Hunting Human Mendelian Disease-Associated Genes

Using Matrix Completion with Side information

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

1. INTRODUCTION

Discovering novel phenotype genes has received increasing attention in recent years. It is yet one of the most challenging problems in modern biology. Even hunting the Mendelian disease-related genes is a very demanding process and plays a major role in understanding the relationship between Mendelian phenotype diseases and genes. It has been shown that while a single data source might not be effective enough, integrating several complementary genomic data sources results in more accurate prediction [1]. Furthermore, integrating the similarity of diseases and sharing information about known disease-associated genes across both genes and diseases through a multi-task setting, allows us to handle gene prioritization for diseases with very few known genes and genes with limited available information. More specifically, whereas the typical approaches for hunting disease-related genes are limited to the use of genomic information [1], our proposed framework for gene prioritization can integrate both genomic data sources and phenotypic information simultaneously. Moreover, in order to detect and understand the common patterns in the data, we formulate the hunting disease-related genes task as a factorization of a partially filled gene-disease-matrix where the goal is to predict missing values in the Online Mendelian Inheritance in Man (OMIM) database [2]. This is in contrast to typical gene prioritization approaches [1] which model each disease separately.

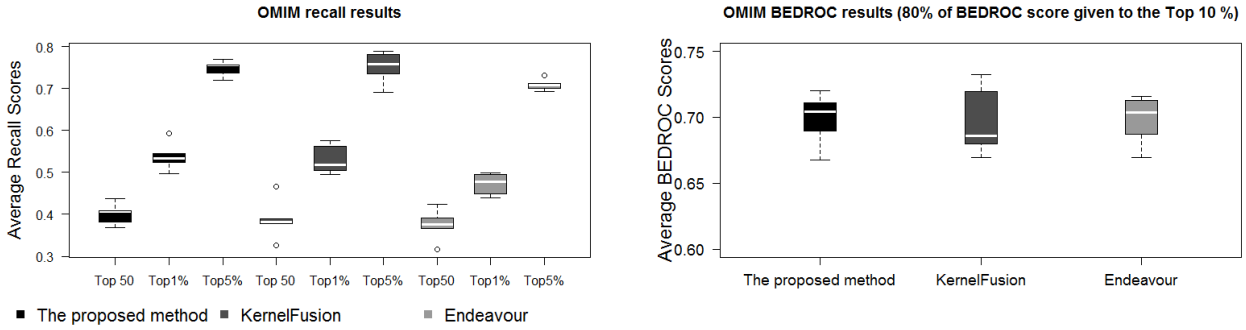
2. METHODS

Gene disease databases such as OMIM can be seen as an incomplete matrix M×N where each row is a gene and each column is a phenotype (disease). In our work, we consider OMIM database which is a human Mendelian phenotype disease specific association database. When the OMIM database is represented as an incomplete matrix, the OMIM matrix can be modeled by the two smaller matrices G and P, which, when multiplied, approximately reconstruct the OMIM matrix. 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; D can be significantly less than both M and N. Then, the goal is to find a factorization that minimize the mean square error on the observed data used for training the model while keeping a good performance on the test data. Bayesian matrix factorization (BPMF) [3] is a state-of-the-art method to fill such an incomplete matrix. Like BPMF, we apply a fully Bayesian treatment to deal with the matrix factorization. This modification is done by introducing common multivariate Gaussian priors for latent variables. But, BPMF uses no side information which leads to a poor gene-disease-matrix completion. We propose an extended version of BPMF with added the advantage of working with multiple genomic and phenotypic side sources for completing gene-disease-matrix [4]. This extension is done by incorporating genomic and phenotypic features to the corresponding latent variables.

In this study, we use several genomic data sources including annotation-based data sources such as Gene Ontology annotation, literature-based genomic information, and InterPro (IP) domains data

sources for genes, and as well the literature-based phenotypic information for diseases. After integrating the side genomic and phenotypic information to the OMIM matrix, we factorize the OMIM matrix using our proposed model. Then, we can complete matrix using the low-dimensional gene factor and the phenotype factor (40 hidden variables). Then, we sort values in each phenotype vector of the OMIM matrix (or each column of the OMIM matrix) in decreasing order and rank all genes in each column of OMIM matrix separately. To this end, we determine the rank of test genes in each disease vector to evaluate our model.

Figure. 1 The proposed approach vs Endeavor and KernelFusion: (left) Recall scores, (right) BEDROC scores



3. RESULTS & DISCUSSION

The performance of our models are evaluated using five folds cross-validation over Mendelian diseases with no less than 10 positive annotations in the OMIM database. We report the average recall results, when considering the top 50, 1%, and 5% of the prioritized genes. Moreover, we report the average BEDROC, a performance metric proposed to address the early discovery problems. It assigns different weights to the early and late part of ROC curve. In this study, we report BEDROC score which corresponds to 80% of the BEDROC being assigned to the top 10% of ranked genes. We compared the performance of our methods to Endeavour and KernelFusion approach for gene prioritization. Endeavour has been considered as a powerful tool for gene prioritization. We already developed and used KernelFusion model for the second Critical Assessment of Functional Annotation (CAFA2) to predict Human Phenotype Ontology (HPO) [5]. We have succeeded in obtaining the promising results among the participant groups in that challenge. For instance, our model ranked first in terms of Overall evaluation using the averaged AUC over terms with no less than 10 positive annotations [5]. As shown in Figure 1, experimental result on the updated version of Endeavour [1] benchmark indicates that our proposed model can boost the accuracy of the state-of-the-art gene prioritization model.

4. REFERENCES

1. Tranchevent, LC., et al. 2008. Endeavour update: a web resource for gene prioritization in multiple species. Nucleic Acids Res. 35: W377-W384.

2. Amberger, J., et al. 2011. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). Hum Mutat., 32, 564-567.

3. Salakhutdinov R. and Mnih A. 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo. *Proceedings of the 25th international conference on Machine learning ACM* pp. 880- 887.

4. Jaak S., et al. 2015. Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC. arXiv preprint arXiv:1509.04610.

5. Jiang, Y., et al., 2016, An expanded evaluation of protein function prediction methods shows an improvement in accuracy. arXiv.org > q-bio > arXiv:1601.00891.