A Novel Drug Repositioning Approach Based on Collaborative Metric Learning

Huimin Luo, Jianxin Wang, Cheng Yan, Min Li, Fang-Xiang Wu, and Yi Pan

Abstract—Computational drug repositioning, which is an efficient approach to find potential indications for drugs, has been used to increase the efficiency of drug development. The drug repositioning problem essentially is a top-K recommendation task that recommends most likely diseases to drugs based on drug and disease related information. Therefore, many recommendation methods can be adopted to drug repositioning. Collaborative metric learning (CML) algorithm can produce distance metrics that capture the important relationships among objects, and has been widely used in recommendation domains. By applying CML in drug repositioning, a joint metric space is learned to encode drug's relationships with different diseases. In this study, we propose a novel drug repositioning computational method using Collaborative Metric Learning to predict novel drug-disease associations based on known drug and disease related information. Specifically, the proposed method learns latent vectors of drugs and diseases by applying metric learning, and then predicts the association probability of one drug-disease pair based on the learned vectors. The comprehensive experimental results show that CMLDR outperforms the other state-of-the-art drug repositioning algorithms in terms of precision, recall, and AUPR.

Index Terms—Drug repositioning, collaborative metric learning, latent vectors

1 Introduction

Drug discovery and development is a complicated, time consuming and expensive process involving the identification of candidates, synthesis, characterization, screening, evaluation of therapeutic efficacy and clinical trials [1], [2], [3]. The total average cost for bringing a new drug to marketing approvals is rising rapidly [1], and is estimated to be $2.6 billion in 2013 by the Tufts Center for the Study of Drug Development group [4]. However, the success rates of drug development is extremely low. In addition, many investigational drugs have failed due to inadequate efficacy, safety concerns and commercial reasons in clinical phrases [5].

Drug repositioning, which is an alternative drug development strategy, aims at identifying novel uses for existing drugs, and can reduce risk and costs of development of new drugs [6], [7]. With ever-increasing of large-scale biological and chemical information, this offers new opportunities for researchers to develop computational drug repositioning methods [8], [9]. In recent years, many methods based on various computational models have been proposed to perform drug repositioning efficiently [10], [11], [12]. For instance, based on the observation that similar drugs tend to indicate for similar diseases, Gottlieb et al. [13] proposed a drug repositioning method, PREDICT, to identify potential indications for drugs. The proposed method constructed features of drug-disease associations based on multiple drug-drug and disease-disease similarity measures, and trained a logistic regression classifier to predict new drug-disease associations. Wang et al. [14] formalized drug-disease prediction as a binary classification problem, and constructed a predictor for drug repositioning (PreDR) to infer potential drug-disease associations by fusing heterogeneous data. Drug similarity profiles were computed by integrating chemical structures, target proteins, and side-effects. In addition, disease similarity profiles were computed based on phenotype data. Then, similarity between two drug-disease pairs were calculated as Kronecker product kernel, and SVM was applied to identify novel drug-disease associations.

Many network-based drug repositioning methods have been presented to infer new uses for existing drugs by applying biological networks which can model associations between biological concepts [11]. For example, Wang et al. [15] constructed a disease-drug-target heterogeneous network consisting of interaction and similarity network by integrating disease, drug and target related information. Based on the constructed network, a Triple Layer Heterogeneous Graph Based Inference method, TL_HGBI, was proposed to predict candidate drugs for diseases. Inspired by the network-based random walk algorithm, Liu et al. [16]

- *H. Luo is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China and also with the School of Computer and Information Engineering, Henan University, KaiFeng 475001, China. E-mail: luohuimin@csu.edu.cn.*
- *J. Wang, C. Yan, and M. Li are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: {jxwang, yancheng01, limin}@mail.csu.edu.cn.*
- *F. Wu is with the Division of Biomedical Engineering, Department of Mechanical Engineering, University of Saskatchewan, Saskatoon SKS7N5A9, Canada. E-mail: faw341@mail.usask.ca.*
- *Y. Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302 USA. E-mail: yipan@gsu.edu.*

Manuscript received 29 Jan. 2019; revised 22 May 2019; accepted 18 June 2019. Date of publication 2 July 2019; date of current version 1 Apr. 2021. (Corresponding author: Jianxin Wang.)
Digital Object Identifier no. 10.1109/TCBB.2019.2926453

TABLE 1
Statistics of the Benchmark Standard Dataset Used in this Study

Dataset	Drugs	Diseases	Drug-disease interactions
	2,159	573	5,943

proposed a two-pass random walk method, TP-NRWRH, to predict novel indications for drugs based on the drug-disease heterogeneous network. The association probability of one drug-disease pair was predicted by performing drug-centric and disease-centric random walks. Shahreza et al. [17] integrated drug, disease, and target information to develop heterogeneous label propagation algorithm (HeterLP), and used it to infer associations between drugs, targets and diseases. HeterLP improved similarities by using the projection technique, and applied the label propagation on each constructed sub-network to predict association probabilities between different entities. Many computational drug repositioning methods are developed based on the assumption that similar drugs tend to associate with similar diseases. Therefore, similarity measures are very important for these similarity-based methods. Luo et al. [18] developed comprehensive similarity measures to improve the accuracy of drug similarities and disease similarities, and then applied bi-random walk algorithm on the constructed drug-disease network to infer novel drug-disease associations.

Essentially, the drug repositioning problem can be formulated as a recommendation task aiming to recommend potential indications to drugs. By constructing a recommendation system, the association probability of one drug-disease pair can be predicted, and then the most likely diseases are recommended to one drug based on their association probabilities. In recent studies, collaborative filtering (CF) methods, such as matrix factorization and low rank matrix completion approaches, have been widely used in drug-target prediction [19], [20], [21], [22], [23], drug-disease prediction [24], [25], and other prediction domains [26], [27], [28]. For example, Gönen et al. [19] proposed a kernelized Bayesian matrix factorization with twin kernels method, KBMF2K, to predict novel drug-target interactions. Based on the kernels for drug compounds and target proteins, KBMF2K projected them into a unified subspace. Then low-dimensional representations of drugs and targets were used to infer drug-target interactions. Yang et al. [20] constructed weighted causal network to compute chemical-disease association scores. The latent variables of chemicals and diseases were learned by applying probabilistic matrix factorization model. Based on these latent variables, chemical-disease association types could be predicted. Then, computed association scores and types were used for drug repositioning predictions. Liu et al. [21] have proposed a neighborhood regularized logistic matrix factorization method, NRLMF, to predict novel drug-target interactions. NRLMF applied logistic matrix factorization to learn latent vectors of drugs and targets, and computed interaction probability of drug-target pairs. Luo et al. [23] have proposed a network integration approach, DTINet, to learn low-dimensional feature representations for drugs and targets. Based on these learned features, DTINet then predicted novel drug-target interactions by applying inductive matrix completion method [29]. By formulating the drug repositioning problem as a matrix completion problem, Luo et al. [24] constructed drug-disease heterogeneous network by incorporating drug similarity, disease similarity and drug-disease associations, and developed a drug repositioning recommendation system (DRRS), to recommend potential diseases to drugs.

According to studies [30], matrix factorization approaches may be unsuccessful in capturing finer-grained preferences of users when violating the triangle inequality which stated that for any three objects, the distance between any two objects should not be larger than the sum of distances of the other two pairs. To solve the problem for matrix factorization, collaborative metric learning (CML) has been proposed and applied in recommendation domains successfully [30]. CML could learn a joint metric to uncover the underlying relationships between objects, and has achieved superior accuracy over the state-of-the-art collaborative filtering algorithms in recommendation domains.

In this study, we proposed a novel drug repositioning method, collaborative metric learning based drug repositioning (CMLDR), which can recommend potential indications for known drugs having validated disease associations, and new drugs without known associations. The main contributions of this paper involves: (1) our proposed method can capture drug's hidden relationships with different diseases in a more intuitive way; (2) for recommending potential diseases to known drugs, our proposed method can perform recommendation effectively by using only known drug-disease associations; (3) for recommending potential diseases to new drugs, CMLDR only needs drug-related data, and can outperform the state-of-the-art drug repositioning methods which utilize both drug-drug similarity, disease-disease similarity and drug-disease association information.

2 Materials and Methods

In this study, we propose a novel drug repositioning computational approach, CMLDR, to identify potential indications for drugs. First, we give a brief description of the collected benchmark dataset. Then, a novel drug repositioning method based on the collaborative metric learning algorithm [30] is designed and utilized to infer potential drug-disease associations.

2.1 Dataset

The Comparative Toxicogenomics Database (CTD) is a publicly available database involving substantial data describing chemicals, genes, and human diseases, as well as relationships between them [31]. We extracted all chemical-disease associations with therapeutic type, which indicates that a chemical has therapeutic role in a disease, from CTD. Then, the collected chemical-disease associations are filtered through choosing chemicals with DrugBank [32] identifiers and diseases with OMIM identifiers. The filtered associations are integrated with data obtained from paper [13] to construct the benchmark dataset. Table 1 summarizes the dataset in terms of numbers of drugs, diseases and drug-disease interactions.

In this study, drug similarity is calculated based on chemical structural information, and we collect structural similarities between drugs from published literature [33].

For two drugs, A and B denote their chemical fingerprints, the structural similarity between them is calculated by Tanimoto coefficient defined as $|A \cap B|/|A \cup B|$, that is

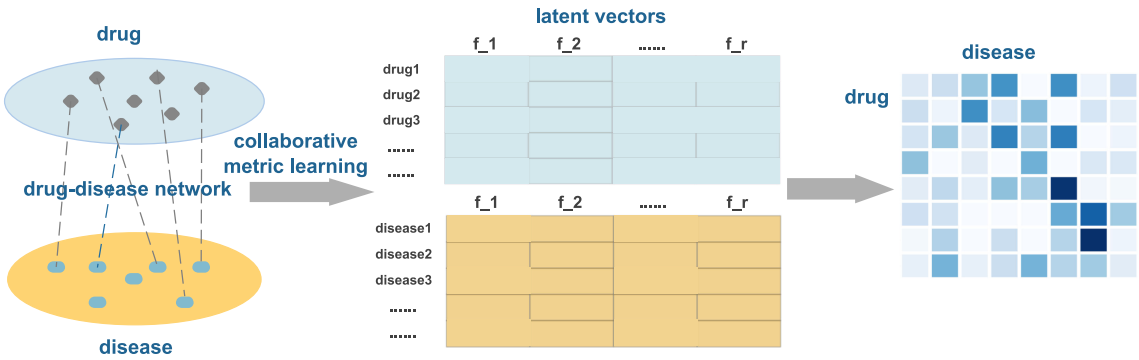

Fig. 1. Based on known/positive associations in the drug-disease heterogenous network, latent vectors of drugs and disease can be learned by applying metric learning. Then, the association probability of one drug-disease pair is predicted according to their latent vectors.

the number of common chemical fingerprints divided by the number of total chemical fingerprints of the given two drugs [33].

2.2 Problem Formalization

The drug-disease network is modeled as a bipartite graph $G_{rd}(D_r, D_d, E)$, where $D_r = \{r_1, r_2, \ldots, r_p\}$ denotes p drugs, $D_i = \{d_1, d_2, \ldots, d_t\}$ denotes t diseases, $E(G) \subseteq D_r \times D_i$, $E(G) = \{e_{ij}\}$ contains edges between drug r_i and disease d_j, the weight of e_{ij} is initially set to 1 if there exists a known association between drug r_i and disease d_j, otherwise 0. Let $S = \{(r_i, d_j)\}$ be a collection of positive drug-disease pairs which have known associations in benchmark dataset.

Metric learning aims to learn a metric function that keeps similar objects closer and dissimilar objects far apart [30]. In this study, for recommending potential diseases to drugs, its objective is to learn one distance metric that makes the distances between positive drug-disease pairs smaller, and distances between other pairs larger. Based on the learned latent vectors of drugs and diseases, the association probability of a drug and a disease can be predicted by computing the squared Euclid distance between them. A schematic description of the whole process is described in Fig. 1. For a given drug, all candidate diseases having no validated associations with it are ranked according to the predicted probability scores in descending order. The diseases ranked in top-K are considered more likely to associate with the given drug.

2.3 Collaborative Metric Learning based Drug Repositioning

Based on the set of known drug-disease pairs S with positive associations, collaborative metric learning attempts to learn a drug-disease joint metric to encode these associations. Specifically, the learned metric which observes the known positive relationships, pulls the set of positive drug-disease pairs having known associations closer and pushes the other pairs away.

2.3.1 Model Formulation

All drugs and diseases are projected into the drug-disease joint space with a low dimensionality n. Each drug and each disease are described by a drug vector $u_i \in R^n$ and a disease vector $v_j \in R^n$, respectively. Then, the association probability of drug r_i and disease d_j is represented by the squared euclidean distance between them.

To learn these latent vectors of drugs and diseases, the euclidean distance between drug r_i and disease d_j, i.e., $d(i,j) = \|u_i - v_j\|$, needs to respect the relationships of r_i with diseases. That is, for drug r_i, it should be closer to its known associated diseases than diseases without associations with it. By using loss function described in [30], the constraint is converted into the following loss function:

$$L_m = \sum_{(i,j)\in S} \sum_{(i,k)\notin S} w_{ij}[m + d(i,j)^2 - d(i,k)^2]_+, \tag{1}$$

where S denotes the set of known drug-disease associations, j is one disease associated with drug i, k is one disease having no association with drug i, $[z]_+ = \max(z, 0)$ represents the standard hinge loss, $m(m > 0)$ denotes the safety margin size, and w_{ij} is ranking loss weight.

For a given drug r_i, its disease profile indicates the presence or absence of association with every disease, diseases having known associations with it are regarded as positive diseases, negative diseases are randomly chosen from diseases that have no known associations. The loss function defined above would pull r_i's positive diseases closer to it, and push it's negative diseases away from it until they are beyond the defined safety margin m.

A rank-based weighting scheme, called Weighted Approximate-Rank Pairwise (WARP) loss, has been proposed to approximate ranks of items, and the trained models with it have achieved better performance [34]. Therefore, WARP loss is applied in training prediction model, which could penalize positive diseases predicted to be at a lower rank. After performing prediction, all diseases are sorted by their predicted association probabilities with drug r_i in descending order. Then, positive disease d_j is penalized based on its rank as follows

$$w_{ij} = \log\left(rank(i,j) + 1\right), \tag{2}$$

where $rank(i,j)$ is the rank of disease d_j under consideration of drug r_i. Specially, for drug r_i, the lower rank of d_j means that d_j ranks behind more diseases. From this equation, we can see that one positive disease with a lower rank would be penalized heavily, while one positive disease with a higher rank gets slighter penalty.

The procedure of estimating $rank(i,j)$ is performed as follows: (1) For each positive drug-disease pair (i,j), M negative diseases are sampled from diseases without associations with drug r_i, thus M negative drug-disease pairs (i,k)

are produced, k is one negative disease; (2) According to positive pair (i, j) and M negative pairs (i, k), the loss values of M negative pairs are obtained by computing the hinge loss defined in Equation (1); (3) Let P denote the number of nonzero loss values, then $rank(i, j)$ is approximated as $\lfloor \frac{t \times P}{M} \rfloor$, t is the number of diseases in the benchmark dataset.

The regularization scheme is defined as in [30]. Specifically, let u_* denote the latent vector of one drug, v_* denote the latent vector of one disease, U represent the latent vectors of all drugs, and V represent the latent vectors of all diseases. To ensure the feasibility and robustness of the proposed recommendation model [30], u_* and v_* are bounded within a sphere.

$$\|u_*\|_2 \leq l \; and \; \|v_*\|_2 \leq l,$$

where $\|*\|_2$ represents L2 norm, and l controls the size of the sphere. Then we define the objective function of the recommendation model as follows:

$$\begin{aligned} \min_{u_*, v_*} \quad & L_m \\ s.t. \quad & \|u_*\|_2 \leq l \; and \; \|v_*\|_2 \leq l. \end{aligned}$$

The constrained objective function is minimized with Mini-Batch Stochastic Gradient Descent (SGD). The learning rating is controlled using AdaGrad [35], which is a gradient-based optimization algorithm having the adaptive learning rate.

2.3.2 CMLDR

The prediction model is constructed based on drug-disease associations and drug-drug similarity data, and the top-ranked predictions are considered as candidate indications and recommended to drugs. All known/positive drug-disease associations are split into training set *Pset* and validation set *Vset*. In the training procedure, N positive drug-disease associations are sampled from *Pset*. For each positive association (r_i, d_j), we sample M negative diseases having no associations with drug r_i. These positive and negative samples are used to compute the gradients and update latent vectors of drugs and diseases. This procedure is repeated until convergence, which means that the prediction performance on *Vset* becomes stable. The parameters safety margin size m, N, and M are set to 2.0, 20 and 10 by default in this study.

CMLDR can be used to identify potential diseases for approved drugs having validated associations and for new drugs without any known disease associations. For one new drug, we learn its latent vector using the latent vectors of its kn nearest neighbors which are drugs in the training set. The set of kn nearest neighbors of drug r_i is denoted by $Neighbor_{kn}(r_i)$.

The drug-disease recommendation method based on collaborative metric learning is illustrated in Algorithm 1. Once the latent vectors U and V of all drugs and diseases have been learned, the association probability score of any unknown drug-disease pair (r_i, d_j) can be predicted by calculating the squared euclidean distance between the r_i and d_j.

Algorithm 1. Identify Potential Indications for Drugs Using CMLDR

Input: drug-disease association matrix A_{rd}, drug-drug similarity matrix S^r.
Output: drug latent vectors U, disease latent vectors V, predicted drug-disease distance matrix R.
*/*extract positive associations to construct training set and validation set;*/*
$[Pset, Vset]$ = ConstructSamples (A_{rd})
Initialize U, V with random normal values
*/*normalize drug similarity matrix;*/*
*/*diagonal matrix D^r, D^r_{ii} is the sum of row i of S^r;*/*
$S'^r = (D^r)^{-1/2} S^r (D^r)^{-1/2}$
while *has not achieved stable performance on Vset* **do**
  */*sample N associations from Pset;*/*
  Tset ← Sample(*Pset*, *N*)
  for *association in Tset* **do**
    */*sample M negative diseases;*/*
    Nset* ← Sample (r_i, M)
    */*keep the closest negative disease;*/*
    Nset ← KeepClosest(*Nset**)
  end
  */*compute ranking loss weight for each positive association in Pset;*/*
  W ← RankWeight()
  */*compute gradients and update U, V;*/*
  Update(U, V)
end
for r_i *in* D_r **do**
  if *new* (r_i) **then**
    */*update latent vector of r_i based on that of its neighbors in Pset;*/*
    $U_{r_i} = \frac{\sum_{r_j \in Neighbor_{kn}(r_i)} S'^r_{ij} U_{r_j}}{\sum_{r_j \in Neighbor_{kn}(r_i)} S'^r_{ij}}$
  end
end
*/*compute distances between drugs and diseases;*/*
$R_{ij} = \|U_i - V_j\|^2$
return U, V, R

3 Experiments and Results

We conduct comprehensive experiments to evaluate CMLDR's performance. The results demonstrate CMLDR's superior accuracy over the state-of-the-art drug repositioning approaches on the benchmark dataset.

3.1 Evaluation Methodology

To systematically evaluate the ability of CMLDR in recommending candidate diseases to drugs, precision and recall rates for Top-K recommendations are used as the performance evaluation metrics. For drug-disease association prediction, the top-ranked candidate diseases of drugs are more interested in practice. Therefore, for one drug, it makes more sense to evaluate prediction results of the Top-K diseases instead of all the diseases.

According to [30], known associations for each drug are split into training, validation, and test sets, which contained 60, 20, and 20 percent of associations, respectively. In this case, for drugs with less than 5 known associations, their

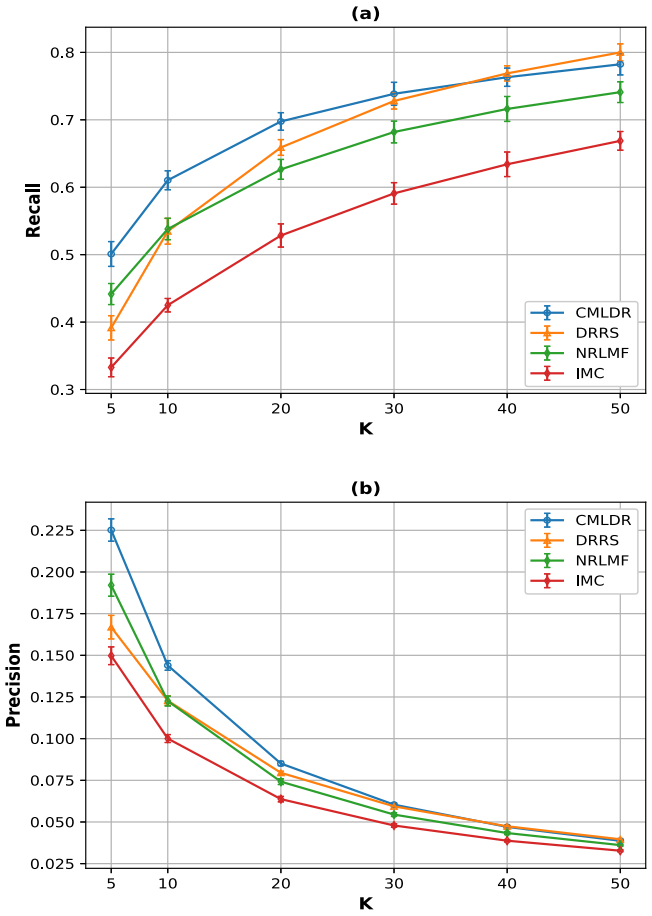

Fig. 2. Prediction evaluation results in terms of Recalls (a) and Precisions (b) for different methods in recommending potential diseases for known drugs having validated interactions under various K.

associations are added to the training set. Those unknown drug-disease associations are considered as candidate associations. After learning based on the training and validation sets, we obtain the latent vectors of drugs and diseases. We utilize a drug's latent vector and a disease's latent vector to compute their squared euclidean distance, which denotes the association probability of this drug-disease pair. For one drug, the test associations are ranked together with candidate associations, and are sorted in descending order according to their predicted distance values. For one specific ranking threshold K, true positive (TP), false negative (FN), and false positive (FP) are calculated based on the ranking results of associations. For one test association predicted with a higher rank than K, it is regarded as a correctly identified positive sample. For one candidate association predicted with a lower rank than K, it is regarded as a correctly identified negative sample. The number of positive samples identified correctly is denoted by TP, the number of positive samples identified incorrectly is denoted by FP, and the number of negative samples identified incorrectly is denoted by FN. Thus the precision and recall at top K can be calculated as follows:

$$Precison = TP/(TP + FP)$$

$$Recall = TP/(TP + FN).$$

To obtain reliable results, the above experiment procedure is repeated ten times. In addition, the area under the Precision-Recall curve (AUPR) and area under the receiver operating characteristic curve (AUC) are also used to evaluate the performance of the methods. It has been reported that AUPR can provide a more informative assessment than AUC for the highly imbalanced dataset [36]. The benchmark dataset used in this study is imbalanced, for the number of known drug-disease associations is far less than the number of unknown associations. Therefore, we put more emphasis on AUPR than on AUC in the following experiments.

3.2 Comparison with other Methods

To validate the prediction performance of CMLDR, we compare it with three state-of-the-art methods: IMC [29], NRLMF [21], and DRRS [24], which have been applied in drug repositioning. Inductive matrix completion (IMC) generates the association matrix by applying related feature vectors to a low-rank matrix, and has been used successfully for disease-gene prediction and drug-target prediction. Neighborhood regularized logistic matrix factorization (NRLMF) method focuses on modeling the probability that a drug would interact with a target by logistic matrix factorization, where the properties of drugs and targets are represented by drug-specific and target-specific latent vectors, respectively. Although NRLMF is originally developed for drug-target association prediction, it can be applied in predicting candidate diseases for drugs. DRRS adopts a fast Singular Value Thresholding (SVT) algorithm to complete the drug-disease adjacency matrix, and recommends potential diseases to drugs based on the completed scores.

By applying evaluation metrics introduced in previous section, we evaluate the performance of all methods in predicting novel drug-disease associations. The experiment results in terms of precision rates, recall rates and standard deviations at various top-K predictions are depicted in Fig. 2.

These experiment results show that our proposed CMLDR method outperforms the other methods in terms of recalls and precisions. Specifically, CMLDR has achieved an average recall value of 0.501 at top-5, while DRRS, NRLMF, and IMC obtain inferior results of 0.392, 0.442 and 0.333, respectively. The precision results show that CMLDR achieves the best precision with 0.225 at top-5, indicating that 22.5 percent true drug-disease associations are successfully ranked in the top 5. DRRS, NRLMF, and IMC have precision results of 0.167, 0.192 and 0.150, respectively. In addition, the average AUPR values and AUC values obtained by all methods are reported in Table 2. We find that CMLDR has achieved the best average AUPR value compared with the other methods. The average AUPR value achieved by CMLDR is 0.340, and NRLMF obtains the second best AUPR of 0.290. In terms of AUC, although CMLDR shows a ~ 3 percent decrease of AUC compared with DRRS, it achieves a ~ 9 percent increase of AUPR. In this study, AUPR is considered to be more important than AUC for the imbalanced benchmark dataset.

Moreover, we compared CMLDR with other methods in terms of running time on the benchmark dataset. The experimental results are reported in Supplementary Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2926453. The results show CMLDR is faster than

TABLE 2
The Average AUPR and AUC Values Obtained by Various Methods in Predicting Diseases for Known Drugs Having Verified Associations

	CMLDR	DRRS	NRLMF	IMC
AUPR	**0.340**	0.253	0.290	0.219
AUC	0.896	**0.925**	0.888	0.853

The best result is in bold.

matrix completion method DRRS, and takes a similar amount of time compared with other methods.

3.3 Parameter Sensitivity Analysis for Dimension n

The objective of CMLDR is to project drugs and diseases to a joint n-dimensional space. In this section, we focused on the sensitivity analysis of the latent space dimension n in recommending potential diseases to known drugs. The impact of the dimensionality of the latent space n on the performance of CMLDR, in terms of recalls and precisions, is shown in Fig. 3. We find that larger n generally achieves better results, and the prediction performance of CMLDR becomes stable when n is greater than 50. Thus, the parameter n is recommended to be set in the range [50, 100]. In this study, we set the parameter n to 100.

3.4 Predicting Indications for New Drugs

Drugs with at least 5 known disease associations are selected to evaluate the prediction performance of CMLDR for new drugs. Specially, 80 percent of selected drugs are used to train and validate prediction model, and 20 percent of selected drugs are used as the test set. The drugs in test set are considered as new drugs without any known associations. This evaluation process is repeated ten times to obtain reliable results.

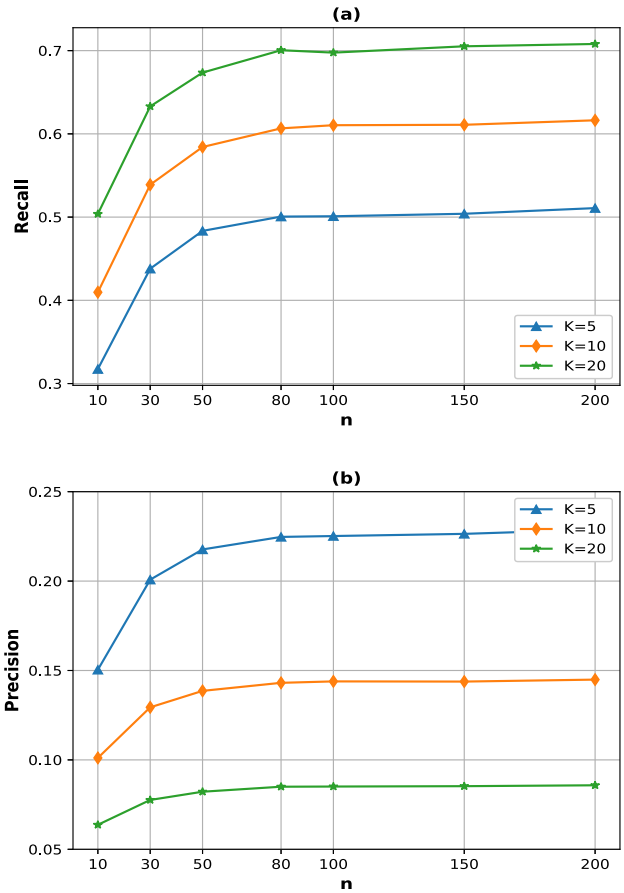

Fig. 3. Performance trend of CMLDR on the benchmark dataset measured by recalls and precisions with different settings of dimension n.

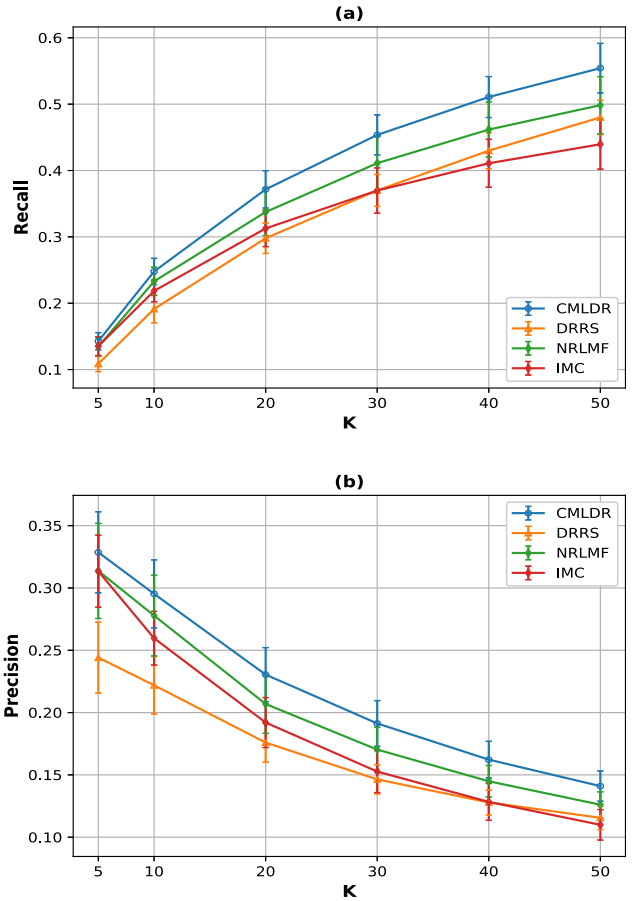

Fig. 4. Prediction evaluation results in terms of Recalls (a) and Precisions (b) for different methods in recommending potential diseases for new drugs without validated associations under various K.

For CMLDR, the latent vectors of new drugs are learned by utilizing the latent vectors of its kn nearest neighbors, and these neighbors are drugs in the training set. The parameter kn is set to 5 defaultly in this study. After prediction, the results in terms of Recall as well as Precision are reported in Figs. 4a and 4b, respectively. These results show that CMLDR has achieved superior performance over the other methods at various K. For example, CMLDR achieves an average recall value of 0.143 at $K = 5$, while DRRS, NRLMF, and IMC obtain inferior recall values of 0.109, 0.135 and 0.135, respectively. At $K = 50$, CMLDR achieves a recall value of 0.554, while DRRS, NRLMF, and IMC obtain inferior recall values of 0.48, 0.499 and 0.44, respectively. Moreover, CMLDR achieves precision values of 0.329 at $K = 5$ and 0.141 at $K = 50$, respectively. In comparison, the second best method NRLMF obtains inferior precision values of 0.314 at $K = 5$ and 0.126 at $K = 50$, respectively. As shown in Table 3, the average AUPR value achieved by CMLDR is 0.295, which is higher than that obtained by other methods. For the AUC measure, DRRS has achieved the best average AUC value of 0.827, and CMLDR obtains the second best AUC of 0.816. Although CMLDR shows a ~ 1 percent decrease of AUC compared with DRRS, it achieves a ~ 8 percent increase of AUPR.

TABLE 3
The Average AUPR and AUC Values Obtained by Various Methods in Predicting Diseases for New Drugs without known Associations

	CMLDR	DRRS	NRLMF	IMC
AUPR	**0.295**	0.210	0.265	0.242
AUC	0.816	**0.827**	0.771	0.703

The best result is in bold.

TABLE 4
The Number of Retrieved Drug-Disease Associations by the Methods for 18 Drugs in Various Top-Ranked Predictions

	Top5	Top10	Top20	Top30	Top40	Top50
CMLDR	62	112	203	274	317	353
DRRS	31	50	85	120	162	196
NRLMF	55	106	176	227	280	311
IMC	54	97	149	199	237	258

TABLE 5
Retrieved Diseases in Top-20 Predictions for Risperidone (DrugBank: DB00734)

Rank	Disease (OMIM Id)	Retrieved	Rank	Disease (OMIM Id)	Retrieved
1	**600511**	**Y**	11	**608516**	**Y**
2	**181500**	**Y**	12	118700	-
3	**603175**	**Y**	13	**605419**	**Y**
4	**603013**	**Y**	14	600116	-
5	**604906**	**Y**	15	**164230**	**Y**
6	**603206**	**Y**	16	**143465**	**Y**
7	**600850**	**Y**	17	**209850**	**Y**
8	**603176**	**Y**	18	168100	-
9	**181510**	**Y**	19	**167870**	**Y**
10	607834	-	20	**607373**	**Y**

** The retrieved known drug-disease interactions are in bold.*

3.5 Case Studies

For case studies, considering that drugs with more disease associations can provide extensive supporting data, we choose drugs with comprehensive association information to validate the prediction ability of CMLDR in practice.

Specifically, in the benchmark dataset, 637 drugs have more than 0 and less than or equal to 10 associations, 130 drugs have more than 10 and less than or equal to 20 associations, 31 drugs have more than 20 and less than or equal to 30 associations, and 18 drugs have more than 30 associations. Therefore, the 18 drugs with more associations are selected and regarded as test drugs, then we examine if their associations could be identified successfully by applying CMLDR.

In the inference process, all known drug-disease associations except those related with the test drugs in the benchmark dataset are used to train and validate the prediction model. After the model has been trained, all diseases were sorted in descending order of how likely they would associate with the test drugs. The number of retrieved drug-disease associations for 18 drugs in various top-ranked predictions are given in Table 4. It can be observed that 353 ($353/767 \approx 46$ percent) known drug-disease associations have been retrieved in top-50 predictions successfully.

TABLE 6
Retrieved Diseases in Top-20 Predictions for Cyclophosphamide (DrugBank: DB00531)

Rank	Disease (OMIM Id)	Retrieved	Rank	Disease (OMIM Id)	Retrieved
1	**114480**	**Y**	11	**109800**	**Y**
2	**236000**	**Y**	12	**259500**	**Y**
3	**607893**	**Y**	13	114500	-
4	**608935**	**Y**	14	601626	-
5	**605027**	**Y**	15	**260350**	**Y**
6	276300	-	16	608812	-
7	**176807**	**Y**	17	**211980**	**Y**
8	**256700**	**Y**	18	**607248**	**Y**
9	**254500**	**Y**	19	**182280**	**Y**
10	**603956**	**Y**	20	**606856**	**Y**

** The retrieved known drug-disease interactions are in bold.*

TABLE 7
The Number of Identified Drug-Disease Associations of 471 Recently Added Drug-Disease Associations in CTD

	Top5	Top10	Top20	Top30	Top40	Top50
CMLDR	95	159	202	235	261	278
DRRS	24	44	65	85	103	118
NRLMF	78	141	182	212	228	237
IMC	71	123	165	186	200	215

We choose Risperidone (DrugBank: DB00734) and Cyclophosphamide (DrugBank: DB00531) as examples, and list the predicted results of the top-20 candidate diseases for them in Tables 5 and 6. Risperidone is indicated for the treatment of schizophrenia and mood disorders annotated in public database DrugBank [32]. There are thirty-nine diseases associated with Risperidone in the benchmark dataset, out of which sixteen associations are predicted successfully in top-20 predictions. For Cyclophosphamide, it has been used in the treatment of lymphoma and leukemia annotated in DrugBank. There are forty-four validated diseases in the benchmark dataset, and sixteen diseases have been retrieved in top-20 results.

To further validate the prediction power of CMLDR in practical application, we train and validate prediction model based on all known drug-disease associations in the benchmark dataset. The benchmark dataset is derived from CTD database of May 2018 and the published literature [13]. CTD is constantly updated, and the latest data release of April 2019 has been expanded to contain more chemical-disease relationships. By searching the latest version of CTD for novel associations of drugs and diseases, we collect 471 recently added associations with therapeutic evidence which are not contained in the benchmark dataset. Then, we analyze the top-20 results predicted by the methods for all drugs, and summarize the number of identified associations of 471 recently added drug-disease associations in CTD. As shown in Table 7, CMLDR has identified 202 associations, while DRRS, NRLMF, and IMC only identified 65, 182, and 165 associations, respectively. These results demonstrate the practical ability of CMLDR in predicting novel indications for drugs. Moreover, the 202 associations identified by CMLDR are reported in Supplementary Table S2, available online.

4 Conclusion

In this study, we applied the metric learning approach to address the problem of drug repositioning. Specifically, a novel computational method for drug repositioning, named CMLDR, was developed to recommend potential disease indications to given drugs. CMLDR applied collaborative metric learning algorithm to learn the features of drugs and diseases which are represented by two latent vectors in the joint drug-disease space. Compared with matrix factorization, collaborative metric learning algorithm can capture drug-disease relationships in a more intuitive way and can better propagate such information through drug-disease pairs. Therefore, CMLDR can achieve better performance than other drug repositioning methods. Drugs in the benchmark dataset can be grouped into approved drugs with validated diseases, and new drugs having no known/validated associations. For recommending novel diseases to known drugs, CMLDR does not need similarity data and only uses known drug-disease associations. Moreover, CMLDR can effectively perform recommendation for new drugs by utilizing known associations and drug-drug similarity information.

The performances of CMLDR were empirically evaluated on our collected benchmark dataset. We conducted comprehensive experiments involving the recommendation of drug indications to approved drugs and new drugs, and case study analysis. The experiment results have demonstrated that, under different experimental settings, CMLDR consistently outperforms other competing methods in terms of precision and recall rate at different top ranked predictions, AUPR and AUC values. In future studies, we can collect and fuse more useful information to improve the performance of CMLDR and expand its application scope. For example, drug targets and drug-drug interactions can be incorporated in prediction model to represent drug features. The proposed method can also be extended and utilized to address other problems, such as effective drug combination prediction [37], circrna-disease association identification [38] and microbe-disease association identification [39], [40]. Moreover, considering that it's not reasonable that negative samples are selected from unknown associations randomly, we would design efficient method for generating reliable negative samples and incorporate it into our model.

Acknowledgments

This study is supported in part by the Natural Science Foundation of China (No. 61828205, No. 61772557, No. 61802113, No. 61420106009 and No. 61772552), Project (No. B18059) and Hunan Provincial Science and Technology Program (No. 2018WK4001).

References

[1] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, et al., "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nature Rev. Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010.
[2] N. Prakash, and P. Devangi, "Drug discovery," *J. Antivirals Antiretrovirals*, vol. 2, no. 4, pp. 63–68, 2010.
[3] H. Xue, J. Li, H. Xie, et al., "Review of drug repositioning approaches and resources," *Int. J. Biol. Sci.*, vol. 14, no. 10, pp. 1232–1244, 2018.
[4] A. Mullard, "New drugs cost US $2.6 billion to develop," *Nature Rev. Drug Discovery*, vol. 13, p. 877, 2014.
[5] T. J. Hwang, D. Carpenter, J. C. Lauffenburger, et al., "Failure of investigational drugs in late-stage clinical development and publication of trial results," *JAMA Internal Med.*, vol. 176, no. 12, pp. 1826–1833, 2016.
[6] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nature Rev. Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.
[7] G. Jin and S. T. C. Wong, "Toward better drug repositioning: Prioritizing and integrating existing methods into efficient pipelines," *Drug Discovery Today*, vol. 19, no. 5, pp. 637–644, 2014.
[8] T. Cheng, M. Hao, T. Takeda, et al., "Large-scale prediction of drug-target interaction: A data-centric review," *The AAPS J.*, vol. 19, no. 5, pp. 1264–1275, 2017.
[9] E. March-Vila, L. Pinzi, N. Sturm, et al., "On the integration of in silico drug design methods for drug repurposing," *Frontiers Pharmacology*, vol. 8, 2017, Art. no. 298.
[10] J. Li, S. Zheng, B. Chen, et al., "A survey of current trends in computational drug repositioning," *Briefings Bioinf.*, vol. 17, no. 1, pp. 2–12, 2016.
[11] M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, et al., "A review of network-based approaches to drug repositioning," *Briefings Bioinf.*, vol. 19, no. 5, pp. 878–892, 2017.
[12] A. Ezzat, M. Wu, X. L. Li, et al., "Computational prediction of drug-target interactions using chemogenomic approaches: An empirical survey," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bby002, 2018.
[13] A. Gottlieb, G. Y. Stein, E. Ruppin, et al., "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, 2011, Art. no. 496.
[14] Y. Wang, S. Chen, N. Deng, et al., "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data," *PloS One*, vol. 8, no. 11, 2013, Art. no. e78518.
[15] W. Wang, S. Yang, X. Zhang, et al., "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinf.*, vol. 30, no. 20, pp. 2923–2930, 2014.
[16] H. Liu, Y. Song, J. Guan, et al., "Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks," *BMC Bioinf.*, vol. 17, no. 17, 2016, Art. no. 539.
[17] M. L. Shahreza, N. Ghadiri, S. R. Mousavi, et al., "Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning," *J. Biomed. Informat.*, vol. 68, pp. 167–183, 2017.
[18] H. Luo, J. Wang, M. Li, et al., "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm," *Bioinf.*, vol. 32, no. 17, pp. 2664–2671, 2016.
[19] M. Gönen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinf.*, vol. 28, no. 18, pp. 2304–2310, 2012.
[20] J. Yang, Z. Li, X. Fan, et al., "Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization," *J. Chemical Inf. Model.*, vol. 54, no. 9, pp. 2562–2569, 2014.
[21] Y. Liu, M. Wu, C. Miao, et al., "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS Comput. Biol.*, vol. 12, no. 2, 2016, Art. no. e1004760.
[22] A. Ezzat, P. Zhao, M. Wu, et al.,"Drug-target interaction prediction with graph regularized matrix factorization," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 646–656, 2017.
[23] Y. Luo, X. Zhao, J. Zhou, et al., "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Commun.*, vol. 8, no. 1, 2017, Art. no. 573.
[24] H. Luo, M. Li, S. Wang, et al., "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinf.*, vol. 34, no. 11, pp. 1904–1912, 2018.
[25] W. Zhang, X. Yue, W. Lin, et al., "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinf.*, vol. 19, no. 1, 2018, Art. no. 233.
[26] N. Natarajan and I. S. Dhillon, "Inductive matrix completion for predicting gene-disease associations," *Bioinf.*, vol. 30, no. 12, pp. i60–i68, 2014.
[27] C. Yan, J. Wang, P. Ni, et al., "DNRLMF-MDA: Predicting microRNA-disease associations based on similarities of microRNAs and diseases," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 233–243, Jan./Feb. 2019.
[28] C. Lu, M. Yang, F. Luo, et al., "Prediction of lncRNA-disease associations based on inductive matrix completion," *Bioinf.*, vol. 34, no. 19, pp. 3357–3364, 2018.

[29] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *arXiv preprint arXiv:1306.0626*, 2013.
[30] C. K. Hsieh, L. Yang, Y. Cui, et al., "Collaborative metric learning," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 193–201.
[31] A. P. Davis, C. J. Grondin, R. J. Johnson, et al., "The comparative toxicogenomics database: Update 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D972–D978, 2016.
[32] D. S. Wishart, Y. D. Feunang, A. C. Guo, et al., "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2017.
[33] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug-drug and drug-food interactions," *Proc. Nat. Acad. Sci. United States America*, vol. 115, no. 18, pp. E4304–E4311, 2018.
[34] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: Learning to rank with joint word-image embeddings," *Mach. Learn.*, vol. 81, no. 1, pp. 21–35, 2010.
[35] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.
[36] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS One*, vol. 10, no. 3, 2015, Art. no. e0118432.
[37] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC Bioinf.*, vol. 18, no. 1, p. 18, 2017.
[38] C. Yan, J. Wang, and F.-X. Wu, "Dwnn-rls: regularized least squares method for predicting circrna-disease associations," *BMC Bioinf.*, vol. 19, no. 19, p. 520, 2018.
[39] C. Yan, D. Guihua, F. X. Wu, Y. Pan, and J. Wang, "Brwmda:predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2907626, 2019.
[40] C. Yan, D. Guihua, F. X. Wu, Y. Pan, and J. Wang, "Mchmda: Predicting microbe-disease associations based on similarities and low-rank matrix completion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2019.2926716, 2019.

Huimin Luo is working toward the PhD degree in the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her current research interests include bioinformatics and systems biology.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the dean and a professor in School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 150 papers in various International journals and refereed conferences. He is a senior member of the IEEE.

Cheng Yan is working toward the PhD degree in the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include bioinformatics and machine learning.

Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently the vice dean and a professor with the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her research interests include computational biology, systems biology and bioinformatics. She has published more than 80 technical papers in refereed journals such as the *Bioinformatics*, the *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, the *Proteomics*, and conference proceedings such as BIBM, GIW and ISBRA.

Fang-Xiang Wu (M'06-SM'11) received the BSc and the MSc degrees in applied mathematics, both from the Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from North-western Polytechnical University, Xi'an, China, in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a postdoctoral fellow with the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a professor of the Division of Biomedical Engineering and the Department of Mechanical Engineering, U of S. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, applications of control theory to biological systems. He has published more than 260 technical papers in refereed journals and conference proceedings. He is serving as the editorial board member of five international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals. He is a senior member of the IEEE.

Yi Pan received the BEng and MEng degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and the PhD degree in computer science from the University of Pittsburgh, in 1991. He is a Regents' professor of computer science and an Interim associate dean and chair of Biology, Georgia State University. He joined Georgia State University, in 2000 and was promoted to full professor, in 2004, named a Distinguished University professor, in 2013 and designated a Regents' professor (the highest recognition given to a faculty member by the University System of Georgia), in 2015. He served as the chair of Computer Science Department from 2005-2013. His profile has been featured as a distinguished alumnus in both Tsinghua Alumni Newsletter and University of Pittsburgh CS Alumni Newsletter. His research interests include parallel and cloud computing, wireless networks, and bioinformatics. He has published more than 300 papers including over 180 SCI journal papers and 60 IEEE/ACM Transactions papers. In addition, he has edited/authored 40 books. His work has been cited more than 10000 times. He has served as an editor-in-chief or editorial board member for 15 journals including 7 IEEE Transactions. He is the recipient of many awards including IEEE Transactions Best Paper Award, 4 other international conference or journal Best Paper Awards, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized many international conferences and delivered keynote speeches at over 50 international conferences around the world.