非計量多次元尺度構成法によるゲノムコピー数解析

片岡史彬[1]，田口善弘[1,2]

[1]中央大学理工学部物理学科，[2]中央大学理工学研究所

ゲノムコピー数多型は様々な遺伝性疾患の一因である。ゲノムコピー数を調べる実験には多くの生物学的なノイズが載っているが、何の先入観も持たずにこれらのデータからデータの本質を抽出するのは容易ではない。今研究では、非計量多次元尺度構成法(nMDS)を用い、データ間の数値の「距離」に注目することで、フィルタリングせずにコピー数の正常と異常を分けられることを示している。

Analysis of DNA copy number alterations with non-metric multidimensional scaling method

Fumiaki Kataoka[1]　Y-h. Taguchi[1,2]

[1]Dept. Phys., Chuo. Univ., fumiaki108@gmail.com
[2] Dept. Phys., Chuo. Univ., tag@granular.com

DNA copy number alterations are the cause of many genetic diseases. However, there are much noise in the observation of DNA copy number variation (CNV). Thus, it is not easy to estimate there profiles. In this paper, we apply non-metric multidimensional scaling method (nMDS) to estimate the DNA copy number alterations. And to use only copy number distances, we can separate normal genes from anomalous ones without any pre-filtering.

Keywords: CNV (copy number vatiation), nMDS (non-metric multidimensional scaling)

1. Introduction

DNA copy number alterations are the cause of many genetic diseases. However, there are much noise in the observation of DNA copy number variation (CNV). Thus, it is not easy to separate normal genes from anomalous ones, or to characterize anomalous genes. In this paper, we apply non-metric multidimensional scaling method (nMDS) to estimate the DNA copy number alterations observed by array-based comparative genomic hybridization (aCGH).

2.Method and Results

In order to test the ability of nMDS on the estimation of DNA copy number alterations observed by aCGH, we have applied nMDS to those in colorectal cancer [1]. Raw data is taken from ACTuDB [2]. nMDS implemented by us [3] is used for analysis. Euclidean distances between log2 ratio transformed 2074 BAC array gene expression profiles for 125 clients are computed. Then, we get 2D embeddings by nMDS of 2074 BAC array genes (Fig.1). Missing observations are substituted by zero value : normal copy number in plofiles transformed into log2 ratio. Fig.1 shows that normal genes concentrate into the origin, and anomalous genes are lying around.

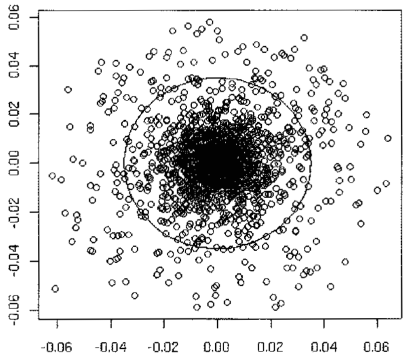

Figure 1 : 2D embeddings by nMDS of 2074 BAC array genes,
and a cercle with radius R=0.035 centered at origin.
Genes within the circle are normal and those outside are abnormal.

To analyze only anomalous genes, we set a circle with radius R centered at origin in Fig.1, and exclude genes within the circle from 2074 BAC array genes. First, to decide this radius R, we draw distribution of distances from the origin in embedded space (Fig.2). Fig.2 looks like the combination of exponential distribution and Gaussian distribution.

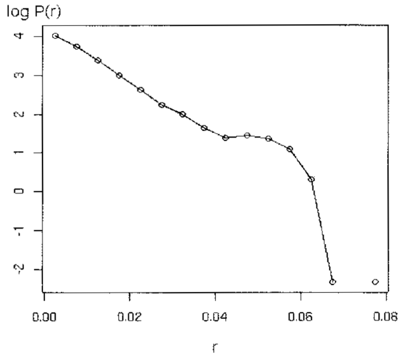

Figure 2 : Distribution of distances from the origin in the embedded space.

Second, we set the distribution function $P_0(r)$ as follows, where r is the distance from the origin, and apply maximum likelihood estimation about parameters “p”, “a”, “m”, and “s” : paramater “p” is ratio of each distributions, “a” is slope of exponetial distribution, “m” is mean of Gaussian distribution, and “s” is standard deviation of Gaussian distribution.

$$P_0(r)=(1-p)a\,e^{(-a|r|)}+p\frac{1}{(\sqrt{2\pi}\,s)}e^{(\frac{-(r-m)^2}{2s^2})}$$

Then, Fig.2 is fitted by this function (Fig.3). Fig.4 shows the first and second terms of $P_0(r)$.

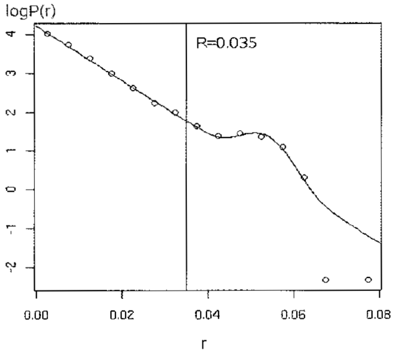

Figure 3 : Fig.2 (P(r) : open circles) fitted by $P_0(r)$ (solid line).
$p=3.28\times10^{-2}$, $a=70.0$
$m=5.24\times10^{-2}$, $s=5.23\times10^{-3}$

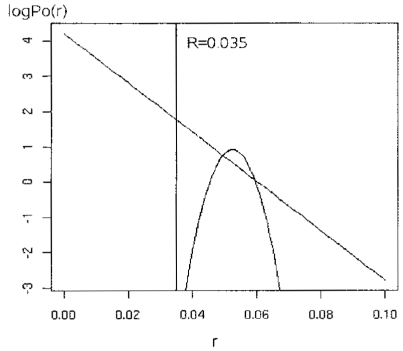

Figure 4 : Exponential distribution part and Gaussian distribution part in $P_0(r)$.

Radius R is determined as 0.035, to take all the Gaussian distribution part. And finally, we get only 225 BAC array genes. About BAC array genes estimated to be with anomalies, we get 2D embeddings only for these 225 genes (Fig.5).We can classify those anomalous genes into four clusters with Mclust in mclust package [4] for R [5].

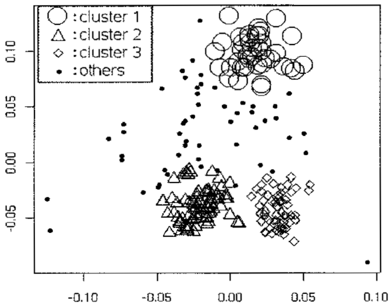

Figure 5 : 2D embeddings by nMDS of 225 BAC array genes.

28 genes out of all 40 genes in the cluster 1 belong to chromosome 18, 56 out of all 74 genes in the cluster 2 belong to chromosome 20, and 55 out of 57 genes in the cluster 3 belong to chromosome 8 (Table 1). In addition, cluster 2 and cluster 3 include more genes with DNA gain than those with DNA loss, while cluster 1 includes more genes with DNA loss than those with DNA gain (Table 2) : DNA gain (loss) is greater (less) than 0.225 (-0.225) [1], and low gain (less) is between 0 and 0.225 (-0.225 and 0). Fig.6 shows distribution of mean DNA copy number of BACs in each cluster.

cluster 1

Chromosome	2	5	6	9	15	16	17	18
Number	1	3	1	1	1	3	2	28

cluster 2

Chromosome	2	7	9	11	13	18	20	23
Number	1	3	1	1	5	1	56	6

cluster 3

Chromosome	3	6	8
Number	1	1	55

Table 1 : Numbers of chromosome in each cluster.

	cluster 1	cluster 2	cluster 3
gain	0	35	11
low gain	0	38	44
low loss	25	1	2
loss	15	0	0
total	40	74	57

Table 2 : Number of DNA gain and loss.

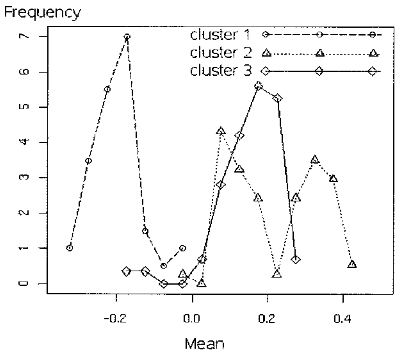

Figure 6 : distribution of mean DNA copy number of BACs in each cluster.

The cluster 2 in Fig.6 has double peaks. This is because mean DNA copy number of BACs in chromosome 20 have difference between the first half and the second half area in chromosome (Fig.7). The horizontal axis of Fig.7 is the order of genes along chromosome 20 : (position of genes [bp])×100)/(full length of those chromosome 20[bp]). We can also see this difference in Fig.5 (Fig.8). In Fig.8, points marked "・" is the first half of means of chromosome 20 in anomalous genes, and marked "+" is means of the second half of them.

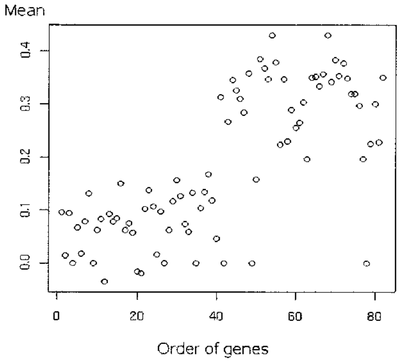

Figure 7 : Mean DNA copy number of BACs in all of chromosome 20.

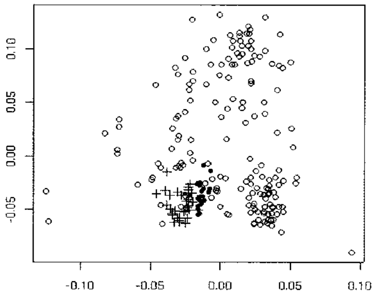

Figure 8 : Difference of means in cluster 2.

Moreover, 225 BAC array genes distribute in cromosome's area as Fig.9. Fig.9 shows that edge areas of chromosome include more anomalous genes than center area.

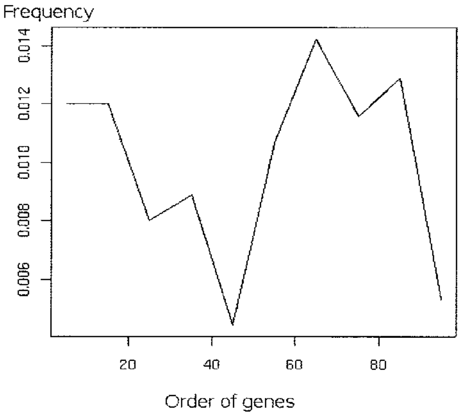

Figure 9 : Frequency of number of anomalous genes in chromosome's area.

Horizontal axis : (position of genes [bp])×100/(full length of those chromosome[bp])

3.Conclusion

Without any pre-filtering, nMDS can correctly distinguish anomalous genes from normal genes, and characterize those anomalous genes. So, it is clear that nMDS has ability to estimate DNA copy number alterations with resistance to noise included into BAC array data.

References

[1] Nakao, K, Mehta, K. R., Fridlyand, J., Moore, D. H., Jain., A. N., Lafuente, A., Wiencke, J. W., Terdiman, J. P., and Waldman, F. M., High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization, Carcinogenesis, 25:1345-1357, 2004.

[2] http://bioinfo-out.curie.fr/actudb/

[3] Taguchi, Y-h., and Oono, Y., Relational patterns of gene expression via non-metric multidimensional scaling analysis, *Bioinformatics*, 21:730-740, 2005.

[4] http://www.stat.washington.edu/mclust

[5] http://www.r-project.org