Author’s response to reviews

Title: Metrics to Estimate Differential Co-Expression Networks

Authors:

Elpidio-Emmanuel Gonzalez-Valbuena (emmanuel.valbuena@gmail.com)

Victor Trevino (vtrevino@itesm.mx)

Version: 1 Date: 06 Sep 2017

Author’s response to reviews:

Response to reviewers’ comments for the manuscript BIDM-D-17-00030 entitled “Metrics to Estimate Differential Co-Expression Networks” by Elpidio-Emmanuel Gonzalez-Valbuena and Victor Trevino, which was sent to the BioData Mining scientific journal.

In the following, the original comments from reviewers are shown in black while the responses are shown in blue and text modifications in red.

Reviewer reports:

Reviewer #1: In this manuscript, the authors analyze 6 different measures of differential co-expression to detect differences in gene expression data. They consider 4 metrics known in literature and 2 novel metrics and propose a methodology to generate controlled datasets from real data for evaluate the performance of differential co-expression measures. The methods are also compared on TCGA breast tumor data. Moreover, they provide a package in R to implement differential co-expression analyses.

The authors should correct fundamental language errors. For example, in the definition of the metrics 4, they have to correct: "All metrics above selects" in "All metrics above select" and there are many other errors of this type in the text. In other cases, the uncorrect language makes difficult to understand the meaning of the sentences, for example:” The generated dataset will maintain the internal correlation structure than experimental cancer data but absent of differential expression."

-------- Response ---------

For the specific examples shown, we have made the corrections as follows. The former was corrected removing the “s” at the end of the word selects. The later was clarified removing “than experimental cancer data”, so the sentence is now “In this way, the generated dataset will maintain the internal correlation structure but absent of differential expression”.

In addition, we have made several English corrections, provided by a professional editing service (Scribendi), which have been highlighted in blue in the new submission. We highly appreciate the effort of the reviewer to improve the quality of the manuscript.

---------------

The methods are clearly described except minor concerns that I list in the following:

1) In the description of the 3 categories of methods, the authors define the third category as "differential gene co-expression". This definition is to be modified to show the centrality of the gene in these approaches. For example, they can call this class "genes with differential co-expression patterns".

------- Response -------

We welcome the suggestion from reviewer to improve the definition of this category. For simplicity, clarity, and consistency with the other categories, we re-defined the mentioned category as “differential co-expressed genes”.

---------------

2) When the authors report the metrics in the reference [15] in the background section, they do not clarify which is the exponential beta. The beta parameter is described in the methods section but the authors have to clarify the terms in the formula when they introduce it in the background section. Moreover, they have to clarify that the sum is on the index j and sqrt is for the squared root.

------- Response -------

As suggested, we amended the introductory sentence as follows:

“For example, a method focuses only on differential connections for gene i estimating \( \Delta C_i = |d_{i0} - d_{i1}| \) where \( d_{ix} \) is the number of co-expressed genes above a correlation threshold in each condition [16]; while another method considers all correlations using \( \Delta C_i = \sum(\sqrt{|\text{sign}(A_{ij})\cdot A_{ij}^2 - \text{sign}(B_{ij})\cdot B_{ij}^2|}^\beta) \) where \( A_{ij} \) and \( B_{ij} \) represent the correlation coefficients for gene i with all genes j in the A and B experimental conditions, \( \sqrt{\cdot} \) is the square root function, \( \beta \) weights for large correlation differences, and the sum function sums over all j genes [15].”

---------------

However, I have major concerns about the procedure to generate controlled datasets. The authors build covariance matrices in order to generate data of genes linked in a network. They did not check for the positive definiteness of the covariance matrices. A criterium to ensure positive
definiteness is the belonging of the covariance matrices to the class of diagonally dominant matrices.

-------- Response --------

By definition, all correlation matrices generated are positive definitive. For example the correlation matrix of \( v=0.7 \) (10x10, all values \( =0.7 \) except diagonal=1) generates eigenvalues of 7.3, 0.3, 0.3, 0.3, \( \ldots \). For \( v=0.5 \), the eigenvalues are 5.5, 0.5, 0.5, 0.5, \( \ldots \). For \( v=0.2 \), the eigenvalues are 2.8, 0.8, 0.8, 0.8, \( \ldots \). All these matrices are positive definitive. Note that the correlation matrix of the generated data matrices using the package mvtnorm are approximated to these correlation matrices. So although the matrices have unique eigenvalues, the generated matrices of data (having number of columns equal to samples or patients) are random with correlation matrix very similar to the given correlation matrix. This property is used to evaluate how well the metrics detect the expected differences in correlations.

So, in conclusion, there is no need to check for positive definitiveness. We nevertheless appreciate the comment of the reviewer.

--------

Moreover, it is not clear how the addition of Gaussian noise of standard deviation \( s/3 \) for a fixed number of genes and \( s/10 \) for a different number of genes (where \( s \) is the standard deviation of the cancer data set) can preserve the correlation structure of the cancer data set.

-------- Response --------

The inspiration is that “positive genes” need to have correlation structure similar to that observed in cancer data. So, in one type of “positive genes”, in the stage T1 they will have \( s/3 \) variance, in T2 will accumulate \( 2s/3 \), and in T3 will aggregate to \( 3s/3 = s \). Thus, the “positive genes” in T3 will resemble the correlation structure of those observed in cancer data. Nevertheless, we agree that to resemble the whole dataset of cancer data, \( s \) would need to be applied to all genes. We did not apply \( s \) to all genes because there will be uncertainty about which genes would be finally affected making difficult the evaluation of metrics. So, in order to provide a controlled dataset having clear positives but also variability in all genes, we used \( s/3 \) for positive genes and \( s/10 \) for negative genes. To us, this provide clear differences between the two sets of genes but still leaving room for random effects.

To clarify these ideas, the following sentence was added in the corresponding section:

“So far, the correlation structure of positive genes will resemble that of an observed tumor dataset while the correlation structure of negative genes will be clearly less similar helping to distinguish between both types of genes.”

We thank the reviewer that helped to improve the manuscript.
Finally, I am not convinced of the need to generate two datasets that do not have differentially expressed genes but show only differences in gene correlations. In gene expression data sets, it happens that there are genes that change their correlations with many genes and we can indicate these as core genes. It could happen that the core gene is not differentially expressed but the genes that change their correlations with it result differentially expressed. It is important to consider differentially expressed and differentially co-expressed core genes at the same time. The differentially expressed genes represent deregulated functions in the cells. The differentially co-expressed core genes could represent key cancer driver gene as a result of mutations or other cancer modifications.

Response

We agree that real datasets include differential expressed genes. Still, differential expression, as such, does not affect the numeric estimation of correlation used in differential co-expression and therefore real data can be analyzed using our framework. For simulations however, if differential expression is included, it needs more complex models to ensure the control over correlations (apart of using a trivial model of adding a constant value, which will have clearly no effect on correlations). Thus, adding differential expression in the simulation would inevitably increase the complexity of the methodology unnecessarily. We attempted to keep the simulations as simple as possible while still reflecting changes in correlations that can be assessed with the metrics in an objective manner.

We appreciate all comments from the reviewers.