PEKORA: High-Performance 3D Genome Reconstruction Using K-th Order Spearman's Rank Correlation Approximation

Advances in high-throughput sequencing technologies have enabled the use of genomic information to better understand biological processes through studies such as genome-wide association studies, polygenic risk score estimation and chromosome conformation capture. The study of spatial chromosome organization of the human genome plays an important role in understanding gene regulation. Chromosome conformation capture techniques, such as Hi-C, are able to simultaneously capture long-range interactions between all possible pairs of loci on all chromosomes. These interactions are then quantified as interaction frequency and represented as contact matrices which are quantized at a specific genomic resolution, i.e., interaction frequencies are accumulated for all 5 kb bins. These techniques have revealed structures of genome organization, such as A/B compartments, topologically associated domains (TADs), chromatin loops and frequently interacting regions (FIRE).

Although the advancement of Hi-C techniques enables the generation of massive amounts of high-resolution data, we still face several challenges, such as a high proportion of missing data and noisy observed interaction frequencies. To address these problems, we can first predict the spatial structure of the genome in three-dimensional space, using the interaction frequency as a proxy for the distance between two loci. Second, given a predicted structure, we can then infer the missing interaction frequencies. Unfortunately, it is computationally expensive to accurately and efficiently reconstruct high-resolution genome structures using the existing state-of-the-art methods such as FLAMINGO [5], H3DG [4], and SuperRec [6]. This is due to the fact that the number of interactions increases quadratically with an increase in resolution. In this work, we present a High-**PE**rformance 3D Genome Reconstruction using **K**-th **O**rder Spearman's **R**ank Correlation **A**pproximation method to reconstruct high-resolution 3D chromosome models at 5 kb. It exploits the sparse matrix property, uses an approximation of Spearman correlation as the loss function, and adjusts automatically the step size of gradient descent method at each iteration.

Methods: In our work, we focus on predicting the spatial organization of each chromosome at high resolution. Let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be a contact matrix representing all *in-cis* interactions of a chromosome. The number of rows or columns n is determined by the length of a chromosome and the resolution of the contact matrix. Due to missing data, we only have partial observations of $\mathbf{C}$ over an index set $\Omega \subset \{1,2,\ldots,n\} \times \{1,2,\ldots,n\}$. To describe the contact matrix concisely, we define the observation operator $\mathcal{P}_\Omega : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ as follows:

$$[\mathcal{P}_\Omega(\mathbf{C})]_{ij} = \begin{cases} \mathbf{C}_{ij}, & (i,j) \in \Omega, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where the number of observations is $m = |\Omega| \ll n^2$.

A chromosome is modeled as a string of beads $\mathbf{P} \in \mathbb{R}^{n\times 3}$ in 3-dimensional (3D) space, where each bead represents the center of a DNA fragment at a specific resolution. The Euclidean distance or "wish" distance between two beads is calculated by converting the interaction frequency following $\mathbf{D}_{ij} = \mathbf{C}_{ij}^{-\alpha}$ with the conversion factor α. $\mathbf{C}_{ij}$ and $\mathbf{D}_{ij}$ are the interaction frequency and Euclidean distance between beads i and j, respectively. Given $\mathbf{P}$ we can compute the Euclidean distance matrix: $\mathbf{D} = \mathbf{1}\text{diag}(\mathbf{G})^T + \text{diag}(\mathbf{G})\mathbf{1}^T - 2\mathbf{G}$, where $\mathbf{G} = \mathbf{P}\mathbf{P}^T$ is the Gram matrix.

Unlike the competitors Hierarchical3DGenome (H3DG) [4] and FLAMINGO [5], our method does not create hierarchical structures. Instead, we exploit the sparseness of the contact matrix and optimize its strucure using a gradient-descent-based alternating-minimization method [3]. In each iteration, we optimize the predicted DNA beads $\mathbf{P}$ by minimizing the prediction error using the objective function

$$L = L_1 + L_2 \tag{2}$$

with

$$L_1 = \frac{1}{m}\sum \|\mathbf{D}_{ij} - \hat{\mathbf{D}}_{ij}\|_2, \forall (i,j) \in \Omega, \tag{3}$$

where $\mathbf{D}_{ij}$ and $\hat{\mathbf{D}}_{ij}$ are the Euclidean distance and the predicted Euclidean distance of beads i and j, respectively, and

$$L_2 = \sum_{o=1}^{k} \frac{1}{m-2-o} \sum_{l=1}^{m-2-o} \sigma(\hat{\mathbf{D}}_l - \hat{\mathbf{D}}_{l+1+o}), \tag{4}$$

where $\sigma(\cdot)$ is the Rectified Linear Unit function. $\hat{\mathbf{D}}_l$ is the l-th predicted Euclidean distance sorted according to the order of the corresponding Euclidean distance $\mathbf{D}_l$. Using the Rectified Linear Unit function, the objective function penalizes the predictions only when $\hat{\mathbf{D}}_l > \hat{\mathbf{D}}_{l+1+o}$.

In our experiments, we found that only using L_1 as objective function is not sufficient to improve the accuracy of the prediction, as the Euclidean distance is inversely correlated with the interaction frequency, i.e. smaller interaction frequencies result in larger distances. These values come from very-long interactions between i and j where $i \ll j$ (or vice versa), which tend to be noisier. Since the squared error favors values with greater magnitude, it will prioritize the optimization of noisy observations. Assuming that the data $\mathbf{D}_{ij}$ are sorted and the corresponding predicted Euclidean distances $\hat{\mathbf{D}}_{ij}$ are sorted accordingly, the largest Spearman correlation is obtained when $\hat{\mathbf{D}}_l \le \hat{\mathbf{D}}_{l+1}, \forall\ 0 \le l \le m$ where $\hat{\mathbf{D}}_l$ is the l-th prediction after sorting. The Spearman correlation does not take into account the difference between $\hat{\mathbf{D}}_l$ and $\hat{\mathbf{D}}_{l+1}$. Therefore, we introduce the term L_2 being the k-th order Spearman's rank order approximation.

We use the Adam [2] optimizer, an extension of stochastic gradient descent, to update the DNA beads $\mathbf{P}$. Unlike stochastic gradient descent, which maintains a single step size to update $\mathbf{P}$, Adam updates the step size for each individual point. This is done through the first and the second moments of the gradients. In addition, the low memory footprint of Adam allows it to optimize structures with a larger number

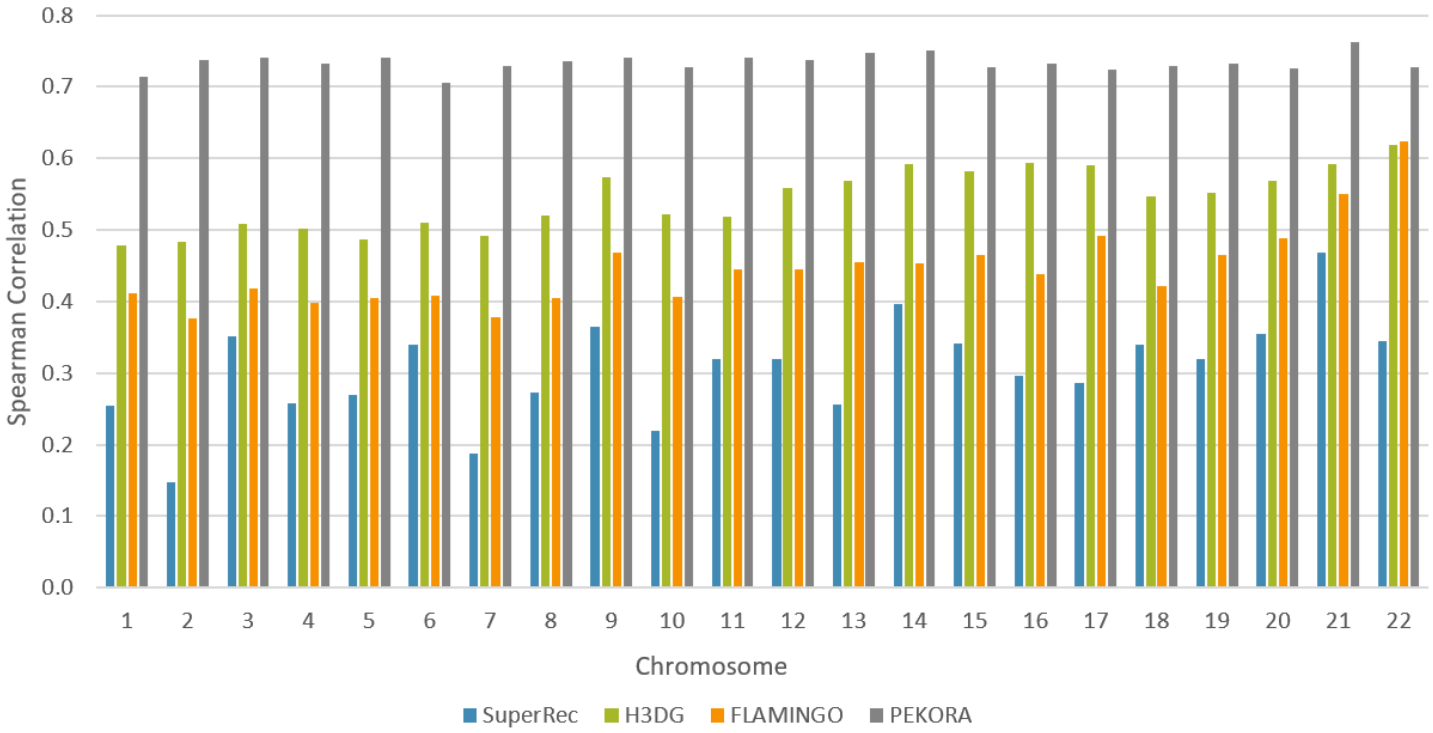

Figure 1: Spearman correlation obtained for the 3D reconstructions of each individual chromosome generated by PEKORA, FLAMINGO, H3DG and SuperRec on the GM12878 data at a resolution of 5 kb.

of observations m. Although the step size is set individually for each point, it still depends on the learning rate, and choosing an appropriate learning rate during optimization is non-trivial. Therefore, the step size of Adam itself is computed automatically based on the Barzilai-Borwein method [1], which is an approximation of Newton's method. To summarize the optimization process, the DNA beads at iteration t are computed as

$$\mathbf{P}^{(t)} = \mathbf{P}^{(t-1)} - \eta^{(t-1)} \cdot \hat{\mathbf{m}}^{(t-1)} / (\sqrt{\hat{\mathbf{v}}^{(t-1)} + \epsilon}) \qquad (5)$$

where $\eta^{(t-1)}$ is the learning rate determined by the Barzilai-Borwein method, $\hat{\mathbf{m}}^{(t-1)}$ is the first moment, and $\hat{\mathbf{v}}^{(t-1)}$ is the second moment.

Results: To quantify the accuracy of the entire predicted structure, we use the Spearman correlation, because it is independent of the conversion between interaction frequency and Euclidean distance. For our analysis, we used a Hi-C dataset of the GM12878 cell line at a resolution of 5 kb (GEO Accession Number: GSE63525). We compared PEKORA to the state-of-the-art methods FLAMINGO, H3DG and SuperRec [6]. For the 3D reconstruction, we set the order k to 20. The data are preprocessed by normalizing it using the Knight-Ruiz method to remove bias. In Figure 1, we show the obtained Spearman correlations for the 3D constructions of each individual chromosome generated by PEKORA, FLAMINGO, H3DG and SuperRec on the GM12878 data at a resolution of 5 kb. On average, PEKORA outperforms the state of the art by a large margin of 35%.

Conclusion: We present PEKORA, a high-performance 3D genome reconstruction using k-th order Spearman's rank correlation approximation. We have shown that PEKORA outperforms the state of the art by 35% on average. In the future, we will include more analyses on other cell lines with more evaluation methods, and at different resolutions.

References

[1] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[4] Tuan Trieu, Oluwatosin Oluwadare, and Jianlin Cheng. Hierarchical reconstruction of high-resolution 3d models of large chromosomes. *Scientific reports*, 9(1):4971, 2019.

[5] Hao Wang, Jiaxin Yang, Yu Zhang, Jianliang Qian, and Jianrong Wang. Reconstruct high-resolution 3d genome structures for diverse cell-types using flamingo. *Nature Communications*, 13(1):2645, 2022.

[6] Yanlin Zhang, Weiwei Liu, Yu Lin, Yen Kaow Ng, and Shuaicheng Li. Large-scale 3d chromatin reconstruction from chromosomal contacts. *BMC genomics*, 20:129–141, 2019.