RESEARCH

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

YaoGong Zhang[1,2], YingJie Xu[1,2], Xin Fan[1,2], YuXiang Hong[1,2], ZhiCheng He[1,2], YaLou Huang[1,2] and MaoQiang Xie[1,2*]

[*] Correspondence: xiemq@nankai.edu.cn
[1] College of Software, NanKai University, 300350 TianJin, China
Full list of author information is available at the end of the article

Abstract

Background: An increasing amount of genome-phenome association data, has provided us a great chance to globally explore the underlying genetic mechanisms and understand the regularization between genes and diseases with a deep sight perspective. Gene modules mining, which reveal the interactions between genes and help researchers to identify candidate genes as drug targets, have always been a significant and valuable problem by mining genome-phenome association data. Nevertheless, the hierarchical structure of phenotype ontology has been rarely leveraged by previous gene clustering studies. The relationships between genes and disease phenotypes 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 hierarchical character of phenotype ontology to gain understanding of biological system.

Results: We propose a novel method, Consistent Multiple Nonnegative Matrix Factorization (CMNMF), to factorize genome-phenome association data with the help of hierarchical structure information of phenotype ontology, which can cluster genes well for further mining gene functional modules. The CMNMF constrains the gene cluster matrix to remain consistent while it interacts with different phenotype ontology levels in decomposing the genome-phenome association matrix, meanwhile it restricts the similarity of adjacent phenotype ontology pairs to satisfy the hierarchical structure. CMNMF and seven baseline methods are conducted on mouse gene-phenotype associations. The performance of clustering and GO enrichment analysis show the effectiveness of CMNMF. Additionally, a preliminary supervised CMNMF is proposed for showing its generalizing ability.

Conclusions: Our work show that the proposed CMNMF can identify gene functional modules with biological significance over conventional methods. CMNMF and its supervised version provides a new perspective to mining gene functional modules and disease genes.

Availability: `https://github.com/nkiip/CMNMF`

Keywords: NMF; Gene modules mining; Hierarchical information

Background

Studies on genome-phenome association data have been a key research area recently. It is of great significance to explore the interactions between genes and (or) phenotypes for drug development and disease treatment. With the development of technology, biomedical researchers have 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 human disease studies. Thereby it is necessary and essential for researchers to have a deep sight investigation on mouse data. The international database resources such as Mouse Genome Informatics (M-GI) [1], Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] etc. 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 challenge, which has drawn growing attention in the literatures.

It is well known that genes in mammal genomes are usually organized into groups functionally associate with phenotype groups [3, 4]. Mining of functionally related genes modules is a crucial primary 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.

Phenotype ontology was created to serve as a standardized vocabulary of phenotypic abnormalities that have been seen in diseases [5]. 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 gene module clustering. The traditional way to cluster gene modules is based on network clustering algorithms. Ahn [6] constructed communities as groups of links rather than nodes, link communities naturally incorporate overlap while revealing hierarchical organization. Gopalan [7] develop a scalable approach that is based on a Bayesian model of networks, it allows nodes to participate in multiple communities to community detection. Didier [8] assessed aggregation, consensus and multiplex-modularity approaches to detect communities from multiple network sources, it showed that taking into account the multiplexity of biological networks could lead to better-defined functional modules.

Another natural way to cluster genes is the NMF based method. Compared to network based clustering algorithms, NMF based methods can express the nodes in a low latent space, which provide us a good opportunity to add more specific constraints between homogenous or heterogenous nodes. Some procedures relating to gene expression profiles (such as gene expression, miRNA expression, and copy number variation etc.), were seeking to cluster genes through NMF based methods. Zhang [9] focused on integrating multiple type genomic data to identify microRNA-gene regulatory modules for cancers by sharing a common space of biological sets. Liu [10] proposed Hessian regularization based NMF (HR-NMF) algorithm to gene expression data clustering task. Zhang [11] proposed a NMF-based constrained clustering framework which enforced the similarity of must-link and cannot-link and applied it to deal with clustering of gene expression data. Wang [12] extended NMF to a joint version (jNMF), which factorized multiple transcriptomics data matrices into one common submatrix plus multiple individual submatrices to identify differentially expressed genes. Additionally, Hwang [13] added penalty and regularization terms to keep the final results consistent with clusters obtained from prior knowledge on the disease phenotype similarity network.

Since some data contains specific structures behind it, some structure based methods are proposed and all achieve good results. In order to explore and model the structure correlations among users and items, Wang [14] designed a hierarchical group matrix factorization (HGMF) method for item recommendation. Mashhoori [15] found that items can be treated as different groups and then incorporated hierarchical information between groups into matrix factorization for collaborative filtering. Shan [16] focuses on predicting missing traits for plants which incorporates hierarchical phylogenetic information into matrix factorization. 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 Nonnegative 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 adjacent phenotype ontology pairs, which has parent-child relationships, should have a high similarity score in the hierarchical structure. To demonstrate the approach, we conduct our proposed method on a sampled data first, and series of measurements with other baselines on gene clustering task. The performance of our CMNMF on F_1, $measure$, $Rand\ Index$, $Jaccard\ Index$, and M_{Sim} outperform other baselines, including Kmeans, Kernel-Kmeans [17], HAC [18], ColNMF [19], NMF [20], LDA [21]. Additionally, we extend CMNMF with gene pathways prior which we call Supervised CMNMF (S-CMNMF) on multi-label gene signal pathway classification task, the AUC score demonstrates the advantages of S-CMNMF over LP [22], supervised NMF and supervised ColNMF.

Method

The notations and definitions used in the article are specified in Table 1. We denote $\boldsymbol{A}_{(n \times m)}$ as a binary matrix for storing gene-phenotype associations by n genes and m phenotypes, where $\boldsymbol{A}_{ij}$ is set to 1 for a known association and 0 otherwise. The goal of factorizing matrix $\boldsymbol{A}$ is to derive gene functional clusters $\boldsymbol{G}_{(n \times k)}$ based on gene-phenotype associations. The loss of it can be defined as $min_{\boldsymbol{P},\boldsymbol{G}} ||\boldsymbol{A} - \boldsymbol{G}\boldsymbol{P}||_F^2$, where $\boldsymbol{G}$ and $\boldsymbol{P}$ denotes gene clusters and phenotype clusters, respectively.

However, the NMF in this form cannot make full use of the hierarchical mapping information of phenotype ontology. To address this problem, we design a loss function with two components. The first component is used for penalizing the inconsistence of factorizations on gene-phenotype associations from different phenotype levels (See Figure 1(a)). In other words, the gene clustering results on parent and child phenotype ontology levels should be consistent (See Figure 1(b)). The second component is a hierarchical mapping constraint with phenotype ontologies from parent level and child level, as shown in Figure 1(c). With it, the clusters of phenotype ontologies from different levels should fit the hierarchical architecture. By

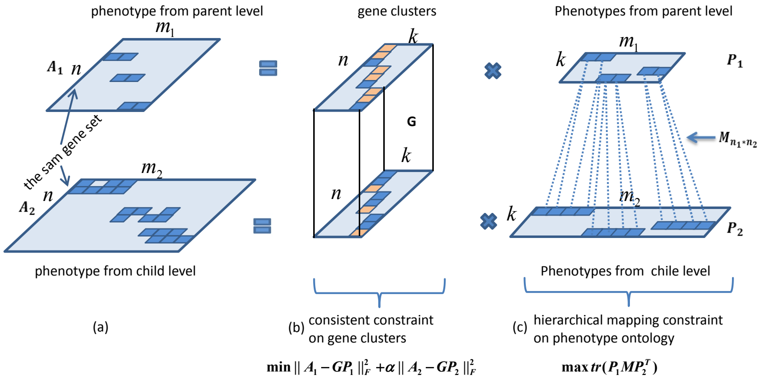

phenotype from parent level
m_1
A_1 n
the sam gene set
m_2
n
A_2
phenotype from child level
gene clusters
k
n
G
k
n
Phenotypes from parent level
k
m_1
P_1
$M_{n_1*n_2}$
m_2
k
P_2
Phenotypes from chile level
(a)
(b) consistent constraint on gene clusters
(c) hierarchical mapping constraint on phenotype ontology
$\min \| A_1 - GP_1 \|_F^2 + \alpha \| A_2 - GP_2 \|_F^2$
$\max tr(P_1 M P_2^T)$

Figure 1: Illustration of the CMNMF framework. (a) The gene-phenotype associations are separated to 2 matrices by the level of phenotype ontologies. These 2 matrices share the same gene set as their first dimensions. (b) Consistent constraint on factorized gene clusters. (c) Hierarchical mapping constraints on the phenotype ontologies from parent and child levels.

Table 1: Summary of notations.

Notations	Explanations
$\boldsymbol{A}$	Genome-phenome association matrix with phenotype from parent and child level
$\boldsymbol{A}_1$	Genome-phenome association matrix with phenotype ontology from parent level
$\boldsymbol{A}_2$	Genome-phenome association matrix with phenotype ontology from child level
$\boldsymbol{G}$	Gene cluster membership
$\boldsymbol{P}$	Phenotype cluster membership
$\boldsymbol{G}_0$	Annotated gene cluster membership
$\boldsymbol{P}_1$	Phenotype cluster membership of parent level
$\boldsymbol{P}_2$	Phenotype cluster membership of child level
$\boldsymbol{M}$	Phenotype ontologies relationship
n	Number of disease genes
m	Number of phenotype ontology in parent and child level
k	Number of latent clusters(e.g. classes)
m_1	Number of phenotype ontology in parent level
m_2	Number of phenotype ontology in child level

optimizing these two components, a CMNMF (Consistent Multiple Non-negative Matrix Factorization) algorithm is proposed to obtain consistent gene clusters by factorizing multiple related gene-phenotype association matrices.

Loss Functions for Penalizing Inconsistence

Motivated by above consistent assumption, we extract phenotype ontologies from two adjacent levels, and two gene-phenotype association matrices $(\boldsymbol{A}_1)_{n\times m_1}$ and $(\boldsymbol{A}_2)_{n\times m_2}$ are extracted from gene-phenotype text file. We assume that the factorizations on $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ 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 $\boldsymbol{G}_{n\times k}$ to achieve this goal. The representation of the data can be derived by optimizing the following quadratic objective function:

$$L_C = \min_{\boldsymbol{G},\boldsymbol{P}_1,\boldsymbol{P}_2} \|\boldsymbol{A}_1 - \boldsymbol{G}\boldsymbol{P}_1\|_F^2 + \alpha \|\boldsymbol{A}_2 - \boldsymbol{G}\boldsymbol{P}_2\|_F^2 , \qquad (1)$$

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

For the hierarchical mapping constraints on phenotype ontologies, loss function in Equation (2) is used to encourage the interactions between phenotypes from parent level and child level.

$$L_H = \sum_{ij} \boldsymbol{M}_{ij} (\boldsymbol{P}_1^{(i)})^T \boldsymbol{P}_2^{(j)} = tr(\boldsymbol{P}_1 \boldsymbol{M} \boldsymbol{P}_2^T), \tag{2}$$

where $\boldsymbol{M}_{m_1 \times m_2}$ denotes the hierarchical mapping relation matrix between phenotype ontologies from adjacent levels. $\boldsymbol{M}_{ij} = 1$, if there is a parent-child relationship between phenotype i and phenotype j, otherwise 0. We enforce hierarchical mapping constraints by maximizing the mapping between the phenotype ontologies in gene-phenotype network $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

By combining these two components, the loss function can be given as follows:

$$\begin{aligned} L &= \|\boldsymbol{A}_1 - \boldsymbol{G}\boldsymbol{P}_1\|_F^2 + \alpha \|\boldsymbol{A}_2 - \boldsymbol{G}\boldsymbol{P}_2\|_F^2 - \beta tr(\boldsymbol{P}_1 \boldsymbol{M} \boldsymbol{P}_2^T) \\ &\text{s.t.} \quad \sum_j \boldsymbol{G}_{ij} = 1, \quad \sum_i (\boldsymbol{P}_1)_{ij} = 1, \quad \sum_i (\boldsymbol{P}_2)_{ij} = 1 \end{aligned} \tag{3}$$

where α, β are parameters to balance the tradeoff between the two components.

The CMNMF Algorithm

To minimize the loss function in Equation (3), an alternative iterative schema is adopted for proposed CMNMF algorithm. It solves the problem with respect to one variable while fixing the other variables alternatively. As in the original NMF, the loss function in Equation (3) is not convex in $\boldsymbol{G}$, $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ jointly, but it is convex in $\boldsymbol{G}$ for fixed $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, vice versa. In this subsection, the computations of $\boldsymbol{G}$, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are given, the CMNMF algorithm is outlined in **Algorithm 1**.

Computation of G in CMNMF

We fix variables $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, the partial derivative of Equation (3) with respect to $\boldsymbol{G}$ is:

$$\frac{\partial L}{\partial \boldsymbol{G}} = -2(\boldsymbol{A}_1 \boldsymbol{P}_1^T - \boldsymbol{G}\boldsymbol{P}_1\boldsymbol{P}_1^T) - 2\alpha(\boldsymbol{A}_2\boldsymbol{P}_2^T - \boldsymbol{G}\boldsymbol{P}_2\boldsymbol{P}_2^T)$$

the multiplicative update rule is:

$$\boldsymbol{G}_{ij} \leftarrow \boldsymbol{G}_{ij} \frac{(\boldsymbol{A}_1\boldsymbol{P}_1^T + \alpha \boldsymbol{A}_2 \boldsymbol{P}_2^T)_{ij}}{(\boldsymbol{G}\boldsymbol{P}_1\boldsymbol{P}_1^T + \alpha \boldsymbol{G}\boldsymbol{P}_2\boldsymbol{P}_2^T)_{ij}}$$

To satisfy the equality constraint, we normalize $\boldsymbol{G}_{ij}$ as $\boldsymbol{G}_{ij} \leftarrow \frac{\boldsymbol{G}_{ij}}{\sum_j \boldsymbol{G}_{ij}}$.

Computation of P_1 and P_2 in CMNMF

When $\boldsymbol{G}$ is computed, the partial derivative of Equation (3) with respect to $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are:

$$\begin{aligned} \frac{\partial L(\boldsymbol{P}_1)}{\partial \boldsymbol{P}_1} &= -2(\boldsymbol{G}^T\boldsymbol{A}_1 - \boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_1) - \beta \boldsymbol{P}_2 \boldsymbol{M}^T \\ \frac{\partial L(\boldsymbol{P}_2)}{\partial \boldsymbol{P}_2} &= -2\alpha(\boldsymbol{G}^T\boldsymbol{A}_2 - \boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_2) - \beta \boldsymbol{P}_1 \boldsymbol{M} \end{aligned}$$

Algorithm 1 CMNMF

```
Input: gene-phenotype association matrix $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, number of cluster dimensions $K$, and parameter $\alpha$,$\beta$,$\gamma$,$\lambda_1$,$\lambda_2$,
Output: model parameters $\boldsymbol{G}, \boldsymbol{P}_1, \boldsymbol{P}_2$
$\boldsymbol{G}, \boldsymbol{P}_1, \boldsymbol{P}_2 \leftarrow$ random values
repeat
    Update $\boldsymbol{G}_{ij} \leftarrow \boldsymbol{G}_{ij} \frac{(\boldsymbol{A}_1\boldsymbol{P}_1^T + \alpha\boldsymbol{A}_2\boldsymbol{P}_2^T)_{ij}}{(\boldsymbol{G}\boldsymbol{P}_1\boldsymbol{P}_1^T + \alpha\boldsymbol{G}\boldsymbol{P}_2\boldsymbol{P}_2^T)_{ij}}$
    Normalize $\boldsymbol{G}_{ij} \leftarrow \frac{\boldsymbol{G}_{ij}}{\sum_j \boldsymbol{G}_{ij}}$
    Update $(\boldsymbol{P}_1)_{ij} \leftarrow (\boldsymbol{P}_1)_{ij} \frac{(\boldsymbol{G}^T\boldsymbol{A}_1 + \frac{1}{2}\gamma\boldsymbol{P}_2\boldsymbol{M}^T)_{ij}}{(\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_1)_{ij}}$, $(\boldsymbol{P}_2)_{ij} \leftarrow (\boldsymbol{P}_2)_{ij} \frac{(\alpha\boldsymbol{G}^T\boldsymbol{A}_2 + \frac{1}{2}\gamma\boldsymbol{P}_1\boldsymbol{M})_{ij}}{(\alpha\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_2)_{ij}}$
    Normalize $(\boldsymbol{P}_1)_{ij} \leftarrow \frac{(\boldsymbol{P}_1)_{ij}}{\sum_i (\boldsymbol{P}_1)_{ij}}$,   $(\boldsymbol{P}_2)_{ij} \leftarrow \frac{(\boldsymbol{P}_2)_{ij}}{\sum_i (\boldsymbol{P}_2)_{ij}}$
until convergence
return $\boldsymbol{G}, \boldsymbol{P}_1, \boldsymbol{P}_2$
```

the multiplicative update rule is (note that when we compute $\boldsymbol{P}_1$, we take $\boldsymbol{P}_2$ fixed, vise versa):

$$
\begin{aligned}
(\boldsymbol{P}_1)_{ij} &\leftarrow (\boldsymbol{P}_1)_{ij} \frac{(\boldsymbol{G}^T\boldsymbol{A}_1 + \frac{1}{2}\beta\boldsymbol{P}_2\boldsymbol{M}^T)_{ij}}{(\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_1)_{ij}} \\
(\boldsymbol{P}_2)_{ij} &\leftarrow (\boldsymbol{P}_2)_{ij} \frac{(\alpha\boldsymbol{G}^T\boldsymbol{A}_2 + \frac{1}{2}\beta\boldsymbol{P}_1\boldsymbol{M})_{ij}}{(\alpha\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{P}_2)_{ij}}
\end{aligned}
$$

To satisfy the equality constraint, we normalize $(\boldsymbol{P}_1)_{ij}$ as $(\boldsymbol{P}_1)_{ij} \leftarrow \frac{(\boldsymbol{P}_1)_{ij}}{\sum_i (\boldsymbol{P}_1)_{ij}}$ and $(\boldsymbol{P}_2)_{ij}$ as $(\boldsymbol{P}_2)_{ij} \leftarrow \frac{(\boldsymbol{P}_2)_{ij}}{\sum_i (\boldsymbol{P}_2)_{ij}}$.

Results and Discussion

We first demonstrated proposed CMNMF and NMF on sampled MGI mouse gene-phenotype association matrix. We then executed CMNMF and seven baseline methods on the whole association matrix for comparing the performance of mining mouse gene functional modules. A statistical analysis of clustering result shows the improvement by using the hierarchical information of phenotype ontologies. Moreover, a preliminary supervised CMNMF is proposed for showing the generalization of proposed model. Finally, gene ontology enrichment analysis is conducted to evaluate the biological significance of mined gene modules.

Data Preparation and Evaluation

The mouse gene-phenotype ontology associations are extracted from file "MGI_Geno_Disease.rpt" (December-2015) downloaded from MGI [1], in which we have 7029 gene-phenotype associations at level 4 and 8588 associations at level 5. 2342 ontology terms at level 4 (parent level) and 3257 ontology terms at level 5 (child level) are selected from the 11471 MP terms in the "MPheno_OBO.ontology" (November-2015) [2]. 3583 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 phenotypes, which neither has relationships with any disease genes in gene-phenotype association matrix nor has relationships with any phenotypes in parent or child levels in hierarchial mapping relations, then we generate a dataset containing 1354 disease genes and 5245 phenotypes.

To evaluate the clustering performance for CMNMF, we crawled 280 mouse gene pathways from KEGG[3] and selected 225 of them as the "ground truth" of gene modules. We intersect each pathway with genes in gene-phenotype association matrix, if the common genes for this pathway are less than 3, we hold this pathway out. At last, we get 225 pathways in total. Genes in a signal pathway can be considered as a gene functional module (or a gene cluster).

In the leave-one-out cross-validation experiment for S-CMNMF, after removing genes not present in the gene pathways, and choosing the phenotypes that has relationships with selected genes and the phenotypes that have parent-child relationships, we generated a dataset with 760 genes and 5245 phenotypes.

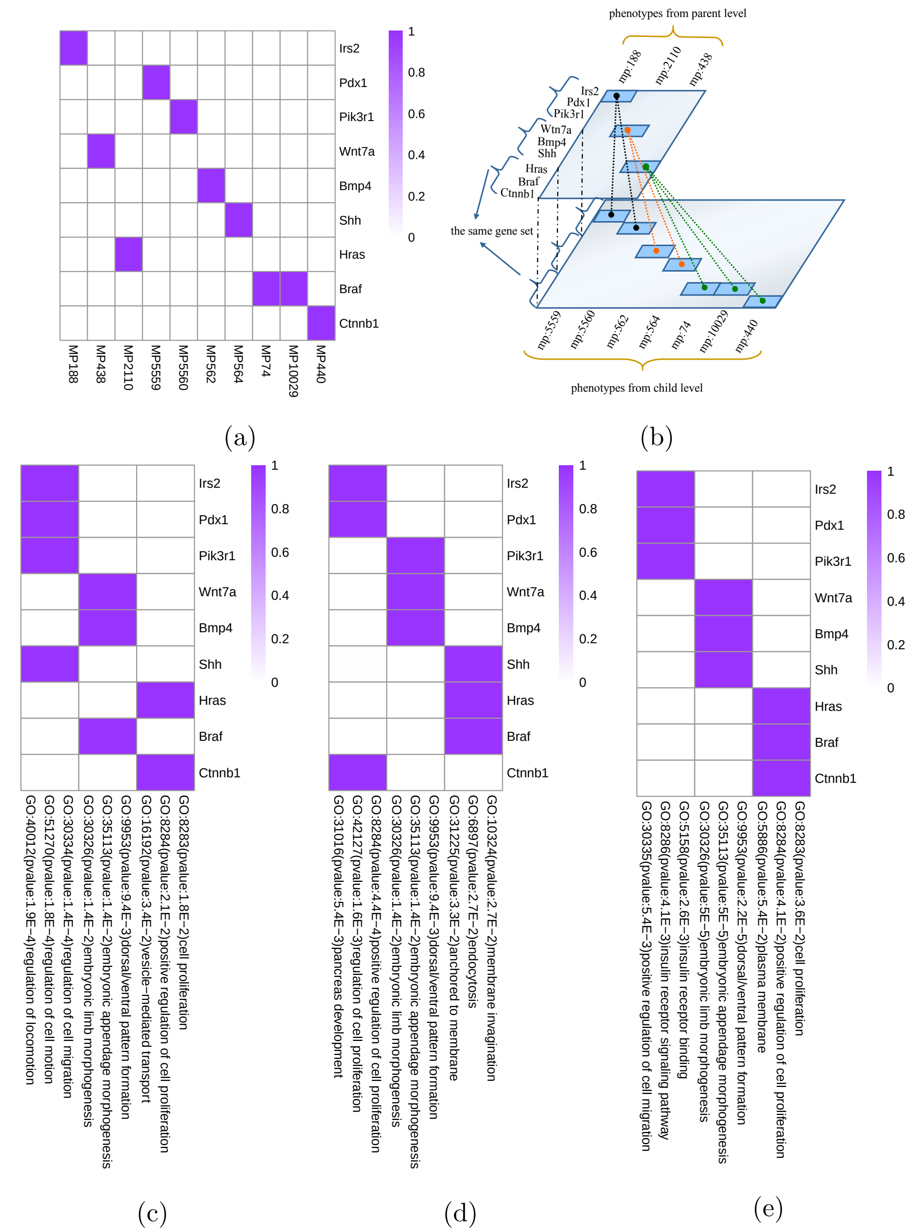
CMNMF on Sampled Data

To illustrate the effect of the consistent constraint and hierarchical mapping constraint, we demonstrate proposed CMNMF, NMF and ColNMF on a small sampled gene-phenotype association matrix from MGI dataset. The overview of sampled data matrix is shown in Figure2 (a), in which 9 genes from 3 pathways are selected. Specifically, Irs2, Pdx1, Pik3r1 are from pathway MMU7930 (Type-II diabetes mellitus). Hras, Braf, Ctnnb1 are from MMU5212 (Prostate cancer). Wnt7a, Bmp4, Shh are from MMU5217 (Basal cell carcinoma). The hierarchical relations between 9 selected genes' associated phenotype ontologies are shown in Figure2 (b). We expect that the genes in one cluster tend to be the members of one gene signal pathway with the help of two constraints.

The clustering results from factorization by NMF, ColNMF and CMNMF are presented in Figure2 (c-e), respectively. Note that NMF with the consistent constraint (*i.e.* ColNMF), which is a special case with $\gamma = 0$ in our CMNMF, can improve the clustering result significantly. 3 pairs are right. For CMNMF, 9 pairs are right. is better than original NMF on combined association matrix.

Figure2 (b) illustrates the model we used in CMNMF, the upper matrix means the associations between genes and phenotypes from parent level, the lower matrix means the associations between genes and phenotypes from child level, the different color dash lines means parent-child phenotype relationships in different groups. To be more specific, we sampled the common nine genes from three pathways (or groups) (For each group, there are two genes associating with phenotypes in the child level (like in Figure 2(b), Pdx1 and Pik3r1, from pathway mmu7930, are interacting with MP5559, MP5560, which are in the child level of the phenotype ontology hierarchical tree), and another gene associating with the parent phenotypes (Irs2, from pathway mmu7930, is interacting with MP188, which is in the parent level). The task is to cluster the three genes, which are coming from the same pathway, into a same group.

We apply NMF, ColNMF and CMNMF to the sampled data, show the results in Figure 2(c), (d) and (e) respectively. Figure 2(c) tells that NMF method gives a random clustering results, Figure 2(d) tells the fact that we use a common basis matrix G while factorizing A_1 and A_2 simultaneously, we can see that in each gene group module, there exists misclassified genes (such like, Ctnnb1 shouldn't be grouped with Irs2 and Pdx1, it's supposed to classified with Hras and Braf). After we adding the hierarchical information to our model, just as Figure 2(e) shows,



(a) (b)

(c) (d) (e)

Figure 2: Illustration of sampled data, rows stand for genes, columns stand for phenotype. If there is a association between a gene and a phenotype, the corresponding position is colored, otherwise it's uncolored. (a) a overview of association matrix of sampled data; (b) Illustration of hierarchical information of phenotype ontology we used in our model; (c) Clustering result of NMF; (d) Clustering result of ColNMF(γ=1); (e) Clustering result of CMNMF. In (c), (d) and (e), each gene group is characterized with three most significant GO terms at the bottom.

all the genes are classified into three groups perfectly. It demonstrates that the two constraints working together can get the best performance. The reasons for

Table 2: Performance results of different methods on F_1, *Jaccard Index*, *Rand Index*, $\mathcal{M}_{sim}$

	F_1 measure	Jaccard Index	Rand Index	$\mathcal{M}_{sim}$
HAC	0.0738	0.0383	0.9688	1.5523
Kmeans	0.0536	0.0290	0.9683	1.4207
Kernel-Kmeans	0.0594	0.0306	0.9677	1.3514
LDA	0.0978	0.0525	0.9618	1.1609
NMF	0.093	0.0488	**0.9740**	1.3048
HMF	0.0904	0.0474	0.9686	0.8982
ColNMF (CMNMF with $\gamma = 0$)	0.0942	0.0495	0.9738	1.3479
CMNMF ($\alpha = 1$ and $\gamma = 10$)	**0.1190**	**0.0585**	0.9701	**1.6453**

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

Comparison with Baseline Methods by Mining Gene Modules

CMNMF is compared to HAC (Hierarchical Agglomerative Clustering) [18], K-Means, Kernel-Kmeans [17], LDA [21], NMF [20], HMF (Hierarchical Matrix Factorization) [15] and ColNMF (Collective NMF) [19] (in fact, CMNMF degrades to ColNMF when $\gamma = 0$) by clustering genes. To be fair, all matrix factorization-based methods are implemented with sparse and non-negative constraints.

For CMNMF, HMF and ColNMF, we need to separate gene-phenotype assocations matrix to two matrices according to the levels of phenotype ontologies for CMNMF, HMF and ColNMF (Only phenotype ontologies in level 4 and 5 are adopted in the experiment). The combined matrix is used for the other baselines. The details of parameter tuning for baselines are given in *Supporting Information (SI) Parameter Tuning*.

For methods LDA, NMF, HMF, ColNMF and CMNMF, the gene clustering results are row-normalized by using z-score. For each gene, it can belong to multiple clusters, as it can participate in a couple of pathways, so we take the following strategy to predict the clusters a gene belongs to: first, for cluster distributions of a gene, we sort the values in a descending order; then we choose the top k clusters, of which the sum of the values is closest to 0.9, as the modules this gene belongs to.

The $\mathcal{M}_{sim}$ [23], an internal clustering evaluation measure (*see SI* $\mathcal{M}_{sim}$), is used for evaluating the performance of clustering genes by factorizing gene-phenotype associations matrix. The higher the $\mathcal{M}_{sim}$, the better the performance. For $\mathcal{M}_{sim}$, genes are featured with annotated GO terms, and $FS_{simGIC}(G_k, G_l)$ [24] is used to measure the similarity between gene G_k and G_l by GO terms. In other words, genes are clustered well if the GO annotations of genes in a cluster are similar.

F_1 measure, *Jaccard Index*, and *Rand index* (see *SI Measurement Definition*) are selected to validate whether the clustering results can make good biological sense. 225 mouse gene pathways from KEGG are selected as test criteria although there is unconformity between MGI gene-phenotype associations and KEGG pathways. For these three external evaluation measurements, the higher values, the better results.

The results are reported in Table 2, which are produced by the best parameters of each method and we take the average value of 10 times of experiments for each method. Clearly, CMNMF performs significantly better than all the other methods at significance level 0.01 with F_1 measure, $Jaccard\ Index$, and $\mathcal{M}_{sim}$ (see *SI Significant Analysis*). In particular, the consistent constraint works well by comparing convenient NMF and ColNMF. Moreover, CMNMF ($\alpha = 1$ and $\gamma = 10$) beats ColNMF (*i.e.* CMNMF with $\gamma = 0$) because of the adoption of hierarchical mapping constraint. However, NMF outperforms other methods with $Rand\ Index$ measure. We notice that true negative gene pairs are used when $Rand\ Index$ value is calculated, that is to say $Rand\ Index$ considers the number of gene pairs that do not appear in any pathways, which is not the situation we focus on.

Parameter Tuning

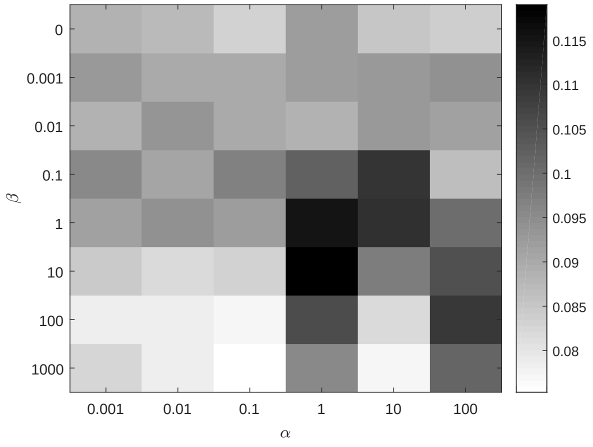


Figure 3: F_1 score under different α and β combination

In this part, we give discussions on how parameters α and β affect the clustering results. We choose F_1 measure as a criteria to tune parameters, because F_1 measure is a harmonic average of precision and recall, it can avoid grouping genes into a single cluster that behaves differently than expected.

The α in our CMNMF model, balances the contributions from different levels is an important parameter. When α is close to 0, the model degrades into NMF on level 4 and when α is big enough information from level 5 dominates the model. The β controls the effects of hierarchical information of phenotype ontologies, when β is set to 0, the models does not take hierarchical information into consideration, in this situation, our model degrades to ColNMF model, the higher β value, the more influence of hierarchical information has. The performance of CMNMF with different α and β combination is shown in Figure 3. We search α in {0.001, 0.01, 0.1, 1, 10, 100, 1000} and β in {0, 0.001, 0.01, 0.1, 1, 10, 100}, darker colour stands the bigger F_1 score under the corresponding α and β combination.

It can be seen from Figure 3, the colour is getting darker with β growing under almost every α value and the F_1 scores peak at $\beta = 10$, which demonstrates the hierarchical information has general positive effects. After that, as β increasing, F_1 score decreases for most α. In our model, the best performance of F_1 measure is achieved when $\alpha = 1$ and $\beta = 10$. We also give other baselines' parameter tuning processes in *SI Parameter Tuning*.

Biological Analysis on Gene Modules

We used the DAVID [25] to investigate the enrichment of functional annotations of genes selected in each gene modules found by our model CMNMF. DAVID[4] starts by reading the input file that contains a list of genes, and estimates the statistical significance of the enrichment of GO terms. Table 3 shows the enrichment of GO terms. We present some significant biological processes for certain gene modules.

Table 3: Enrichment of GO categories in gene modules selected by CMNMF

No	Genes in Module	Most Related GO Terms	P-value
204	GCK, PDX1, INS2, E2F1, HNF1A,TNF, LEP, RPS6KA3, AKT2, SLC5A1, *etc*	response to organic substance; glucose metabolic process; positive regulation of macromolecule-metabolic process;	1.5E-08 9.5E-07 1.6E-06

201	AR, NFKBIA, CREBBP, CEBPA, DCC,VHL, SOS1, FLT3, FASL, *etc*	hemopoietic or lymphoid organ development;
		immune system development;
		hemopoiesis;
			3.3E-17
			7.5E-17
			1.5E-16

138	GCK, SLC2A4, IRS1, INS2, IRS2, LDLR, LIPA, PCSK1, POMC, *etc*	response to insulin stimulus; response to hormone stimulus; response to endogenous stimulus;	2.8E-10 7.2E-10 1.9E-09

127	CDK4, ERCC1, G6PC, GAD2,GHR, GPI1, PDX1,MAFA, *etc*	glucose metabolic process;
		hexose metabolic process;
		monosaccharide metabolic process;
			1.5E-07
			5.4E-07
			1.2E-06

41	FCGR2B, C1QA, C4B, TROVE2, LYN, POLB, WT1,LAMB2, *etc*	immune effector process;
		negative regulation of immune system process;
		immune response;
			7.5E-06
			2.2E-05
			5.9E-05

13	MSH2,TSC2,PRKAR1A, VHI,CDKN2A,KRAS, FHIT,TRP53, *etc*	regulation of cell proliferation; negative regulation of cell proliferation; regulation of apoptosis;	2.9E-07 4.3E-07 5.9E-06

The first column is the index of gene modules got by CMNMF, for each gene module, we present some genes in the second column, and we put three most related GO terms and their corresponding P-values in the third and fourth column respectively.

It is clear that in Table 3, gene modules found by CMNMF are statistical significant in terms of certain common GO terms (like: regulation of cell proliferation, immune response, et al.). Besides, we explore the relationships between genes in the modules and diseases in literature to demonstrate the biological meaning of modules found by CMNMF.

In gene module No.204, a number of genes are validated to be associated with diabetes. Haeusler [26] found that, in diabetic subjects with HbA1c > 7.0, gluconeogenic enzymes were expressed normally, but GCK was suppressed more than 60%. Moreover, HbA1c and fasting glucose were negatively correlated with GCK, but showed no correlation with G6PC, PCK1, or PCK2. Through the studies of [27], Munich INS2 (C95S) mutant mice are considered a valuable model to study the mechanisms of beta-cell dysfunction and death during the development of diabetes. In the study of [28], which revealed a significant correlation between TNF-α levels and BMI (p-value=0.006), the correlation being stronger in males when compared to females. A significant correlation was found between per cent β cell function and

TNF-α (p-value=0.008). TNF-α correlated significantly with HOMA IR, HOMA B and insulin, in type 2 diabetes.

In gene module No.201, we have found genes in this module have a strong relation to prostate cancer. To be more specific, Han [29] found that with the exception of NFKBIA 3' UTR polymorphism, the heterozygous and mutant genotypes of the other polymorphisms were significantly associated with prostate cancer risk, polymorphisms in NFKB1 and NFKBIA genes may modulate the risk of developing prostate cancer. Ding [30] developed functional evidence for CBP (Crebbp) and PTEN interaction in prostate cancer based on findings of their correlate expression in the human disease. Timofeeva [31] established primary prostate cancer epithelial cells from 14 AA and 13 EA men, to identify cancer-specific gene expression patterns in AA men. Tan [32] gave An overview of AR structure and activity, its actions in prostate cancer, and how structural information and high-throughput screening have been or can be used for drug discovery are provided herein.

We can observe the similar phenomena in other gene modules, it could guide the researchers to study the relations between gene modules and some certain diseases. Our model CMNMF is likely to inspire researchers to discover some underlying patterns behind the sophisticated biological system.

Gene-Phenotype Association Prediction with Supervised CMNMF

When we consider adding the gene pathways as a prior, the supervised CMNMF can be used for multi-label gene signal pathway classification prediction. In our experiments, we collected 225 mouse pathways from KEGG in total. We choose 21 common KEGG disease pathways (*e.g.* alzheimer, diabetes, prostate cancer, *et al.*) on mouse, there are 145 member genes in these 21 KEGG disease pathways. In the leave-one-out cross validation, each of the 145 member gene was held out in all pathways and then classified into the 225 pathways as a multi-label classification problem, the other 144 member genes was used as the initialization of to classify the held-out gene. The higher the target pathways in the ranking of the 145 member

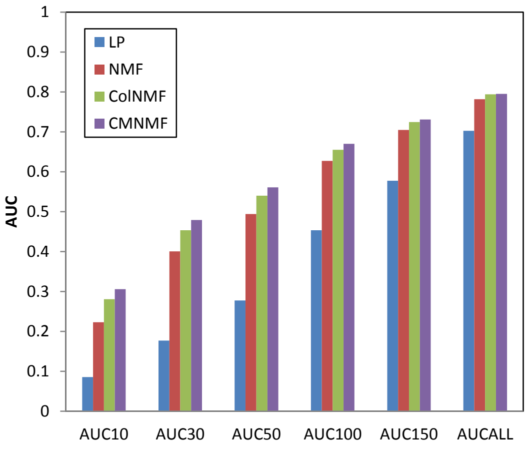


Figure 4: AUC score of different method

genes, the better the performance. We measured the performance by the AUC (*see SI AUC*). The baselines we use here is LP, NMF and ColNMF. LP was applied to the gene similarity network, in which the data was measured by genome-phenome association matrix. NMF and ColNMF are the methods we have discussed above with gene pathways as a prior. The average AUC score in the leave-one-out cross validation across the member genes by all methods are reported in Figure 4. The results clearly show that our methods CMNMF takes advantage of LP, NMF and ColNMF. The Optimization framework of Supervised CMNMF can be found in *SI Supervised CMNMF*.

Conclusions

In this paper, we introduce a novel method, Consistent Multiple Nonnegative Matrix Factorization (CMNMF) based on genome-phenome association data factorization. Our approach was applied to an integration of hierarchical information of phenotype ontology, as far as we know, which has been rarely used in previous studies. The modules found by our method express notable biological meaning.

There are several merits of CMNMF. First, CMNMF keeps the gene cluster matrix consistent while it interacts with different phenotype ontology levels in decomposing data matrix; on the other hand, it restricts the similarity of phenotype pairs expressed in phenotype cluster matrix to satisfy the hierarchical structure. To achieve a better clustering results, we introduce sparse regularization on gene and phenotype decomposing matrix. The experiments result on our proposed method and other baselines (including HAC, Kmeans, kernel-Kmeans, LDA, NMF, HMF, CMNMF) show the effectiveness of our method. Furthermore, we give a more detailed discussion on the gene modules found by our method in biological aspects. Besides, we conduct a supervised CMNMF for gene signal pathway prediction, the preliminary comparison of our model with LP, NMF and ColNMF demonstrates the good extension of CMNMF to gene classification problem.

Although we have a trial on integration of hierarchical information of phenotype ontology on mouse for gene module discovery problem, there still exists a number of challenging issues to be solved. Such like: 1) in this work, we didn't give much talk on how to use the phenotype clustering results, because neither there has a clear definition of relationship between phenotypes, nor we have found a meaningful classification of phenotypes as a prior that we can utilize in our model; 2) at the moment, we are doing this research on mouse data, because there are much common characteristics between human and mouse, a key issue in the future is how to align the data between two species or even more in a framework to explore the understanding patterns of human biological mechanism; 3) as we know, gene clustering can also be conducted on a variety of gene expression data, here comes a fascinating question, is there possible to find an effective way to cluster genes by integrating genome-phenome data and expression data? Our work is just a snapshot of the biology system, we provide all our codes on Github web site for interested researchers for future further investigations.

Endnotes

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

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

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

[4] https://david.ncifcrf.gov/summary.jsp

Declarations

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. 14JCQNJC00600), and the Science and Technology Planning Project of Tianjin (No. 13ZCZDGX01098).

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

Author's contributions
MQX and YGZ originally design the model. YGZ worked on the method, experiment, analyses, and writing of the manuscript. YJX contributed on method and analyses. XF and YXH contributed on the experiment. ZCH contributed on the method. MQX and YLH contributed on writing of the manuscript. All authors read and approved the final manuscript.

Additional Files
Additional file 1 — Supporting Information (SI)
This supplement file includes some additional detailed illustrative text and tables we have mentioned in our paper. We give clearly explanatory text in Supporting Information, please refer to it for more details.

Author details
[1]College of Software, NanKai University, 300350 TianJin, China. [2]College of Computer and Control Engineering, NanKai University, 300350 TianJin, China.

References

1. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E.: The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic acids research **43**(Database issue), 726–36 (2015)
2. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
3. Xuan H, Li X, Ren S, Z.S.: Modular organization of the human disease genes: a text-based network inference. (2015). http://www.ncbi.nlm.nih.gov/pubmed/26527852 Accessed 2015-11-19
4. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., Brunak, S.: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology **25**(3), 309–16 (2007). doi:10.1038/nbt1295
5. Köhler, S., Doelken, e.a.: The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research **42**(Database issue), 966–74 (2014). doi:10.1093/nar/gkt1026
6. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
7. Gopalan, P.K., Blei, D.M.: Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences of the United States of America **110**(36), 14534–9 (2013)
8. Didier, G., Brun, C., Baudot, A.: Identifying communities from multiplex biological networks. PeerJ **3**, 1525 (2015)
9. Zhang, S., Li, Q., Liu, J., Zhou, X.J.: A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. Bioinformatics **27**(13), 401–409 (2011). doi:10.1093/bioinformatics/btr206
10. Xiao Liu, Jun Shi, Congzhi Wang: Hessian regularization based non-negative matrix factorization for gene expression data clustering. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), vol. 2015, pp. 4130–4133. IEEE, Milan, Italy (2015). http://www.ncbi.nlm.nih.gov/pubmed/26737203
11. Zhang, X., Zong, L., Liu, X., Luo, J.: Constrained Clustering With Nonnegative Matrix Factorization. IEEE Transactions on Neural Networks and Learning Systems **PP**(99), 1–1 (2015)
12. Wang, H.-Q., Zheng, C.-H., Zhao, X.-M.: jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics (Oxford, England) **31**(4), 572–80 (2015)
13. Hwang, T., Atluri, G., Xie, M., Dey, S., Hong, C., Kumar, V., Kuang, R.: Co-clustering phenome-genome for phenotype classification and disease gene discovery. Nucleic Acids Research **40**(19), 1–16 (2012). doi:10.1093/nar/gks615

14. Wang, X., Pan, W., Xu, C.: HGMF: Hierarchical Group Matrix Factorization for Collaborative Recommendation. In: Proceeding CIKM '14 Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 769–778. ACM Press, New York, New York, USA (2014). doi:10.1145/2661829.2662021. http://dl.acm.org/citation.cfm?id=2661829.2662021
15. Ali Mashhoori, S.H.: Incorporating Hierarchical Information Into the Matrix Factorization Model for Collaborative Filtering. Lecture Notes in Computer Science, vol. 7198, pp. 504–531. Springer, Berlin, Heidelberg (2012). http://www.springerlink.com/index/10.1007/978-3-642-28493-9
16. Shan, H., Kattge, J., Reich, P., Banerjee, A., Schrodt, F., Reichstein, M.: Gap Filling in the Plant Kingdom—Trait Prediction Using Hierarchical Probabilistic Matrix Factorization. ICML, 1303–1310 (2012). 1206.6439
17. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04, p. 551. ACM Press, New York, New York, USA (2004). http://dl.acm.org/citation.cfm?id=1014052.1014118
18. Ward Jr., J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58, 236-244. http://dx.doi.org/10.1080/01621459.1963.10500845 - Open Access Library. http://www.oalib.com/references/7992081 Accessed 2015-11-25
19. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08, p. 650. ACM Press, New York, New York, USA (2008). http://dl.acm.org/citation.cfm?id=1401890.1401969
20. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–91 (1999)
21. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
22. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(3), 036106 (2007)
23. Bordino, I., Castillo, C., Donato, D., Gionis, A.: Query similarity by projecting the query-flow graph. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10, p. 515. ACM Press, New York, New York, USA (2010). doi:10.1145/1835449.1835536. http://dl.acm.org/citation.cfm?id=1835449.1835536
24. Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., Xuan, P.: Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics **29**(11), 1424–32 (2013). doi:10.1093/bioinformatics/btt160
25. Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology **4**(5), 3 (2003)
26. Haeusler, R.A., Camastra, S., Astiarraga, B., Nannipieri, M., Anselmino, M., Ferrannini, E.: Decreased expression of hepatic glucokinase in type 2 diabetes. Molecular metabolism **4**(3), 222–6 (2015)
27. Herbach, N., Rathkolb, B., Kemter, E., Pichl, L., Klaften, M., de Angelis, M.H., Halban, P.A., Wolf, E., Aigner, B., Wanke, R.: Dominant-negative effects of a novel mutated Ins2 allele causes early-onset diabetes and severe beta-cell loss in Munich Ins2C95S mutant mice. Diabetes **56**(5), 1268–76 (2007)
28. Swaroop, J.J., Rajarajeswari, D., Naidu, J.N.: Association of TNF-α with insulin resistance in type 2 diabetes mellitus. The Indian journal of medical research **135**, 127–30 (2012)
29. Han, X., Zhang, J.-J., Yao, N., Wang, G., Mei, J., Li, B., Li, C., Wang, Z.-A.: Polymorphisms in NFKB1 and NFKBIA Genes Modulate the Risk of Developing Prostate Cancer among Han Chinese. Medical science monitor : international medical journal of experimental and clinical research **21**, 1707–15 (2015)
30. Ding, L., Chen, S., Liu, P., Pan, Y., Zhong, J., Regan, K.M., Wang, L., Yu, C., Rizzardi, A., Cheng, L., Zhang, J., Schmechel, S.C., Cheville, J.C., Van Deursen, J., Tindall, D.J., Huang, H.: CBP loss cooperates with PTEN haploinsufficiency to drive prostate cancer: implications for epigenetic therapy. Cancer research **74**(7), 2050–61 (2014)
31. Timofeeva, O.A., Zhang, X., Ressom, H.W., Varghese, R.S., Kallakury, B.V.S., Wang, K., Ji, Y., Cheema, A., Jung, M., Brown, M.L., Rhim, J.S., Dritschilo, A.: Enhanced expression of SOS1 is detected in prostate cancer epithelial cells from African-American men. International journal of oncology **35**(4), 751–60 (2009)
32. Tan, M.H.E., Li, J., Xu, H.E., Melcher, K., Yong, E.-l.: Androgen receptor: structure, role in prostate cancer and drug discovery. Acta pharmacologica Sinica **36**(1), 3–23 (2015)