P.O.Box 8718, Beijing 100080, China	Journal of Software, Sept. 2005,16(9):1591-1598
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2005 by The Editorial Department of *Journal of Software*

一种基于主集分割的基因芯片聚类算法

滕 莉, 付旭平, 李宏宇, 李 瑶, 陈文斌, 李荣宇, 沈一帆

滕 莉1, 付旭平2, 李宏宇1, 李 瑶2, 陈文斌3, 李荣宇4, 沈一帆1
1(复旦大学 计算机科学与工程系,上海 200433)
2(复旦大学 生命科学学院 遗传研究所,上海 200433)
3(复旦大学 数学系,上海 200433)
4(上海博星基因芯片有限责任公司,上海 200092)
作者简介: 滕莉(1980－),女,山东莒南人,硕士,主要研究领域为生物信息,聚类算法;付旭平(1975－),男,博士,讲师,主要研究领域为分子生物学;李宏宇(1980－),男,博士生,主要研究领域为图像分割,数据聚类;李瑶(1965－),女,博士,教授,博士生导师,主要研究领域为基因芯片技术及其功能基因组;陈文斌(1970－),男,博士,副教授,主要研究领域为并行计算,快速算法;李荣宇(1973－),男,硕士,工程师,主要研究领域为基因芯片生物学;沈一帆(1965－),男,博士,教授,博士生导师,主要研究领域为科学计算,可视化.
联系人: 滕 莉 Phn: +86-852-60801741, E-mail: tengli.hust@263.net, http://www.cse.cuhk.edu.hk/~lteng/

Received 2004-05-31; Accepted 2005-04-04

Abstract
Clustering algorithms are wildly used in the research of microarray data to extract groups of genes or samples that are tightly coexpressed. In most of them, some parameters should be predefined artificially, however, it is very difficult to determine them manually without prior domain knowledge. To handle this problem, an iterative clustering algorithm is proposed. Firstly, by sorting the original data by dominant set, similar genes would be aligned together. It's hard to specify the cluster boundary. A criterion is presented to partition a cluster from the sorted data according to the property that the distances between the inside elements are smaller than that of outside elements. The idea is to remove the cluster form the current data set, repeat the process, and stop the algorithm when the stop criterions are satisfied. The new clustering algorithm is analyzed on several aspects and tested on the published yeast cell-cycle microarray data. The results of the application confirm that the method is very applicable, efficient and has good ability to resist noise.

Teng L, Fu XP, Li HY, Li Y, Chen WB, Li RY, Shen YF. A microarray cluster algorithm based on dominant set segmentation. *Journal of Software*, 2005,16(9):1591-1598.
DOI: 10.1360/jos161591
http://www.jos.org.cn/1000-9825/16/1591.htm

摘要
聚类算法广泛应用于生物芯片数据分析中,用于寻找表达相似的基因或样本.大多数已有算法都需要人为地给出一些参数,然而在没有先验知识的情况下,人为地确定这些参数是十分困难的.为了解决这一难题,提出了一种迭代的聚类算法.首先用主集方法对原有基因进行重新排序,使高度相似的基因排列在特定区域.类的分割界线通常难于确定.提出一种标准,根据类内元素间的距离远小于类外元素间的距离的性质,从排序后的数据集中划分出一个类.将找到的类从当前数据集中排除以后,对剩下的数据重复以上处理,直到满足所提出的循环停止条件为止.从多方面分析了该算法的性能,并将该算法应用于酵母细胞周期的芯片表达谱数据聚类.理论分析和应用结果都表明,该算法是实用、有效的,并且有很好的抗噪性能.

基金项目：Supported by the National Natural Science Foundation of China under Grant No.60473104 (国家自然科学基金)

References:

[1] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Gloub TR. Interpreting patterns of gene expression with

self-organizing maps: Methods and application to hematopoietic differentiation. Proc. of the National Academy of Sciences, USA, 1999,96:2907-2912.

[2] Carr DB, Somogyi R, Michaels G. Templates for looking at gene expression clustering. Statistical Computing & Statistical Graphics Newsletter, 1997,8:20-29.

[3] Eisen MB, Spellman PT, Brown PO, Bottstein D. Cluster analysis and display of genome-wide expression patterns. Proc. of the National Academy of Sciences, USA, 1998,95:14863-14868.

[4] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patters. Bioinformatics, 2001,17:126-136.

[5] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nature Genetics, 1999,22:281-285.

[6] Lukashin AV, Fuchs R. Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics, 2001,17(5):405-414.

[7] Ben-Dor A, Yakhini Z. Clustering gene expression patterns. Journal of Computational Biology, 1999,6:281-297.

[8] Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Research, 1999,9(11):1106-1115.

[9] de Risi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 1997, 278:680-686.

[10] Lander ES. Array of hope. Nature Genetics, 1999,21:3-4.

[11] Schena M, Shalon D, Davis R, Brown P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995,270:467-470.

[12] Sherlock G. Analysis of large-scale gene expression data. Brief Bioinformatics, 2001,2(4):350-362.

[13] Pavan M, Pelillo M. A new graph-theoretic approach to clustering and segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Computer Society, 2003. 98-104. http://www.dsi.unive.it/~pelillo/papers/cvpr03.pdf

[14] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis, RW. A genome wide transcriptional analysis of the mitotic cell cycle. Molecular Cell, 1998,2(1):65-73.

[15] Getz G, Levin E, Domany E, Zhang MQ. Super-Paramagnetic clustering of yeast gene expression profiles. Physics A, 2000,279: 457-464.