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

Кафедра информационных систем и технологий

Бутомов Артем Сергеевич

Разработка программного компонента для проведения сравнительного анализа биологических данный 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 and Technologies

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

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

Изучать пространственную структуру ДНК организма важно для понимания механизмов регуляции жизнидеятельности 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 расплетения региона, чем больше значение элемента вектора, тем с большей вероятностью, что регион, соответствующий элементу, был расплетен.

Однако, рассматриваемый протокол не исключает возможности наличия ошибок в результатах биологического эксперемнта. 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 эксперемента появляется шум, данные FAIRE-seq удобно анализировать с помощью вероятностых моделей.

Цель данной работы - разработать математическую модель для проведения сравнительного анализа нескольких эксперементов биологический 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

1.2. Смесь многомерных распределений Пуассона

Таким образом с помощью наибольшего правдоподобия можно найти оценку последовательности скрытых состояний.

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

Моделировать количество прочтений вдоль генома с помощью смеси многомерных распределений Пуасона

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 EM-алгоритма, в котором правдоподобие оптимизируется до сходимости. Последовательность действий формируется 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 EM-алгоритм для смеси многомерных Пуасоновских испусканий.

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

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

$$\pi_k^* = \frac{N_k}{N} \tag{3}$$

Шаг M

$$\frac{\partial}{\partial \lambda_k} E[\log p(x|z; \theta)]$$

$$= \frac{\partial}{\partial \lambda_k} \sum_{n=1}^{N} \sum_{k=1}^{K} E[z_{nk}] \{\log \pi_k + \log \mathcal{P}(x_n|\theta)\}$$

$$= \sum_{n=1}^{N} E[z_{nk}] \frac{\partial}{\partial \lambda_k} \log \mathcal{P}(x_n|\theta)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} \log \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} (\log {\lambda_k}^{x_n} + \log e^{-\lambda_k} - \log x_n!)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \lambda_k} (x_n \log \lambda_k - \lambda_k - \log x_n!)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) (\frac{x_n}{\lambda_k} - 1) = 0$$

получаем

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

где

$$N_k = \sum_{n=1}^{N} \gamma(z_n k) \tag{5}$$

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

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

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

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \pi_k} \{\log \pi_k + \log \mathcal{P}(x_n|\theta)\} + \frac{\partial}{\partial \pi_k} \lambda (\sum_{j=1}^{K} \pi_j - 1)$$

$$= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial}{\partial \pi_k} \{\log \pi_k\} + \frac{\partial}{\partial \pi_k} \lambda (\sum_{j=1}^{K} \pi_j - 1)$$

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

Домножив на π_k, получаем

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

Просуммируем вдоль k

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

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,

$$\pi_k^* = \frac{N_k}{N} \tag{8}$$

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

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 E мы определяем для каждого наблюдения ближайший кластер. На шаге M 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}

1.2.1. Предсказание модели

В результате работы алгоритма, необходимо предскзать наиболее вероятную послеовательность 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.

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

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

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.

Отличие (null) от $(-)$ заключаются в полном отсутсвии сигнала.

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

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

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/.