Reviewer's report

**Title:** Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia

**Version:** 2  **Date:** 22 April 2013

**Reviewer:** Maykel Cruz-Monteagudo

**Reviewer's report:**

The present manuscript provides new insights on the application to an unsolved clinical problem (etiology determination and/or identification of early predictive biomarkers of pre-eclampsia) of co-expression network analysis in combination with genetic algorithms to the prioritization of genes associated to pre-eclampsia. By taking advantage of publicly available genomic resources the authors not only outline the use of the above mentioned approaches for gene prioritization, but also identify genes already known to be related with pre-eclamptic conditions, which supports the reliability of the approach proposed. The approach proposed also allowed identifying a new set of genes less explored or unknown in the context of pre-eclampsia, opening new experimental opportunities.

In my opinion the medical genomics community will be interested on a paper describing how to take advantage of publicly available genomic information as well as systems biology, machine learning and evolutionary optimization tools for pre-eclampsia-related genes prioritization. Also of great interest for the readers of BMC Medical Genomics can be the new genes identified as a result of the study. So, I consider that the quality and scope of the present manuscript warrant a great interest by the readers of BMC Medical Genomics, just like a previous work of the authors on pre-eclampsia (Reference 8 in the manuscript) which is a highly accessed article on BMC Systems Biology.

Anyhow, some discretionary and minor essential revisions should be addressed by the authors:

**Discretionary Revisions:**

Although the authors put more emphasis on the results obtained by the GADA algorithm, in my opinion the very nature of the GANN algorithm should provide much more reliable results in terms of identification of genes related to a certain condition (pre-eclampsia in this case). The nearest-neighbor classification method is based on an instance-based learning process, where each new instance (represented as a vector of features) is compared with existing ones (represented on the same reference space) using a distance metric, and the closest existing instance is used to assign the class to the new one. Considering this, if a NN classifier is wrapped (and used as evaluation function) into a search method based on GA optimization algorithm; and the features used to represent each instance (case or patient in this case) is simply the expression value of the respective gene; what it is actually doing by the algorithm is to compare the
“expression profiles” determined by the subset of genes selected by the GA algorithm (the vector of genes used as reference space) and to assign the class of the case left out from training (LOO cross-validation) to the same class of the closest training instance (the case with the closest expression profile) according to a certain distance metric. So, the subset of genes identified by the GA algorithm that best differentiate a normal pregnant woman from a pre-eclamptic one should be significantly related to the pre-eclamptic condition. However, since the class assignment (and consequently the evaluation of the significance of the subset of genes identified) in the Fisher Discriminant Analysis is based on a linear combination of genes, the intuitive idea provided by the NN approach of an “expression profile” directly determined by the vector of gene expression values is lost. So, although perfectly valid for genes prioritization, the biophysical coherence with the phenomenon under study of the GADA approach is much less apparent compared with the GANN approach. I consider that this issue should be mentioned or highlighted in the manuscript to provide the reader with all the necessary information at the time to apply the approach proposed in the manuscript to a different clinical problem, which I consider is an added value of the present work (the extension to different related problems).

Minor Essential Revisions:
- Authors should mention which distance metric was used (Euclidean, Manhattan, Chebichev…?) on the GANN approach since it is well known that the results of the algorithm are highly dependent on the metric used.
- The codes used to identify the normal and pre-eclamptic groups or classes (N and PRE, respectively) must be defined prior to their use in the manuscript. In the current manuscript they are used prior to any definition which affects the correct understanding of the results described.
- The code “PRE” is used indistinctly to account for both, the pre-eclamptic class as well as to refer to the pre-eclampsia condition. It is confusing and so, different codes should be used.
- Finally, although essentially correct, the final presentation of the manuscript would be benefited of a new revision of the English and grammar.

1. Is the question posed by the authors well defined? Yes
2. Are the methods appropriate and well described? Yes
3. Are the data sound? Yes
4. Does the manuscript adhere to the relevant standards for reporting and data deposition? Yes
5. Are the discussion and conclusions well balanced and adequately supported by the data? Yes
6. Are limitations of the work clearly stated? This point could be improved.
7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished? Yes
8. Do the title and abstract accurately convey what has been found? Yes
9. Is the writing acceptable? Yes

RECOMMENDATION: PUBLISH AFTER MINOR CHANGES NOTED.

**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:**

I declare that I have no competing interests.