

## **Reviewer's report**

**Title:** A Biclustering Algorithm Based On A Bicluster Enumeration Tree : Application To DNA Microarray Data

**Version:** 4 **Date:** 26 August 2009

**Reviewer:** Federico Divina

### **Reviewer's report:**

#### MAJOR COMPULSORY REVISIONS

This paper introduces a new biclustering algorithm and a new evaluation function for biclusters. The paper is well organized.

However it presents several weak points.

First of all, no motivations for the introduction of a new biclustering algorithm and of a new evaluation function are given (there are more than the MSR around, so why don't use of those?).

The proposed evaluation function should be better explained, and authors should also provide some examples of simple biclusters with their ASR. For instance, give a perfectly correlated bicluster and show that its ASR is 1, and so on.

Authors should also justify why they decided to propose a new data structure for representing biclusters. Several solutions are described in state of the art literatures, so why using a completely new one?

The fact then that a new evaluation function is used in a new algorithm which uses a new representation model, render then difficult to assess the effectiveness of each new element.

For instance, it is hard to prove the effectiveness of an evaluation function if the function is incorporated in a new algorithm. I believe that using the ASR within an existing biclustering algorithm, for example the algorithm proposed by Cheng and Church, would be a more effective way to show that ASR can lead to more interesting biclusters. Moreover, the comparison of other measures would then be fair.

The preprocessing phase is a bit unclear to me. Authors suggest removing values with a procedure that should be better explained. This phase can be quite important, since eliminating values from the expression matrix could lead to a loss of information.

The description of the algorithms should also be improved, as it is now it is rather difficult to read.

The section that compares ASR with MSR and ACV is potentially interesting. However, I would not include such a section in the results section; it is rather a study on the function. Moreover, table 3 is unclear, what kind of biclusters are M1, M2....? As it is now, I cannot see the use of this proposed study.

Authors also claim that ASR is less sensitive to the presence of noise in the data, however they do not perform a specific study to show this property.

The experimentation should also be improved, especially for the part where real data are used. Only a dataset has been used, which, in my opinion is not enough to draw any interesting conclusions. The comparison with the other algorithms is rather superficial. The discussion of the gene expression profile is also rather useless as it is at the moment. Also, why don't authors show biclusters obtained with other algorithms? Also, visually, the biclusters shown do not seem to be very interesting. Figure 8 is not very clear either, I would suggest using a table instead. And how many biclusters are extracted by the various algorithms? Are the differences statistically significant? Interesting data about the biclusters found would also be their volume, which is not shown anywhere. GoTermFinder was used only on two biclusters found by the algorithm proposed in the paper. Why aren't the same results for the other algorithms shown? Various things could be added to this section.

#### MINOR ESSENTIAL REVISIONS

Abstract: instead of attributes, I would suggest to write "genes", since usually rows represent genes .

In the introduction authors should clearly state what a bicluster is (a subset of genes and conditions of the original expression matrix). The description of what a bicluster can be a bit confused to a reader not familiar with the problem. Also authors should explain that what one typically wants from a bicluster is that its genes present a coherent behavior under all the experimental conditions contained in the bicluster.

Pag 4 "most solution algorithms" -> most of the algorithms used to discover biclusters...

Pag 6, proposition 1 "let  $(I,J)$  a bicluster" -> let  $(I,J)$  be a bicluster

Pag 7 "attributes/individuals of the bicluster is strongly" -> genes of the bicluster are strongly

Pag 8, "the unmissing values"??????????

Pag 9, "threshold used on equation" -> threshold used in equation

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

'I declare that I have no competing interests