

Алгоритм поиска схожих последовательностей переменных частей Т-клеточных рецепторов с использованием префиксного дерева

Подсытник М. А. (ИТМО)

Научный руководитель – Власова Е. К. (ИТМО)

Введение. Т-клеточные рецепторы (TCR) играют ключевую роль в адаптивном иммунном ответе, обеспечивая распознавание антигенов за счёт своей вариабельности. Регион, имеющий наибольшую силу связывания с эпитопом антигена, называется CDR3 (Complementarity Determining Region 3). Именно эта область определяет уникальность взаимодействия TCR с эпитопом, что делает анализ её последовательностей наиболее важным в иммунологии. Сравнение CDR3 последовательностей позволяет выявлять схожие TCR, потенциально реагирующие на одни и те же антигены. Утверждается, что похожие по последовательности CDR3 имеют схожую пространственную структуру, в результате чего связываются с теми же эпитопами. В нашей работе предложен алгоритм поиска схожих CDR3 последовательностей на основе префиксного дерева (Trie).

Основная часть. Префиксное дерево создаётся на основе набора CDR3 последовательностей, где каждая аминокислотная последовательность представляется как путь от корня до листа дерева. Такое представление позволяет быстро находить общие префиксы, что позволяет значительно сократить время поиска совпадений на наборе и затраты памяти на хранение.

Поиск основан на обходе дерева с учётом заданного количества допустимых замен в последовательности одинаковой длины. Данный подход позволяет выполнять поиск всех схожих последовательностей на наборе в 1 миллион последовательностей по заданному клонотипу с одной заменой в среднем за 0,1 мс, а с двумя заменами в среднем за 2 мс. Наивный алгоритм перебора на тех же данных требует 140 и 170 мс соответственно. Алгоритм также поддерживает поиск с использованием матрицы BLOSUM, что позволяет более точно учитывать биологическую близость аминокислот при сравнении последовательностей. С этой точки зрения алгоритм является универсальным, так как может поддерживать любую заданную матрицу замен.

Для повышения эффективности реализована многопоточная обработка большого количества запросов поиска по одному дереву. Данная модификация позволила обрабатывать любое количество запросов в 5 раз быстрее последовательного выполнения для 8 потоков.

Зачастую решается задача поиска наличия соответствия, а не всех вхождений из набора. Данный функционал также был добавлен на основе уже существующего алгоритма.

Выводы. В работе предложен эффективный алгоритм поиска схожих CDR3 последовательностей на основе префиксного дерева. Алгоритм поддерживает многопоточную обработку и использование матриц замен, что делает его универсальным инструментом для анализа TCR-репертуаров в иммунологических исследованиях. В дальнейшем планируется интеграция реализованного алгоритма в библиотеку *miru*, а также реализация алгоритма, которая позволит проводить поиск соответствий внутри различных длин CDR3. Помимо этого, дальнейшие планы по проекту включают в себя создание собственных матриц замен с учетом вставок/удалений и частоты аминокислот на конкретных позициях CDR3.

Список использованных источников:

1. Mikhail Shugay et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D419–D427.
2. Anna Postovskaya et al. tcrBLOSUM: an amino acid substitution matrix for sensitive alignment of distant epitope-specific TCRs
3. Goncharov, M. et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. Nat. Methods 19, 1017–1019 (2022)
4. <https://github.com/antigenomics/mirpy>