Vol.2021-MPS-133 No.27
Vol.2021-BIO-66 No.27
2021/6/30

Identification of Enhancers and Promoters in the Genome by Multi-Dimensional Scaling

Ryo Ishibashi[1,a)] Y-h,Taguchi[1,b)]

Abstract: The position of the enhancers on genomic DNA remains poorly understood. During phase of cell division cycle, since genome forms chromatin structure and spreads within the nucleus, can not observe chromosomes. There is method called Hi-C (High - throughput chromosome conformation capture) that measures physical interaction of genomes. In previous studies, DNA extrusion loops were directly derived from Hi-C heat maps. In this paper, we show that MDS (Multi-Dimensional Scaling) can be used to locate enhancers and promoters more precisely.

1. Introduction

In order for cells to utilize genetic information, many genes must be expressed in a coordinated manner. The accessibility of genomic information depends on how the DNA is packed into chromatin. Chromatin is the basis for a variety of biological processes, including the regulation of the cell cycle and the replication, repair, and maintenance of DNA [2]. There is a region called euchromatin that consists of DNA that has a relatively loose genomic structure. The loose structure allows RNA polymerase and other proteins involved in transcription to DNA. It is also known that enhancers and promoters can be approached in the euchromatin region in order to form DNA loops. Gene expression is controlled by promotors located near the gene and gene regulatory sites called enhancers located from the gene. However, how these interact with each other to regulate gene expression remain understood. Hi-C analyze the 3D structure of a genome by detecting genomic regions that are spatially close to each other using Next Generation Sequencing (NGS) [3]. The conventional method led to an approximation of the genome structure from the Hi-C heat map [4]. In this paper, we show that it is possible to locate enhancers and promotors by applying Multi-Dimensional Scaling (MDS) .

2. Data and Methods

2.1 Multi-Dimensional Scaling

MDS is the method that reproduces the original location of objects from the distance data between objects. The principle is as follow. Consider an $N \times P$ data matrix $\boldsymbol{X} = (x_{ij})$, N data $\boldsymbol{o}_i = (x_{i1}, x_{i2}, ..., x_{iP})$. Then consider the $N \times N$ matrix $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}^t$ and the $N \times N$ distance matrix $\boldsymbol{D} = (||\boldsymbol{o}_i - \boldsymbol{o}_j||)$ created from $\boldsymbol{X}$. Futhermore, we define $\boldsymbol{D}^{(2)}$ as the matrix of all components of the distance matrix D squared, and multiply $\boldsymbol{D}^{(2)}$ by the $N \times N$ centralization matrix $\boldsymbol{J}(= \boldsymbol{E} - \frac{1}{N}\boldsymbol{1})$ from both sides.

$$\begin{aligned} -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}^{(2)}\boldsymbol{J}^t &= -\frac{1}{2}\boldsymbol{J}\{\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^t)\boldsymbol{1} - 2\boldsymbol{X}\boldsymbol{X}^t - \boldsymbol{1}\,\mathrm{diag}(\boldsymbol{X}\boldsymbol{X}^t)\}\boldsymbol{J}^t \\ &= \boldsymbol{J}\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{X}^t = \boldsymbol{J}\boldsymbol{X}(\boldsymbol{J}\boldsymbol{X})^t \\ &= \boldsymbol{X}^*\boldsymbol{X}^{*t} = \boldsymbol{B}_{cen} \end{aligned} \tag{1}$$

$\boldsymbol{X}^*$ is the centered data matrix and $\boldsymbol{B}_{cen}$ is the inner product matrix obtained from the centered data matrix.

$$\boldsymbol{G}^t\boldsymbol{B}_{cen}\boldsymbol{G} = \begin{pmatrix} \lambda_1 & & & & & \\ & \lambda_2 & & 0 & & \\ & & \ddots & & & \\ & & & \lambda_P & & \\ & & & & 0 & \\ & 0 & & & & \ddots \\ & & & & & & 0 \end{pmatrix} \tag{2}$$

$$\begin{aligned} \boldsymbol{B}_{cen} = \boldsymbol{G}^t &\begin{pmatrix} \sqrt{\lambda_1} & & & & & \\ & \sqrt{\lambda_2} & & 0 & & \\ & & \ddots & & & \\ & & & \sqrt{\lambda_P} & & \\ & & & & 0 & \\ & 0 & & & & \ddots \\ & & & & & & 0 \end{pmatrix} \\ &\times \begin{pmatrix} \sqrt{\lambda_1} & & & & & \\ & \sqrt{\lambda_2} & & 0 & & \\ & & \ddots & & & \\ & & & \sqrt{\lambda_P} & & \\ & & & & 0 & \\ & 0 & & & & \ddots \\ & & & & & & 0 \end{pmatrix} \boldsymbol{G} \end{aligned} \tag{3}$$

[1] Chuo University Faculty of Science and Engineering
[a)] a17.t34a@g.chuo-u.ac.jp
[b)] tag@granular.com

ⓒ 2021 Information Processing Society of Japan

$\boldsymbol{G}$ is the orthogonal matrix of the inner matrix $\boldsymbol{B}_{cen}$. From the (1) and (3)

$$\boldsymbol{X}* = \boldsymbol{G}^t \begin{pmatrix} \sqrt{\lambda_1} & & & & & \\ & \sqrt{\lambda_2} & & 0 & & \\ & & \ddots & & & \\ & & & \sqrt{\lambda_P} & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{pmatrix} \tag{4}$$

From the above, we can find the original coordinates.

2.2 Hi-C data

The Hi-C dataset is a useful and powerful tool for understanding how chromatin is organized in the nucleus to effectively carry out its biological functions; that is, it allows us to examine the physical interaction of DNA loops. We downloaded 8 Hi-C data during cell cycles : 0 min (Prometaphase) and 35 min (Anaphase/Telophase), 60 min (Cytokinesis), 90 min, 120 min, 180 min, 240 min, 360 min (G1) from Series GSE141067 [3]. In this study, we analyzed Hi-C data with a resolution of 50 kbp, and found that the genomic long-range interactions were reshaped from 60 min and completed in 90 min - 120 min [3]. We assigned to the missing values the average number of Hi-C detections for each distance between genomic coordinates. In addition, because the Hi-C data handled in this study had a large difference between the number of detections with a far distance between coordinates and the number of detections with a close distance, the following processing was performed dependent upon the distance between nucleotide.

$$d_{ij}\text{new} = \begin{cases} d_{ij} \quad , \mid i-j \mid \leq 5 \\ d_{ij} \times \log \mid i-j \mid , \ \mid i-j \mid > 5 \end{cases} \tag{5}$$

where i, j are coordinates and d_{ij} is the number of Hi-C detections (Figs. 1 and 2). Figure 3 shows the heat map of the Hi-C data, where the genomes are close to each other if the number of Hi-C detections is large. In order to apply the multi-dimensional scale construction method, the Hi-C data was transformed as follows.

$$D_{ij} = \frac{1}{d_{ij}\text{new}} \tag{6}$$

2.3 Applying MDS to the Hi-C data

I applied MDS to the distance matrix D

$$\boldsymbol{D} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^t \tag{7}$$

It is not clear that which column of the orthogonal matrix $\boldsymbol{U}$ presents the desired structure. But in many cases, it is believed that columns 2 and 3 of the orthogonal matrix $\boldsymbol{U}$ present the desired structure. That is why, I selected columns 2 and 3 of the orthogonal matrix $\boldsymbol{U}$ as the desired structure. In this study, we call the structure of the second and third columns of orthogonal matrix $\boldsymbol{U}$ a pseudo

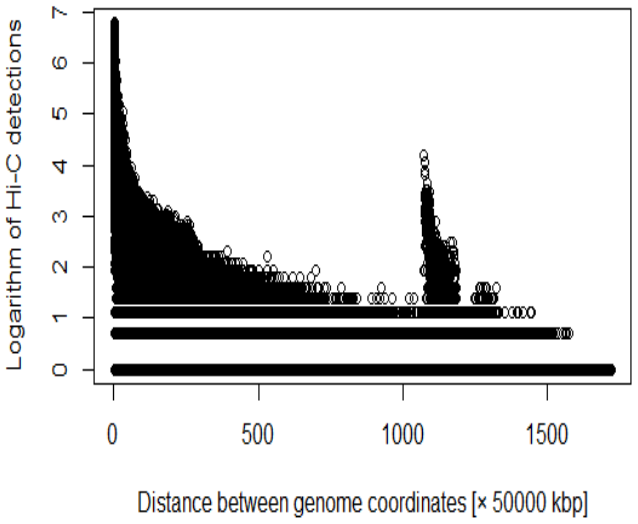

Fig. 1 Plot of distance between coordinates versus logarithm of Hi-C detections before adding weights

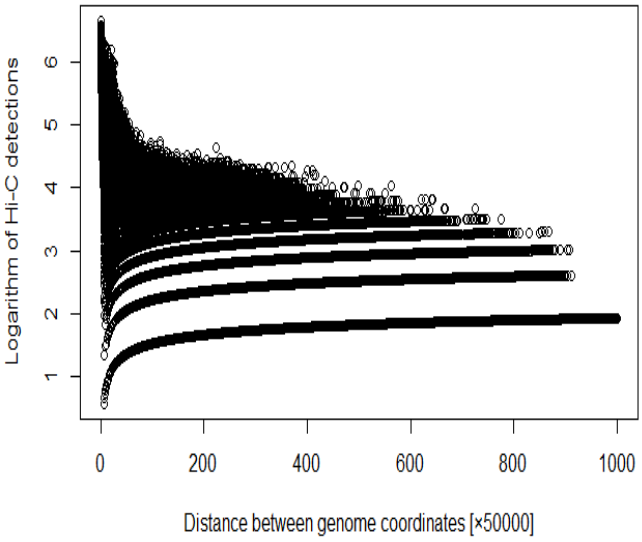

Fig. 2 Plot of distance between coordinates versus logarithm of Hi-C detections after adding weights

chromosome.

It then acquires the euchromatin region where the enhancer and promoter are thought to be located. Find the distance between the coordinates of the hypothetical chromosomes by Euclidean distance.

$$E_i = \sqrt{(U_{i+1,1} - U_{i1})^2 + (U_{i+1,2} - U_{i2})^2} \tag{8}$$

The length of the DNA loops in the euchromatin region is not well understood, but in this study, we averaged the E_i every 50 kbp and used it again as the distance between

ⓒ 2021 Information Processing Society of Japan

coordinates. The distance between coordinates is reduced because it is averaged every 50 kbp. For this reason, we included averages taken every 45 kbp, 40 kbp, 40 kbp, and 35 kbp from both sides.

$$E_i\text{new} = \begin{cases} \frac{\Sigma_{j=1}^{k} E_{i+j}}{k} & , i = 0, k = 5, 6, ..., 9 \\ \frac{\Sigma_{j=0}^{9} E_{i+j}}{10} & , i = 1, 2, ..., N-1 \\ \frac{\Sigma_{j=0}^{k} E_{i+j}}{k+1} & , i = N-9, k = 8, 7, ...4 \end{cases} \quad (9)$$

Five times the average of the distance between these genomic coordinates, was set as the threshold, and the coordinates of $E_i$new above the threshold were acquired as euchromatin regions (Fig. 5).

2.4 Enrichment analysis

I used BiomaRt to retrieve genes from the obtained coordinates. The gene list obtained by BiomaRt is enriched with g:profiler [11] to find functions, processes, and transcription factors related to enhancers and promoters. The coordinates acquired as euchromatin regions were subjected to enrichment analysis using g:profiler.

3. Result

3.1 Hypothetical chromosomes.

The heat map of the Hi-C data after organizing the Hi-C data by (5) is shown in Fig. 3. Pairs with large values in the matrix mean that they are region pairs with high contact probability. MDS was applied to the Hi-C data, and the resulting tentative chromosomes are shown in Fig. 4. Then, the euchromatin region was acquired.

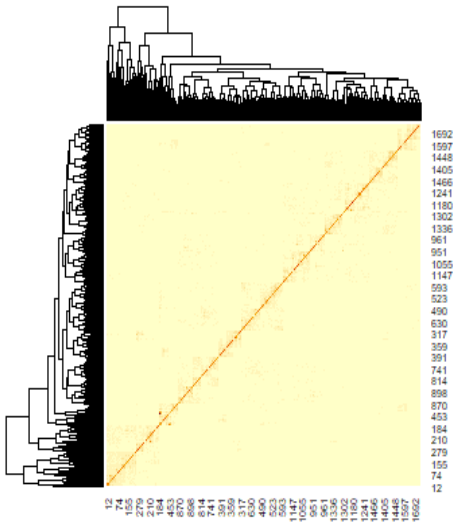

Fig. 3 Heat map of Hi-C data after adding weighting

Finally, the obtained euchromatin regions were subjected to enrichment analysis using g:Profiler. The results are shown in Table 1. The functions and processes involved in transcription were obtained.

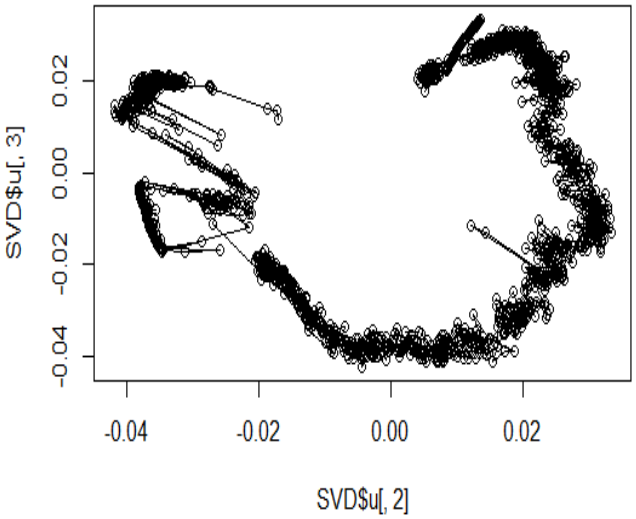

Fig. 4 Plot of hypothetical chromosomes 18 (0 bp - 86000 kbp)

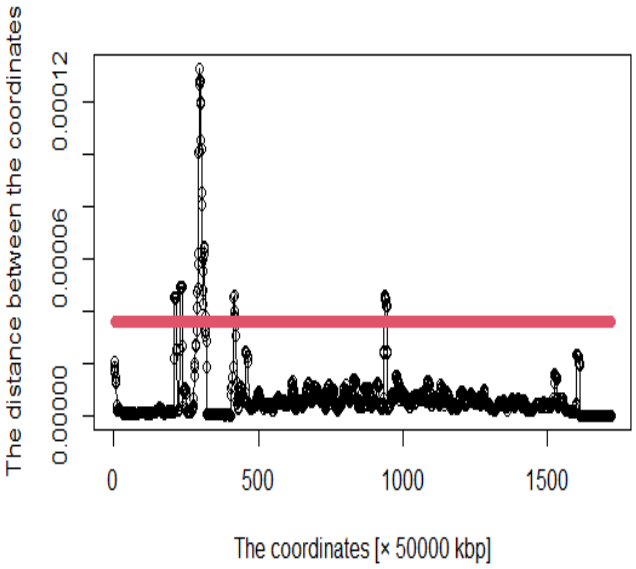

Fig. 5 Distance plot between coordinates (The red line is the threshold.)

CREB3 (CAMP Responsive Element Binding Protein 3) encodes a transcription factor that is a member of the leucine zipper family of DNA binding proteins. This protein binds to the cAMP-responsive element and regulates cell proliferation. The mRNA expression of CREB3 was higher in OS tissue [1]. FOXM1 is known to play an important role in cell cycle progression, and the expression of endogenous FOXM1 peaks at S phase and G2/M phase [5]. FOXM1 upregulation was subsequently found in the majority of solid human cancers [6]. RARA and RAR γ are genes involved in osteosarcoma and germ cell tumors, respectively;

ⓒ 2021 Information Processing Society of Japan

Table. 1 Results of enrichment analysis of 90 min Hi-C data by g:Profiler

source	term_name	term_id	adjusted_p_value
GO:BP	transcription preinitiation complex assembly	GO:0070897	8.80E-06
GO:BP	RNA polymerase II preinitiation complex assembly	GO:0051123	0.000171455
GO:CC	immunoglobulin complex	GO:0019814	2.59E-40
GO:CC	immunoglobulin complex, circulating	GO:0042571	1.13E-08
GO:CC	transcription factor TFIID complex	GO:0005669	0.000436621
GO:CC	RNA polymerase II, holoenzyme	GO:0016591	0.008361059
GO:CC	nuclear DNA-directed RNA polymerase complex	GO:0055029	0.011479167
GO:CC	DNA-directed RNA polymerase complex	GO:0000428	0.012838369
GO:CC	RNA polymerase complex	GO:0030880	0.019811323
TF	Factor: CREB3; motif: NTGCCACGTCAYCN	TF:M04207	0.000406132
TF	Factor: Foxm1; motif: NTGTTTRT	TF:M07255	0.001192314
TF	Factor: RARA; motif: GAGGTCAAAAGGTCAAKK; match class: 1	TF:M08018_1	0.018729697
TF	Factor: PR; motif: NNNNNNRGNACNNKNTGTTCTNNNNNN; match class: 1	TF:M00957_1	0.028864188
TF	Factor: MafG; motif: CMATGACTCAGCAGA; match class: 1	TF:M07048_1	0.033770646
TF	Factor: ER-alpha; motif: TGACCYN; match class: 1	TF:M03547_1	0.039840168
TF	Factor: AML2; motif: TGTGGTNNN	TF:M07372	0.041987127

RAR binds to ligands as heterodimers to target response elements.On ligand binding, the corepressors dissociate from the receptors and associate with the coactivators leading to transcriptional activation. In the absence of ligand, the RXR-RAR heterodimers associate with a multiprotein complex containing transcription corepressors that induce histone deacetylation, chromatin condensation and transcriptional suppression [7]. MAFG (v-maf avian musculoaponeurotic fibrosarcoma oncogene homolog G) is a small MAF protein belonging to the basic leucine zipper (bZIP) family of transcription factors. It is known that MAFG overexpression is important for osteosarcoma cell growth [9]. AML2 (RUNX Family Transcription Factor 3) is a Protein Coding gene. It is known that down-regulation of AML2 occurs in the early stages of the carcinogenic process [10].

4. Discussion

In existing studies, the 3D structure of the genome and DNA extrusion loops have been obtained only from Hi-C heat maps [4]. It is obvious that the higher the resolution of Hi-C, the more difficult it is to obtain complex DNA interactions from heat maps. MDS can reproduce DNA extrusion loops with simple matrix factorization. As evidence of this, as shown in the results, we have obtained many transcription factors involved in transcription and cancer, including osteosarcoma. However, in this study, we projected the genome to two dimensions instead of three. This is a fault of MDS in that it is not clear which column of the orthogonal matrix reproduces the desired structure. Of course, since DNA extrusion loops are thought to exist in the orthogonal direction of the two-dimensional plane we considered, we believe that analyzing them in three dimensions is a future task.

References

[1] Wu, Y., Xie, Z., Chen, J., Chen, J., Ni, W., Ma, Y., Huang, K., Wang, G., Wang, J., Ma, J., Shen, S., & Fan, S. (2019). Circular RNA circTADA2A promotes osteosarcoma progression and metastasis by sponging miR-203a-3p and regulating CREB3 expression. Molecular cancer, 18(1), 73. available from ⟨https://doi.org/10.1186/s12943-019-1007-1⟩

[2] Dekker, J., Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell, 164(6), 1110–1121. available from ⟨https://doi.org/10.1016/j.cell.2016.02.007⟩

[3] Kang H, Shokhirev MN, Xu Z, Chandran S, Dixon JR, Hetzer MW. Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. Genes Dev. 2020 Jul 1;34(13-14):913-930. doi: 10.1101/gad.335794.119. Epub 2020 Jun 4. PMID: 32499403; PMCID: PMC7328517.

[4] Mota-Gómez, I., Lupiáñez, D. G. (2019). A (3D-Nuclear) Space Odyssey: Making Sense of Hi-C Maps. Genes, 10(6), 415. available from ⟨https://doi.org/10.3390/genes10060415⟩

[5] Wierstra I, Alves J. FOXM1, a typical proliferation-associated transcription factor. Biol Chem. 2007 Dec;388(12):1257-74. doi: 10.1515/BC.2007.159. PMID: 18020943.

[6] Liu M, Dai B, Kang SH, Ban K, Huang FJ, Lang FF, Aldape KD, Xie TX, Pelloski CE, Xie K, Sawaya R, Huang S. FoxM1B is overexpressed in human glioblastomas and critically regulates the tumorigenicity of glioma cells. Cancer Res. 2006 Apr 1;66(7):3593-602. doi: 10.1158/0008-5472.CAN-05-2912. PMID: 16585184.

[7] Luo, Yi et al. "Pivotal regulatory network and genes in osteosarcoma." Archives of medical science : AMS vol. 9,3 (2013): 569-75. available from ⟨doi:10.5114/aoms.2012.30956⟩

[8] Dohi, O., Hatori, M., Suzuki, T., Ono, K., Hosaka, M., Akahira, J. - i., Miki, Y., Nagasaki, S., Itoi, E. and Sasano, H. (2008), Sex steroid receptors expression and hormone - induced cell proliferation in human osteosarcoma. Cancer Science, 99: 518-523. available from ⟨https://doi.org/10.1111/j.1349-7006.2007.00673.x⟩

[9] Hua-jian Shan, Lun-qing Zhu, Chen Yao, Zhi-qing Zhang, Yuan-yuan Liu, Qin Jiang, Xiao-zhong Zhou, Xiao-dong Wang, Cong Cao, MAFG-driven osteosarcoma cell progression is inhibited by a novel miRNA miR-4660, Molecular Therapy - Nucleic Acids, Volume 24, 2021, Pages 385-402, ISSN 2162-2531, available from ⟨https://doi.org/10.1016/j.omtn.2021.03.006⟩

[10] Qing-Lin Li, Kosei Ito, Chohei Sakakura, Hiroshi Fukamachi, Ken-ichi Inoue, Xin-Zi Chi, Kwang-Youl Lee, Shintaro Nomura, Chang-Woo Lee, Sang-Bae Han, Hwan-Mook Kim, Wun-Jae Kim, Hiromitsu Yamamoto, Namiko Yamashita, Takashi Yano, Toshio Ikeda, Shigeyoshi Itohara, Johji Inazawa, Tatsuo Abe, Akeo Hagiwara, Hisakazu Yamagishi, Asako Ooe, Atsushi Kaneda, Takashi Sugimura, Toshikazu Ushijima, Suk-Chul Bae, Yoshiaki Ito, Causal Relationship between the Loss of RUNX3 Expression and Gastric Cancer, Cell, Volume 109, Issue 1, 2002, Pages 113-124, ISSN 0092-8674, available from ⟨https://doi.org/10.1016/S0092-8674(02)00690-6⟩

[11] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo: g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) Nucleic Acids Research 2019; available from ⟨https://doi.org/10.1093/nar/gkz369⟩

ⓒ 2021 Information Processing Society of Japan