Скрытые Марковские модели переменного порядка для анализа данных ChIP-seq

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

3 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 наличие 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 объёме, 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.

```
CAAAAGACAAATAGTGATGTCACCAATCGAGC
--------------------------------
     GACA ATA     GTCA   AATG
    AGAC   TAGTG TGTC
     GACA   AGTG TGTCA   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 испусаниями. 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 искомых моделей, активное приминение 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) завист 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. Последннее 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 $\forall t \in Z$

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

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

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 $P(y_t|y_{t-1}..y_{t-m+1}) = P(x_t|x_t..x_{t-m+1})P(y_t|x_t)$

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

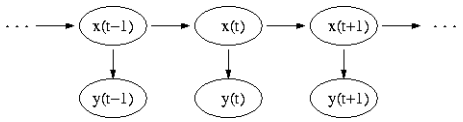

Рис. 2: HMM order 2

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 $Y = (y_1, ...y_T)$ найти параметры модели Λ, которые бы максимизировали правдоподобие при максимально сжатых контекстах [1]

Алгоритм:
Параметры алгоритма: m - максимальная длина контекста, ϵ - барьер для обрезания

1. Инициализация контекстов.
   $C_0 = \{c|c \in S^m\}$

[1] Параметр алгоритма ϵ определяет допустимое отклонение распределений

Начальное распределение переходов произвольное. [2]

2. EM (Expectation–Maximization algorithm).
   Пересчет производится подобно алгоритму Baum-Welch для HMM
   Вводятся дополнительные параметры
   $\alpha_t(c) = P(y_1^t, c(x_t) = c|\Lambda)$
   $\beta_t(c) = P(y_{t+1}^T|c(x_t) = c, \Lambda))$
   $\gamma_t(c) = P(x_t = c|Y, \Lambda)$
   $\xi_t(q, c) = P(c(x_t) = c, x_{t+1} = q|Y, \Lambda)$
   с помощью которых итеративно пересчитываются параметры модели
   $\alpha_0(c) = p(c)b(y_0, c)$, $\alpha_{t+1}(c) = \sum_{q \in S, c' = C(cq)} \alpha_t(c')a(c[0], c')b(y_{t+1}, c)$
   $\beta_T(c) = 1$, $\beta_t(c) = \sum_{q \in S, c' = C(qs)} a(q, c)b(y_{t+1}, c')\beta_{t+1}(c')$
   $p = P(Y|\Lambda) = \sum_{c \in C} \alpha_T(c)$
   $\gamma_t(c) = \frac{\alpha_t(c)\beta_t(c)}{p}$
   $p(c) = \sum_t \gamma_t(c)$
   $\xi_t(q, c) = \frac{\alpha_t(c)a(q,c)b(y_{t+1},c)\beta_{t+1}(qc)}{p}$
   $a(q, c) = \frac{\sum_t \xi_t(q,c)}{p(c)}$
   Пересчет B зависит от принятого семейства моделей испусканий. Производится с помощью γ в точности также как и в алгоритме Baum-Welch.

3. Обрезание дерева.
   Если существует внутренний лист s такой, что $\forall q \in S\ kl(sq, s) < \epsilon$ (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 заканчивет реботу, 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 верятности могут очень близко подходить к нулю, что негативно сказывается на точность расчета. Для избежания этой проблемы все расчеты проходят не с вероятностями, а с логарифмми от них.

Замечание. EM может застревать в локальных максимумах функции правдоподобия.

3 Simulation

Проиллюстрируем работу модели на искусственных данных.
Из дерева 3 была просэмплирована выборка размером $T = 4000$ с Гауссовскими распределениями испусканий.
Обучение проходило, начиная с контекстов длины 4.
Обученное дерево представлено на рисунке 4
Из рисунка 5 видно, что модель сначала 24 итерации обучалась на 16 контекстах длины 4, потом шаг обрезания сократил количество контекство до 3, и модель еще 16 итераций обучалась на них
Здесь инициализация распределений переходов была равномерная. Если ее задавать с помощью алгоритма

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

k-means, то сходимость на первом наборе контекстов будет заметно быстрее. Это видно из рисунка 6. [3] Общее падение правдоподобия незначительно.

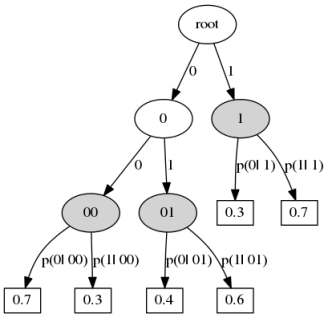

Рис. 3: Real tree

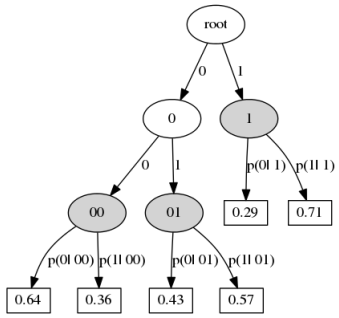

Рис. 4: Predicted tree

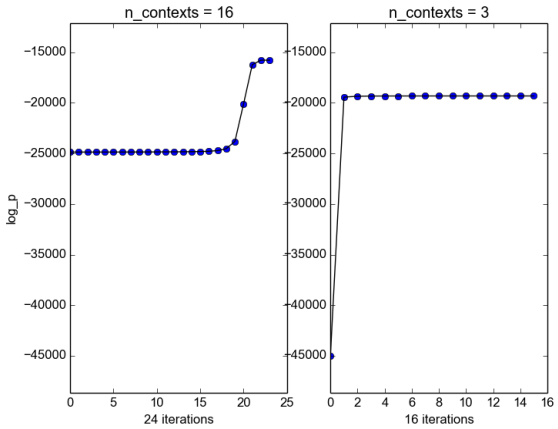

Рис. 5: Log likelihood

4 Оценка модели

5 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.

[3] Здесь можно подумать, что последние итерации лишние. Да, возможно так и есть. Барьер для остановки EM в эксперименте стоял $1e-2$

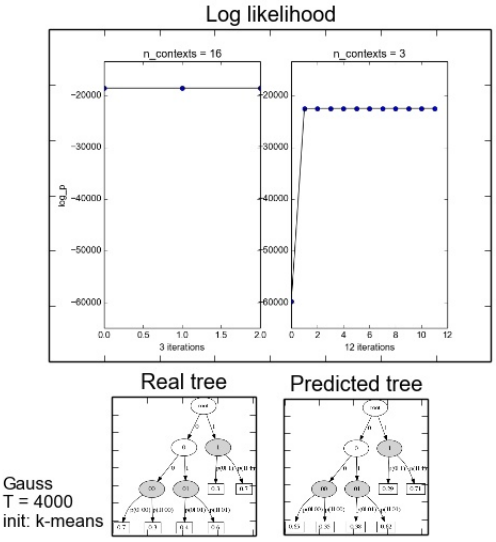

Рис. 6: Общий график для теста с инициализацией алгоритмом k-means