GC^{2}NMF: A Novel Matrix Factorization Framework for Gene-phenotype Association Prediction

Yaogong Zhang[‡], Jiahui Liu[‡], Yuxiang Hong[‡], Xin Fan[‡], Xiaohu Liu[‡], Yuan Wang[§] Yalou Huang[‡] and Maoqiang Xie[‡]
[‡]College of Software, NanKai University, TianJin, China, 300071
Email: {ygzhang@mail., jiahui@mail., hongyuxiang@mail., nkufanxin@mail., liuxiaohu@mail., huangyl@, xiemq@}nankai.edu.cn
[§]Computer Science and Information Engineering, Tianjin University of Science and Technology, 300222, China
Email: wangyuan23@tust.edu.cn

Abstract—Gene-phenotype association prediction can be applied to reveal the inherited basis of human diseases and facilitate drug development. Gene-phenotype associations are related to complex biological processes and influenced by various factors, such as relationship between phenotypes and that among genes. While due to sparseness of curated gene-phenotype associations and lack of integrated analysis of the joint effect of multiple factors, existing applications are limited to prediction accuracy and potential gene-phenotype association detection. In this paper, we propose a novel method by exploiting weighted graph constraint learned from hierarchical structures of phenotype data and group prior information among genes by inheriting advantages of Non-negative Matrix Factorization (NMF), called Weighted **G**raph **C**onstraint and **G**roup **C**entric **N**on-negative **M**atrix **F**actorization (GC^{2}NMF). Specifically, firstly we introduce the depth of parent-child relationships between two adjacent phenotypes in hierarchical phenotypic data as weighted graph constraint for a better phenotype understanding. Secondly, we utilize intra-group correlation among genes in a gene group as group constraint for gene understanding. Such information provides us with the intuition that genes in a group probably result in similar phenotypes. The model not only allows us to achieve a high-grade prediction performance, but also helps us to learn interpretable representation of genes and phenotypes simultaneously to facilitate future biological analysis. Experimental results on biological gene-phenotype association datasets of mouse and human demonstrate that GC^{2}NMF can obtain superior prediction accuracy and good understandability for biological explanation over other state-of-the-arts methods.

1 Introduction

Currently, predicting new gene-phenotype associations has long been an important goal in computational biology. It is of great significance to explore the interactions between genes and phenotypes for drug development and disease treatment. The increasing biological data, such as Mouse Genome Informatics (MGI) [9], Human Phenotype Ontology (HPO) [14], Kyoto Encyclopedia of Genes and Genomes (KEGG) [12], provides us with a great opportunity to discover underlying patterns of diseases and their biological mechanisms.

Recently, different network-based methods for gene-phenotype association prediction have been proposed [23], [13], [15], [22], [10], [24]. Among all these methods, BiRW, proposed by Xie[24], achieved the best prediction results. It performs random walk on PPI network and phenotype similarity network alternatively to enrich genome-phenome association matrix, then prioritizes disease genes based on the enriched association matrix. However, there are only few known associations available for training. For example, in HPO, more than half of the phenotypes are annotated with no or only one gene association. Lacking of integrated analysis of the joint effect of multiple factors and severe sparseness of the curated gene-phenotype association matrix limit the performances of these methods on gene (phenotype) representations and even degrade their prediction accuracy.

To tackle the problem, we propose Weighted **G**raph **C**onstraint and **G**roup **C**entric **N**on-negative **M**atrix **F**actorization (GC^{2}NMF) to infer gene-phenotype relationships. We show the illustration of data relations of GC^{2}NMF in Fig. 1. We introduce hierarchical relation information in phenotypic data (i.e. Human Phenotype Ontology shown in Fig. 1(a)) and gene group information in gene data (i.e. Gene Group shown in Fig. 1(b)) to alleviate the problem.

Firstly, biologists have confirmed that two phenotypes of the parent-child relationship in adjacent levels from the phenotype ontology are similar [14]. For a better understanding of phenotypes, we should keep them having a short Euclidean distance in latent space if a parent-child relationship exists between two phenotypes. However, such plain graph constraint [11] neglects an important characteristic, the sig-

[‡] *Maoqiang Xie is corresponding author.*

nificance of adjacent phenotypes' relationship varies with their depths. The phenotype ontology at a higher level is a more general taxonomy of its child phenotype ontology, and the phenotype ontology at deeper levels is more specific. Thus two adjacent general phenotype ontologies are loosely associated at higher levels, while two adjacent specific phenotype ontologies are tightly associated at deeper levels. Based on this fact, we adjust Euclidean distance loss of parent-child phenotype pairs according to locations of edges and propose the weighted graph constraint to capture its hierarchical structure for a better phenotype representation in latent space.

Secondly, genes in a gene group are functionally related, because they participate in some specific biological process cooperatively. This tells the fact that genes in a group should keep close relationships, so we utilize gene group information to constrain any two genes close to each other in the same group in latent space. Because of using pairwise constraints in group relationship (commonly achieved by graph Laplacian) is quite time-consuming, the time complexity is quadratic to the number of genes in the group. We propose the group centric constraint to use the gene groups in a more efficient way. Specifically, we introduce a group centric point to every single gene group. This point is the geometrical centric point of all genes in a group in latent representation space. Then we keep genes within a group near to their corresponding centric point. We constrain the intra-group distance between a gene and its corresponding group centric point instead of comparing every pair of genes in a group, so the model can have a linear time complexity of the maximum number of genes in a group and have a good scalability when dealing with large groups.

The two aspects mentioned above extend pairwise relationships between one gene and one phenotype to multiple genes and multiple phenotypes. Adjacent phenotypes could be affected by related genes and genes in a group probably result in similar phenotypes. We define such knowledge as "group to group" relations. This knowledge benefits us greatly by helping to overcome the sparseness problem of the curated gene-phenotype association matrix during modeling.

The major contributions of this paper are summarized as follows:

- We alleviate the sparseness problem of gene-phenotype association matrixes in NMF modeling by extending pairwise relations to "group to group" relations.
- We introduce weighted graph constraint into NMF framework to predict gene-phenotype relations. Meanwhile, it benefits the capturing of the hierarchical structure of phenotypic data and the learning of a better phenotype representation in latent space.
- A novel group centric constraint is proposed to strengthen the representation of genes in latent space, which enables a linear time complexity as the number of group members increases.
- Empirical studies on mouse and human data experiments demonstrate the effectiveness of GC^{2}NMF. The proposed model GC^{2}NMF outperforms its state-of-the-art counterparts and achieves the best results in both quality and quantity evaluation.

2 Related Work

Making accurate identifications of gene-phenotype associations is the first step toward a systematic understanding of the molecular mechanisms of a complex disease. Also, it is essential to know disease-related genes for diagnosis and drug development [6]. These years, many works have been conducted on gene-phenotype association prediction. Kohler prioritizes candidate genes by the use of random walk from known genes for a given disease [13]. Li extends random walk with restart algorithm to the heterogeneous network, it makes better use of the phenotypic data by using the target phenotypes and corresponding genes as seed nodes simultaneously [15]. Vanunu uses the known disease relationships to decide an initial set of genes that are associated with a query disease phenotype, then it performs label propagation on the PPI network to prioritize disease genes [22]. Xie performs random walk on PPI network and phenotype similarity network alternatively to enrich genome-phenome association matrix, then prioritizes disease genes based on the enriched association matrix [24].

The NMF framework achieves a great success on relationship prediction and element representation in recent years. Due to flexibility of NMF, many complicated variants and extensions have been proposed to adapt to the complex relationships. Zhang [25] introduces Expectation-Maximization (EM) procedure based NMF and weighted NMF to learn a low dimensional linear model to describe the user rating matrix. Benzi [2] formulates a novel song recommender system as a matrix completion problem that benefits from collaborative filtering through Non-negative Matrix Factorization and content-based filtering via total variation on graphs. These methods demonstrate the effectiveness of NMF to relationship prediction and completion problems.

Besides these accomplishment of NMF, the combination of graph constraint and NMF framework also achieves greatness. Cai [4] proposes Graph Regularized Non-negative Matrix Factorization (GNMF) for data representation, which constructs an affinity graph to encode the geometrical information via matrix factorization to uncover the hidden semantics and simultaneously respects the intrinsic geometric structure. Rajabi [17] examines the applicability of graph regularized nonnegative matrix factorization for estimating end members and abundance fractions in a mixed pixel. Although these works use the graph constraint on data representation, it becomes different when we face the situation that a hierarchical structure exists in the data. In order to have a further recognition on data representation in latent space and expect to get a better result, we explore the characteristic of the hierarchical structure and propose the weighted graph constraint.

Group information in NMF has been proven to be an effective way to solve problems. In some cases, elements in input can be divided into groups. The relationship between elements in groups can be used to improve the model. AliMashhoor in [1] finds that items could be treated as different groups and then constructed the (user, item group) by using auxiliary information such as tags and temporal

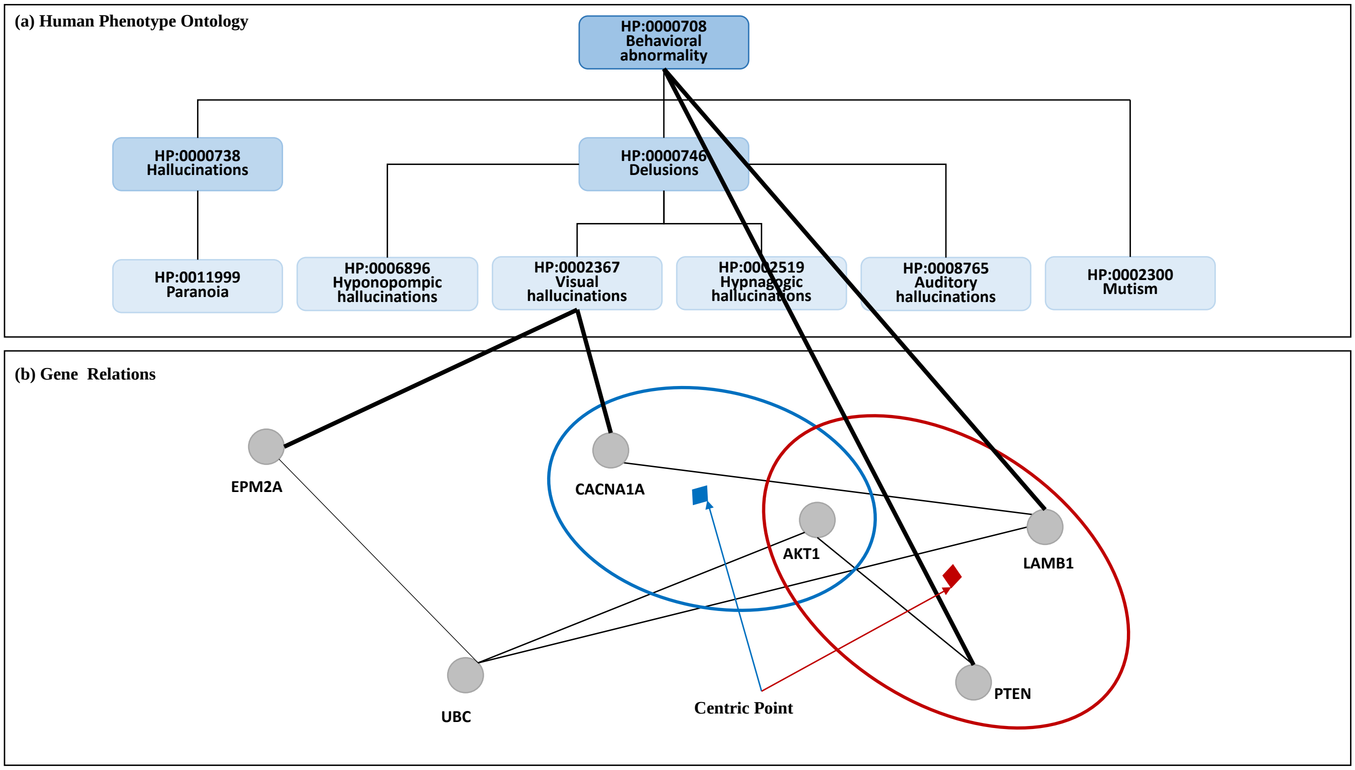
(a) Human Phenotype Ontology
HP:0000708 Behavioral abnormality
HP:0000738 Hallucinations
HP:0000746 Delusions
HP:0011999 Paranoia
HP:0006896 Hyponopompic hallucinations
HP:0002367 Visual hallucinations
HP:0002519 Hypnagogic hallucinations
HP:0008765 Auditory hallucinations
HP:0002300 Mutism
(b) Gene Relations
EPM2A
CACNA1A
AKT1
LAMB1
UBC
PTEN
Centric Point

Fig. 1. The illustration of relations between biological elements. (a) Human Phenotype Ontology, the hierarchical phenotype ontology on human. Each rectangle stands for a human phenotype ontology, the fine line linking two adjacent phenotype ontologies (such as the line between HP:0000708 and HP:0000738) means a parent-child relationship. (b) Gene Relations, including Gene Group relations shown in circles and Gene Network relations shown with fine lines, and the diamond points stand for the geometrical group centric point in corresponding circles in latent representation space. Thick lines mean curated associations between genes and phenotypes.

information. Ma [16] mainly focuses on (user group, item) correlations from social networks or rating similarities. Both of them show that incorporating group structure into matrix factorization can result in better recommendation performance. In our paper, we introduce the group centric constraint to use the gene groups to overcome the sparseness of gene-phenotype association data and learn a more precise data representation in latent space.

3 Methods

In this section, we firstly present notation definitions and basic concepts used in this paper. Secondly, we introduce the weighted graph constraint of hierarchical phenotype ontology. Thirdly, we define group centric constraint mathematically, discuss the learning and inference methods, and analyze the computational complexity of GC^{2}NMF.

3.1 Notations

The notations and definitions used in our model are specified in Table 1. Let $\boldsymbol{R}_{n\times m}$ be a binary matrix for storing gene-phenotype associations of n genes and m phenotypes, where $\boldsymbol{R}_{ij} = 1$ for a known association between gene i and phenotype j and $\boldsymbol{R}_{ij} = 0$ otherwise. $\mathcal{G}$ is the set of all gene groups. Let G be a gene group in $\mathcal{G}$. In biology, phenotype ontologies are organized in a tree structure. $\boldsymbol{S}_1$ is the relationship between phenotype ontologies, where $(\boldsymbol{S}_1)_{ij} = 1$ if there is a parent-child relationship between phenotype i and phenotype j, and $(\boldsymbol{S}_1)_{ij} = 0$ otherwise. The Protein Protein Interaction (PPI) network [8] can be considered as an interaction network of genes. Let $\boldsymbol{S}_2$ be the relationships between genes in PPI network, where $(\boldsymbol{S}_2)_{ij} = 1$ if an interaction exists between gene i and gene j in PPI network, and $(\boldsymbol{S}_2)_{ij} = 0$ otherwise. $\boldsymbol{U}_{n\times K}$ and $\boldsymbol{V}_{K\times m}$ are the output of our model.

TABLE 1
Notations

PART	DESCRIPTION
n	the number of genes
m	the number of phenotypes
K	the dimension of latent space
$\boldsymbol{R}_{n\times m}$	gene-phenotype association matrix
$\boldsymbol{U}_{n\times K}$	gene representation in latent space
$\boldsymbol{V}_{K\times m}$	phenotype representation in latent space
$\boldsymbol{X}_{i\cdot}$	the ith row of matrix $\boldsymbol{X}$
$\boldsymbol{X}_{\cdot j}$	the jth column of matrix $\boldsymbol{X}$
$\boldsymbol{S}_i$	an adjacent relation matrix, $i = 1$ for phenotypes, $i = 2$ for genes
$\boldsymbol{Y}$	a binary indicator matrix
G	a gene group
$\mathcal{G}$	a set of gene groups
$\bar{\boldsymbol{U}}_G$	centric point vector of gene group $\boldsymbol{G}$

3.2 Brief Review of NMF

The goal of factorizing gene-phenotype association matrix $\boldsymbol{R}$ is to derive a gene representation $\boldsymbol{U}$ and phenotype representation $\boldsymbol{V}$ in latent space. Then for any gene-phenotype pairs, we could use the dot product value of two corresponding vectors to predict the association between them. The objective function of NMF is defined as:

$$min_{\boldsymbol{U},\boldsymbol{V}}||\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V}||_F^2. \tag{1}$$

Lee&Seung [7] presented iterative update algorithms for parameter inference by following update rules:

$$\boldsymbol{U}_{ik} \leftarrow \boldsymbol{U}_{ik}\frac{(\boldsymbol{R}\boldsymbol{V})_{ik}}{(\boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{V})_{ik}},\ \ \boldsymbol{V}_{kj} \leftarrow \boldsymbol{V}_{kj}\frac{(\boldsymbol{R}^T\boldsymbol{U})_{kj}}{(\boldsymbol{V}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U})_{kj}}. \tag{2}$$

The rules in Eq. (2) have been proven [7] to find a local minima of the objective function in Eq. (1).

3.3 Weighted Graph Constraint

Phenotype ontology serves as a standardized vocabulary of phenotypic abnormalities that have been seen in diseases on mouse [21] and human [14], it is organized into a hierarchical tree structure. The higher the phenotypes locate, the more general they are; the deeper the phenotypes locate, the more specific they are. The edge between two phenotypes from two adjacent levels indicates a parent-child relationship between them. Because of such relationship, we keep two adjacent phenotypes near in latent spaces with constraint on their representation vectors:

$$\begin{aligned}\mathcal{R}_1 &= \frac{1}{2}\sum_{ij}(\boldsymbol{S}_1)_{ij}||\boldsymbol{V}_{\cdot i}-\boldsymbol{V}_{\cdot j}||^2\\ &= \sum_i \boldsymbol{V}_{\cdot i}^T\boldsymbol{V}_{\cdot i}(\boldsymbol{D}_1)_{ii} - \sum_{ij}\boldsymbol{V}_{\cdot i}^T\boldsymbol{V}_{\cdot j}(\boldsymbol{S}_1)_{ij}\\ &= tr(\boldsymbol{V}\boldsymbol{D}_1\boldsymbol{V}^T) - tr(\boldsymbol{V}\boldsymbol{S}_1\boldsymbol{V}^T)\\ &= tr\big(\boldsymbol{V}\boldsymbol{L}_1\boldsymbol{V}^T\big),\end{aligned} \tag{3}$$

where $tr(\cdot)$ denotes the trace of a matrix, $\boldsymbol{S}_1$ is the symmetrical relationship matrix of phenotype ontology. We define $(\boldsymbol{D}_1)_{ii} = \sum_j(\boldsymbol{S}_1)_{ij}$, which is the diagonal matrix with the row-sum of $\boldsymbol{S}_1$ on the diagonal entries. We define $\boldsymbol{L}_1 = \boldsymbol{D}_1 - \boldsymbol{S}_1$, which is called graph Laplacian [11].

By considering the hierarchical tree structure of phenotype ontology (see Fig. 1(a)), general phenotype ontologies are loosely associated at higher levels, while specific phenotype ontologies are tightly associated at deeper levels. To better capture this characteristic of phenotype data, we introduce variable $depth_{(i,j)}$ and an auxiliary function $C(depth_{(i,j)})$. $depth_{(i,j)}$ equals to the depth of the parent node i, we define $depth_{(i,j)} = 1$ if parent node i is the root node. $C(depth_{(i,j)})$ is the weight of the adjacent relationship when the depth of the edge varies.

We define $C(depth_{(i,j)})$ as follows:

$$C(depth_{(i,j)}) = 1 + log(depth_{(i,j)}) \tag{4}$$

$$depth_{(i,j)} = \begin{cases} depth_{(p)}+1 & : \ others \\ 1 & : \ node\ i\ is\ the\ root\end{cases}$$

where $depth_{(p)}$ is the depth of the parent node of node i, then the new relation matrix of phenotype ontologies can be expressed as:

$$(\boldsymbol{S}_1^{'})_{ij} = (\boldsymbol{S}_1)_{ij} * C(depth_{(i,j)}).$$

By replacing $(\boldsymbol{S}_1)_{ij}$ with $(\boldsymbol{S}_1^{'})_{ij}$ into Eq. (3):

$$\begin{aligned}\mathcal{R}_1^{'} &= \sum_{ij}(\boldsymbol{S}_1^{'})_{ij}||\boldsymbol{V}_{\cdot i}-\boldsymbol{V}_{\cdot j}||^2\\ &= tr\big(\boldsymbol{V}\boldsymbol{L}_1^{'}\boldsymbol{V}^T\big),\end{aligned} \tag{5}$$

where $\boldsymbol{D}_1^{'} = \sum_i(\boldsymbol{S}_1^{'})_{ij}$, $\boldsymbol{L}_1^{'} = \boldsymbol{D}_1^{'} - \boldsymbol{S}_1^{'}$.

Actually the graph Laplacian is a special case of our proposed weighted graph Laplacian when we set $C(depth_{(i,j)}) = 1$ for any two adjacent node i and node j. As there is no hierarchical structure in PPI network for genes, we just apply the general graph Laplacian in PPI network:

$$\begin{aligned}\mathcal{R}_2 &= \frac{1}{2}\sum_{ij}(\boldsymbol{S}_2)_{ij}||\boldsymbol{U}_{i\cdot}-\boldsymbol{U}_{j\cdot}||^2\\ &= tr(\boldsymbol{U}^T\boldsymbol{D}_2\boldsymbol{U}) - tr(\boldsymbol{U}^T\boldsymbol{S}_2\boldsymbol{U})\\ &= tr\big(\boldsymbol{U}^T\boldsymbol{L}_2\boldsymbol{U}\big),\end{aligned} \tag{6}$$

where $\boldsymbol{S}_2$ is the symmetrical adjacent matrix of genes in PPI networks. $(\boldsymbol{D}_2)_{ii} = \sum_j(\boldsymbol{S}_2)_{ij}$, which is the diagonal matrix with the row-sum of $\boldsymbol{S}_2$ on the diagonal entries, $\boldsymbol{L}_2 = \boldsymbol{D}_2 - \boldsymbol{S}_2$.

By minimizing $\mathcal{R}_2$ $(\mathcal{R}_1^{'})$, we expect that if two genes (phenotypes) are close on the graph $(\boldsymbol{S}_2)_{ij} \neq 0$ $\Big((\boldsymbol{S}_1^{'})_{ij} \neq 0\Big)$ after the factorization, Euclidean distance between $\boldsymbol{U}_{i\cdot}$ and $\boldsymbol{U}_{j\cdot}$ $(\boldsymbol{V}_{\cdot i}$ and $\boldsymbol{V}_{\cdot j})$ are also close to each other. By incorporating Eq. (6) and Eq. (5) into Eq. (1) and introducing Frobenius norm $||\cdot||_F^2$ of $\boldsymbol{U}$ and $\boldsymbol{V}$ to avoid overfitting, the objective function of NMF with weighted graph Laplacian constraints on genes and phenotype ontology is formulated as follows:

$$\begin{aligned}\mathcal{O}_1 =& \frac{1}{2}||\boldsymbol{Y}\odot(\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V})||_F^2\\ &+ \frac{\lambda_1}{2}(tr(\boldsymbol{U}^T\boldsymbol{L}_2\boldsymbol{U}) + tr\big(\boldsymbol{V}\boldsymbol{L}_1^{'}\boldsymbol{V}^T\big))\\ &+ \frac{\lambda_2}{2}(||\boldsymbol{U}||_F^2 + ||\boldsymbol{V}||_F^2)\\ &\text{s.t.}\quad \boldsymbol{U}\geq 0, \boldsymbol{V}\geq 0,\end{aligned} \tag{7}$$

where $\odot$ is the Hadamard product known as the entrywise product between matrices. $\boldsymbol{Y}$ is a binary indicator matrix. $\boldsymbol{Y}_{ij} = 1$ if there is an association between gene i and phenotype j, otherwise $\boldsymbol{Y}_{ij} = 0$.

Let ϕ_{ik} and ψ_{kj} be the Lagrange multiplier for constraint $\boldsymbol{U}_{ik} \geq 0$ and $\boldsymbol{V}_{kj} \geq 0$, respectively, where $\Phi = [\phi_{ik}]$ and $\Psi = [\psi_{kj}]$. The Lagrange function $\mathcal{L}_1$ of Eq. (7) is:

$$\begin{aligned}\mathcal{L}_1 =& \frac{1}{2}||\boldsymbol{Y}\odot(\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V})||_F^2\\ &+ \frac{\lambda_1}{2}(tr(\boldsymbol{U}^T\boldsymbol{L}_2\boldsymbol{U}) + tr\big(\boldsymbol{V}\boldsymbol{L}_1^{'}\boldsymbol{V}^T\big))\\ &+ \frac{\lambda_2}{2}(||\boldsymbol{U}||_F^2 + ||\boldsymbol{V}||_F^2) + tr(\Phi\boldsymbol{U}^T) + tr(\Psi\boldsymbol{V}^T).\end{aligned} \tag{8}$$

The partial derivatives of $\mathcal{L}_1$ with respect to $\boldsymbol{U}_{ik}$ and $\boldsymbol{V}_{kj}$ are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial(\boldsymbol{U}_{ik})} =& -\Big(\boldsymbol{Y}_{i\cdot} \odot (\boldsymbol{R}_{i\cdot} - \boldsymbol{U}_{i\cdot}\boldsymbol{V})\Big)(\boldsymbol{V}^T)_{\cdot k} \\ & + \lambda_1((\boldsymbol{D_2} - \boldsymbol{S_2})\boldsymbol{U})_{ik} + \lambda_2(\boldsymbol{U})_{ik} + \Phi_{ik}, \\ \frac{\partial \mathcal{L}}{\partial(\boldsymbol{V}_{kj})} =& -(\boldsymbol{U}^T)_{k\cdot}\Big(\boldsymbol{Y}_{\cdot j} \odot (\boldsymbol{R}_{\cdot j} - \boldsymbol{U}\boldsymbol{V}_{\cdot j})\Big) \\ & + \lambda_1(\boldsymbol{V}(\boldsymbol{D_1^{'}} - \boldsymbol{S_1^{'}}))_{kj} + \lambda_2(\boldsymbol{V})_{kj} + \Psi_{kj}. \end{aligned} \tag{9}$$

Using the KKT conditions $\boldsymbol{\phi}_{ik}\boldsymbol{U}_{ik} = 0$ and $\boldsymbol{\psi}_{kj}\boldsymbol{V}_{kj} = 0$, we can get the following equations for $\boldsymbol{U}_{ik}$ and $\boldsymbol{V}_{kj}$:

$$\begin{aligned} &-\Big(\boldsymbol{Y}_{i\cdot}\odot(\boldsymbol{R}_{i\cdot} - \boldsymbol{U}_{i\cdot}\boldsymbol{V})\Big)(\boldsymbol{V}^T)_{\cdot k}\boldsymbol{U}_{ik} + \\ &\quad \lambda_1((\boldsymbol{D_2} - \boldsymbol{S_2})\boldsymbol{U})_{ik}\boldsymbol{U}_{ik} + \lambda_2(\boldsymbol{U})_{ik}\boldsymbol{U}_{ik} = 0, \end{aligned} \tag{10}$$

$$\begin{aligned} &-(\boldsymbol{U}^T)_{k\cdot}\Big(\boldsymbol{Y}_{\cdot j} \odot (\boldsymbol{R}_{\cdot j} - \boldsymbol{U}\boldsymbol{V}_{\cdot j})\Big)\boldsymbol{V}_{kj} \\ &\quad + \lambda_1(\boldsymbol{V}(\boldsymbol{D_1^{'}} - \boldsymbol{S_1^{'}}))_{kj}\boldsymbol{V}_{kj} + \lambda_2(\boldsymbol{V})_{kj}\boldsymbol{V}_{kj} = 0. \end{aligned} \tag{11}$$

The updating rules for $\boldsymbol{U}_{ik}$ and $\boldsymbol{V}_{kj}$ can be derived from Eq. (10) and Eq. (11):

$$\boldsymbol{U}_{ik} \leftarrow \boldsymbol{U}_{ik}\frac{(\boldsymbol{Y}_{i\cdot} \odot \boldsymbol{R}_{i\cdot})(\boldsymbol{V}^T)_{\cdot k} + \lambda_1(\boldsymbol{S_2}\boldsymbol{U})_{ik}}{\big(\boldsymbol{Y}_{i\cdot} \odot (\boldsymbol{U}_{i\cdot}\boldsymbol{V})\big)(\boldsymbol{V}^T)_{\cdot k} + \lambda_1(\boldsymbol{D_2}\boldsymbol{U})_{ik} + \lambda_2\boldsymbol{U}_{ik}} \tag{12}$$

$$\boldsymbol{V}_{kj} \leftarrow \boldsymbol{V}_{kj}\frac{(\boldsymbol{U}^T)_{k\cdot}(\boldsymbol{Y}_{\cdot j} \odot \boldsymbol{R}_{\cdot j}) + \lambda_1(\boldsymbol{V}\boldsymbol{S_1^{'}})_{kj}}{(\boldsymbol{U}^T)_{k\cdot}\big(\boldsymbol{Y}_{\cdot j} \odot (\boldsymbol{U}\boldsymbol{V}_{\cdot j})\big) + \lambda_1(\boldsymbol{V}\boldsymbol{D_1^{'}})_{kj} + \lambda_2\boldsymbol{V}_{kj}} \tag{13}$$

3.4 Group Centric Constraint

Gene groups, such as KEGG pathways give us a basis to constrain genes within groups. A gene group reveals a biological process at gene levels. All genes in a group participate in the same biological process. Thus it could been hypothesized that all genes within a group probably lead to similar or even the same phenotypes.

In this article, instead of making any pair elements close in the group, the time complexity would be quadratic to the number of elements. We introduce the group centric constraint to make any two genes in the same group close to each other. We use a geometrical centric point of all genes in the latent space as the group centric point, and all group members in this group should be close to the group centric point. In each new iteration, we use $\boldsymbol{U}$ and $\boldsymbol{V}$ in the last iteration to calculate each groups' centric points, which are used as fixed variables in current iteration. The group centric constraint can be formulated as follows.

$$\sum_{G\in\mathcal{G}}\sum_{j\in G}||\boldsymbol{U}_{G_j\cdot} - \bar{\boldsymbol{U}}_G||^2, \tag{14}$$

where $\boldsymbol{U}_{G_j\cdot}$ is the jth element in gene group G, and $\bar{\boldsymbol{U}}_G = \sum\limits_{i\in G}\boldsymbol{U}_{i\cdot}/\sum\limits_{i\in G}1$, which is the geometric center of group G. $||\boldsymbol{U}_{G_j\cdot} - \bar{\boldsymbol{U}}_G||^2$ means the Euclidean distance between a group member and the corresponding group centric point. By incorporating Eq. (14) into Eq. (7), we get the loss objective function of Weighted Graph Constraint and Group Centric Non-negative Matrix Factorization (GC^{2}NMF).

$$\begin{aligned} \mathcal{O}_2 =& \frac{1}{2}||\boldsymbol{Y} \odot (\boldsymbol{R} - \boldsymbol{U}\boldsymbol{V})||^2 + \frac{\lambda_0}{2}\sum_{G\in\mathcal{G}}\sum_{j\in G}||\boldsymbol{U}_{G_j\cdot} - \bar{\boldsymbol{U}}_G||^2 \\ & + \frac{\lambda_1}{2}(tr(\boldsymbol{U}^T(\boldsymbol{D_2} - \boldsymbol{S_2})\boldsymbol{U}) + tr(\boldsymbol{V}(\boldsymbol{D_1^{'}} - \boldsymbol{S_1^{'}})\boldsymbol{V}^T)) \\ & + \frac{\lambda_2}{2}(||\boldsymbol{U}||^2 + ||\boldsymbol{V}||^2) \\ & \text{s.t.} \quad \boldsymbol{U} \geq 0, \ \boldsymbol{V} \geq 0. \end{aligned} \tag{15}$$

By introducing the Lagrange multiplier for ϕ_{ik} and ψ_{kj} constraints $\boldsymbol{U}_{ik} \geq 0$ and $\boldsymbol{V}_{kj} \geq 0$, $\Phi = [\phi_{ik}]$ and $\Psi = [\psi_{kj}]$, the Lagrange function $\mathcal{L}_2$ of Eq. (15) is

$$\begin{aligned} \mathcal{L}_2 =& \frac{1}{2}||\boldsymbol{Y} \odot (\boldsymbol{R} - \boldsymbol{U}\boldsymbol{V})||_F^2 + \frac{\lambda_0}{2}\sum_{G\in\mathcal{G}}\sum_{j\in G}||\boldsymbol{U}_{G_j\cdot} - \bar{\boldsymbol{U}}_G||^2 \\ & + \frac{\lambda_1}{2}(tr(\boldsymbol{U}^T\boldsymbol{L}_2\boldsymbol{U}) + tr\big(\boldsymbol{V}\boldsymbol{L}_1^{'}\boldsymbol{V}^T\big)) \\ & + \frac{\lambda_2}{2}(||\boldsymbol{U}||_F^2 + ||\boldsymbol{V}||_F^2) \\ & + tr(\Phi\boldsymbol{U}^T) + tr(\Psi\boldsymbol{V}^T). \end{aligned} \tag{16}$$

The partial derivative of $\mathcal{L}_2$ with respect to $\boldsymbol{U}_{ik}$ can be expressed as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial(\boldsymbol{U}_{ik})} =& -\Big(\boldsymbol{Y}_{i\cdot} \odot (\boldsymbol{R}_{i\cdot} - \boldsymbol{U}_{i\cdot}\boldsymbol{V})\Big)(\boldsymbol{V}^T)_{\cdot k} \\ & + \lambda_0 \sum_{G\in\mathcal{G}\&i\in G}(\boldsymbol{U}_{i\cdot} - \bar{\boldsymbol{U}}_G)_k \\ & + \lambda_1((\boldsymbol{D_2} - \boldsymbol{S_2})\boldsymbol{U})_{ik} + \lambda_2(\boldsymbol{U})_{ik} + \Phi_{ik}. \end{aligned} \tag{17}$$

The partial derivative of $\mathcal{L}_2$ with respect to $\boldsymbol{V}_{kj}$ can be expressed as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial(\boldsymbol{V}_{kj})} =& -(\boldsymbol{U}^T)_{k\cdot}\Big(\boldsymbol{Y}_{\cdot j} \odot (\boldsymbol{R}_{\cdot j} - \boldsymbol{U}\boldsymbol{V}_{\cdot j})\Big) \\ & + \lambda_1(\boldsymbol{V}(\boldsymbol{D_2} - \boldsymbol{S_2}))_{kj} + \lambda_2(\boldsymbol{V})_{kj} + \Psi_{kj}. \end{aligned} \tag{18}$$

Using the KKT conditions $\boldsymbol{\phi}_{ik}\boldsymbol{U}_{ik} = 0$ and $\boldsymbol{\psi}_{kj}\boldsymbol{V}_{kj} = 0$, we can get the following multiplicative update equations for $\boldsymbol{U}_{ik}$ and $\boldsymbol{V}_{kj}$:

$$\begin{aligned} \boldsymbol{U}_{ik} \leftarrow \boldsymbol{U}_{ik}&\frac{(\boldsymbol{Y}_{i\cdot} \odot \boldsymbol{R}_{i\cdot})(\boldsymbol{V}^T)_{\cdot k} + \lambda_0\boldsymbol{x}_1 + \lambda_1(\boldsymbol{S_2}\boldsymbol{U})_{ik}}{\boldsymbol{\Psi} + \lambda_0\boldsymbol{x}_2 + \lambda_1(\boldsymbol{D_2}\boldsymbol{U})_{ik} + \lambda_2\boldsymbol{U}_{ik}} \\ \boldsymbol{\Psi} =& \big(\boldsymbol{Y}_{i\cdot} \odot (\boldsymbol{U}_{i\cdot}\boldsymbol{V})\big)(\boldsymbol{V}^T)_{\cdot k} \end{aligned} \tag{19}$$

$$\boldsymbol{V}_{kj} \leftarrow \boldsymbol{V}_{kj}\frac{(\boldsymbol{U}^T)_{k\cdot}(\boldsymbol{Y}_{\cdot j} \odot \boldsymbol{R}_{\cdot j}) + \lambda_1(\boldsymbol{V}\boldsymbol{S_1^{'}})_{kj}}{(\boldsymbol{U}^T)_{k\cdot}\big(\boldsymbol{Y}_{\cdot j} \odot (\boldsymbol{U}\boldsymbol{V}_{\cdot j})\big) + \lambda_1(\boldsymbol{V}\boldsymbol{D_1^{'}})_{kj} + \lambda_2\boldsymbol{V}_{kj}}, \tag{20}$$

where $\boldsymbol{x}_1 = \lambda_0 \sum\limits_{G\in\mathcal{G},i\in G}(\bar{\boldsymbol{U}}_G)_k, \ \boldsymbol{x}_2 = \sum\limits_{G\in\mathcal{G},i\in G}\boldsymbol{U}_{ik}$.

After we get the new updated $\boldsymbol{U}$, the group center $(\bar{\boldsymbol{U}}_G)_k$ can be updated with the following rule:

$$(\bar{\boldsymbol{U}}_G)_k \leftarrow \frac{\sum\limits_{i\in G}\boldsymbol{U}_{ik}}{\sum\limits_{i\in G}1} \tag{21}$$

The computation processes of GC^{2}NMF are summarized in Algorithm 1.

Algorithm 1 GC^{2}NMF

```
Input: $\boldsymbol{R}$: the gene-phenotype association matrix;
    $\boldsymbol{S}_1$, $\boldsymbol{S}_2$: symmetrical proximity matrices;
    $\lambda_1$, $\lambda_2$: parameters;
    $Max\_ites$, $\varepsilon$: the maximum iterations, tolerance;
Output: $\boldsymbol{U}$, $\boldsymbol{V}$: gene and phenotype representations in latent space
$\boldsymbol{U}$, $\boldsymbol{V}$: initialize with random values [0,1];
$ite \leftarrow 1$,
Calculate objective function $\mathcal{O}_2$ according to Eq. (15)
repeat
    $\mathcal{O}_2^{'} \leftarrow \mathcal{O}_2$
    Calculate each group centre: $(\bar{\boldsymbol{U}}_G)_k \leftarrow \frac{\sum_{i \in G} \boldsymbol{U}_{ik}}{\sum_{i \in G} 1}$
    Update $\boldsymbol{U}$ and $\boldsymbol{V}$ with Eq. (19) and Eq. (20), respectively;
    Calculate new objective function $\mathcal{O}_2^{'}$ according to Eq. (15)
    if $|\mathcal{O}_2^{'} - \mathcal{O}_2| < \varepsilon$ then
        exit the loop
    end if
    $ite \leftarrow ite + 1$,
until $ite \geq Max\_ites$
return $\boldsymbol{U}$ and $\boldsymbol{V}$
```

3.5 Computational Complexity Analysis

In this subsection, we discuss the computational time cost of our proposed GC^{2}NMF in comparison to standard NMF. The formulation of learning in GC^{2}NMF is decomposable and can be processed in parallel.

The indicator matrix $\boldsymbol{Y}$ is a sparse matrix, we use p_1 and p_2 to denote the average number of nonzero elements on each row and column of $\boldsymbol{Y}$, respectively; and T as the iteration times in the algorithm. In NMF, for each element in $\boldsymbol{U}$, the time cost is: $O(Km + 4p_1)$, where Km is for the calculation of $(\boldsymbol{U}_{i\cdot}\boldsymbol{V})$, p_1 is for Hadamard product, since $\boldsymbol{U}$ has $m * n$ elements in total, the time cost for updating $\boldsymbol{U}$ in NMF is: $O(TmK(Km + 4p_1))$, in the same way, we can get the time complexity for updating $\boldsymbol{V}$ in NMF: $O(TKn(Kn + 4p_2))$.

Specifically, we assume that there are averagely q1 elements in each group, the average number of groups in which each element appears is q2. Thus, the total time for updating $\boldsymbol{U}$ is $O(TmK(Km + 4P_1 + q_1 + 2q_2 + 2n))$, where $q_1 + 2q_2$ is used for calculating $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in Eq. (19), $2n$ is for $(\boldsymbol{S}_2\boldsymbol{U}_{ik})$ and $(\boldsymbol{D}_2\boldsymbol{U})$. On the other hand, the time complexity for updating $\boldsymbol{V}$ in GC^{2}NMF is $O(TKn(Kn + 4p_2 + 2m))$.

In most cases, the dimension of latent space K is a constant number, and $K \lll min(m, n)$. The max number of iteration times is also a fixed number in the experiments, then the total complexity presented in a general way is $O(m^2)$ and $O(n^2)$ for updating $\boldsymbol{U}$ and $\boldsymbol{V}$ in NMF, $O(m^2 + nm)$ and $O(n^2 + mn)$ for $\boldsymbol{U}$ and $\boldsymbol{V}$ in GC^{2}NMF. The time complexity for GC^{2}NMF and NMF is summarized in Table 2.

3.6 Software Package

A MATLAB software package is available through GitHub at `https://github.com/nkiip/GC2NMF`. It contains all the source code and data used to run GC^{2}NMF.

TABLE 2
Time Complexity Analysis

	NMF	**GCNMF**
Update $\boldsymbol{U}$	$O(TmK(Km + 4p_1))$	$O(TmK(Km + 4P_1 + q_1 + 2q_2 + 2n))$
Update $\boldsymbol{V}$	$O(TKn(Kn + 4p_2))$	$O(TKn(Kn + 4p_2 + 2m))$
in total:	$O(m^2 + n^2)$	$O(m^2 + n^2 + nm)$

TABLE 3
Data Description

Dataset	Relations (A-B)	Number of A	Number of B	Number of (A-B)
mouse	Gene-Phenotype	670	5420	16122
	Gene-Gene	670	670	605
	Phenotype-Phenotype	4481	4481	8084
	Gene-Pathway	670	292	2614
human	Gene-Phenotype	3193	7044	333616
	Gene-Gene	3193	3193	33823
	Phenotype-Phenotype	7044	7044	17272
	Gene-Pathway	3193	296	8791

4 Results and Discussion

4.1 Data preparation

We extract 1376 genes and 5420 phenotypes to build binary gene-phenotype association matrix for mouse from MGI [3]. 8158 pairs of mouse hierarchical phenotype-phenotype relationship are extracted from MGI as well. 3495 genes and 7143 phenotypes are used to build binary gene-phenotype relation matrix for human from HPO [14]. 17408 pairs of human phenotype-phenotype relationship are also extracted from HPO. Gene-gene interaction data are extracted from BIOGRID [5] for both mouse and human. We extract gene group data (i.e. KEGG pathways) from KEGG [12]. The mouse pathway data contains 292 pathways and 7754 genes in total, and the human pathway data contains 296 pathways and 6990 genes in total.

There are some isolated genes and phenotypes in the data sources described above, after the intersection of genes sets and phenotype ontology sets between corresponding data sources, we summarize the details of data used in our experiments in Table 3.

4.2 Evaluation Metric

We use several metrics to evaluate the prediction performance, Area under the Curve (AUC), Normalized Discounted Cumulative Gain (NDCG) and F_1.

$NDCG$ is defined as:

$$NDCG@k = \sum_{u \in U^{test}} NDCG_u@k / |U^{test}|, \tag{22}$$

where $NDCG_u@k = \frac{1}{Y_u}\sum_{i=1}^{k}\frac{2^{t_i}-1}{\log_2(i+1)}$, Y_u is the maximum $DCG_u@k$ score for gene u, and t_i is 1 if the phenotype at i is related to gene u and 0 otherwise, $|U^{test}|$ is the number of genes in test data.

We define $LI_u^{'}$ as the top-k predicted phenotypes for gene u and LI_u as the phenotype list of gene u in test data. $Precision_u@k = |LI_u \cap LI_u^{'}|/|LI_u^{'}|$ and $Recall_u@k = |LI_u \cap LI_u^{'}|/|LI_u|$, then Precision and Recall on the whole test data are defined as follows:

$$\begin{aligned} Precision@k &= \sum_{u \in U^{test}} Precision_u@k/|U^{test}|, \\ Recall@k &= \sum_{u \in U^{test}} Recall_u@k/|U^{test}|. \end{aligned} \tag{23}$$

The F_1 score is defined as:

$$F_1@k = \frac{2 \times Precision@k \times Recall@k}{Precision@k + Recall@k}. \tag{24}$$

4.3 Comparison Methods

In order to show the effectiveness of GC^{2}NMF, we compare GC^{2}NMF with following baselines.

- PMF [19]: It is the basic matrix factorization method, using only gene-phenotype association matrix for prediction. It assumes a diagonal covariance matrix for the Gaussian prior, implying independent latent features.
- BPMF [18]: A fully Bayesian treatment of the Probabilistic Matrix Factorization (PMF) model in which model capacity is controlled automatically by integrating over all model parameters and hyper parameters.
- PPMF [20]: It's a variant and generalization of PMF. A variational inference method is used for the learning process.
- BiRW [24]: It's a network based state-of-the-art method for gene-phenotype association prediction.
- NMF [7]: We also introduce Frobenius norm $||\cdot||_F^2$ of $\boldsymbol{U}$ and $\boldsymbol{V}$ to avoid overfitting in original NMF.

Besides the above baselines, we consider two variants of GC^{2}NMF, one is weighted graph constraint NMF, another is group centric constraint NMF, these two methods are denoted as GC^{2}NMF($\lambda_0 = 0$) and GC^{2}NMF($\lambda_1 = 0$), respectively.

4.4 Parameter Settings

In our model GC^{2}NMF, we need to tune parameters λ_0, λ_1, λ_2 , where λ_0 and λ_1 balance the impacts of group centric constraint and weighted graph constraint, λ_2 controls model complexity to avoid overfitting. λ_2 is chosen from set {1, 10, 100, 1000}. For the range of λ_0 and λ_1, we choose the value in {0, 0.001,0.01,0.1,1,10,100,1000} for λ_0 and λ_1, then we can search in grid to find the best parameters for the model. The hyper-parameters in baselines are tuned according to corresponding paper described.

4.5 Experiment Results

As we have mentioned in previous section, the most notable contributions of our work is that we propose a general framework that combines weighted graph constraint and group centric constraint to alleviate the sparseness problem of gene-phenotype association matrix. Therefore, we examine the effectiveness of weighted graph constraint, group centric constraint and both of these two constraint as a whole in this part. More specifically, we compare the performance of our framework with current state-of-the-art approaches to predict the gene-phenotype association.

We perform our experiments on two biological datasets, one is on mouse species, the other is on human species. We randomly choose 80% data in gene-phenotype association matrix as training data, the remaining 20% data is used as test data to evaluate the predicting results. The experiment for each method is repeated 10 times independently and the average results are reported.

TABLE 4
Experiment on Weighted Graph constraint

Dataset	Evaluation	NMF	GNMF	GC^{2}NMF (λ_0=0)
Mouse	AUC$_{@200}$	0.0312	0.0874	0.0881
	NDCG$_{@200}$	0.0177	0.0470	0.0475
	F1$_{@200}$	0.0016	0.0040	0.0046
Human	AUC$_{@200}$	0.0031	0.1287	0.1343
	NDCG$_{@200}$	0.0039	0.1003	0.1016
	F1$_{@200}$	0.0008	0.0159	0.0166

4.5.1 Effectiveness of Weighted Graph Constraint

In order to show the effectiveness of weighted graph constraint, we compare the experiment results on NMF, NMF with general graph constraint (denoted as GNMF [4] in Table 4), and NMF with weighted graph constraint (denoted as GC^{2}NMF ($\lambda_0 = 0$) in Table 4). From the results in Table 4, we observe that when we consider the general graph constraints on phenotype ontology and gene network in NMF, GNMF performs better than NMF model to a large extent. This means graph constraint does help to figure out the association between similar genes and similar phenotypes. Comparing with GNMF, results on GC^{2}NMF ($\lambda_0 = 0$) demonstrate by taking hierarchical structure of phenotype ontology into consideration, weighted graph constraint helps us learn a better data representation in latent space and get a more accurate performance.

TABLE 5
Experiment on Group Centric Constraint

Dataset	Evaluation	NMF	GC^{2}NMF (λ_1=0)
Mouse	AUC$_{@200}$	0.0312	0.0326
	NDCG$_{@200}$	0.0177	0.0183
	F1$_{@200}$	0.0016	0.0018
Human	AUC$_{@200}$	0.0031	0.0037
	NDCG$_{@200}$	0.0039	0.0046
	F1$_{@200}$	0.0008	0.0013

4.5.2 Effectiveness of Group Centric Constraint

We show the effectiveness of group centric constraint by comparing NMF and NMF with group centric constraint, the latter method is denoted as GC^{2}NMF ($\lambda_1 = 0$). The results are reported in Table 5. GC^{2}NMF ($\lambda_1 = 0$) performs better than NMF. It validates that related genes in a group often associate with similar phenotypes.

It is worth mentioning that comparing with NMF, the improvement brought by group centric constraint is not as much as weighted graph constraint. This may because gene group prior information is not sufficient now. Genes are manually summarized into a group when these genes participate a particular biological process cooperatively. It will cost biologists lots of efforts to validate whether the genes are related to a group with extensive experiments. With the pre-genomic era characterized by the effort to sequence the human genome just being completed, we are entering the post-genomic era that concentrates on harvesting the fruits hidden in the genomic sequences. More gene functions are being revealed day by day, and gene groups are being enriched gradually. In the future, those gene groups will play a more important role in gene-phenotype association prediction task.

4.5.3 *Effectiveness of GC^{2}NMF*

To validate the effectiveness of GC^{2}NMF, which combines both weighted graph constraint and group centric constraint, we do extensive experiments on different evaluation metrics with other baselines and results are shown in Table 6. We observe that GC^{2}NMF beats all baselines in most cases. When we consider weighted graph constraint NMF (GC^{2}NMF with λ_0=0), it has a significant improvement in almost all evaluation metrics compared with NMF. While for NMF with group centric constraint (GC^{2}NMF with λ_1=0), some improvements can be observed on AUC@200, AUC@600, NDCG@200 etc., but the enhancement is not as impressive as weighted graph constraint NMF (GC^{2}NMF with λ_0=0) comparing with NMF. It demonstrates that weighted graph constraint has more impact than group constraint on the predicting performance, this is because graph constraint includes direct gene-gene interactions and phenotype-phenotype relationships. Each interaction and relationship is captured by the model through graph constraint (see Eq. 3 and Eq. 6). However, genes in a group participate in a biological process cooperatively, and a biological process is extremely complex as the result of a series of reactions. Each gene in a gene group only participates a small part of the whole process. Thus the relationship of each gene pair in a group is obviously much weaker than gene-gene interactions in PPI networks. When we consider both constraints together, GC^{2}NMF achieves the best performance on both mouse data and human data.

Network-based method BiRW shows a strong performance over other NMF based baselines. BiRW achieves the best performance in some metrics (like NDCG$_{@600}$ and F1$_{@600}$), because it considers relations of genes and relations of phenotypes (i.e. the relation matrix $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$), and uses them to propagate the relations out to the whole network. Propagation in the global network is one effective way to alleviate the sparseness. However, in prediction problem, the elements in top positions are more important than last positions. In fact, our model GC^{2}NMF has an advantage over all baselines on top position metrics AUC@200, F_1@200, $NDCG$@200, except AUC@200 on human.

4.6 Sparseness Study

In much areas of bioinformatics, the known data only take a small part of the whole. There is a need to explore the performances of each methods with data sparseness constraint. To exploit the sparsity-proof ability of GC^{2}NMF, we conduct an experiment to observe its performance when sparse degree of training data varies. We compare our model with NMF on AUC, NDCG and F1.

For each algorithm, firstly we randomly choose 20% data in gene-phenotype association data matrix as test dataset, the remaining 80% used as training dataset; secondly we choose 20%, 40%, 60%, 80% 100% of the training dataset as the real training dataset in the experiment. Note that the test dataset for different part of training dataset experiment is the same one. For different partition of the training dataset in each method, we repeat the experiment 10 times independently and the average results are reported in Table 7.

An increasingly satisfying performance can be observed as the partition of training data increases in each algorithm on mouse in Table 7, GC^{2}NMF has the best performance in all cases. It is obviously superior to other models especially when the training dataset is only 20%. This demonstrates robustness of our model, which has the ability to learn a representation of data even with very sparse data. For each fixed partition of training data, GC^{2}NMF beats all the compared methods on both mouse and human data, which proves the robustness of GC^{2}NMF. Whereas, we find three metrics less favorably on our model GC^{2}NMF as the partition of training data increases on human. The biological mechanisms on human are more complex than other species, the abnormal results on human demonstrate the good learning ability of GC^{2}NMF under sparse data. Besides, the abnormal results also tell that there is a need to design novel prediction methods to mine more biological underlying patterns on human to reveal intrinsic insight into gene-phenotype association.

5 Conclusion

Gene-phenotype association prediction is a vital task for bioresearch and biomedical. However, known associations only take a small portion. The task suffers from severe sparseness problems. In this paper, we propose a novel method GC^{2}NMF for gene-phenotype association prediction. Taking account of complex biological processes and curated relationship from genes to phenotypes, we combine the effects of biological hierarchical relations of phenotypes with those of gene group to overcome the sparseness problem of gene-phenotype associations. Proposed model GC^{2}NMF applies weighted strategy that measures discriminatory similarity according to phenotype levels and introduces group centric points to keep time complexity linear to the maximum size of gene groups. The encouraging results suggest that our method GC^{2}NMF can not only help to identify novel gene-phenotype associations but also to guide biological experiments for scientific research. Besides, GC^{2}NMF achieves a reliable element representation in latent space and operates efficiently.

Though the weighted graph constraint and group centric constraint are proposed to solve the gene-phenotype asso-

TABLE 6
Experiment Result

Dataset	Evaluation	PMF	BPMF	PPMF	BiRW	NMF	GC^{2}NMF (λ_1=0)	GC^{2}NMF (λ_0=0)	GC^{2}NMF
Mouse	AUC$_{@200}$	0.0306	0.0209	0.0280	0.0566	0.0312	0.0326	0.0881	**0.0923**
	AUC$_{@600}$	0.1724	0.0735	0.1735	0.1732	0.1302	0.1307	0.2627	**0.2630**
	AUC$_{@1000}$	0.1677	0.1235	0.1651	0.1905	0.2703	0.2699	0.3673	**0.3680**
	NDCG$_{@200}$	0.0169	0.0114	0.0165	0.0338	0.0177	0.0183	0.0475	**0.0484**
	NDCG$_{@600}$	0.1161	0.0459	0.1159	**0.1160**	0.0694	0.0692	0.0936	0.0945
	NDCG$_{@1000}$	0.0714	0.0457	0.0709	0.0858	0.1047	0.1045	0.1145	**0.1150**
	F1$_{@200}$	0.0016	0.0010	0.0016	0.0022	0.0016	0.0018	0.0046	**0.0052**
	F1$_{@600}$	0.0058	0.0027	0.0058	**0.0058**	0.0027	0.0027	0.0032	0.0043
	F1$_{@1000}$	0.0018	0.0011	0.0017	0.0018	0.0027	0.0026	0.0026	**0.0033**
Human	AUC$_{@200}$	0.0584	0.0146	0.0584	0.0590	0.0031	0.0037	0.1343	**0.1352**
	AUC$_{@600}$	0.1724	0.0735	0.1735	0.1732	0.0957	0.0969	**0.2902**	0.2897
	AUC$_{@1000}$	0.1724	0.0735	0.1735	0.1732	0.2038	0.2048	0.3985	**0.3987**
	NDCG$_{@200}$	0.0537	0.0127	0.0534	0.0538	0.0039	0.0046	0.1016	**0.1025**
	NDCG$_{@600}$	0.1161	0.0459	0.1159	0.1160	0.0767	0.0772	0.1738	**0.1749**
	NDCG$_{@1000}$	0.1161	0.0459	0.1159	0.1160	0.1265	0.1265	0.2147	**0.2163**
	F1$_{@200}$	0.0066	0.0023	0.0067	0.0067	0.0008	0.0013	0.0166	**0.0171**
	F1$_{@600}$	0.0058	0.0027	0.0058	0.0058	0.0076	0.0076	0.0133	**0.0139**
	F1$_{@1000}$	0.0058	0.0027	0.0058	0.0058	0.0082	0.0082	0.0111	**0.0117**

TABLE 7
Experiment Result under Different Sparseness

Dataset	Train	NMF			GC^{2}NMF(λ_0=0)			GC^{2}NMF(λ_1=0)			GC^{2}NMF		
		AUC	NDCG	F1	AUC	NDCG	F1	AUC	NDCG	F1	AUC	NDCG	F1
Mouse	20%	0.0691	0.0275	0.0007	0.2479	0.0890	0.0022	0.0716	0.0284	0.0007	**0.2497**	**0.0897**	**0.0023**
	40%	0.0725	0.0287	0.0007	0.2728	0.0940	0.0023	0.0747	0.0293	0.0007	**0.2754**	**0.0948**	**0.0024**
	60%	0.0815	0.0323	0.0008	0.2988	0.0994	0.0024	0.0825	0.0324	0.0008	**0.3007**	**0.0995**	**0.0026**
	80%	0.1204	0.0677	0.0018	0.3298	0.1058	0.0025	0.1244	0.0700	0.0018	**0.3299**	**0.1059**	**0.0027**
	100%	0.2665	0.1035	0.0026	0.3673	0.1145	0.0026	0.2699	0.1045	0.0026	**0.3674**	**0.1147**	**0.0028**
Human	20%	0.0054	0.0039	0.0002	0.4914	0.2492	0.0129	0.0058	0.0039	0.0002	**0.4916**	**0.2493**	**0.0130**
	40%	0.0120	0.0224	0.0015	0.4710	0.2407	0.0125	0.0120	0.0224	0.0015	**0.4727**	**0.2437**	**0.0126**
	60%	0.0498	0.0637	0.0043	0.4486	0.2338	0.0120	0.0498	0.0637	0.0043	**0.4509**	**0.2340**	**0.0121**
	80%	0.1141	0.0959	0.0064	0.4257	0.2242	0.0116	0.1162	0.0966	0.0064	**0.4261**	**0.2248**	**0.0117**
	100%	0.2038	0.1265	0.0082	0.4038	0.2188	0.0111	0.2648	0.1265	0.0082	**0.4040**	**0.2189**	**0.0112**

* The metrics AUC, NDCG, F$_1$ are at 1000.

ciation prediction problem, we should notice that these two constraints can be widely applied in other fields, where the data contain hierarchical graphs and (or) element groups.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61300166 and No. 61105049), the Natural Science Foundation of Tianjin (No. 14JC-QNJC00600), and the Science and Technology Planning Project of Tianjin (No. 13ZCZDGX01098).

References

[1] Sattar Hashemi Ali Mashhoori. *Incorporating Hierarchical Information into the Matrix Factorization Model for Collaborative Filtering*, volume 7198 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[2] Kirell Benzi, Vassilis Kalofolias, Xavier Bresson, and Pierre Vandergheynst. Song Recommendation with Non-Negative Matrix Factorization and Graph Total Variation. ICASSP, March 2016.

[3] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and J. A. Blake. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Research*, 36(Database):D724–D728, dec 2007.

[4] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–60, aug 2011.

[5] A. Chatr-aryamontri and Breitkreutz. The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823, jan 2013.

[6] Yang Chen and Li Li. Phenome-driven disease genetics prediction toward drug discovery. *Bioinformatics*, 31(12):i276–i283, jun 2015.

[7] H. Sebastian Seung Daniel D. Lee. Algorithms for Non-negative Matrix Factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.

[8] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, jun 2010.

[9] Janan T Eppig, Judith A Blake, Carol J Bult, James A Kadin, and Joel E Richardson. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic acids research*, 43(Database issue):D726–36, jan 2015.

[10] Taehyun Hwang and Rui Kuang. A heterogeneous label propagation algorithm for disease gene discovery. *SIAM*, page 12, 2010.

[11] Aref Jeribi. *Spectral Graph Theory*. American Mathematical Society, 1997.

[12] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, jan 2016.

[13] Sebastian Köhler and Sebastian Bauer. Walking the Interactome for

Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82(4):949–958, apr 2008.

[14] Sebastian Köhler and et al. Doelken. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(Database issue):D966–74, jan 2014.

[15] Yongjin Li and Jagdish C Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–24, may 2010.

[16] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. SoRec: social recommendation using probabilistic matrix factorization. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, page 931, New York, New York, USA, oct 2008. ACM Press.

[17] Roozbeh Rajabi, Mahdi Khodadadzadeh, and Hassan Ghassemian. Graph Regularized Nonnegative Matrix Factorization for Hyperspectral Data Unmixing. In *2011 7th Iranian Conference on Machine Vision and Image Processing*, pages 1–4. IEEE, nov 2011.

[18] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 880–887, New York, New York, USA, jul 2008. ACM Press.

[19] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

[20] Hanhuai Shan and Arindam Banerjee. Generalized Probabilistic Matrix Factorizations for Collaborative Filtering. In *2010 IEEE International Conference on Data Mining*, pages 1025–1030. IEEE, dec 2010.

[21] Cynthia L Smith and Janan T Eppig. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley interdisciplinary reviews. Systems biology and medicine*, 1(3):390–9, jan 2009.

[22] Oron Vanunu and Oded Magger. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Compu Bio*, 6(1):e1000641, jan 2010.

[23] Xuebing Wu and Rui Jiang. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4:189, may 2008.

[24] MaoQiang Xie, YingJie Xu, YaoGong Zhang, TaeHyun Hwang, and Rui Kuang. Network-based Phenome-Genome Association Prediction by Bi-Random Walk. *PloS one*, 10(5):e0125138, jan 2015.

[25] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. *Proceedings of the 2006 SIAM International Conference on Data Mining*, volume 2006. Society for Industrial and Applied Mathematics, Philadelphia, PA, apr 2006.