

Скрытые Марковские модели переменного порядка для анализа данных ChIP-seq

Атаманова Анна, кафедра системного программирования СПбГУ, anne.atamanova@gmail.com

5 апреля 2015 г.

Аннотация

Здесь нужно кратко описать суть работы и результаты. Цели работы:

Целью данной работы стоит изучение марковской модели переменного порядка, ее реализация и применение на данных ChIP-seq

1 Введение

ДНК (дезоксирибонуклеиновая кислота) — длинная двухцепочечная молекула, являющаяся носителем генетической информации в биологических организмах. В клетках эукариот ДНК находится в упакованном состоянии. Упаковка ДНК реализована с участием специальных белковых комплексов — нуклеосом. Химические модификации субъединиц нуклеосомы, гистонов, могут влиять на плотность упаковки ДНК. Увеличение плотности ДНК влияет на доступность соответствующих участков ДНК для внутренней машинерии клетки.

Иммунопреципитация хроматина с последующим секвенированием (chromatin immunoprecipitation sequencing, ChIP-seq) — это биологический протокол, позволяющий получить информацию о наличии или отсутствии некоторой химической модификации гистонов вдоль генома [1]. Суть метода заключается в использовании антитела для отбора фрагментов ДНК, связанных с гистонами, имеющими изучаемую химическую модификацию с последующим секвенированием. В ходе секвенирования случайные фрагменты ДНК, читаются секвенатором в объеме, достаточном для того, чтобы с большой вероятностью каждый фрагмент был прочитан несколько раз. Затем для каждого полученного прочтения ищется соответствующий ему участок последовательности генома (рис. 1). Обычно прочтения, которым может соответствовать более одного участка в геноме, исключают из рассмотрения.

```
CAAAAGACAAATAGTGATGTCCCAATCGAGC
-----
      GACA ATA      GTCA  AATG
AGAC  TAGTG TGTC
      GACA  AGTG TGTC  ATCG

00001100001110000110000001000000
```

Рис. 1: Схематическое изображение выравнивания прочтений секвенатора (под чертой) на известную последовательность генома (над чертой).

Результаты эксперимента представляют в виде вектора длины генома, в котором стоит 1, если в соответствующей позиции генома начинается хотя бы одно прочтение и 0 в обратном случае.

Протокол хроматин-иммунопреципитации, как и большинство биологических протоколов, не исключает наличие в результатах эксперимента ошибок. Недостаточная специфичность антитела, наличие ошибок секвенирования и нестабильность положения гистонов на ДНК приводят к возникновению сигнала не зависящего от наличия изучаемой модификации гистонов. Использование вероятностных моделей позволяет провести анализ результатов хроматин-иммунопреципитации с учётом наличия ошибок.

Большинство существующих моделей (TODO: ref) для данных хроматин-иммунопреципитации основано на аппарате скрытых Марковских моделей второго порядка с Пуассоновскими испусканиями. Использование распределения Пуассона для покрытия, опирается на предположение о том, что в каждой позиции генома в среднем начинается одинаковое количество прочтений. Марковский процесс, как правило, имеет два состояния + — сигнал есть и — — сигнала нет. Второй порядок модели означает, что состояние некоторого окна зависит только от состояния его прямого предшественника.

Использование моделей второго порядка объясняется тем, что количество параметров модели, а также сложность её обучения и использования экспоненциально зависят от порядка, то есть, модель порядка m требует оценки 2^m параметров.

В настоящее время, в качестве семейства искоемых моделей, активное применение находит HMM (Hidden Markov Model)[2] второго порядка с Пуассоновским испусканием. Данное семейство допускает предположение о том, что каждое состояние (наличие/отсутствие белка в заданной части генома) зависит только от одного предыдущего. Можно ограничиться и более лояльным допущением о том, что состояние зависит от n предыдущих состояний, однако такое допущение резко увеличивает сложность модели ($O(2^n)$ параметров). Также, сложность заключается в подборе этого n и переобучении в случае, если не все состояния имеют одинаковые длины контекстов зависимости. Последнее замечание подводит к идее использования VOHMM (Variable Order Hidden Markov Model)[3]

2 Скрытые марковские модели переменного порядка

Определение. Последовательность случайных величин $\{X_i\}_{i \in Z}$ называется *цепью Маркова порядка m* , если $\forall t \in Z$

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2} \dots X_{-\infty} = x_{-\infty}) = P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2} \dots X_{t-m+1} = x_{t-m+1})$$

Определение. Марковская цепь является *однородной*, если вероятностное распределение переходов $P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2} \dots X_{t-m+1} = x_{t-m+1})$ едино для всех t . Далее будем обозначать просто $P(x_t | x_{t-1} \dots x_{t-m+1})$

Определение. *Марковской моделью (Markov Model (MM)) порядка m* называют вероятностную модель, описывающую однородный марковский процесс порядка m . Параметрами модели являются множество состояний $S = \{1..n\}$ и множество переходов $A = \{a(q, x^m)\}_{q \in S, x^m \in S^m}$, где $a(q, x^m) = P(q | x^m)$.

Определение. *Скрытая Марковская модель (Hidden Markov Model (HMM)) порядка m* - вероятностная модель, параметрами которой являются множество скрытых состояний $S = \{1..n\}$, множество переходов $A = \{a(q, x^m)\}_{q \in S, x^m \in S^m}$ и множество распределений испусканий $B = \{b(y, x)\}_{y \in R^l, x \in S}$, где $b(y, x) = P(y | x)$. Такая модель описывает цепь $\{Y\}_{i \in Z}$, если ее состояния были испущены из состояний марковской цепи $\{X_i\}_{i \in Z}$ с параметрами A согласно распределению $P(y | x)$, и $P(y_t | y_{t-1} \dots y_{t-m+1}) = P(x_t | x_{t-1} \dots x_{t-m+1})P(y_t | x_t)$

На рисунке (Рис 2) схематично представлена скрытая марковская модель порядка 2.

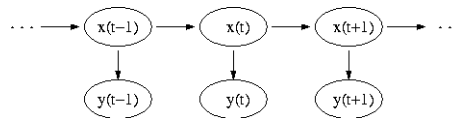


Рис. 2: HMM order 2

Определение. *Контекстное дерево* - дерево (бор), в котором каждая внутренняя вершина имеет n ребер соответствующих состояниям $\{1..n\}$ и метку, которая является конкатенацией метки на ее родителе и метки ребра от него. Корень помечен пустой строкой.

Определение. *Скрытая марковская модель переменного порядка (Variable Order Hidden Markov Model (VOHMM))* - вероятностная модель, параметрами которой являются множество скрытых состояний $S = \{1..n\}$, конечное множество контекстов $C = \{c_i\}_i$, где c_i - листья некоторого контекстного дерева, множество переходов $A = \{a(q, c)\}_{q \in S, c \in C}$ и множество распределений испусканий $B = \{b(y, x)\}_{y \in R^l, x \in S}$, где $b(y, x) = P(y | x)$.

Обучение модели VOHMM:

Задача:

По цепи наблюдений $Y = (y_1, \dots, y_T)$ найти параметры модели Λ , которые бы максимизировали правдоподобие при максимально сжатых контекстах ¹

Алгоритм:

Параметры алгоритма: m - максимальная длина контекста, ϵ_{EM} - барьер для остановки EM, ϵ_{prune} - барьер для обрезания дерева

¹Параметр алгоритма ϵ определяет допустимое отклонение распределений

1. Инициализация контекстов.

$$C_0 = \{c | c \in S^m\}$$

Начальное распределение переходов произвольное. ²

2. EM (Expectation–Maximization algorithm).

Пересчет производится подобно алгоритму Baum-Welch для HMM

- (a) Expectation

Вводятся дополнительные параметры:

$$\alpha_t(c) = P(y_t^1, c(x_t) = c | \Lambda)$$

$$\beta_t(c) = P(y_{t+1}^T | c(x_t) = c, \Lambda)$$

$$\gamma_t(c) = P(x_t = c | Y, \Lambda)$$

$$\xi_t(q, c) = P(c(x_t) = c, x_{t+1} = q | Y, \Lambda)$$

с помощью которых итеративно пересчитываются параметры модели

$$\alpha_0(c) = p(c)b(y_0, c), \alpha_{t+1}(c) = \sum_{q \in S, c' = C(qc)} \alpha_t(c')a(c[0], c')b(y_{t+1}, c[0])$$

$$\beta_T(c) = 1, \beta_t(c) = \sum_{q \in S, c' = C(qc)} a(q, c)b(y_{t+1}, c'[0])\beta_{t+1}(c')$$

$$p = P(Y | \Lambda) = \sum_{c \in C} \alpha_T(c)$$

$$\gamma_t(c) = \frac{\alpha_t(c)\beta_t(c)}{p}$$

$$p(c) = \sum_t \gamma_t(c)$$

- (b) Maximization

$$\xi_t(q, c) = \frac{\alpha_t(c)a(q, c)b(y_{t+1}, c)\beta_{t+1}(qc)}{p}$$

$$a(q, c) = \frac{\sum_t \xi_t(q, c)}{p(c)}$$

Пересчет B зависит от принятого семейства моделей испусканий. Производится с помощью γ в точности также как и в алгоритме Baum-Welch.

Пересчет EM проходит до тех пор пока разница правдоподобий между итерациями не будет меньше ϵ_{EM}

3. Обрезание дерева.

Если существует внутренний лист s такой, что $\forall q \in S kl(sq, s) < \epsilon_{prune}$ (дети не уточняют родителя), то s становится листом, а все его потомки обрезаются.

$kl(u, w) = \sum_{q' \in S} P(q' | u) \log \frac{P(q' | u)}{P(q' | w)}$ - расстояния Кульбака-Лейблера для апостериорных распределений.

4. Если на третьем шаге ничего не произошло, то алгоритм заканчивает работу, иначе происходит обновление матрицы a для новых контекстов

$$a(q, c) = P(q | c)$$

и алгоритм переходит на второй шаг.

Обозначения:

$c(x_t)$ - контекст состояния x_t

$C(s)$ - листья, являющиеся потомками s , если s принадлежит дереву

$C'(s)$ - контекст максимальной длины, являющийся префиксом s , если s не принадлежит дереву

Замечание. Вероятностные переходы на листьях задают вероятностные переходы на всем дереве

$$p(q | s) = \frac{\sum_{c \in C(s)} p(q | c)}{\sum_q \sum_{c \in C(s)} p(q | c)}$$

Замечание. При пересчете вероятности могут очень близко подходить к нулю, что негативно сказывается на точность расчета. Для избежания этой проблемы все расчеты проходят не с вероятностями, а с логарифмами от них.

Замечание. EM может застревать в локальных максимумах функции правдоподобия.

3 Simulation

Ниже приведено несколько примеров результата обучения алгоритма VONMM по выборке, построенной по фиксированному дереву.

²В определенных случаях (Gauss, Poisson) частотное распределение, полученное из цепи алгоритмом k-means (k=m), ускоряет работу

Приведено сравнение реальных деревьев и деревьев предсказанных алгоритмом, и график роста правдоподобия на всех итерациях. В качестве распределений испусканий выбраны двумерное распределение Гаусса и одномерное распределение Пуассона.

1. Смесь. Рисунки 3 и 4

Начальное дерево: единственный пустой контекст. Длина сэмпированной выборки $T = 4000$.

Параметры алгоритма: максимальная длина контекстов $m = 2$, барьер для обрезания $\epsilon_{prune} = 0.004$, барьер для останова EM $\epsilon_{EM} = 0.01$ (сравнение идет по логарифму правдоподобия)

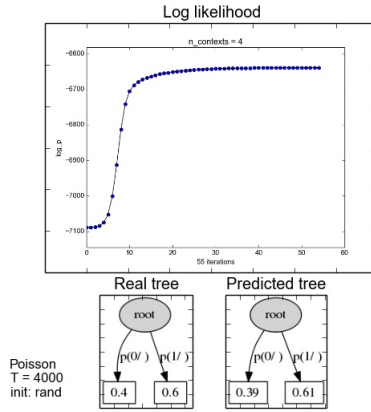


Рис. 3: Смесь двух распределений Пуассона

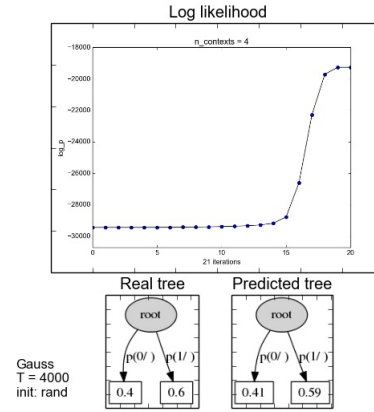


Рис. 4: Смесь двух Гауссиан

2. Более интересный случай. Рисунки 5, 6, 7

Начальное дерево глубины 3, распределения испусканий - двумерное гауссовское, $T = 400$

Параметры алгоритма: $m = 4$, остальные параметры аналогичны параметрам предыдущих примеров

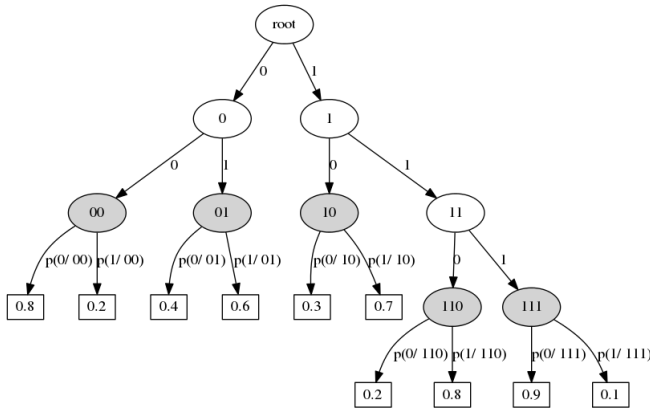


Рис. 5: Реальное дерево

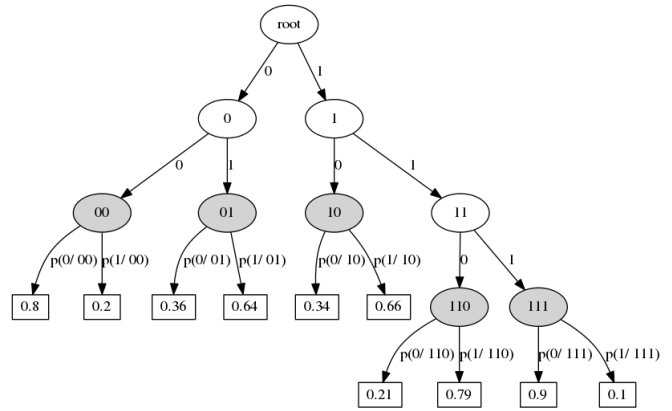


Рис. 6: Предсказанное дерево

Замечание. Взяв в качестве инициализации алгоритм k-means, можно было сойтись быстрее (например, за 6 итераций), но она годится не для всех распределений испусканий

4 Chip-seq, реальные данные

В ходе работы была рассмотрена 21-ая хромосома *кого-то там*.

Данные - просуммированные индикаторы начальных позиций ридов, объединенные в бины размером 10000.

Применение модели VONMM к такой выборке дало следующие результаты:

1. Случай в котором рассматриваются два состояния - наличие/отсутствие сигнала. Рисунок 8
2. Случай в котором рассматриваются три состояния - наличие сигнала/шум/отсутствие сигнала. Рисунок 9

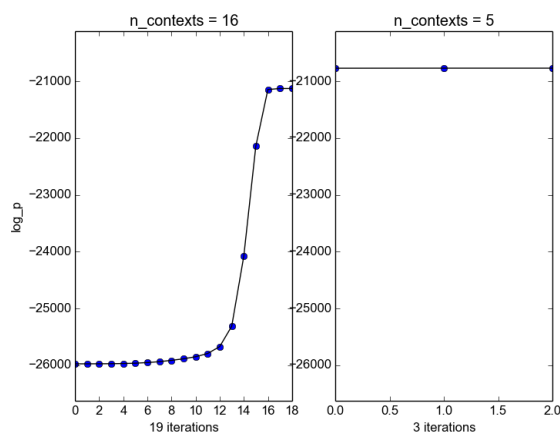


Рис. 7: График роста логарифма правдоподобия

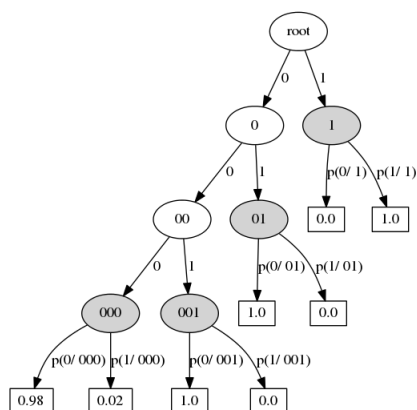


Рис. 8: Дерево с двумя состояниями

5 Оценка модели

6 Заключение

Список литературы

- [1] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.
- [2] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [3] Y Wang, Lizhu Zhou, and Jianhua Feng. Mining complex time-series data by learning Markovian models. In *International Conference on Data Mining*, 2006.

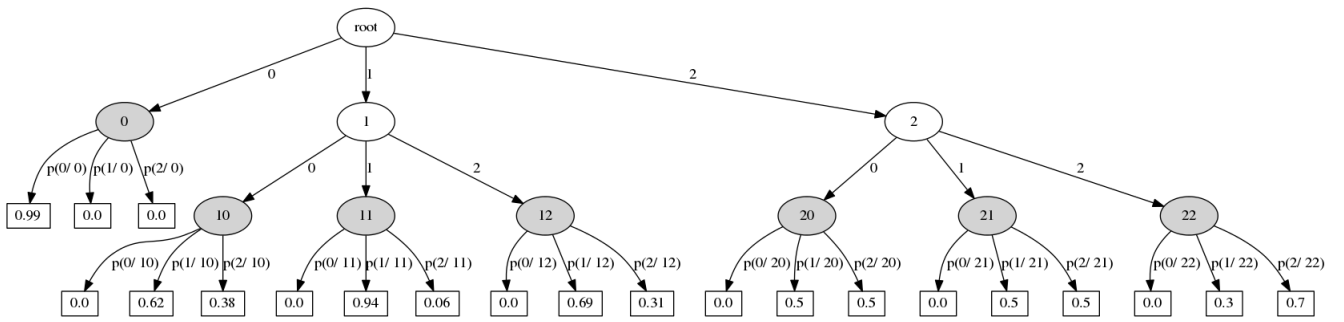


Рис. 9: Дерево с двумя состояниями