МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ)"

ФИЗТЕХ-ШКОЛА ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

КАФЕДРА ДИСКРЕТНОЙ МАТЕМАТИКИ

Выпускная квалификационная работа по направлению 01.03.02
ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА
НА ТЕМУ:

Вариационный байесовский вывод в графических моделях в задаче восстановления клональной структуры опухоли

Студент _____ Иванов В.В.
Научный руководитель к.ф-м.н. _____ Yuanhua Huang.
Зав. кафедрой д.ф-м.н., профессор _____ Райгородский А.М.

МОСКВА, 2020

1 Аннотация

Содержание

2 Обозначения, сокращения, основные определения

В силу междисциплинарного характера данной дипломной работы, автор счел уместным определить все понятия из биологии, без которых понимание работы будет затруднено или невозможно, не вдаваясь по возможности в технические детали. Для терминов, не имеющих устоявшегося перевода на русский язык, были использованы принятые в научном сообществе транслитерации.

Определение 2.1 (*Центральная догма молекулярной биологии*)**.** Наблюдаемая в природе закономерность передачи генетической информации: она распространяется от нуклеиновых кислот к белкам, вначале от ДНК к РНК в процессе **транскрипции**, а затем от РНК к белкам в процессе **трансляции**. Правило было впервые сформулировано Френсисом Криком в 1958 году.

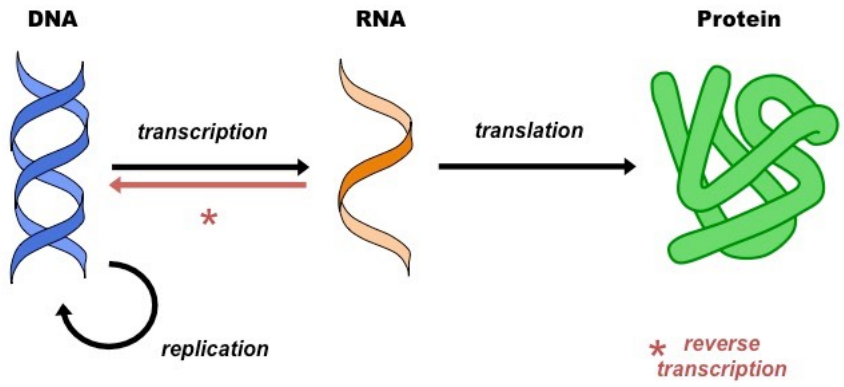

Рис. 2.1: Центральная догма молекулярной биологии

В упрощённом понимании, в процессе транскрипции участок ДНК преобразуется в т.н. **пре-матричную РНК (пре-мРНК)**, которая после **сплайсинга** — вырезания **интронов**, участков, не кодирующих белковые последовательности, — превращается в **матричную РНК (мРНК)**, которая транслируется в белковую последовательность в рибосомах.

Определение 2.2 (*Геном, экзом, транскриптом*)**.**

1. **Геном** — генетический материал клетки, нуклеотидная последовательность ДНК организма.

2. **Экзом** — набор **экзонов** организма — участков генома, кодирующих белковые последовательности.

3. **Транскриптом** — совокупность всех транскриптов, синтезируемых одной клеткой или группой клеток, включая мРНК и некодирующие РНК. Представляет собой ту часть экзома, которая преобразуется в белки в момент наблюдения, и зависит от типа клетки, стадии клеточного цикла, условий внешней среды и т.д.

Определение 2.3 (*Референсный геном*)**.** Консенсусная последовательность ДНК конкретного организма.

Определение 2.4 (*Секвенирование*)**.** Процесс определения первичной последовательности нуклеиновых кислот в клетке — ДНК или РНК. Прибор, осуществляющий секвенирование, называют **секвенатором**. Большая часть современных секвенаторов вначале дробит входную последовательность нуклеотидов на небольшие фрагменты, которые затем **амплифицируются** (многократно воспроизводятся) мащиной в ходе **ПЦР** — полимеразной цепной реакции.

Определение 2.5 (*Прочтение (рид)*)**.** Короткий нуклеотидный фрагмент, распознанный секвенатором после ПЦР. В научном сообществе чаще используется транслитерация ***рид*** от английского *sequencing read*. Набор ридов, извлечённых из образца, является основным конечным продуктом секвенирования.

Определение 2.6 (NGS)**.** Next Generation Sequencing — общее название современных методов секвенирования, позволяющих, в отличие от исторических предшественников, получать полный геном, экзом или транскриптом в ходе одного эксперимента.

Определение 2.7 (*Oxford Nanopore Sequencing*)**.** Так называемое **секвенирование нанопорами** — метод, не требующий применения ПЦР.

В контексте данной работы важно, что длина ридов, полученных по этой технологии, заметно больше, что компенсируется меньшей пропускной способностью.

Определение 2.8 (*Секвенирование одиночных клеток*)**.** Совокупность новых методов секвенирования, позволяющих извлекать нуклеотидные последовательности из каждой из клеток образца в отдельности.

Определение 2.9 (*10X Genomics*)**.** На момент написания данного текста, основной производитель технологий и программного обеспечения в нише секвенирования одиночных клеток. Представленная в работе статистическая модель проектировалась совместимой с ПО и форматом данных от 10X Genomics.

Определение 2.10 (*ОНП (снип)*)**.** Однонуклеотидный полиморфизм — позиция в геноме, на которой в статистически значимых долях популяции встречаются несколько различных вариантов нуклеотидов. Могут существенно влиять на фенотип, в том числе быть причиной патологий. В сообществе более принята траслитерация ***снип*** от английского *SNP — single nucleotide polymorphism.*

Определение 2.11 (*Гетеро- и гомозиготные ОНП*)**.** ОНП называется **гомозиготным**, если в родительских хромосомных наборах на соответствующей позиции находится один и тот же нуклеотид, и **гетерозиготным** в противном случае.

Определение 2.12 (*Гаплотипирование ОНП*)**.** Общее название набора методов для определения **гаплотипов** в геноме — непрерывных участков ДНК, содержащих полиморфизмы, обычно наследуемые вместе. В англоязычной литературе это называют *SNP phasing.*

Определение 2.13 (*Ген*)**.** Последовательность ДНК, составляющие сегменты которой не обязательно должны быть физически смежными. Эта последовательность ДНК содержит информацию об одном или нескольких продуктах в виде белка или РНК. Продукты гена функционируют в составе генетических регуляторных сетей, результат работы которых реализуется на уровне фенотипа.

Определение 2.14 (*Аллель*)**.** Вариант фрагмента ДНК, встречающийся в статистически значимой доле популяции. Частные случаи — ОНП, варианты генов.

Определение 2.15 (*Аллельный дисбаланс*)**.** Ситуация, когда один из аллелей доминирует над остальными — например, экспрессируется сильнее, более представлен в данных секвенирования и т.д.

Определение 2.16 (*CNV — структурные вариации генома*)**.** *CNV — copy number variation* — масштабные структурные модификации генома, такие как:

- **Loss events**:

  Делеция — удаление фрагмента;

- **Gain events**:

  Дупликация — удвоение фрагмента (может происходить более одного раза и порождать больше двух копий);

  Удвоение генома — удвоение числа хромосомных наборов;

- **Инверсия** — обращение непрерывного подотрезка;

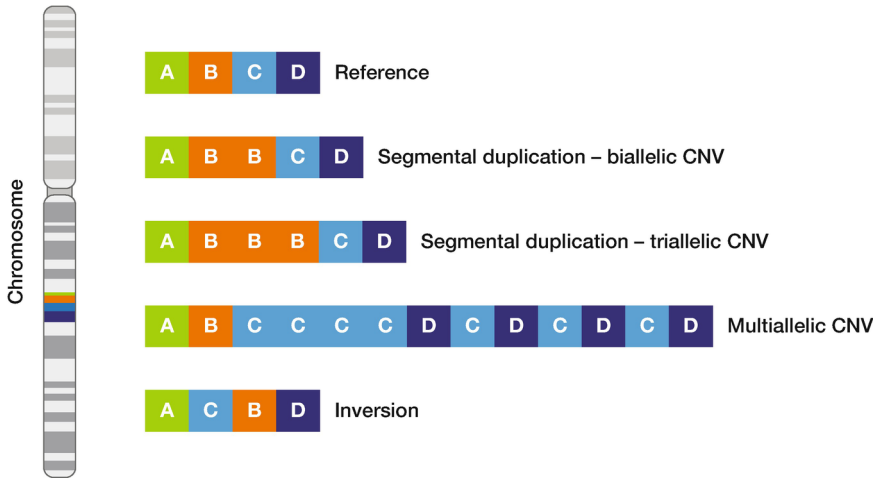

Рис. 2.2: Основные типы структурных вариаций

Разновидности структурных вариантов на этом не исчерпываются, но в контексте данной работы наибольший интерес представляет число копий

крупных фрагментов генома. Такого рода структурные вариации характерны для опухолевых клеток.

Определение 2.17 (*Медуллобластома*)**.** Самый распространённый тип педиатрической опухоли мозга. Поражает мозжечок.

Определение 2.18 (*Хромотрипсис*)**.** Мутационный процесс, в ходе которого тысячи локальных структурных вариаций случаются в небольших фрагментах генома, локализованных в одной или нескольких хромосомах. Играет важную роль в онкогенезе в отдельных типах рака и в появлении некоторых врождённых заболеваний.

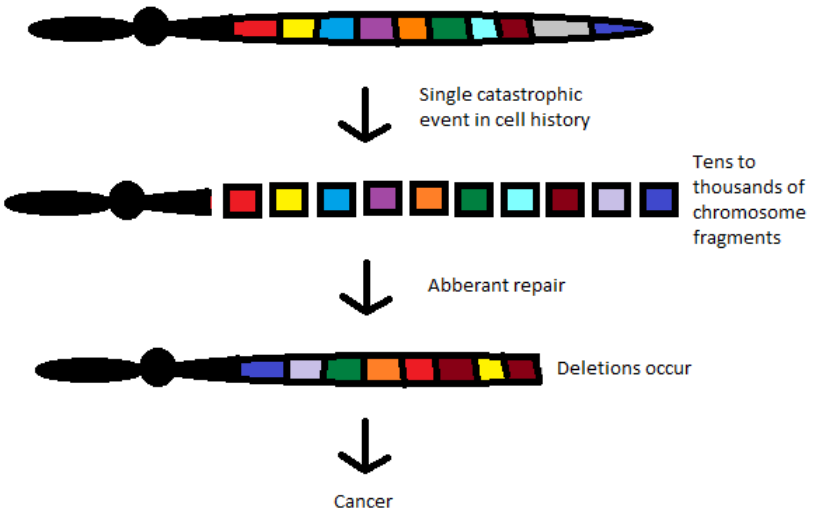

Рис. 2.3: Хромотрипсис

3 Введение

3.1 Неформальная постановка задачи, её актуальность

3.2 Краткое описание предложенного метода

3.3 Дальнейшие планы и перспективы

4 Материалы и методы

4.1 Использованные данные

4.2 Алгоритмы предобработки данных

Профессия вычислительного биолога подразумевает рутинную обработку больших гетерогенных данных, особенно что касается single-cell технологий. В связи с этим был реализован протокол предобработки данных секвенирования, основные шаги которого разобраны в данном разделе.

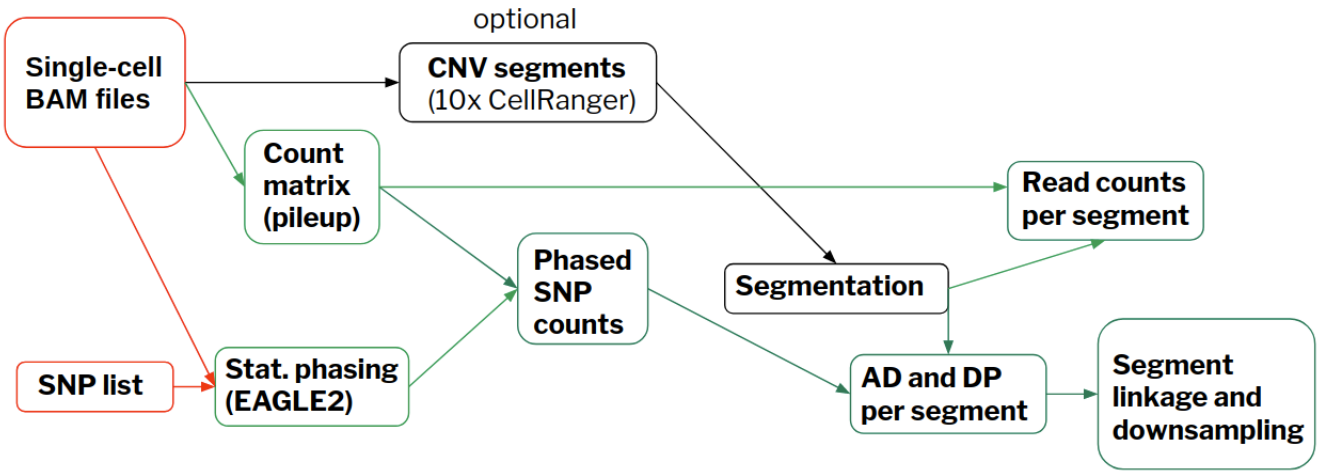

Рис. 4.1: Граф протокола предобработки данных для алгоритма XClone. Красным обозначены входные данные, чёрным — опциональные шаги, зелёным — реализованные стадии.

4.2.1 Извлечение данных из BAM-файлов

BAM — *binary SAM* — *binary sequence alignment/map format* — общепринятый формат сжатого хранения данных секвенирования. BAM-файл, полученный по протоколам 10x Genomics, занимает до нескольких терабайтов дискового пространства, потому эффективное извлечение информации из BAM-файлов это нетривиальная инженерная задача. Главные входные файлы XClone — матрицы прочтений. Таких матриц требуется три:

- матрица RD всех прочтений достаточного качества;
- матрица DP всех прочтений, накрывающих хоть один ОНП в пределах сегментов;
- матрица AD всех прочтений, накрывающих хоть альтернативный аллель ОНП в пределах сегментов.

Для получения матрицы RD из данных scRNA-seq был использован протокол **count** из **CellRanger**. Для всех остальных матриц во всех остальных случаях был использован **CellSNP**[1].

4.2.2 Статистическое фазирование гаплотипов

Гаплотипирование — определение того, от какого родителя унаследован каждый алллель в геноме — одна из ключевых задач генетики человека. Сложность её решения обусловлена контекстом, в котором она возникает в современных исследованиях, когда секвенируются порядка $2 \cdot 10^4$ — 10^6 позиций в геномах тысяч человек. Если прочтения короткие и не накрывают много позиций одновременно, то нужно секвенировать обоих родителей каждого участника эксперимента, что непрактично и не всегда возможно. Следовательно, нужно разрабатывать статистические методы гаплотипирования. Они основаны на наблюдении, что некоторые группы аллелей часто наследуются совместно. Это явление называется **неравновесной сцепленностью**. Если прогаплотипировано достаточное количество представителей популяции, то можно построить приближённые таблицы сцепленности и гаплотипировать новые образцы методом максимизации правдоподобия.

На момент написания этого текста, стандартом статистического гаплотипирования считается алгоритм **EAGLE2**[3][2]. Этот алгоритм основан на скрытых марковских моделях и использует 32,470 образца из базы данных **Haplotype Reference Consortium**[2].

[1]https://github.com/single-cell-genetics/cellSNP
[2]https://data.broadinstitute.org/alkesgroup/Eagle/

Алгоритм EAGLE2 обладает существенным недостатком: его метки имеют только локальный смысл. В пределах окна в 20-50 килобаз любые два ОНП с одинаковой наследуются совместно, но при сдвиге окна смысл меток может спонтанно поменяться на противоположный, это так называемая **ошибка смены цепи**. Т.е. два ОНП с разных концов хромосомы, помеченые одной меткой, могут быть унаследованы от разных родителей. Из-за этого в матрицах прочтений размывается сигнал аллельного дисбаланса: чтобы сделать данные менее разреженными, прочтения соседних небольших сегментов суммируются, в том числе и аллель-специфичные. Ясно, что если среди двух соседних сегментов с одинаковой меткой один полностью унаследован от отца, а второй — от матери, то при сложении их аллель-специфичные сигналы скомпенсируют друг друга. Это, в свою очередь, приводит к неправильному предсказанию аллель-специфичных структурных вариаций и неправильной кластеризации клеток. Авторы EAGLE2 в переписке явно дали понять, что в общем случае детектировать и исправлять такого рода ошибки их подход не позволяет. Но в контексте модели XClone удалось разработать статистический метод, показавший хорошие результаты при устранении ошибок переключения. Его подробное описание можно найти в одноимённом разделе.

4.2.3 Подходы к сегментации генома

Одной из основных задач XClone является предсказание **ASCNV** — аллель-специфических структурных вариаций генома. Это происходит в несколько этапов: (1) вначале определяются диапазоны, подозрительные на ASCNV, (2) затем их глубина покрытия сравнивается с эталонной для подсчёта RDR, (3) откуда получается оценка общего числа копий каждого конкретного диапазона, (4) которая затем уточняется при помощи сигналов аллельного дисбаланса. Наименее тривиальным из этих трёх шагов является первый: ни количество CNV, ни их границы заранее не известны. Тем не менее, на точность предсказания влияют ещё и технические факторы: **bin mappability** и **GC-contamination**.

Bin mappability неформально стоит понимать как долю k-меров из заданного диапазона, которые однозначно выравниваются на этот же диапазон, где k подчиняется Пуассоновской модели данных секвенирования. Если диапазон состоит из повторов одного короткого участка, то его mappability будет низкой, так как однозначно выравниваться будут только риды длиной больше половины от размера этого диапазона, вероятность которых будет мала. При заданной сегментации, эту величину можно с заданной точностью посчитать аналитически, но обычно для этого используют метод Монте-Карло.

Для отфильтровывания участков низкого качества используется **CellRanger DNA**, алгоритм[3] от 10X Genomics. Краткий обзор этого алгоритма представлен ниже:

Найденные участки накрывают некоторое подмножество референсного генома, которое затем подразделяется на сегменты размера 20-50 килобаз, в пределах которых вероятность ошибки смены цепи невелика, а потому сигнал аллельного дисбаланса статистически достоверный. Стоит отметить, что CellRanger DNA сам по себе является алгоритмом поиска CNV. Тем не менее, он сообщает, в том числе, и об участках без структурных вариантов. Благодаря этому можно гарантировать, что все участки генома, пригодные для надёжного определения ASCNV, войдут в итоговую сегментацию.

[3]https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/algorithms/cnv_calling

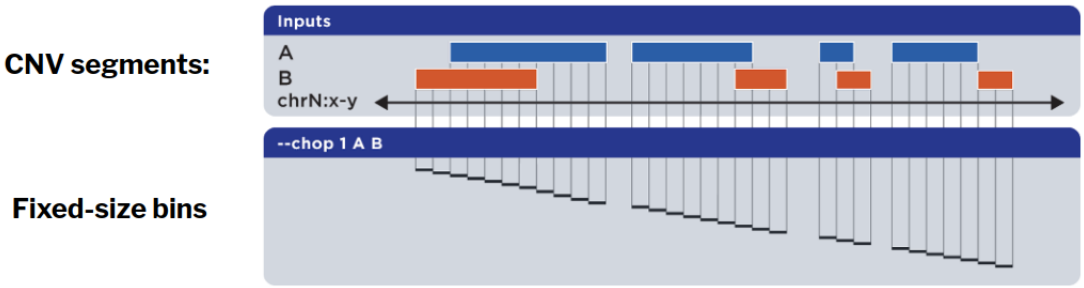

Рис. 4.2: Иллюстрация алгоритма сегментирования генома. Длина индивидуальных сегментов задаётся заранее и выбирается из диапазона 20-50 килобаз. Подряд идущие фрагменты

В силу того, что структурные вариации обычно охватывают участки генома размеров хотя бы в несколько мегабаз, перед началом предсказания уместно аггрегировать подряд идущие сегменты в блоки фиксированного размера (обычно 1-5 мегабаз), чтобы получить менее шумные BAF и RDR. Тем не менее, наивно аггрегировать содержимое сегментов внутри блока — просуммировать числа прочтений — не получится, т.к. можно потерять аллель-специфический сигнал из-за ошибок смены цепи. В связи с этим был разработан алгоритм суммирования с коррекцией ошибок, который разобран в следующем разделе.

4.2.4 Исправление ошибок смены цепи

Поскольку одной из главных задач XClone является предсказание *аллель-специфичных* структурных вариаций в геноме, матрицы AD и DP аллель-специфичных прочтений должны отражать биологию аллельного дисбаланса в клетках образца. Для этого нужно понимать, к какому гаплотипу принадлежит каждый ОНП. В разделе про статистическое гаплотипирование ОНП был сделан акцент на том, что существующие алгоритмы гарантируют только локальную корректность: при использовании алгоритма EAGLE2, следует ожидать, что при разбиении хромосомы на непересекающиеся окна длины 20-50 килобаз все гетерозиготные ОНП в пределах одного окна будут иметь одинаковый гаплотип, если это на самом деле так. Тем не менее, гаплотипы соседних сегментов с точки

зрения алгоритма могут не совпадать даже тогда, когда на самом деле должны. К этому приводят так называемые **ошибки смены цепи** — спонтанная и неявная замена гаплотипических меток на противоположные внутри алгоритма. Классификацию ошибок переключения можно найти в статье [1], цитата из которой приведена ниже:

"Phasing accuracy is typically measured by counting the number of ***'switches'*** *between known maternal and paternal haplotypes that should not occur if individual maternal and paternal chromosomal nucleotide sequence content has been accurately characterized. If an inconsistency is identified, then it is called a 'switch error.' These switch errors manifest themselves as induced and false recombination events in the inferred haplotypes compared with the true haplotypes. To identify* ***switch errors****, the phase of each site is compared with upstream neighboring phased sites. The switch error rate (SER) is defined as the number of switch errors divided by the number of opportunities for switch errors. Switch errors were further classified into three categories:* ***long****,* ***point****, and* ***undetermined****. A long switch appears as a large-scale pseudo recombination event; that is, there are no other switches in the local neighborhood around the long switch (e.g., no other switches within three consecutive heterozygous sites). On the contrary, a small-scale switch error appearing as two neighboring switch errors is considered as a point switch (e.g., two switches within three consecutive heterozygous sites, with the pair of switches counted as a point switch). The remaining switches are considered undetermined (e.g., only two sites phased in a small phasing block, so the switch error could not be classified into long or point)."*

Тем не менее, разбиение генома на фрагменты по 20-50 килобаз непрактично: в силу разреженности данных, в каждом таком сегменте может оказаться всего несколько ридов. Это даёт очень слабый и шумный сигнал аллельного дисбаланса. В связи с этим был разработан метод, одновременно решающий обе описанные проблем. На первом шаге алгоритма происходит разбиение генома на непересекающиеся сплошные сегменты длины L. Затем каждые N подряд идущих сегментов объединяются в блок длины NL. В пределах блока переключения моделируют-

ся бернуллиевскими случайными величинами, по одной на каждый сегмент. Параметры этих распределений, в свою очередь, выводятся **EM-алгоритмом**. После исправления ошибок, прочтения сегментов внутри блока суммируются, что даёт более стабильный сигнал. Эта идея была сформулирована в [5], но технические детали были осознанно исключены авторами CHISEL из препринта.

Прежде чем приступать к рассмотрению метода, сформулируем необходимые определения:

Определение 4.1 (*EM-алгоритм*)**.**
EM-алгоритм (от английского *"EM" — "Expectation Maximization"*) — метод поиска оценок максимального правдоподобия (ОМП) или оценок апостериорного максимума (ОАП) параметров статистических моделей, содержащих скрытые переменные.

Algorithm 1: EM-алгоритм в общем виде

Result: $\boldsymbol{\Theta}^*$, $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^*)$
$t = 0$;
$\boldsymbol{\Theta}^{(0)}$ инициализируется случайно;
while $Q(\boldsymbol{\Theta}^{(t+1)} \mid \boldsymbol{\Theta}^{(t+1)}) - Q(\boldsymbol{\Theta}^{(t)} \mid \boldsymbol{\Theta}^{(t)}) > \varepsilon$ **do**
  $\mathcal{L}(\boldsymbol{\Theta}^{(t)}; \boldsymbol{Z}, \boldsymbol{X}) := p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}^{(t)})$;
  $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) := \mathbb{E}_{\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\Theta}^{(t)}} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{X})$ `// E-шаг`
  $\boldsymbol{\Theta}^{(t+1)} := \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ `// M-шаг`
  $t = t + 1$
end
$\boldsymbol{\Theta}^* := \boldsymbol{\Theta}^{(t)}$

Здесь $\boldsymbol{Z}$ — дискретные скрытые переменные, $\boldsymbol{\Theta}$ — параметры статистической модели, $\boldsymbol{X}$ — выборка, $\varepsilon > 0$. Каждая итерация алгоритма состоит из двух основных шагов:

1. **E-шаг**, на котором устраняется явная зависимость от скрытых переменных посредством взятия матожидания логарифма совместной функции правдоподобия по условному распределению $\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}$;

2. **M-шаг**, на котором параметры нового апостериорного распределения $\boldsymbol{\Theta}^{(t+1)}$ выбираются таким образом, чтобы максимизировать $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$

— функцию правдоподобия "в среднем".

С теоретическим обоснованием и формальным доказательством корректности EM-алгоритма можно ознакомиться в ([4], стр. 363-365). В контексте решаемой задачи $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}$ имеют следующий смысл:

- $\boldsymbol{Z} = \{z_1, \ldots, z_N\}$ — независимые в совокупности индикаторы корректности гаплотипов сегментов

$$\forall i : z_i \sim \text{Bern}(p_i)$$

$$\forall q \in \{0,1\}^N : p(\boldsymbol{Z} = q \mid p_1, \ldots, p_n) = \prod_{i=1}^{N} p(z_i = q_i \mid p_i) = \prod_{i=1}^{N} p_i^{q_i}(1-p_i)^{1-q_i}$$

  Если $z_i = 1$, то будем говорить, что сегмент i имеет корректный гаплотип, иначе — инвертированный. Эти обозначения имеют смысл только в пределах одного блока, в соседних блоках метки могут иметь противоположный смысл. Из этого наблюдения становится ясно, что алгоритм не решает проблему переключения полностью, но уменьшает число ошибок за счёт аггрегации сегментов в блоки.

- Обозначим через M число клеток образца, тогда $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M)$, $X_c := (\boldsymbol{a}_c, \boldsymbol{b}_c)$ — вектора прочтений для каждой из клеток, по компоненте на сегмент. $\boldsymbol{a}_c = (a_{c,1}, \ldots, a_{c,N})$ — число прочтений аллеля А (альтернативный аллель), $\boldsymbol{b}_c = (b_{c,1}, \ldots, b_{c,N})$ — аллеля Б (референсный аллель).

- $\forall c \in \overline{1, M} : \boldsymbol{r}_c := \boldsymbol{a}_c + \boldsymbol{b}_c$ — вектора прочтений обоих аллелей вместе.

- $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_M; p_1, \ldots, p_N)$, где θ_c — пропорция ридов гаплотипа 1 в блоке в клетке c. Алгоритм предполагает, что пропорция гаплотипа 1 одинакова во всех сегментах внутри блока с точностью до переключения.

В этих обозначениях можно сформулировать и доказать следующее утверждение:

Утверждение 4.1. Правила пересчёта параметров апостериорного распределения на M-шаге EM-алгоритма имеют вид:

$$\begin{aligned} p_i^{(t+1)} &= \frac{p_i^{(t)} \prod_{c=1}^{M} (\theta_c^{(t)})^{a_{c,i}} (1-\theta_c^{(t)})^{b_{c,i}}}{p_i^{(t)} \prod_{c=1}^{M} (\theta_c^{(t)})^{a_{c,i}} (1-\theta_c^{(t)})^{b_{c,i}} + (1-p_i^{(t)}) \prod_{c=1}^{M} (\theta_c^{(t)})^{b_{c,i}} (1-\theta_c^{(t)})^{a_{c,i}}} \\ \theta_c^{(t+1)} &= \frac{\sum_{i=1}^{N} a_{i,c}\gamma_{i,1}^{(t)} + b_{i,c}\gamma_{i,0}^{(t)}}{\sum_{i=1}^{N} r_{i,c}} \end{aligned} \tag{1}$$

где $\forall j \in \{0,1\} : \gamma_{i,j}^{(t)} := P(z_i = j \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)})$.

Доказательство. Вектора прочтений в клетках независимы в совокупности, потому:

$$P(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\Theta}) = \prod_{c=1}^{M} p(\boldsymbol{X}_c \mid \boldsymbol{Z}, \boldsymbol{\Theta}) = \prod_{c=1}^{N} \theta_c^{\widehat{a}_c(\boldsymbol{Z})} (1-\theta_c)^{\widehat{b}_c(\boldsymbol{Z})}$$

Где

$$\begin{cases} \widehat{a}_c(\boldsymbol{Z}) := \sum_{i=1}^{N} \left[z_i a_{c,i} + (1-z_i) b_{c,i} \right], \\ \widehat{b}_c(\boldsymbol{Z}) := \sum_{i=1}^{N} \left[(1-z_i) a_{c,i} + z_i b_{c,i} \right], \\ c \in \overline{1, M} \end{cases}$$

Тогда функция правдоподобия и её логарифм принимают вид

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}) = p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\Theta}) p(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \\ &l(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \\ &= \log \prod_{\boldsymbol{q} \in \{0,1\}^N} \left(\prod_{c=1}^{M} \theta_c^{\widehat{a}_c(\boldsymbol{q})} (1-\theta_c)^{\widehat{b}_c(\boldsymbol{q})} \prod_{i=1}^{N} p_i^{q_i} (1-p_i)^{1-q_i} \right)^{\mathbb{I}\{\boldsymbol{Z}=\boldsymbol{q}\}} = \\ &= \sum_{\boldsymbol{q} \in \{0,1\}^N} \mathbb{I}\{\boldsymbol{Z} = \boldsymbol{q}\} \left(\sum_{c=1}^{M} \sum_{i=1}^{N} \widehat{a}_{c,i}(\boldsymbol{q}) \log \theta_c + \widehat{b}_{c,i}(\boldsymbol{q}) \log(1-\theta_c) \right) + \\ &+ \sum_{\boldsymbol{q} \in \{0,1\}^N} \mathbb{I}\{\boldsymbol{Z} = \boldsymbol{q}\} \left(\sum_{i=1}^{N} q_i \log p_i + (1-q_i) \log(1-p_i) \right) \end{aligned}$$

Изменением порядка суммирования можно показать, что каждая из этих двух сумм распадается на N сумм поменьше, по одной на каждую из

скрытых переменных. В следствие этого и того, что компоненты случайного вектора $\boldsymbol{Z}$ независимы в совокупности, шаги EM-алгоритма имеют вид:

E-шаг:

$$p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}) \propto p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\Theta}^{(t)}) p(\boldsymbol{Z} \mid \boldsymbol{\Theta}^{(t)}) \implies$$

$$\implies \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta}^{(t)}} l(\boldsymbol{\Theta}; \boldsymbol{Z}, \boldsymbol{X}) = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{z}_i|\boldsymbol{X}_i,\boldsymbol{\Theta}^{(t)}} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{z}_i, \boldsymbol{X}_i) =$$

$$= \sum_{i=1}^{N} \sum_{q_i=0}^{1} p(\boldsymbol{z}_i = q_i \mid \boldsymbol{X}_i, \boldsymbol{\Theta}^{(t)}) \left(\begin{array}{l} \sum\limits_{c=1}^{M} \left[\widehat{a}_{c,i}(q_i) \log \theta_c + \widehat{b}_{c,i}(q_i) \log(1-\theta_c) \right] + \\ + \log p(\boldsymbol{z}_i = q_i \mid \boldsymbol{\Theta}) \end{array} \right) =$$

$$= \sum_{i=1}^{N} \left[\begin{array}{l} \gamma_{i,1}^{(t)} \left(\sum\limits_{c=1}^{M} \left[a_{c,i} \log \theta_c + b_{c,i} \log(1-\theta_c) \right] + \log p_i \right) + \\ + \gamma_{i,0}^{(t)} \left(\sum\limits_{c=1}^{M} \left[b_{c,i} \log \theta_c + a_{c,i} \log(1-\theta_c) \right] + \log(1-p_i) \right) \end{array} \right] = Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$$

M-шаг:

$$p_i^{(t+1)} = \arg\max_{p_i} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) \iff \frac{\gamma_{i,1}^{(t)}}{p_i^{(t+1)}} - \frac{\gamma_{i,0}^{(t)}}{1 - p_i^{(t+1)}} = 0 \iff p_i^{(t+1)} = \gamma_{i,1}^{(t)}$$

$$\theta_c^{(t+1)} = \arg\max_{\theta_c} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)}) \iff \frac{\sum_{i=1}^{N} \gamma_{i,1}^{(t)} a_{c,i} + \gamma_{i,0}^{(t)} b_{c,i}}{\theta_c^{(t+1)}} - \frac{\sum_{i=1}^{N} \gamma_{i,1}^{(t)} b_{c,i} + \gamma_{i,0}^{(t)} a_{c,i}}{1 - \theta_c^{(t+1)}} = 0$$

$$\iff \theta_c^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{i,1}^{(t)} a_{c,i} + \gamma_{i,0}^{(t)} b_{c,i}}{\sum_{i=1}^{N} a_{c,i} + b_{c,i}}$$

Где необходимое условие локального экстремума является также достаточным в силу выпуклости функции $Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t)})$ ([4], стр. 363-364). □

Стоит отметить, что на практике $p_i^{(t+1)}$ следует считать по эквивалетной, но уже численно устойчивой формуле:

$$p_i^{(t+1)} = \left(1 + \exp \left[\log(1 - p_i^{(t)}) - \log(p_i^{(t)}) + \sum_{c=1}^{M} \Delta_{c,i} (\log(\theta_c^{(t)}) - \log(1 - \theta_c^{(t)})) \right] \right)^{-1}$$

Где $\Delta_{c,i} := b_{c,i} - a_{c,i}$, а показатель экспоненты стоит искусственно приводить к диапазону $[-C; C]$ для некоторого $C > 0$ (авторами было выбрано $C = 100$). В противном случае $\prod_{c=1}^{M}(\theta_c^{(t)})^{a_{c,i}}(1-\theta_c^{(t)})^{b_{c,i}}$ может представлять собой произведение тысяч или даже миллионов очень маленьких величин в больших степенях. Стандартной реализации чисел с плавающей запятой двойной точности недостаточно для хранения результатов промежуточных вычислений при использовании наивной формулы.

4.3 Первоначальная версия XClone: только ASE-модуль

4.3.1 Plate notation

4.3.2 Семплирование по Гиббсу

4.3.3 Предложенная модель, её недостатки

4.3.4 Unidentifiability problem, её решение в частном случае

4.4 Заключительная версия XClone: ASE- и RDR-модули

4.4.1 Вариационный байесовский вывод

4.4.2 Структура ASE-модуля

4.4.3 Структура RDR-модуля

4.4.4 Известные недостатки и планы по их исправлению

5 Благодарности

Данная работа посвящена приложениям байесовских методов машинного обучения к актуальной задаче вычислительной онкологии — задаче восстановления клональной структуры опухоли по данным высокопроизводительного секвенирования одиночных клеток. Такой выбор темы отражает научные интересы автора — статистическое моделирование на больших медицинских данных, — сформировавшиеся за время работы над проектами по вычислительной системной биологии под руководством к.ф.-м.н. Юрия Львовича Притыкина[4]. Автор благодарен ему за время, уделённое на протяжении двух лет работы под его руководством, и за возможность уже на младших курсах приобщиться к высоким стандартам современной академической науки.

Работа над дипломом велась под руководством Оливера Штегле[5], профессора Heidelberg University[6] (Университет Хайдельберга, Германия) а также действующих и бывших сотрудников его научных групп в DKFZ[7] (Немецкий Центр Онкологических Исследований), EMBL Heidelberg[8] (Европейская Лаборатория Молекулярной Биологии) и HKU[9] (Университет Гонконга). Формальным научным руководителем следует считать к.ф.-м.н. Yuanhua Huang[10], заведующего группой вычислительной биологии в Университете Гонконга. Автор признателен ему за профессионализм и тщательность, с которой он на протяжении многих месяцев руководил разработкой метода. Кроме того, автор выражает личную благодарность профессору Штегле и к.ф.-м.н. Ханне Сьюзак, Николе Казирахи, Родриго Гонцало Парра, а также Д. О. Бредихину и В. А. Огородникову за ценные замечания и советы, придавшие многим аспектам метода завершённый, логически стройный вид.

[4] https://scholar.google.com/citations?user=Arx56RkJBrYC&hl=en
[5] https://scholar.google.com/citations?user=ClSXZ4IAAAAJ&hl=en
[6] https://www.uni-heidelberg.de/en
[7] https://www.dkfz.de/en/index.html
[8] https://www.embl.de/
[9] https://www.hku.hk/
[10] https://www.sbms.hku.hk/staff/yuanhua-huang

Список использованных источников

[1] Y. Choi и др. «Comparison of phasing strategies for whole human genomes». В: *PLOS GENETICS* (апр. 2018). DOI: 10.1371/journal.pgen.1007308. URL: https://doi.org/10.1371/journal.pgen.1007308.

[2] Haplotype Reference Consortium. «A reference panel of 64,976 haplotypes for genotype imputation». В: *Nature Genetics* (48 авг. 2016), с. 1279—1283. DOI: 10.1038/ng.3643. URL: https://doi.org/10.1038/ng.3643.

[3] P.R. Loh и др. «Reference-based phasing using the Haplotype Reference Consortium panel». В: *Nature Genetics* (48 окт. 2016), с. 1443—1448. DOI: 10.1038/ng.3679. URL: https://doi.org/10.1038/ng.3679.

[4] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 978-0-262-01802-9.

[5] S. Zaccaria и B.J. Raphael. «Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with CHISEL». В: *bioRxiv* (нояб. 2019). DOI: 10.1101/837195. URL: https://doi.org/10.1101/837195.