УДК 575.112

РАЗРАБОТКА ПЛАГИНА ПОИСКА ПЕРЕСТРОЕК ДЛЯ ПЛАТФОРМЫ ПРОВЕДЕНИЯ СРАВНИТЕЛЬНЫХ ИССЛЕДОВАНИЙ ТРЁХМЕРНЫХ СТРУКТУР ГЕНОМОВ.

Дравгелис В.А. (Университет ИТМО)
Научный руководитель – аспирант ФИТИП Замятин А.А.
(Университет ИТМО)

Введение. Структура генома живых организмов является сложной и многогранной. Трехмерная организация генома живых организмов является одним из важных аспектов их биологии. Она играет ключевую роль в регуляции экспрессии генов [1], репликации ДНК, репарации ДНК [2] и других процессах. Для её изучения используются различные методы, в том числе метод Hi-C [3]. Этот метод позволяет создать матрицу контактов между различными регионами генома, что дает представление о его трехмерной структуре.

Результатом работы метода Hi-C является квадратная матрица, размерность которой равна длине исследуемого участка ДНК. Такие матрицы отображаются в виде тепловых карт и называются картами Hi-C. Карты Hi-C содержат много информации о трехмерной организации генома, но их анализ, к примеру поиск геномных перестроек, является сложной задачей.

Для облегчения анализа карт Hi-C предлагается использовать специальные программные средства. В рамках данной работы был разработан метод обнаружения геномных перестроек на картах Hi-C с использованием свёрточных нейронных сетей.

Основная часть. Hi-C — это метод, который используется для изучения трехмерной организации хромосом внутри клеток. Он основан на принципе лигаз-опосредованной ПЦР [4], которая позволяет выявлять контакты между различными регионами генома. Результатом работы данного метода является квадратная матрица, размерность которой равна длине исследуемого участка ДНК (чаще всего одной или нескольким хромосомам). Такие матрицы отображаются в виде тепловых карт и называются картами Hi-C. Обычно матрицы имеют разрешение от 1000 (1 Kb – kilobases, килобазы) до 1000000 (1 Mb – megabases, мегабазы), это значит, что одна ячейка или один пиксель содержит информацию о взаимном расположении 1000–1000000 нуклеотидных оснований. В данной работе рассматривались матрицы размерности 500 000–100 000 Kb, что достаточно немного, потому что в них содержалось по одной хромосоме, зачастую работы ведутся с матрицами, содержащими несколько хромосом. Из-за этого рассматривать и анализировать матрицы вручную становится затруднительно.

Для облегчения анализа и поиска объектов интереса на картах Hi-C предлагается использовать специальное ПО, поставляемое отдельно или интегрированное в ПО для просмотра Hi-C карт (HICT).

В основе разрабатываемой системы принято решение использовать свёрточные нейронные сети, по описанному в [5] принципу. Он заключается в использовании свёрточных нейронных сетей, для классификации отдельных маленьких участков карты Hi-C, в зависимости от нахождения или отсутствия в них геномных перестроек. Это возможно благодаря тому, что каждый вид перестроек имеет характерные паттерны на карте, по которым их можно найти. В отличии от описанной в статье системы, разрабатываемая, имеет другую архитектуру, использующее меньшее количество ресурсов, за счёт оптимизации использования полносвязных слоёв. Для формирования обучающей выборки были использованы аугментированные карты Hi-C нескольких видов комаров. За счёт этого нейросеть обучилась находить перестройки в геномах любых видов живых организмов.

На данном этапе существует обученная модель, которая достаточно точно обнаруживает геномные перестройки в разрешении 50, 10, 5 и 1 Kb. Система по поиску и детектированию геномных перестроек реализована в виде отдельной утилиты. Ведутся работы по её интеграции в платформу для интерактивного ручного скаффолдинга геномных сборок и

визуализации карт Hi-C “HiCT”.

Выводы. В рамках этой работы было проведено исследование об использовании различных методов из статистики и машинного обучения для анализа данных, получаемых методом Hi-C. Изначально были рассмотрены методы из статистики и типовые алгоритмы кластеризации данных. Однако, не получив на простых методах удовлетворительного результаты был выполнен поэтапный переход к более сложным. В конце концов, основываясь на проведённых ранее исследованиях, был совершён выбор в пользу свёрточных нейронных сетей. Несмотря на сложности, возникающие при работе с ними, и необходимости самостоятельного создания наборов данных для их обучения, конечный результат оправдывал затраченные усилия. На тестовой выборке данных модель показала точность 99.97%. Это позволило использовать её в основе разработанной системы и определять местоположение геномных перестроек с высокой точностью.

Список использованных источников:

1. Pai DA, Engelke DR. Spatial organization of genes as a component of regulated expression. // Chromosoma. – 2010 – 119(1):13-25.
2. Sivakumar A, de Las Heras JI, Schirmer EC. Spatial Genome Organization: From Development to Disease. // Front Cell Dev Biol. – 2019 – 7:18.
3. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. // Science. – 2009 – 326(5950):289-293.
4. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. // Methods. – 2012 – 58(3):268-276.
5. Xiaotao Wang et al. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. // Sci. Adv. – 2022