3D Construction of chromosomes using Reinforcement Learning

Neeta Kannu

Proposal Paper: Spring 2024
CS 5080- Reinforcement Learning

Abstract

This research presents a methodology for reconstructing 3D chromosomes through reinforcement learning (RL). We start by collecting and processing Hi-C data to ensure its quality. Chromosome structures are represented using spatial coordinates and other derived features from input data. RL algorithms explore a parameter space to optimize structures based on experimental data. We use Deep Deterministic Policy Gradient (DDPG) or Proximal Policy Optimization (PPO) for training, followed by parameter tuning and optimization. Evaluation involves assessing the accuracy and biological plausibility through correlation coefficients, RMSE, and visual inspection. This methodology provides a comprehensive approach to 3D chromosome reconstruction using RL.

Introduction

Chromosome structure plays a fundamental role in regulating gene expression, DNA replication, and genome stability, thereby influencing various cellular processes and organismal development. Understanding the three-dimensional (3D) organization of chromosomes is essential for unraveling the complexities of genome function and regulation. Recent advancements in genomic techniques, particularly Hi-C data, have provided unprecedented insights into the spatial organization of chromatin within the nucleus.

The reconstruction of chromosome structures from Hi-C data is a challenging task that requires sophisticated computational methods and algorithms. Traditional approaches often rely on optimization techniques to infer 3D chromatin architectures from the interaction frequencies derived from Hi-C experiments. However, these methods may face limitations in accurately capturing the intricate spatial arrangements of chromosomal regions.

In this context, reinforcement learning (RL) presents a promising avenue for advancing the reconstruction of chromosome structures. RL focuses on enabling agents to learn optimal behavior by interacting with their environment and receiving feedback in the form of rewards. By applying RL techniques to chromosome modeling, researchers can leverage the power of adaptive learning to iteratively refine and improve the accuracy of reconstructed 3D chromatin structures.

This research paper aims to explore state-of-the-art techniques and methodologies for constructing chromosome structures using reinforcement learning. We will review existing methods, such as the 3DMax algorithm, the GEM framework, and the Orca model architecture, and analyze their strengths and limitations in reconstructing chromosome architectures from Hi-C data. Furthermore, we will propose a novel RL-based approach for 3D chromosome modeling, discussing its potential advantages and implications for advancing our understanding of genome organization and function.

With the latest developments in reinforcement learning and integrating them into the task of chromosome structure reconstruction, this research holds the promise of unlocking new insights into the spatial organization of the genome and its impact on gene regulation and cellular processes.

Background

The three-dimensional (3D) organization of chromosomes within the nucleus plays a fundamental role in regulating gene expression, DNA replication, and other cellular processes. Advances in genomic technologies, particularly high-throughput chromosomal conformation capture (Hi-C), have revolutionized our ability to investigate chromatin architecture by capturing genome-wide chromosomal interactions.

Traditional methods for reconstructing 3D chromosome structures from Hi-C data often rely on optimization algorithms, such as gradient ascent, to infer spatial arrangements of chromatin regions based on interaction frequencies. While effective, these methods have limitations in accurately capturing the complex and dynamic nature of chromosomal organization.

In recent years, there has been a growing interest in machine learning techniques, particularly reinforcement learning (RL), to improve the reconstruction of chromosome structures from Hi-C data. RL offers a promising approach by allowing computational agents to learn optimal strategies for interacting with their environment and receiving feedback as rewards, thereby enabling adaptive learning and refinement of 3D chromatin models.

Current state-of-the-art methodologies for reconstructing chromosome structures include the 3DMax algorithm, the genomic organization reconstructor based on the conformational energy and manifold learning (GEM) framework, and the Orca model architecture. These approaches incorporate

various optimization and modeling techniques to infer 3D chromosomal configurations from Hi-C interaction data.

Despite these advancements, challenges remain in accurately reconstructing high-resolution and biologically relevant chromosome structures. Incorporating reinforcement learning into chromosome modeling presents an exciting opportunity to address these challenges by enabling agents to iteratively refine 3D chromatin models based on feedback from the environment.

This research aims to build upon the current state of the art in chromosome structure reconstruction by proposing a novel approach that integrates reinforcement learning techniques. By RL, we aim to develop a more adaptive and accurate method for constructing 3D chromosome models from Hi-C data, ultimately advancing our understanding of genome organization and its functional implications.

Related Work

Several significant methodologies and algorithms have been developed for reconstructing chromosome structures from Hi-C data. One notable approach is the Oluwadare et al. (2018), which utilizes gradient ascent optimization to reconstruct chromosome structures. This algorithm optimizes the log-likelihood objective function by incorporating factors such as contact maps, resolution, and interaction frequencies derived from the Hi-C data. Another notable study by Trieu et al. (2014) focuses on the large-scale reconstruction of the 3D structures of human chromosomes from chromosomal contact data. This approach involves preprocessing Hi-C data to reduce biases, removing contacts with low likelihood ratios, and representing chromosome structures based on observed contacts and non-contacts. These methodologies provide essential insights into the reconstruction of chromosome structures from Hi-C data, contributing to advancements in genomics and biomedicine research.

The GEM framework, as outlined by Zhou (2022), offers a method for reconstructing 3D spatial organizations of chromosomes through manifold learning techniques. The framework embeds neighboring affinities from Hi-C space into 3D Euclidean space, aiming to preserve local structure while mapping from Hi-C space. The Orca model architecture, as described by Zhu et al. (2018), presents a hierarchical sequence encoder and multilevel cascading decoder for multiscale 3D genome prediction. Trained on processed micro-C datasets for H1-ESCs and HFF cells, the model predicts interaction matrices representing pairwise genome interactions at varying resolutions. These diverse methodologies and algorithms contribute to the growing body of research in chromosome structure reconstruction, providing valuable insights and applications for understanding genomic organization and function.

The research paper J Li et al. (2016) outlines a method to reconstruct chromosome 3D structures from Hi-C data using computational techniques and reinforcement learning principles. The dataset includes both simulated and real Hi-C data from mouse embryonic stem cells (mESC) and human GM06990 cells. The methodology entails representing states using 3D coordinates, implementing actions like the shortest-path method and multidimensional scaling (MDS), and optimizing rewards by minimizing the discrepancy between predicted and experimental contact frequencies.

In Gong et al. (2023), the NeRV-3D-DC framework is presented for reconstructing 3D chromosome structures from Hi-C data. The dataset encompasses simulated and actual Hi-C data from GM12878 and IMR90 cell lines. The methodology includes converting contact matrices into distance matrices, reconstructing structures using divide-and-conquer strategies, and evaluating structures using metrics such as root mean square error (RMSE) and Pearson correlation coefficient (PCC).

Zhang et al. (2024) explores the analysis of 3D genomic mapping data, particularly focusing on Hi-C data to identify structural features like compartments, topologically associating domains (TADs), and loops. The paper emphasizes the challenges of achieving consistent feature identification across different methods and datasets.

Wang et al. (2023) introduces the EVRC algorithm for reconstructing chromosome 3D structures from Hi-C data. The dataset comprises simulation data and published Hi-C datasets. The methodology involves converting interaction frequencies into spatial distances, integrating the co-clustering coefficient, and refining structure fit using reinforcement learning.

Song et al. (2021) presents the DQN x-drop algorithm for local sequence alignment utilizing deep reinforcement learning. The dataset consists of DNA sequence pairs. The methodology includes representing alignment states, selecting actions based on alignment directions, and enhancing alignment scores using reinforcement learning.

Z Li et al. (2023) discusses methods for reconstructing 3D chromosome structures from Hi-C data, considering various techniques such as distance-based and contact-based methods. The paper underscores rewards based on the accuracy of reconstructed structures compared to experimental data and outlines methodologies involving state initialization, action selection, reward calculation, policy update, iteration, and validation.

Methodology

Data Collection and Preprocessing:

The dataset used in this research includes both simulated and actual Hi-C data from the GM12878 and IMR90 datasets. These datasets are valuable for reconstructing 3D chromosome structures using reinforcement learning techniques. Simulated data provide controlled environments for algorithm testing, while actual Hi-C data offer insights into real biological systems.

State Representation:

The chromosome structure is represented by integrating spatial coordinates, chromatin loop arrangements, and other relevant structural features. These features are derived from the input data and the model parameters. This detailed representation ensures an accurate depiction of the chromosome structure's current state at each reconstruction stage, laying a strong foundation for subsequent analysis and optimization.

Reinforcement Learning Setup:

The RL environment is designed to enable algorithms to effectively navigate a parameter space of chromatin structures. Experimental data, such as Hi-C data, is a crucial input to the environment. The state space includes the current representation of the chromosome structure, allowing for dynamic adjustments and modifications. Actions within the environment involve fine-tuning chromosomal region coordinates, optimizing genomic locus connectivity, and adjusting model parameters to better match experimental data. The reward system is based on the fidelity of the reconstructed structure compared to experimental data or a predefined objective function. This ensures that the RL agent generates biologically meaningful chromosome models.

Algorithm Selection and Training:

Deep Deterministic Policy Gradient (DDPG) or Proximal Policy Optimization (PPO) can efficiently optimize objective functions, work with experimental constraints, and create biologically plausible 3D chromosome models. These algorithms offer robust learning frameworks capable of adapting and refining the reconstruction process based on environmental feedback. This ensures the creation of high-quality chromosome structures. While DDPG or PPO are promising for chromosome structure reconstruction, exploring alternative approaches such as Deep Q-Networks (DQN) or actor-critical methods may provide valuable insights. However, PPO stands out due to its stability, sample efficiency, simplicity, and proven performance across various domains, making it the optimal choice for reconstructing 3D chromosome structures using reinforcement learning techniques.

Parameter Tuning and Optimization:

Fine-tuning algorithm parameters and RL hyperparameters will enhance reconstruction performance. Systematically adjusting these parameters allows the algorithm to effectively navigate the parameter space and optimize the chromosome structure reconstruction process. Optimization efforts focus on optimizing computational resources and training procedures to ensure efficiency and scalability, enabling the approach to be applied to large-scale datasets and complex chromosome structures.

Evaluation

Quantitative Evaluation:

Correlation Coefficients: We compute Pearson and Spearman's correlation coefficients to analyze how closely the reconstructed chromosome structures align with the ground-truth data. These coefficients offer insights into the fidelity of the reconstructed structures compared to the actual data.

Root Mean Square Error (RMSE): We calculate RMSE to measure the overall discrepancy between the reconstructed chromosome structures and the ground-truth structures. This metric helps quantify the accuracy of the reconstruction process in terms of spatial fidelity.

Qualitative Evaluation:

Visual Inspection: We visually compare the reconstructed chromosome structures with the ground-truth data, using tools like 3D visualization software such as UCSF Chimera or PyMOL and measurable techniques like voxel-based analysis. Specifically, we analyze structural features like chromatin domain arrangement and genomic locus positioning. This evaluation approach allows us to thoroughly evaluate the accuracy and biological plausibility of the reconstructed models, utilizing visualization tools and techniques for detailed examination and comparison.

Combination of Metrics:

Integration of Quantitative and Qualitative Assessment: Combining quantitative metrics such as correlation coefficients and RMSE with qualitative evaluation methods like visual inspection ensures a thorough assessment of the RL agent's performance. This integrated approach provides a complete understanding of the accuracy and biological significance of the reconstructed 3D chromosome models.

Timeline

Task	Date
Data Cleaning	29-Feb-24
Data Pre-Processing	11-Mar-24
Methodology Development and RL Preparation	30-Mar-24
Parameter Tuning and Optimization	30-Apr-24
Test Results and Analysis	10-May-24

Conclusion

In summary, our research employs reinforcement learning (RL) for 3D chromosome reconstruction, utilizing data preprocessing to ensure the quality and representation of chromosome structures. We train RL algorithms such as DDPG or PPO to optimize objective functions for reconstructing biologically plausible models. Parameter tuning and optimization improve reconstruction performance, which we evaluate using correlation coefficients, RMSE, and visual inspection. This methodology presents a promising approach to understanding chromatin organization and gene regulation.

References

Gong, Haiyan et al. (2023). "A 3D Chromosome Structure Reconstruction method with High Resolution Hi-C Data using Nonlinear Dimensionality Reduction and Divide-and-conquer strategy". In: *IEEE Transactions on NanoBioscience*.

Li, Jiangeng, Wei Zhang, and Xiaodan Li (2016). "3D genome reconstruction with ShRec3D+ and Hi-C data". In: *IEEE/ACM transactions on computational biology and bioinformatics* 15.2, pp. 460–468.

Li, Zilong, Stephanie Portillo-Ledesma, and Tamar Schlick (2023). "Techniques for and challenges in reconstructing 3D genome structures from 2D chromosome conformation capture data". In: *Current Opinion in Cell Biology* 83, p. 102209.

Oluwadare, Oluwatosin, Yuxiang Zhang, and Jianlin Cheng (2018). "A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data". In: *BMC genomics* 19.1, pp. 1–17.

Song, Yong-Joon and Dong-Ho Cho (2021). "Local alignment of DNA sequence based on deep reinforcement learning". In: *IEEE open journal of engineering in medicine and biology* 2, pp. 170–178.

Trieu, Tuan and Jianlin Cheng (2014). "Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data". In: *Nucleic acids research* 42.7, e52–e52.

Wang, Xiao et al. (2023). "EVRC: reconstruction of chromosome 3D structure models using error-vector resultant algorithm with clustering coefficient". In: *Bioinformatics* 39.11, btad638.

Zhang, Yang et al. (2024). "Computational methods for analysing multiscale 3D genome organization". In: *Nature Reviews Genetics* 25.2, pp. 123–141.

Zhou, Jian (2022). "Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale". In: *Nature genetics* 54.5, pp. 725–734.

Zhu, Guangxiang et al. (2018). "Reconstructing spatial organizations of chromosomes through manifold learning". In: *Nucleic acids research* 46.8, e50–e50.