Live-streamed preprint Journal Club on "EMT network-based feature selection improves prognosis prediction in lung adenocarcinoma" – October 23, 2018

Daniela Saderi, Ph.D.[1] and Dariusz Murakowski[2]

[1]Oregon Health & Science University
[2]MIT

April 28, 2020

Abstract

This is a review of the bioRxiv preprint "EMT network-based feature selection improves prognosis prediction in lung adenocarcinoma" by Borong Shao, Maria Bjaanæs, Åslaug Helland, Christof Schütte, Tim Conrad, doi:10.1101/410472. This review was compiled from a discussion during the live-streamed Bioinformatics preprint journal club as part of an Open Access Week effort organized by the PREreview team and PLOS. Event details can be found here, and the collaborative Etherpad showing all the journal club notes can be found here.

In addition to those named as authors above, the participants who wished to be acknowledged for their contributions to this review are as follows: Samantha Hindle, Paul Goetsch, and Bradly Alicea.

Summary

The goal of this preprint is to demonstrate the utility of using a phenotype relevant network-based feature selection (PRNFS) framework to improve prediction of cancer prognosis from multiple sets of high-dimensional omics data. The proposed network described biological interactions pertaining to epithelial-to-mesenchymal transition (EMT), with the goal of improving the prognosis prediction of lung adenocarcinoma.

All participants found the research very interesting and reported that, for the most part, the results supported the conclusions. However, one third of the participants reported having problems with understanding the methods because they appeared incomplete or not sufficiently clear for another researcher to replicate the findings. The major problem reported by two thirds of the participants was related to the figures and tables being hard to read and interpret.

Major comments

Many journal club participants recognized the importance of the study and the applicability of the results to research questions beyond cancer prognosis. Throughout the preprint, the authors make connections between gene expression networks and phenotypic processes in a novel way using a variety of methods. Many participants suggested the inclusion of a figure showing the network and a diagram to help the reader navigate

the comparisons between methods and appreciate the advantages and improvements of the proposed approach over alternative ones. Given that the leading author was present during the journal club, we learned that more details on the network are available on the author's GitHub repository and the link is in the manuscript – however, it currently leads to a 404 page. Given that many readers missed this, it was suggested that the authors emphasize this more in the manuscript. Additionally, in order for the GitHub repository to be useful and accessible, it would help to have a short description of its content in a README.md page (see GitHub guide).

Furthermore, it would be helpful for the reader to be able to tease apart the differences between feature selection alone and classification methods. For example, this would help address how selecting features based on the EMT-based phenotype GRN would improve predictions compared to random signature. One participant suggested using a framework developed by Venet et al. (2011) to rapidly assess comparisons between networks and validate improvements of one over another.

Other comments from the participants were also related to suggesting ways to improve the readability of the manuscript and help readers to more easily understand the main takeaways of the results. For example, it was suggested that the authors combined Figures S5, S6, S7 into one figure with three panels for a more direct comparison between GE+DM and GE or DM independently. For similar reasons, it was also suggested to select a fewer number of "essential" figures and tables for the main manuscript, and move the remaining figures and tables to supporting information. Additional suggestions are listed below.

Minor comments, suggestions, and typos

- It would be useful to add a paragraph to the Discussion highlighting (1) how this approach could be utilized to better target a set of prognostic markers from patient samples and (2) the potential generalization of this approach to other cancers.
- To avoid ambiguity, it would be better to use "AUROC" or "ROCAUC" instead of just "AUC". Moreover, this abbreviation and "AUPR " should be defined at their first usage (line 185). "ROC-PR" should be "AUPR."
- TCGA LUAD data should be cited according to guidelines given by the Broad Institute.
- For a network visualization, the networkD3 package in R is very useful.
- To improve readability of the preprint, it would be helpful to (1) print the supplementary table numbers and/or captions directly above their corresponding table contents, and (2) print the figure numbers on the same page as each figure (pages 20-30).
- Depending on the style guide, Ref. 37 should state that it is a PhD dissertation. It would also be helpful to refer to specific chapters or even sections.
- NetRank should have a capitalized R throughout.
- The authors should clarify how the EMT features were binarized using their means (line 280).
- The authors should clarify which steps were performed elsewhere and which were performed for this manuscript. For example, they *previously* "tested the EMT signatures" (line 344).
- Minor editing and proofreading by a generalist reader would be helpful. Some examples are listed below:
- Line 4 typo: caner –> cancer
- Line 9 typo: details –> detail
- Line 132 typo: being –> are
- Line 137 typo: should use a semicolon, as in "samples; we thus"
- Line 139: should be 74, 455, 123 nodes, respectively
- Lines 206-207: "association rule mining" can be cited earlier than line 269. Some of the details of this approach are currently found in the Results section; they may be more appropriate in the Experiments

and/or Introduction sections.

- Line 282: textit should be preceded by \ in the (presumably LaTeX) source code
- Lines 288-304: the interpretation of these rules would benefit from citations supporting the "established findings in cancer research"

References

David Venet, Jacques E. Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, 7(10):e1002240, Oct 2011. doi: 10.1371/journal.pcbi.1002240. URL https://doi.org/10.1371%2Fjournal.pcbi.1002240.