**УДК 51-76**

# РАЗРАБОТКА ИНСТРУМЕНТА ДЛЯ ИНТЕРАКТИВНОЙ ВИЗУАЛИЗАЦИИ И СКАФФОЛДИНГА ГЕНОМНЫХ СБОРОК ПО ДАННЫМ HI-C

**Автор – 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. 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. На данный момент последний этап скаффолдинга проводится и проверяется человеком, что значительно затрудняет общую автоматизацию процесса геномной сборки хромосомного уровня.

На момент создания инструмента HiCT в 2022 году проблема заключалась в отсутствии альтернатив среди ПО с открытым исходным кодом, эффективно решающего задачу интерактивного скаффолдинга для геномов, сравнимых по размеру с геномом человека, на основе данных Hi-C. Единственной альтернативой с открытым исходным кодом является инструмент JBAT, разработанный в лаборатории Aiden Lab. Данное ПО обладает рядом недостатков, среди которых можно отметить нерациональный расход оперативной памяти, ограниченный набор инструментов автоматического скаффолдинга и долгий процесс финализации конечной сборки. Несмотря на то, что в последующих версиях инструмента JBAT некоторые из этих проблем были устранены, разработка альтернативного приложения с открытым исходным кодом, реализующего иной подход к модели хранения данных, является важной задачей. Основная библиотека HiCT, реализующая операции для работы с моделью данных, разрабатывается с учётом дальнейшей автоматизации процесса, в том числе с помощью применения методов ИИ. Также она упрощает интеграцию с существующими инструментами анализа данных на Python за счёт использования совместимых интерфейсов.

**Основная часть.** На текущем этапе работы мы разработали набор библиотек на Python и пользовательское приложение для интерактивной визуализации Hi-C карт для скаффолдинга и валидации геномных сборок, использующее модель данных, предложенную в весеннем семестре 2022 года и концепцию веб-интерфейса.

Проект HiCT состоит из четырёх частей: библиотеки HiCT для языка Python, сервера HiCT и веб-интерфейса HiCT, а также набора вспомогательных утилит. Библиотека HiCT реализует функционал для работы с файлами и моделью данных, а именно: запросы произвольной прямоугольной области Hi-C карты при текущем состоянии сборки, операции скаффолдинга и операции экспорта сборки и геномного контекста области. Данная библиотека впоследствии может быть использована в качестве источника данных для алгоритмов автоматизированного скаффолдинга и глубокого анализа Hi-C карт. Конечный пользователь взаимодействует с HiCT при помощи веб-интерфейса, который был значительно доработан за данный период. Пользовательские элементы преимущественно реализованы с использованием библиотеки Bootstrap, интерактивная область Hi-C карты реализована при помощи картографической библиотеки OpenLayers, которая способна загружать карты различных разрешений по фрагментам, называемым *тайлами*, таким образом загружая

только те фрагменты Hi-C карты, которые попадают в область видимости пользователя. Для сопряжения веб-интерфейса с библиотекой также был разработан сервер HiCT, использующий фреймворк Flask. Набор вспомогательных утилит включает в себя конвертер данных Hi-C из формата Cooler в формат HiCT.

Данная реализация использовалась при сборке геномов девяти комаров семейства *Anopheles*, был составлен список замечаний к пользовательскому интерфейсу. В одном из девяти случаев исходные данные были сравнительно низкого качества, из-за чего был выявлен ряд недостатков в модели данных. Их устранению и добавлению поддержки многопоточности посвящена значительная часть работы.

**Выводы.** Проделанная работа над инструментом HiCT позволила перевести его из состояния концепта в состояние минимального жизнеспособного продукта. Разработанные программные библиотеки и инструменты позволили применить HiCT в реальной задаче сборки генома. Обновлённая модель повысила производительность запросов тайлов, а также сделала возможным добавление многопоточности в приложение, устранив проблему «разрушающего чтения» из структуры Декартова дерева по неявному ключу с отложенными операциями. Новая структура модели позволила существенно упростить и ускорить процесс конвертации данных из формата Cooler в формат HiCT. Программный интерфейс библиотеки HiCT унифицирован с Cooler в части запросов произвольной прямоугольной области для упрощения разработки и тестирования.

Среди дальнейших планов присутствуют как краткосрочные: добавление 1D-треков в веб-интерфейс HiCT, а также перенос сервера HiCT с фреймворка Flask на FastAPI, так и долгосрочные, а именно создание режима «валидации» сборки, в котором будут оптимизированы операции запроса тайлов за счёт отключения операций скаффолдинга, а также разработка модуля для поиска геномных перестроек при помощи алгоритмов глубокого машинного обучения и ИИ с последующей интеграцией данного модуля в веб-интерфейс HiCT.

**Список использованных источников:**


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. Abdennur N., Mirny L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays // Bioinformatics. — 2019. — URL: https://doi.org/10.1093/bioinformatics/btz540.
3. 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.
4. OpenLayers [Электронный ресурс]. — URL: https://openlayers.org/ (дата обращения 17.12.2022).
5. Robinson J., Thorvaldsdottir H., Turner D., Mesirov J. IGV.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV) // Bioinformatics // Bioinformatics. — 2023. — URL: https://doi.org/10.1093/bioinformatics/btac830


Сердюков А.Н. (автор) Подпись

Замятин А.А. (научный руководитель) Подпись