Author's response to reviews

Title: A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data

Authors:

Li Li (maple_525@163.com)
Yang Guo (guoyang1024@gmail.com)
Wenwu Wu (wenwu_wu@yahoo.com.cn)
Youyi Shi (shyouyi@gmail.com)
Jian Cheng (chengjian2007@126.com)
Shiheng Tao (shihengt@nwsuaf.edu.cn)

Version: 2 Date: 6 May 2012

Author's response to reviews: see over
Response to comments on revised manuscript 1457193402662778

Dear editor,

This manuscript is a revised version of our earlier manuscript entitled "A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data". Thank you for the opportunity to invite us to revise the manuscript. We have done the revision thoroughly in the light of the reviewer’s comments and suggestions.

We are grateful to the two reviewers for their helpful and constructive comments. Basically, we have re-written and re-organized the paper according to comments and suggestions made by the reviewers. The major changes are highlighted in colored font (blue) in the paper and point-by-point responses to the concerns are given in red typeface in this letter.

To Referee 1:

Major comments:

[Comment 1]
I suggest the authors perform a scientific check before resubmission. There are many grammar errors in the manuscript.

[Response 1]
The manuscript has been checked by English native speaker and grammar errors have been corrected.

[Comment 2]
For PPI metric, the authors failed to describe the method clearly. How PPI is calculated for each bicluster is not clear.

[Response 2]
The detailed descriptions of PPI and the calculation of PPI scores have been added in the part of Methods.

[Comment 3]
The authors picked two relative small gene expressions to evaluate the performance of those five algorithms. And authors mentioned some of those algorithms are data-dependent. But one may still have no idea about which algorithm will work better for a specific dataset. The author also pointed out some of algorithm dependent on number of genes or number of subjects. If so, a simulation study base one of these real dataset (e.g. randomly draw subset of the real data with different genes and different number of conditions) may explore more properties of those algorithms.
[Response 3]
Sincerely thank you for your very helpful comment, especially the suggestion about simulation study. We have rewritten the comparison results in the abstract as “(1) ISA is the most effective one among the five algorithms on the dataset of GDS1620 and BIMAX outperforms the other algorithms on the dataset of pathway. (2) Both ISA and BIMAX are data-dependent. The former one does not work well on the datasets with few genes, while the latter one holds well for the datasets with more conditions. (3) FABIA and QUBIC perform poorly in this study and they may be suitable to large datasets with more genes and more conditions. (4) SAMBA is also data-independent as it performs well on two given datasets”, which will provide useful information for researchers to choose the rational algorithm according to datasets they use. Simulation study based on real dataset is a good idea. But these data-dependent algorithms are considered to be more suitable for datasets of large size. Using subset of the given real dataset is not able to explore more properties of those algorithms insofar as the datasets we used are not large enough. However, your comment enlightened us to make a comparison with random gene groups.

[Comment 4]
The authors use two metrics to evaluate the performance of each algorithm, how many biclusters have un-consistent scores (e.g., high en-score low PPI score)?

[Response 4]
This is a very constructive comment. Kendall tau rank correlation coefficient was used to test whether the two metrics are statistically consistent. And the test result revealed that the two lists of scores are correlated positively with the p-value being $4.714e-11$.

[Comment 5]
Does the PPI score not depend bicluster size? Numbers of “singular” genes (does not belong to any network).

[Response 5]
We have recalculated the PPI scores and removed the influence of “singular” genes that do not belong to any network in the calculation.

Minor comments:
1. On page 3, line 5, “may be suit to large dataset”, change “suit” to “suitable” since “suit” is either a noun or a verb.
2. On page 3, line 15 “…but fare poorly when the number of ….” change “fair” to “fairly”
3. Page 10, line 8 it is hard to understand this sentence “because data GDS1620 which has fewer conditions than pathway data obtains 22 balusters”.


5. Many other grammar errors and sentences that is hard to understand. Please do a scientific check.

[Response]

All spelling errors mentioned above have been corrected according to your comments. Additionally, other grammar and spelling errors in the manuscript have been carefully and thoroughly checked and revised.

To Referee 2:

Major comments:

The whole Gene Ontology Weighted Enrichment Score section (on page 6, 7, 8) is poorly presented and has some fundamental problems. It is true that Li’s method (using the best P-value of GO term as the P-value of one bicluster which includes multiple GO terms) is not appropriate, since the P-value of a bicluster obtained in this way is size-dependent. But this problem could be easily solved by normalizing with the number of genes which have GO terms within one bicluster. Instead, the authors proposed to assign more weights to GO terms that include more genes. The reason for such a weight assignment is not specified in the paper, and it seems hard to justify. The P-value of one GO term obtained by a hyper-geometric test already reflects the size effect. Why should one GO term including more genes in one bicluster receive more weight? To justify the weight assignment, the author should make a comparison with the simple normalization solution, i.e., calculating the enrichment score of one bicluster as the sum of the enrichment scores of all GO terms divided by the total number of genes that have GO terms in this bicluster.

[Response]

Sincerely thank you for your helpful comments. The Gene Ontology Weighted Enrichment Score section has been rewritten, and the reason for the weight assignment of GO terms has been specified in this section as “the GO term associated with more genes may not have higher enrichment score, but it accounts for more proportions of genes in the bicluster. So we consider this term contribute more to the enrichment score of this bicluster”. Besides, the sum of all weight in one bicluster is equal to the denominator of the WE score. We have considered about the normalization solution as you suggested. This solution enlarged the range of the scores according to the biclusters’ size. Under this normalization solution, the scores of small size biclusters tend to be higher. For two biclusters of the same size, the one associated with more GO terms tends to
have higher score; the other associated with fewer GO terms which are even more significantly enriched tends to have lower score. Therefore, we consider weighted enrichment scoring will be better.

Minor comments:

[Comment 1]
Page 4 and page 5 have redundant full names of the five algorithms.
[Response 1]
We have corrected it following your comment.

[Comment 2]
The interpretation of P-value on page 7 is very confusing.
[Response 2]
The detailed interpretation has been added in this section.

[Comment 3]
The results part needs to be rewritten.
[Response 3]
We have rewritten the results part following your comment.

[Comment 4]
The use of the past tense through the paper would be preferred.
[Response 4]
We have done following your suggestion.