

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

1	Аннотация	1
2	Обозначения, сокращения, основные определения	4
3	Введение	9
3.1	Неформальная постановка задачи, её актуальность	9
3.2	Краткое описание предложенного метода	9
3.3	Дальнейшие планы и перспективы	9
3.4	Благодарности	9
4	Материалы и методы	11
4.1	Использованные данные	11
4.2	Алгоритмы предобработки данных	11
4.2.1	Извлечение данных из BAM-файлов	11
4.2.2	Статистическое гаплотипирование ОНП	11
4.2.3	Подходы к сегментации генома	11
4.2.4	Pileup — подсчёт ридов, выравнивающихся на сегменты	11
4.2.5	Исправление ошибок переключения	11
4.3	Первоначальная версия XClone: только ASE-модуль	17
4.3.1	Plate notation	17
4.3.2	Семплирование по Гиббсу	17
4.3.3	Предложенная модель, её недостатки	17
4.3.4	Identifiability problem, её решение в частном случае через сведение к задаче поиска совершенного паросочетания	17
4.4	Заключительная версия XClone: ASE- и RDR-модули	17
4.4.1	Вариационный байесовский вывод	17
4.4.2	Структура ASE-модуля	17
4.4.3	Структура RDR-модуля	17
4.4.4	Сравнение модели с аналогами — CellRanger и CHISEL	17

4.4.5	Известные недостатки и планы по их исправлению .	17
-------	--	----

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.

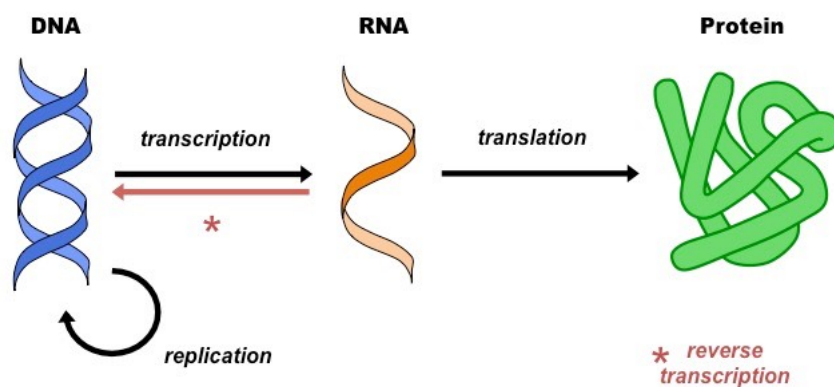


Рис. 1: Центральная догма молекулярной биологии

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 экзона, 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** (множественно воспроизводятся) машиной в ходе **ПЦР** — полимеразной цепной реакции.

Определение 2.5 (*Прочтение (rid)*). Короткий нуклеотидный фрагмент, распознанный секвенатором после ПЦР. В научном сообществе чаще используется транслитерация **rid** 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 (snip)*). 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 транслитерация **snip** 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;

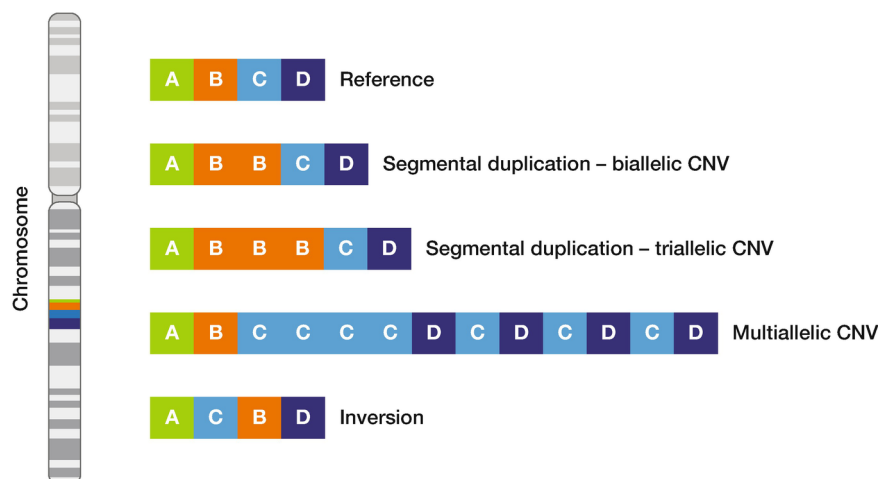


Рис. 2: Основные типы структурных вариаций

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.

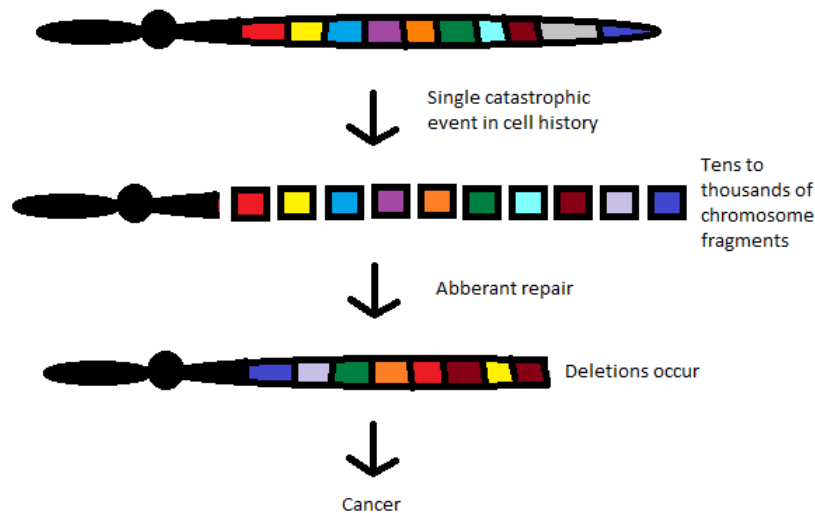


Рис. 3: Хромотрипсис

3 Введение

3.1 Неформальная постановка задачи, её актуальность

3.2 Краткое описание предложенного метода

3.3 Дальнейшие планы и перспективы

3.4 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 Притыкина¹. 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 Штегле², профессора Heidelberg University³ (Университет Хайдельберга, Германия) а также действующих и бывших сотрудников его научных групп в DKFZ⁴ (Немецкий Центр Онкологических Исследований), EMBL Heidelberg⁵ (Европейская Лаборатория Молекулярной Биологии) и HKU⁶ (Университет Гонконга). Формальным научным руководителем следует считать к.ф.-

¹<https://scholar.google.com/citations?user=Arx56RkJBrYC&hl=en>

²<https://scholar.google.com/citations?user=ClSXZ4IAAAAJ&hl=en>

³<https://www.uni-heidelberg.de/en>

⁴<https://www.dkfz.de/en/index.html>

⁵<https://www.embl.de/>

⁶<https://www.hku.hk/>

м.н. Yuanhua Huang⁷, заведующего лабораторией вычислительной биологии в Университете Гонконга. Автор признателен за тщательность, с которой он на протяжении многих месяцев руководил разработкой метода, а также за неиссякаемое терпение и оптимизм, благодаря которому мы преодолевали трудности, неизбежные при научно-исследовательской деятельности. Кроме того, автор выражает личную благодарность профессору Штегле и к.ф.-м.н. Ханне Сьюзак, Николе Казирахи, Родриго Гонцало Парра, а также Даниле Бредихину и Владимиру Огородникову за ценные замечания и советы, придавшие многим аспектам метода завершённый, логически стройный вид.

⁷<https://www.sbms.hku.hk/staff/yuanhua-huang>

4 Материалы и методы

4.1 Используемые данные

4.2 Алгоритмы предобработки данных

4.2.1 Извлечение данных из BAM-файлов

4.2.2 Статистическое гаплотипирование ОНП

4.2.3 Подходы к сегментации генома

4.2.4 Pileup — подсчёт ридов, выравнивающих на сегменты

4.2.5 Исправление ошибок переключения

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 переключения** — 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 моделируются бернуллиевскими случайными величинами, по одной на каждый сегмент. Параметры этих распределений, в свою очередь, выводятся **ЕМ-алгоритмом**. После исправления ошибок, прочтения сегментов внутри блока суммируются, что даёт более стабильный сигнал. Эта идея была сформулирована в [3], но технические детали были осознанно исключены авторами CHISEL из препринта.

Прежде чем приступить к рассмотрению метода, сформулируем необходимые определения:

Определение 4.1 (*ЕМ-алгоритм*).

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

Algorithm 1: ЕМ-алгоритм в общем виде

Result: Θ^* , $p(\mathbf{Z} \mid \mathbf{X}, \Theta^*)$

$t = 0$;

$\Theta^{(0)}$ инициализируется случайно;

while $Q(\Theta^{(t+1)} \mid \Theta^{(t+1)}) - Q(\Theta^{(t)} \mid \Theta^{(t)}) > \varepsilon$ **do**

$\mathcal{L}(\Theta^{(t)}; \mathbf{Z}, \mathbf{X}) := p(\mathbf{X}, \mathbf{Z} \mid \Theta^{(t)})$;

$Q(\Theta \mid \Theta^{(t)}) := \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}} \log \mathcal{L}(\Theta; \mathbf{Z}, \mathbf{X})$ // Е-шаг

$\Theta^{(t+1)} := \arg \max_{\Theta} Q(\Theta \mid \Theta^{(t)})$ // М-шаг

$t = t + 1$

end

$\Theta^* := \Theta^{(t)}$

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

1. **Е-шаг**, на котором устраняется явная зависимость от скрытых переменных посредством взятия математического ожидания логарифма совместной функции правдоподобия по условному распределению $\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}$;
2. **М-шаг**, на котором параметры нового апостериорного распределения $\Theta^{(t+1)}$ выбираются таким образом, чтобы максимизировать $Q(\Theta, \Theta^{(t)})$ — функцию правдоподобия "в среднем".

С теоретическим обоснованием и формальным доказательством корректности ЕМ-алгоритма можно ознакомиться в ([2], стр. 363-365). В контексте решаемой задачи $\mathbf{X}, \mathbf{Z}, \Theta$ имеют следующий смысл:

- $\mathbf{Z} = \{z_1, \dots, z_N\}$ — независимые в совокупности индикаторы кор-

ректности гаплотипов сегментов

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

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

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 агрегации сегментов в блоки.

- Обозначим через M число клеток образца, тогда $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M)$, $X_c := (\mathbf{a}_c, \mathbf{b}_c)$ — вектора прочтений для каждой из клеток, по компоненте на сегмент. $\mathbf{a}_c = (a_{c,1}, \dots, a_{c,N})$ — число прочтений аллеля А, $\mathbf{b}_c = (b_{c,1}, \dots, b_{c,N})$ — аллеля В.
- $\forall c \in \overline{1, M} : \mathbf{r}_c := \mathbf{a}_c + \mathbf{b}_c$ — вектора прочтений обоих аллелей вместе.
- $\Theta = (\theta_1, \dots, \theta_M; p_1, \dots, p_N)$, 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 М-шаге ЕМ-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 $\forall j \in \{0, 1\} : \gamma_{i,j}^{(t)} := P(z_i = j \mid \mathbf{X}, \Theta^{(t)})$.

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

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

Где

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

Тогда функция правдоподобия и её логарифм принимают вид

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

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 \mathbf{Z} независимы в совокупности, шаги ЕМ-алгоритма имеют вид:

Е-шаг:

$$\begin{aligned}
p(\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}) &\propto p(\mathbf{X} \mid \mathbf{Z}, \Theta^{(t)})p(\mathbf{Z} \mid \Theta^{(t)}) \implies \\
&\implies \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)}} l(\Theta; \mathbf{Z}, \mathbf{X}) = \sum_{i=1}^N \mathbb{E}_{\mathbf{z}_i \mid \mathbf{X}_i, \Theta^{(t)}} \log \mathcal{L}(\Theta; \mathbf{z}_i, \mathbf{X}_i) = \\
&= \sum_{i=1}^N \sum_{q_i=0}^1 p(\mathbf{z}_i = q_i \mid \mathbf{X}_i, \Theta^{(t)}) \left(\sum_{c=1}^M \left[\hat{a}_{c,i}(q_i) \log \theta_c + \hat{b}_{c,i}(q_i) \log(1 - \theta_c) \right] + \right. \\
&\quad \left. + \log p(\mathbf{z}_i = q_i \mid \Theta) \right) = \\
&= \sum_{i=1}^N \left[\gamma_{i,1}^{(t)} \left(\sum_{c=1}^M [a_{c,i} \log \theta_c + b_{c,i} \log(1 - \theta_c)] + \log p_i \right) + \right. \\
&\quad \left. + \gamma_{i,0}^{(t)} \left(\sum_{c=1}^M [b_{c,i} \log \theta_c + a_{c,i} \log(1 - \theta_c)] + \log(1 - p_i) \right) \right] = Q(\Theta \mid \Theta^{(t)})
\end{aligned}$$

М-шаг:

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

Где необходимое условие локального экстремума является также достаточным в силу выпуклости функции $Q(\Theta \mid \Theta^{(t)})$ ([2], стр. 363-364). \square

Стоит отметить, что на практике $p_i^{(t+1)}$ следует считать по эквивалентной, 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 Identifiability problem, её решение в частном случае через сведение к задаче поиска совершенного паросочетания

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 Сравнение модели с аналогами — CellRanger и CHISEL

4.4.5 Известные недостатки и планы по их исправлению

Список литературы

- [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] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 978-0-262-01802-9.
- [3] 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>.