ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ АКАДЕМИЧЕСКИЙ УНИВЕРСИТЕТ РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Диссертация допущена к защите
Зав. кафедрой
_____А.В. Омельченко
“ ”_____2015 г.

Д И С С Е Р Т А Ц И Я

НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ

М А Г И С Т Р А

Тема: **Построение и оценка качества репертуара антител**

Направление: 03.04.01 – Прикладные математика и физика

Выполнил студент *Е.В. Старостина*
(подпись)

Руководитель:
м.н.с. *Я.Ю. Сафонова*
(подпись)

Рецензент:
м.н.с. *П.В. Добрынин*
(подпись)

Санкт-Петербург
2015 г.

Реферат

С. 39, рис. 16, табл. 3.

В данной работе представлен алгоритм построения репертуара антител с полноразмерным восстановлением последовательностей антител по данным иммуносеквенирования. Предложенный алгоритм основан на использовании графов Хэмминга. Также предложен набор метрик для оценки качества построенного репертуара.

Алгоритм построения репертуара антител реализован в программном пакете IgRepertoireConstructor. Вычисление метрик для оценки качества реализовано в виде приложения IgQUAST.

Ключевые слова: иммуносеквенирование, репертуар антител, графы Хэмминга, метрики оценки качества репертуара антител

Содержание

Введение

Мощность адаптивной иммунной системы в основном опирается на возможность генерации большого и разнообразного репертуара антигенных рецепторов (также называемых **антителами**), продуцируемых B-лимфоцитами. Антитела отвечают за распознавание антигенов, которые представляют собой части бактериальных клеток, вирусов и других микроорганизмов. Множество всех антител человека называется его репертуаром антител. В последнее время все больше ученых занимается изучением репертуара антител с использованием методов секвенирования следующего поколения, что постепенно приводит к изменению и совершенствованию нашего понимания иммунного ответа.

Информация, полученная методами секвенирования антител, может использоваться для исследований по целому ряду направлений [1]. В качестве одного из таких направлений можно называть поиск антител, специфичных к антигенам. Поиск специфичных к антигенам антител - стандартная задача при разработке лекарств. Сейчас существует множество лекарственных средств на основе моноклональных антител. За открытие процесса получения моноклональных антител Жорж Кёлер и Сезар Мильштейн получили Нобелевскую премию в 1984 году. Как пример исследования можно привести [2], где производится изоляция моноклональных антител методами секвенирования, вместо обычно используемого скрининга. В данном исследовании осуществляется иммунизация мышей антигеном, через 7 дней секвенируется репертуар антител плазматических клеток их костного мозга и осуществляется спаривание наиболее часто встречающихся V-генов тяжелой и легкой цепей. На основании этого изготавливается библиотека рекомбинантных антител и проверяется на специфичность к антигену, показано, что специфичность составляет 78%.

Также, иммуносеквенирование дает информацию для изучения адаптивного иммунного ответа, вызванного патогеном или вакцинацией. Существуют исследования, изучающие изменение репертуара антител под воздействием вакцинации

[3] или инфекций. Секвенирование репертуара антител также используется для изучении ВИЧ-инфекции [4], [5].

Еще одним важным направлением исследований является анализ репертуара антител пациентов с аутоиммунными заболеваниями, у которых вырабатывается иммунный ответ на собственные антигены организма. Изучение таких случаев может привести к лучшему пониманию механизмов возникновения нарушений иммунной регуляции, что в свою очередь, послужит шагом на пути к появлению эффективных методов лечения таких заболеваний.

Антитела формируются как результат ряда соматических генных перестроек. Этот механизм уникален для развития B-клеток и продолжается на протяжении всей жизни организма. Сформированное антитело состоит из двух цепей - тяжелой и легкой, связанных бисульфидными связями, на нем можно выделить вариабельные и константные регионы. Связывание с антигеном происходит на участке вариабельного региона. Вариабельный регион формируется как результат рекомбинации набора последовательно упорядоченных V(variable), D(diversity) и J(joining) генных сегментов. Этот процесс носит название V(D)J-рекомбинации, также в ходе этого процесса могут происходить удаление и вставка нуклеотидов на стыках между генными сегментами. Получившаяся ДНК последовательность вариабельного региона антитела имеет длину около 330-400 нуклеотидов. Сформированная B-клетка, не связавшаяся с антигеном, называется "наивной". Образовавшись в костном мозге, она покидает его и мигрирует в периферические лимфоидные ткани, такие, как лимфатические узлы. Оказавшись в лимфатическом узле, B-клетка может быть «представлена» тому или иному антигену, который она способен распознать, т.к. на поверхности B-клетки расположены антигенные рецепторы, способные к связыванию. Когда B-клетка связывается с антигеном, она активируется. Далее эта клетка претерпевает ряд процессов и может начать активно делиться либо стать клеткой памяти. В ходе деления, рецепторный локус B-клетки подвергается огромному числу мутаций с частотой в $10^5 - 10^6$ раз превышающей нормальную частоту мутаций в гено-

ме. Этот процесс носит название соматического гипермутагенеза. В основном, мутации затрагивают специальные регионы, называемые гипервариабельными (complementarity-determining region - CDR), эти регионы соответствуют участкам, отвечающим за непосредственное связывание с антигеном.

В результате описанных механизмов, разнообразие антигенных рецепторов может достигать огромных значений, для человека теоретически до 10^{13} [6]. Это число превышает общее число B-лимфоцитов в теле человека, которое составляет $\sim 1 - 2 \times 10^{11}$ [7]. Также необходимо отметить высокую повторность репертуара, которая образуется как за счет того, что число генных сегментов ограничено, так и потому, что многие уникальные антитела образованы B-клетками, произошедшими от общего предка.

Такое высокое разнообразие антител делает использование секвенирования методом Сэнгера слишком дорогим и трудоемким для использования при секвенировании репертуара антител. Развитие технологий секвенирования следующего поколения сделало возможным полномасштабное изучение репертуара антител во всем их многообразии.

Теперь вкратце рассмотрим основные технологии секвенирования следующего поколения и их особенности.

Традиционно до 2014 года ([8], [9], [10]), для секвенирования антител использовалась, в основном, технология 454. Она обладает следующими особенностями, данные приведены для GS FLX Titanium XL+ [11] :

- Ошибки вставки/удаления/замены
- Средняя длина ридов - 700 нуклеотидов
- Производительность - до 10^6 ридов за запуск
- Точность - 99.997% (для консенсуса, собранного с 15x покрытием)

Можно заметить, что производительность технологии недостаточна для покрытия всего репертуара антител. Также, известно, что обработка ошибок встав-

ки/удаления более сложна и менее эффективна, чем обработка ошибок замены. Но благодаря длине ридов, позволяющей прочитать вариабельный регион антитела целиком, до недавнего времени, исследователи повсеместно работали с данными, полученными с использованием этой технологии.

Вторая популярная технология секвенирования - Illumina. Ее особенностями являются (данные приведены для Illumina Miseq [12]):

- Ошибки замены
- Длина ридов - 300 нуклеотидов (до 2013 года - до 150)
- Производительность - до $44 - 50 \times 10^6$ ридов за запуск
- $> 70\%$ баз с вероятностью ошибки < 0.001
- Возможность генерации парных ридов, то есть ридов с известным расстоянием между ними, называемым расстоянием вставки

Таким образом, длина ридов до 300 в сочетании с использованием парных библиотек с расстоянием вставки около 500 достаточна для покрытия всего вариабельного региона антитела, что позволяет использовать подобные данные для секвенирования антител. Также стоит отметить, что технология 454 официально не поддерживается с конца 2013 года. Вкупе с лучшей производительностью и меньшей частотой ошибок по сравнению с 454 технологией, это делает Illumina наиболее подходящей технологией для секвенирования антител.

Далее секвенирование антител с использованием технологии Illumina будет упоминаться под названием иммуносеквенирование.

В данной работе представлен метод сборки репертуара антител по данным иммуносеквенирования, а также предложены метрики для оценки качества собранного репертуара.

Глава 1

Построение репертуара

1.1. Обзор существующих методов

Задача построения репертуара антител может быть сформулирована различным образом в зависимости от характера имеющихся данных и предполагаемого дальнейшего применения результатов анализа.
Можно выделить 3 основных формулировки данной задачи:

- Задача определения VDJ-классификации

  В данном случае целью является определить для каждого рида, из комбинации каких V, D и J генных сегментов было сформировано соответствующее ему антитело. Таким образом, для человека риды распределяются на 225 x 30 x 13 кластеров (т.к. для человеческого генома существует 225 - V, 30 - D и 13 - J функциональных геннов сегментов для тяжелой цепи антитела). VDJ-классификация является первым шагом анализа для многих задач, включая задачу синтеза антител на конкретную мишень. На данный момент, существует множество программ решающих эту задачу, например IgBLAST [13] и IMGT-VQUEST [14].

- Задача классификации последовательностей CDR

  Целью является кластеризация ридов по последователям CDR-регионов соответствующим им антител, в частности по CDR3, т.к. он является наиболее вариабельным и определяет специфичность антитела в наибольшей степени. Это позволяет лучше сгруппировать риды со схожей специфичностью, также такой анализ позволяет изучать динамику изменения репертуара антител человека, например, в ходе лечения какого-либо заболевания. В качестве примера приложения, решающего эту задачу, можно привести инструмент MiGEC [15].

- Задача полноразмерного восстановления последовательности антител
  Целью данной задачи является кластеризация ридов, соответствующих антителам, принадлежащих клонам одной и той же B-клетки, и восстановление последовательностей исходных антител. Результатом подобного анализа является множество всех различных последовательностей антител в образце и размер каждого клона антитела. Такие данные можно использовать, например, для изучения эволюции антител ([5], [16], [17]). Эта задача достаточно нова, т.к. до появления данных иммуносеквенирования от Illumina, исследователи были вынуждены не восстанавливать последовательности целиком, а использовать V(D)J-классификацию и последовательностями CDR-регионов для анализа репертуара. Лишь в конце мае появились статьи, описывающие инструменты MiXCR [18] и IMSEQ [19], решающую данную задачу. Данная работа не включает подробного анализа этих инструментами, поскольку они были выпущены незадолго до окончания настоящего магистерского проекта. Детальное сравнение с инструментами MiXCR и IMSEQ является планом дальнейшей работы и входит в сотрудничество лаборатории центра алгоритмической биотехнологии СПБГУ и лаборатории адаптивного иммунитета Института Биоорганической Химии.

В данной работе решается задача полноразмерного восстановления последовательностей антител, т.к. на момент начала работы не существовало ни одного решения данной задачи, а также потому что решение подобной задачи открывает широкие возможности для дальнейшего анализа развития и эволюции антител.

1.2. Постановка задачи

В данной работе представлен метод сборки репертуара с восстановлением последовательностей антител целиком. В качестве входных данных предполага-

ется использование ридов Illumina длинной около 250 с размером вставки около 400, покрывающих вариабельный регион антитела.
Собранный репертуар представляется двумя файлами:

- Файл с последовательностями клонов антител и их размером в формате FASTA с идентификаторами в формате - cluster___(идентификатор)___size___(число ридов, относящихся к антителу)

```
>cluster___21244___size___11
GAGGTGCAGCTGGTGGAGTCTGGGGGAGGCGTGGTCCAGCCTGGGAGGTCCCTGAGACTC
TCCTGTGCAGCCTCTGGATTCACCTTCAGTAGCTATGCTATGCACTGGGTCCGCCAGGCT
CCAGGCAAGGGGCTGGAGTGGGTGGCAGTTATATCATATGATGGAAGCAATAAATACTAC
GCAGACTCCGTGAAGGGCCGATTCACCATCTCCAGAGACAATTCCAAGAACACGCTGTAT
CTGCAAATGAACAGCCTGAGAGCTGAGGACACGGCTGTGTATTACTGTGCGAGAGAAGTA
GTGGGAGCTACAGGAGGCTACTACGGTATGGACGTCTGGGGCCAAGGGACCACGGTCACC
GTCTCCTCA
>cluster___21245___size___1
GAGGTGCAGCTGGTGGAGTCTGGGGGAGGCTTGGTACAGCCTGGGGGGTCCCTGAGACTC
TCCTGTGCAGCCTCTGGATTCACCTTTAGCAGCTATGCCATGAGCTGGATCCGCCAGCCC
CCAGGGAAGGGGCTGGAGTGGATTGGGAGTATCTATTATAGTGGGAGCACCTACTACAAC
CCGTCCCTCAAGAGTCGAGTCACCATATCCGTAGACACGTCCAAGAACCAGTTCTCCCTG
AAGCTGAGCTCTGTGACCGCCGCGGACACGGCCGTGTATTACTGTGCGAGAGGCAGACAT
GCTTTTGATATCTGGGGCCAAGGGACAATGGTCACCGTCTCT
```

Рис. 1.1. Пример файла с последовательностями антител

- Файл с информацией о классификации каждого рида в формате - (идентификатор рида)\t(идентификатор клона антитела)

```
262944_merged_read_MISEQ2:53    1
326457_merged_read_MISEQ2:53    2
```

Рис. 1.2. Пример файла с с информацией о классификации каждого рида

1.3. Проблемы существующих методов исправления ошибок в данных иммуносеквенирования

Фактически, задача, которая решается в данной работе, сводится к задаче исправления ошибок в ридах и кластеризации перекрывающихся ридов с отсечкой на длину минимального перекрытия. Задача исправления ошибок в ридах хорошо известна и часто выполняется как первый этап сборки генома. Существуют широко распространенные приложения для решения задачи исправления ошибок секвенирования, такие как Quake [20] и BayesHammer [21]. К сожалению, эти приложения плохо работают для данных иммуносеквенирования, поскольку идея их алгоритмов состоит в выделении коротких k-меров и поиске среди них доверенных k-меров, используемых для дальнейшей коррекции

При исправлении ошибок QUAKE опирается на глубину покрытия k-меров и считает “доверенными” k-меры, покрытие которых превышает некоторое значение. К сожалению, в результате процессов, участвующих в формировании адаптивного иммунного ответа, данным иммуносеквенирования свойственно крайне неравномерное покрытие, что приводит к некорректным результатам работы QUAKE.

BayesHammer, в свою очередь, разработан для исправления данных секвенирования одной клетки, для которых также характерно неравномерное покрытие. Он выбирает “доверенные” k-меры внутри каждой группы “похожих” k-меров, считая что внутри каждой такой группы скорее всего есть истинный, а остальные получились из него из-за ошибок. К сожалению, в данных иммуносеквенирования наблюдается много истинных похожих k-меров, что приводит к тому, что BayesHammer совершает много некорректных исправлений.

1.4. Разработанный метод

1.4.1. Подготовка данных

В качестве входных данных приложению подаются парные риды длиной около 250 нуклеотидов с расстоянием вставки около 370, как показано на рис. 1.3. В первую очередь, такие пары ридов склеиваются по наибольшему перекрытию. При наличии несовпадений в последовательностях двух ридов на перекрытии, выбирается буква с большим значением качества, что позволяет скорректировать значительную часть ошибок. В дальнейшем под ридами будут подразумеваться именно такие склеенные риды.

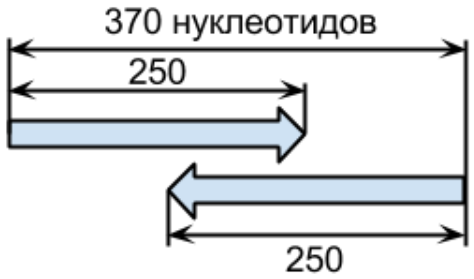

Рис. 1.3. Парные риды

1.4.2. Использование графов Хэмминга

Графом Хэмминга называют граф, в вершинах которого стоят k-меры. Расстоянием Хэмминга называется число замен, требуемое для превращения одного k-мера в другой. Ребро между двумя вершинами графа проводится в случае, если расстояние Хэмминга между k-мерами в вершинах не превышает некоторого фиксированного τ, которое задается при построении графа в качестве параметра. Вес ребра равен значению расстояния Хэмминга между вершинами. Примеры графов Хэмминга для k-меров длины 5 с $\tau = 1$ и $\tau = 2$ приведены на рис. 1.4.

Графы Хэмминга используются для коррекции ошибок в ридах в приложе-

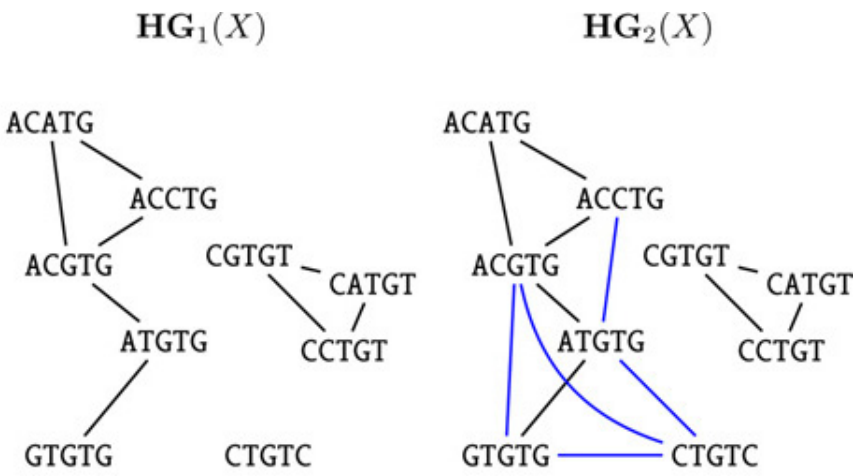

Рис. 1.4. Примеры графов Хэмминга с $k = 5, \tau = 1, 2$

нии BayesHammer [21] обычно с небольшими значениями k порядка 21. Как уже было сказано, использование коротких k-меров приводит к перекоррекции при работе с данными иммуносеквенирования. К сожалению, при увеличении размера k до значений порядка 300, что позволяет избежать этой проблемы, появляется другая сложность - BayesHammer опирается на предположение о том, что истинный k-мер обязательно должен точно присутствовать в ридах, что справедливо для небольших значений k, но с увеличением размера k это перестает быть правдой из-за наличия ошибок, так что подобная адаптация BayesHammer некорректна.

В рассматриваемом методе используются графы Хэмминга, построенные на ридах целиком. То есть, в данном случае, расстояние Хэмминга считается для такого перекрытия ридов, при котором значение расстояния меньше τ, причем существует отсечка на длину перекрытия, для тяжелых цепей антител она составляет 320. Пример такого графа можно увидеть на рис. 1.5.

Граф Хэмминга строится на $\tau = 3$, такое значение было выбрано как среднее число ошибок в контаминированных ридах (ридах, прочитанных с бактериальных геномов, попавших в образец как загрязнения).

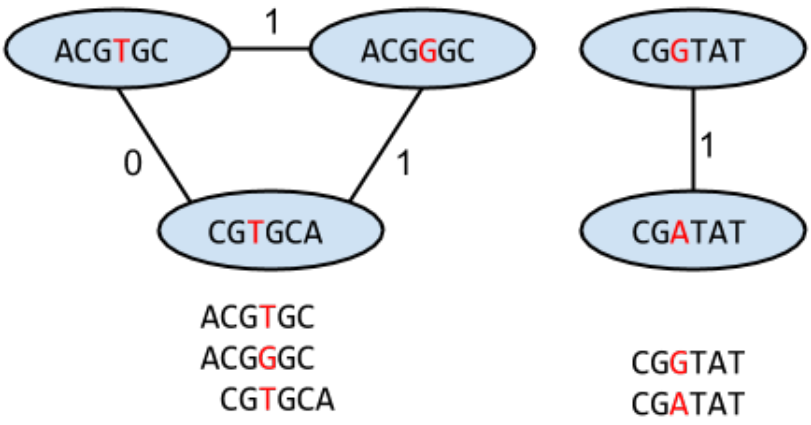

Рис. 1.5. Пример графа Хэмминга на ридах

1.4.3. Построение репертуара

Рассмотрим граф Хэмминга, построенный на данных иммуносеквенирования, как описано выше. Сделаем несколько предположений о структуре этого графа:

- Риды в разных компонентах связности графа относятся к разным клонам антител. Это предположение опирается на то, что число ошибок в риде не больше, чем значение τ, использованное при построении графа.

- Одна компонента графа может содержать риды, пришедшие из разных клонов антител. Предполагается, что разбиение ридов по компонентам графа Хэмминга довольно грубо и требуется дальнейшая более аккуратная кластеризация ридов внутри каждой компоненты графа.

- Компонента графа состоит из почти полных клик, соединенных небольшим числом ребер. Каждая почти полная клика состоит из ридов, относящихся к одному антителу. Данное предположение было сделано на основе анализа компонент связности Хэмминг графа и множественного выравнивания последовательностей внутри каждой компоненты. Примеры таких компонент можно увидеть на рис. 1.6.

Таким образом, для построения репертуара антител требуется:

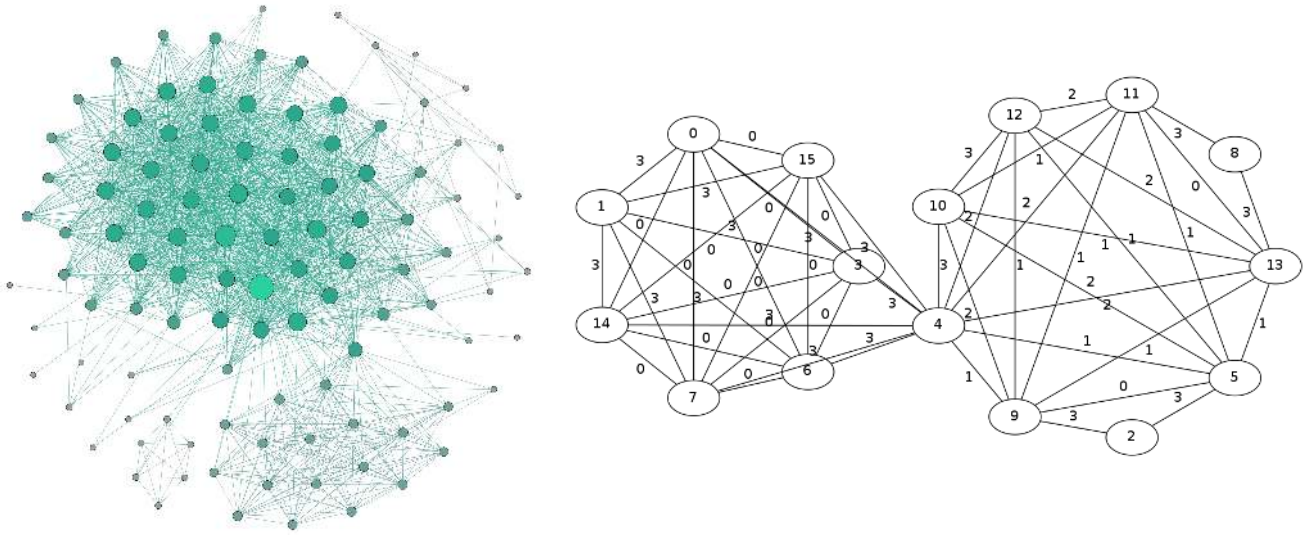

Рис. 1.6. Примеры компонент графов хэмминга

1. Построить граф Хэмминга на последовательностях ридов

2. Выделить компоненты графа Хэмминга

3. Найти внутри каждой компоненты почти полные клики. Алгоритм поиска почти полных клик использует идею о том, что построенные Хэмминг графы требуют добавления сравнительно небольшого количества ребер для минимальной триангуляции, что можно использовать для выделения максимальных клик и, как следствие, почти полных клик. Данный алгоритм не является частью данной работы, поэтому не приводится в этом тексте. Полное описание можно найти в статье [22] Алгоритм поиска почти полных клик не является частью данной работы.

4. Для каждой найденной почти полной клики построить консенсус по последовательностям. Этот консенсус и будет итоговой последовательностью антитела.

5. Представленность антитела считается как число ридов, относящихся к нему. Пример такого консенсуса приведен на рис. ниже.

6. Обработав таким образом все компоненты графа Хэмминга, получим искомый репертуар антител

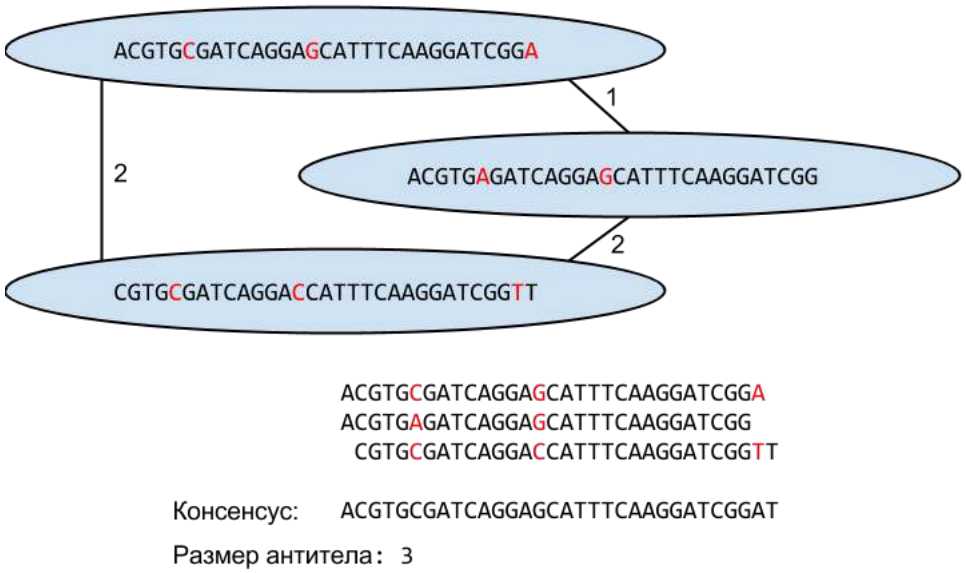

Рис. 1.7. Пример построения последовательности антитела для группы ридов

Таким образом, был разработан алгоритм, позволяющий по входным данным восстановить последовательности и размеры клонов антител. Алгоритм был реализован на C++ и Python в виде приложения IgRepertoireConstructor.

Глава 2

Оценка качества сборки репертуара антител

2.1. Подходы к решению проблемы

Задача оценивания качества сборки геномных последовательностей давно известна. Для ее решения существуют различные приложения, такие как QUAST [23] и GAGE [24]. Однако, метрики, используемые для оценки геномных сборок, не подходят для оценки сборки репертуара антител.

Задача сборки репертуара антител достаточно нова и еще не существует приложений, оценивающих ее качество, равно и как устоявшегося набора метрик.

Напомню, что результатом сборки репертуара антител является набор последовательностей клонов антител с их размерами (числом ридов, к ним относящихся) и информация о том, к какому антителу относится каждый рид из исходных данных. При этом длина последовательности одного антитела составляет порядка 330-400 нуклеотидов.

В зависимости от имеющихся данных, можно выделить три основных подзадачи в задаче оценки качества:

1. Оценка качества единичного репертуара
2. Сравнение нескольких репертуаров
3. Сравнение с идеальным репертуаром

Также отмечу, что так как, фактически, задача оценки качества репертуара антител сводится к оценке качества кластеризации ридов, то, антитела далее будут рассматриваться как кластеры ридов к ним относящихся, и называться кластерами в названиях метрик, а размер клона антитела будет называться размером кластера.

Целью данной части работы является выработать набор метрик для каждой подзадачи и реализовать их в приложении IgQUAST.

2.2. Оценка единичного репертуара

2.2.1. Постановка задачи

Допустим, имеется только один собранный репертуар. В таком случае можно лишь оценить размер репертуара и распределение размеров клонов антител.

2.2.2. Анализ соматических мутаций

Известно, что мутации в антителах происходят не равномерно и независимо, а распределены специальным образом, причем распределены так, что участки, соответствующие CDR-регионам, особенно CDR3, мутируют наиболее часто. IgQUAST строит графики с распределением мутаций по длине антитела (рис. 2.1), а также позволяет наблюдать как отличаются последовательности клонов антител друг относительно друга (рис. 2.2).

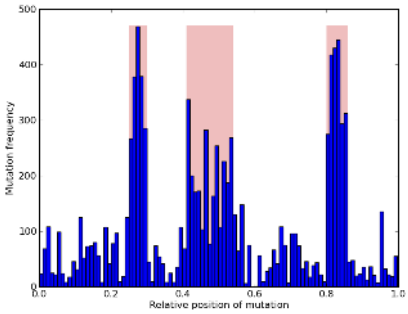

Рис. 2.1. Распределение мутаций по длине антитела, розовым выделены теоретические позиции CDRов

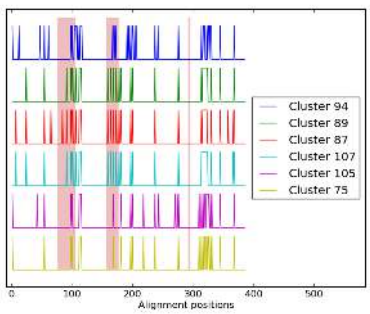

Рис. 2.2. Выравнивание группы антител относительно другого антитела, пиками показаны отличающиеся позиции, розовые столбцы - усредненные позиции CDRов из вывода IgBLAST

2.2.3. Список метрик и графиков

Метрики

- # **clusters** - число клонов антител в оцениваемом репертуаре
- # **singletons** - число клонов антител единичного размера в оцениваемом репертуаре
- **Max cluster** - размер максимального клона антитела
- **Avg cluster** - средний размер клона антитела
- # **clusters (>=N)** - число клонов антител размером не меньше N. IgQUAST использует несколько значений N: 10, 50, 100, 500 и 1000.

Показано, что размеры кластеров антител распределены по степенному закону (power law distribution) [25], а значит можно предположить, что число клонов антител единичного размера составляет более 90% от их общего числа, а средний размер клона антитела - не сильно превышает 1. Размер же максимального клона антитела может составлять до 10^6.

Графики

- Распределение длин последовательностей антител
- Распределения размеров клонов антител для всех антител/антител не единичного размера/только для достаточно больших антител (по умолчанию, размером больше 10)

2.3. Сравнение нескольких репертуаров

2.3.1. Постановка задачи

Допустим, имеется несколько разных сборок антител, построенных разными приложениями по одним входным данным. В данном случае, стоит задача

соотнести эти репертуары друг с другом, выявить и посчитать некоторые метрики их “похожести”. Это может позволить выявить проблемы каждой сборки по сравнению с другой.

2.3.2. Используемые методы

В первую очередь, для каждого репертуара считается базовый набор метрик, описанный выше. Далее, необходимо сопоставить антитела в каждом репертуаре, это можно сделать анализируя их последовательности и относящиеся к ним риды. Так IgQUAST получает группы антител, последовательности которых схожи или, которые составлены из одних и тех же ридов и вычисляет некоторые метрики для таких групп.

Сначала, IgQUAST ищет наиболее схожие последовательности для каждой пары сравниваемых репертуаров. Наиболее схожими считаются последовательности, расстояние Хэмминга между которыми минимально. Для ускорения работы приложения, используется хэширование k-меров и кандидаты на роль близких последовательностей антител выбираются среди тех, у которых есть хотя бы один общий k-мер (пример работы алгоритма показан на рис. 2.3). Размер k-мера выбирается из соображения, что при существовании отсечки на максимально рассматриваемое расстояние Хэмминга τ и средней длине последовательностей l, у двух последовательностей будет хотя бы один общий k-мер длины $\frac{l}{d+1}$, что для длины l=350 и отсечки на расстояние d=4 даст k=70.

Далее строятся графы следующим образом:

- В вершинах - клоны антител. Вершины, относящиеся к разным репертуарам, окрашены по-разному.
- Существуют ребра двух видов: первые проводятся, если последовательности антител наиболее схожи, вторые проводятся, если два антитела построены на пересекающихся множествах ридов, в таком случае ребра имеют

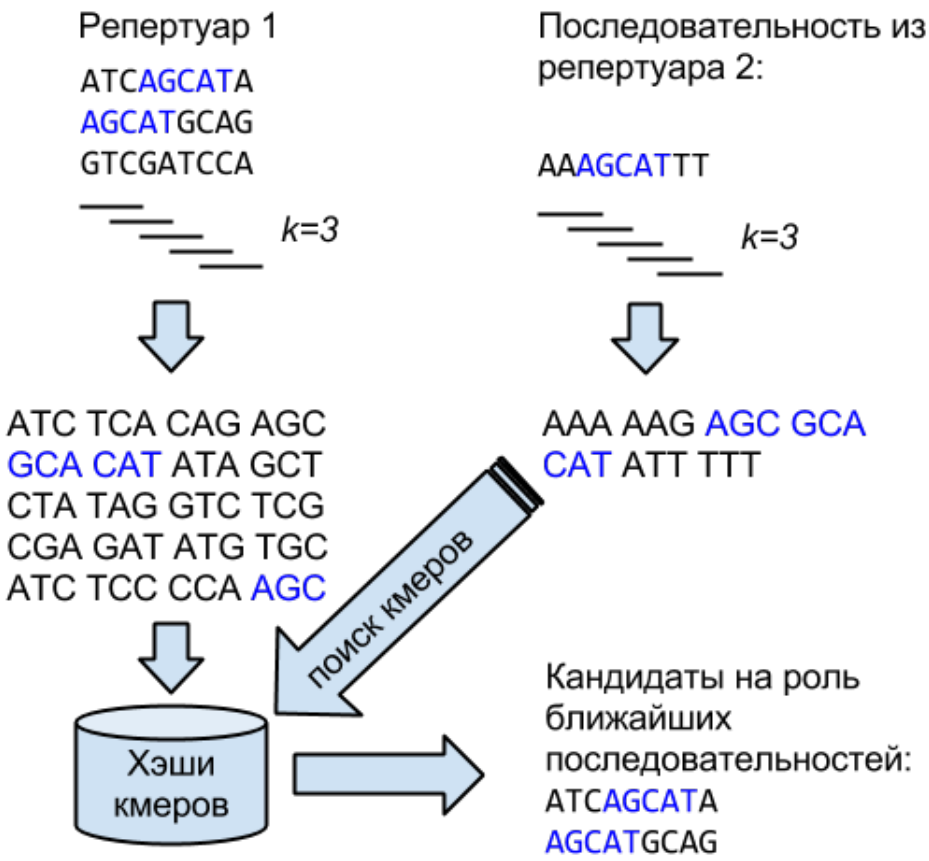

Рис. 2.3. Пример поиска кандидатов на роль близких последовательностей с использованием хэшей

вес равный размеру пересечения.

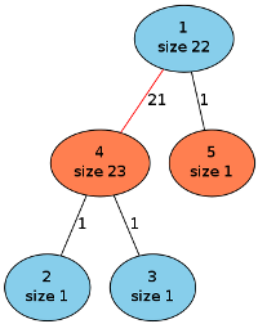

Рис. 2.4. Пример графа сравнения двух репертуаров

Пример такого графа изображен на рис. 2.4. Голубые кластеры относятся к первому репертуару, красные - ко второму. Красное ребро между вершинами 1 и 4 означает, что последовательности в вершинах близки (расстояние Хэмминга меньше отсечки d=4, по умолчанию), черные - что кластеры в вершинах построены на пересекающихся множествах ридов, с размером пересечения

равным весу ребра, но при этом последовательности кластеров не близки. На таких графах можно выделить пары вершин, построенные на одинаковых либо очень сильно пересекающихся множествах ридов. Число таких пар показывает, насколько близка кластеризация ридов в сравниваемых репертуарах. Также возможен случай, когда множества ридов, на которых построены антитела, похожи, но последовательности при этом разные, что говорит об ошибке в одном из сравниваемых репертуаров. Такие пары клонов антител называются:

- **Идеальной группой** (рис. 2.5), если клоны антител построены на одинаковом наборе ридов и их последовательности наиболее схожи
- **Доверенной группой** (рис. 2.6), если клоны антител построены на сильно пересекающихся (пересечение >90% от размера меньшего антитела) наборах ридов и их последовательности наиболее схожи
- **Недоверенной группой** (рис. 2.7), если клоны антител построены на сильно пересекающихся (пересечение >90% от размера меньшего антитела) наборах ридов и их последовательности не являются наиболее схожими

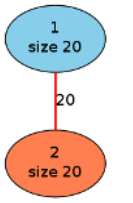

Рис. 2.5. Пример идеальной группы (1 и 2)

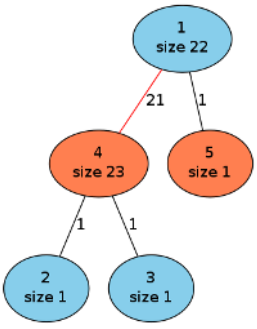

Рис. 2.6. Пример доверенной группы (1 и 4)

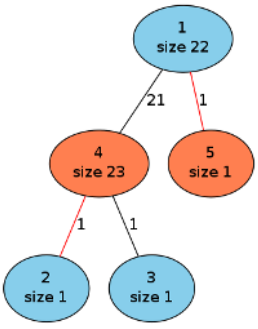

Рис. 2.7. Пример недоверенной группы (1 и 4)

Так же выделяются **изолированные клоны антител** - те, которые не имеют достаточно близкой (ближе по расстоянию Хэмминга, чем заданная отсечка) последовательности во втором репертуаре.

2.3.3. Список метрик и графиков

Общие метрики

- # **ideal groups** - число идеальных групп
- # **trusted groups** - число доверенных групп
- # **untrusted groups** - число недоверенных групп
- # **non-trivial ideal/trusted/untrusted groups** - число идеальных/доверенных/недоверенных групп не единичного размера
- # **big untrusted groups** - число недоверенных групп достаточно большого размера

В случае сравнения очень похожих репертуаров, число идеальных и доверенных групп будет примерно одинаково и его значение будет стремиться к общему числу клонов антител в наименьшем из сравниваемых репертуаров, а число не тривиальных идеальных и доверенных групп будет стремиться к числу не тривиальных клонов антител.
Не нулевое значение числа недоверенных групп говорит о том, что в одном из репертуаров скорее всего ошибочно восстановлены последовательности антител, относящиеся к этим группам.

Индивидуальные метрики для каждого из сравниваемых репертуаров

- # **isolated clusters** - число изолированных клонов антител
- # **short clusters** - число клонов антител с длиной последовательности меньше 300
- # **short isolated clusters** - число изолированных клонов антител с длиной последовательности меньше 300

- **avg/max isolated cluster size** - средний/максимальный размер изолированных клонов антител
- **# trivial isolated clusters** - число изолированных клонов антител единичного размера

Для похожих репертуаров, значение числа изолированных клонов антител для каждого репертуара будет стремиться к 0. Присутствие клонов с последовательностями короче 300 нуклеотидов может свидетельствовать о наличии ошибок в сборке.

Графики

- Распределения размеров изолированных клонов антител для изолированных всех антител/антител не единичного размера/только для достаточно больших антител(по умолчанию, размером больше 10)
- Распределения групп антител всех/включающих хотя бы один нетривиальный клон антитела

На рис. 2.8 показан пример такого графика. По оси X - степень доверия - процент ридов, которые должны иметь общими два антитела, чтобы образовывать группу (от размера меньшего кластера). Например, если есть два антитела из разных репертуаров - одно размером 10, другое - 15 и есть 7 ридов, относящихся и к первому и ко второму антителу. Тогда они будет считаться группой при степени доверия меньше или равной 70%, а на 80% - уже нет. В идеале, на таких графиках не видны оранжевые столбцы, это означает, что все группы доверенные, т.е. из пересекающихся наборов ридов собираются очень близкие последовательности.

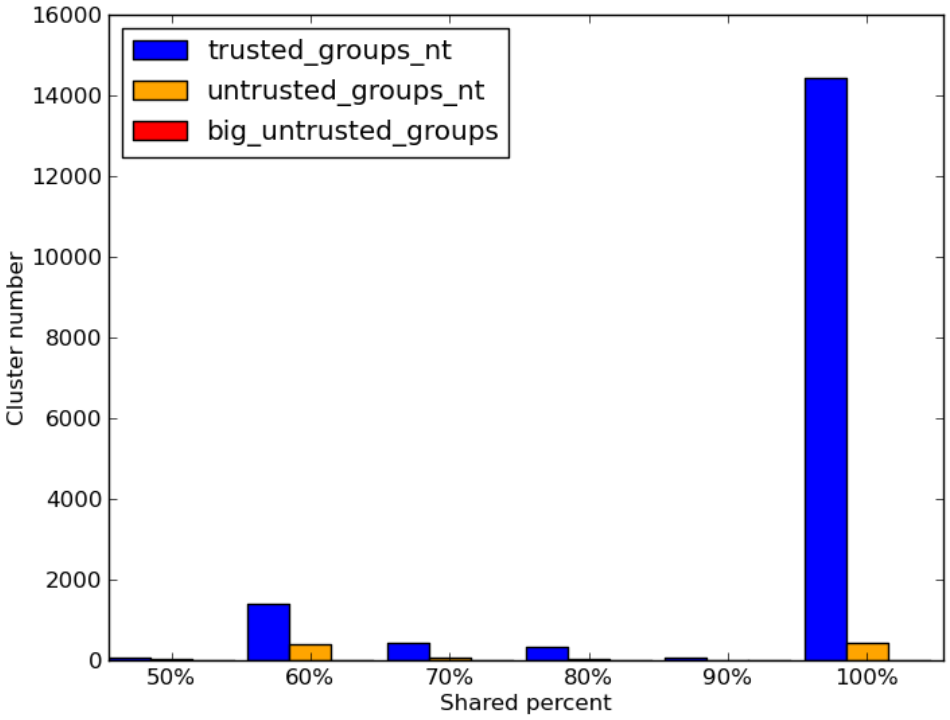

Рис. 2.8. Пример распределение групп антител

2.4. Оценка репертуара при известном идеальном

2.4.1. Постановка задачи

Допустим, идеальный репертуар известен. В таком случае, можно наилучшим образом оценить, насколько качественно собран оцениваемый репертуар. К сожалению, как правило, идеальный репертуар известен только для симулированных данных.

2.4.2. Используемые методы

Идеальный репертуар антител задает правильную кластеризацию ридов. Можно оценить качество построенного репертуара в терминах чувствительности как доли представленности “правильных” клонов антител в построенном репертуаре и в терминах специфичности как доли ошибок типа некорректно объединенных в одно антитело ридов. Также, оценивается идентичность последовательностей между антителами идеального репертуара и соответствующими

им (построенными на тех же множествах ридов) антителами построенного репертуара.

Рис. 2.9 демонстрирует различные типы отношения между клонами антител идеального и построенного репертуара. Можно видеть, что желтый клон антитела исходного репертуара распался на три разных антитела в оцениваемом, такие клоны антитела идеального репертуара считаются не объединенными и их число отражено в метрике # **not merged**. Фиолетовый клон антитела оцениваемого репертуара ошибочен, т.к. он включает риды из нескольких клонов антител идеального репертуара, число таких клонов антитела отражено в метрике # **errors**. Можно найти соответствующие друг другу антитела в идеальном и оцениваемом репертуаре так, что клону антитела идеального репертуара соответствует клон антитела в оцениваемом репертуаре, разделяющий с ним наибольшее число ридов. Так, антитело C1 на рис. ниже соответствует антителу C2. Для пар соответствующих друг другу клонов антител можно посчитать метрики идентичности и наполненности. Идентичность последовательностей двух антител считается как $\frac{o-d}{o}$, где o - длина перекрытия последовательностей, d - расстояние Хэмминга. Наполненность (**fill-in**) антитела идеального репертуара считается как отношение размера соответствующего ему антитела построенного репертуара, к размеру исходного антитела. Так, для желтого антитела идеального репертуара, наполненность будет равна N2/N1 = 4/6. Средняя наполненность (**avg fill-in**) считается как среднее по наполненности всех антител идеального репертуара, не образующих ошибок в оцениваемом репертуаре, так, для примера ниже, антитела E1 и E2 не будут учитываться при расчете этой метрики.

2.4.3. Список метрик и графиков

Метрики

- # **original clusters** - число клонов антител в идеальном репертуаре

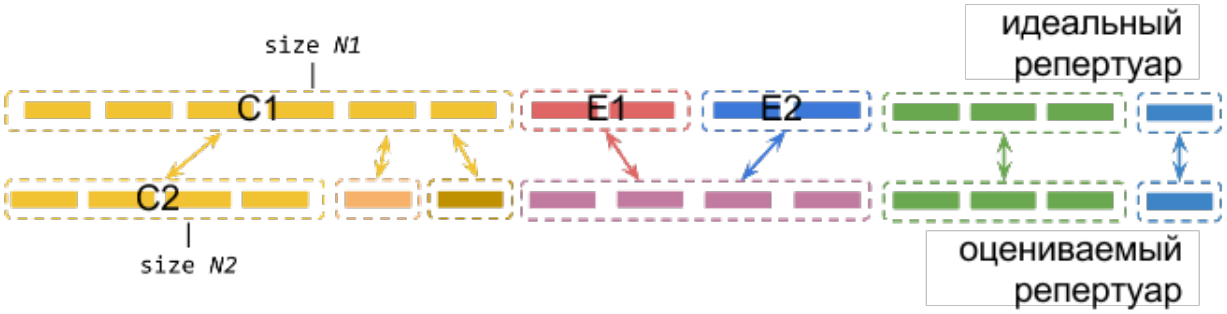

Рис. 2.9. Пример отношений между исходным репертуаром и оцениваемым

- # **not merged** - число клонов антител идеального репертуара, которые содержат несколько клонов антител в построенном. Для корректно построенного репертуара значение этой метрики будет 0
- # **not merged (not trivial + singletons)** - число клонов антител идеального репертуара, которые включают одно антитело не единичного размера и некоторое число антител единичного размера в построенном репертуаре. Большое значение этой метрики может означать, что, возможно, стоит более агрессивно объединять риды в один клон антитела
- # **original singletons** - число клонов антител единичного размера в идеальном репертуаре
- **max original cluster** - максимальный размер клона антитела в идеальном репертуаре
- # **constructed clusters** - число клонов антител в оцениваемом репертуаре
- # **errors** - число клонов антител оцениваемого репертуара, содержащих риды из более чем одного клона антитела идеального репертуара. Для корректно построенного репертуара значение этой метрики будет 0
- # **constructed singletons** - число клонов антител единичного размера в оцениваемом репертуаре

- **max constructed cluster** - максимальный размер клона антитела в оцениваемом репертуаре
- **avg fill-in** - средняя наполненность клонов антител идеального репертуара
- **fill-in of max cluster** - наполненность наибольшего клона антитела идеального репертуара
- **correct singletons (%)** - число истинных клонов антител единичного размера в построенном репертуаре относительно общего числа кластеров единичного размера
- **used reads (%)** - процент ридов использованных при построении репертуара. Возможно, алгоритм построения репертуара отфильтровал некоторые риды из входного набора данных
- **# lost clusters** - число клонов антител идеального репертуара, никак не представленных в построенном. Ненулевое значение этой метрики свидетельствует о том, что репертуар построен не полностью и были потеряны некоторые клоны антител, вероятно из-за слишком агрессивной фильтрации входных данных
- **min/avg/max percentage of identity (%)** - минимальный/средний/-максимальный процент идентичности между соответствующими последовательностями идеального и построенного репертуаров. Низкое значение этих метрик может свидетельствовать об ошибках в восстановлении последовательностей антител. При хорошем качестве сборки репертуара, минимальный процент идентичности будет близок к 100%

В случае, если оцениваемый репертуар собран хорошо, число клонов антител в нем будет стремиться к числу клонов антител в идеальном репертуаре, максимальный и средний размеры клона антитела так же будут близки для иде-

ального и оцениваемого репертуаров. Метрики наполненности в таком случае будут стремиться к 1.

Графики

- Распределение процента идентичности между соответствующими друг другу последовательностями идеального и построенного репертуаров
- Зависимость между длиной кластера и процентом идентичности

Глава 3

Результаты

3.1. Результаты сравнения с репертуаров, собранным без коррекции ошибок

Допустим, возьмем данные иммуносеквенирования и будем относить к одному клону антитела риды с достаточно большим точным перекрытием, без какого-либо исправления ошибок. Последовательности антител построим как консенсус ридов, относящихся к одному антителу. Полученный таким образом репертуар назовем наивным репертуаром.

В таблицах 3.1 и 3.2 приведены результаты сравнения наивного репертуара, с результатами IgRepertoireConstructor. В качестве входных данных использовался один из наборов данных, предоставленный компанией Genentech.

Из таблиц 3.1 и 3.2 видно, что число клонов антител, собранных IgRepertoireConstructor в полтора раза меньше, чем число клонов антител в наивном репертуаре, а средний размер клона в полтора раза больше, что означает, что исправление ошибок помогло построить отличающийся от наивного репертуар. Также, можно заметить по числу идеальных и доверенных групп, которое довольно велико относительно размера репертуара, что в целом репертуары довольно похожи, хотя большинство таких групп (кроме 997 антител, входящих в нетривиальные идеальные и доверенные группы) состоят из клонов антител единичного размера. Также видно, что в наивном репертуаре довольно много клонов антител изолированы, т.е. не имеют близкой по последовательности пары в репертуаре IgRepertoireConstructor, хотя большинство таких антител единичного размера.

На рис. 3.1 и 3.2 представлены некоторые из графики, построенных IgQUAST по итогам сравнения репертуаров.

Таблица 3.1. Общие метрики

#ideal groups	2256793
#trusted groups	2256793
#untrusted groups($>$ 4 errors)	6
#non-trivial ideal groups	997
#non-trivial trusted groups	997
#non-trivial untrusted groups($>$ 4 errors)	0
#big untrusted groups($>$ 4 errors, size $>$ 20)	0

Таблица 3.2. Метрики для репертуаров по-отдельности

	IgRepertoireConstructor	NaiveRepertoire
#clusters	2328773	3099967
#singletons	2267863	3027123
max constructed cluster size	2203	2203
avg constructed cluster size	1.445	1.086
#isolated clusters	1234	383212
#short isolated clusters($<$300bp)	1	0
min isolated cluster size	1	1
avg isolated cluster size	247.140	1.264
max isolated cluster size	33021	921
#trivial isolated clusters(size = 1)	649	361392

Видно, что размеры клонов антител IgRepertoireConstructor больше, чем в наивном репертуаре. Также видно, что характер распределения длин антител не изменился.

Из графиков 3.3 и 3.4, что практически все группы доверенные для любой степени доверия. Это говорит о том, что те исправления, которые вносятся в

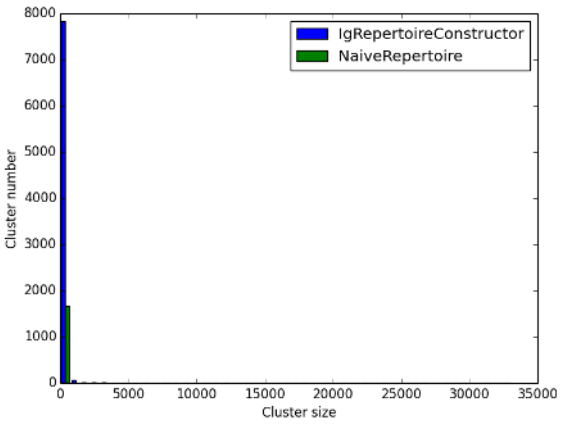

Рис. 3.1. Распределение размеров кластеров

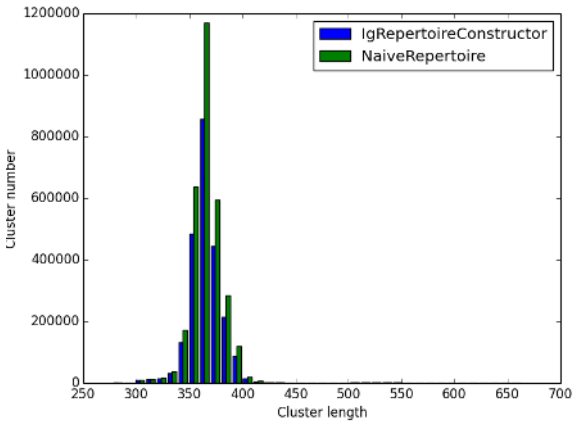

Рис. 3.2. Распределение длин кластеров

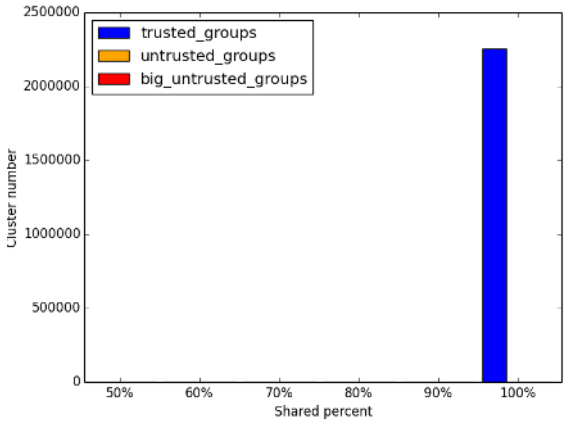

Рис. 3.3. Распределение числа доверенных / недоверенных групп

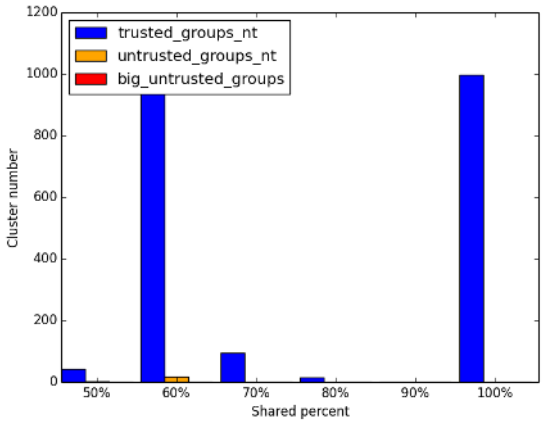

Рис. 3.4. Распределение числа доверенных / недоверенных групп не тривиального размера

риды, достаточно аккуратны, чтобы не превратить последовательности антител, построенные на сильно пересекающихся (степень доверия - отсечка на то число ридов, при котором можно считать два множества ридов сильно пересекающимися - 50, 60, 70, 100% от размера наименьшего клона антитела), что, в свою очередь, означает, что использованный подход достаточно неагрессивен и не подвержен перекоррекции.

3.2. Результаты сравнения для симулированных данных

С помощью IgSimulator [26] был сгенерирован тестовый набор данных(со следующими параметрами запуска: число базовых пар - 2000, число мутировавших последовательностей - 4000, размер репертуара - 20000). Далее, этот набор данных был использован в качестве входного для IgRepertoireConstructor и полученный результат был оценен с помощью IgQUAST по сравнению с идеальным репертуаром, сгенерированным симулятором. Результаты сравнения представлены в таблице 3.3.

Таблица 3.3. Результаты сравнения с идеальным репертуаром

#original clusters	3342
#not merged	1191
#not merged (not trivial + singletons)	699
#original singletons	1825
max original cluster	1035
#constructed clusters	8009
#errors	19
#constructed singletons	6998
max constructed cluster	848
avg fill-in	0.558
max cluster fill-in	0.819
#correct singletons	1805
used reads (%)	100
#lost clusters	0
lost clusters size (%)	0
min percentage of identity (%)	83.1
avg percentage of identity (%)	99.9
max percentage of identity (%)	100

Видно, что число построенных кластеров более чем в раза превышает истинное значение (3342 вместо 8009), что означает, что коррекция является недостаточной и есть довольно много ридов на расстоянии больше используемого по умолчанию. Значения метрик #not merged также подтверждают, что есть довольно много построенных клонов антител, которые должны быть объединены. Также, есть 6998 ридов, которые образуют антитело единичного размера, хотя корректно это только в 1805 случаях, для остальных по каким-то причинам не получилось объединить несколько ридов в одно антитело, скорее всего так случилось потому что используемая отсечка на расстояние между ридами слишком мала, либо перекрытие ридов оказалось недостаточно для отнесения их к одному клону антитела. Тем не менее, максимальный кластер восстановлен достаточно полно, почти 82% ридов, в него входящих, образуют одно антитело и в оцениваемом репертуаре.

Видно, что в большинстве случаев, последовательности восстановлены довольно точно, хотя есть клоны антител, идентичность которых в идеальном и оцениваемом репертуаре составляет около 83.1%, что может являться ошибкой алгоритма построения репертуара. Также видно, что несмотря на то, что число клонов антител велико по сравнению с идеальным, присутствуют некоторые ошибки, когда риды из разных антител некорректно отнесены к одному. Число таких ошибок - 19. При построении репертуара использованы все риды из входных данных.

Как видно из рис. 3.5, распределения размеров клонов антител для идеального и оцениваемого репертуаров - похожи. По рис. 3.6 и 3.7 видно, что хотя присутствуют антитела с низким процентом идентичности между оцениваемым и идеальным репертуарами, их немного. Также видно, что идентичность не зависит от длины последовательности антитела.

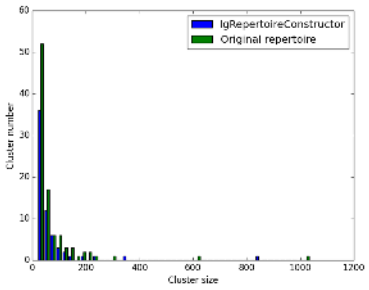

Рис. 3.5. Распределение размеров клонов антител для симулированных данных

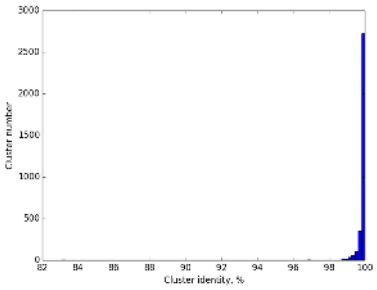

Рис. 3.6. Распределение процента идентичности между соответствующими последовательностями идеального и построенного репертуаров

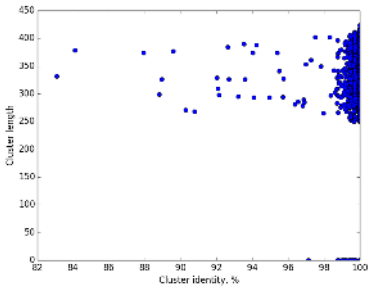

Рис. 3.7. Зависимость между длиной кластера и процентом идентичности

Заключение

В данной работе представлен новый подход к построению репертуара антител. Для восстановления последовательностей антител по данным иммуносеквенирования предлагается использовать графы Хэмминга, построенные на ридах. Данный подход обеспечивает аккуратную коррекцию ошибок и кластеризует вместе риды, относящиеся к одному клону антитела. Метод реализован на C++ и python в виде приложения IgRepertoireConstructor. Также, в данной работе предложены метрики оценки репертуара антител как отдельно, так и в сравнении с идеальным репертуаром, если он известен, либо с другими репертуарами, построенными по тем же входным данным. Подсчет метрик реализован на python в виде приложения IgQUAST. По итогам данной работы приняты статья и постер на конференцию ISMB 2015.

Литература

1. Georgiou G., Ippolito G. C., Beausang J. et al. The promise and challenge of high-throughput sequencing of the antibody repertoire // Nat Biotechnol. 2014. Vol. 32, no. 2. P. 158–168.

2. Reddy S. T., Ge X., Miklos A. E. et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells // Nat Biotechnol. 2010. Vol. 28, no. 9. P. 965–969.

3. J.D. G., A.J. P., J. T., D.F. K. Studying the antibody repertoire after vaccination: practical applications // Trends Immunol. 2014. Vol. 35, no. 7. P. 319–31.

4. J. Z., G. O., Y. Y. et al. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains // Proc Natl Acad Sci U S A. 2013. Vol. 110, no. 16. P. 6470–5.

5. Wu X., Zhou T., Zhu J. et al. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing // Science. 2011. Vol. 333, no. 6049. P. 1593–1602.

6. Murphy K. P. Janeway's immunobiology. 8th edition. Garland Science, 2012.

7. Apostoaei A., Trabalka J. Review, Synthesis, and Application of Information on the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia // SENES Oak Ridge, Inc. 2010.

8. Tan. Y., Blum L., Kongpachith S. et al. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination // Clin Immunol. 2014. Vol. 151. P. 55–65.

9. Jiang N., He J., Weinstein J. et al. Lineage structure of the human antibody

repertoire in response to influenza vaccination // Sci Transl Med. 2013. Vol. 5. P. 171ra19.

10. Baum P., Venturi V., Price D. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors // Eur J Immunol. 2012. Vol. 42. P. 2834–2839.

11. Roche. GS FLX+ System overview. 2015. URL: http://454.com/products/gs-flx-system/index.asp (дата обращения: 11.06.2015).

12. Illumina. Miseq system. 2015. URL: http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_miseq.pdf (дата обращения: 11.06.2015).

13. Ye J., Ma N., Madden T., Ostell J. IgBlast: an immunoglobulin variable domain sequence analysis tool // Nucleic Acids Res. 2013. Vol. 41. P. W34–40.

14. Brochet X., Lefranc M., Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis // Nucleic Acids Res. 2008. Vol. 36. P. W503–8.

15. Shugay M., Britanova O., Merzlyak E. et al. Towards error-free profiling of immune repertoires // Nat Methods. 2014. Vol. 11. P. 653–655.

16. W.H. R. Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery // Nat Rev Rheumatol. 2015. Vol. 11, no. 3. P. 171–82.

17. M. B., S Z. N., H E. et al. IgTree: creating Immunoglobulin variable region gene lineage trees // J Immunol Methods. 2008. Vol. 338, no. 1-2. P. 67–74.

18. Bolotin D., Poslavsky S., I. Mitrophanov M. S. et al. MiXCR: software for comprehensive adaptive immunity profiling // Nat Methods. 2015. Vol. 12. P. 380–381.

19. Kuchenbecker L., Nienen M., Hecht J. et al. IMSEQ - a fast and error aware approach to immunogenetic sequence analysis // Bioinformatics. 2015.

20. Kelley D. R., Schatz M. C., Salzberg S. L. Quake: quality-aware detection and correction of sequencing errors // Genome Biology. 2010. Vol. 11.

21. Nikolenko S., Korobeynikov A., Alekseyev M. BayesHammer: Bayesian clustering for error correction in single-cell sequencing // BMC Genomics. 2013. Vol. 14. P. S7.

22. Safonova Y., Bonissone S., Kurpilyansky E. et al. IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis // Bioinformatics. 2015.

23. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies // Bioinformatics. 2013. Vol. 29. P. 1072–1075.

24. Salzberg S. L., Phillippy A. M., Zimin A. et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms // Genome Research. 2011. Vol. 22. P. 557–567.

25. Weinstein J., Jiang N., White R. et al. High-throughput sequencing of the zebrafish antibody repertoire // Science. 2009. Vol. 324. P. 807–10.

26. Safonova Y. L. J., Lapidus A. IgSimulator: a versatile immunosequencing simulator // Bioinformatics. 2015. P. 1367–4811.