

Incorporating prior knowledge into clustering of gene expression profiles

Daisuke Komura¹

komura@hal.rcast.u-tokyo.ac.jp

Hiroshi Nakamura¹

nakamura@hal.rcast.u-tokyo.ac.jp

Shuichi Tsutsumi¹

tsutsumi@genome.rcast.u-tokyo.ac.jp

Hiroyuki Aburatani²

haburata-tky@umin.ac.jp

Sigeo Ihara¹

ihara@genome.rcast.u-tokyo.ac.jp

¹ Research Center for Advanced Science and Technology, The University of Tokyo.
4-6-1 Komaba, Meguro, Tokyo 153-8904, Japan

² Genome Science Div., Center for Collaborative Research, The University of Tokyo,
4-6-1, Komaba, Meguro, Tokyo 153-8904, Japan

Keywords: gene expression profiles, prior knowledge, semi-supervised clustering

1 Introduction

Clustering genes with similar expression profiles in order to collect coexpressed genes is one of the most important tasks in revealing the complex biological regulatory networks[1]. Most clustering algorithms need many parameters to be specified a priori, such as the number of clusters, cluster diameters and similarity measures to obtain accurate clusters. However, the appropriate parameters depend on each dataset and inappropriate parameters lead to wrong clusters. Unfortunately, choosing the appropriate parameters automatically is still a challenging problem.

Recently, large amount of biological information, including biological literatures and public microarray datasets, has been accumulated. In this paper, we present a new search-based clustering algorithm which incorporates such biological knowledge. As depicted in Figure 1, in order to obtain better clusters, the knowledge is transformed into additional pairwise constraints for choosing appropriate parameters. This algorithm is categorized in semi-supervised clustering algorithms often presented in the text mining field[2][3]. We indicate that our clustering algorithm provides more meaningful clustering results than conventional ones, and help us to decipher the gene regulatory networks.

2 Method

We introduce a new semi-supervised clustering algorithm which utilizes biological knowledge. Such knowledge is provided as two types of constraints between pairs of instances (genes): *must-link* constraints specifying that two instances (genes) have to be in the same cluster, and *cannot-link* constraints specifying that two instances (genes) cannot be in the same cluster. For example, a known fact "genes A and B are co-regulated." is transformed into a *must-link* constraint between gene A and gene B. In our algorithm, parameters to be tuned with *must-link* and *cannot-link* constraints consist of the distance matrices \mathbf{A} of each cluster, the diameter of each cluster $\delta(\mathcal{Z}_k) = \max_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{Z}_k} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}_k}$, where $\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}_k} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}_k (\mathbf{x}_i - \mathbf{x}_j)}$. In other words, the cluster number and cluster structures are determined automatically by adding pairwise constraints. We show brief description of our algorithm in Figure 2. Clusters are created to satisfy pairwise constraints. First, fine-grained clusters, or core clusters with small diameters are created. Then, each cluster is expanded to satisfy *must-link* constraints. Finally, distance matrices of each cluster are updated to satisfy *cannot-link* constraints. Detail of the algorithm will be shown in our poster.

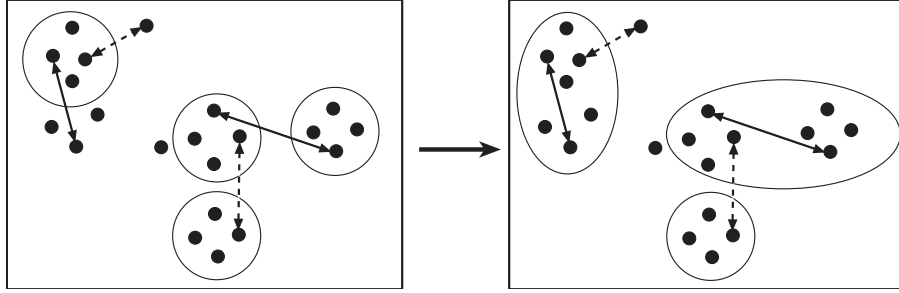


Figure 1: Improvement of the clustering results with pairwise constraints (left/right figure: without/with constraints). *Must-links* and *cannot-links* are expressed as solid lines and dotted lines, respectively.

```

Algorithm: Search-based semi-supervised clustering
Input: Set of instances (genes)  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ 
       set of must-link constraints  $\mathcal{M} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$ , set of cannot-link constraints  $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$ 
       upper-bound of diameter of final cluster  $d_u$ , threshold value of diameter of core cluster  $d_c$ 
Output: set of  $k$  clusters  $\mathcal{Z} = \{\mathcal{Z}_i\}_{i=1}^k$  and  $k$  weight matrices  $\mathcal{A} = \{\mathbf{A}_i\}_{i=1}^k$ 
Method:
0. Initialize clusters
  0a.  $\mathcal{Z}_i = \{\mathbf{x}_i\}$  for each  $\mathbf{x}_i$  in  $\mathcal{X}$ 
1. Create core clusters
  1a. Create new clusters  $\mathcal{Z}_{new} = \{\mathcal{Z}_i, \mathcal{Z}_j\}$  whose merger results in the smallest diameter and  $\delta(\mathcal{Z}_{new}) < d_c$ 
  1b.  $\mathcal{Z} = \mathcal{Z} \cup \mathcal{Z}_{new} - \{\mathcal{Z}_i, \mathcal{Z}_j\}$ 
2. Merge clusters to satisfy must-link constraints as long as the diameter of merger clusters does not exceed  $d_u$ 
  2a. Create new clusters  $\mathcal{Z}_{new} = \{\mathcal{Z}_i, \mathcal{Z}_j\}$  if  $\{(\mathbf{x}_k, \mathbf{x}_l) \in \mathcal{M} | \mathbf{x}_k \in \mathcal{Z}_i, \mathbf{x}_l \in \mathcal{Z}_j\} \neq \phi$  and  $\delta(\mathcal{Z}_{new}) < d_u$ 
  2b.  $\mathcal{Z} = \mathcal{Z} \cup \mathcal{Z}_{new} - \{\mathcal{Z}_i, \mathcal{Z}_j\}$ 
3. Learn the distance measure of each cluster to satisfy cannot-link constraints
  3a. Update matrix  $\mathbf{A}_i$  if  $\{(\mathbf{x}_k, \mathbf{x}_l) \in \mathcal{C} | \mathbf{x}_k, \mathbf{x}_l \in \mathcal{Z}_i\} \neq \phi$ 

```

Figure 2: Search-based semi-supervised clustering algorithm.

3 Results and Discussions

The proposed algorithm has many advantages over other semi-supervised clustering algorithms which utilize pairwise constraints. First, the number of clusters is determined automatically. This property is especially beneficial when there are many clusters. Second, because not all genes are allocated to cluster, clustering results are not affected by outlier or singleton. Moreover, users can modify the initial results by giving additional pairwise constraints with low computational costs.

The results of numerical experiments to validate the proposed algorithm will be shown in our poster.

References

- [1] Segal, E., Stuart, J., Koller, D. and Kim, S., A Gene Co-Expression Network for Global Discovery of Conserved Genetics Modules, *Science*, 302:249-255, 2003
- [2] Bilenko, M., Basu, S. and Mooney, R.J., Integrating Constraints and Metric Learning in Semi-Supervised Clustering, *Proc. 21st International Conference on Machine Learning (ICML2004)*, 81-88, 2004
- [3] Cohn, D., Caruana, R. and McCallum, A., Semi-supervised Clustering with User Feedback, *Technical Report TR2003-1892, Cornell University*, 2003