Author's response to reviews

Title: Research prioritization through prediction of future biomedical impact: a position paper

Authors:

Madhavi K. Ganapathiraju (madhavi@pitt.edu)
Naoki Orii (norii@cs.cmu.edu)

Version: 2 **Date:** 28 May 2013

Author's response to reviews: see over

May 28, 2013

Dear Dr. Scott Edmunds,

We thank you and the reviewers Dr. Matthias Erwin Futschik and Dr. Chaoming Song, for reviewing our paper and providing valuable feedback!

We inserted our responses into the reviewers' comments below. We carried out additional study as suggested by Dr. Futschik and presented the results for information. For reasons described, they are not inserted into the manuscript.

Please let us know if you would like to see any further changes to the manuscript.

I would like to add that our intention with this paper is to bring the attention of the scientific community to the fact that today there is a need to prioritize research based on which one will have a large impact on future science. Impact prediction of PPIs is meant to serve only as an example to demonstrate this. We hope that this point is coming across in our write up. We appreciate any comments about this.

Best Regards,
Madhavi Ganapathiraju

http://tonks.dbmi.pitt.edu/w

Reviewer: Matthias Erwin Futschik

Reviewer's report:

In the submitted work, Ganapathiraju and Orii have posed an intriguing question: Can we predict the scientific impact of an interaction between proteins? Whereas many studies have addressed the importance of proteins and interactions within a network context, the attempt to predict, whether or not the discovery of an interaction will evoke follow-up research, is novel. As a measure of impact of an interaction, the authors have used citations recorded for the publication linked to this interaction.

Notably, the authors showed that the impact can be predict using a set of network features (i.e. without any additional information regarding the function of proteins), although the accuracy of this prediction remains low. In fact, it is somewhat surprised that there has not been a stronger correlation detected with basic features such as the number of interaction that proteins have. Although the final accuracy remains limited, the work have addressed an interesting task and outlined a procedure how this can be tackled.

Major Compulsory Revisions:

1. The authors used a PPI network which presents the knowledge of today to assess the accuracy of their method. However, a more realistic set up would be to use a network which presents the knowledge at the point of time when a PPI was discovered, as such a discovery could lead to research activity and the identification of new interactions in the context of the interacting proteins. Thus, the identification of a

PPI itself might lead to a change of local network features over time which might bias the approach that the authors took. In principle, the use of older version of the human interactome should be possible as both HPRD and BioGRID are versioned. Such procedure would present a more faithful assessment of the performance of the method.

We agree strongly with the reviewer's point. We had already considered this point during our design phase, but did not carry out computations for the following reason: the network features are of very low reliability as it is, because only 40,000 interactions are currently known out of an estimated 600,000 interactions. Although our goal is to predict which PPI is biologically most relevant based on its network features, we had to do this with a partial view of the interactome.

As the reviewer expressed interest in this study, we did carry out the analysis, taking network features as available at the time of publication of that interaction. A formal description is given below, along with a figure showing precision versus recall values.

Previously, the random forest model used features that were calculated based on currently-existing PPI network. Now, for each interaction, features are calculated based on a network which presents the knowledge at the point of time when that interaction was discovered. For example, when calculating node and edge centralities for interaction i, we do not use the currently-existing PPI network but rather calculate features on an "old" network that was known when i was discovered. The following diagram shows average precision-recall curves at thresholds of 5, 10, 30, and 50. The blue solid line corresponds to the random forest model. The red dashed line corresponds to the random probability assignment. While both predictors are weak, our model improves over the baseline, especially for the top few predictions. At high recall, our model shows similar performance to the baseline, but at low recall, the proposed model shows significant increase in precision.

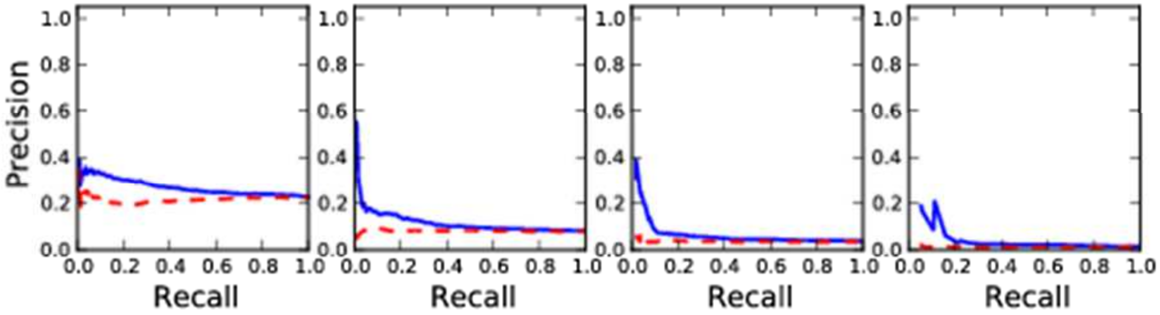

Figure RR1 – Precision recall curve with time dependent features

2. Table 4 seems to contain several errors. For example, PCM1 - KIAA0368 refers to Pubmed 16189514, which is a large Y2H screen. The authors however noted that the excluded such publications and used only one-to-one relationships of interactions and publications. Furthermore, the Pubmed id 12928435 was listed for two interactions, again violating their filtering procedure. The authors need to carefully check this issue.

This is not an error. The model was trained with all the interactions that have 1-to-1 relation with publication, but the prediction was carried out on all the PPIs. Based on the reviewer's comment, we understand that our wording was not clear. Therefore we edited the paragraph to make it clear.

The paragraph now reads as follows:

High-impact edges in the human interactome: After carrying out the evaluations described earlier, the method was applied to identify high impact edges from amongst all the PPIs in the human interactome. A final model was trained with all available interactions that have a one-to-one relationship with a paper (i.e. including those that were originally left out for evaluation purposes). The model was then applied to identify high-impact PPIs from amongst *all* the PPIs in the interactome. Note here, that the model was applied on all the PPIs without restricting to those that have one-to-one relationship with publications; such dataset is required only for training the model accurately and to evaluate the model reliably, whereas, the final *prediction* of whether a PPI is of high impact is carried out based on its network features alone without dependence on the number of times it has been reported. Table 4 lists the top 10 PPIs that are predicted to be of high impact. As can be seen, most of them indeed resulted in high impact on biomedical science, resulting in up to 413 citations. A longer list is available in Supplementary File 1.

Minor Essential Revisions1. The authors simplified their approach by using only interaction which were referenced by one publication. This excludes both publications with many interactions reported and interaction referred to by many publications. The authors may discuss how this exclusion is potentially influencing the performance of their method.

There are some papers such as those that presented yeast 2-hybrid studies that are typically cited for the technology. When such papers receive citations, they are usually not about any one interaction but about the technology itself. Conversely, it is possible for a paper to report a handful of interactions of which only one interaction is deemed "important" and receives majority of the citations if not all. In this case, we do not know which interaction contributed to the number of citations unless each of the citations is manually assigned to each interaction. Such manual curation is out of the scope of this work. Therefore, there was no way to include such data in our model which studies impact of each PPI separately.

Similarly, if a PPI is reported by multiple methods, in a way it is likely an important interaction. However it is quantitatively problematic to include it in the current model except by using some heuristic such as "maximum citations received by any publication reporting that interaction". As these methods are not accurate descriptors, we did not include them in the current quantitative modeling.

Please note that after the model is developed and evaluated with 1-to-1 dataset, it is applied on *all* PPIs and high impact PPIs are identified from amongst all of them.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.

Reviewer: CHAOMING SONG

Reviewer's report:

The manuscript by Ganapathiraju et al. presents a new methodological framework towards understanding the impact of PPI based on large scale citation datasets. Basically the authors provided data analysis on the datasets that offers a global view of biomedical impact. They further offered an inference model and test the predictive power of their algorithm. The manuscript is, in general, well written, technically sound, and interesting.

Discretionary Revisions

My only concern is that the new framework so far is based only on the scientific impact (that is the citation) without involving contexts of the references. Therefore, a high PPI involving many citations might be just a controversial one. Further validations based on semantic analysis have to be addressed somewhere.

Thank you.

About the concern, it is true that a paper may be cited for controversial reasons. Our reasoning is that such papers are not going to be many to influence the overall model and its performance. These, if any, would be few and may be treated as noise in the dataset.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests: no