P69. HUNTING HUMAN PHENOTYPE-ASSOCIATED GENES USING MATRIX FACTORIZATION

Pooya Zakeri[1,2,*], *Jaak Simm*[1,2], *Adam Arany*[1,2], *Sarah Elshal*[1,2] & *Yves Moreau*[1,2].
Department of Electrical Engineering, STADIUS, KU Leuven, Leuven 3001, Belgium[1]; *iMinds Medical IT, Leuven 3001, Belgium*[2]. [*]*pooya.zakeri@esat.kuleuven.be*

In the last decade, the phenotype-genes identification has received growing attention. It is yet one of the most challenging problem in biology. In particular, determining disease-associated genes is a demanding process and plays a crucial role in understanding the relationship between phenotype disease and genes. Typical approaches for gene prioritization often models each diseases individually, that fails to capture the common patterns in the data. This motivates us to formulate the hunting phenotype-associated genes problem as a factorization of an incompletely filled gene-phenotype-matrix where the objective is to predict unknown values. Experimental result on the updated version of Endeavour benchmark demonstrates that our proposed model can effectively improve the accuracy of the state-of-the-art gene prioritization model.

INTRODUCTION

In biology, there is often the need to discover the most promising genes among large list of candidate genes to further investigate. While a single data source might not be effective enough, fusing several complementary genomic data sources results in more accurate prediction. Moreover, fusing the phenotypic similarity of diseases and sharing information about known disease genes across both diseases and genes through a multi-task approach, enable us to handle gene prioritization for diseases with very few known genes and genes with limited available information. Typical strategies for hunting phenotype-associated genes often models each phenotype individually [1, 2, 3, 4], that fails to capture the common patterns in the data. This motivates us to formulate the hunting phenotype-associated genes task as a factorization of an incompletely filled gene-phenotype-matrix where the objective is to predict unknown values.

METHODS

We consider OMIM database which is a human phenotype disease specific association databases. OMIM focuses on the relationship between human genotype and associated diseases. OMIM database can be seen as an incomplete matrix where each row is a gene and each column is a phenotype (disease).

The idea behind the factorizing the M×N OMIM matrix is to represent each row and each column by a latent vector of size D. Then, the OMIM matrix can be modeled by product of an N×D gene matrix G and an M× D disease matrix P.

Bayesian matrix factorization (BPMF) [5] is a famous method to fill such an incomplete matrix. But BPMF uses no side information which results in an inaccurate gene-phenotype-matrix completion.

We propose an extended version of BPMF with an ability to work with multiple side information sources for completing gene-phenotype-matrix [6], which allows to make out-of-genes-phenotype-matrix ranking. In our proposed framework we are also able to integrate both genomic data sources and phenotypes information, whereas earlier approaches for hunting phenotype associated genes are limited to only fuse genomic information. This modification is done by adding genomic and phenotypic features to the corresponding latent variables [6]. In this study, we consider several genomic data sources including annotation-based data sources such as UniProt annotation, literature-based data sources on each genes, and as well the literature-based phenotypic information on each diseases, as just as in [1, 4, 9]. The framework of our Bayesian data fusion model for gene prioritization is illustrated in Figure 1.



***FIGURE 1**. The framework of our Bayesian data fusion model for gene prioritization.*

RESULTS & DISCUSSION

We report the average TPR results, when considering the top 1%, 5%, 10%, and 30% of the ranked genes. Experimental result on the updated version of Endeavour [3] benchmark demonstrates that our proposed model can effectively improve the accuracy of the state-of-the-art gene prioritization model.

REFERENCES

Aerts, S. *et al*. *Nat Biotech*, 24(5), 537–544, (2006).
De Bie T, Tranchevent LC, van Oeffelen LMM, Moreau Y, *Bioinformatics*, 23(13):i125-i132, (2007).
Tranchevent LC1, et. al. *NAR*, (35) W377-W384(2008) .
ElShal S, *et al*. Davis J. Moreau Y. *NAR*, (2015).
R. Salakhutdinov and A. Mnih. *25th ICML*, 880–887. ACM, (2008).
SIMM J, *et al*. arXiv:1509.04610 [stat.ML], (*2106)*.