УДК 57.085:004.415

Регион-специфичные матрицы замен для выравнивания CDR3 Т-клеточных рецепторов

Подсытник М. А. (ИТМО)
Научный руководитель – Шугай М. А. (РНИМУ им. Пирогова)
Научный руководитель – Власова Е. К. (ИТМО)

Введение. Т-клеточные рецепторы (TCR) обеспечивают распознавание эпитопов и характеризуются большим разнообразием в пределах одного организма. Наиболее вариабельной частью TCR является участок CDR3, который в наибольшей степени определяет специфичность рецептора, поскольку участвует в непосредственном контакте с эпитопом.

CDR3 формируется в результате V(D)J-рекомбинации: края последовательности происходят из V- и J-генов, а центральная N(D)N-область возникает за счёт случайных нуклеотидных вставок. Эти участки формируются разными механизмами и, как следствие, могут обладать различными свойствами.

Одна из ключевых задач анализа TCR — поиск сходных CDR3, поскольку предполагается, что похожие последовательности могут распознавать один и тот же эпитоп. Традиционно сходство оценивали с помощью редакционного расстояния или универсальных аминокислотных матриц замен, таких как BLOSUM62. Однако такие матрицы не учитывают специфические особенности формирования CDR3. В связи с этим была поставлена задача построения специализированных матриц замен для CDR3 с отдельным учётом регионов V, N(D)N и J, что позволяет более точно моделировать сходство Т-клеточных рецепторов.

Основная часть. Для построения регион-специфичных матриц замен использовались CDR3-последовательности из базы VDJdb, содержащей аннотации эпитопной специфичности и разметку V- и J-сегментов.

На первом этапе был построен граф сходства: каждый уникальный CDR3-клонотип рассматривался как вершина, а ребро проводилось между двумя последовательностями, если редакционное расстояние между ними не превышало 2 при ограничениях не более двух замен, одной вставки и одного удаления. Для выделения плотных групп применялся алгоритм Брона–Кербоша, позволяющий находить максимальные клики — полностью связные подграфы. Из найденных клик отбирались только те, которые соответствовали одному эпитопу и содержали не менее четырёх клонотипов, что позволяло сосредоточиться на устойчивых эпитоп-специфичных группах.

В каждой выделенной группе последовательности размечались на V-, N(D)N- и J-регионы. Для каждого региона отдельно строились матрицы замен по схеме, аналогичной BLOSUM: выполнялось глобальное попарное выравнивание последовательностей внутри группы, подсчитывались частоты аминокислотных замен, после чего вычислялись log-odds оценки как логарифм отношения наблюдаемой вероятности к ожидаемой. Учитывая различие длин CDR3, глобальное выравнивание позволяло корректно учитывать не только подстановки, но и события вставок и удалений, что обеспечивало получение регион-специфичных штрафов за пропуски.

Практическая применимость матриц была оценена на тестовом наборе. В качестве положительных примеров использовались CDR3 из VDJdb, специфичные к эпитопу желтой лихорадки LLWNGPMAV, а в качестве фона — последовательности с аналогичным распределением вероятности генерации (Pgen), сгенерированные моделью OLGA в объёме, превышающем исходный набор в 50 раз. Качество поиска оценивалось по метрикам Precision и Recall с построением PR-кривых. При сопоставимых значениях Recall использование регион-специфичных матриц повышало Precision с ~0,2–0,3 до ~0,5–0,7, что свидетельствует о существенном росте точности выявления эпитоп-специфичных последовательностей по сравнению с базовыми подходами.

Выводы. Экспериментальная валидация продемонстрировала, что использование регион-специфичных матриц повышает точность выявления эпитоп-специфичных CDR3, что подтверждает их преимущество над существующими методами в задачах поиска сходства.

Список использованных источников:

1. Shugay M, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificities. *Nucleic Acids Res.* 2018;46(D1):D419–D427. doi:10.1093/nar/gkx760.
2. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(22):10915–10919. doi:10.1073/pnas.89.22.10915.
3. Glanville J, et al. Identifying specificity groups in the T cell receptor repertoire. *Nat Immunol.* 2017;18(5):613–623.
4. Dash P, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature.* 2017;547(7661):89–93.
5. Sethna Z, Elhanati Y, Callan CG Jr, Walczak AM, Mora T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*. 2019;35(17):2974–2981. doi:10.1093/bioinformatics/btz035.