Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Systems biology

Group Based Gene-phenotype Association prediction

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

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

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Clinical diseases are characterized by distinct phenotype ontologies. To identify disease genes is to elucidate the gene-phenotype relationships. Mutations in functionally related genes may result in similar phenotypes. It is reasonable to predict disease-causing genes by integrating phenotypic data and genomic data. Gene-Phenotype association prediction has been an essential biological problem these years, Gene groups (such like: KEGG pathways) give us a priori to constrain genes within groups. However, such gene group information rarely be used for gene-phenotype association prediction problem. As we know a gene group (like KEGG pathways) reveals a biological process in gene levels, all genes in such group participate the biological process. So an intuitive inference is that all genes within a group should be similar. Thereby It is challenging to utilize the consistence of genes within a group to get a better association prediction.

Results: In this paper, we propose a novel method, Group-based Nonnegative Matrix Factorization (GNMF) to utilize the group information of genes and phenotype to in factoring genome-phenome association data. In GNMF, we constrain the genes within a group to keep similar to each other. Through GNMF, we can get a better gene and phenotype representation in latent space. We conduct a statistic test and the result show that group do help the prediction significantly. We evaluate the GNMF algorithm with 10-fold cross-validation on MGI data and OMIM data. In the experiments, the GNMF achieved best overall rankings compared with the baseline.

Availability: The code used in this paper is provided on: https://github.com/nkiip/****.

Contact: xiemq@nankai.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phenotypes denote the observable physical or biological traits of an organism. Understanding the relations between genes and gene functions (or related phenotypes) is one of the main objectives of genetics in the post-genome era [1] [2] [3]. With the advent of OMICS techniques, the number of uncovered gene-phenotype associations has increased significantly over the last several decades. However, the number of genes with identified pheno- types has not been able to reach the genomic scale yet, due to some technical challenges such as the multi-functionality of genes and heterogeneity of diseases [4 6]. At this moment, various types of proteomic and/or genomic data (such as protein-protein interaction (PPI) data [6 12], sequence data [13 14] and function annotations [15 19]) have been used to identify gene-phenotype associations. Previous studies showed that products of different genes tend to physically interact with each other if these genes are involved in causation of similar disorders [20 21]. Similar phenotypes are determined by genes with related functions, too [22]. Researchers used this information to predict phenotypes by the interactome [6 8] or by the topology of the PPI network [7]. Moreover, sequence information, together with function annota- tions, has been used to prioritize candidate gene-phenotype associations. For example, the features of sequence data were used to build a model, which

Table 1. Notations

PART	DESCRIPTION
n	the number of genes
m	the number of phenotypes
K	the dimension of latent space
$\boldsymbol{R}$	Genome-phenome association matrix
$\boldsymbol{U_{n\times k}}$	gene distribution
$\boldsymbol{V}_{k\times m}$	phenotype distribution
$\boldsymbol{X}_{i\cdot}$	the ith row of matrix $\boldsymbol{X}$
$\boldsymbol{X}_{\cdot j}$	the jth column of matrix $\boldsymbol{X}$

was then trained by the function annotations [13 14 23]. Researchers also employed machine learning approaches and function annotations to construct models[16] [24] [14].

2 Methods

2.1 Problem Formulation

2.2 Biological Interpretation

2.3 Group constraint NMF

$$\sum_{ij} \boldsymbol{S}_{ij}||\boldsymbol{U}_{i\cdot} - \boldsymbol{U}_{j\cdot}||^2 = tr\Big(\boldsymbol{U}^T(\boldsymbol{D}-\boldsymbol{S})\boldsymbol{U}\Big) \tag{1}$$

objective function:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}||\boldsymbol{Y}\odot(\boldsymbol{R}-\boldsymbol{U}\boldsymbol{V})||^2 + \frac{\lambda_0}{2}\sum_{G\in\mathcal{G}}\sum_{j\in G}||\boldsymbol{U}_{G_j\cdot} - \bar{\boldsymbol{U}}_{G\cdot}||^2 \\ &+ \frac{\lambda_1}{2}(tr(\boldsymbol{U}^T(\boldsymbol{D_1}-\boldsymbol{S_1})\boldsymbol{U}) + tr(\boldsymbol{V}(\boldsymbol{D_2}-\boldsymbol{S_2})\boldsymbol{V}^T)) \\ &+ \frac{\lambda_2}{2}(||\boldsymbol{U}||^2 + ||\boldsymbol{V}||^2) \\ \text{s.t.}\quad & \boldsymbol{U}>0\ \boldsymbol{V}>0\end{aligned} \tag{2}$$

The gradient of Eq. (2) with respect to $\boldsymbol{U}_{i\cdot}$ can be expressed as follows,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial(\boldsymbol{U}_{i\cdot})} =& -\Big(\boldsymbol{Y}_{i\cdot}\odot(\boldsymbol{R}_{i\cdot}-\boldsymbol{U}_{i\cdot}\boldsymbol{V})\Big)\boldsymbol{V}^T + \lambda_0\sum_{G\in\mathcal{G}\&i\in G}(\boldsymbol{U}_{i\cdot}-\bar{\boldsymbol{U}}_{G\cdot}) \\ &+ \lambda_1((\boldsymbol{D_1}-\boldsymbol{S_1})\boldsymbol{U})_{i\cdot} + \lambda_2(\boldsymbol{U})_{i\cdot}\end{aligned} \tag{3}$$

The gradient of Eq. (2) with respect to $\boldsymbol{V}_{\cdot j}$ can be expressed as follows,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial(\boldsymbol{V}_{\cdot j})} =& -\boldsymbol{U}^T\Big(\boldsymbol{Y}_{\cdot j}\odot(\boldsymbol{R}_{\cdot j}-\boldsymbol{U}\boldsymbol{V}_{\cdot j})\Big) + \lambda_1(\boldsymbol{V}(\boldsymbol{D_2}-\boldsymbol{S_2}))_{\cdot j} \\ &+ \lambda_2(\boldsymbol{V})_{\cdot j}\end{aligned} \tag{4}$$

2.3.1 Computation of $\boldsymbol{U}$ and $\boldsymbol{V}$

the multiplicative update rule is:

$$\boldsymbol{U}_{ij} \leftarrow \boldsymbol{U}_{ij}\frac{\Big((\boldsymbol{Y}_{i\cdot}\odot\boldsymbol{R}_{i\cdot})\boldsymbol{V}^T + \lambda_0\boldsymbol{X}_1 + \lambda_1(\boldsymbol{S_1}\boldsymbol{U})_{i\cdot}\Big)_j}{\Big((\boldsymbol{Y}_{i\cdot}\odot(\boldsymbol{U}_{i\cdot}\boldsymbol{V}))\boldsymbol{V}^T + \lambda_0\boldsymbol{X}_2 + \lambda_1(\boldsymbol{D_1}\boldsymbol{U})_{i\cdot} + \lambda_2\boldsymbol{U}_{i\cdot}\Big)_j} \tag{5}$$

Algorithm 1 GNMF

Input:
G: hierarchical HIN;
$\mathcal{P}$: the set of meta paths connecting genes;
λ_1,λ_2: adapting parameter;
α: step size;
ϵ: convergence tolerance;
Output: $\boldsymbol{W}$: the weight matrix of all genes on all paths
for each meta path $\mathcal{P}_l \in \mathcal{P}$ **do**
  Evaluate user similarity $\boldsymbol{S}^l$
  Calculate predicting score $\hat{\boldsymbol{R}}^l$ with Eq.()
end for
$\boldsymbol{W} \leftarrow$ random values(between [0,1])
repeat
  Calculate $\frac{\partial\mathcal{L}}{\partial(\boldsymbol{W})_{ij}}$ with Eq.()
  Update $\boldsymbol{W}_{ij} \leftarrow max(0, \boldsymbol{W}_{ij} - \alpha\frac{\partial\mathcal{L}}{\partial(\boldsymbol{W})_{ij}})$
until convergence
return $\boldsymbol{W}$

where $\boldsymbol{X}_1 = \lambda_0\sum_{G\in\mathcal{G}\&i\in G}\bar{\boldsymbol{U}}_{G\cdot}$, $\boldsymbol{X}_2 = \sum_{G\in\mathcal{G}\&i\in G}\boldsymbol{U}_{i\cdot}$

$$\boldsymbol{V}_{ij} \leftarrow \boldsymbol{V}_{ij}\frac{\Big(\boldsymbol{U}^T(\boldsymbol{Y}_{\cdot j}\odot\boldsymbol{R}_{\cdot j}) + \lambda_1(\boldsymbol{V}\boldsymbol{S_2})_{\cdot j}\Big)_i}{\Big(\boldsymbol{U}^T(\boldsymbol{Y}_{\cdot j}\odot(\boldsymbol{U}\boldsymbol{V}_{\cdot j})) + \lambda_1(\boldsymbol{V}\boldsymbol{D_2})_{\cdot j} + \lambda_2\boldsymbol{V}_{\cdot j}\Big)_i} \tag{6}$$

update rule for $\bar{\boldsymbol{U}}_{G\cdot}$:

$$\bar{\boldsymbol{U}}_{G\cdot} \leftarrow \frac{\sum_{i\in G}\bar{\boldsymbol{U}}_{i\cdot}}{\sum_{i\in G}1} \tag{7}$$

2.4 Optimization

3 Result and Discussion

3.1 Simulation study

3.2 Data preparation

Gene and protein data Gene and protein data of the six species were obtained from BIOMART of Ensembl (http://www.ensembl.org/biomart/martview). As we focused mainly on PPIs and orthologous proteins, the genes retrieved were restricted to the protein-coding genes. The corresponding Ensembl Protein ID was considered because it would be cross-linked to the PPI and orthology data.

PPI and orthology data The PPI and orthology data were retrieved from the online database resource, Search Tool for the Retrieval of Interacting Genes (STRING) database [47]. The experimentally validated PPI in Human Protein Reference Database (HPRD) [48] were also incorporated by assigning a solid high score of 0.9. The combined score was calculated using the same strategy of STRING [47]. Each interaction was assigned by a combined score of various sources, indicating the reliability of the interaction. Since the majority of interactions in STRING were derived from computations based on prediction algorithms or interolog inference, we abandoned the interactions with a combined score less than 0.5. We also obtained orthologous proteins data from the STRING database, and scanned the domains by PfamScan[49] for further domain composition calculation (see Prioritization of gene-phenotype associations).

Phenotypes and known gene-phenotype associations The majority of the phenotypes were downloaded from the Open Biological and Biomedical Ontologies (http://www.obofoundry.org/). Known gene-phenotype associations were retrieved from the database of each

corresponding species (Table S1). For humans, we incorporated two databases, the Human Phenotype Ontology (HPO)[50] and the Online Mendelian Inheritance in Man (OMIM), into our database [51], and connected OMIM to HPO by annotations from http://www.human-phenotype-ontology.org.

PhenomeNET PhenomeNET is a cross-species phenotype network, in which the similarity between the nodes was calculated based on the information content of ontology terms [17]. We employed the information of the node pairs with a similarity score â‰¥0.5. With this network, the phenotypes from different species are available to be compared. PhenomeNET is available at http://phenomebrowser.net/availability.html.

4 Conclusion

Acknowledgements

Funding

This work has been 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).