УДК 575.112

РАЗРАБОТКА ИНСТРУМЕНТА HiCT: ПЕРЕХОД ОТ ИНТЕРАКТИВНОГО ПРОТОТИПА К ЭКСПЛУАТАЦИОННО-УСТОЙЧИВОЙ РЕАЛИЗАЦИИ

Сердюков А. Н.[1]
Научный руководитель – канд. техн. наук, доцент Муравьёв С. Б.[1]
[1]Университет ИТМО
anserdiukov@itmo.ru

Введение

Скаффолдинг является одним из заключительных этапов сборки генома и заключается в упорядочивании и ориентировании контигов – однозначно определённых последовательностей ДНК, полученных на выходе автоматического сборщика, в последовательности большего размера – скаффолды – которые должны соответствовать истинной последовательности нуклеотидов в молекуле ДНК [1].

Hi-C – это метод молекулярной биологии, позволяющий получить информацию о взаимном расположении участков ДНК в трёхмерном пространстве. Hi-C данные часто визуализируются в виде тепловых карт и долгое время являлись одним из главных источников дополнительных данных для скаффолдинга геномных сборок. С развитием технологий секвенирования, скаффолдинг всё реже становится необходимым для получения геномных сборок высокого качества, однако Hi-C данные нашли применение в задаче поиска и анализа геномных перестроек.

В 2022 году существовало единственное программное решение, позволяющее производить интерактивный ручной скаффолдинг – это инструмент JBAT [2], разработанный в лаборатории Aiden Lab. По состоянию на 2026 год, существующие решения с открытым исходным кодом, такие как Cooler [3], HiGlass [4] или hictk [5] ориентированы на визуализацию и не позволяют производить операции скаффолдинга, а некоторые из них требуют длительной предобработки или большого объёма ОЗУ.

Для решения этих задач в конце 2021 года была начата разработка инструмента HiCT для интерактивного ручного скаффолдинга. После применения первой версии инструмента для сборки геномов восьми комаров, было принято решение улучшить модель данных, а кодовая база была перенесена из экосистемы Python на Java.

Для практического применения HiCT было недостаточно корректного рендера и базовых операций над сборкой, требовалось также улучшить пользовательский опыт, исправить недочёты пользовательского интерфейса, а также интегрировать вспомогательные компоненты в кодовую базу на Java. Решению этих задач посвящена работа на данном этапе развития проекта.

Основная часть

Основной задачей данного этапа стал перевод конвертера формата файлов с Python на Java. Инструмент HiCT использует собственный формат файлов на основе контейнера HDF5 [6], в котором Hi-C матрицы для каждого разрешения сохраняются в блочном разреженном представлении, а для каждого контига сохраняется информация о сопоставлении его с блоками. Для работы с этим представлением в HiCT реализована модель данных на основе Декартовых деревьев по неявному ключу со случайными приоритетами (с модификацией для персистентности), а также системой кэширования для ускорения работы параллельных запросов. Формат файла и реализация структуры данных изменяются для улучшения производительности на геномных сборках с различными особенностями, поэтому формат файлов для инструмента HiCT не применяется для долговременного хранения, и важной задачей является возможность быстрого преобразования между ним и более стабильными форматами библиотеки Cooler [3] и инструмента Juicebox [2]. Для этого существующая реализация конвертера,

располагавшаяся в Python-проекте HiCT_Utils, была переведена на язык Java и включена в основной репозиторий HiCT_JVM. Таким образом, вся необходимая для работы с инструментом инфраструктура стала располагаться в одном Java-проекте, и распространяться единым jar-файлом. Отдельной задачей стала реализация необходимых методов на языке Java, так как в ней отсутствуют пакеты для практически-применимой замены Python-библиотек h5py и numpy, активно применявшихся в Java-версии конвертера. Помимо создания Java-реализации конвертера с консольным интерфейсом, элементы для взаимодействия с ним также были интегрированы в веб-интерфейс HiCT.

Другим значимым направлением изменений стала оптимизация архитектуры доменных объектов и методов. В частности, акцент был сделан на снижение необходимости повторных обращений к файлу, большему разделению на независимые подзадачи, которые можно было бы исполнять параллельно, а также применению кэширования состояний, для снижения времени отклика и повышения пропускной способности решения. Помимо этого, были исправлены ошибки в методах закрытия файлов, а также добавлена возможность отслеживания хода выполнения длительных операций, таких как конвертация файла или открытие файлов с большими Hi-C картами.

Значительные изменения претерпел также пользовательский веб-интерфейс, в котором появилась возможность переименования контигов и скаффолдов, экспорта всей Hi-C карты в текущем разрешении в нескольких форматах (SVG, PNG и PDF) с наложением разметки, сохранения сессии, а также экспорта и импорта всех настроек визуализации. По запросам пользователей добавилась возможность «бесконечного» приближения и отдаления Hi-C карты.

Выводы

По итогам данного этапа HiCT демонстрирует качественный переход к более зрелому состоянию: конвертация стала частью единого jar-пакета, длительные операции получили возможность наблюдения за ходом выполнения, сессии устойчивее к прерываниям клиентской стороны, а конфигурации именования и визуализации стали переносимым и воспроизводимым состоянием проекта. Данные изменения значительно улучшают пользовательский опыт и повышают интерес к переходу на инструмент HiCT. В дальнейшем планируется зафиксировать API серверной части и реализовать Python-библиотеку для реализации DataLoader в задачах машинного обучения.

Литература

1. Comprehensive mapping of long-range interactions reveals folding principles of the human genome / E. Lieberman-Aiden [et al.] // Science. — 2009. — Vol. 326, no. 5950. — P. 289-293
2. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom / N. C. Durand [et al.] // Cell Syst. — 2016. — Vol. 3, no. 1. — P. 99–101.
3. Abdennur N., Mirny L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays // Bioinformatics. 2020. Vol. 36, no. 1. P. 311–316.
4. Kerpedjiev P., Abdennur N., Lekschas F. et al. HiGlass: web-based visual exploration and analysis of genome interaction maps // Genome Biology. 2018. Vol. 19, no. 1. Art. 125.
5. Rossini R., Paulsen J. hictk: blazing fast toolkit to work with .hic and .cool files // Bioinformatics. 2024. Vol. 40, no. 7. Art. btae408.
6. The HDF Group. Hierarchical Data Format, version 5 [Электронный ресурс]. — Режим доступа: https://www.hdfgroup.org/HDF5/. (дата обращения 02.02.2026).