \in K; \\ 0, & \text{if } (i, j) \in \Omega. \end{cases} \tag{1}$$

With the known drug-disease associations described by X, our goal is to reconstruct an $n \times m$ scoring matrix Y that predicts potential associations for those $(i, j) \in \Omega$ in X. Specifically, the entry y_{ij} indicates the probability of association between drug i and disease j to some extent. The matrix Y can be defined as below.

$$y_{ij} \in \begin{cases} 1, & \text{if } (i, j) \in K; \\ [0, 1], & \text{if } (i, j) \in \Omega. \end{cases} \tag{2}$$

2.3 Drug Repositioning with DrPOCS

To predict potential drug-disease associations as described above, we employed projection onto convex sets based approach here. The main idea of DrPOCS is to construct a series of convex sets with pieces of prior knowledge and obtain an optimal solution from the intersection of all possible convex sets by a convergence-guaranteed iterative procedure. Before applying the DrPOCS approach, one must know how to project onto the different convex sets. The detail of the convex sets and the corresponding projections (as shown in Fig. 1B) can be found as follows.

2.3.1 *Projections Based on Local Pair-Wise Correlations*

With the assumption that similar drugs can be used for the same diseases and *vice versa*, we explored the row-wise and column-wise correlations of the matrix X to construct the convex sets. Given a drug i with its vector denoted as $X_i = \{x_{i1}, x_{i2}, \cdots x_{im}\}$, the corresponding association score of this drug with unknown disease j can be estimated based on the average similarity between disease j and other diseases known associated with the drug as follows.

$$S_{ij} = \frac{\sum_{l=1}^{n} S_D(D_j, D_l) x_{il}}{\sum_{l=1}^{n} x_{il}}, \tag{3}$$

where $S_D(D_j, D_l)$ is the disease similarities between diseases j and l, and $x_{il} = 1$ if the given drug i was known to be associated with disease l and $x_{il} = 0$ otherwise. Furthermore, the corresponding estimation error denoted as ε_1 can be defined as the standard deviation of the similarities between disease j with diseases known associated with drug i as below.

$$\varepsilon_1 = \sqrt{\frac{\sum_{l \in K_i} (S_D(D_j, D_l) - S_{ij})^2}{|K_i|}}, \tag{4}$$

where K_i denotes the disease set known associations with drug i, and $|K_i|$ denotes the number of elements.

By considering the possible estimation error, we obtained a convex set for each missing value as follows.

$$\begin{aligned} \Phi_1 = \{Y : S_{ij} - \varepsilon_1 \le Y_{ij} \le S_{ij} + \varepsilon_1, (i, j) \in \Omega; \\ Y_{ij} = 1, (i, j) \in K\}. \end{aligned} \tag{5}$$

The projection onto convex set Φ_1 is then given by the following equation.

$$P_1(Y_{ij}) = \begin{cases} S_{ij} - \varepsilon_1, & \text{for } \mathrm{Y}_{ij} < S_{ij} - \varepsilon_1; \\ S_{ij} + \varepsilon_1, & \text{for } \mathrm{Y}_{ij} > S_{ij} + \varepsilon_1; \\ Y_{ij}, & \text{otherwise.} \end{cases} \tag{6}$$

Likewise, for a given disease j, we can also define a similar projection $P_2(\cdot)$ based on the similarity between a drug with drugs known associated with the disease.

2.3.2 *Projections Based on Singular Value Decomposition*

After exploring the local pair-wise correlations, we further attempt to construct the convex sets by exploiting the global information of the matrix. By assuming the drug-disease associations as a low-rank matrix, we performed the low-rank approximation based on singular vector decomposition (SVD), which can be used to find the dominant components summarizing the matrix and then to predict the missing value by regressing against the dominant components. The SVD can be written as

$$Y \approx Y^{(r)} = \sum_{k=1}^{r} \sigma_k u_k v_k^T \tag{7}$$

where r is the approximated rank of Y, $U = \{u_1, \ldots, u_r\}$ is the left singular vectors, $V = \{v_1, \ldots, v_r\}$ is the right singular vectors, $D = \mathrm{diag}(\sigma_1, \cdots \sigma_r)$ is a diagonal matrix with positive singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. In fact, $Y^{(r)}$ gives the closest rank r matrix approximation to Y. Here, we used the SVD to impute the missing values where the r was selected based on the following inequality.

$$\frac{\sum_{t=1}^{r} \sigma_t^2}{\sum_{t=1}^{\min(n,m)} \sigma_t^2} \geq \alpha, \tag{8}$$

where $\alpha \in [0, 1]$ is the tolerance parameter.

Taking into the possible errors, we obtained a series of convex sets Φ_3^k for the missing values based on SVD where k is the iteration steps.

$$\begin{aligned}\Phi_3^k = \{Y : {Y_{ij}^{(r)}}^k - \varepsilon_3^k \leq Y_{ij} \leq {Y_{ij}^{(r)}}^k + \varepsilon_3^k, (i,j) \in \Omega;\\ Y_{ij} = 1, (i,j) \in K\},\end{aligned} \tag{9}$$

where the estimation ε_3^k is defined as

$$\varepsilon_3^k = \frac{\left\| Y - Y^{(r)} \right\|_F}{|\Omega|}. \tag{10}$$

The projection onto the convex set is then given by the following equation:

$$P_3^k(Y_{ij}) = \begin{cases} {Y_{ij}^{(r)}}^k - \varepsilon_3^k, & \text{if } \mathrm{Y}_{ij} < {Y_{ij}^{(r)}}^k - \varepsilon_3^k; \\ {Y_{ij}^{(r)}}^k + \varepsilon_3^k, & \text{if } \mathrm{Y}_{ij} > {Y_{ij}^{(r)}}^k + \varepsilon_3^k; \\ Y_{ij}, & \text{otherwise}. \end{cases} \tag{11}$$

2.3.3 Drug Repositioning with DrPOCS

The DrPOCS algorithm provides a convenient framework to utilize multiple pieces of prior to get an optimal solution. In this paper, by formulating the drug-disease associations as a low-rank matrix, we proposed a novel method to identify candidate indications of old drugs based on the iterative projections onto different convex sets as shown in Fig. 1C. With the integration of drug structure and disease phenotype information, the DrPOCS approach is able to predict potential associations between drugs and diseases with matrix completion. Specifically, with the projections described above, at the *k*th iteration, Y^{k+1} can be obtained with DrPOCS as follows:

$$Y^{k+1} = \frac{1}{N} \sum_{i=1}^{N} P_i(Y^k), \tag{12}$$

where N is the number of projections. $P_i(\cdot)$ are projections onto corresponding convex sets. The DrPOCS method can be summarized in the following algorithm.

In the above procedure, the DrPOCS algorithm provides a framework to adaptively select the preferred integration of chemical structure and disease phenotype information. In addition, it also considered the global information of the matrix. In this manner, a good solution between different prior information can be obtained.

3 Results

In this section, we applied the DrPOCS approach to predict potential drug-disease associations, and compared our DrPOCS approach with those popular existing approaches on real drug-disease association datasets. The details can be found below.

Algorithm 1. DrPOCS Based Method for Drug Repositioning

```
Input: Drug-disease association matrix, disease similarity matrix
       SimD, drug similarity matrix SimC, and parameter $\alpha$.
Repeat
  If $k = 0$,
    1. Select an initial matrix ($Y^0 = X$ by default).
    2. $k = k + 1$
  Else:
    1. For each kind of correlation $i$, compute the convex sets
       and the corresponding projections $P_i(\cdot)$.
    2. $Y^{k+1} = \frac{1}{N}\sum_{i=1}^{N} P_i(Y^k)$,
    3. $k = k + 1$
  End if
  Repeat unti $Y^{k+1} = Y^k$.
End
Output: Predicted scoring matrix.
```

3.1 Benchmark Results on Known Drug-Disease Associations

With the initial binary matrix X constructed as mentioned in Methods, we aimed to predict potential drug-disease associations (denoted by a matrix Y) based on the DrPOCS. Especially, we used the known associations from CTD database as the gold standard, where 80972 known drug-disease associations involving 656 approved human drugs and 1912 diseases were used as positive set while the other possible drug-disease pairs were treated as negative set.

3.1.1 Competitive Methods for the Performance Evaluation

To evaluate the performance of our DrPOCS approach, we compared it with four popular existing methods based on 10-fold cross-validation, including the nearest profile methods based on the chemical similarity $(\mathrm{NP}_{\_\mathrm{C}})$ and disease similarity $(\mathrm{NP}_{\_\mathrm{D}})$ [9], DrugNet [31] and NBI [32]. DrugNet is a network-based drug repositioning method, which integrates the information of diseases, drugs and targets to perform drug-disease prioritization [31]. NBI has been originally developed for predicting drug-target interactions, while it is also applicable for the prediction of drug-disease associations based on a bipartite graph [32]. All the five computational approaches were compared with respect to TPR (true positive rate), FPR (false positive rate) and AUC (area under ROC curve) score. It should be pointed out that, in each cross-validation trial, the projections onto different convex sets will be constructed again, without using the information about the test drug–disease associations. All methods are compared with the same data resources and definitions of drug and disease similarities.

3.1.2 Matrix Vectors Based Cross-Validation

To evaluate the performance of different methods systematically, we performed 10-fold cross-validation based on the vectors of the matrix. As shown in Fig. 2A, the drug-disease

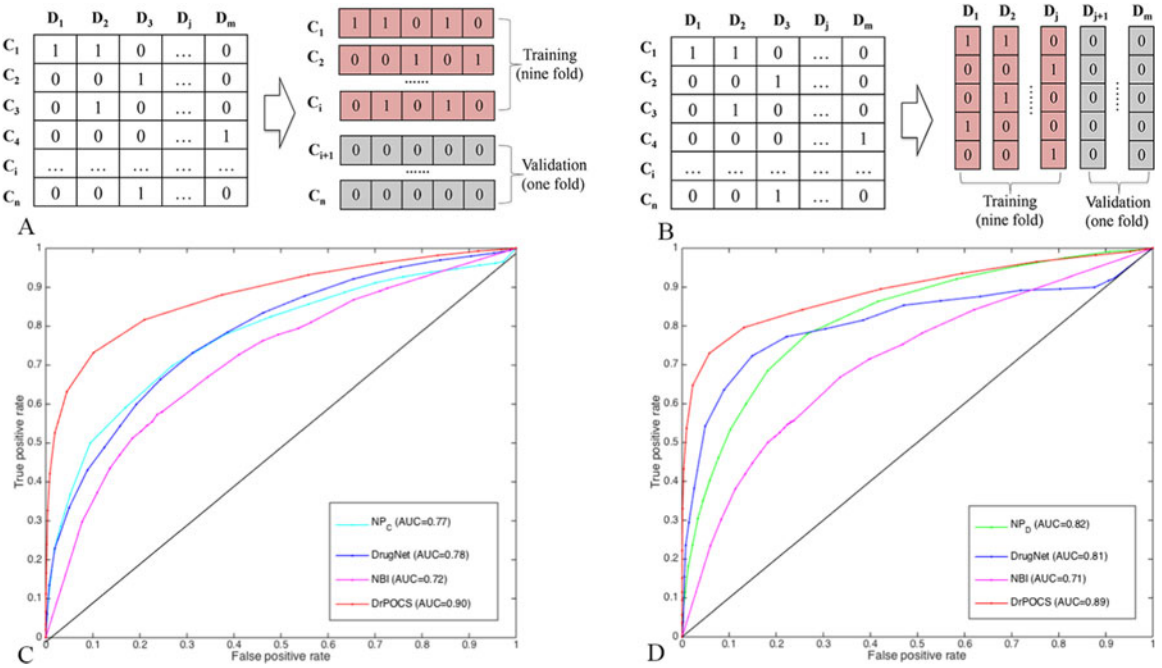

Fig. 2. The performance of different methods based on the 10-fold cross-validation performed with respect to matrix vectors. A. The illustration of 10-fold cross-validation based on rows, where each row was denoted as a drug vector and the entry in the vector represent the associations between the drug and all the diseases, with known associations were labeled by 1 and 0 otherwise. B. The illustration of 10-fold cross-validation based on columns where each column was denoted as a disease vector and the entry in the vectors represent the associations between the disease and all the drugs. C. The performance of different prediction approaches with respect to matrix rows, where DrPOCS is our proposed method, $\mathrm{NP}_{\_C}$ denotes the nearest profile method based on chemical similarity. D. The performance of different prediction approaches with respect to matrix columns, where $\mathrm{NP}_{\_D}$ represents the nearest profile method based on disease similarity. DrugNet and NBI are another two methods for drug repositioning based on known drug-disease associations.

association matrix was split into 10 subsets based on the rows. Here, each row was denoted as a drug vector where the entry in the vector represent the associations between the drug and all the diseases, with known associations were labeled by 1 and 0 otherwise. In each cross-validation trial, each fold of rows with unknown disease associations were taken in turn as the validation set while the rest subsets were treated as the training set. Then, DrPOCS was compared with the nearest profile method based on chemical similarity ($\mathrm{NP}_{\_C}$), NBI and DrugNet. The ROC curves for the four computational methods were shown in Fig. 2C. It can be seen that the POCS based method, with the highest AUC of 0.90, significantly outperforms all others, followed by DrugNet with AUC of 0.78. The performance of NBI is the worst with AUC of 0.72. In this comparison, the DrPOCS method demonstrated the performance for predicting new indications for known drugs.

Furthermore, by splitting the matrix into 10 subsets based on columns as shown in Fig. 2B, we compared DrPOCS method with the nearest profile method based on disease similarity ($\mathrm{NP}_{\_D}$), NBI and DrugNet. In this experiment, each column was denoted as a disease vector where the entry in the vectors represent the associations between the disease and all the drugs. Then the drug-disease association matrix was split into 10 subsets based on the columns while each fold of column with unknown drug associations were taken in turn as the validation set and the rest subsets with known drug associations were treated as the training set. The results were shown in Fig. 2D, from which we can see that our DrPOCS approach achieves the highest AUC of 0.89.

3.1.3 Matrix Elements Based Cross-Validation

In the above experiments, we either predicted new indications for known drugs or identified novel treatments for known diseases. We further performed 10-fold cross-validations based on the elements of the matrix, where the drug-disease associations were randomly split into 10 subsets with equal size, and each subset of associations was taken in turn as the test set while the other nine subsets were used to train our method. The five computational approaches, including $\mathrm{NP}_{\_C}$, $\mathrm{NP}_{\_D}$, DrugNet, NBI and DrPOCS method, were compared with the 10-fold cross-validation. Fig. 3 shows the results for the five methods, where our DrPOCS approach achieved the highest AUC of

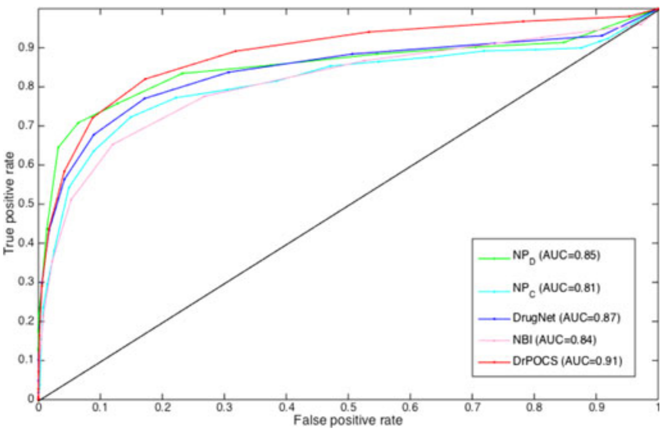

Fig. 3. The performance of different methods based on the 10-fold cross-validation performed with respect to matrix elements.

TABLE 1
The Performance of Different Methods Obtained by 10-Fold Cross-Validation Which was Performed with Respect to Matrix Elements

Method	AUC	Precision	Recall	F1 score
NP_C	0.8108	0.0921	**0.6332**	0.1608
NP_D	0.8523	0.3112	0.4876	0.3799
DrugNet	0.8668	0.4165	0.3522	0.3816
NBI	0.8391	0.4069	0.3461	0.3740
DrPOCS	**0.9109**	**0.6310**	0.4366	**0.5161**

AUC - Area under ROC curve; Precision - TP/(TP+FP); Recall - TP/(TP+FN); F1 score – Harmonic mean of precision and recall.

0.91, followed by DrugNet with AUC of 0.87. The performance of nearest profile method based on chemical similarity is the worst with AUC of 0.81. The detail results can be found in Table 1, from which we can see that DrPOCS has the highest AUC (0.9109), Precision (0.6310), F1 (0.5161) and significantly outperforms the other approaches. The results suggest that the DrPOCS is indeed effective for predicting new associations between drugs and diseases. Moreover, by comparing the 10-fold cross-validation under different scenarios, it is found that DrPOCS can integrate different kind of data very well and the results are not subject to the data perturbation, which further confirmed the stability of our method.

In our DrPOCS based approach, given an initial $n \times m$ binary matrix X describing known drug-disease associations, we predicted potential drug-disease associations with a low-rank matrix approximate to X, where the rank is smaller than $\min(n, m)$. The low-rank matrix was obtained with singular vector decomposition (SVD) [33]. Notably, the performance of DrPOCS depends on the choice of rank r, where the larget the rank r we chose, the smaller the squared Frobenius norm between X and $X^{(r)}$ will be obtained according to the Eckart–Young–Mirsky theorem [34]. However, it is undesirable when the rank r is too close to $\min(n, m)$ since it may obtain a strong approximation to the original matrix which is useless to uncover relationships. In other words, only the r most significant singular vectors were used here while the others were treated as noise. To determine the best rank r during SVD, the signal-to-noise ratio (SNR) based on Equation (8) was adopted in the 10-fold cross-validation. Table 2 shows the results achieved with different parameter α, from which we can see that the best results can be reached with $\alpha = 0.8$. Accordingly, the best rank r was determined with $\alpha = 0.8$.

3.2 The Novel Drug Indications Predicted by DrPOCS Approach

With the 10-fold cross-validation results above, a threshold of 0.4 was selected for DrPOCS approach to predict potential drug-disease associations, where the predictions with scores higher than the threshold were treated as candidate associations, with which resulting 354 drug-disease associations composed of 132 diseases and 213 drugs. The detail prediction results can be found in Supplementary File I.

TABLE 2
The 10-Fold Cross-Validation Results of DrPOCS Approach with Different Values of Parameter α in svd, Where the Cross-Validation was Performed with Respect to Matrix Elements

α	0.5	0.6	0.7	0.8	0.9
AUC	0.8818	0.8931	0.8952	**0.9109**	0.9007

3.2.1 *Validation of Potential Associations Based on Relationship between Targets and Disease Genes*

For the novel predictions, we first investigated the targets of drugs and genes associated with diseases according to DrugBank and CTD databases, respectively. In the past decades, it is well accepted that the novel indications can be identified for a drug if the drug can target a new molecule that is related to certain disease and numerous studies based on the 'new target new therapy' principle were proposed to reposition drugs [35], [36]. Thus, for a pair of drug and disease, if at least one disease gene is the target of the drug of interest, we assumed that the drug may be indeed useful for the disease and the prediction is confident [37]. For example, we have predicted Trandolapril to treat lung cancer. ACE is a gene related to neoplasms and is highly expressed in lung cancer [38], while Trandolapril is an ACE inhibitor used to treat high blood pressure. By targeting the gene ACE, the drug Trandolapril may have potential to be used for lung neoplasm.

In addition, previous studies have shown that the drug targets tend to be enriched in disease gene neighborhood across the protein-protein interaction network and vice versa, especially in cancer, cardiovascular and immune disorders [39]. It is also reported that the drug targets and their indicated disease genes with smaller distance (less than 3) in PPI network will tend to cause less side effects [40]. Therefore, the short distance between drug targets and disease gene can validate our novel associations to some extent. For example, it is found that the mutations in gene STAT3 are associated with hyperige recurrent infection syndromes (HIES) [41] while the drug Flutamide, acting as a selective antagonist of the androgen receptor (AR), is a nonsteroidal antiandrogen primarily used to treat prostate cancer [42]. Based on the interaction between STAT3 and AR, the Flutamide may be a candidate drug for HIES identified in our paper.

Furthermore, studies have also shown that drug treatment tend to perturb the activities of certain pathways rather than a single genes [43], thus the common pathways of drugs and disease can help us to identify the potential associations [44]. For a novel association, if the drug targets and disease genes participate in at least one common pathway, this association will be confident. For instance, the dysfunction of primary immunodeficiency pathway (hsa05340) may cause increased susceptibility to infection, autoimmune disease and malignancy. By affecting this pathway, the drug Triprolidine, originally used to relieve symptoms of allergy, hay fever and the common cold, may be demonstrated the new indications for autoimmune disease.

Taken together, 138 out of 354 associations can be validated by different types of relationship between drug target and disease genes.

3.2.2 Validation of Potential Associations Based on Similar Mode of Action

Moreover, it is well known that drugs with similar chemical structures may have similar mode of action (MoA) thus can be used to treat the same diseases. Based on this idea, we investigated whether our repositioned drugs have similar structures with the drugs approved for the certain diseases. For a predicted drug-disease association, if the drug has similar chemical structure with those approved for the disease, we can say that the drug may be used for the cancer, where the drugs were regarded as chemically similar if their structure similarity is above 0.7. Among our predictions, 164 associations were found to have similar structures with the known drugs, implying their possible effects of the certain disease. For example, the drug Propranolol, a beta-blocker used to affect the heart and circulation, has high chemical similarity of 0.841 with the drug Bisoprolol, indicating its new potential indication for the treatment of mycobacterium tuberculosis.

Besides, the drugs targeting the same protein may also have same MoA and therapy. Therefore, for a pair of drug and disease, if the drug has at least one target with those drugs approved for the disease, this drug may be indeed effective for the disease. For example, by targeting the same protein NR3C1, the drugs Methylprednisolone, Betamethasone, Prednisolone and Prednisone have all been used against the psoriatic arthritis. Among our predicted associations, the drug Amcinonide was found to target NR3C1, indicating this drug may be really effective for psoriatic arthritis. In accordance with this phenomenon, 154 associations are validated to be effective.

Except for the drug-centered approach, we can also infer that the drug may be used for the disease if it is similar to the one for which the drug is used to, where the diseases were regarded as similar if their phenotype similarity is above 0.5. Hence, it is not surprising that 101 associations can be validated to be effective. For example, based on the disease similarity of 0.602 between stroke and atrial fibrillation, the drug Aliskiren previously used to treat stroke, was inferred to be used for atrial fibrillation. As a whole, 207 associations can be validated based on the similar mode of actions.

3.2.3 Validation of Potential Associations Based on Text Mining

Except for the evidence mentioned above, to validate the compounds we repositioned for the diseases, we further performed text mining by querying the PubMed database to determine whether these compounds have been reported effective for corresponding diseases. As a result, 10 out of them have already been reported to be associated with the corresponding diseases. For instance, it is reported that the decreased level of Testosterone may lead to galactosemia which demonstrated the effectiveness of Testosterone in the treatment of galactosemia [45]. Except the effective associations, it is not surprising that DrPOCS method can also identify the side effects of the drugs which is another way for drug repositioning. For example, it is reported that the drug Metaraminol can increase the risk of the myocardial infarction caused by coronary artery vasospasm secondary [46].

To sum up, the evidences described above demonstrate that our repositioned compounds for the diseases are effective and are good candidates for future clinical trials. The novel predicted indications validated with various types of evidences can be found in Supplementary File I.

4 Discussion and Conclusion

Due to the time-consuming and high-risk in the traditional drug development, drug repositioning holds great potential for precision medicine in the post-genomic data. In this paper, we proposed DrPOCS, a set theoretic approach based on the projection onto convex sets, to predict the new indications of drugs. Specifically, DrPOCS based method exploited the global correlation structure of the drug-disease matrix based on the singular value decomposition. Furthermore, it also conveniently utilizes the biological constraints by considering the local correlation with the chemical structure and disease phenotype information. Compared with other existing methods by 10-fold cross-validation, DrPOCS has provided a superior performance. The results with different evidence have further shown that our proposed approach is able to successfully identify new indications for drugs, which demonstrate the effectiveness of our proposed approach for drug repositioning.

In addition, the success of our proposed method also provides expanded applications for the prediction of other biological associations, such as drug-gene, gene-disease and drug-side effect associations, by integrating additional similarity measures among diseases, genes, side effects and drugs. Furthermore, with the emergency and explainable of multidimensional data, DrPOCS can expand to predict the missing values of tensor data with calculating the relations between horizontal, lateral and frontal slices as well as the tensor low-rank decomposition.

We also noticed that there is much room to improve our method. The initial matrix used in the DrPOCS depends on the collected drug-disease associations while current known associations is often incomplete. Here, we only selected the quantitative associations rather than the qualitative associations. With more associations validated, the performance of our method will be improved. Moreover, it is expected that more convex sets with strong biological meanings could be constructed, thus the estimated values would be more relevant. We believe our methodology provides an alternative way to identify new indications for old drugs, which in turn can help to provide insights into the prediction of other associations.

Acknowledgments

This work was partly supported by the National Nature Science Foundation of China (61772368, 61572363, 91530321, 61602347), Natural Science Foundation of Shanghai (17ZR1445600), China Postdoctoral Science Foundation Funded Project (2016M601647), City University of Hong Kong (Project 7004862), and Hong Kong Research Grants Council (RGC) of Hong Kong (Project C1007-15G).

References

[1] A. Kamb, S. Wee, and C. Lengauer, “Why is cancer drug discovery so difficult?” *Nat. Rev. Drug Discovery*, vol. 6, pp. 115–120, 2007.

[2] K. I. Kaitin, "Deconstructing the drug development process: The new face of innovation," *Clin. Pharmacology Therapeutics*, vol. 87, no. 3, pp. 356–361, Mar. 2010.

[3] J. H. Kim and A. R. Scialli, "Thalidomide: The tragedy of birth defects and the effective treatment of disease," *Toxicol. Sci.*, vol. 122, no. 1, pp. 1–6, Jul. 2011.

[4] H. Pijl, S. Ohashi, M. Matsuda, Y. Miyazaki, A. Mahankali, V. Kumar, R. Pipek, P. Iozzo, J. L. Lancaster, A. H. Cincotta, and R. A. DeFronzo, "Bromocriptine: A novel approach to the treatment of type 2 diabetes," *Diabetes Care*, vol. 23, no. 8, pp. 1154–1161, Aug. 2000.

[5] Y. Y. Wang, J. C. Nacher, and X. M. Zhao, "Predicting drug targets based on protein domains," *Mol. Biosyst.*, vol. 8, no. 5, pp. 1528–1534, Apr. 2012.

[6] M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Sci.*, vol. 321, no. 5886, pp. 263–266, Jul. 11, 2008.

[7] L. Yang, J. Chen, L. Shi, M. P. Hudock, K. Wang, and L. He, "Identifying unexpected therapeutic targets via chemical-protein interactome," *PLoS One*, vol. 5, no. 3, 2010, Art. no. e9568.

[8] P. Imming, C. Sinning, and A. Meyer, "Drugs, their targets and the nature and number of drug targets," *Nat. Rev. Drug Discovery*, vol. 5, no. 10, pp. 821–834, Oct. 2006.

[9] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinf.*, vol. 24, no. 13, pp. i232–i240, Jul. 01, 2008.

[10] F. J. Azuaje, L. Zhang, Y. Devaux, and D. R. Wagner, "Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs," *Sci. Rep.*, vol. 1, 2011, Art. no. 52.

[11] X. D. Zhang, J. Song, P. Bork, and X. M. Zhao, "The exploration of network motifs as potential drug targets from post-translational regulatory networks," *Sci. Rep.*, vol. 6, Feb. 08, 2016, Art. no. 20558.

[12] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease," *Sci. Trans. Med.*, vol. 3, no. 96, Aug. 17, 2011, Art. no. 96ra76.

[13] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Sci. Trans. Med.*, vol. 3, no. 96, Aug. 17, 2011, Art. no. 96ra77.

[14] D. Shigemizu, Z. Hu, J. H. Hung, C. L. Huang, Y. Wang, and C. DeLisi, "Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer," *PLoS Comput. Biol.*, vol. 8, no. 2, Feb. 2012, Art. no. e1002347.

[15] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proc. Nat. Academy Sci. United States America.*, vol. 107, no. 33, pp. 14621–14626, Aug. 17, 2010.

[16] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork, "Drug-induced regulation of target expression," *PLoS Comput. Biol.*, vol. 6, no. 9, 2010, Art. no. e1000925.

[17] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, Jun. 07, 2011, Art. no. 496.

[18] C. Wu, R. C. Gudivada, B. J. Aronow, and A. G. Jegga, "Computational drug repositioning through heterogeneous network clustering," *BMC Syst. Biol.*, vol. 7, no. Suppl 5, 2013, Art. no. S6.

[19] J. Yang, Z. Li, X. Fan, and Y. Cheng, "Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization," *J. Chem. Inf. Model.*, vol. 54, no. 9, pp. 2562–2569, Sep. 22, 2014.

[20] W. Dai, X. Liu, Y. Gao, L. Chen, J. Song, D. Chen, K. Gao, Y. Jiang, Y. Yang, J. Chen, and P. Lu, "Matrix factorization-based prediction of novel drug indications by integrating genomic space," *Comput. Math. Methods Med.*, vol. 2015, Art. no. 275045, 2015.

[21] M. E. Wall, P. A. Dyck, and T. S. Brettin, "SVDMAN–Singular value decomposition analysis of microarray data," *Bioinf.*, vol. 17, no. 6, pp. 566–568, Jun. 2001.

[22] M. Gonen, "Predicting drug-target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinf.*, vol. 28, no. 18, pp. 2304–2310, Sep. 15, 2012.

[23] Y. H. Taguchi, M. Iwadate, and H. Umeyama, "Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease," *BMC Bioinf.*, vol. 16, Apr 30, 2015, Art. no. 139.

[24] X. Gan, A. W. Liew, and H. Yan, "Microarray missing data imputation based on a set theoretic framework and biological knowledge," *Nucleic Acids Res.*, vol. 34, no. 5, pp. 1608–1619, 2006.

[25] B. J. van Wyk, and M. A. van Wyk, "A POCS-based graph matching algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1526–1530, Nov. 2004.

[26] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: Shedding new light on drug metabolism," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D1091–D1097, Jan. 2014.

[27] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Comparative toxicogenomics database: A knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D786–D792, Jan. 2009.

[28] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Hum. Genet.*, vol. 14, no. 5, pp. 535–542, May 2006.

[29] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 52–55, Jan. 01, 2002.

[30] S. Kim, L. Han, B. Yu, V. D. Hahnke, E. E. Bolton, and S. H. Bryant, "PubChem structure-activity relationship (SAR) clusters," *J. Cheminform.*, vol. 7, 2015, Art. no. 33.

[31] V. Martinez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-based drug-disease prioritization by integrating heterogeneous data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 41–49, Jan. 2015.

[32] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol*, vol. 8, no. 5, 2012, Art. no. e1002503.

[33] Y. Wang, D. Yang, and M. Deng, "Low-rank and sparse matrix decomposition for genetic interaction data," *Biomed. Res. Int.*, vol. 2015, 2015, Art. no. 573956.

[34] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank.," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[35] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, and P. E. Bourne, "Drug discovery using chemical systems biology: Repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis," *PLoS Comput. Biol.*, vol. 5, no. 7, Jul. 2009, Art. no. e1000423.

[36] J. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker, and R. A. Houghten, "Shifting from the single to the multitarget paradigm in drug discovery," *Drug Discovery Today*, vol. 18, no. 9-10, pp. 495–501, May 2013.

[37] R. R. Brinkman, M. P. Dube, G. A. Rouleau, A. C. Orr, and M. E. Samuels, "Human monogenic disorders - a source of novel drug targets," *Nat. Rev. Genet.*, vol. 7, no. 4, pp. 249–260, Apr. 2006.

[38] G. Ucar, Z. Yildirim, E. Ataol, Y. Erdogan, and C. Biber, "Serum angiotensin converting enzyme activity in pulmonary diseases: Correlation with lung function parameters," *Life Sci.*, vol. 61, no. 11, pp. 1075–1082, 1997.

[39] J. Sun, K. Zhu, W. Zheng, and H. Xu, "A comparative study of disease genes and drug targets in the human protein interactome," *BMC Bioinf.*, vol. 16, no. Suppl 5, 2015, Art. no. S1.

[40] X. Wang, B. Thijssen, and H. Yu, "Target essentiality and centrality characterize drug side effects," *PLoS Comput. Biol.*, vol. 9, no. 7, 2013, Art. no. e1003119.

[41] A. F. Freeman, and S. M. Holland, "The hyper-IgE syndromes," *Immunol. Allergy Clin. North Am.*, vol. 28, no. 2, pp. 277–291, May 2008.

[42] B. R. Goldspiel, and D. R. Kohler, "Flutamide: An antiandrogen for advanced prostate cancer," *DICP*, vol. 24, no. 6, pp. 616–623, Jun. 1990.

[43] N. Pratanwanich, and P. Lio, "Pathway-based bayesian inference of drug-disease interactions," *Mol. Biosyst.*, vol. 10, no. 6, pp. 1538–1548, Jun. 2014.

[44] Y. Pan, T. Cheng, Y. Wang, and S. H. Bryant, "Pathway analysis for drug repositioning based on public database mining," *J. Chem. Inf. Model.*, vol. 54, no. 2, pp. 407–418, Feb. 24, 2014.

[45] C. S. Gubbels, C. K. Welt, J. C. Dumoulin, S. G. Robben, C. M. Gordon, G. A. Dunselman, M. E. Rubio-Gozalbo, and G. T. Berry, "The male reproductive system in classic galactosemia: Cryptorchidism and low semen volume," *J. Inherit Metab. Dis.*, vol. 36, no. 5, pp. 779–786, Sep. 2013.

[46] A. Khavandi, J. J. Gatward, J. Whitaker, and P. Walker, "Myocardial infarction associated with the administration of intravenous ephedrine and metaraminol for spinal-induced hypotension," *Anaesthesia*, vol. 64, no. 5, pp. 563–566, May 2009.

Yin-Ying Wang received the PhD degree from Shanghai University, Shanghai, China. She is a postdoctoral in the Department of Computer Science, Tongji University. Her research focuses on data mining and computational systems biology. She has published more than 10 journal and conference papers.

Chunfeng Cui received the PhD degree from the Chinese Academy of Sciences, Beijing, China, in 2016. Her research interests include numerical optimization, tensor eigenvalues, tensor computations, and bioinformatics.

Liqun Qi received the PhD degree from the University of Wisconsin-Madison, in 1984. He is the chair professor and head of the Department of Applied Mathematics, The Hong Kong Polytechnic University. He has published more than 150 research papers in international journals. His research interests include multilinear algebra, tensor computation, and nonlinear numerical optimization and applications.

Hong Yan (F'06) received the PhD degree from Yale Universiy, New Haven, CT, USA. He was a professor of imaging science of computer engineering with the City University of Hong Kong, Kowloon, Hong Kong. His research interests include image processing, pattern recognition, and bioinformatics. He is a Fellow of the Internation for Pattern Recognition and the IEEE.

Xing-Ming Zhao received the PhD degree from the University of Science and Technology of China, Hefei, China. He is a professor of the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University. His research focuses on data mining and computational systems biology. He has published more than 50 journal papers. He is an editorial board member of several jourals and senior member of the IEEE.