Vol.2020-MPS-128 No.2
Vol.2020-BIO-62 No.2
2020/6/29

Conjecturing human genes that are easy to be double strand breaks by tensor decomposition.

Nobuo Hosaka[1,a)] Yoshihiro Taguchi[1,b)]

Abstract:
Double-strand break (DSB) is the phenomenon of cutting double-stranded DNA, which can cause gene malfunction. Therefore, it is important to obtain knowledge about DSB because it becomes a great step for the treatment of various diseases caused by gene dysfunction. In this study, we estimated the genes that are easy to be DSBs by tensor decomposition applied to the data about the number of DSBs and histone modifications in the genomes with and without DSB-causing drug.

1. Introduction

Double strand break (DSB) is a phenomenon where double-stranded DNA is cut, which causes mutations such as "insertion", insertion is that other DNA fragments enter the cutting site, "deletion", deletion is that two DNA fragments are broken and the genome in between them is lost, and "fusion", fusion is that the broken site attaches to other chromosomes[1]. If DSBs occur at the positions within the genome where protein coding genes are encoded, and cause mutations, the genes will malfunction and various diseases will occur dependent on the role of the genes. For example, if the gene is involved in cell division and proliferation, it may cause cancer. Nevertheless, we don't know much about DSB, which can result in these dangerous mutations. Therefore, we report here that we have identified genes that are easily to be DSBs. We hope that this report can help future studies on DSB.

2. Dataset used in the analysis

2.1 The first dimension of the tensor:histone modifications

In order to identify DSB-prone genes by tensor analysis, the histone modifications attached to the genomes are used as the first dimension of the tensor. Nevertheless, there are innumerable histone modifications, among which we have to select the one related to DSB because we are interested in the genes that are likely to cause DSB. Therefore, we selected histone modifications that help or promote DSB repair itself[2]. In addition, we added the protein ubiquitin and the enzyme DNA ligase, both of which are involved in DSB[2]. Chip-seq could identify them in human geome. Histone modifications data were downloaded from E-MTAB-5817-ChIPSeq mapping of histone modifications at multiple DNA Double Strand Breaks on the ArrayExpress site*1. Two types of data are present in the histone modifications: proteins, and enzymes. One of which is deliberately caused by the addition of a drug (see below) and another of them is without the drug. The selected histone modifications are described as follows:

(1) ubiquitin
(2) Histone H1.2
(3) Histone H2A.Z
(4) Histone H2A.Z(acetyl K4+K7+K11)
(5) acetyl-Histone H2B (K120)
(6) ubiquityl-Histone H2B (K120)
(7) Histone H3
(8) Histone H3K36me2
(9) Histone H3K36me3
(10) dimethyl-Histone H3 (Lys4)
(11) Histone H3 (acetyl K56)
(12) Histone H3K79me2
(13) Histone H3K9me2
(14) Histone H3 (trimethly K9)
(15) Histone H4 (acetyl K12)
(16) acetyl-Histone H4 (K16)
(17) Histone H4K20me1
(18) Histone H4 (dimethyl K20)
(19) Histone H4 [P S1]
(20) DNA ligase4
(21) Histone MacroH2A.1

The DSB data were mapped to the entire human genome by the BLESS method. The DSB mapping data play an important role in the analysis because it shows whether a genome is prone to DSBs or not by looking at the differences in DSBs depending on the presence or absence of drug in the genome, and it also shows that the 21 types of data de-

[1] Chuo University Faculty of Science and Engineering
[a)] nobulukun@gmail.com
[b)] tag@granular.com

scribed above may be associated with DSBs if the mapping data have large values, regardless the positive or negative values. In order to compare the 21 types of data described above with each genome, 22 types of elements including the DSB mapping data are considered as the first dimension of the tensor. This data is also downloaded from the same URL[*1] as the above data.

2.2 The second dimension of the tensor:drug 4-hydroxy Tamoxifen

DSBs are caused by irradiation, chemicals, and mistakes in DNA replication. 4-hydroxy Tamoxifen is one of the main metabolites of Tamoxifen; and Tamoxifen causes DSB[3]. Thus we use the data of DSB induced by 4-hydroxy Tamoxifen. The set with two elements, one with this drug and another without this drug, is the second dimension of the tensor.

2.3 The third dimension of the tensor:genome

The third dimension is one of 22 chromosome: about three billion bases, which is described below.

3. How to analysis

3.1 Making data

We download 22 data in fastq format, both with and without drugs, from the URL[*1]. Using "Bowtie2",we convert from fastq to sam file. Use index file H.sapiens,NCBI GRCh38 that is downloaded from the index of Bowtie2's homepage[*1] for the item index in "bowtie2". "Bowtie2" is as shown below, but it is slightly different command from other data because only BLESS(DSB data) is paired-end read. We use "samtools" to convert from sam to bam files and sort them with "samtools" to sort the genome in order. Finally, we use "bedtools2" to convert from bam to bed file. This bed file is so large that it can't fit into the memory of R used for analysis so we divide this file into each chromosomes using "grep" command. In order to know how many histone modifications are attached to each base of genome, we use "genomeCoveragebed" command in bedtools2. Enter the reference genome GRCh38 in the [genome] field of the command "genomeCoveragebed". Because this is still large,we use "split" to divide the chromosomes further. The same chromosome, 22 kinds of histone modifications and DSBs with the same division number are read into R by using the "read.table". Delete the first column (chromosome name) and the second column (number of genomes) and leave the third column (number of histone modifications or DSBs). Use the "array" to create a third-order tensor and substitute the number of histone modifications and the number of DSBs into it.

[command line]

```
bowtie2 -x [INDEX] -U FK2_mOHT.fastq -S
    FK2_mOHT.sam
(mOHT=not contain compound. pOHT=contain
    compound.)
bowtie2 -x [INDEX] -1 BLESS_mOHT1.fastq -2
    BLESS_mOHT2.fastq -S BLESS_mOHT.sam
(BLESS only)
samtools view -bS BLESS_mOHT.sam>BLESS_mOHT.bam
samtools sort BLESS_mOHT.bam BLESS_mOHT.sort
bedtools2 bamtobed -i BLESS_mOHT.sort >
    BLESS_mOHT.bed
grep -w "chrXX" BLESS_mOHT.bed >
    BLESS_mOHTchrXX.bed
genomeCoverageBed -d -i BLESS_mOHTchrXX.bed -g
     [genome] > BLESS_mOHTchrXX
split -d -l (N) BLESS_mOHTchrXX
    BLESS_mOHTchrXX_
(N is number that splits chromosomeXX.)
x<-read.table("BLESS_mOHTchrXX_00")
xx<-x[,-c(1,2)]
aaaXX<-array(0,c(22,2,N))
aaaXX[i,j,]<-xx
(First dimention i change to number of 1-22
    depending on proteins or DSB. Second
    dimention j change to number of 1-2
    depending on containing compound or not.)
```

[*1] http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

3.2 Tensor data

The tensors used for the analysis are the third order tensors that are the box composed of histone modification + DSB (22 types), presence or absence of chemicals (2 patterns) and number of genomes (number of divisions N).However, although the size of the histone modifications, enzymes, and proteins themselves are larger than one base, counting the number of modifications per base would be redundant because the position of the histone modifications may be altered by a slight impact or something, so we add the histone modifications, DSBs, etc. in sequential 100 bases to form one block.

[command line]

```
aa=N/100
c<-rep(1:aa,each=100,length.out=N)
bbb<-array(0,c(22,2,aa))
for(i in 1:22){
a1<-split(aaaXX[i,1,],c)
a2<-split(aaaXX[i,2,],c)
a11<-lapply(a1,sum)
a22<-lapply(a2,sum)
for(j in 1:aa){
beta[i,1,j]<-a11[[j]]
beta[i,2,j]<-a22[[j]]}}
```

3.3 Feature extraction by unsupervised learning using tensor decomposition[4]

Tucker decompositon is used and used algorithm is "HOSVD"[4].

$$x_{ijk} = \sum_{l_1=1}^{N} \sum_{l_2=1}^{M} \sum_{l_3=1}^{K} G(l_1, l_2, l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k} \qquad (1)$$

Because tensor data is the third order tensor,Tucker decomposition decomposes tensor data x_{ijk} to core tensor $G(l_1, l_2, l_3)$ and orthogonal three matrices $u_{l_1 i}$, $u_{l_2 j}$, and

u_{l_3k}. In this analysis, variable i expresses histone modification and DSB,variable j expresses whether drug was used or not and variable k expresses the size of genome with the size of 100 bases. The dimension of matrices u_{l_1i}, u_{l_2j}, and u_{l_3k} is decided by the parameters(N,M,K)=(22,2,20).

The goal is to find genomes that are prone to DSB. If we find the parameter l_3 that has the correlation to DSB in matrix u_{l_3k}, we substitute it for matrix u_{l_3k} and get genomic regions k associated with DSB.

To find l_3 mentioned in the above, we first consider the matrix u_{l_1i} obtained by HOSVD (list 1 in our study), then select the DSB mapping data for its parameter i (DSB's parameter is $i = 1$ in our study), and find l_1 that represents DSB well, i.e., the one with the highest absolute value among the l_1 of 1-22. As a result, l_1 is considered to be associated with DSB, since it represents DSB well. Additionally we want to find the parameter that represents bases that changes drastically with the drug or not, because the drug causes DSB so the bases are likely to be DSB. We compare u_{l_21} and u_{l_22},and then choose l_2 that has bigger matrix element's gap than other. Selecting a core tensor G from the list [@data] in Z given previously selected l_1 and l_2, we get a vector $G(l_3)$. Because this vector represents the strength of the relation to u_{l_21} and u_{l_22}, it is representing the occurrence of DSB as well as cause of DSB. Here, the core tensor represents the weights, and the larger the absolute value, the stronger the matrix elements are reflected associated with DSB. Therefore, we can select l_3, which has strong DSB-prone features, and u_{l_3k}, which is assigned the selected l_3, becomes a DSB-prone genome.

We perform a hypothesis testing to confirm whether this genome is a DSB-prone genome or not. The values of normalized u_{l_3k} is assumed follow a normal distribution independently(null hypothesis), and we compute p values attributed to k using a chi-squared distribution (using the "pchisq").

$$p_k = p_{\chi^2}\left[> \sum_{l_3} \left(\frac{u_{l_3k}}{\sigma_{l_3}} \right)^2 \right] \tag{2}$$

p_k is the cumulative probability of the χ-squared distribution if it is greater than or equal to $(\frac{u_{l_3k}}{\sigma_{l_3}})^2$, and σ_{l_3} represents the standard deviation.

Furthermore we apply multiple comparison correction (Benjamini–Hochberg method)[5] by "p.adjust" to above-getted p-values and select k associated with p-values less than or equal to 0.01. From those p-values, we select sequential genomic regions, k, to which signifncat p-values are assigned. Using "diff", we pick up the genes that have two or more consecutive p-values in the order of p-values and identify them as DSB-prone genes. The flowchart of the analysis is shown at the end of this document1.

[command line]

```
hosvdXX<-hosvd(as.tensor(beta),c(22,2,20))
hosvdXX$U[[1]][1,]
hosvdXX$U[[2]]
plot(colSums(hosvdXX$Z@data[chosenl_1,chosenl_2,]**2),t="h")
p<-pchisq(rowSums(scale(hosvdXX$U[[3]][,chosenl_3])**2),degree of freedom,lower.tail=F)
data<-c(1:aa)[p.adjust(p,"BH")<0.01]
r<-diff(data)==1
r1<-which(a==TRUE)
r2<-r1+1
r3<-c(r1,r2)
r4<-sort(r3)
r5<-r4[!duplicated(r4)]
r6<-0+100*data[a5]
(The number 0 changes other number depending on split number and N.)
exmaple:I split chromosome1 to ten equal parts .N is 24,895,643.
split number1 : r6<-0+100*data[a5]
split number2 : r6<-24,895,644+100*data[a5]
```

3.4 Confirmation whether this analysis can choose genes that are prone to DSB or not

In order to confirm whether genes that are prone to DSB are selected or not, Fisher's exact test[6] is performed by comparing the position of the genome selected by this analysis using two data: the data covering the entire human genome of mRNAs where mutations, deletions and insertions that occur after DSB have occurred, and the data of all mRNAs in the entire human genome.

Before perform Fisher's exact test, we download mRNA data; and then we try to make it easy to use. Mutation data are obtained from the mutation description[*2] in the COSMIC mutation data file of the COSMIC site. From the COSMIC data, mRNAs associated with Deletion and Insertion are obtained. At first, we read the COSMIC data into R using the "fread". Next, extract the 22nd line that contains the mutation information. In the case of Deletion, use the "grep" to find the mRNA numbers that contains word Deletion. Insertion is also performed in the same way, and then mutated mRNA's ENST names are obtained each mutations by entering mRNA numbers into column of the object in which the COSMIC data was entered by "fread".

[command line]

```
ab<-fread("COSMIC data")
ab1<-ab[,22]
ab2<-grep("Deletion",ab1)
ab3<-grep("deletion",ab1)
ab4<-c(ab2,ab3)
ab5<-ab1[ab4,2]
ab6<-str_sub(ab5,start=1,end=15)
```

We have got mutated mRNA's ENST names; and then we make the original mutation data by entering it for the all mRNA data. The data includes four elements; ENST name, chromosome number, beginning and ending of gene. To compare with them, we also make the mRNA data that is almost same above data; but here has all mRNAs. At first,

[*2] https://cancer.sanger.ac.uk/cosmic/download

we download the all mRNA data from the table browser[*3] of the UCSC genome browser, and we read the UCSC data into R by using "fread". For getting the original mutation data; at first, we get mutated mRNA's ENST names number in UCSC data(using "match"); and then we enter it into UCSC data object; consequently, we get four elements about mutated mRNA. We also get the mRNA data but this is UCSC data itself.

[command line]

```
ac<-fread("UCSC data")
ac1<-ac[,1]
ac2<-str_sub(ac1,start=1,end=15)
ac3<-match(ab6,ac2)
ac4<-ac3[!is.na(ac3)]
deletion<-ac[ac4,c(1,2,4,5)]
insertion<-ac[vv,c(1,2,4,5)]
rna<-ac[,c(1,2,4,5)]
```

Using the original mutation data and the mRNA data obtained above, the following a,b,c,d can be obtained. Let a denote the amount of mRNAs in the analysed region, b denote the amount of mRNAs in the region that have undergone either deletion or insertion, c denotes the amount of mRNAs that have one or more genome numbers k selected by feature extraction between beginning and ending of a mRNA about all mRNAs, d is almost same for c but investigation range of mRNA is different; d's range changes all mRNAs to mutated mRNAs. Using these four parameters, we give the p value by the following Fisher's exact test.

$$p = \frac{(a-b)!(b)!(a-c)!(c)!}{(a)!(a-b-c+d)!(c-d)!(b-d)!(d)!} \tag{3}$$

In order to reduce calculation time and errors, we split two datum each chromosomes to make short (using "which"). We try to get c. At first, we choose a mRNA and we find out genomic regions matching k between beginning and ending of it(using "match"). Next, "match" function gives NA when matching nothing so we delete NA(using "is.na") from the object of "match". As a result, we find out whether feature extraction catches the mRNA or not because sum of the object bigger than zero when feature extraction catches the mRNA. We perform these operations for all mRNAs(using "for" sentence) then we get c. If we do for all mutated mRNAs, then we get d. But the mRNAs related to insertion and deletion have duplication each other. So we perform above operations indivisually for each mRNAs related to insertion and deletion and sum of them is d.

Next, we try to get a and b. At the beginning, we splitted data to read R and that splitted range is the analysed region. And it represents that there is mRNA in the region when the region include either mRNA's beginning or ending. So we search whether include either beginning or ending or not about all mRNA(using "which"), we get a. Here, if we do about all mutated mRNA, we get b. But, as with d, we must perform that individually and sum of them. Using a, b, c, and d, we construct a matrix m and perform Fisher's exact test with the "fisher.test".

$$m = \begin{bmatrix} a-b-c+d & c-d \\ b-d & d \end{bmatrix} \tag{4}$$

[command line]

```
d1<-deletion[,2]
d2<-which(d1=="chrXX")
del<-deletion[d2,]
t1<-rna[,2]
t2<-which(t1=="chrXX")
total<-rna[t2,]
s<-numeric("total mRNAs")
for(i in 1:"total mRNAs"){
x<-total[i,3]
y<-total[i,4]
z<-match(r6,c(x:y))
s[i]<-sum(z[!is.na(z)])}
s1<-which(s!=0)
sss<-as.numeric(total[,3])
sss1<-which(BEGIN<sss&sss<END)
sss<-as.numeric(total[,4])
sss2<-which(BEGIN<sss&sss<END)
sss3<-c(sss1,sss2)
sss4<-sss3[!duplicated(sss3)]
m<-matrix(c(a-b-c+d,c-d,b-d,d),ncol=2)
fisher.test(m)
```

4. Results

4.1 Result of fisher's exact test

Because of the large number of bases, we divided chromosomes 1-22 into 10 equal parts for 1-9 and 5 equal parts for 10-22. Since we want to see the DSB, we exclude the three data for which all parameters l_1 for the DSB are significantly low. Consequently we obtained 152 Fisher's exact test results. Of the 152 cases, 67 cases have p=1, 48 cases have $0.05 < p < 1$, and 37 cases have $p < 0.05$. When we corrected for multiple comparisons again for the 152 cases, we found that 24 cases out of the 37 cases that have p-value less than 0.05 are not coincidental, but chosen by the analysis. Since it is not coincidental that 24 events occur less than 5 percent of the 152 cases, we could infer whether DSB is likely to occur in mRNA-making genes by unsupervised learning using tensor decomposition.

5. Discussions

In Japan, malignant tumors (cancer) have been the leading cause of death among Japanese since 1981. Therefore, it is very important to establish a cure for cancer from the point of view of public health. In order to establish a treatment method for cancer, it is essential to obtain detailed knowledge about the disease. We hope to find as many cancer-prone genes as possible because we have found that DSB is one of the main causes of genetic mutations that cause cancer. In this study, however, we limited our guesses to the genes that generate mRNAs. Therefore, we believe that it is important to further extend this study to include non-coding RNAs and to make inferences on all genes. Fur-

[*3] https://genome.ucsc.edu/cgi-bin/hgTables

thermore, by comparing 22 protein combinations in the first matrix $u_{l_1 i}$ of 27 p values for which the genes with low p-values and frequent DSB predisposition are selected by Fisher's exact test, we can see the protein patterns when DSB predisposition occurs. Depending on the results of the comparison, we may be able to infer whether DSB is likely to occur or not just by looking at the protein combinations, and we would like to do this in the future.

References

[1] Cannan, W. J. and Pederson, D. S.: Mechanisms and consequences of double-strand DNA break formation in chromatin, *Journal of cellular physiology*, Vol. 231, No. 1, pp. 3–14 (2016).

[2] Clouaire, T., Rocher, V., Lashgari, A., Arnould, C., Aguirrebengoa, M., Biernacka, A., Skrzypczak, M., Aymard, F., Fongang, B., Dojer, N. et al.: Comprehensive mapping of histone modifications at DNA double-strand breaks deciphers repair pathway chromatin signatures, *Molecular cell*, Vol. 72, No. 2, pp. 250–262 (2018).

[3] Wozniak, K., Kolacinska, A., Blasinska-Morawiec, M., Morawiec-Bajda, A., Morawiec, Z., Zadrozny, M. and Blasiak, J.: The DNA-damaging potential of tamoxifen in breast cancer and normal cells, *Archives of toxicology*, Vol. 81, No. 7, pp. 519–527 (2007).

[4] Taguchi, Y.-H.: Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes, *BMC bioinformatics*, Vol. 19, No. 4, p. 99 (2018).

[5] Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300 (1995).

[6] Fisher, R. A.: On the interpretation of χ 2 from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society*, Vol. 85, No. 1, pp. 87–94 (1922).

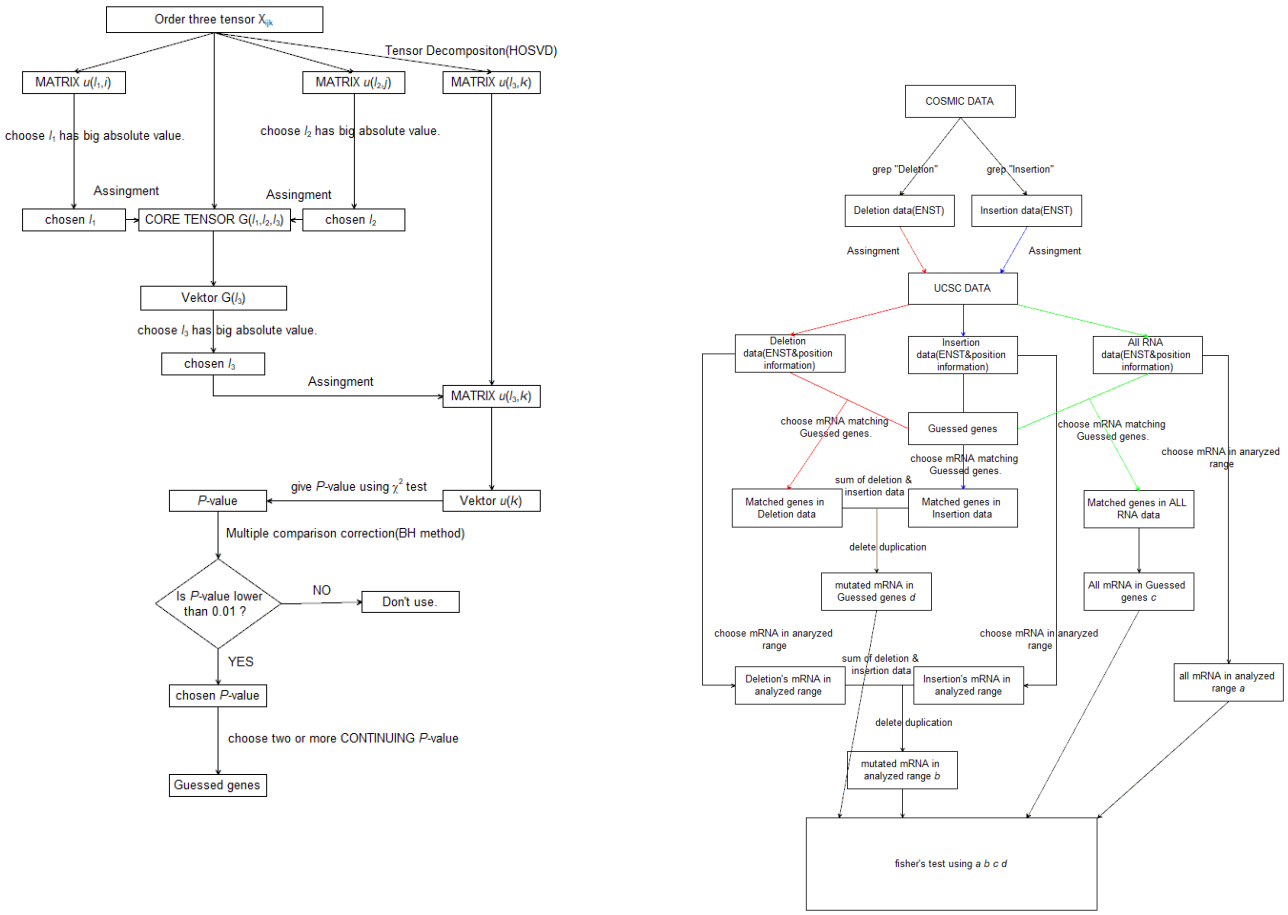

Fig. 1 Flow chart about analysis