SUPPLEMENTARY MATERIAL

Computing the maximum similarity bi-clusters of gene expression data

(Bioinformatics, 2006)

Xiaowen Liu and Lusheng Wang
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

1 PROOF OF THEOREM 1

First, we introduce a lemma that is crucial in the proof.

LEMMA 1. *Let $S(I, J)$ be a similarity matrix. $S(I_1, J_1)$ and $S(I_2, J_2)$ are two bi-clusters of $S(I, J)$ with $I_1 \subseteq I_2 \subseteq I$ and $J_1 \subseteq J_2 \subseteq J$. For each row $i \in I_1$, we have $s(i, J_1) \leq s(i, J_2)$. For each column $j \in J_1$, we have $s(I_1, j) \leq s(I_2, j)$.*

PROOF. In bi-cluster $S(I_2, J_2)$, J_2 contains all columns in J_1. For row $i \in I_1$, we have

$$s(i, J_2) - s(i, J_1) = \sum_{j \in J_2} s_{ij} - \sum_{j \in J_1} s_{ij} = \sum_{j \in J_2 - J_1} s_{ij} \ .$$

In addition, $\forall i \in I_2, \forall j \in J_2, s_{ij} \geq 0$. Therefore, $s(i, J_2) - s(i, J_1) \geq 0$ and $s(i, J_1) \leq s(i, J_2)$.

Similarly, I_2 contains all rows in I_1. For column $j \in J_1$,

$$s(I_2, j) - s(I_1, j) = \sum_{i \in I_2} s_{ij} - \sum_{i \in I_1} s_{ij} = \sum_{j \in I_2 - I_1} s_{ij} \geq 0 \ .$$

Therefore, $s(I_1, j) \leq s(I_2, j)$. $\Box$

THEOREM 1. *The MSB algorithm runs in $O((n+m)^2)$ time and outputs an optimal solution for the Maximum Similarity Bi-cluster problem.*

PROOF. We will prove the theorem by contradiction. Suppose the obtained bi-cluster of the MSB algorithm is $S(I_A, J_A)$ and there is another bi-cluster $S(I_{opt}, J_{opt})$ such that $s(I_{opt}, J_{opt}) > s(I_A, J_A)$ and $S(I_{opt}, J_{opt}) \neq S(I_A, J_A)$. In the MSB algorithm, we obtain $n + m - 1$ different bi-clusters $S(I_1, J_1)$, $S(I_2, J_2)$, ..., $S(I_{n+m-1}, J_{n+m-1})$. From Step 6 of the MSB algorithm, for any k' with $1 \leq k' \leq n + m - 1$, $S(I_{opt}, J_{opt}) \neq S(I_{k'}, J_{k'})$. Since $S(I_{opt}, J_{opt}) \neq S(I_{n+m-1}, J_{n+m-1})$, at least one row $i' \in I_{opt}$ or one column $j' \in J_{opt}$ is removed in Step 5 of the MSB algorithm.

Without loss of generality, we assume that row $i' \in I_{opt}$ is the first in I_{opt} or J_{opt} that is removed in the algorithm. (It is similar to show that case that a column $j' \in J_{opt}$ is the first to be removed.) Thus, we can assume that row i' is removed from I_k to get I_{k+1}, for some $1 \leq k \leq n+m-2$ and no row or column of $S(I_{opt}, J_{opt})$ is removed in the first $k - 1$ loops of the MSB algorithm. Therefore, $I_{opt} \subseteq I_k$ and $J_{opt} \subseteq J_k$. From Lemma 1,

$$s(i', J_{opt}) \leq s(i', J_k). \tag{1}$$

From the definition of $s(I, J)$ and the fact that $i' \in I_{opt}$, we have

$$s(I_{opt}, J_{opt}) \leq s(i', J_{opt}). \tag{2}$$

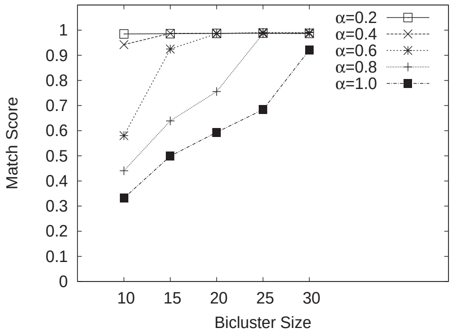
In the MSB algorithm, row i' is removed from $S(I_k, J_k)$ in step 5. Thus, we know row i' has the minimum similarity score in all rows and columns of $S(I_k, J_k)$. That is,

$$s(i', J_k) = \min\{\min_{i \in I_k} s(i, J_k), \min_{j \in J_k} s(I_k, j)\} = s(I_k, J_k). \tag{3}$$

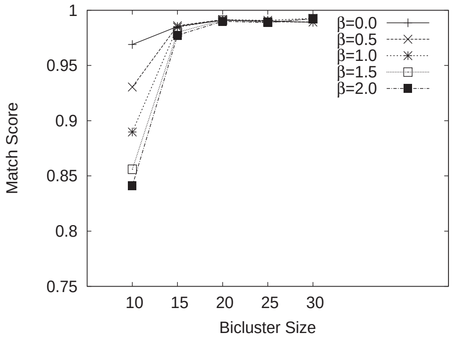
From equations (1, 2, 3), we have

$$s(I_{opt}, J_{opt}) \leq s(I_k, J_k) \leq s(I_A, J_A).$$

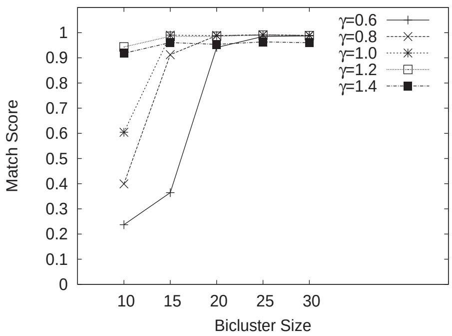
It is a contradiction with the assumption that $s(I_{opt}, J_{opt}) > s(I_A, J_A)$ and $S(I_{opt}, J_{opt}) \neq S(I_A, J_A)$. $\Box$



(a) Test of α.



(b) Test of β.



(c) Test of γ.

Fig. 1. Results for parameters selection.

2 PARAMETER SELECTION

Let us study the effects of different parameters settings. We implanted non-overlapped square additive bi-clusters with sizes ranging from 10×10 to 30×30. The noise level is $\delta = 0.1$. To show the effects of the three parameters α, β, γ to RMSBE, we tested the performance of RMSBE with three parameter settings: (1) $\beta = 0.5$, $\gamma = 1.2$ and α is in the range $[0.2, 1.0]$, (2) $\alpha = 0.4$, $\gamma = \beta + 0.7$ and β is in the range $[0.0, 2.0]$, (3) $\alpha = 0.4$, $\beta = 0.5$ and γ is in the range $[0.6, 1.4]$. The results are shown in Figure 1 (a), (b) and (c).

The value of α determines the threshold for similarity score. (See Equation (1) in the paper.) If α is small, only the expression values very close to the reference gene are considered. Increasing the value of α allows the algorithm to consider more expression values. Figure 1(a) shows that RMSBE obtains the best match scores when α is in $[0.2, 0.4]$.

β is the bonus for the similarity score. When the reference gene distance of an element is greater than the threshold $\alpha \cdot d_{avg}$, β affects the similarity score of the element. If β is big, all elements with reference gene distance greater than the threshold $\alpha \cdot d_{avg}$ have similar scores. Otherwise, the similarity score is closely related to the term $\frac{d_{ij}}{\alpha \cdot d_{avg}}$ in Equation (1). Figure 1(b) shows that for the values of β in $[0.0, 0.5]$, the algorithm performs well.

The role of γ is to filter out large low average similarity bi-clusters. When γ is big, the algorithm outputs bi-clusters with smaller sizes and higher qualities. On the contrary, bi-clusters with larger sizes and lower qualities are discovered when γ is small. Figure 1(c) shows that when $\gamma = 1.2$ and $\gamma = 1.4$, RMSBE has good performance for the implanted bi-clusters with different sizes.

3 TEST RESULTS USING MATCH SCORE IN PRELIĆ *ET AL.*

Prelić *et al.* (2006) use a different match score

$$S(M_1, M_2) = \frac{1}{|M_1|} \sum_{A(I_1,J_1) \in M_1} \max_{A(I_2,J_2) \in M_2} \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}$$

in the experiment. We also test RMSBE and other methods using the match score in Prelić *et al.* (2006). The test results are similar with the results in the original paper.

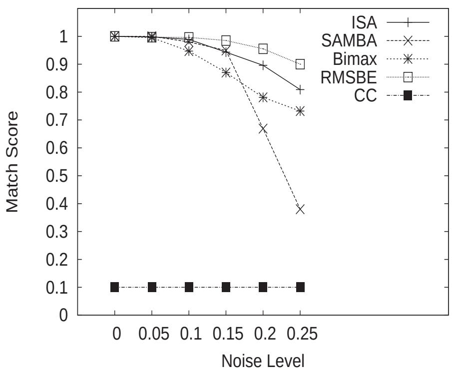


Fig. 2. Results for constant bi-clusters (corresponding to Figure 4 in original paper) using match score in Prelić *et al.* (2006).

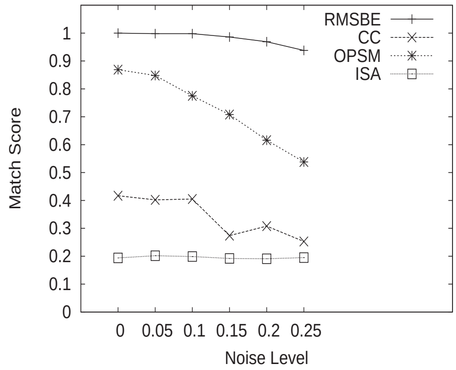


Fig. 3. Results for additive bi-clusters (corresponding to Figure 5(a) in original paper) using match score in Prelić *et al.* (2006).

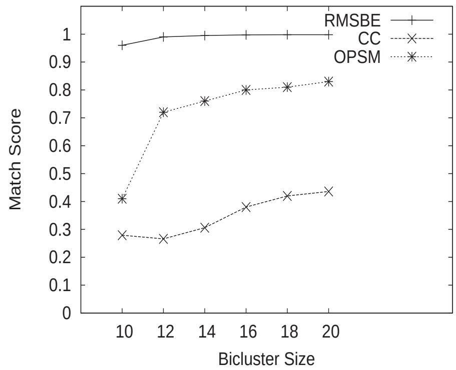


Fig. 4. Results for additive bi-clusters with different sizes (corresponding to Figure 5(b) in original paper) using match score in Prelić *et al.* (2006).

4 TEST RESULTS USING DIFFERENT DATA DISTRIBUTIONS

We further test the performance of RMSBE and other methods on the data fitting different normal distributions. The results with mean 0 and SD$= 0.5$ are shown in Figure 6. The results with mean 7 and SD$= 1$ are shown in Figure 7. We can see that Figure 6 and Figure 7 are identical to Figure 5 (a) in the original paper.

REFERENCES

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

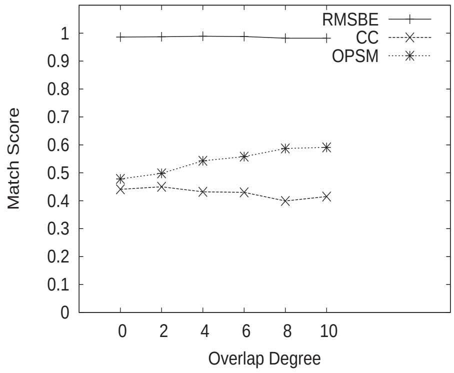


Fig. 5. Results for overlap additive bi-clusters (corresponding to Figure 5(c) in original paper) using match score in Prelić *et al.* (2006).

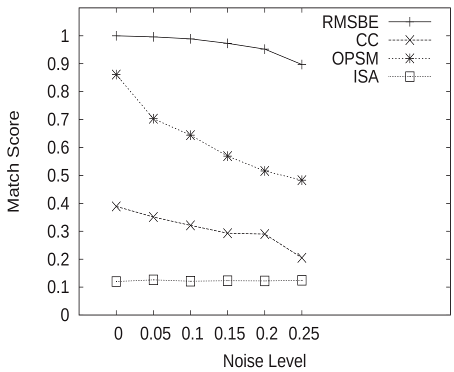


Fig. 6. Results for additive bi-clusters using data fitting the normal distribution with the mean of 0 and SD$=0.5$. The parameters for CC are changed to $\delta = 0.0005$, $\alpha = 1.2$ to get better result. Other parameters are the same with those used in Figure 5 (a) in the original paper.

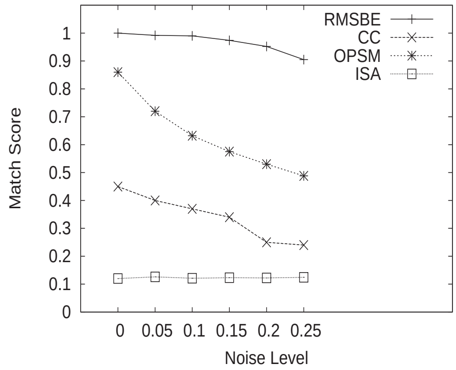


Fig. 7. Results for additive bi-clusters using data fitting the normal distribution with the mean of 7 and SD$=1$. The parameters are the same with those used in Figure 5 (a) in the original paper.