Author’s response to reviews

Title: Predicting metabolite-disease associations based on KATZ model

Authors:

Xiujuan Lei (xjlei@snnu.edu.cn)
Cheng Zhang (chengzhang@snnu.edu.cn)

Version: 2 Date: 01 Sep 2019

Author’s response to reviews:

Re: Manuscript reference No. BIDM-D-19-00052_R2

Dear Editor,

Thank you for your letter. I am very pleased to resubmit the revised version of “Predicting metabolite-disease associations based on KATZ model”. The comments of the reviewer were highly insightful and encouraged us to greatly improve the quality of our manuscript. According to the reviewer’s suggestions, the revisions were addressed point-by-point below. We are also sending the revised manuscript.

We hope that the revisions in the manuscript and our accompanying responses will be sufficient to make our manuscript suitable for publication in BioData Mining.

We look forward to hearing from you at your earliest convenience.

Yours sincerely,

Xiujuan Lei (Corresponding Author)
School of Computer Science, Shaanxi Normal University, Xi’an 710119, China
E-mail: xjlei@snnu.edu.cn
Response to Reviewers

Reviewer #2:

Comments:

"1) In the Background section authors stated:

"some relevant methods about predicting have been delivered for genomics such as gene-disease correlations [9-11], transcriptomics like circRNA-disease associations [12, 13] and proteomics such as identification of essential proteins [14-16], but the computational methods for predicting metabolite-disease associations can be counted on the fingers of a hand." This is a very vague statement regarding competing methods. It is unclear whether those methods exist or not and if they exist, they should be mentioned explicitly.

Reply: Thanks for your comments. These methods in metabolomics has been existed but are very few compared with other fields. And the first method paper in this field as an example has been mentioned in paper."

REPLY: It is still unclear for a reader how is your method different, better or worse.

Reply: Thanks for your comments. In their paper, they only consider the metabolite similarity when using their method. In order to make full use of the known data, we will add the disease similarity in our method. AUC (area under curve) is defined as the area under the roc curve. We tend to use the value of AUC as the evaluation criterion for the method because the roc curve does not clearly explain which computational method work better in many case. The computational method are more effective when its AUC is higher than other methods. In our paper, we use three common validation methods to calculate the value of AUC. In Leave-one-out cross validation (LOOCV), the value of AUC in our method is 0.9181 while the RWR is 0.7633 and PAGERANK is 0.8242. In 5-fold cross-validation, the value of AUC in our method is 0.8897 while the RWR is 0.6692 and PAGERANK is 0.7951. In 10-fold cross-validation, the value of AUC in our method is 0.9029 while the RWR is 0.7266 and PAGERANK is 0.8113. According to these validation methods, KATZMDA can obtain higher AUC value. It means that KATZMDA is more effective than those compared methods and has a latent capability to explore more novel metabolite-disease associations.

"2) Section "Comparison with other methods":

Benchmarking of the results should be done with respect to the existing methods while authors just mention something about a classic random walk method and page rank that they have implemented themselves and did not give source code.
Reply: Thanks for your comments. Random walk has been applied in the first paper in this field and we use our data sets in random walk to compare with our method."

REPLY: Did you use the implementation from Hu et. al? If so it should be stated, otherwise more details on your implementation of benchmarking methods are needed. For instance, Hu et. al had 2 hyperparameters in their random-walk based method. Did you use the same values for your random-walk based comparison method?

Reply: Thanks for your comments. Because of different data, the networks which we and Hu et. Al build are different. For the sake of comparability of methods, our data is used in the compared methods, but we use the same values for random-walk based comparison method. The threshold epsilon is set 10^-6 and the parameter gamma is set 0.85.

3) No information about computational complexity or run-time. This information is also missing in benchmarking with other algorithms.

Reply: Thanks for your comments. Because the size of our data sets is larger, we don't calculate the computational complexity and we will consider it in the future work. In our paper, we set AUC as a judging standard to compare with other methods and don't consider it, either. We will consider it in the future work, too."

REPLY: Lack of publicly available source code and even approximate estimate on how long the computations might take, makes it very difficult for other members of the community to use the suggested method.

Reply: Thanks for your comments. After the publication of this paper, the source code during the current study will be obtained from the corresponding author on reasonable request. Because the configuration of the computer is different, the running time will be affected. When using our computer, we obtain the latent associations between 216 diseases and 2262 metabolites which may spend 5 minutes. The more effectiveness of the method has, the more useful the latent associations is. So we need use some validation methods to evaluate accuracy of our method which spend about 3 days. The compared methods also spend 3 days. So we can also get the compared results at the same time.