УДК 575.112

ПРОГРАММНЫЙ КОНВЕЙЕР, НА ОСНОВЕ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ, ДЛЯ ОБНАРУЖЕНИЯ МЕЖВИДОВЫХ ГЕНОМНЫХ ПЕРЕСТРОЕК НА ОСНОВЕ ДАННЫХ HI-C

Дравгелис В. А.[1]
Научный руководитель – канд. техн. наук, доцент Муравьёв С. Б.[1]
[1]Университет ИТМО
vitdrav@itmo.ru

Введение

Структурные вариации (СВ) представляют собой крупномасштабные изменения генома, включающие делеции, дупликации, инверсии и транслокации фрагментов длиной от 50 пар нуклеотидов и более. Они оказывают существенное влияние на архитектуру хроматина, фенотипическое разнообразие и эволюцию видов. Сравнительные геномные исследования демонстрируют, что крупные хромосомные перестройки нередко сопровождают процессы видообразования [1]. Технология Hi-C, основанная на методе захвата конформации хромосом, позволяет анализировать пространственную организацию хроматина в масштабе всего генома и выявлять структурные перестройки по характерным паттернам взаимодействий вблизи точек разрыва [2]. Тем не менее большинство существующих инструментов ориентированы на внутривидовые сравнения и недостаточно эффективны при межвидовом анализе — как по полноте обнаружения, так и по точности. Вероятной причиной является неоптимальная предобработка матриц Hi-C и специфика паттернов взаимодействий вблизи межвидовых точек разрыва.

Основная часть

В настоящей работе предложен вычислительный метод обнаружения точек разрыва геномных перестроек на основе данных Hi-C. Подход сочетает статистическую предобработку матриц контактов с ансамблем из десяти свёрточных нейронных сетей (СНС), обученных распознавать паттерны, характерные для СВ.

Задача решается как бинарная классификация подматриц Hi-C: каждая из них относится либо к регулярному региону, либо к региону, содержащему точку разрыва. Подматрицы размером 48 × 48 бинов извлекаются скользящим окном вдоль главной диагонали хромосомной карты. Обнаружение осуществляется в два этапа: на первом этапе модель сканирует диагональ, на втором – для каждого положительно классифицированного окна выполняется горизонтальный поиск по соответствующей строке матрицы.

Поскольку аннотированные достаточные по объёму межвидовые наборы данных отсутствуют, обучающие данные были сгенерированы синтетически: СВ моделировались непосредственно в референсном геноме Gorilla gorilla, после чего Hi-C прочтения выравнивались на модифицированный геном, что позволило получить карты с известными координатами точек разрыва. Валидация проводилась на опубликованном наборе данных сравнительного анализа между гориллой и человеком [2].

Реализованы два режима работы: режим двух карт, когда одновременно используются Hi-C карты двух видов, и режим одной карты, когда доступна лишь одна карта. Первый режим усиливает контраст паттернов вблизи точек разрыва за счёт вычитания фонового сигнала.

Предобработка матриц включает дистанционно-зависимую нормализацию: для каждой пары индексов (i, j) вычисляется ожидаемое значение контакта в зависимости от геномного расстояния $|i - j|$, после чего применяется логарифмическое преобразование.

Это позволяет подавить шум и выявить структурно значимые сигналы. Тестирование девяти конфигураций предобработки показало, что режим двух карт с дистанционной нормализацией обеспечивает наилучшие результаты: точность (precision) 0.48, полнота (recall) 0.67, F1-мера 0.56 на тестовой выборке. В качестве сравнения лучшая из существующих моделей – EagleC2 [3], имеет F1-меру на тех же данных равную 0.24.

Выводы

Предложенный метод на основе ансамбля CHC демонстрирует высокую эффективность обнаружения точек разрыва межвидовых геномных перестроек в данных Hi-C. Установлено, что дистанционно-зависимая нормализация и совместное использование карт двух видов являются ключевыми факторами достижения высокой полноты и точности. Разработанный подход успешно обобщается с синтетических обучающих данных на реальные межвидовые наборы и может быть применён в сравнительной и эволюционной геномике для автоматизированного поиска структурных перестроек.

Литература

1. Mahmoud M., Gobet N., Cruz-Davalos D. I. et al. Structural variant calling: the long and the short of it // Genome Biology. 2019. Vol. 20. P. 246. https://doi.org/10.1186/s13059-019-1828-7.
2. Yoo D. et al. Complete sequencing of ape genomes // *Nature* . 2025. Vol. 641. P. 401-418
3. Wang X., Luan Y., Yue F. et al. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps // Science Advances. 2022. Vol. 8. P. eabn9215. doi:10.1126/sciadv.abn9215.