Содержание

1 Введение **6**
1.1 Мотивация . . . 6
1.2 Актуальность . . . 6
1.3 Постановка задачи . . . 7
1.4 Биологические сведения . . . 7
1.4.1 ДНК и гистоны . . . 7
1.4.2 Хроматин-иммуннопреципитация . . . 8
1.4.3 ChIP-seq . . . 9
1.5 Статистические сведения . . . 10
1.5.1 Классический и байесовский вывод . . . 10
1.5.2 Используемые распределения . . . 10
1.5.3 Упоминаемые распределения . . . 11
1.5.4 Математическое ожидание . . . 12
1.5.5 Сопряжённое априорное семейство распределений . . . 13
1.5.6 Дивергенция Кульбака-Лейблера . . . 13
1.5.7 Информационный критерий Акаике . . . 14
1.6 Обзор литературы . . . 14
1.6.1 HHMM . . . 14
1.6.2 HPeak . . . 15
1.6.3 BayesPeak . . . 15
1.6.4 ChIPDiff . . . 15
1.6.5 Анализ подходов . . . 16
1.7 Данные . . . 16

2 Преобразование данных **17**

3 Пуассоновская смесь **18**
3.1 Мотивация . . . 18
3.2 Определение . . . 18
3.3 ОМП-вывод . . . 19
3.4 Байесовский вывод . . . 21
3.4.1 VBEM . . . 22
3.4.2 Оценка правдоподобия с помощью семплинга . . . 24
3.5 Применение . . . 25

3.5.1 ОМП-обучение 25
3.5.2 Байесовское обучение 26

4 Скрытая марковская модель с пуассоновскими наблюдениями 28
4.1 Мотивация 28
4.2 Определение 28
4.3 ОМП-вывод 29
4.3.1 Алгоритмы прямого-обратного хода и Витерби 29
4.3.2 Алгоритм Баума-Велша 30
4.4 Байесовский вывод 31
4.4.1 VBEM 31
4.4.2 Алгоритм прямого-обратного хода для VBEM 32
4.5 Применение 33
4.5.1 ОМП-обучение 33
4.5.2 Байесовское обучение 34

5 Скрытая марковская модель с многомерными пуассоновскими наблюдениями 35
5.1 Определение 35
5.2 ОМП-вывод 36
5.2.1 Алгоритмы прямого-обратного хода и Витерби 36
5.2.2 Алгоритм Баума-Велша 36
5.3 Байесовский вывод 36
5.3.1 VBEM 37
5.4 Применение 37
5.4.1 ОМП-обучение 37
5.4.2 Байесовское обучение 38

6 Ограниченная скрытая марковская модель с многомерными пуассоновскими наблюдениями 39
6.1 Мотивация 39
6.2 Определение 39
6.3 ОМП-вывод 39
6.3.1 Алгоритм Баума-Велша 40
6.4 Байесовский вывод 40

6.4.1 VBEM 40
6.5 Применение 41
6.5.1 ОМП-обучение 41
6.5.2 Байесовское обучение 42

7 BCZIPHMM **43**
7.1 Мотивация 43
7.2 Определение 44
7.3 Обучение 46

8 Результаты **47**
8.1 Реализация обучения моделей 48
8.2 Сравнение с аналогами 49

9 Заключение **49**

10 Глоссарий **52**

1 Введение

1.1 Мотивация

Связывание ДНК и белков является важнейшим биологическим явлением, которое участвует во многих процессах в клетке. Так, синтез новых биополимеров (ДНК, РНК, белков) происходит в результате взаимодействия ДНК с полимеразами, транскрипционные факторы, активаторы и репрессоры управляют активностью генов, модификации гистонов могут приводить к изменению пространственной структуры ДНК и её доступности для других белков и т. п. [1]

Одним из общепринятых методов исследования взаимодействия между ДНК и белком является хроматин-иммунопреципитационное секвенирование (*англ.* chromatin immunoprecipitation sequencing, ChIP-seq). Суть метода заключается в преципитации раствора ДНК с антителом, специфичным к интересующему нас белку, и последующим секвенированием получившихся фрагментов. Более подробно см. в разделе 1.4.

Мотивацией для этой работы послужило изучение гистонных модификаций. Гистонные модификации – важный эпигенетический маркер; их связь с клеточным ростом, процессом дифференциации и раковым поражением хорошо известна (см. [5] и многие другие статьи).

1.2 Актуальность

Известно, что все клетки организма (за исключением половых) имеют один и тот же набор наследственной информации (геном). Тем не менее, структура и функции разных клеток могут очень сильно отличаться: между нейроном головного мозга, энтероцитом кишечника и эритроцитом больше различий, чем сходств. Одна из причин такого различия – изменение уровня экспрессии определённых генов, которое, в свою очередь, может быть вызвано эпигенетическим ремоделированием хроматина или изменением активности того или иного регуляторного белка.

Таким образом, для изучения различий между клетками разных клеточных линий важно знать, как изменяется связывание белков с ДНК. Связывание белков с ДНК обычно изучается при помощи метода ChIP-seq, что приводит нас к задаче отыскания различий между двумя экспериментами

ChIP-seq. На данный момент существует много алгоритмов, исследующих один эксперимент ChIP-seq, в то время как научной базы для сравнения двух экспериментов практически нет – большинство работ, в которых производится такое сравнение, пользуется универсальными статистическими тестами. Алгоритм ChIPDiff, ставящий перед собой аналогичную цель отыскания различий, использует сомнительный критерий кратного насыщения (см. 1.6). В нашей же работе находятся качественно отличающиеся области.

1.3 Постановка задачи

Целью работы было построение адекватной математической модели, подходящей для сравнения данных двух экспериментов ChIP-seq, с целью выявления различающихся паттернов модифицирования. Очевидными требованиями к такой модели являются: правдоподобие (то есть согласованность с реальными данными), биологическая интерпретируемость и достаточно быстрое время обсчёта.

Для достижения цели были поставлены следующие задачи:

1. Изучить существующие аналоги и подходы к решению задачи (раздел 1.6).
2. Построить несколько подходящих моделей (главы 3 – 7) и произвести отбор (глава 8).
3. Сравнить выбранную модель с аналогами (глава 8).

Было изучено несколько вариантов, из которых наиболее пригодной была признана ограниченная байесовская скрытая марковская модель с пуассоновскими наблюдениями и подкачкой нуля (*англ.* Bayesian constrained zero-inflated Poisson hidden Markov model, BCZIPHMM).

1.4 Биологические сведения

1.4.1 ДНК и гистоны

Носитель наследственной информации всех клеточных форм жизни, ДНК (дезоксирибонуклеиновая кислота), представляет собой биологический полимер со сложной пространственной структурой, изменяющейся в ходе кле-

точного цикла. ДНК состоит из двух противонаправленных нитей, образующих двойную спираль. У эукариот эта двойная спираль связывается с специальными белковыми структурами (нуклеосомами), и вместе с ними образует хроматин. ДНК большинства эукариот разделена на отдельные хромосомы. [1, стр. 230]

Нуклеосомы являются белковыми гетерооктамерами, субъединицы которых называются гистонами. Гистоны – одни из самых консервативных белков, их последовательность практически одинакова у всех эукариот. Однако, гистоны часто подвергаются посттрансляционным изменениям аминокислотных остатков (напр., ацетилирование или метилирование), которые, как показано, могут влиять на структуру хроматина, экспрессию генов и другие клеточные процессы. Гистонные модификации не являются непосредственно наследуемыми, в отличие от последовательности нуклеотидов ДНК, поэтому их паттерны относятся к эпигенетической информации. [5]

1.4.2 Хроматин-иммуннопреципитация

Антителом называется белок, характерный специфическим связыванием с данным биополимером (как правило, тоже белком). Хроматин-иммуннопреципитация (*англ.* chromatin immunoprecipitation, ChIP) использует антитела, чтобы выявить участки ДНК, связанные с данным белком (гистоном, транскрипционным фактором и т. п.). Метод можно разделить на несколько логических шагов.

1. Кросс-линкинг: обработка клетки, после которой непрочные водородные связи между белками и ДНК заменяются прочными ковалентными.

2. Лизис клеток, дробление: клеточные мембраны растворяются, очищенная ДНК дробится на небольшие фрагменты с помощью рестриктаз или ультразвука.

3. Иммуннопреципитация: полученный раствор преципитируют с антителами к интересующему нас белку, закреплёнными на подложке. После этого несвязавшиеся фрагменты отмываются.

4. Отбор по размеру: из оставшихся фрагментов отбираются те, размер которых лежит в некоторых наперёд заданных границах (как правило,

около 200 пар оснований).

Полученные фрагменты требуют дальнейшей обработки, позволяющей получить о них какую-либо информацию. Такой обработкой может быть полимеразная цепная реакция (ПЦР), исследование с помощью ДНК-чипа (ChIP-chip) или секвенирование (ChIP-seq). В недавних исследованиях отдаётся предпочтение последнему варианту, так как он позволяет получить наиболее точные результаты. [1, стр. 394]

1.4.3 ChIP-seq

Классическим методом для исследования взаимодействия белков и ДНК является хроматин-иммуннопреципитационное секвенирование (*англ.* chromatin immunoprecipitation sequencing, ChIP-seq). Этот метод в качестве обработки хроматин-иммуннопреципитационных фрагментов применяет секвенирование нового поколения (*англ.* next-generation sequencing, NGS). Секвенатор читает начало случайного фрагмента, получая короткую последовательность нуклеотидов, называющуюся ридом (*англ.* read); длина рида зависит от конкретной технологии секвенирования. Эта операция повторяется достаточное количество раз, чтобы прочитать большую часть фрагментов. Риды затем выравниваются на известную референсную геномную последовательность организма; как правило, при выравнивании допускаются ошибки (максимальное количество зависит от длины рида), но при этом риды, выравнявшиеся на несколько позиций в геноме, отбрасываются, так как невозможно с точностью установить место их происхождения. После этого вместо ридов исследователи имеют дело с тегами (*англ.* tag), то есть наборами из хромосомы, позиции и ориентации цепи, на которую выравнялся рид, например, `chr16 1347124 +`. Как правило, при наличии нескольких идентичных тегов отбрасываются все, кроме одного (см., например, [2]), так как повторные идентичные риды могут быть ошибкой секвенатора.

Следует заметить, что, как и любой биологический эксперимент, ChIP-seq не даёт полностью достоверных данных. Так, нельзя гарантировать, что антитело абсолютно специфично, несвязавшиеся фрагменты были полностью отмыты, что отбор по размеру прошёл безупречно, что секвенатор не выдаёт ложных или ошибочных ридов и т. п. Именно поэтому к результатам ChIP-seq, как правило, применяются статистические методы, позволяющие

отделить значимые результаты от ошибок эксперимента.

1.5 Статистические сведения

1.5.1 Классический и байесовский вывод

Пусть модель объясняет данные d с помощью параметров θ, то есть задано распределение вероятностей на пространстве возможных данных $\mathcal{P}(d|\theta)$. Плотность этого распределения (или функция, пропорциональная ей) называется правдоподобием ($\mathcal{L}(d;\theta)$). Классический статистический вывод изучает именно эту величину; одной из стандартных задач классического вывода является нахождение оценки максимума правдоподобия (ОМП, *англ.* maximum likelihood estimate, MLE), то есть $\arg\max_\theta \mathcal{L}(d;\theta)$. Байесовский вывод, с другой стороны, изучает распределение

$$\mathcal{P}(\theta|d) = \frac{\mathcal{P}(d|\theta)\mathcal{P}(\theta)}{\int \mathcal{P}(d|\theta)\mathcal{P}(\theta)d\theta},$$

то есть апостериорное распределение параметров. Следует заметить, что в такой постановке задачи нам необходимо знать т. н. априорное распределение параметров, $\mathcal{P}(\theta)$, которое выражает наши знания о параметрах до начала эксперимента. Как правило, апостериорное распределение параметров имеет сложную форму, поэтому одной из стандартных задач байесовского вывода является приближение апостериорного распределения параметров более простым. [6]

1.5.2 Используемые распределения

В этой работе используются следующие распределения вероятностей [6] (приведены плотности распределений относительно меры Лебега или считающей меры, в зависимости от области определения):

- Биномиальное распределение:

$$\mathrm{Bin}(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k};\ k,n \in \mathbb{N}_0,\ p \in [0,1].$$

- Распределение Пуассона:

$$\mathrm{Pois}(d;\lambda) = \frac{\lambda^d e^{-\lambda}}{d!};\ d \in \mathbb{N}_0,\ \lambda \in \mathbb{R}_{\geqslant 0}.$$

Известно, что

$$\lim_{n\to+\infty, np\to\lambda} \mathrm{Bin}(d;n,p) = \mathrm{Pois}(d;\lambda),$$

более того, распределения становятся неотличимо близкими при параметрах $n \geqslant 100$, $np \leqslant 10$.

- Гамма-распределение:

$$\Gamma(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x};\ x \in \mathbb{R}_{\geqslant 0},\ \alpha,\beta \in \mathbb{R}_{>0}.$$

- Распределение Дирихле:

$$\mathrm{Dir}(x_1,\ldots,x_n;\mu_1,\ldots,\mu_n) = \frac{\Gamma(\sum_i \mu_i)}{\prod_i \Gamma(\mu_i)} x_1^{\mu_1-1}\cdots x_n^{\mu_n-1};$$

$$x_1,\ldots,x_n \in \mathbb{R}_{\geqslant 0},\ x_1+\cdots+x_n = 1,\ \mu_1,\ldots,\mu_n \in \mathbb{R}_{>0}.$$

- Бета-распределение:

$$\mathrm{B}(x;\alpha,\beta) = \mathrm{Dir}(x,1-x;\alpha,\beta);\ x \in [0,1],\ \alpha,\beta \in \mathbb{R}_{>0}.$$

- Категориальное распределение:

$$\mathbf{P}(X \in A) = \sum_{k=1}^{n} p_k \chi_A(x_k),\ p_1,\ldots,p_n \in \mathbb{R}_{\geqslant 0},\ p_1+\cdots+p_n = 1,$$

и, кроме того, все $x_1,\ldots,x_n$ различны. Здесь χ_A – характеристическая функция множества A. Исходом категориального испытания. тем самым, служит один из объектов x_k.

1.5.3 Упоминаемые распределения

В этой работе также упоминаются (но не используются в рассуждениях) следующие распределения (приведены плотности распределений относительно считающей меры):

- Обобщённое распределение Пуассона (*англ.* generalized Poisson distribution, GP):

$$\mathrm{GP}(d;\lambda,\varphi) = \left(\frac{\lambda}{1+\varphi\lambda}\right)^d \frac{(1+\varphi d)^{d-1}}{d!} \exp\left(\frac{-\lambda(1+\varphi d)}{1+\varphi\lambda}\right),$$

  см., например, [8].

- Распределение Пуассона с подкачкой нуля (*англ.* zero-inflated Poisson distribution, ZIP):

$$\mathrm{ZIP}(d;\lambda,\pi) = \begin{cases} (1-\pi) + \pi\,\mathrm{Pois}(0;\lambda), & d = 0, \\ \pi\,\mathrm{Pois}(d;\lambda), & d > 0, \end{cases}$$

  см., например, [8].

1.5.4 Математическое ожидание

В данной работе используются следующие обозначения, связанные с математическим ожиданием:

- Математическое ожидание:

$$\mathbb{E}f(X) = \int f(X)d\mathcal{P}(X) = \int f(x)p(x)dx,$$

  где $p(x)$ – плотность распределения случайной величины X.

- Математическое ожидание по распределению:

$$\mathbb{E}_{X\sim q(x)}f(X) = \int f(x)q(x)dx,$$

  где $q(x)$ – некоторая плотность распределения.

- Условное математическое ожидание:

$$\mathbb{E}_{X|Y=y}f(X) = \int f(X)d\mathcal{P}(X|Y=y) = \int f(x)\frac{p(x,y)}{\int p(x,y)dx}dx,$$

  где $p(x,y)$ – плотность совместного распределения случайных величин X и Y. В случае, когда это не вызовет путаницы, мы будем обозначать

соответствующую величину сокращённо $\mathbb{E}_{X|y}$.

1.5.5 Сопряжённое априорное семейство распределений

Пусть дано распределение с параметрами $\mathcal{P}(d|\theta)$. Семейство распределений $\mathcal{P}(\theta|\nu)$ с гиперпараметрами ν называется сопряжённым априорным к $\mathcal{P}(d|\theta)$, если

$$\forall \nu \;\; \exists \widehat{\nu} \;\; \mathcal{P}(\theta|d,\nu) = \frac{\mathcal{P}(d|\theta,\nu)\mathcal{P}(\theta|\nu)}{\mathcal{P}(d|\nu)} = \mathcal{P}(\theta|\widehat{\nu}),$$

то есть если апостериорное распределение параметров принадлежит тому же семейству. Заметим, что, в силу нормированности распределений, для этого необходимо и достаточно, чтобы

$$\forall \nu \;\; \exists \widehat{\nu} \;\; \mathcal{P}(d|\theta,\nu)\mathcal{P}(\theta|\nu) \propto \mathcal{P}(\theta|\widehat{\nu}),$$

где знак пропорциональности скрывает мультипликативную константу, не зависящую от θ. Например, гамма-распределения образуют сопряжённое априорное семество распределений к распределению Пуассона:

$$\mathrm{Pois}(d;\lambda)\Gamma(\lambda;\alpha,\beta) \propto \Gamma(\lambda;\alpha+d,\beta+1).$$

Кроме того, распределения Дирихле являются сопряжённым априорным семейством к категориальному распределению (при фиксированном наборе исходов), а бета-распределения – к биномиальному (при фиксированном n) или категориальному с двумя исходами. [6]

1.5.6 Дивергенция Кульбака-Лейблера

В статистике существует много способов определить расхождение между двумя распределениями. В данной работе используется дивергенция Кульбака-Лейблера:

$$KL(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx,$$

где p и q – плотности двух распределений. Величина $KL(q \parallel p)$ считается не определённой, если $\int_{x:q(x)=0} p(x)dx > 0$, то есть, если носитель $q(x)$ не покрывает полностью носитель $p(x)$. Известно, что $KL(q \parallel p) \geqslant 0$ для любых двух распределений и $KL(q \parallel p) = 0$ в том и только том случае, когда распределения p и q совпадают почти всюду. Дивергенция Кульбака-Лейблера

часто интерпретируется как оценка количества потерянной информации при замене распределения p на распределение q, что соответствует её применению в этой работе. [6]

1.5.7 Информационный критерий Акаике

Информационный критерий Акаике (*англ.* Akaike information criterion, AIC) применяется для выбора из нескольких моделей, обученных на одних и тех же данных методом максимального правдоподобия. Смысл критерия заключается в вычислении для каждой модели следующей величины:

$$\mathrm{AIC} = 2k - 2\ln L,$$

где k – количество степеней свободы (т. е. количество численных параметров), $L = \mathcal{L}(d; m)$ – правдоподобие данных с точки зрения выбранной модели. Критерий Акаике показывает относительную предпочтительность моделей; чем ниже вычисленная величина, тем лучше модель. Критерий Акаике штрафует модели с большим количеством параметров и тем самым защищает от переобучения, стандартной проблемы в машинном обучении. [6]

1.6 Обзор литературы

Идея использовать скрытые марковские модели для анализа разбитых на окна данных ChIP-seq сама по себе не нова. В качестве примера существующих работ на эту тему можно привести [4, HHMM], [8, HPeak], [9, BayesPeak], [10, ChIPDiff]. Далее мы кратко описываем их особенности и использованные ими подходы. Следует отметить, что все модели, кроме последней, изучают результаты одного эксперимента.

1.6.1 HHMM

Авторы [4] используют одновременно данные ChIP-seq и ChIP-chip (другой хроматин-иммуннопреципитационный метод), моделируя их с помощью двухуровневой иерархической скрытой марковской модели. Главная скрытая марковская модель отвечает за истинное состояние насыщенности или ненасыщенности региона. Две подчинённых скрытых марковских модели отвечают

за наблюдения, полученные в ходе экспериментов ChIP-chip и ChIP-seq, соответственно. Состояния подчинённых моделей являются наблюдениями главной. Алгоритм сначала обучает подчинённые модели, а затем использует результаты обучения, чтобы обучить главную. Данные ChIP-seq моделируются пуассоновским распределением с подкачкой нуля (*англ.* zero-inflated Poisson distribution, ZIP) и обобщённым распределением Пуассона (*англ.* generalized Poisson distribution, GP) (см. 1.5.3). Обучение модели производится методом максимального правдоподобия, используются различные нетрадиционные эвристики (мотивация не приводится, но, предположительно, они ускоряют сходимость).

1.6.2 HPeak

Модель [8] – скрытая марковская модель с двумя состояниями (насыщенное и ненасыщенное), причём наблюдения в насыщенном случае распределены согласно GP, а в ненасыщенном – согласно ZIP. Обучение модели производится методом максимального правдоподобия, оно основано на итеративном применении алгоритма Витерби и различных эвристик.

1.6.3 BayesPeak

В работе [9], скрытая марковская модель имеет четыре состояния, так как использует данные об обнаружении фрагментов на положительной или отрицательной цепи ДНК. Для корректной работы модели тщательно подбирается нужная ширина окна (около половины средней длины фрагмента, чтобы 5'- и 3'-концы фрагмента, покрывающего данный нуклеотид, оказались в соседних окнах). Как следует из названия, модель байесовская; оценка апостериорного распределения параметров производится с помощью семплинга алгоритмами Гиббса и Метрополиса-Гастингса. Наблюдения моделируются распределением Пуассон-гамма (*англ.* Poisson-Gamma distribution), которое идентично негативному биномиальному распределению.

1.6.4 ChIPDiff

Наконец, [10] посвящён отысканию в двух экспериментах регионов с различной насыщенностью данной модификацией (дифференциально насыщенные регионы). У модели три состояния (одинаковая насыщенность, регион более

насыщен в первом эксперименте, регион более насыщен во втором эксперименте). Для определения различий используется критерий кратной насыщенности (если отношение предполагаемых насыщенностей больше, чем пороговое значение, регион считается дифференциально насыщенным). Наблюдения моделируются с помощью усечённого бета-биномиального распределения. Модель обучается методом максимального правдоподобия с помощью алгоритма Баума-Велша (см. 4.3.2).

1.6.5 Анализ подходов

Все модели, кроме последней, исследуют результаты одного эксперимента ChIP-seq. Для отыскания различий между двумя экспериментами можно, разумеется, определить места связывания в каждом эксперименте независимо и затем найти, где они отличаются. Однако такой подход предполагает, что паттерны связывания в двух клеточных линиях независимы, что крайне сомнительно, учитывая консервативность большинства биологических явлений. Таким образом, нам необходима модель, изначально учитывающая оба эксперимента. [10, ChIPDiff] удовлетворяет этому требованию, однако критерием различия в нём является кратность отличия покрытий; например, изменение покрытия с 1% до 3% трактуется этим алгоритмом точно так же, как изменение с 30% до 90%. Мы же поставили себе цель отыскивать качественные различия в покрытии, то есть, например, изменение с высокого покрытия на низкое и наоборот.

1.7 Данные

В этой работе использовались данные, полученные из базы данных NCBI GEO [11], а именно, результаты серии экспериментов GSE25308 [12], подробно описанной в статье [2]. Эксперименты состояли из хроматин-имуннопреципитационного секвенирования по разным модификациям гистонов клеточных линий миобластов и миоцитов мыши (*M. musculus*). Поскольку миоциты являются дифференцированной формой миобластов, это исследование позволяет изучить связь между гистонными модификациями и дифференцированием клетки. Как и все эксперименты из базы данных NCBI GEO, GSE25308 находится в открытом доступе по ссылке [12].

2 Преобразование данных

Нашей целью было исследование результатов эксперимента ChIP-seq. Заключительный шаг физической части метода, секвенирование, производит большое количество коротких ридов, которые потом преобразуются в теги с помощью выравнивания. Мы ищем участки генома с высокой степенью связывания с интересующим нас белком. Мы будем называть такие участки *насыщенными.*

Сделаем несколько упрощающих предположений. Во-первых, мы будем изучать *агрегированное покрытие* генома. Для этого мы разобъём каждую хромосому на неперекрывающиеся окна одинакового размера и подсчитаем количество тегов, попавших в каждое окно. Таким образом, входными данными для нашего алгоритма будет последовательность неотрицательных целых чисел.

Далее, мы предполагаем, что наличие тега в каждой данной позиции не зависит от наличия тега в другой позиции. Это предположение опирается на равномерность фрагментации ДНК на стадии дробления и на равномерность выбора фрагмента секвенатором. Наконец, мы предполагаем, что вероятность наблюдения тега в данной позиции одинакова для всех нуклеотидов одного окна, так как они расположены физически близко. Оба этих предположения, вероятнее всего, в целом верны.

Из сделанных нами предположений легко видеть, что покрытие каждого окна подчиняется биномиальному распределению $\mathrm{Bin}(L, p)$, где L – размер окна. Как мы помним, при больших L и маленьких p биномиальное распределение можно заменить на более простое распределение Пуассона. Таким образом, мы моделируем покрытие каждого окна распределением $\mathrm{Pois}(\lambda)$, причём распределения различных окон независимы в совокупности. Последнее предположение, которого мы придерживаемся в большей части работы (но не в итоговой модели) – что множество возможных λ для данного эксперимента ограничено двумя элементами: меньший соответствует ненасыщенным регионам, а больший – насыщенным. Состояние насыщенности или ненасыщенности данного окна мы будем называть его *скрытым состоянием,* потому что мы не можем непосредственно проверить наличие интересующего нас белка на указанной позиции.

Корректность вышеприведённых рассуждений, разумеется, открыта для

критики. Однако многие исследования показывают, что агрегированное покрытие обычного секвенирования подчиняется распределению Пуассона (см., например, [7]).

3 Пуассоновская смесь

3.1 Мотивация

Для построения самой простой модели мы предполагаем, что априорное распределение скрытых состояний одинаково и независимо для всех окон. Такая постановка задачи естественным образом даёт в качестве модели смесь распределений Пуассона с двумя компонентами. В этом подразделе мы будем обозначать насыщенное состояние числом 1, а ненасыщенное – числом 0.

3.2 Определение

Пусть $\lambda_0 > 0, \lambda_1 > \lambda_0, \pi \in [0, 1], T \in \mathbb{N}_{>0}$ – параметры модели. Тогда состоянием модели является пара $((s_t)_{0 \leqslant t < T}, (d_t)_{0 \leqslant t < T}) \in \{0, 1\}^T \times \mathbb{N}^T$, а распределение вероятностей определяется как:

$$\mathbf{P}(\overrightarrow{d}, \overrightarrow{s} | \lambda_0, \lambda_1, \pi) = \prod_{0 \leqslant t < T} \mathrm{Pois}(d_t; \lambda_{s_t}) \pi^{s_t} (1 - \pi)^{1 - s_t}.$$

Здесь T – количество окон, $\overrightarrow{s}$ – последовательность скрытых состояний, $\overrightarrow{d}$ – последовательность наблюдений, π – априорная вероятность насыщенного состояния (т. е. $\mathbf{P}(s_t = 1)$ для любого t) и λ_0, λ_1 – меньшая и большая интенсивности распределений Пуассона для наблюдений, соответственно.

Полученное распределение можно трактовать следующим образом: для каждого окна t независимо выбирается скрытое состояние ($s_t = 0$ с вероятностью $1 - \pi$ или $s_t = 1$ с вероятностью π), после чего из соответствующего распределения $\mathrm{Pois}(\lambda_{s_t})$ выбирается наблюдение d_t.

3.3 ОМП-вывод

Задача о поиске оценки максимума правдоподобия для этой модели выглядит следующим образом:

$$\arg\max_{\lambda_0,\lambda_1,\pi} \mathbf{P}(\overrightarrow{d}\,|\lambda_0,\lambda_1,\pi) = ?$$

Эта задача, к сожалению, не имеет точного решения из-за нетривиальной формы распределения, как и большинство реальных задач ОМП. В самом деле, указанное в формуле правдоподобие представляет собой сумму по всем возможным последовательностям скрытых состояний, причём каждое слагаемое представляет собой произведение с большим количеством множителей. Для приближённого решения задачи мы будем использовать один из наиболее стандартных подходов, а именно алгоритм EM (*англ.* Expectation-Maximization algorithm, см., например, [3, стр. 48]). Это итеративный алгоритм, который постепенно улучшает целевую функцию (правдоподобие), то тех пор, пока параметры не сойдутся. Алгоритм EM применим в тех случаях, когда необходимо максимизировать правдоподобие, являющееся интегралом или суммой по каким-либо скрытым состояниям. Пусть θ^k – k-ое приближение набора параметров. Тогда $k+1$-ая итерация алгоритма выглядит следующим образом:

E Вычислить:

$$f^k(\theta) := \mathbb{E}_{\overrightarrow{s}|\overrightarrow{d},\theta^k} \log \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}|\theta)$$

M Найти следующее приближение:

$$\theta^{k+1} := \arg\max f^k(\theta)$$

Сходимость алгоритма EM хотя бы к локальному максимуму функции правдоподобия, вообще говоря, не гарантирована (хотя каждая итерация гарантированно увеличивает правдоподобие). При применении EM в этой работе производился мультистарт (перезапуск алгоритма несколько раз с разными начальными приближениями) и валидация полученных приближений из практических соображений (например, согласованность для разных размеров окна).

В нашем случае вычисление матожидания на шаге E не составляет труда:

$$\begin{aligned}
\mathbb{E}_{\overrightarrow{s}|\overrightarrow{d},\theta^k} \log \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}|\theta) = \mathbb{E}_{\overrightarrow{s}|\overrightarrow{d},\theta^k} \sum_{0\leqslant t<T} \Big(& s_t \log \pi + (1-s_t)\log(1-\pi) \\
& + d_t \log \lambda_{s_t} - \log(d_t!) - \lambda_{s_t} \Big) \\
= \log \pi \sum_{0\leqslant t<T} & \mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k) \\
+ \log(1-\pi) \sum_{0\leqslant t<T} & \mathbf{P}(s_t = 0|\overrightarrow{d}, \theta^k) \\
+ \log \lambda_1 \sum_{0\leqslant t<T} & d_t \mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k) \\
+ \log \lambda_0 \sum_{0\leqslant t<T} & d_t \mathbf{P}(s_t = 0|\overrightarrow{d}, \theta^k) \\
- \lambda_1 \sum_{0\leqslant t<T} & \mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k) \\
- \lambda_0 \sum_{0\leqslant t<T} & \mathbf{P}(s_t = 0|\overrightarrow{d}, \theta^k) \\
- \sum_{0\leqslant t<T} & \log(d_t!).
\end{aligned}$$

Остаётся только вычислить значения $\mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k)$ и $\mathbf{P}(s_t = 0|\overrightarrow{d}, \theta^k)$, что несложно:

$$\begin{aligned}
\mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k) &= \frac{\mathbf{P}(\overrightarrow{d}, s_t = 1|\theta^k)}{\mathbf{P}(\overrightarrow{d}|\theta^k)} \\
&= \frac{\mathbf{P}(d_t, s_t = 1|\theta^k)}{\mathbf{P}(d_t|\theta^k)} \\
&= \frac{\pi^k \mathrm{Pois}(d_t; \lambda_1^k)}{\pi^k \mathrm{Pois}(d_t; \lambda_1^k) + (1-\pi^k)\mathrm{Pois}(d_t; \lambda_0^k)},
\end{aligned}$$

и, очевидно,

$$\mathbf{P}(s_t = 0|\overrightarrow{d}, \theta^k) = 1 - \mathbf{P}(s_t = 1|\overrightarrow{d}, \theta^k).$$

На шаге M для нахождения точки максимума мы берём частные производные f^k по параметрам:

$$\frac{\partial}{\partial\pi}f^k = \frac{1}{\pi}\sum_{0\leqslant t<T}\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k) - \frac{1}{1-\pi}\sum_{0\leqslant t<T}\mathbf{P}(s_t=0|\overrightarrow{d},\theta^k),$$
$$\frac{\partial}{\partial\lambda_0}f^k = \frac{1}{\lambda_0}\sum_{0\leqslant t<T}d_t\mathbf{P}(s_t=0|\overrightarrow{d},\theta^k) - \sum_{0\leqslant t<T}\mathbf{P}(s_t=0|\overrightarrow{d},\theta^k),$$
$$\frac{\partial}{\partial\lambda_1}f^k = \frac{1}{\lambda_1}\sum_{0\leqslant t<T}d_t\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k) - \sum_{0\leqslant t<T}\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k).$$

Легко видеть, что каждая производная зависит только от одного параметра и имеет ровно одну точку смены знака, так что следующее приближение имеет вид:

$$\pi^{k+1} = \frac{1}{T}\sum_{0\leqslant t<T}\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k),$$
$$\lambda_0^{k+1} = \frac{\sum_{0\leqslant t<T}d_t\mathbf{P}(s_t=0|\overrightarrow{d},\theta^k)}{\sum_{0\leqslant t<T}\mathbf{P}(s_t=0|\overrightarrow{d},\theta^k)},$$
$$\lambda_1^{k+1} = \frac{\sum_{0\leqslant t<T}d_t\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k)}{\sum_{0\leqslant t<T}\mathbf{P}(s_t=1|\overrightarrow{d},\theta^k)}.$$

3.4 Байесовский вывод

Для байесовской постановки задачи необходимо выбрать априорное распределение параметров, $\mathcal{P}(\theta)$. Из соображений вычислимости мы выбираем следующие распределения:

$$\lambda_0 \sim \Gamma(\alpha_0,\beta_0)$$
$$\lambda_1 \sim \Gamma(\alpha_1,\beta_1)$$
$$\pi \sim \mathrm{B}(\mu_1,\mu_0)$$

для некоторых гиперпараметров $\alpha_1,\beta_1,\alpha_0,\beta_0,\mu_1,\mu_0$. Напомним, что гамма-распределение является сопряжённым априорным к распределению Пуассона, а бета-распределение – к категориальному распределению с двумя исходами (то есть $\mathcal{P}(s_t)$). Однако, апостериорное распределение параметров даже для такой несложной модели имеет слишком сложный вид. Для того,

чтобы получить о нём информацию, мы пользуемся вариационным приближением, то есть приближаем истинное апостериорное распределение $\mathcal{P}(\overrightarrow{s}, \theta)$ распределением из фиксированного семейства. Одним из других возможных подходов является семплинг, использованный, например, в работе [9].

Следуя указаниям из работы [3], мы устанавливаем в качестве метрики нашего приближения дивергенция Кульбака-Лейблера (KL), а в качестве семейства приближений – семейство распределений вида $q_{\overrightarrow{s}}(\overrightarrow{s})q_\theta(\theta)$, то есть, таких распределений, для которых случайные величины $\overrightarrow{s}$ и θ независимы. Для нахождения таких $q_{\overrightarrow{s}}$ и q_θ мы используем вариационный байесовский алгоритм EM.

3.4.1 VBEM

Вариационный байесовский EM (*англ.* variational Bayes expectation-maximization algorithm, VBEM), см. [3, стр. 54], как и обычный EM, является итеративно сходящимся алгоритмом. Мы начинаем с некоторого произвольного начального приближения q_θ^0. Каждая итерация состоит из двух шагов:

VBE Зафиксируем $q_{\overrightarrow{s}}^{k+1} \propto \exp(\mathbb{E}_{\theta \sim q_\theta^k} \log \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}, \theta))$.

VBM Зафиксируем $q_\theta^{k+1} \propto \exp(\mathbb{E}_{\overrightarrow{s} \sim q_{\overrightarrow{s}}^{k+1}} \log \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}, \theta))$.

Это фактически покомпонентная оптимизация целевой функции (т. е. дивергенции KL). Заметим, что q_s и q_θ, как распределения, нормализованы, поэтому пропорциональность определяет их полностью.

Можно убедиться в том, что приближения q_θ^k всегда (кроме, возможно, нулевого шага, на котором мы можем выбрать произвольное распределение) будут иметь вид:

$$q_\theta^k = \Gamma(\lambda_1; \alpha_1^k, \beta_1^k)\Gamma(\lambda_0; \alpha_0^k, \beta_0^k)\mathrm{B}(\pi; \mu_1^k, \mu_0^k),$$

где

$$\begin{aligned}
\alpha_1^{k+1} &= \alpha_1 + \sum_{0 \leqslant t < T} d_t \mathbb{E} s_t, \\
\alpha_0^{k+1} &= \alpha_0 + \sum_{0 \leqslant t < T} d_t \mathbb{E}(1 - s_t), \\
\beta_1^{k+1} &= \beta_1 + \sum_{0 \leqslant t < T} \mathbb{E} s_t, \\
\beta_0^{k+1} &= \beta_0 + \sum_{0 \leqslant t < T} \mathbb{E}(1 - s_t), \\
\mu_1^{k+1} &= \mu_1 + \sum_{0 \leqslant t < T} \mathbb{E} s_t, \\
\mu_0^{k+1} &= \mu_0 + \sum_{0 \leqslant t < T} \mathbb{E}(1 - s_t).
\end{aligned}$$

Кроме того, легко видеть, что распределение $q^k_{\overrightarrow{s}}$ раскладывается на множители соответственно отдельным окнам:

$$q^k_{\overrightarrow{s}}(\overrightarrow{s}) = \prod_{0 \leqslant t < T} q^k_{s_t}(s_t),$$

и что:

$$\begin{aligned}
q^k_{s_t}(1) &= \frac{1}{Z} \left(\mathbb{E}(\log \pi) + d_t \mathbb{E}(\log \lambda_1) - \mathbb{E}\lambda_1 \right) \\
&= \frac{1}{Z} \left(\gamma(\mu_1^k) - \gamma(\mu_1^k + \mu_0^k) + d_t (\gamma(\alpha_1^k) - \log \beta_1^k) - \frac{\alpha_1^k}{\beta_1^k} \right), \\
q^k_{s_t}(0) &= \frac{1}{Z} \left(\mathbb{E}(\log(1 - \pi)) + d_t \mathbb{E}(\log \lambda_0) - \mathbb{E}\lambda_0 \right) \\
&= \frac{1}{Z} \left(\gamma(\mu_0^k) - \gamma(\mu_1^k + \mu_0^k) + d_t (\gamma(\alpha_0^k) - \log \beta_0^k) - \frac{\alpha_0^k}{\beta_0^k} \right),
\end{aligned}$$

где Z – некоторая константа нормализации.

Предыдущие рассуждения дают нам непосредственные правила для построения приближений.

3.4.2 Оценка правдоподобия с помощью семплинга

К сожалению, вычисление маргинального правдоподобия данных

$$\mathbf{P}(\overrightarrow{d}) = \sum_{\overrightarrow{s}} \int_{\theta} \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}, \theta) d\theta,$$

не поддаётся упрощению: нетрудно взять интеграл или свернуть сумму, но невозможно выполнить и то, и другое. Проблема может быть частично решена с помощью численных методов:

$$\begin{aligned}\mathbf{P}(\overrightarrow{d}) &= \int_{\theta} \mathbf{P}(\overrightarrow{d}, \theta) d\theta \\ &= \int_{\theta} q_{\theta}(\theta) \frac{\mathbf{P}(\overrightarrow{d}, \theta)}{q_{\theta}(\theta)} d\theta \\ &= \mathbb{E}_{\theta \sim q_{\theta}} \left[\frac{\mathbf{P}(\overrightarrow{d}, \theta)}{q_{\theta}(\theta)} \right],\end{aligned}$$

где q_{θ} – произвольное распределение, при условии, что отношение $\dfrac{\mathbf{P}(\overrightarrow{d}, \theta)}{q_{\theta}(\theta)}$ корректно определено для всех допустимых θ, то есть

$$\forall \theta \;\; q_{\theta}(\theta) = 0 \Rightarrow \mathbf{P}(\overrightarrow{d}, \theta) = 0.$$

Но матожидание в последнем выражении можно вычислить стохастически с помощью семплинга $\theta \sim q_{\theta}$ и замены матожидания на выборочное среднее. Заметим, что величина

$$\begin{aligned}\mathbf{P}(\overrightarrow{d}, \theta) &= \sum_{\overrightarrow{s}} \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}, \theta) \\ &= \prod_{0 \leqslant t < T} (\mathbf{P}(d_t, s_t = 0, \theta) + \mathbf{P}(d_t, s_t = 1, \theta))\end{aligned}$$

поддаётся вычислению, т. к. априорные распределения состояний независимы. Из практических соображений в качестве q_{θ} мы, как правило, выбираем плотность вариационного приближения апостериорного распределения параметров, полученное в результате работы алгоритма VBEM: скорее всего, оно сосредоточено там же, где и истинное распределение $\mathcal{P}(\overrightarrow{d}, \theta)$, а семплинг из него легко реализуем. Следует, однако, отметить, что в случаях, когда

вариационное приближение распределения становится слишком узким, оказывается выгоднее брать более широкое распределение с таким же матожиданием, так как узкое распределение делает семплируемую величину слишком неравномерной для получения достоверной оценки из разумного количества семплов.

Кроме стохастического приближения, можно использовать нижнюю оценку правдоподобия:

$$\begin{aligned}
\log \mathbf{P}(\overrightarrow{d}) &= \int_\theta q_\theta(\theta) \log \mathbf{P}(\overrightarrow{d}) d\theta \\
&= \int_\theta q_\theta(\theta) \log \frac{\mathbf{P}(\overrightarrow{d}, \theta)}{\mathbf{P}(\theta | \overrightarrow{d})} d\theta \\
&= \int_\theta q_\theta(\theta) \log \mathbf{P}(\overrightarrow{d}, \theta) d\theta - \int_\theta q_\theta(\theta) \log q_\theta(\theta) d\theta \\
&\quad + \int_\theta q_\theta(\theta) \log q_\theta(\theta) d\theta - \int_\theta q_\theta(\theta) \log \mathbf{P}(\theta | \overrightarrow{d}) d\theta \\
&= \mathbb{E}_{\theta \sim q_\theta(\theta)} \Big[\log \mathbf{P}(\overrightarrow{d}, \theta) - \log q_\theta(\theta) \Big] \\
&\quad + KL(q_\theta \parallel \mathbf{P}(\theta | \overrightarrow{d}));
\end{aligned}$$

остаётся заметить, что дивергенция Кульбака-Лейблера в последней строчке неотрицательна, а предыдущее слагаемое легко вычислимо, так что оно и является упомянутой нижней оценкой.

3.5 Применение

3.5.1 ОМП-обучение

Мы обучаем модель PM на агрегированном покрытии, полученном в ходе одного эксперимента ChIP-seq. В качестве начального приближения пуассоновских интенсивностей λ_1^0, λ_0^0 мы используем удвоенное среднее покрытие и половину среднего покрытия, соответственно. В качестве начального приближения распределения вероятностей компонентов смеси мы используем равномерное распределение ($\pi^0 = 0.5$). Мы запускаем EM с этими начальными приближениями и итерируемся до тех пор, пока сумма модулей изменений всех параметров (то есть расстояние L_1) не станет меньше 0.001. В качестве валидации модели мы сравниваем её с обычным распределением Пуассона, используя критерий Акаике.

размер окна	$\ln \mathcal{L}_{PM}$	$\mathrm{AIC}_{PM} - \mathrm{AIC}_{P}$
50	-2128431	-1812942
100	-1666184	-1950481
500	-901460	-2040378
1000	-683143	-2025185
5000	-362976	-1822519
10000	-274753	-1610145
50000	-152864	-1013327
100000	-130562	-803113
500000	-95986	-490199
1000000	-86645	-360439
5000000	-44599	-183230

Таблица 1: Логарифм правдоподобия модели PM и её преимущество перед распределением Пуассона.

Таблица 1 показывает полное лог-правдоподобие (логарифм правдоподобия) моделей PM, разность лог-правдоподобий между моделью PM и распределением Пуассона и разность между критериями Акаике. Использован трек H3K4me3 первой хромосомы клеточной линии MB из серии экспериментов GSE25308 (см. 1.7).

Легко видеть, что разность между значениями AIC более чем достаточна, чтобы объявить модель PM существенно более правдоподобной.

Чтобы учесть, что EM не всегда сходится к глобальному максимуму, мы несколько раз перезапускали обучение с разными начальными оценками. В каждом случае EM сходился к одному и тому же результату (с точностью до погрешности), что позволяет предположить, что EM находит правильную ОМП.

3.5.2 Байесовское обучение

Мы обучаем байесовскую модель PM на тех же данных. В качестве гиперпараметров априорного распределения параметров мы выбираем следующие

размер окна	$\ln \mathcal{L}_{PM}$	$\ln \mathcal{L}_{BPM}$
50	-2128431	-2128460
100	-1666184	1666213
500	-901460	-901490
1000	-683143	-683173
5000	-362976	-363009
10000	-274753	-274788
50000	-152864	-152907
100000	-130562	-130607
500000	-95986	-96035
1000000	-86645	-86691
5000000	-44599	-44649

Таблица 2: Логарифм правдоподобия ОМП- и байесовского варианта модели PM.

величины:

$$\mu_1 = 1 \qquad \mu_0 = 2$$
$$\alpha_1 = 9 \qquad \alpha_0 = 1$$
$$\beta_1 = \frac{N}{12} \qquad \beta_0 = \frac{N}{4},$$

потому что в этом случае

$$\mathbb{E}(\lambda_1) = \frac{3N}{4}, \quad \sigma\lambda_1 = \frac{N}{4},$$

$$\mathbb{E}(\lambda_0) = \frac{N}{4}, \quad \sigma\lambda_0 = \frac{N}{4},$$

поскольку мы предполагаем, что λ_0 занимает первую половину отрезка $[0, N]$ ($\frac{N}{4} \pm \frac{N}{4}$), а λ_1 – вторую ($\frac{3N}{4} \pm \frac{N}{4}$), и π, вероятнее всего, мало ($\mu_1 < \mu_0$).

Мы используем VBEM для приближения апостериорного распределения параметров и семплинг для оценки правдоподобия. Мы эмпирически установили, что для получения устойчивой оценки хватает всего $100 - 200$ семплов; оценка при этом получается точной до четвёртого знака.

Из таблицы 2 можно видеть, что правдоподобие байесовской модели ниже,

чем у ОМП-варианта (что не только ожидаемо, но и неизбежно, так как:

$$\max_{\theta} \mathbf{P}(d|\theta) \geqslant \int_{\theta} \mathbf{P}(d|\theta)\mathbf{P}(\theta)d\theta = \mathbf{P}(d).$$

Однако различие между правдоподобиями несущественно, так что байесовскую модель можно также считать адекватной.

4 Скрытая марковская модель с пуассоновскими наблюдениями

4.1 Мотивация

Предположение о независимости распределения скрытых состояний соседних окон, сделанное при построении предыдущей модели, для реальных данных выглядит неправдоподобным. В самом деле, протяжённые участки одного и того же состояния встречаются чаще, чем это предсказывает независимая модель. Предположительно, близкие участки генома будут иметь одинаковое состояние с большей вероятностью. Для того, чтобы включить это наблюдение в нашу модель, при этом не сильно увеличив её сложность, мы будем использовать скрытую марковскую модель (первого порядка) с пуассоновскими наблюдениями (*англ.* Poisson hidden Markov model, PHMM).

4.2 Определение

Скрытые состояния в скрытой марковской модели моделируются марковской цепью: распределение каждого следующего состояния полностью определяется значением предыдущего состояния. Наблюдения в нашей модели подчиняются распределению Пуассона с интенсивностью, зависящей от соответствующего скрытого состояния. Более строго, для данного множества возможных скрытых состояний $\mathfrak{S}$ и параметров модели

$$\begin{aligned} \overrightarrow{\pi} &= (\pi_s)_{s\in\mathfrak{S}} \in [0,1]^{\mathfrak{S}} \text{ (начальные вероятности)}, \\ A &= (a_{rs})_{r,s\in\mathfrak{S}} \in [0,1]^{\mathfrak{S}\times\mathfrak{S}} \text{ (вероятности перехода)}, \\ \overrightarrow{\lambda} &= (\lambda_s)_{s\in\mathfrak{S}} \in \mathbb{R}^{\mathfrak{S}}_{\geqslant 0} \text{ (интенсивности наблюдений)}, \end{aligned}$$

таких, что

$$\sum_{s\in\mathfrak{S}} \pi_s = 1,$$
$$\forall r \in \mathfrak{S} \quad \sum_{s\in\mathfrak{S}} a_{rs} = 1,$$

и при данном количестве наблюдений $T \in \mathbb{N}_{>0}$, векторе наблюдений $\overrightarrow{d} \in \mathbb{N}^T$ и векторе скрытых состояний $\overrightarrow{s} \in \mathfrak{S}^T$, правдоподобие определяется следующим образом:

$$\mathbf{P}(\overrightarrow{d}, \overrightarrow{s} | \overrightarrow{\pi}, A, \overrightarrow{\lambda}) = \pi_{s_0} \prod_{0<t<T} a_{s_{t-1}s_t} \prod_{0\leqslant t<T} \mathrm{Pois}(d_t; \lambda_{s_t}).$$

4.3 ОМП-вывод

4.3.1 Алгоритмы прямого-обратного хода и Витерби

Как и раньше, нас интересует маргинальное распределение скрытых состояний модели $\mathcal{P}(s_t|\overrightarrow{d}, \theta)$ и оценка максимального правдоподобия на скрытые состояния $\arg\max_{\overrightarrow{s}} \mathbf{P}(\overrightarrow{s}|\overrightarrow{d}, \theta)$. В данном случае, эти задачи можно решить при помощи схожих методов: алгоритма прямого-обратного хода (*англ.* forward-backward algorithm) и алгоритма Витерби (*англ.* Viterbi algorithm), соответственно.

Алгоритм прямого-обратного хода динамически вычисляет следующие величины:

$$\begin{aligned}
\alpha_t(s) &:= \mathbf{P}(s_t = s, d_0, \ldots, d_t|\theta) \\
&= \mathbf{P}(d_t|s_t = s, \theta) \cdot \sum_{r\in\mathfrak{S}} \mathbf{P}(s_t = s|s_{t-1} = r, \theta)\alpha_{t-1}(r), \\
\alpha_0(s) &= \mathbf{P}(d_0|s_0 = s, \theta)\mathbf{P}(s_0 = s|\theta), \\
\beta_t(s) &:= \mathbf{P}(d_{t+1}, \ldots, d_{T-1}|s_t = s, \theta) \\
&= \sum_{r\in\mathfrak{S}} \beta_{t+1}(r)\mathbf{P}(d_{t+1}|s_{t+1} = r, \theta)\mathbf{P}(s_{t+1} = r|s_t = s, \theta), \\
\beta_{T-1}(s) &= 1.
\end{aligned}$$

После их вычисления, маргинальные распределения вероятностей состояний

и переходов легко находятся:

$$\mathbf{P}(s_t = s|\overrightarrow{d}, \theta) = \frac{\alpha_t(s)\beta_t(s)}{\sum_{r\in\mathfrak{S}} \alpha_t(r)\beta_t(r)},$$
$$\mathbf{P}(s_{t-1} = r, s_t = s|\overrightarrow{d}, \theta) = \frac{\alpha_{t-1}(r)a_{rs}\mathbf{P}(d_t|s_t = s, \theta)\beta_t(s)}{\sum_{r',s'\in\mathfrak{S}} \alpha_{t-1}(r')a_{r's'}\mathbf{P}(d_t|s_t = s', \theta)\beta_t(s')}.$$

Алгоритм Витерби предназначен для нахождения оценки максимального правдоподобия для совместного распределения состояний. Он динамически вычисляет следующие величины:

$$\varepsilon_t(s) := \max_{s_0,\ldots,s_{t-1}} \mathbf{P}(s_t = s, s_0, \ldots, s_{t-1}|\overrightarrow{d})$$
$$= \mathbf{P}(d_t|s_t = s, \theta) \max_{s_{t-1}} \varepsilon_{t-1}(s_{t-1})a_{s_{t-1}s}.$$

Наибольшая из величин ε_{T-1} даёт максимальное правдоподобие; пройдя по цепочке вычислений обратно, мы можем узнать саму последовательность состояний, которая привела к этому результату.

Заметим наконец, что, с точки зрения конечного результата, обоим алгоритмам достаточно знать плотность распрделения вероятностей с точностью до мультипликативной константы. Это связано с тем, что алгоритм прямого-обратного хода на последнем этапе производит нормализацию (в ходе которой мультипликативная константа сокращается), а алгоритм Витерби оперирует только сравнениями, результат которых также не зависит от мультипликативной константы.

4.3.2 Алгоритм Баума-Велша

Как уже упоминалось ранее, задача нахождения оценки максимального правдоподобия для параметров модели, как правило, не имеет точного решения; вместо этого обычно используются итеративно сходящиеся методы. Здесь мы, как и прежде, будем применять метод EM (см. 3.3). Специализация метода EM на скрытые марковские модели традиционно называется алгоритмом Баума-Велша (*англ.* Baum-Welch algorithm).

Очередная оценка параметров модели вычисляется следующим образом:

$$\pi_s^{k+1} = \mathbf{P}(s_0 = s|\,\overrightarrow{d}\,, \theta^k),$$
$$a_{rs}^{k+1} = \frac{\sum_{0<t<T} \mathbf{P}(s_t = s, s_{t-1} = r|\,\overrightarrow{d}\,, \theta^k)}{\sum_{0<t<T} \mathbf{P}(s_{t-1} = r|\,\overrightarrow{d}\,, \theta^k)},$$
$$\lambda_s^{k+1} = \frac{\sum_{0\leqslant t<T} d_t \mathbf{P}(s_t = s|\,\overrightarrow{d}\,, \theta^k)}{\sum_{0\leqslant t<T} \mathbf{P}(s_t = s|\,\overrightarrow{d}\,, \theta^k)}.$$

Оставшиеся вероятности, как мы знаем, можно найти с помощью алгоритма прямого-обратного хода.

4.4 Байесовский вывод

Для байесовского вывода мы предполагаем следующие априорные распределения на параметрах:

$$\begin{aligned} & \overrightarrow{\pi} \sim \mathrm{Dir}\,(\overrightarrow{\mu})\,, \\ \forall r \in \mathfrak{S} \quad & \overrightarrow{a}_r = (a_{rs})_{s\in\mathfrak{S}} \sim \mathrm{Dir}\,(\overrightarrow{\nu}_r)\,, \\ \forall r \in \mathfrak{S} \quad & \lambda_r \sim \Gamma(\alpha_r, \beta_r) \end{aligned}$$

для некоторых гиперпараметров $\overrightarrow{\mu}, N = (\overrightarrow{\nu_r})_r, \overrightarrow{\alpha} = (\alpha_r)_r, \overrightarrow{\beta} = (\beta_r)_r$. Выбор распределений, как обычно, мотивирован соображениями вычислимости: распределения параметров принадлежат сопряжённым апиорным семействам к соответствующим распределениям наблюдений (см. 1.5).

4.4.1 VBEM

Обучение этой модели мы тоже производим с помощью алгоритма VBEM (см. 3.4.1). Точно так же, как и в модели PM, оценки на распределение параметров будут иметь ту же форму, что и априорное распределение; таким образом, для описания оценки распределения достаточно вычислить гиперпараметры. Выписывая соответствующие матожидания, мы получаем правила

вычисления очередных гиперпараметров оценок:

$$\begin{aligned}
\mu_s^{k+1} &= \mu_s + \mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_0 = s],\\
\nu_{rs}^{k+1} &= \nu_{rs} + \sum_{0<t<T} \mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_{t-1} = r, s_t = s],\\
\alpha_s^{k+1} &= \alpha_s + \sum_{0\leqslant t<T} d_t \mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_t = s],\\
\beta_s^{k+1} &= \beta_s + \sum_{0\leqslant t<T} \mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_t = s],
\end{aligned}$$

где [.] – характеристическая функция логического выражения. Все оставшиеся матожидания могут быть вычислены с помощью модифицированного алгоритма прямого-обратного хода.

4.4.2 Алгоритм прямого-обратного хода для VBEM

Для вычисления оценок в алгоритме VBEM нам необходимо уметь вычислять следующие матожидания: $\mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_{t-1} = r, s_t = s]$, $\mathbb{E}_{\overrightarrow{s}\sim q_{\overrightarrow{s}}^{k+1}}[s_t = s]$. Как легко заметить, такие матожидания численно равны вероятностям соответствующих событий в соответствующем распределении. Из описания VBEM имеем:

$$\begin{aligned}
q_{\overrightarrow{s}}^{k+1} &\propto \exp \mathbb{E}_{\theta\sim q_\theta^k} \ln \mathbf{P}(\overrightarrow{d}, \overrightarrow{s}, \theta)\\
&= \exp\Big(\mathbb{E}\ln\pi_{s_0} + \sum_{0<t<T} \mathbb{E}\ln a_{s_{t-1}s_t} + \sum_{0\leqslant t<T} (d_t \mathbb{E}\ln\lambda_{s_t} - \mathbb{E}\lambda_{s_t}) + C\Big),
\end{aligned}$$

где C – величина, не зависящая от $\overrightarrow{s}$ (для краткости во второй строчке опущено распределение q_θ^k, по которому берутся матожидания). Таким образом, полученное распределение очень похоже на обычное распределение для ОМП-скрытой марковской модели, если в качестве значения параметров использовать экспоненту матожидания логарифма этих параметров. Заметим, что непосредственная подстановка таких значений приводит к субнормализации: например, в общем случае $\sum_{s\in\mathfrak{S}} \exp\mathbb{E}\ln\pi_s < 1$. Соответственно, постоянная C в этом распределении оказывается больше, чем аналогичная постоянная в ОМП-модели; кроме того, значение C не имеет замкнутой формы для вычисления. Однако, как было подчёркнуто в 4.3.1, наличие неизвестной мультипликативной константы не делает алгоритм прямого-обратного хода

некорректным. Таким образом, применяя алгоритм прямого-обратного хода для описанных выше параметров, мы получаем искомые матожидания. Осталось заметить, что, поскольку распределение q_θ^k имеет известную форму, мы можем воспользоваться готовыми результатами для вычисления матожиданий:

$$\mathbb{E}_{\theta\sim q_\theta^k} \ln \pi_s = \psi(\mu_s^k) - \psi(\sum_{r\in\mathfrak{S}} \mu_r^k),$$
$$\mathbb{E}_{\theta\sim q_\theta^k} \ln a_{rs} = \psi(\nu_{rs}^k) - \psi(\sum_{p\in\mathfrak{S}} \nu_{rp}^k),$$
$$\mathbb{E}_{\theta\sim q_\theta^k} \ln \lambda_s = \psi(\alpha_s) - \ln(\beta_s),$$
$$\mathbb{E}_{\theta\sim q_\theta^k} \lambda_s = \frac{\alpha_s}{\beta_s},$$

где $\psi(.)$ – дигамма-функция.

4.5 Применение

4.5.1 ОМП-обучение

Мы обучаем модель PHMM на агрегированном покрытии, полученном в ходе одного эксперимента ChIP-seq. Мы предполагаем, что модель имеет два состояния (насыщенное и ненасыщенное окно). Мы используем те же начальные приближения для интенсивностей, что и для модели PM. Для начальных вероятностей и вероятностей перехода в качестве начального приближения мы используем равномерные распределения. Так же, как и в случае модели PM, мы производим итерации алгоритма Баума-Велша до тех пор, пока расстояние L_1 между предыдущей и следующей оценкой параметров не станет меньше 0.001. Для валидации модели мы используем критерий Акаике; в качестве базовой модели используется модель PM.

Из таблицы 3 можно заметить, что с увеличением размера окна разница в значениях AIC существенно падает. Этот результат согласуется с ожиданиями: чем крупнее геномный регион, тем меньше его зависимость от соседей. Так, на приведённом примере для окна в 5 миллионов нуклеотидов критерий Акаике отдаёт предпочтение более простой модели PM. Однако, в остальных случаях, особенно при небольших размерах окна, более сложная модель PHMM существенно лучше, чем PM.

размер окна	$\ln \mathcal{L}_{PHMM}$	$\mathrm{AIC}_{PHMM} - \mathrm{AIC}_{PM}$
50	-1967016	-322826
100	-1576575	-179213
500	-882018	-38880
1000	-674478	-17325
5000	-362284	-1380
10000	-274503	-495.0
50000	-152764	-196.9
100000	-130447	-224.1
500000	-95952	-64.88
1000000	-86632	-21.49
5000000	-44599	2.878

Таблица 3: Логарифм правдоподобия модели PHMM и её преимущество перед моделью PM.

4.5.2 Байесовское обучение

Мы обучаем байесовский вариант модели на тех же данных. Мы выбираем такие же гиперпараметры априорного распределения интенсивностей наблюдений λ_s, что и в модели PM. Мы выбираем следующие гиперпараметры априорного распределения: $\mu_1 = 1$, $\mu_0 = 2$, $\nu_{01} = \nu_{10} = 1$, $\nu_{11} = \nu_{00} = 2$. Это отражает наши предположения о том, что ненасыщенных окон больше, чем насыщенных, и что состояние окна имеет тенденцию скорее сохраняться, чем изменяться.

Для начальной оценки апостериорного распределения интенсивностей наблюдений мы принимаем те же значения, что и в модели PM. Гиперпараметры для начальных вероятностей и вероятностей перехода мы принимаем такими же, как гиперпараметры для априорного распределения (из тех же соображений).

Мы валидируем нашу модель, сравнивая её маргинальное правдоподобие данных с таким же правдоподобием для модели PM. Правдоподобие оценивается с помощью семплинга по важности, аналогично 3.4.2.

Из таблицы 4 видно, что байесовский вариант PHMM также существенно лучше, чем более простая модель PM.

размер окна	$\ln \mathcal{L}_{BPHMM}$	$\ln \mathcal{L}_{BPM}$
50	-1972807	-2128460
100	-1578479	1666213
500	-882208	-901490
1000	-674568	-683173
5000	-362340	-363009
10000	-274594	-274788
50000	-152863	-152907
100000	-130572	-130607
500000	-96023	-96035
1000000	-85661	-86691
5000000	-42958	-44649

Таблица 4: Логарифм правдоподобия байесовских вариантов моделей PHMM и PM.

5 Скрытая марковская модель с многомерными пуассоновскими наблюдениями

Мы выдвигаем гипотезу, что связывание белков на треках двух экспериментов ChIP-seq не является независимым. В частности, для важных эпигенетических маркеров следует ожидать большей консервативности паттернов, чем получилось бы при независимом распределении треков. Для включения этого предположения в модель мы переходим к рассмотрению скрытой марковской модели с многомерными пуассоновскими наблюдениями (*англ.* multi-Poisson hidden Markov model, MPHMM).

5.1 Определение

Отличие модели MPHMM от PHMM состоит в природе и распределении наблюдений. Зафиксируем размерность наблюдений $N \in \mathbb{N}_{>0}$. Тогда, для данного множества возможных скрытых состояний $\mathfrak{S}$ и параметров модели

$$\overrightarrow{\pi} = (\pi_s)_{s\in\mathfrak{S}} \in [0,1]^{\mathfrak{S}} \text{ (начальные вероятности)},$$
$$A = (a_{rs})_{r,s\in\mathfrak{S}} \in [0,1]^{\mathfrak{S}\times\mathfrak{S}} \text{ (вероятности перехода)},$$
$$\Lambda = (\lambda_{sn})_{s\in\mathfrak{S},1\leqslant n\leqslant N} \in \mathbb{R}_{\geqslant 0}^{\mathfrak{S}\times[N]} \text{ (интенсивности наблюдений)}$$

(где $[N] = \{1, 2, \ldots, N\}$), таких, что

$$\sum_{s \in \mathfrak{S}} \pi_s = 1,$$
$$\forall r \in \mathfrak{S} \sum_{s \in \mathfrak{S}} a_{rs} = 1,$$

и при данном количестве наблюдений $T \in \mathbb{N}_{>0}$, матрице наблюдений $D = (d_{tn})_{0 \leqslant t < T, 1 \leqslant n \leqslant N} \in M_{\mathbb{N}}(T, N)$ и векторе скрытых состояний $\overrightarrow{s} \in \mathfrak{S}^T$, правдоподобие модели описывается формулой

$$\mathbf{P}(D, \overrightarrow{s} | \overrightarrow{\pi}, A, \Lambda) = \pi_{s_0} \prod_{0 < t < T} a_{s_{t-1} s_t} \prod_{\substack{0 \leqslant t < T, \\ 1 \leqslant n \leqslant N}} \mathrm{Pois}(d_{tn}; \lambda_{s_t n}).$$

5.2 ОМП-вывод

5.2.1 Алгоритмы прямого-обратного хода и Витерби

Алгоритмы прямого-обратного хода и Витерби не зависят от конкретной формы распределения наблюдений и, таким образом, остаются такими же, как в 4.3.1.

5.2.2 Алгоритм Баума-Велша

Алгоритм Баума-Велша должен включать в себя обновление оценок интенсивностей наблюдений. Легко видеть, что:

$$\lambda_{sn}^{k+1} = \frac{\sum_{0 \leqslant t < T} d_{tn} \mathbf{P}(s_t = s | D, \theta^k)}{\sum_{0 \leqslant t < T} \mathbf{P}(s_t = s | D, \theta^k)}.$$

5.3 Байесовский вывод

Форма априорных распределений параметров остаётся такой же, как в модели PHMM, за исключением того, что для каждого состояния мы предполагаем N независимых гамма-распределений для параметров интенсивности распределений с гиперпараметрами α_{sn}, β_{sn} вместо одного.

5.3.1 VBEM

Алгоритм VBEM также практически не изменяется, кроме оценок гиперпараметров распределения интенсивностей наблюдений:

$$\alpha_{sn}^{k+1} = \alpha_{sn} + \sum_{0 \leqslant t < T} d_{tn} \mathbb{E}_{\overrightarrow{s} \sim q_{\overrightarrow{s}}^{k+1}} [s_t = s],$$
$$\beta_{sn}^{k+1} = \beta_{sn} + \sum_{0 \leqslant t < T} \mathbb{E}_{\overrightarrow{s} \sim q_{\overrightarrow{s}}^{k+1}} [s_t = s].$$

В частности, гиперпараметры β_{sn}^k, относящиеся к одному состоянию s, равны с точностью до разности между соответствующими априорными гиперпараметрами β_{sn} (так как второе слагаемое не зависит от n).

Модифицированный алгоритм прямого-обратного хода, как и обычный, не изменяется.

5.4 Применение

5.4.1 ОМП-обучение

Мы обучаем модель MPHMM на агрегированном покрытии, полученном в ходе двух экспериментов ChIP-seq на клеточных линиях одного организма. Мы предполагаем, что множество возможных скрытых состояний содержит четыре элемента: LOW (на обоих треках ненасыщенные регионы), HIGH (на обоих треках насыщенные регионы), INCREASED (на первом треке ненасыщенный регион, на втором – насыщенный) и DECREASED (наоборот). Начальные приближения параметров выбираются следующим образом:

$$\forall s \quad \mu_s = 0.25,$$
$$\forall r, s \quad \nu_{rs} = 0.25,$$
$$\lambda_{LOW,1} = \lambda_{INCREASED,1} = \frac{\overline{d_1}}{2},$$
$$\lambda_{LOW,2} = \lambda_{DECREASED,2} = \frac{\overline{d_2}}{2},$$
$$\lambda_{HIGH,1} = \lambda_{DECREASED,1} = 2\overline{d_1},$$
$$\lambda_{HIGH,2} = \lambda_{INCREASED,2} = 2\overline{d_2},$$

размер окна	$\ln \mathcal{L}_{MPHMM}$	$\mathrm{AIC}_{MPHMM} - \mathrm{AIC}_{2\times PHMM}$
50	-3947291	-360789
100	-3108959	-352293
500	-1638569	-334567
1000	-1216376	-315621
5000	-632005	-216708
10000	-476905	-173704
50000	-268840	-99563
100000	-213481	-111314
500000	-145816	-110634
1000000	-118090	-125417
5000000	-60401	-63533

Таблица 5: Логарифм правдоподобия модели MPHMM и её преимущество перед независимыми моделями PHMM.

где $\overline{d_1}, \overline{d_2}$ – средние покрытия первого и второго трека, соответственно. Мы производим валидацию модели с помощью критерия Акаике; в качестве базовой модели мы выбираем две независимые модели PHMM для каждого трека.

Из таблицы 5 легко видеть, что модель с многомерными измерениями гораздо лучше подходит под данные, чем независимые модели для каждого трека, причём для всех рассмотренных размеров окон.

5.4.2 Байесовское обучение

Мы обучаем байесовский вариант модели на тех же данных. Однако, к сожалению, получающаяся оценка апостериорного распределения имеет неподходящую (для наших целей) форму: гиперпараметры интенсивностей для скрытых состояний не соответствуют смыслу, который мы вкладываем в эти состояния. В частности, отношение ожидаемых интенсивностей на двух треках практически одинаково для всех состояний, в то время как отношение должно быть меньшим для DECREASED и большим для INCREASED. Возможно, такая оценка действительно хорошо приближает апостериорное распределение параметров, но для целей работы (обнаружение дифференциально насыщенных регионов) она не подходит. Эта проблема устраняется в следующей модели.

6 Ограниченная скрытая марковская модель с многомерными пуассоновскими наблюдениями

6.1 Мотивация

При обучении модели MPHMM, описанной в предыдущей главе, мы обнаружили, что

$$\lambda_{\text{INCREASED},1} > \lambda_{\text{LOW},1},$$
$$\lambda_{\text{INCREASED},2} < \lambda_{\text{HIGH},2}$$

и т. д. для метода ОМП. Это, возможно, понижает способность модели обнаруживать дифференциально насыщенные регионы. Кроме того, байесовское обучение модели MPHMM вообще не привело к удовлетворительным результатам. Чтобы исправить эти недостатки, мы рассматриваем ограниченную скрытую марковскую модель с многомерными пуассоновскими наблюдениями (*англ.* constrained multi-Poisson hidden Markov model, CMPHMM).

6.2 Определение

Фактически, семейство моделей CMPHMM является подмножеством семейства MPHMM. Зафиксируем отношение эквивалентности (которое мы будем называть ограничением) $\sim$ на множестве $\mathfrak{S} \times [N]$. Тогда модель CMPHMM – это модель MPHMM, интенсивности наблюдений которой удовлетворяют следующим условиям:

$$\forall r, s \in \mathfrak{S} \quad \forall m, n \in [N] \quad (r, m) \sim (s, n) \Rightarrow \lambda_{rm} = \lambda_{sn}.$$

Правдоподобие модели не изменяется.

6.3 ОМП-вывод

Алгоритмы прямого-обратного хода и Витерби остаются неизменными.

6.3.1 Алгоритм Баума-Велша

Алгоритм Баума-Велша при построении очередных оценок должен учитывать ограничение. К счастью, соответствующий условный максимум найти несложно:

$$\lambda_{sn}^{k+1} = \frac{\sum\limits_{0 \leqslant t < T} \sum\limits_{(r,m) \sim (s,n)} d_{tm} \mathbf{P}(s_t = r | D, \theta^k)}{\sum\limits_{(r,m) \sim (s,n)} \mathbf{P}(s_t = r | D, \theta^k)}.$$

Легко убедиться, что новая оценка удовлетворяет ограничению.

6.4 Байесовский вывод

Для байесовской постановки задачи мы предполагаем такие же априорные распределения, как и для модели MPHMM, за исключением того, что в случае CMPHMM каждому классу эквивалентности соответствует единственное априорное распределение (так как все соответствующие интенсивности идентичны). Таким образом, гиперпараметры распределения интенсивностей – это $(\alpha_c)_{c \in \mathfrak{S} \times [N]/\sim}, (\beta_c)_{c \in \mathfrak{S} \times [N]/\sim}$.

6.4.1 VBEM

Как и в предыдущих случаях, оценки апостериорного распределения параметров имеют ту же форму. Гиперпараметры вероятностей обновляются так же, как в модели MPHMM. Очередное значение гиперпараметров интенсивностей находится следующим образом:

$$\alpha_c^{k+1} = \alpha_c + \sum_{0 \leqslant t < T} \sum_{(s,n) \in c} d_{tn} \mathbb{E}_{\vec{s} \sim q_{\vec{s}}^{k+1}} [s_t = s],$$

$$\beta_c^{k+1} = \beta_c + \sum_{0 \leqslant t < T} \sum_{(s,n) \in c} \mathbb{E}_{\vec{s} \sim q_{\vec{s}}^{k+1}} [s_t = s].$$

Модифицированный алгоритм прямого-обратного хода не претерпевает изменений.

размер окна	$\ln \mathcal{L}_{CMPHMM}$	AIC_{CMPHMM}	AIC_{MPHMM}	$\mathrm{AIC}_{2\times PHMM}$
50	-4088175	8176387	7894629	8255418
100	-3262673	6525385	6217964	6570256
500	-1800375	3600788	3277183	3611750
1000	-1370185	2740408	2432799	2748420
5000	-737605	1475248	1264056	1480764
10000	-561536	1123110	953857	1127560
50000	-317510	635057	537725	637288
100000	-268490	537018	427008	538322
500000	-200971	401980	291679	402313
1000000	-180741	361520	236226	361643
5000000	-92274	184586	120847	184380

Таблица 6: Логарифм правдоподобия и AIC модели CMPHMM и её преимущество перед MPHMM и независимыми моделями PHMM.

6.5 Применение

6.5.1 ОМП-обучение

Мы обучаем модель CMPHMM на тех же данных, что и модель MPHMM. Мы предполагаем, что модель имеет те же четыре состояния. В качестве ограничения мы устанавливаем отношение эквивалентности, порождённое следующими парами:

$$(\mathrm{LOW}, 1) \sim (\mathrm{INCREASED}, 1),$$
$$(\mathrm{HIGH}, 1) \sim (\mathrm{DECREASED}, 1),$$
$$(\mathrm{LOW}, 2) \sim (\mathrm{DECREASED}, 2),$$
$$(\mathrm{HIGH}, 2) \sim (\mathrm{INCREASED}, 2).$$

Таким образом, в данной модели $\lambda_{\mathrm{LOW},1} = \lambda_{\mathrm{INCREASED},1}$ и т. п., что логично с точки зрения предметной области – для первого трека оба состояния соответствуют одному и тому же ненасыщенному окну. Мы используем те же начальные приближения для алгоритма Баума-Велша (легко убедиться, что они удовлетворяют ограничению). Мы валидируем модель с помощью критерия Акаике, сравнивая её с двумя независимыми моделями PHMM для каждого трека, а также с моделью MPHMM.

Из таблицы 6 можно видеть, что модель CMPHMM проигрывает в сравнении с моделью MPHMM (что ожидаемо, так как оптимизация в случае

размер окна	$\log \mathcal{L}_{BCMPHMM}$	$\log \mathcal{L}_{2\times BPHMM}$	$\log \mathcal{L}_{CMPHMM}$
50	-3874820	-5012256	-4088175
100	-3121768	-3822349	-3262673
500	-1762844	-1881438	-1800375
1000	-1352232	-1413289	-1370185
5000	-726938	-750993	-737605
10000	-550647	-571878	-561536
50000	-307852	-322999	-317510
100000	-264422	-272540	-268490
500000	-195169	-202204	-200971
1000000	-172444	-180451	-180741
5000000	-86401	-91239	-92274

Таблица 7: Логарифм правдоподобия байесовского и ОМП-вариантов модели CMPHMM, а также независимых байесовских моделей PHMM.

CMPHMM идёт по строгому подмножеству моделей MPHMM), но выигрывает в сравнении с независимыми моделями. Однако, модель CMPHMM выглядит более адекватной с точки зрения биологической интерпретируемости: как уже упоминалось, в этой модели состояния LOW и INCREASED неразличимы с точки зрения первого трека, что логично, так как мы ищем именно области, где низкое покрытие меняется на высокое; определять низкое покрытие по-разному в зависимости от того, произошло ли в итоге изменение, кажется неоправданным.

6.5.2 Байесовское обучение

Мы обучаем байесовский вариант модели на тех же данных, с теми же скрытыми состояниями и тем же ограничением, что и ОМП-вариант. Мы используем те же начальные приближения для VBEM, что и в случае MPHMM (они удовлетворяют ограничению). Мы сравниваем нашу модель с двумя независимыми байесовскими моделями PHMM для каждого трека и с ОМП-вариантом.

Любопытно, что байесовский подход показывает, что EM не сходится к глобальному максимуму при поиске ОМП (причиной этого, скорее всего, является сложная форма функции правдоподобия). Это следует из того, что (см. таблицу 7) правдоподобие байесовской модели оказывается выше, чем

правдоподобие ОМП-модели, обученной с помощью EM, а, очевидно,

$$\max_{\theta} \mathbf{P}(d|\theta) \geqslant \int_{\theta} \mathbf{P}(d|\theta)\mathbf{P}(\theta)d\theta = \mathbf{P}(d).$$

Из этого можно сделать вывод, что байесовский подход более пригоден для анализа модели, тем более что матожидание полученного приближения апостериорного распределения весьма близко к оценкам EM.

То, что CMPHMM и в байесовском случае оказывается лучше, чем независимые PHMM для каждого трека, валидирует сделанные нами предположения о зависимости треков.

7 BCZIPHMM

7.1 Мотивация

Из всех изученных нами моделей для сравнения треков наиболее адекватной оказалась байесовская CMPHMM (MPHMM обладает недостаточной интерпретируемостью, а ОМП-вариант CMPHMM имеет меньшее правдоподобие). Многие исследования [4, 8, 9] указывают, что распределение Пуассона имеет слишком маленькую дисперсию для корректного моделирования результатов ChIP-seq. Для борьбы с этой проблемой используются разные приёмы. Один из них – подкачка нуля (*англ.* zero inflation), то есть, замена обычного распределения на смесь этого распределения с постоянной (нулём). Так, распределение Пуассона с подкачкой нуля (*англ.* zero-inflated Poisson distribution, ZIP) определяется следующим образом:

$$\mathbf{P}(X = k) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & k = 0; \\ (1-\pi)\dfrac{\lambda^k}{k!}e^{-\lambda}, & k > 0. \end{cases}$$

Здесь π – параметр смеси, λ – интенсивность распределения Пуассона. Распределение ZIP, тем самым, придаёт больший вес нулю; если интенсивность существенно отлична от нуля, то такое распределение будет бимодальным. Такой эффект является привлекательным с практической точки зрения: в ходе эксперимента ChIP-seq нередко получаются регионы, имеющие нулевое покрытие по объективным причинам. Такими причинами могут являться:

отсутствие референсной последовательности (как следствие, отсутствие выравниваний на этот регион), большое количество геномных повторов (как следствие, отсутствие уникальных выравниваний на этот регион) и другие.

Мы модифицируем нашу имплементацию модели CMPHMM, добавляя к ней состояние NULL, в котором оба наблюдения гарантированно нулевые. Это играет ту же роль, что и замена распределения Пуассона на ZIP в приведённых выше примерах. Будем называть эту имплементацию ограниченной байесовской скрытой марковской моделью с пуассоновскими наблюдениями и подкачкой нуля (*англ.* Bayesian constrained zero-inflated Poisson hidden Markov model, BCZIPHMM).

7.2 Определение

Определение BCZIPHMM легко восстановить из предыдущих глав, однако для удобства мы полностью воспроизводим его здесь.

BCZIPHMM является скрытой марковской моделью. Наблюдениями является последовательность пар неотрицательных целых чисел, соответствующих агрегированному покрытию какого-либо геномного региона (как правило, хромосомы), полученному в результате двух экспериментов ChIP-seq. Скрытое состояние для каждого окна может принимать одно из пяти значений: LOW, HIGH, INCREASED, DECREASED, NULL. Компоненты наблюдений независимы и подчиняются распределению Пуассона с интенсивностью, зависящей от скрытого состояния. Интенсивности наблюдений, начальные вероятности и вероятности перехода – случайные величины, подчиняющиеся соответствующим априорным распределениям (гамма-распределениям и распределениям Дирихле). Правдоподобие модели описывается следующей формулой:

$$\begin{aligned}\mathbf{P}(D, \overrightarrow{s}, \overrightarrow{\pi}, A, \Lambda) = \pi_{s_0} \prod_{0<t<T} a_{s_{t-1}s_t} \prod_{\substack{0 \leqslant t < T, \\ 1 \leqslant n \leqslant 2}} \mathrm{Pois}(d_{tn}; \lambda_{s_t,n}) \\ \times \mathrm{Dir}(\overrightarrow{\pi}; \overrightarrow{\mu}) \prod_{r \in \mathfrak{S}} \mathrm{Dir}(\overrightarrow{a_{r*}}; \overrightarrow{\nu_r}) \\ \times \prod_{\substack{\mathrm{sign} \in \{+,-\}, \\ 1 \leqslant n \leqslant 2}} \Gamma(\lambda_{\mathrm{sign},n}; \alpha_{\mathrm{sign},n}, \beta_{\mathrm{sign},n}),\end{aligned}$$

где

$$
\begin{aligned}
T \in \mathbb{N}_{>0} &\text{ – количество наблюдений,}\\
D = (d_{tn})_{0\leqslant t<T, 1\leqslant n\leqslant 2} \in \mathbb{N}^{T\times\{1,2\}} &\text{ – матрица наблюдений,}\\
\mathfrak{S} := \left\{\begin{matrix}\text{NULL, LOW, HIGH,}\\ \text{INCREASED,}\\ \text{DECREASED}\end{matrix}\right\} &\text{ – множество возможных состояний,}\\
\overrightarrow{s} = (s_t)_{0\leqslant t<T} \in \mathfrak{S}^T &\text{ – последовательность состояний,}\\
\Lambda = (\lambda_{-,1}, \lambda_{+,1}, \lambda_{-,2}, \lambda_{+,2}) \in \mathbb{R}_{>0}^{\{+,-\}\times\{1,2\}} &\text{ – интенсивности наблюдений,}\\
\left.\begin{aligned}
\lambda_{NULL,1} = \lambda_{NULL,2} &:= 0\\
\lambda_{LOW,1} = \lambda_{INCREASED,1} &:= \lambda_{-,1}\\
\lambda_{HIGH,1} = \lambda_{DECREASED,1} &:= \lambda_{+,1}\\
\lambda_{LOW,2} = \lambda_{DECREASED,2} &:= \lambda_{-,2}\\
\lambda_{HIGH,2} = \lambda_{INCREASED,2} &:= \lambda_{+,2}
\end{aligned}\right\} &\text{ – псевдонимы интенсивностей,}\\
\left.\begin{aligned}
(\alpha_{-,1}, \alpha_{+,1}, \alpha_{-,2}, \alpha_{+,2}) &\in \mathbb{R}_{>0}^{\{+,-\}\times\{1,2\}}\\
(\beta_{-,1}, \beta_{+,1}, \beta_{-,2}, \beta_{+,2}) &\in \mathbb{R}_{>0}^{\{+,-\}\times\{1,2\}}
\end{aligned}\right\} &\text{ – гиперпараметры интенсивностей,}\\
\overrightarrow{\mu} = (\mu_s)_{s\in\mathfrak{S}} \in \mathbb{R}_{>0}^{\mathfrak{S}} &\text{ – гиперпараметры начальных вероятностей,}\\
N = (\overrightarrow{\nu_r})_{r\in\mathfrak{S}} \in \mathbb{R}_{>0}^{\mathfrak{S}\times\mathfrak{S}} &\text{ – гиперпараметры вероятностей перехода.}
\end{aligned}
$$

Нас интересует исследование апостериорного распределения вероятностей $\mathcal{P}(\overrightarrow{s}, \overrightarrow{\pi}, A, \Lambda|D)$. В частности, маргинальное распределение $\mathcal{P}(\overrightarrow{s}|D)$ позволяет нам делать выводы о дифференциально насыщенных регионах.

Гиперпараметры фиксируются следующим образом:

$$\begin{aligned}
\alpha_{-,1} &= \alpha_{-,2} = 1, \\
\alpha_{+,1} &= \alpha_{+,2} = 9, \\
\beta_{-,1} &= \beta_{-,2} = 4/W, \\
\beta_{+,1} &= \beta_{+,2} = 12/W, \\
\overrightarrow{\mu} &= \mathbf{1}, \\
N &= \mathbf{1} + E,
\end{aligned}$$

где W – размер окна, а $\mathbf{1}$ означает (в зависимости от контекста) вектор или матрицу, все элементы которых равны единице. Мотивацию этих значений гиперпараметров см. 3.5.2.

7.3 Обучение

Мы приближаем искомое апостериорное распределение, используя алгоритм вариационного приближения, а именно VBEM, описанный в 3.4.1.

Очередные оценки для гиперпараметров апостериорного распределения считаются следующим образом:

$$\mu_s^{k+1} = \mu_s + \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_0 = s],$$

$$\nu_{rs}^{k+1} = \nu_{rs} + \frac{\sum\limits_{0<t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_{t-1} = r, s_t = s]}{\sum\limits_{0<t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_{t-1} = r]},$$

$$\alpha_{-,1}^{k+1} = \alpha_{-,1} + \sum_{0\leqslant t<T} d_{t,1}\mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = LOW \wedge s_t = INCREASED],$$

$$\alpha_{+,1}^{k+1} = \alpha_{+,1} + \sum_{0\leqslant t<T} d_{t,1}\mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = HIGH \wedge s_t = DECREASED],$$

$$\alpha_{-,2}^{k+1} = \alpha_{-,2} + \sum_{0\leqslant t<T} d_{t,2}\mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = LOW \wedge s_t = DECREASED],$$

$$\alpha_{+,2}^{k+1} = \alpha_{+,2} + \sum_{0\leqslant t<T} d_{t,2}\mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = HIGH \wedge s_t = INCREASED],$$

$$\beta_{-,1}^{k+1} = \beta_{-,1} + \sum_{0\leqslant t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = LOW \wedge s_t = INCREASED],$$

$$\beta_{+,1}^{k+1} = \beta_{+,1} + \sum_{0\leqslant t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = HIGH \wedge s_t = DECREASED],$$

$$\beta_{-,2}^{k+1} = \beta_{-,2} + \sum_{0\leqslant t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = LOW \wedge s_t = DECREASED],$$

$$\beta_{+,2}^{k+1} = \beta_{+,2} + \sum_{0\leqslant t<T} \mathbb{E}_{\vec{s}\sim q_{\vec{s}}^{k+1}}[s_t = HIGH \wedge s_t = DECREASED],$$

Мы оцениваем правдоподобие с помощью семплинга по важности (см. 3.4.2).

8 Результаты

Мы обучаем модель на агрегированном покрытии одного и того же геномного региона (как правило, хромосомы), полученном в результате двух экспериментов ChIP-seq (как правило, на двух родственных клеточных линиях). Мы производим валидацию модели, сравнивая её с двумя независимыми моделями PHMM для каждого трека, а также с моделью CMPHMM, отличающейся отсутствием состояния NULL. Оба сравнения показывают, что для большинства размеров окон BCZIPHMM имеет большее правдоподобие (см. таблицу

размер окна	$\log \mathcal{L}_{BCZIPHMM}$	$\log \mathcal{L}_{2\times BPHMM}$	$\log \mathcal{L}_{BCMPHMM}$
50	$*-3867261$	-5012256	-3874820
100	-3220008	-3822349	$*-3121767$
500	$*-1621935$	-1937743	-1762844
1000	$*-1226367$	-1442737	-1352232
5000	$*-675293$	-759971	-726937
10000	$*-521599$	-578464	-550646
50000	$*-288290$	-326742	-307852
100000	$*-242900$	-275045	-264421
500000	$*-181984$	-203096	-195168
1000000	$*-163032$	-180921	-172443
5000000	-91068	-91247	$*-86401$

Таблица 8: Логарифм правдоподобия BCZIPHMM и её преимущество перед CMPHMM и независимыми моделями PHMM.

8).

То, что для очень маленьких и очень больших окон BCMPHMM более правдоподобна, можно объяснить тем, что для маленьких окон нулевое покрытие проще объяснять как частный случай низкого покрытия (тем самым, дополнительное состояние не нужно), а при большом размере окна нулевое покрытие вообще не встречается, так что состояние NULL не появляется.

8.1 Реализация обучения моделей

Алгоритмы обучения моделей (а также предобработки данных) были реализованы на языке Java в рамках биоинформатического проекта компании JetBrains. Время обучения существенно зависит от сложности модели (классическая PM обучается быстрее всего, BCZIPHMM – медленнее всего), от размера окна (модели с маленькими окнами обучаются существенно медленнее) и от размера хромосомы. Время обучения на первой хромосоме домашней мыши *M. musculus*, в частности, составляет от нескольких миллисекунд (классическая PM с размером окна в 5 миллионов нуклеотидов) до суток (BCZIPHMM с размером окна в 50 нуклеотидов). Обучение модели для окна в 500 нуклеотидов занимает не более 15 минут даже для BCZIPHMM. Приведённые оценки времени работы получены для 64-битного настольного компьютера с процессором Intel Core 2 Quad CPU Q6700 2.66GHz $\times$ 4 и оперативной памятью объёмом 8 ГБ.

Следует отметить, что оптимизация алгоритмов обучения не входила в за-

размер окна	ChIPDiff	BCZIPHMM	пересечение
50	76108	10632	5076
100	7496	7490	3226
500	1479	2580	812
1000	692	1452	418
5000	134	307	98

Таблица 9: Количество окон, где ChIPDiff и BCZIPHMM предсказывают наличие изменений.

дачи данной работы и, в основном, не выполнялась, за исключением параллельного обучения моделей на нескольких хромосомах. Тем самым, можно предположить, что время обучения можно существенно улучшить.

8.2 Сравнение с аналогами

Самым близким аналогом нашей модели является алгоритм ChIPDiff (см. 1.6). Следует отметить, что ChIPDiff использует принципиально другой критерий для выявления различий в треках (кратное изменение покрытия), поэтому мы не ожидали совпадения результатов.

Мы обучили нашу модель и ChIPDiff на одних и тех же данных для разных размеров окна (ChIPDiff не удалось запустить для окон больше, чем 5000 нуклеотидов, из-за нехватки памяти). В таблице 9 приведено количество окон, в которых были предсказанны различия тем или иным алгоритмом, а также количество окон, в которых различия были предсказаны обоими алгоритмами.

Легко видеть, что пересечение, тем не менее, имеет существенный размер. Кроме того, можно сделать вывод о большей стабильности наших результатов: значения ChIPDiff для окон в 50 и 100 нуклеотидов различаются в 10 раз, тогда как естественнее было бы ожидать обратной пропорциональности размеру окна (что мы и наблюдаем в нашей модели).

9 Заключение

В ходе выполнения работы были построены и исследованы следующие модели для анализа данных ChIP-seq:

- Пуассоновская смесь (PM, см. 3) – моделирует один трек ChIP-seq как

смесь двух распределений Пуассона. Более правдоподобна, чем обычное распределение Пуассона.

- Скрытая марковская модель с пуассоновскими наблюдениями (PHMM, см. 4) – моделирует один трек ChIP-seq как смесь двух распределений Пуассона, причём выбор предыдущего распределения влияет на выбор следующего. Эта модель более правдоподобна, чем смесь, что подтверждает зависимость между физически близкими участками генома.

- Скрытая марковская модель с многомерными пуассоновскими наблюдениями (MPHMM, см. 5) – моделирует два трека ChIP-seq как смесь четырёх двумерных распределений Пуассона, причём выбор предыдущего распределения влияет на выбор следующего. Эта модель более правдоподобна, чем моделирование треков моделями PHMM по отдельности, что подтверждает зависимость между соответственными участками на разных треках. Байесовский вариант модели, однако, оказывается непригоден для биологического анализа, так как апостериорные распределения параметров теряют свойства, необходимые для нахождения отличающихся участков.

- Ограниченная скрытая марковская модель с многомерными пуассоновскими наблюдениями (CMPHMM, см. 6) – частный случай предыдущей модели с добавочным ограничением: четыре двумерных распределения Пуассона должны быть комбинациями двух распределений для первого трека и двух для второго. Эта модель, очевидно, хуже, чем MPHMM (так как на неё накладывается дополнительное ограничение), зато более биологически интерпретируема (так как независимо определяет высокое и низкое покрытие для каждого трека). Байесовский вариант этой модели, обученный с помощью VBEM, имеет большее правдоподобие, чем классический ОМП-вариант, обученный с помощью EM, что побуждает нас отдать предпочтение байесовскому варианту.

- BCZIPHMM – вариация предыдущей модели, к которой добавляется пятое распределение (точечное распределение, сосредоточенное в $(0, 0)$). Это добавление мотивировано тем, что некоторые области генома никогда не бывают покрыты экспериментом ChIP-seq, например, из-за отсутствия референсной последовательности. Модель BCZIPHMM оказы-

вается более правдоподобна, чем все предыдущие модели, что подтверждает выгодность введения пятого состояния.

Таким образом, мы построили модель, подходящую для сравнения результатов двух экспериментов ChIP-seq. В отличие от существующего аналога (ChIPDiff), использующего кратное изменение покрытия в качестве критерия, наша модель (BCZIPHMM) находит качественные изменения покрытия (замену высокого на низкое или наоборот). Вычислив правдоподобие модели на реальных данных, мы валидируем сделанные в ходе построения модели предположения, как-то: предположение о зависимости состояний на треках, предположение о зависимости физически близких состояний, предположение о необходимости состояния NULL.

Тем самым, поставленную в начале работы цель следует считать достигнутой: построенная нами модель правдоподобна (все сделанные предположения валидированы), биологически интерпретируема (области с изменением связывания соответствуют состояниям INCREASED и DECREASED) и поддаётся обсчёту за реальное время (см. 8). Кроме того, найденные моделью области имеют существенное пересечение с результатами того же ChIPDiff.

Модель применяется для поиска дифференциально насыщенных состояний. Сведения об изменениях в связывании белка между схожими клеточными линиями могут помочь исследованию механизмов, отвечающих за различающееся поведение клеток, причин дифференцирования, особенностей ракового поражения. Планируется применение результатов модели для изучения связи между гистонными модификациями и другими эпигенетическими маркерами (например, метилированием ДНК), экспрессией генов, эффективности методик индуцированной плюрипотентности.

В качестве направления дальнейших исследований можно указать дальнейшее усложнение модели, которое может привести к лучшему соответствию модели реальным данным, перспективы замены распределения Пуассона на другие распределения (отрицательное биномиальное, бета-биномиальное), использование данных обычного секвенирования для учёта неравномерности покрытия, включение в исследование данных о неравномерном распределении тегов по двум цепям ДНК, применение модели для поиска паттернов сочетаний двух белков в пределах одной клеточной линии (вместо использовавшихся в работе паттернов различий в связывании одного белка на двух клеточных линиях).

10 Глоссарий

- B, см. Бета-распределение
- Г, см. Гамма-распределение
- BCZIPHMM, см. 7, стр. 43
- ChIP-chip, см. 1.4, стр. 7
- ChIP-seq, см. Хроматин-иммуннопреципитационное секвенирование
- CMPHMM, см. Ограниченная скрытая марковская модель с многомерными пуассоновскими наблюдениями
- Dir, см. Распределение Дирихле
- EM, см. 3.3, стр. 19
- GP, см. Обобщённое распределение Пуассона
- KL, см. Дивергенция Кульбака-Лейблера
- MPHMM, см. Скрытая марковская модель с многомерными пуассоновскими наблюдениями
- PHMM, см. Скрытая марковская модель с пуассоновскими наблюдениями
- PM, см. Пуассоновская смесь
- Pois, см. Распределение Пуассона
- VBEM, см. Вариационный байесовский EM
- ZIP, см. Распределение Пуассона с подкачкой нуля
- Алгоритм Баума-Велша, см. 4.3.2, стр. 30
- Алгоритм Витерби, см. 4.3.1, стр. 30
- Алгоритм прямого-обратного хода, см. 4.3.1, стр. 29
- Бета-распределение, см. 1.5.2, стр. 10

- Вариационный байесовский EM, см. 3.4.1, стр. 22
- Гамма-распределение, см. 1.5.2, стр. 10
- Дивергенция Кульбака-Лейблера, см. 1.5.6, стр. 13
- Обобщённое распределение Пуассона, см. 1.5.3, стр. 11
- Ограниченная скрытая марковская модель с многомерными пуассоновскими наблюдениями, см. 6, стр. 39
- ОМП, см. Оценка максимума правдоподобия
- Оценка максимума правдоподобия, см. 1.5, стр. 10
- Пуассоновская смесь, см. 3, стр. 18
- Распределение Дирихле, см. 1.5.2, стр. 10
- Распределение Пуассона, см. 1.5.2, стр. 10
- Семплинг по важности, см. 3.4.2, стр. 24
- Скрытая марковская модель с многомерными пуассоновскими наблюдениями, см. 5, стр. 35
- Скрытая марковская модель с пуассоновскими наблюдениями, см. 4, стр. 28
- Хроматин-иммуннопреципитационное секвенирование, см. 1.4.3, стр. 9

Список литературы

[1] Alberts B. *et al. Molecular Biology of the Cell, 4th ed.* Garland Science, ISBN 0-8153-3218-1

[2] Asp P. *et al. Genome-wide remodelling of the epigenetic landscape during myogenic differentiation.* Proc Natl Acad Sci USA 2011 May 31;108(22):E149-58. PMID: 21551099

[3] Beal, Matthew J. *Variational Algorithms for Approximate Bayesian Inference.* University College London, May 2003.

[4] Choi H. *et al. Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data.* Bioinformatics, 2009 Jul 15; 25(14):1715-21. doi: 10.1093/bioinformatics/btp312

[5] Fraga, M. F. *et al., Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer.* Nat Genet. 2005 Apr; 37(4):391-400. Epub 2005 Mar 13.

[6] Gelman, Andrew *et al., Bayesian Data Analysis, 2nd edition.* CRC Press, 2003. ISBN 1-58488-388-X.

[7] Jiang, H. and Wong, W. *Statistical inferences for isoform expression in RNA-Seq.* Bioinformatics 2009, 25 (8), 1026-1032.

[8] Qin, Zhaohui S. *et al. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data.* BMC Bioinformatics 2010, 11:369. doi:10.1186/1471-2105-11-369

[9] Spyrou, Christina *et al. Bayesian analysis of ChIP-seq data.* BMC Bioinformatics 2009, 10:299. doi:10.1186/1471-2105-10-299

[10] Xu H. *et al. Identifying differential histone modification sites from ChIP-seq data.* Methods Mol Biol. 2012; 802:293-303. doi: 10.1007/978-1-61779-400-1_19

[11] https://www.ncbi.nlm.nih.gov/geo/info/overview.html

[12] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25308