

## خوشه‌بندی داده‌های بیان ژنی و کاربرد آن در تحلیل افتراق انواع سرطان خون

محسن واحدی<sup>۱\*</sup> (M.Sc)، حمید علوی مجد<sup>۲</sup> (Ph.D)، یدا... محرابی<sup>۳</sup> (Ph.D)، بهار نقوی<sup>۴</sup> (M.Sc)

-  
-  
-  
-

### چکیده

سابقه و هدف: یکی از شاخه‌های مهم بیوانفورماتیک فناوری ریزآرایه DNA است که امکان بررسی بیان هزاران ژن را به طور هم‌زمان در حداقل زمان ممکن می‌سازد که در سال‌های اخیر موجب تولید حجم انبوهی از داده‌های بیان ژنی شده است. تحلیل آماری این داده‌ها شامل نرمال سازی، خوشه بندی، طبقه بندی و ... از جمله روش‌های مورد استفاده در تحلیل این نوع داده‌ها است.

مواد و روش‌ها: در این مقاله داده‌های بیان ژنی سرطان خون گلوب و همکاران (۱۹۹۹) که بر اساس روش آرایه الیگونوکلوئوتید تولید شده و از طریق اینترنت در اختیار عموم قرار دارد، با استفاده از روش‌های آماری مقیاس‌بندی چند بعدی، خوشه‌بندی سلسله مراتبی و غیر سلسله مراتبی مورد تجزیه و تحلیل قرار گرفته است. مجموعه داده‌ها شامل ۲۰ بیمار مبتلا به سرطان خون لنفوئیدی حاد (ALL) و ۱۴ بیمار مبتلا به سرطان خون میلوئیدی حاد (AML) است. در هر دو روش خوشه‌بندی، داده‌ها به دو خوشه تقسیم شدند. روش‌های مختلف خوشه‌بندی با توجه به گروه‌بندی واقعی نمونه‌ها (AML، ALL) مورد مقایسه قرار گرفتند. نرم افزار R برای تحلیل داده‌ها استفاده شد.

یافته‌ها: ویژگی روش خوشه‌بندی سلسله مراتبی تقسیمی در تشخیص افراد ALL، ۷۵ درصد و حساسیت آن ۹۲ درصد بدست آمد، ویژگی روش افراز کردن اطراف میدوئید در تشخیص افراد ALL، ۹۰ درصد و حساسیت آن ۹۳ درصد بدست آمد که نشان‌دهنده عملکرد خوب این دو روش است. یکی از نمونه‌ها که بر اساس یافته‌های بالینی در گروه AML قرار دارد طبق نتایج تمام روش‌های خوشه‌بندی مورد استفاده در گروه ALL قرار گرفت که از نظر بالینی می‌تواند قابل توجه باشد.

نتیجه‌گیری: با توجه به انطباق قابل توجه نتایج خوشه‌بندی با گروه‌بندی واقعی داده‌ها، می‌توان از این روش‌های آماری در مواردی که اطلاع دقیقی از گروه‌بندی واقعی داده‌ها در دست نیست، استفاده کرد. به علاوه نتایج خوشه‌بندی ممکن است زیرگروه‌هایی از نمونه‌ها را به نحوی متمایز کند که برای انطباق آن با یافته‌های بالینی، پژوهش‌های آزمایشگاهی یا بالینی جدیدی لازم باشد.

واژگان کلیدی: بیوانفورماتیک، ریزآرایه DNA، بیان ژن، خوشه بندی، سرطان خون

### مقدمه

[ ]

( )

(Array)

(Prob)

. mRNA

:

(Complementary

DNA

-

DNA Spotted)

(Oligonucleotide array)

-

. [ ]

)

cDNA

cDNA

(

( )

(Bioinformatics)

(Microarray)

. [ ]

DNA

- ]

. [

. [ ]

(Hierarchical)

(Perfect Match)

PM

. [ ]

MM (MisMatch)

. [ ]

. [ ]

[ ]

MM

/ PM

(ALL)

[ ]

(AML)

(Gene Expression)

[ ]

DNA

R

R

STATA

S-plus

SAS

[ ]

( )

:

:

(

(Class discovery)

- :

- (Gene identification) -

.(Class prediction)

max/min ≤ 5

min max

max - min ≤ 500

[ ]

Archive of SID

(

(AML ALL)

(

[ ]

مواد و روش‌ها

DNA

(

p)

p

Gen Bank

[ ]

(q < p)

q

(Golub)

( )

( )

(Dendrogram)

[ ]

[ ]

(Cophenetic Distances)

(Distance)

(Similarity)

[ ]

(Non

Hierarchical)

$k$

$i$

$X$

$x_{ik}$

$k$

(Partitioning Around

$k$  Medoids)

$k$

PAM

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \xrightarrow{\text{transformed}} D_{n \times n}$$

(Pearson's Correlation Coefficient)

PAM

$k$

$k$

[ ]

$$\rho_{ij} = \frac{\sum_{k=1}^p (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^p (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^p (x_{jk} - \bar{x}_j)^2}}$$

$$d_{ij} = 1 - \rho_{ij}$$

(Divisive).

(Agglomerative)

نتایج

x

)  
(

ALL

/

ALL

ALL

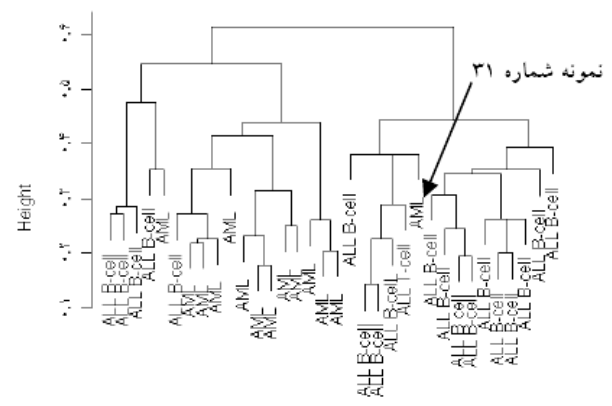
ALL

/

/

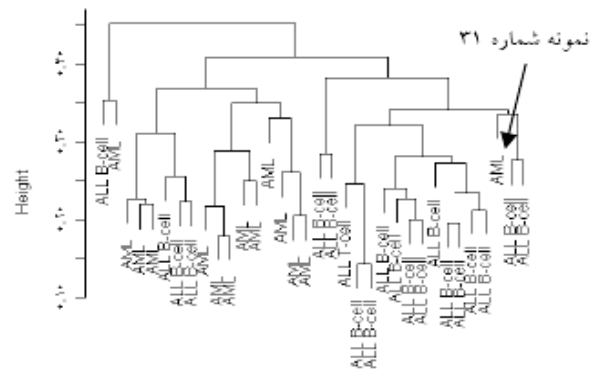
								ALL
								AML
	%	%	%	%	%	%	%	
	%	%	%	%	%	%	%	

Dendrogram for ALL AML data: Coph = 0.98



as dist(d)  
Divisive algorithm, correlation matrix, G= 5.91V genes

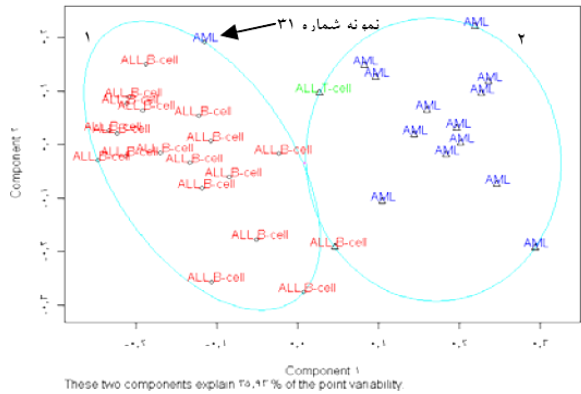
Dendrogram for ALL AML data: Coph = 0.77



as dist(d)  
Average linkage, correlation matrix, G= 6.91V genes

( )

Bivariate cluster plot for ALL AML Correlation matrix, K=2, G= 2,917 genes



( )

**بحث و نتیجه گیری**

)

(

ALL

AML

**تشکر و قدردانی**

AML ALL

**منابع**

[1] Haskell C.M, Berek J. Cancer Treatment. Sounder co, 2001. 5th.Edition PP10-21

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

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

[4] Gershon D. Microarray technology: an array of opportunities. Nature 2002; 416: 885-891.

[5] Bullinger L, Dhner K, Bair E, Frhling S, Schlenk RF, Tibshirani R, and et al. Use of gene expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. The New England Journal of Medicine 2004; 350: 1605-1616.

[6] Valk PJM, Verhaak RGW, Beijen M A, Erpelinck CAJ, Barjesteh S, Waalwijk V, et al. Prognostically Useful Gene-

(

)

AML

ALL

Department of Statistics, Stanford University, 2002. Available from: URL: <http://www-stat.stanford.edu/tibs/research.html>

[13] The R Project for Statistical Computing Available from: URL: <http://www.r-project.org>

[14] National Center for Biotechnology Information Available from: URL: <http://www.ncbi.nlm.nih.gov>

[15] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531–537.

[16] Available from: URL: <http://www.broad.mit.edu/MPR>

[18] Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to cluster Analysis*. Wiley: New York, 2005; 1.Edition : PP 68-279

[19] Sneath P.H. and Sokal R.R. *The principles and practice of numerical classification*. Numerical Taxonomy. W. H. Freeman, San Francisco, 1973: p.278 ff.

Expression Profiles in Acute Myeloid Leukemia. *New Eng J. Med* 2004; 350: 1617-1628.

[7] Satagopan JM, Panageas KS, Tutorial in biostatistics a statistical perspective on gene expression data analysis. *Statist. Med.* 2003; 22:481–499.

[8] Eisen MB, Brown PO. DNA arrays for analysis of gene expression. *Methods Enzymolo* 1999; 303: 179 –205.

[9] Amaratunga D, Cabrera J. *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley & Sons, Ltd, 2004; 1.Edition PP 8-37

[10] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed Optics* 1997; 2: 364–374.

[11] Affymetrix Microarray Suite User Guide. Version 4.0. 2000; Appendix A2, A3.

[12] Efron B, Tibshirani R, Goss V, Chu G. Microarrays and their use in a comparative experiment. Technical report 213,

Archive of SID