1. title

2. Study of spatial conformation of chromosomes is of high importance in the field of computational biology. Although all cell in a living being have the same sequence of genes, it is the 3D positioning of these genes in space that determines how the cell functions. Roughly said, if two genes are close to each other in space, they can interact with each other in order to create a certain protein that regulates a certain task.

3. However, this spatial organization of chromosomes cannot be observed through traditional microscopy. As an alternative, Hi-C has emerged as a powerful method for studying the 3D organization of chromosomes in space.

4. high-throughput chromosome conformation capture, or HiC, captures interactions between chromosomal fragments. In this method, a chromosome is divided into very small equally sized sections called loci, then measures all pair-wise interaction frequencies across all chromosomes. Hi-C data are usually provided as a N × N heatmap or contact matrix where N is the number of loci in the genome.

5. Here is an example of HiC contact matrix.

6. Row and column counts are number of loci.

7. Each cell in the heatmap indicates the number of interactions found between a pair of loci corresponding to the rows and columns.

8. ‘Resolution’ of a Hi-C data is the size of the loci the genome is divided into, which decides the detail level of interaction the contact matrix can capture.

9. Obviously, more high resolution data are desirable. However, the sequencing process is costly. Linear increase of resolution requires quadratic increase of sequencing reads. thus, most of the Hi-C data available have low resolutions. Therefore, a computational method be developed to improve the resolution of currently available Hi-C data.

10. The goal of our CNN model is to enhance the contrast of low resolution data. The idea is that a submatrix of high resolution data has a connection with itself and its neighborhood in low resolution contact matrix, and we use CNN to find it.

11. For example, we have submatrices representing the same part of chromosome in low and high resolution.

12. We will use the low resolution data of the submatrix as feature to predict the center region of its counterpart in high resolution data. We call this “patch to patch” prediction, which is unlike in classification problems returns single response.

13. We have contact matrices of 4 cells. One of them are available in both low and high resolution. We used it as training data, and enhanced the rest with the model we trained.

14. In this model we set submatrix feature size at 40 by 40, and predicated the center region at size 28 by 28, which will be returned as a vector of size 784. To do so we divided contact matrix into 40 by 40 patches. This gave us roughly 20000 training samples.

15. We applied 3 convolutional layers in the model with no pooling and padding. Stride is one and Relu as activation function.

16. The setting of all layers is shown in this table.(stop for 5 secs)

17. This is a visualization of our model.

18. Mean square of differences is used as loss function. Since every response contains 28 by 28 pixels, it will be divided by 784.

19. The model is developed on theano framework with lasagne. Here is a comparison of low resolution, high resolution and predicted contact matrix. The model clearly enhanced the contrast. Some local patterns that is not recognizable in low resolution were effectively captured by our CNN and presented in the response matrix.

20. Then we applied the model into the other cells. As you can see, an obvious increase in contrast is achieved and the contact matrices are smoother.

21. There is no measure of accuracy for such models. All we can contend to is the result be adequately smooth and also not different from the original matrix in terms of pattern.

22. In this project we developed a CNN model to generate high resolution contact matrix of chromosome based on its low resolution version. We trained it with low and high resolution data of one cell and applied the model onto others. This can be considered to be the first phase in the process of comparing 3D structures of malignant cell with normal cells. Higher resolution data allow for better and more reliable comparisons and it is now possible owing to this CNN approach.

23.

24.