УДК 575.112

РАЗРАБОТКА РЕЛИЗНОЙ ВЕРСИИ ПЛАТФОРМЫ HiCT НА JAVA

Сердюков А.Н. (Университет ИТМО)
Научный руководитель – аспирант ФИТиП, Замятин А.А.
(Университет ИТМО)

Введение. Скаффолдинг является последним этапом процесса геномной сборки 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. Современные технологии секвенирования позволяют создавать контиги сравнимые по длине с полными последовательностями хромосом. На данный момент последний этап скаффолдинга проводится и проверяется человеком, что значительно затрудняет общую автоматизацию процесса геномной сборки хромосомного уровня.

Hi-C – метод молекулярной биологии, позволяющий получить информацию о взаимном расположении участков ДНК в трёхмерном пространстве. Эти данные можно использовать в процессе скаффолдинга для того, чтобы правильно упорядочить и ориентировать контиги.

В 2022 году существовало единственное программное решение, позволяющее производить интерактивный ручной скаффолдинг – это инструмент JBAT, разработанный в лаборатории Aiden Lab. Данный инструмент успешно применялся в НОЦ Геномного Разнообразия Университета ИТМО для сборки геномов, однако обладал рядом недостатков, таких как нерациональное использование оперативной памяти, из-за чего его применение для сборок больших геномов было невозможно.

Для решения задач НОЦ Геномного Разнообразия 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 также реализовать релизную версию HiCT, разработке которой посвящён данный проект.

Основная часть. Целью проекта HiCT является разработка программного инструмента, позволяющего в реальном времени визуализировать карты Hi-C контактов и интерактивно выполнять операции скаффолдинга геномных сборок хромосомного порядка, сопоставимых по размеру с геномом человека. Создание набора программных библиотек и инструментов HiCT было начато в 2022 году и изначально платформа реализовывалась в экосистеме языка Python и библиотек. Изначальная реализация позволяла работать с геномами комаров семейства *Anopheles*, однако в процессе её использования был отмечен ряд недостатков, в частности сложность установки и отсутствие одновременного многопоточного исполнения. Данная работа посвящена новой реализации HiCT на языке Java, устраняющей недостатки предыдущей реализации.

Для переноса HiCT с Python на Java необходимо было перевести основную кодовую базу, отвечающую предметной области, найти аналоги библиотек используемых зависимостей для языка Python в экосистеме языка Java. Большая часть кода, относящегося к объектам предметной области, была переведена на язык Java без особых трудностей, так как на Python уже существовала отлаженная их реализация, не использовавшая сторонних библиотек или возможностей, недоступных в Java.

Реализация модели данных также была переведена, при этом соответствующие примитивы синхронизации и многопоточные вызовы заменялись на соответствующие конструкции языка Java. В отличие от большинства реализаций интерпретаторов Python, обладающих глобальной блокировкой интерпретатора (GIL), которая не позволяет нескольким потокам работать с объектом одновременно, стандартные конструкции

многопоточности в Java лишены такого недостатка.

Основная сложность заключалась в переводе кода, активно использующего сторонние библиотеки языка Python. Некоторые зависимости имели подходящие аналоги в Java. Так, например, веб-фреймворк Flask был заменён на Vert.X. Однако, большая часть библиотек не имела прямых аналогов в экосистеме Java и было принято решение заменять их на временные ручные реализации, которые предоставляли минимально необходимое число методов и уровень абстракции для переноса кода с Python. Так, например, библиотека numpy, использовавшаяся для работы с матрицами в версии HiCT на Python, была заменена классом, реализующим матрицу как многомерный массив в Java, и поддерживающим только часть операций. Библиотека для визуализации матриц matplotlib была заменена на ручную реализацию с использованием ImageIO из пакета javax.

Наиболее сложным оказался поиск альтернативы для библиотеки h5py, предоставляющей удобный интерфейс для работы с файлами HDF5, на основе которых построен формат файлов HiCT. Также, в HiCT используются плагины сжатия HDF5, которые позволяют сжимать исходные данные «на лету». Большинство реализаций HDF5 для Java либо не предоставляли необходимый уровень абстракций, либо были реализованы на Java без возможности поддержки плагинов, что делало их применение невозможным. Выбор был сделан на библиотеке SiS jHDF5, разработанной в ETH Zurich. Однако, её интеграция в HiCT также была сопряжена с рядом трудностей. После нескольких попыток отладки, было обнаружено, что конфигурация библиотеки по умолчанию, использовавшаяся для сборки версии jHDF5, доступной в репозиториях, использовала статическое связывание с библиотекой HDF5, что также делало невозможным использование плагинов, так как они требуют наличия динамической библиотеки HDF5. Исправление конфигурации сборки библиотеки HDF5 и скриптов сборки JHDF5 позволило решить данную проблему и собрать модифицированную версию библиотеки JHDF5, предоставляющую достаточный уровень абстракции и имеющую поддержку плагинов сжатия.

Выводы. Создана новая реализация HiCT на языке Java. Данная версия позволяет использовать все возможности обновлённой модели данных HiCT для одновременных чтений из нескольких потоков. Решена проблема разрешения зависимостей, а также собрана модифицированная версия библиотеки JHDF5, имеющей поддержку плагинов сжатия. Новая реализация HiCT поставляется в виде единственного jar-файла, в которого упакованы все зависимости, включая собранные библиотеки. Для запуска HiCT от пользователя требуется только установка Java версии не ниже 19, но не требуется сложный процесс установки дистрибутива библиотеки HDF5 в систему.

Список использованных источников:

1. Durand N.C. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom // Cell Syst. — 2016. — Т. 3, № 1. — С. 99–101.

2. The Java® Language Specification. Java SE 19 Edition [Электронный ресурс]. — URL: https://docs.oracle.com/javase/specs/jls/se19/html/index.html (дата обращения 10.07.2023)

3. H5py documentation [Электронный ресурс]. — URL: https://docs.h5py.org/en/stable/ (дата обращения 29.07.2023)

4. HDF5 Reference Manual (v1.12) [Электронный ресурс]. — URL: https://docs.hdfgroup.org/hdf5/v1_12/_r_m.html (дата обращения 17.09.2023)

5. JHDF (HDF5 for Java) [Электронный ресурс]. — URL: https://unlimited.ethz.ch/pages/viewpage.action?pageId=92865195 (дата обращения 23.10.2023)

6. JHDF5 source code [Электронный ресурс]. — URL: https://sissource.ethz.ch/sispub/jhdf5 (дата обращения 03.10.2023)