

A Novel Clustering Algorithm with Map Energy Minimization

Takatoshi Kawai¹

t-kawai@hhc.eisai.co.jp

Yasuto Yokoi²

yokoi@hydra.mki.co.jp

Yuji Miura²

yuji2@hydra.mki.co.jp

Tsuyoshi Tabata¹

t2-tabata@hhc.eisai.co.jp

Takeshi Nagasu¹

t-nagasu@hhc.eisai.co.jp

Ken Aoshima²

kaku@hydra.mki.co.jp

¹ Laboratory of Seeds Finding Technology, Eisai Co., Ltd., 5-1-3 Tokodai, Tsukuba-shi, Ibaraki 300-2635, Japan

² R&D, Bioscience Division, Mitsui Knowledge Industry Co., Ltd., Harmony tower 21st, 32-2, 1-chome, Nakano-ku, Tokyo 164-8721 Japan

Keywords: clustering algorithm, energy, map arrangement

1 Introduction

Cluster analysis is the fundamental method for DNA microarray gene classification [1], structure activity relationship [2] and any kind of correlation matrix analyses. The hierarchical clustering analysis is the most widely used method providing us clear-separated clusters, however, often distributes closely related clusters into apart positions because the direction of branches at each node is arbitrary. This limitation often makes the correlation matrix look like checkered pattern.

In this work, we propose a new clustering method to overcome this problem. Our new method searches a set of row and column orders in which any pair of the highly coherent rows (or columns) are placed adjacently in the correlation matrix map. The final map is no longer expected to have checkered pattern. We assumed that it could perform this process to minimize the energy of the correlation map pattern. We named our new method “Map Energy Minimization Clustering (Memim clustering)”.

2 Method and Results

A setup of the energy definition and boundary conditions, which determines the final correlation map pattern, and the map arrangement technique, which searches for the minimum energy map, are the important points in Memim clustering.

2.1 Energy Definition Formula

The energy was defined as Formula 1, which satisfies the following conditions, 1) rows or columns could be moved independently each other, 2) cells with similar values were arranged in the neighborhood in the output map, 3) independent of initial row or column orders.

$$\begin{aligned}
 E_{row} &= \sum_{i=1}^{nrows} \left\{ \sum_{j=1}^{ncols} \left(\sum_{l=-w}^w \frac{1}{2} k_{row-i,j,l} \bullet r_{row-i,j,l}^2 \right) \right\} + \alpha_{row} & \text{where} \\
 E_{col} &= \sum_{i=1}^{ncols} \left\{ \sum_{j=1}^{nrows} \left(\sum_{l=-w}^w \frac{1}{2} k_{col-i,j,l} \bullet r_{col-i,j,l}^2 \right) \right\} + \alpha_{col} & k_{row-i,j,l} = K \sqrt{|v_{i,j} - v_{i+l,j}|} \\
 E &= E_{row} + E_{col} & k_{col-i,j,l} = K \sqrt{|v_{i,j} - v_{i,j+l}|} \\
 & & r_{row-i,j,l} = r_{col-i,j,l} = |l|
 \end{aligned}$$

Formula 1: Energy Definition Formula The map energy E is defined as the sum of the potential row energy E_{row} and the potential column energy E_{col} . r represents distance between 2 rows (columns) and k represents inter-row (-column) coefficient defined by square root of cell values difference x constant K . w is window size, α is marginal adjustment function.

2.2 The Map Arrangement

Since the combination number of row and column arrangements increases extremely when the map size is large, therefore we focused on two points. Firstly, only the energy changes between two neighbors will be calculated for minimization step. Secondary, exchanging positions of the group units consisted of two or more

columns (rows) will bring further effect for high-speed calculating. In order to confirm the effect of this new clustering algorithm, we designed a simple 10 x 10 matrix pattern and treated it using Memin clustering method. As shown in Figure 1 Memin succeeded in making a converged map (c) into the same form as the minimum energy map (b). Moreover, when the value of boundary conditions is changed, it also turned out that the energy minimum type changes (d).

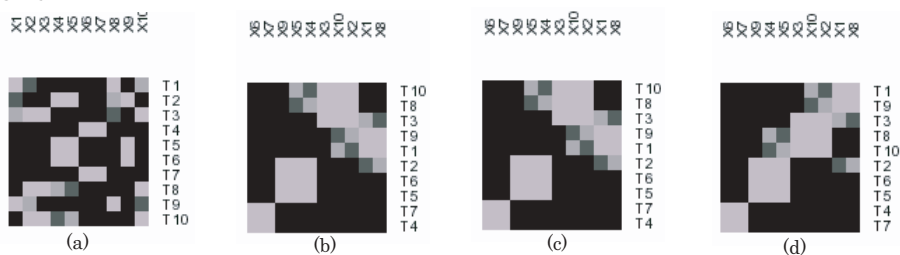


Figure 1: Example of memin simple procession application.

- (a) Initial Map
- (b) The map with the minimum energy (marginal values are -1, w=1)
- (c) Memin convergence result
- (d) The map with the minimum energy (marginal values are 0, w=1)

2.3 Application for the GPCR Homology Clustering

GPCRs (G protein-coupled receptors) are widely expressed in the body and play a key role in physiology. Clustering analysis of GPCRs based on their amino acid sequence homology will show us the functional similarity or ligand type similarity among the same cluster members. Figure 2 (a) shows the similarity score matrix calculated by BLAST. Figure 2 (b) and (c) are the clustering results of average-linkage hierarchical clustering and Memin clustering, respectively. In the hierarchical cluster, galanin and orexin receptors were placed in separated clusters, however, their homology patterns were concatenated via the cluster of NPY receptor in the Memin output.

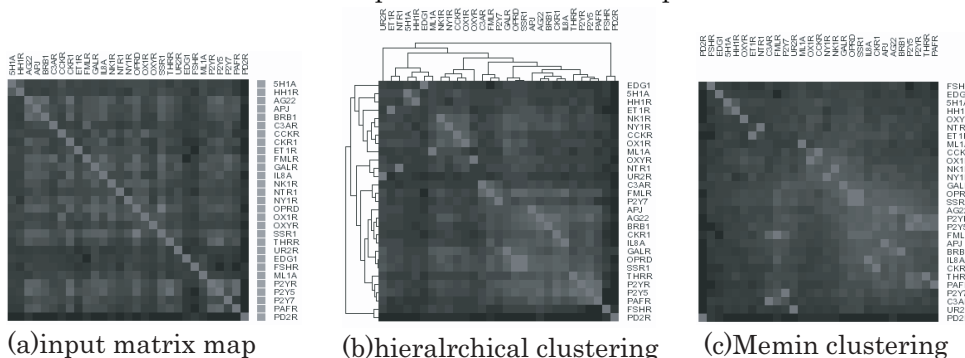


Figure 2: Clustering of GPCR homology score matrix.

3 Discussion

The fact that galanin, orexin, and NPY are all neuropeptides means Memin clustering is biologically meaningful clustering method comparing to the conventional one. This method can also be applied to DNA microarray data analysis. But this method basically has the traveling salesman problem (TSP) with marginal conditions. We consider that some TSP algorithms, filtering spike noises, suitable marginal conditions, window size, and energy definitions should be tested to improve Memin clustering method.

References

- [1] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- [2] Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y., and Weinstein, J.N., Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies, *J. Med. Chem.*, 44:3254–3263, 2001.