УДК 51-76

РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ИНТЕРАКТИВНОГО СКАФФОЛДИНГА И ВАЛИДАЦИИ ГЕНОМНЫХ СБОРОК НА ОСНОВЕ ДАННЫХ HI-C

Автор – Сердюков А.Н. (Университет ИТМО)

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

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №621312 «Разработка программного обеспечения для интерактивного скаффолдинга геномных сборок на основе данных Hi-C»

Работа посвящена созданию программного инструмента, позволяющего в реальном времени визуализировать и интерактивно взаимодействовать с Hi-C данными геномов, сопоставимых по размеру с геномом человека, для решения проблемы скаффолдинга (упорядочивания и ориентации имеющихся последовательностей) и валидации геномных сборок. В рамках данной работы был разработан новый формат хранения данных Hi-C экспериментов, позволяющий эффективно реализовать необходимые запросы, а также утилита, позволяющая преобразовывать существующие данные в этот формат.

Введение. Скаффолдинг является последним этапом геномной сборки и заключается в упорядочивании и ориентировании однозначно определённых последовательностей ДНК – контигов – 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 сравнимых 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.

Проблема состоит в отсутствии открытого ПО, эффективно решающего задачу интерактивного скаффолдинга для геномов, сравнимых по размеру с геномом человека, на основе данных Hi-C. Единственный программный инструмент с открытым исходным кодом – это JBAT, разработанный в лаборатории Айдена. Данное ПО обладает рядом недостатков: нерациональным расходом оперативной памяти, отсутствием возможности сразу же получить нуклеотидную последовательность ДНК, соответствующую произвольной области, а также долгим процессом финализации конечной сборки. Таким образом, актуальной является разработка ПО для интерактивного скаффолдинга геномных сборок, что в дальнейшем может открыть возможности большей формализации и окончательной автоматизации процесса, в том числе с помощью применения методов ИИ.

Основная часть. В рамках данной работы мы разработали новый инструмент с открытым исходным кодом, позволяющий эффективно выполнять запросы произвольной прямоугольной области Hi-C карты, выполнять перемещение и разворот контигов, а также получать последовательность ДНК, соответствующую выделенной области.

Для этого нами был предложен новый формат хранения данных Hi-C экспериментов, для работы с которым используется структура данных «Декартово дерево по неявному ключу со случайными приоритетами и отложенными операциями». В отличие от существующего формата mcool, используемого в Cooler, инструменте визуализации Hi-C данных, матрица взаимодействий не хранится целиком, а разбита на прямоугольные блоки с ограничением по длине каждого из измерений. Границы блоков выровнены на границы контигов, что позволяет эффективно производить операции по перемещению и развороту контигов без необходимости учёта соседних. В отличие от специализированного бинарного формата hic, разработанного в лаборатории Айдена для использования с JBAT, нами в качестве контейнера был выбран HDF5 (он же используется в Cooler, однако структура хранения внутри контейнера иная), что позволяет работать с файлом практически как со сжатой папкой и делегировать операции ввода-вывода библиотекам для работы с HDF5. Данное решение позволяет без труда добавить сжатие и фрагментацию, не изменяя при этом структуру хранения данных. Декартово дерево по неявному ключу активно используется для хранения информации о расположении контигов, а отложенные операции позволяют эффективно реализовать поворот.

Графический интерфейс приложения схож с интерфейсом JBAT, что позволяет перенести пользовательский опыт. Также, были добавлены новые элементы управления для тех функций, которые отсутствуют в JBAT.

Для преобразования форматов mcool и hic в наш формат было разработано дополнительное приложение, позволяющее делать соответствующую.

Выводы. Было разработано приложение, позволяющее работать с Hi-C данными в интерактивном режиме. Поддерживаются все требуемые операции: визуализация произвольной прямоугольной области в заданном разрешении, перемещение и поворот контигов, а также финализация - возвращение готовых скаффолдов геномной сборки в формате FASTA. Также добавлена возможность оперативного вывода последовательности ДНК, соответствующей выделенной области, из связанного файла в формате FASTA. Сравнение во времени работы показало незначительное отставание нашей реализации при работе с небольшими файлами, однако на файлах размером в несколько гигабайт наблюдается сокращение времени работы запросов. При помощи конвертации поддерживается совместимость с существующими данными в форматах mcool и hic. Благодаря использованию HDF5 в качестве контейнера, сохраняется возможность последующего улучшения предложенного нами формата хранения.

В дальнейшем инструмент может быть масштабирован до универсального браузера по работе с данными по 3D-структуре ДНК, а опыт интерактивного скаффолдинга с помощью нашего ПО позволит лучше формализовать задачу и окончательно её автоматизировать с помощью ИИ.

Сердюков А.Н. (автор) Подпись

Замятин А.А. (научный руководитель) Подпись