

Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики
Факультет информационных технологий и программирования

Кафедра информационных систем

Бутомов Артем Сергеевич

Разработка программного компонента для проведения сравнительного анализа биологических данных FAIRE-seq

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:

Научный руководитель:
Лебедев Сергей Андреевич

Рецензент:
аспирант Сергушичев Алексей Александрович

Санкт-Петербург
2015

Saint Petersburg National Research University of Information Technologies,
Mechanics and Optics
Faculty of Information Technology and Software Engineering
Chair of Information Systems

Artiom Butomov

Development of a software component for a
comparative analysis of the biological
FAIRE-seq data

Graduation Thesis

Admitted for defence.
Head of the chair:

Scientific supervisor:
Sergei Lebedev

Reviewer:
postgraduate Alexey Sergushichev

Saint-Petersburg
2015

Оглавление

Введение	4
1. Предлагаемые модели	6
1.1. Описание задачи	6
1.2. Смесь многомерных распределений Пуассона	6
1.2.1. Предсказание модели	9
1.3. Скрытая Марковская Модель	9
1.4. Выбор модели	12
2. Оценка модели/ей	13
2.1. Оценка и контроль FDR	13
Заключение	14
Список литературы	15

Введение

ДНК (дезоксирибонуклеиновая кислота) — длинная двухцепочечная молекула, являющаяся носителем генетической информации в биологических организмах.

Изучать пространственную структуру ДНК организма важно для понимания механизмов регуляции жизнедеятельности клетки.

Формальдегидная изоляция регуляторных элементов с последующим секвенированием (Formaldehyde-Assisted Isolation of Regulatory Elements sequencing, FAIRE-Seq) — Это биологический протокол, позволяющий находить участки, в которых ДНК доступна для связывания белками. Суть работы метода заключается в том, что на ДНК, выделенную из клетки, "прикрепляют" нуклеосомы с помощью формальдегида. Затем ДНК фрагментируют с помощью ультразвука. После этого происходит "разделение" полученных фрагментов ДНК на две группы: участки связанные с белками и "свободные" участки. Далее "свободные" фрагменты "читают" с помощью секвенатора. И наконец, для каждого прочтения секвенатора определяют место в геноме исследуемого организма, откуда он был прочитан.

В контексте данной работы, геном "разбивается" на непересекающиеся отрезки фиксированной длины, называемые бинами. Подсчитывается количество прочтений, начинающихся внутри каждого отрезка. Таким образом, получатся вектор из неотрицательных целых чисел, именуемый вектором покрытия.

Из вектора покрытия можно сделать предположение о вероятности расщепления региона, чем больше значение элемента вектора, тем с большей вероятностью, что регион, соответствующий элементу, был расщеплен.

Однако, рассматриваемый протокол не исключает возможности наличия ошибок в результатах биологического эксперимента. Неточности метода FAIRE-seq обусловлены следующими моментами:

1. Протокол работает с колонией клеток. Таким образом в результатах эксперимента мы видим некоторое среднее состояние по всем клеткам
2. Этап фиксации не обладает 100% КПД, то есть некоторые белки могут "отвалиться"
3. Этап разделения "свободных" и "связанных" фаз также неточен. Вместе со "свободными" вполне могут попасться и связанные фрагменты

Так как в результате эксперимента появляется шум, данные FAIRE-seq удобно анализировать с помощью вероятностных моделей.

Цель данной работы - разработать математическую модель для проведения сравнительного анализа нескольких экспериментов биологических данных FAIRE-seq, научиться оценивать и контролировать число неверных предсказаний модели.

Для достижения цели были поставлены следующие задачи:

1. Изучить предметную область
2. Предложить несколько вероятностных моделей для сравнения экспериментов FAIRE-seq
3. Реализовать модели в виде программы на языке Python
4. Оценить эффективность полученной программы

1. Предлагаемые модели

1.1. Описание задачи

Пусть $\vec{x} = (x_1, \dots, x_N)$ - вектор прочтений, построенный из какого-то ВАР файла. Сопоставим каждому наблюдению некоторую метку-состояние s_n из множества базовых состояний $s = \{1, \dots, S\}$, истинные значения которых не знаем.

Вероятностная модель позволяет найти наиболее правдоподобную последовательность состояний.

$$\hat{s}_{ML} = \arg \max_{s \in 1, \dots, S^N} \mathcal{P}(x, s; \theta)$$

1.2. Смесь многомерных распределений Пуассона

Будем считать, что мы работаем с двумя векторами покрытия. Тогда наблюдение x_n - это пара чисел, каждое из которых соответствует одному вектору $\vec{x} = \{\vec{x}_1, \vec{x}_2\}$.

Пусть $S = \{+, -\}$ — множество состояний, которые мы будем называть "базовыми" состояниями. Каждое из базовых состояний описывает ситуацию в одном образце, например, $(+)$ — сигнал есть, $(-)$ - шум или сигнала нет.

Для задачи сравнения нам нужно множество состояний, описывающее, что происходит в каждом из образцов, то есть $S^2 = \{(+, +), (-, -), (+, -), (-, +)\}$.

Будем моделировать количество прочтений вдоль генома с помощью смеси многомерных распределений Пуассона.

Методом максимума правдоподобия найти решение довольно сложно, поскольку для этого требуется решить систему, где оцениваемые параметры зависят от наблюдаемой выборки x и неизвестных значений скрытых состояний s . Асимптотическая сложность решения возрастет в S^N раз.

Поэтому, будем искать приближенное решение с помощью ЕМ-алгоритма, в котором правдоподобие оптимизируется до сходимости. Последовательность действий сформируется следующим образом[1]:

1. Инициализировать начальные значения параметров
2. Вычислить нижнюю оценку на правдоподобие
3. Найти новое значение параметров модели
4. Вычислить логарифм правдоподобия и проверить на сходимость

Выведем ЕМ-алгоритм для смеси многомерных Пуассоновских распределений.

$$\ln p(X; \pi, \lambda) = \sum_{n=1}^N \ln \sum_{k=1}^K \pi_k \mathcal{P}(x_n; \lambda_k) \quad (1)$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{P}(x_n; \lambda_k)}{\sum_{j=1}^K \pi_j \mathcal{P}(x_n; \lambda_j)} \quad (2)$$

$$\pi_k^* = \frac{N_k}{N} \quad (3)$$

Шаг М

$$\begin{aligned} & \frac{\partial}{\partial \lambda_k} E[\log p(x|z; \theta)] \\ &= \frac{\partial}{\partial \lambda_k} \sum_{n=1}^N \sum_{k=1}^K E[z_{nk}] \{\log \pi_k + \log \mathcal{P}(x_n|\theta)\} \\ &= \sum_{n=1}^N E[z_{nk}] \frac{\partial}{\partial \lambda_k} \log \mathcal{P}(x_n|\theta) \\ &= \sum_{n=1}^N \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} \log \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} \\ &= \sum_{n=1}^N \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} (\log \lambda_k^{x_n} + \log e^{-\lambda_k} - \log x_n!) \\ &= \sum_{n=1}^N \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} (x_n \log \lambda_k - \lambda_k - \log x_n!) \\ &= \sum_{n=1}^N \gamma(z_{nk}) \left(\frac{x_n}{\lambda_k} - 1 \right) = 0 \end{aligned}$$

получаем

$$\lambda_k^* = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) x_n \quad (4)$$

где

$$N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (5)$$

Найдем новые значения априорных вероятностей. Воспользуемся методом множителей Лагранжа для учета ограничений на вектор априорных вероятностей.

$$\frac{\partial}{\partial \pi_k} (E[\log p(x|z; \theta)] + \lambda (\sum_{j=1}^K \pi_j - 1))$$

$$\begin{aligned}
&= \frac{\partial}{\partial \pi_k} \sum_{n=1}^N \sum_{k=1}^K E[z_{nk}] \{\log \pi_k + \log \mathcal{P}(x_n|\theta)\} + \frac{\partial}{\partial \pi_k} \lambda \left(\sum_{j=1}^K \pi_j - 1 \right) \\
&= \sum_{n=1}^N \gamma(z_{nk}) \frac{\partial}{\partial \pi_k} \{\log \pi_k + \log \mathcal{P}(x_n|\theta)\} + \frac{\partial}{\partial \pi_k} \lambda \left(\sum_{j=1}^K \pi_j - 1 \right) \\
&= \sum_{n=1}^N \gamma(z_{nk}) \frac{\partial}{\partial \pi_k} \{\log \pi_k\} + \frac{\partial}{\partial \pi_k} \lambda \left(\sum_{j=1}^K \pi_j - 1 \right) \\
&= \sum_{n=1}^N \gamma(z_{nk}) \frac{1}{\pi_k} + \lambda = 0
\end{aligned}$$

Домножив на π_k , получаем

$$\sum_{n=1}^N \gamma(z_{nk}) + \pi_k \lambda = 0 \quad (6)$$

Просуммируем вдоль k

$$\sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) + \sum_{k=1}^K \pi_k \lambda = 0 \quad (7)$$

Применяя

$$\sum_{k=1}^K \gamma(z_{nk}) = 1$$

и

$$\sum_{k=1}^K \pi_k = 1$$

Из выражения (7) получаем

$$N + \lambda = 0$$

$$\lambda = -N$$

Далее, подставляя найденное λ в выражение (6), получаем

$$\pi_k = -\frac{\sum_{n=1}^N \gamma(z_{nk})}{\lambda} = \frac{N_k}{N}$$

Наконец,

$$\pi_k^* = \frac{N_k}{N}$$

$$s_n = \arg \max_{s \in 1, \dots, S^N} \gamma_{ni}$$

Примечание. Чтобы успешно обучить наши данные, следует правильно проинициализировать начальные значения входных параметров модели. Для этой задачи достаточно использовать алгоритм кластеризации KMeans++, который ”разбивает”

наши наблюдения на S кластеров. Работа алгоритма основана на Методе Максимального Правдоподобия. На шаге Е мы определяем для каждого наблюдения ближайший кластер. На шаге М Вычисляем новое значение кластера, которое принимаем за среднее выборочное наблюдений, относящихся к данному кластеру. Алгоритм итерируется до тех пор, пока изменения логарифма правдоподобия не станет меньше 10^{-3} . В результате, начальные значений параметров Пуассоновского распределения можно принять за значения кластеров.

1.2.1. Предсказание модели

В результате работы алгоритма, необходимо предсказать наиболее вероятную последовательность состояний, породивших наши наблюдения. Для каждого наблюдения выбирается состояние, соответствующее наибольшей апостериорной вероятности.

1.3. Скрытая Марковская Модель

На практике условие о независимости состояний между соседними наблюдениями в предыдущей модели не выполняется.

Поэтому, перейдем к Скрытой Марковской Модели второго порядка, чтобы учесть зависимость между состояниями соседних наблюдений.

$$z_{(n+1),i} \text{ Зависит от } z_{n,i}$$

Введем понятие базовых состояний: $S \in \{+, -, \text{null}\}$. Каждое из базовых состояний описывает ситуацию в одном образце.

Семантика обозначений следующая:

(+) - сигнал есть, (−) - шумовый сигнал, (null) - сигнала нет.

Замечание. Отличие (null) от (−) заключается в полном отсутствии сигнала.

Для задачи сравнения нам нужно множество состояний, описывающее, что происходит в каждом из образцов, то есть

$$S^2 \in (+, +), (-, -), (\text{null}, \text{null}), (+, -), (-, +), (\text{null}, -), (\text{null}, +), (-, \text{null}), (+, \text{null})$$

Состояние S_t с одинаковым базовыми состояниями $(+, +), (-, -), (\text{null}, \text{null})$ означает, что данные сравниваемых образцов наблюдения x_t похожи между собой.

Пусть

$\pi = (\pi_1, \dots, \pi_S)$ - априорные вероятности состояний.

\mathcal{A} - матрица вероятностей перехода между состояниями.

$\lambda = (\lambda_1, \dots, \lambda_S)$ - параметры многомерного распределения Пуассона.

Тогда, функция правдоподобия определяется как

$$p = \dots$$

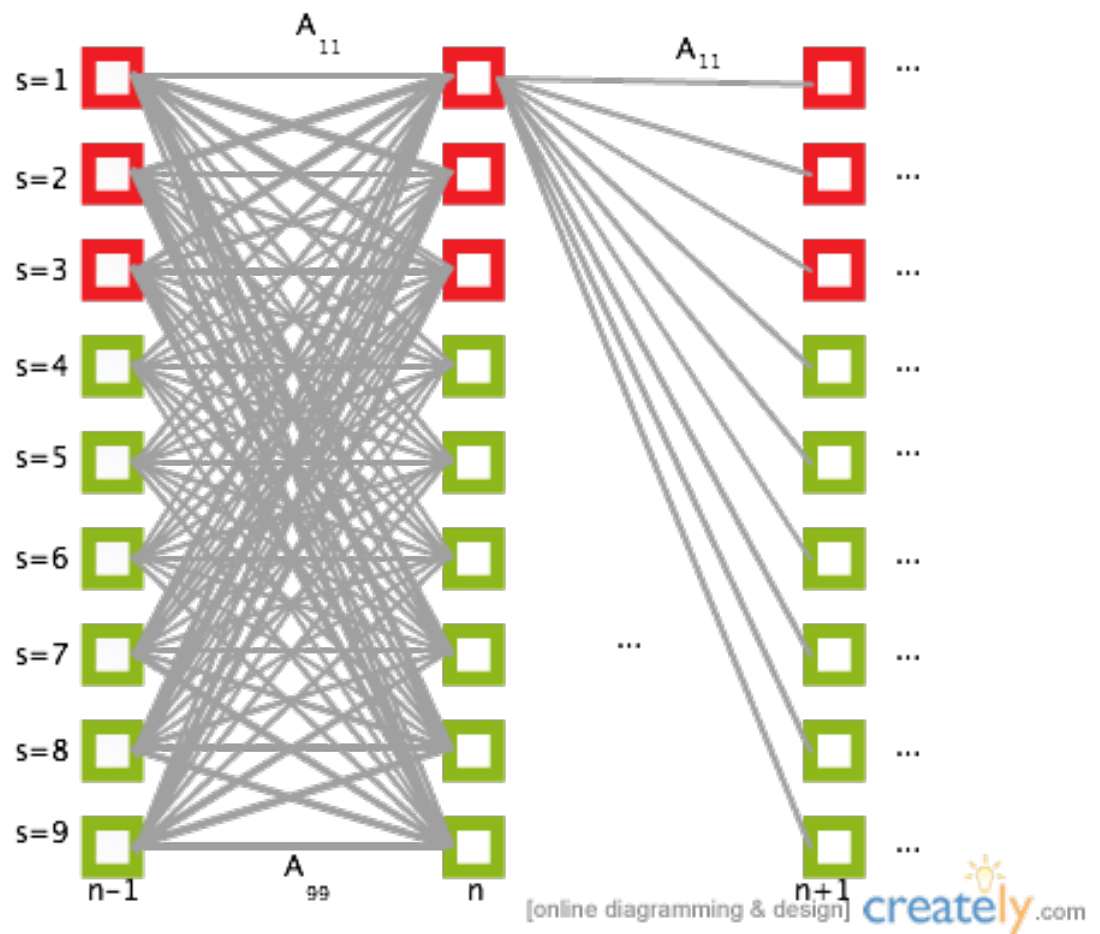


Рис. 1: решетка состояний

Решетка представляет собой диаграмму переходов между скрытыми состояниями модели. Красным светом выделены состояния, в котором базовые состояния похожи, то есть $(+, +), (-, -), (\text{null}, \text{null})$

Шаг Е

$$\gamma_{ni} = \frac{\alpha_{ni}\beta_{ni}}{\sum_{j=1}^S \alpha_{nj}\beta_{nj}}$$

$$\xi_{nij} = \frac{\alpha_{(n-1),i} A_{ij} \mathcal{P}(x_n; \lambda_j) \beta_{nj}}{\sum_{i'=1}^S \sum_{j'=1}^S \alpha_{(n-1),i'} A_{i'j'} \mathcal{P}(x_n; \lambda_{j'}) \beta_{nj'}}$$

Где,

$$\alpha_{ni} = p(s_{ni} = 1, x_1, x_2, \dots, x_n; \theta)$$

$$\beta_{ni} = p(x_{n+1}, \dots, x_N; s_{ni} = 1, \theta)$$

Вычисления α и β производится с помощью алгоритма прямого-обратного хода:

$$\alpha_{1i} = \pi_i \mathcal{P}$$

$$\beta_{Ni} = 1$$

$$\alpha_{ni} = \mathcal{P}(x_n; \lambda_i) \sum_{j=1}^S \alpha_{(n-1),j} A_{ji}$$

$$\beta_{ni} = \sum_{j=1}^S A_{ij} \mathcal{P}(x_{n+1}; \lambda_j) \beta_{(n+1),j}$$

Шаг М

Новые значения параметров модели

$$\pi_i^* = \gamma_i 1$$

$$A_{ij}^* = \frac{\sum_{n=2}^N \xi_{nij}}{\sum_{j'=1}^S \sum_{n=2}^N \xi_{nij} \xi_{nij'}}$$

$$\lambda_i^* = \frac{\sum_{n=1}^N \gamma(z_{ni}) x_n}{\sum_{n=1}^N \gamma(z_{ni})}$$

Замечание. В алгоритме прямого-обратного хода может быть underflow - это значит, что не хватает точности чисел с плавающей точкой. Поэтому удобно проводить вычисления в логарифмах.

1.4. Выбор модели

Для задачи сравнения двух образцов была выбрана Скрытая Марковская Модель. В частном случае, для анализа одного биологического образца СММ более правдоподобнее, чем смесь Пуассоновских испусканий. Это значит, что модель, в которой предполагается зависимость между состояниями соседних наблюдений, более правдоподобная.

2. Оценка модели/ей

2.1. Оценка и контроль FDR

Заключение

Список литературы

- [1] Bishop Christopher. Pattern Recognition and Machine Learning (Information Science and Statistics). — Springer, 2007. — URL: <http://www.ozon.ru/context/detail/id/2978313/>.