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: 1 Date: 18 Aug 2019

Author’s response to reviews:

Re: Manuscript reference NO. BIDM-D-19-00052

Dear editors,

Thank you for your letter. I am very pleased to resubmit for the manuscript “Predicting metabolite-disease associations based on KATZ model” after revision. Your comments and all of the reviewers were highly insightful and encouraged us to greatly improve the quality of our manuscript. In accordance with you and reviewer’s suggestions, the revisions were addressed point-by-point below. At the same time, we also sending the revised manuscript and revised portions can be tracked.

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, Cheng Zhang
Editorial comment:

Please clarify the relationship of this study and the paper published in 2017. https://ieeexplore.ieee.org/document/8046012 Is the current paper just an adaptation from miRNAs towards metabolites?

Reply: Thanks for your comments. KATZ algorithm was first proposed by Leo Katz to evaluate social associations in 1953 and it has been gradually applied in bioinformatics. This algorithm is a path-dependent global measure which directly sums all the possible paths between two nodes in a network and is exponentially damped to give the shorter paths more weight. There are some differences between two papers, which can illustrate the contribution of predicting the associations between metabolites and diseases. Firstly, the difference in ways of constructing similarity network. For example, they integrate the score gained by KATZ about miRNA and miRNA function similarity. However, we use metabolite gaussian interaction profile (GIP) kernel similarity. Secondly, the difference in main propose of KATZ. The main propose of using KATZ in their paper is to calculate the similarity. However, we find KATZ algorithm is suitable in our field when calculating the predicting associations and it is of great significance in understanding disease mechanism’s and advancing biology through integrated interdisciplinary research. So our main propose of using KATZ is to calculate the predicted score between metabolites and diseases. Thirdly, our paper and the paper in 2017 make different contributions in different field. Thanks for your comments again.

For Reviewers

Dear reviewer,

We are truly grateful to your kindly comments on our manuscript entitled “Predicting metabolite-disease associations based on KATZ model”.

Reviewer #1:

1) Human Metabolome Database has abundant information about small molecule metabolites found in the human body. How to extract the known associations. Can you give more details?

Reply: Thanks for your comments. Firstly, we download the data about HMDB and extract the associations between metabolites and diseases. Considering that we need to use disease semantic similarity in method, then we select the diseases with DOID and its relevant metabolites from the associations which has been extracted. Finally, we get 4537 metabolite-diseases associations which consist of 216 diseases and 2262 metabolites.
2) The authors have mentioned that the parameter $\delta$ mentioned by Zou was chosen $< 1/||M||^2$. I'm not sure why set the parameters $\delta$ 0.1 to 0.9 in Parameters analyzing?

Reply: Thanks for your comments. $\delta < 1/||M||^2$ is one of setting value for $\delta$. In our paper, we want to set different values of $\delta$ to explore its effect to our method and we find our method is steadier when changing the value of $\delta$.

3) In this paper, the authors mentioned that they applied a sampling method to ensure that they can obtain balance number of negative and positive samples. Can you provide more details about this sampling algorithm.

Reply: Thanks for your comments. After getting scores of all negative and positive samples by method, we guarantee that the number of positive samples will keep the same. Then we reduce the number of negative samples to the same number of positive samples by randomly selecting negative samples. Finally, we sort the positive and negative samples according to their scores.

4) In this paper, I wonder whether KATZ, RWR, PAGERANK are compared based on the same data set.

Reply: Thanks for your comments. In this paper, we use KATZ, RWR, PAGERANK are compared based on the same data set.

5) Some symbols should use Italics style in this study.

Reply: Thanks for your comments. It has been modified in paper.

Dear reviewer,

We are truly grateful to your kindly comments on our manuscript entitled “Predicting metabolite-disease associations based on KATZ model”.

Reviewer #2:

1) The quality of writing requires improvement:

a) Many sentences are too complicated, unclear and grammatically wrong.

Example: "Thus, it is indispensable to spread computational methods which can save experimental time and fund especially, available prediction results."; "However, only depending on these methods are vulnerable in terms of cost and time."

Reply: Thanks for your comments. It has been modified in paper.

b) There are multiple typos and incoherent spacing before/after dots or commas.

Reply: thanks for your comments. It has been modified in paper.

2) The meaning of acronyms is not given when the acronym was mentioned first (example: GIP kernel, DAG).

Reply: Thanks for your comments. It has been modified in paper.

The method:

3) 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.

4) Meaning and basic intuition behind KATZ framework are missing. Also, all related concepts needed for this model are not introduced.

Example: "Generally, what need to be taken into consideration when computing the potential association between metabolite i and disease j in the known metabolite-disease associations
network are the walks' number of metabolite i and disease j and the different length of different walks [22]." Here authors write about "walks" for the first time and never explain exactly what are those walks, what are they used for and why the length of the walks matter.

Reply: Thanks for your comments. The “walks” means paths between metabolites and diseases. According to the number of paths between each two nodes and the length of each path, KATZ can calculate the score of each two nodes to predict the latent associations. The higher the score is obtained, the greater the potential correlation is.

"Meanwhile, there is a detail cannot be ignored that the longer walks are supposed to have lower influence than shorter." The authors have never elaborated this statement as well even though it is important for algorithms parameters choice.

Reply: Thanks for your comments. KATZ is a path-dependent global measure which directly sums all the possible paths between two nodes in a network and is exponentially damped to give the shorter paths more weight. The longer path between two nodes is, the lower influence between two nodes has. So we need a non-negative coefficient δ control the influence of different-length paths.

5) 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.

6) No available source code.

Reply: Thanks for your comments. When the paper is published, the source code during the current study will be obtained from the corresponding author on reasonable request.

7) 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.