

Author's response to reviews

Title: A Biclustering Algorithm Based on a Bicluster Enumeration Tree:
Application to DNA Microarray Data

Authors:

Wassim Ayadi (wassim.ayadi@gmail.com)
Mourad Elloumi (mourad12345678@yahoo.com)
Jin Kao Hao (hao@info.univ-angers.fr)

Version: 5 **Date:** 9 October 2009

Author's response to reviews: see over

A Biclustering Algorithm Based on a Bicluster Enumeration Tree: Application to DNA Microarray Data

Authors replies to the comments and suggestions of the reviewers

Dear Editor-in chief,

Thank you very much for your decision regarding our manuscript. The authors are grateful to the reviewers for their insightful comments and suggestions.

Please find enclosed the answers to each referee's comments and a revised version of our manuscript according to the referee's comments. We tried to comply with their suggestions as far as possible. Also, an acknowledgement is added at the end of the paper.

We are looking forward to receiving your final decision.

Sincerely yours,

Wassim Ayadi, Mourad Elloumi and Jin-Kao Hao

Detailed replies to the recommendations of the referees:

Reviewer: Mr. Jason Moore

Referee: The description of the data and evaluation measures appear in the results. Please move to methods. This will make the results much easier to read.

Answer: We have moved the evaluation measures to the section "Method" and re-organized the "Results" Section. We hope this latter section is now easier to follow.

Referee: Is the software freely available and/or open-source? Please note in the paper. The impact will be greatly strengthened if the software is publically available.

Answer: The software is available upon request from the authors to academic users. This is indicated in the paper.

Referee: It would be nice to see a more detailed presentation of the other methods.

Answer: A more detailed presentation of the other methods is provided in reference [Madeira and Oliveira 04]. We added a phrase in "Introduction" to state this fact. Indeed, the current paper is already quite long, that's why we take this alternative solution. However, if you think it's better to include such a presentation in the paper, we can do it.

Referee: Were the different approaches statistically compared?

Answer: For the artificial data, the comparison is based on two criteria θ_{Shared} and $\theta_{NotShared}$. Thus only two values are used to compare one algorithm with another algorithm. For the real data, the comparison is done with different p -values and biological annotations.

Referee: Is the performance of *BiMine* on the artificial data 'significantly' better? This would be nice to see or made clearer if I missed it.

Answer: We tried to give more clarification in the subsection “Synthetic Data” and added more comments for Tables 4 and 5. Indeed, the aim to compare *BiMine* on synthetic data is to prove that *BiMine* can extract all implanted biclusters. Following Table 4 (resp. Table 5), *BiMine* can extract 100% (resp. 85.35 %) of implanted biclusters with an extra volume that represent 33.03% (resp. 41.78 %) of implanted biclusters. However, the best of the studied algorithms, i.e., CC, OPSM, ISA and *Bimax*, can extract only 58.18 % (resp. 42.87 %) of implanted biclusters with 21.39 % (resp. 49.31 %) of extra volume. This shows that *BiMine* has a good performance compared to these algorithms.

Referee: Equation 1 image is fuzzy.

Answer: We rewrote Equation 1.

Reviewer: Mr. Federico Divina

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

Answer: We added a small section in “Introduction” to motivate the need of developing new bi-clustering algorithms. Concerning the new evaluation function ASR, please notice that ASR is used by our *BiMine* algorithm to guide the exploration of the search space. Other evaluation functions such as MSR and ACV could be used, but we find our ASR function guides better the search process. This is now discussed at the beginning of Section “Methods - A New Evaluation Function of Biclustering”

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

Answer: In fact, in subsection “Illustrative Example” (resp. “Studies of the ASR Evaluation Function”) we gave some examples of biclusters with their ASR values.

Referee: 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?

Answer: With our *BiMine* algorithm, we need to represent the maximum number of significant biclusters and the links that exist between these biclusters. The BET data structure is specially designed for this purpose. We added an explanation in section “Building Bicluster Enumeration Tree” for this. We find that the data structures used in other biclustering enumerative algorithms [Tanay *et al* 02, Liu and Wang 03, Okada *et al* 07] don’t fulfill this requirement.

Referee: 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.

Answer: The main purpose of the paper is to introduce the *BiMine* algorithm and to present computational results of the WHOLE algorithm. Notice that ASR is used by *BiMine* to obtain biclusters and is NOT used to compare the biclustering results with other algorithms. Indeed, the comparison criteria (θ_{Shared} and $\theta_{NotShared}$ for synthetic data and p -values and biological annotations for the real data) are independent of the evaluation function used.

Nevertheless, your comment is interesting and merits of further studies. In a future work, we’ll particularly investigate the usefulness of our ASR evaluation function within other search algorithms. Already, people can use ASR within their biclustering algorithms.

Referee: 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.

Answer: Our preprocessing phase only eliminates *insignificant* expression values defined by Equation (4) and is explained in detail in Section “Methods-BiMine-Preprocessing”. An example is given in Tables 2 and 3.

Concerning the loss of information in the gene expression matrix, we can say that this loss is minimum and occurs only in two particular cases (missing values and values below a given threshold). Moreover, *BiMine* operates directly on the raw data matrix without resorting to a discretization of data, reducing thus the risk of loss of information. This issue is discussed at the end of this Section.

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

Answer: We rephrased the explanation of the algorithm. The illustrative example helps to understand the algorithm.

Referee: 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.

Answer: We have moved the seven matrices (biclusters M_1 - M_7) and evaluation measures to the “Method” section. The characteristics of these matrices are now explained in the paper. The meaning of Table 3 is also explained here.

Referee: 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.

Answer: In the section of “Conclusion”, we suppressed the mentioned phrase. Notice however that in [Balasubramaniyan *et al* 05], it has been shown that *Spearman’s rank correlation* is less sensitive to the presence of noise in the data. Since our evaluation function ASR is entirely based on *Spearman rank correlation*, ASR would be also less sensitive to the presence of noise in the data.

Referee: 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.

Answer: The yeast dataset used in the experimentation is a very popular one in the gene expression analysis and has been used alone by in many papers for the validation of data mining tools like [Cheng *et al* 08, Madeira and Oliveira 09, Dharan and Nair 09]. Moreover, the yeast dataset is the one of the well known organisms and the functions of each gene are well known.

Referee: 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.

Answer: Actually, our strategy of comparison with the other algorithms is the same as the one adopted in other works [Prelic *et al* 06, Gan *et al* 08, Dharan and Nair 09]. We removed the discussion of the gene expression profile, and added a figure that represents the gene expression profile of biclusters shown in Table 7, in the same way as it has been done in [Dharan and Nair 09]. In Additional File 1 we illustrate the best bicluster obtained by each compared algorithm.

The biclusters constructed thanks to our algorithm are interesting enough compared to those constructed by other biclustering ones [Cheng and Church 00, Ihmels *et al.* 04, Ben-Dor *et al.* 03, Prelic *et al.* 06] (see subsection “Biological relevance”).

Referee: 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?

Answer: We replaced Figure 8 by Table 6 that represents the values of this figure. The number of extracted biclusters by the various algorithms was also added.

The differences between the number of significant biclusters extracted by our algorithm and the ones of those extracted by the other algorithms (OPSM, *Bimax*, ISA, CC) are statistically

significant. Indeed, as it is shown in Table 6 *BiMine* found the best result for all p -values compared to CC, ISA and OPSM. Also, *BiMine* performs well for four cases of p -value (p -value =5%, p -value =1%, p -value =0.5% and p -value =0.1%) over five compared to *Bimax*.

Referee: Interesting data about the biclusters found would also be their volume, which is not shown anywhere.

Answer: We added the volume of biclusters in Table 7.

Referee: *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?

Answer: We included the best bicluster obtained by each compared algorithm with their gene expression profiles (see Additional File 1).

Referee: 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

Answer: These remarks were taken into consideration. A thorough review helped us to improve the quality of the redaction, to clear up grammar and phrasing problems.

References

The following list gives the references used in this letter, but some of them are not cited in the paper.

Balasubramanian, R., Hu llermeier, E., Weskamp, N., Kamper, J: Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics* 2005, 21, 1069–1077.

Ben-Dor, A., B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, New York, NY, USA, 2002. pp. 49–57. ACM.

Cheng, Y. and G. M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 2000 pp. 93–103. AAAI Press.

Cheng K, Law N, Siu W, Liew A: Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. *BMC Bioinformatics* 2008,9: 210

Dharan A. and A.S. Nair. Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. *BMC Bioinformatics*, 10(Suppl 1):S27, 2009.

Gan X., Liew A.W. and Yan H., Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics* 2008, 9:209

Ihmels J., Bergmann, S., , and N. Barkai: Defining transcription modules using large-scale gene expression data. *Bioinformatics* 2004(13), pp. 1993– 2003.

Liu J, Wang W: Op-cluster : Clustering by tendency in high dimensional space. In *Proc.3rd IEEE International Conference on Data Mining* 2003, 187-194

Madeira SC, Oliveira AL: Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2004, 1(1):24–45

Madeira S.C. and A.L. Oliveira. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology* 2009, 4:8.

Okada Y, Okubo K, Horton P, Fujibuchi W: Exhaustive Search Method of Gene Expression Modules and Its Application to Human Tissue Data. *IAENG International Journal of Computer Science* 2007, 34:1-16

Prelic, A., S. Bleuler, P. Zimmermann, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 2006, 22(9): 1122–1129

Tanay A, Sharan R, Shamir R: Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 2002, 18:S136-S144