RESEARCH

Consistent Multiple Nonnegative Matrix Factorization with Hierarchical Information for Gene Functional Modules Mining

Jane E Doe[??*†] and John RS Smith[??,??]

*Correspondence:
jane.e.doe@cambridge.co.uk
[??]Department of Zoology, Cambridge, Waterloo Road, London, UK
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Motivation: An increasing amount of genome-phenome association data, has provided us an unprecedented chance to globally explore the underlying genetic mechanisms and understand the regularization between genes and diseases with a deeper sight perspective. Gene clustering, which reveals the interactions between genes and help researchers to identify candidate genes as drug targets et,al., has always been a significant and valuable problem by using genome-phenome association data. Nevertheless, the hierarchical structure of phenotype ontology has been rarely leveraged by previous gene clustering studies, the properties of genes, diseases and their relationships has not been fully explored, which may result in missing the chance to discover the crucial fact in biology. Thereby It is challenging to utilize this neglected hierarchical character of phenotype ontology to gain understanding of biological system.

Results: We propose a novel method, Consistent Multiple Nonnegative Matrix Factorization (CMNMF), to utilize the hierarchical structure of phenotype ontology to cluster genes in factoring genome-phenome association data on mouse. The CMNMF method, constrains the gene cluster to remain consistent while interacting with different phenotype ontology levels in decomposing the genome-phenome association matrix, meanwhile it restricts the similarity of phenotype pairs to satisfy the hierarchical structure. The performance of our proposed method and other baselines (including NMF, K-means, KK-kernel, HAC) are evaluated on F_1 measure and Gene Ontology similarity measure. The results show the effectiveness of our proposed method compared with other baselines. Additionally, we conduct the regularized Consistent Multiple Nonnegative Matrix Factorization(R-CMNMF) on gene clustering, R-CMNMF beats other baselines as well.

Our work explores the crucial impact of hierarchical structure of phenotype ontology in gene clustering, and we show our proposed method outperforms baselines in both unsupervised and supervised manner, which provides a new perspective to conduct research on biological data exploration.

Availability: Github

Contact: name@bio.comname@bio.com

Second part title: Text for this section.

Keywords: sample; article; author

Introduction

With the development of technology, biomedical researchers has collected a large amount of valuable biological data by these years, especially the genome-phenome

association data on mouse, whose research achievement may transfer to men's disease study, it is of great significance for drug development and disease treatment for humans. Thereby it is necessary and essential for researchers to have a deeper sight investigation on mouse data. The international database resources such as Mouse Genome Informatics(MGI), Kyoto Encyclopedia of Genes and Genomes(KEGG) et al. provide more and more multiple types of mouse biological data. However, how to integrate and utilize those data in a effective way to discover the underlying patterns and biological mechanism is still a challenging issue, which has drawn growing attention in the literature.

It is well known that genes in mammal genomes are usually organized into groups functionally associate with phenotype groups(ref).Identification of modules of functionally related genes is a crucial first step towards dissecting the regulatory circuitry underlying biological processes. Co-regulated or functionally related genes are likely to reveal themselves by associations with different diseases. These modules may provide clues about the main biological processes associated to different physiological states and provide guidance for candidate genes of genetic diseases.

Phenotypes, is the composite of an organism's observable characteristics or traits, results from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two. (1,2). The key to achieving desired gene modules such as favorable disease treatment outcomes lies in the understanding of the relation between phenotypes and the biological roles of genes (3,5). Phenotype ontology was created to serve as a standardized vocabulary of phenotypic abnormalities that have been seen in diseases, can be used as a computational representation of phenotype knowledge based upon a hierarchical structure(ref). This kind of structure provides researchers extra information about the relationships between phenotype pairs. Whereas ,as far as we know, the hierarchical structure of phenotype ontology has not been utilized for gene module identification, the phenotype ontology in different levels reflects underlying associations while interacting with disease genes. With taking advantage of this character, it gives us a new perspective to explore the patterns behind the biological data.

Much research studies have been conducted on integration of multiple sources of biological data to mine the hidden patterns and the relationships between distinct data sources. Some procedures based on gene expression profiles, seeking to map different experimental data types, such as gene expression, miRNA expression, and copy number variation to a common space of known biological pathways or sets(Khatri et al.,2012; Mitrea et al.,2013). (Zhang et al., 2011) focuses on integrating multiple type genomic data to identify microRNA-gene regulatory modules for cancers. Extended Dirichlet mixture model(Lock and Dunson,2013) and principal component analysis(Lock et al., 2013) make the distinction between common and distinct effects across sources. Some studies incorporated clinical phenotype data to increase the ability of identifying new disease-associated genes(Hwang et al., 2012; Lage et al., 2007; Li and Patra, 2010; Vanunu et al., 2010; Wu et al., 2008, 2009)("Phenome-driven disease genetics prediction toward drug discovery"), a key assumption in these methods is that similar disease phenotypes reflect overlapping genetic causes(Houle et at.,2010).("Phenome-driven disease genetics prediction toward drug discovery"). Because some data contains specific structures behind it,

thus some structure based methods are proposed and all achieve good results. In the collaborative recommendation task, some grouping strategies based on structure information are proposed, such as grouping users by social networks, grouping items by defined hierarchy or grouping both of them (SoRec,HMF,HGMF). HPMF focuses on predicting missing traits for plants which incorporates hierarchical phylogenetic information into matrix factorization (HPMF). The results of these methods demonstrate that considering auxiliary structured information can bring a better performance.

In this study, we developed a novel NMF based approach consistent multiple non-negative matrix factorization (CMNMF) to combine structured phenotype ontology data on mouse disease phenotype to predict disease-associated gene modules. The key ideas of our method are that the same gene should be active in the same modules when interacting with phenotypes from different levels of the hierarchical phenotype ontology and the similarity of phenotype pairs, which has parent-child relationships, should keep high in the hierarchical structure. To demonstrate the approach, we conduct our proposed method on a series of measurements with other baselines on gene clustering task in an unsupervised way, the performance of our CMNMF on $F_1measure$, $Randindex$, $Jaccardindex$, and $GOScore$ outperform other baselines, including K-means, PCA K-means KernelK-means, HAC, NMF, LDA. Besides, we conduct regularized CMNMF(R-CMNMF) on gene classification, the AUC score demonstrates the advantages of CMNMF over SVM, LP.

Method

In this section, we describe our framework for the consistent multiple non-negative matrix factorization with hierarchical information of phenotype ontology to identify gene modules. We designed an objective function with three components. The first and second components are based on non-negative genome-phenome association data with respect to distinct level phenotype ontologies. The third one considers the effects of interactions from phenotype ontology. By optimizing this objective function, we obtain a consistent decomposition of two genome-phenome association matrix, which reveals gene functional modules inherent in genome-phenome association data jointly.

Data Preparation

The mouse gene-phenotype ontology associations are extracted from file "MGI_Geno_Disease.rpt" (March-2015) downloaded from MGI [1], in which 5894 gene-phenotype associations at level 4 and 6817 associations at level 5 are kept. 2144 ontology terms at level 4 (parent level) and 2719 ontology terms at level 5 (child level) are selected from the 10748 MP terms in the OBO file[2]. 2866 hierarchial mapping relations(adjacent parent-child relationships) between phenotypes in level 4 and level 5 are kept as M.

In our gene module clustering problem, after removing the phenotypes not present in phenotypes of level 4 and level 5, we generate a dataset containing 1274 disease genes and 4756 phenotypes.

[1] ftp://ftp.informatics.jax.org/pub/reports/MGI_Geno_Disease.rpt
[2] ftp://ftp.informatics.jax.org/pub/reports/MPheno_OBO.ontology

Notations	Explanations
$\boldsymbol{A_1}$	genome-phenome association matrix with phenotype ontology from level 4
$\boldsymbol{A_2}$	genome-phenome association matrix with phenotype ontology from level 5
$\boldsymbol{G}$	Gene cluster membership
$\boldsymbol{P}$	Phenotype cluster membership
$\boldsymbol{M}$	Phenotype ontologies relationship
m	Number of disease genes
K	Number of gene clusters(e.g. classes)
k_1	Number of phenotype ontology in level 4
k_2	Number of phenotype ontology in level 5

To evaluate the clustering performance for CMNMF, we crawled 280 mouse gene pathways from KEGG[3] and selected 229 of them as the " ground truth" of gene modules. Genes in a signal pathway can be considered as a gene functional module (or a gene cluster).

In the leave-one-out cross-validation, after preprocessing(removing the phenotypes not present in gene-phenotype association file and phenotype ontology file, and removing genes not present in the gene pathways), we generated a dataset with 703 genes and 4116 phenotypes.

Problem formulation

The notations and definitions used in the article are specified in Table(**??**). We denote $A_{(m\times n)}$ a binary gene-phenotype association matrix by m genes and n phenotypes, where A_{ij} is set with 1 for known association and 0 otherwise. The goal of factorizing matrix A based on NMF(Lee and Seung,1999) is to derive gene functional clusters . The loss of it can be defined as:

$$min_{P,G}||A - GP||_F^2 \tag{1}$$

Where G and P denotes gene clusters and phenotype clusters , respectively. The notation $||\bullet||_F$ means the $Frobenius$ norm of a matrix.

Based on multiplicative update rules(Lee and seung,2000), we can get a non-negative solution of G and P. The elements of G can be interpreted as latent modules associated with genes, thus a couple of clusters of genes are achieved with a good interpretation on NMF. However, the NMF method in this form has not taken the hierarchical mapping information of phenotype ontology into consideration. To address this problem, we design a loss function with two aspects for formulating the loss with hierarchical information of phenotype ontologies:

- The first aspect is a consistent constraint on gene cluster, the reason is intuitive, one gene interacts with a child phenotype ontology, it should interact with the child ontology's parent phenotype ontology, because parent phenotype ontologies are generalization of child phenotype ontologies, the child phenotype ontologies are specification of parent phenotype ontologies. (e.g. xxxxx)Fig. 1(a) shows The gene clustering results on general and specific phenotype ontology level, on which we put a consistent constrain.

[3]http://www.genome.jp/kegg/pathway.html

- The second component is a hierarchical mapping constraint between phenotype ontologies from parent level and ones from child level. (Fig. 1(b)). Because phenotype ontologies coming from the adjacent level, there should exist strong relationships between these parent-child ontologies pairs, in which we apply hierarchical mapping constrains to demonstrate the relationships between them. The clusters of phenotype ontologies from different levels should be consistent as well.

By optimizing these two components, a CMNMF (Consistent Multiple Non-negative Matrix Factorization) algorithm is proposed to obtain consistent gene clusters.

Loss Functions for Penalizing Inconsistence

Motivated by above consistent assumption, we extract phenotype ontologies from two adjacent levels[4], and two gene-phenotype association matrices A_1 and A_2 are extracted from gene-phenotype text file [5]. We assume that the factorizations on A_1 and A_2 for the gene cluster should be consistent, although the genes are annotated by adjacent level phenotype ontologies. In our work, we use a common basis gene cluster matrix G to achieve this goal. The representation of the data can be derived by optimizing the following quadratic objective function:

$$L_C = \min_{G,P_1,P_2} \|A_1 - GP_1\|_F^2 + \alpha \|A_2 - GP_2\|_F^2 , \tag{2}$$

where α is a parameter to balance the factorization error from different levels.

For the hierarchical mapping constraints on phenotype ontologies, loss function in Eq(3) is used to encourage the interactions between phenotypes from parent level and child level.

$$L_H = \sum_{ij} M_{ij} (P_1^{(i)})^T P_2^{(j)} = tr(P_1 M P_2^T), \tag{3}$$

where M denotes the hierarchical mapping relation matrix between phenotype ontologies from different levels. $M_{ij} = 1$, if phenotype i and phenotype j have a parent-child relationship, otherwise 0. We enforce hierarchical mapping constraints by maximizing the mapping between the phenotype ontologies in gene-phenotype network A_1 and A_2.

Regularization by Sparse Constraint

Since the known gene-phenotype associations are sparse and only cover a part of genes and phenotypes, the regularization with L_2 norm is proposed to control the sparseness of G, P_1 and P_2. The coefficients λ_1 and λ_2 balance the regularization terms. Finally, the framework of regularized CMNMF with sparse constraint is formulated as follows:

$$\begin{aligned} \min_{G,P_1,P_2} & \|A_1 - GP_1\|_F^2 + \alpha \|A_2 - GP_2\|_F^2 - \gamma tr(P_1 M P_2^T) \\ & + \lambda_1 \|G\|_F^2 + \lambda_2 (\|P_1\|_F^2 + \|P_2\|_F^2) \\ & \text{s.t.} \quad \sum_j G_{ij} = 1, \quad \sum_i (P_1)_{ij} = 1, \quad \sum_i (P_2)_{ij} = 1 \end{aligned} \tag{4}$$

[4] file can be downloaded from ...
[5] file can be downloaded from ...

where $\alpha, \gamma, \lambda_1, \lambda_2$ are parameters to balance the trade of each component.

Multiplicative Update Algorithms

we extend the optimization algorithms of the original HMF. The alternative iterative scheme to solve the problem with respect to one variable while fixing the other variables.

Computation of G in CMNMF

If we fix variables P_1 and P_2, solving Eq(4) with respect to G is equivalent to minimizing the following function:

$$\begin{aligned} L(G) &= \|A_1 - GP_1\|_F^2 + \alpha \|A_2 - GP_2\|_F^2 + \lambda_1 \|G\|_F^2 \\ \text{s.t.} \quad & \sum_j G_{ij} = 1, G \geq 0 \end{aligned}$$

Using the Karush-Kuhn-Tucher(KKT) conditions $\phi_{ij} G_{ij} = 0$, the partial derivative of $L(G)$ with respect to G is:

$$\frac{\partial L}{\partial G} = -2(A_1 P_1^T - GP_1 P_1^T) - 2\alpha(A_2 P_2^T - GP_2 P_2^T) + 2\lambda_1 G + \phi$$

The multiplicative update rule is:

$$G_{ij} \leftarrow G_{ij} \frac{(A_1 P_1^T + \alpha A_2 P_2^T)_{ij}}{(GP_1 P_1^T + \alpha GP_2 P_2^T + \lambda_1 G)_{ij}}$$

To satisfy the equality constraint, we normalize G_{ij} as

$$G_{ij} \leftarrow \frac{G_{ij}}{\sum_j G_{ij}}$$

Computation of P_1 and P_2 in CMNMF

When G is computed, solving Eq(4) with respect to P_1 and P_2 is equivalent to minimizing the following function:

$$\begin{aligned} L(P_1) &= \|A_1 - GP_1\|_F^2 - \gamma tr(P_1 M P_2^T) + \lambda_2(\|P_1\|_F^2) \\ \text{s.t.} \quad & \sum_i (P_1)_{ij} = 1, P_1 \geq 0 \\ L(P_2) &= \alpha \|A_2 - GP_2\|_F^2 - \gamma tr(P_1 M P_2^T) + \lambda_2(\|P_2\|_F^2) \\ \text{s.t.} \quad & \sum_i (P_2)_{ij} = 1, P_1 \geq 0 \end{aligned}$$

the partial derivative of $L(P_1)$,$L(P_2)$ with respect to P_1 and P_2 are:

$$\begin{aligned} \frac{\partial L(P_1)}{\partial P_1} &= -2(G^T A_1 - G^T G P_1) - \gamma P_2 M^T + 2\lambda_2 P_1 + \Omega \\ \frac{\partial L(P_2)}{\partial P_2} &= -2\alpha(G^T A_2 - G^T G P_2) - \gamma P_1 M + 2\lambda_2 P_2 + \psi \end{aligned}$$

Using the Karush-Kuhn-Tucher(KKT) conditions $\varphi_{ij}(P_1)_{ij} = 0, \psi_{ij}(P_2)_{ij} = 0$, the multiplicative update rule is(note that when we compute P_1, we take P_2 fixed, vise versa):

$$(P_1)_{ij} \leftarrow (P_1)_{ij} \frac{(G^T A_1 + \frac{1}{2}\gamma P_2 M^T)_{ij}}{(G^T G P_1 + \lambda_2 P_1)_{ij}}$$
$$(P_2)_{ij} \leftarrow (P_2)_{ij} \frac{(\alpha G^T A_2 + \frac{1}{2}\gamma P_1 M)_{ij}}{(\alpha G^T G P_2 + \lambda_2 P_2)_{ij}}$$

To satisfy the equality constraint, we normalize $(P_1)_{ij}$ and $(P_2)_{ij}$ as

$$(P_1)_{ij} \leftarrow \frac{(P_1)_{ij}}{\sum_i (P_1)_{ij}}, \quad (P_2)_{ij} \leftarrow \frac{(P_2)_{ij}}{\sum_i (P_2)_{ij}}$$

Supervised CMNMF

By considering prior gene cluster information, our proposed CMNMF can be utilized for gene-phenotype association problem. Thus, the loss function can be rewrite as:

$$\begin{aligned} \min_{G,P_1,P_2} & \|A_1 - GP_1\|_F^2 + \alpha \|A_2 - GP_2\|_F^2 + \beta \|G - G_0\|_F^2 \\ & - \gamma tr(P_1 M P_2^T) + \lambda_1 \|G\|_F^2 + \lambda_2 (\|P_1\|_F^2 + \|P_2\|_F^2) \\ & \sum_j G_{ij} = 1, \quad \sum_i (P_1)_{ij} = 1, \quad \sum_i (P_2)_{ij} = 1 \end{aligned} \tag{5}$$

Computation of G in Supervised CMNMF

We fix variables P_1 and P_2, solving Eq(5) with respect to G is equivalent to minimize the following function:

$$\begin{aligned} & L_s(G) = \|A_1 - GP_1\|_F^2 + \alpha \|A_2 - GP_2\|_F^2 + \beta \|G - G_0\|_F^2 + \lambda_1 \|G\|_F^2 \\ & \text{s.t.} \quad \sum_j G_{ij} = 1, G \geq 0 \end{aligned}$$

the partial derivative of $L(G)$ with respect to G is:

$$\frac{\partial L_s(G)}{\partial G} = -2(A_1 P_1^T - G P_1 P_1^T) - 2\alpha(A_2 P_2^T - G P_2 P_2^T) + 2\beta(G - G_0) + 2\lambda_1 G + \phi$$

Using the Karush-Kuhn-Tucher(KKT) conditions $\phi_{ij} G_{ij} = 0$, the multiplicative update rule is:

$$G_{ij} \leftarrow G_{ij} \frac{(A_1 P_1^T + \alpha A_2 P_2^T + \beta G_0)_{ij}}{(G P_1 P_1^T + \alpha G P_2 P_2^T + \lambda_1 G + \beta G)_{ij}}$$

To satisfy the equality constraint, we normalize G_{ij} as

$$G_{ij} \leftarrow \frac{G_{ij}}{\sum_j G_{ij}}$$

Algorithm 1 CMNMF

Input: gene-phenotype association matrix A_1,A_2, number of cluster dimensions K, and parameter α,β,γ,λ_1,λ_2,

Output: model parameters G, P_1, P_2

1: $G, P_1, P_2 \leftarrow$ random values
2: **repeat**
3: Update $G_{ij} \leftarrow G_{ij} \frac{(A_1 P_1^T + \alpha A_2 P_2^T)_{ij}}{(G P_1 P_1^T + \alpha G P_2 P_2^T + \lambda_1 G)_{ij}}$
4: Normalize $G_{ij} \leftarrow \frac{G_{ij}}{\sum_j G_{ij}}$
5: Update $(P_1)_{ij} \leftarrow (P_1)_{ij} \frac{(G^T A_1 + \frac{1}{2}\gamma P_2 M^T)_{ij}}{(G^T G P_1 + \lambda_2 P_1)_{ij}}$, $(P_2)_{ij} \leftarrow (P_2)_{ij} \frac{(\alpha G^T A_2 + \frac{1}{2}\gamma P_1 M)_{ij}}{(\alpha G^T G P_2 + \lambda_2 P_2)_{ij}}$
6: Normalize $(P_1)_{ij} \leftarrow \frac{(P_1)_{ij}}{\sum_i (P_1)_{ij}}$, $(P_2)_{ij} \leftarrow \frac{(P_2)_{ij}}{\sum_i (P_2)_{ij}}$
7: **until** convergence
8: **return** G, P_1, P_2

Computation of P_1 and P_2 in Supervised CMNMF

Because the updating rules of supervised CMNMF for P_1 and P_2 are the same as CMNMF above, we would not present it here again.

The CMNMF Algorithm

The complete CMNMF algorithm is represented in **Algorithm 1**. Our algorithm CMNMF can optimize the objective function iteratively by fixing one variable alternatively. As in the original NMF model, the cost of Eq(1) is not convex in G and P jointly, but it is convex in G for fixed P_1 and P_2, vice versa. the Lagrange multiplier method can be applicable here to give an iterative algorithm to guarantee the algorithm to converge to a local minima[**?**]. For supervised CMNMF algorithm, we just need to change the update rule for G as we talked in Computation of G in Supervised CMNMF part.

Experiments and Analysis

Our method is demonstrated on sampled gene-phenotype association matrixes at first by comparing the difference of conventional NMF and proposed CMNMF. We then execute CMNMF and four baseline methods on MGI mouse gene-phenotype ontology associations for mining mouse gene functional modules. Moreover, the evaluation of clustering results and parameter tuning are performed. Finally, gene ontology enrichment analysis is conducted to evaluate the biological significance of mined gene modules.

1 CMNMF on Sampled Data

We sampled gene-phenotype association matrixes A_1, A_2 and hierarchical mapping relations M from real data to illustrate the usage of the hierarchical information. Fig.4(a) describes the structure of the sampled data in which genes are assigned into three groups. For each group, there are two genes associating with phenotypes in the child level and another gene associating with the parent phenotypes. The task is to cluster the three genes into a same group. Due to the third gene doesn't have any associated phenotypes in the same level with that another two genes associate with, neither on A_1 nor on A_2 can NMF group the third gene into a same cluster with another two as shown in Fig.(4)(b-1) and Fig.(4)(b-2). To analyze how the two proposed constraints work, CMNMF is applied to the simulated data to cluster

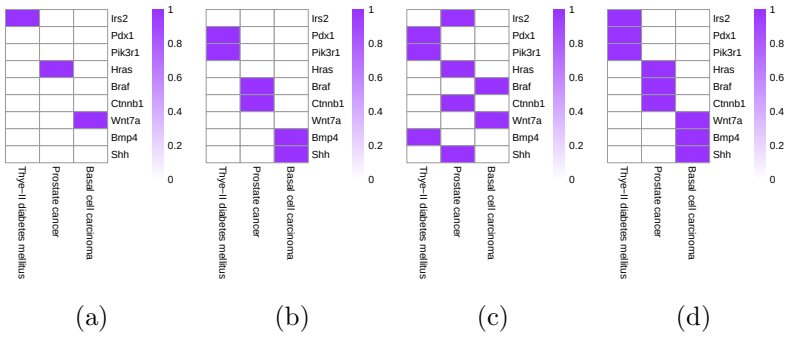

Figure 1 Performance of CMNMF and NMF under different conditions: (a) Cluster result of NMF on parent level, (b) Cluster result of NMF on child level, (c) Cluster result of CMNMF(γ=0), (d) Cluster result of CMNMF(γ=1)

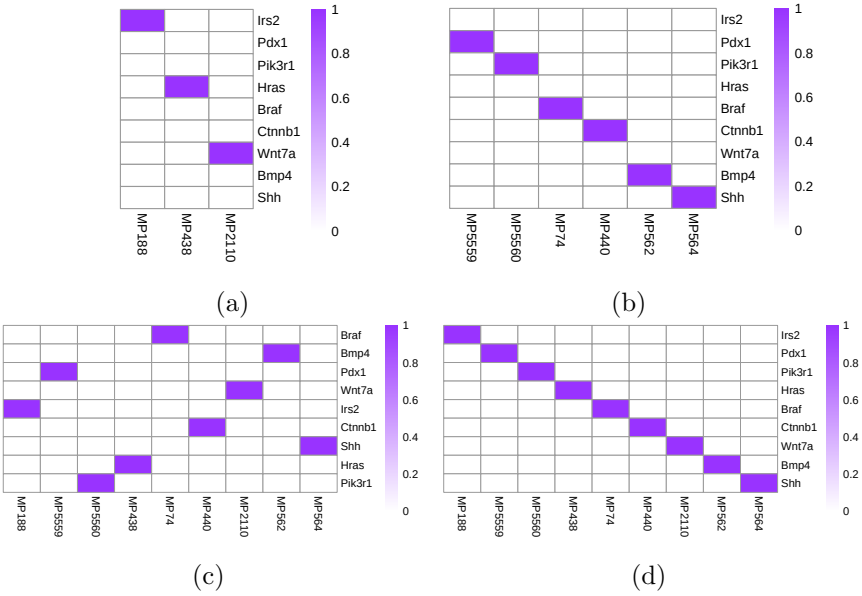

Figure 2 Performance of CMNMF and NMF under different conditions: (a) Cluster result of NMF on parent level, (b) Cluster result of NMF on child level, (c) Cluster result of CMNMF(γ=0), (d) Cluster result of CMNMF(γ=1)

the genes. We first set the parameter γ which controls the effect of hierarchical constraint on CMNMF to zero. As shown in Fig.4(b-3), combing information from two levels and restricting the cluster results of genes to remain consistent indeed help improve the performance. But there still exists little misclassification (the third gene in the first group is wrongly assigned to the second group). When introducing the hierarchical constraint (γ=1) to the model, the CMNMF method groups all the genes into the correct clusters (See Fig.4(b-4)), which demonstrates that the two constraints working together can get the best performance. The reasons for the performance improvement are that with the consistent constraint, CMNMF can incorporate information from two levels and take advantage of the complement characteristic of information from different levels to overcome the shortcoming of lacking enough information from only one level and the hierarchy constraint can

help restrict cluster results according with the structure of data which narrows the range of solution space so that we can get a better solution.

2 Functional gene module similarity analysis

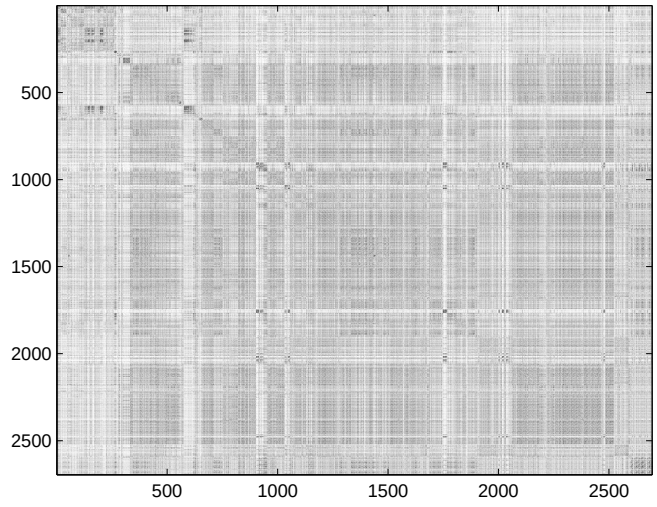

Figure 3 *gene group in pathway's similarity under GO − term*

3 Comparison with Baseline Methods by Mining Gene Modules

In this part, we evaluate the proposed CMNMF and baseline methods on MGI mouse gene-phenotype ontology association data. CMNMF is executed to mine gene functional modules and it is compared with the following clustering methods, K-Means, Kernel-Kmeans, LDA, HAC (Hierarchical Agglomerative Clustering) [**?**], NMF and HMF(Hierarchical Matrix Factorization) [**?**]. To be fair, all matrix factorization-based methods are implemented with sparse and non-negative constraints. We measured the performance by the Jaccard index coefficient [**?**], F_1 measure, GO similarity score, and Rand index. Higher coefficient indicates better clustering result. The results are shown in Table().

The clustering results are column-normalized by using z-score, and the G_{ij} will be set as 0 if it is less than 3σ. All the methods are repeated ten times with different initial values and average Jaccard coefficients are reported in Fig.. Note that the proposed method performs better than other methods except HAC as shown in Fig.**??**(a). But the range of gene clusters size of HAC is much larger than other methods as shown in Fig.**??**(b). In fact, HAC puts 76.5% of genes into one cluster and the other clusters share the left genes which brings it a better Jaccard coefficient, but it cannot help to identify significant gene functional modules. Additionally, the standard deviations of all the methods are small, which demonstrates that these methods including our CMNMF are insensitive to the initial values (Fig.**??**(a)).

The α in our CMNMF model which balances the contributions from different levels is an important parameter. When α is close to 0, the model degenerates into

Table 1 F-Score jaccard-Score GO computed by the cluster result under different clustering function.

	F1-Score	Jaccard-Index	GO	Rand-Index
HAC	0.0777	0.0404	0.7742	0.9646
K-Means	0.0662	0.0343	0.7932	0.9568
Kernel-Kmeans	0.0647	0.0334	0.8148	0.9661
LDA	0.979	0.0515	0.8032	0.9643
NMF	0.093	0.0488	0.67	0.9736
HMF	0.0942	0.0495	0.673	0.9732
CMNMF	**0.1044**	**0.0551**	**0.6226**	**0.9701**

NMF on level 4 and when α is big enough information from level 5 leads the model. The performance of CMNMF with different α is shown in Table **??**. We can find that neither α is close to 0 nor big enough the CMNMF model performs best. When α is near 1, our method obtains a best performance which indicates that associations from different levels are mutual complementation with each other and our CMNMF model successfully captures this characteristic.

4 parameter tuning

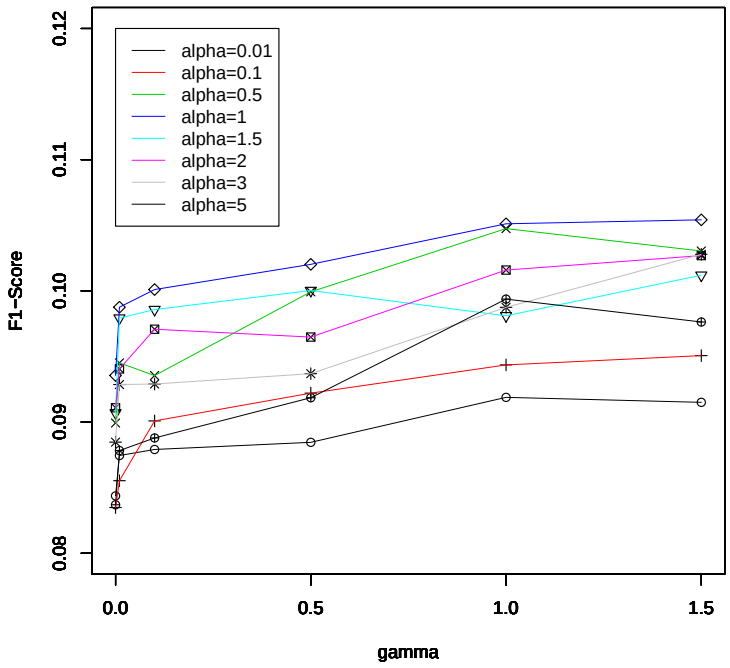

Figure 4 Performance of CMNMF and NMF under different conditions: (a) Cluster result of NMF on parent level, (b) Cluster result of NMF on child level, (c) Cluster result of CMNMF(γ=0), (d) Cluster result of CMNMF(γ=1)

Parameters are tuned by cross-validation for all the competitive methods respectively. For all matrix factorization methods, the balancing parameter of sparse constraints on gene clusters λ_1 and phenotype clusters λ_2 are set by the grid {0.001,0.01,0.1,1}. For CMNMF, the weight of factorization on child level α are set bu the grid {0,0.1,0.2,0.5,0.7,0.9,1,1.1,1.3,1.5,2,3,5,10},γ is set by the range

{0,0.01,0.1,0.2,0.3,0.5,0.7,0.9,1,1.5}. The number of modules k is picked to be 229 for all methods, approximating to the number of KEGG mouse pathways. Initial values of G and P (or P_1 and P_2) are randomized with value between 0 to 1.

Here, we would give more details about how the parameter α and γ affect the performance.

5 Regularized CMNMF

In the experiment of disease gene discovery, we collected the member genes in the 229 pathways from KEGG. In the preprocessed data, there are 145 member genes in 21 KEGG disease pathways such as Alzheimer, diabetes and cancer-related pathways. In the leave-one-out cross-validation, each of the 145 member gene was held out and then classified into the 229 pathways as a multi-label classification problem since some of the disease genes are members of multiple pathways. The higher the target pathways in the ranking of the 145 pathways, the better the performance. We measured the performance by the AUC. LP was applied on the gene similarity network, which was formed by genome-phenome association network, to predict the disease genes as the baseline. The other 144 member genes was used as the initialization of label propagations to classify the held-out gene. another baseline is the LDA model.The average AUC across the member genes by all the methods are reported in Table x. The results clearly show that our methods CMNMF more accurately classified the disease genes compared with LP and LDA, which only uses the genome-phenome association network for disease gene discovery.

6 GO Enrichment Analysis on Gene Clusters

We evaluate the biological significance of the identified gene clusters with GO (Gene Ontology) enrichment analysis by DAVID[**?**], an online functional annotation tool. It is adopted to analyze a set of genes with annotated gene ontologies in "biological process"(BP) branch. In general, a large number of annotations means high-quality modules.

In our work, clusters with more than 300 genes and fewer than 5 genes are dropped as suggested in [**?**]. Table 1 shows that modules identified by CMNMF have the most GO(BP) terms annotations under different p-value cutoff which demonstrates that our method can mine gene functional modules with more biological significance.

Conclusions

In this paper, we introduce a consistent multiple nonnegative matrix factorization for data with hierarchical information. We first analyze the mechanism of our method on a simulated data, then compare the performance of CMNMF with other methods on mining gene modules. We conclude that the CMNMF method is an effective algorithm which can fully utilize the hierarchical structure information behind the data. Experiments on mining gene functional modules show the ability of CMNMF to identify modules with biological significance. In future, we will try to analyze the expansibility of CMNMF on more than just two levels and try to solve other tasks where data has a hierarchical structure.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
This work is supported by the National Natural Science Foundation of China (No. 61300166 and No. 61105049), the Open Project Foundation of Information Technology Research Base of Civil Aviation Administration of China (CAAC-ITRB-201303 and CAAC-ITRB-201408), the Natural Science Foundation of Tianjin (No. 14JCQNJC00600), and the Science and Technology Planning Project of Tianjin (No. 13ZCZDGX01098).

Figures

Figure 5 Sample figure title. A short description of the figure content should go here.

Figure 6 Sample figure title. Figure legend text.

Tables

Table 2 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.