# Скрытые Марковские модели переменного порядка для анализа данных ChIP-seq

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

1 апреля 2015 г.

**Аннотация**

Здесь нужно кратко описать суть работы и результаты.

## 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 некторого 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

Марковский процесс порядка $m$ - это случайный процесс, эволюция которого в каждый момент времени зависит только от $m-1$ предыдущих состояний.

Другими словами, $X = (x_1, x_1, ..., x_T)$ является марковской цепью порядка $m$, если $p(x_t|x_1^{t-1}) = p(x_t|x_{t-m+1}^{t-1})$

Таким образом, каждое следующее состояние определяется контекстом длины $m-1$ и вероятностями перехода из него.

Скрытая марковская модель предполагает, что, помимо всего прочего, каждое состояние задает свою вероятностною модель (например, гауссиан), а в качестве наблюдений поступают не сами состояния, а лишь испускания из них $y_1, ..., y_T$ (2- пример HMM порядка 2).

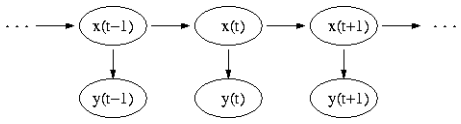


Рис. 2: HMM order 2

Скрытая марковская модель переменного порядка - обобщение HMM разрешающее иметь контекстам разные длины.

Итого, скрытая марковская модель переменного порядка определяется контекстами, вероятностными распределениями переходов и вероястностными распределениями испусканий для каждого состояния.

Перейдем к обучению модели по цепочке наблюдений.

Обычно обучение $n-hmm$ $m$-го порядка сводится к $n^m-hmm$, котрая обучается алгоритмом Баума-Велха (частный случай EM алгоритма, который итеративно максимизирует правдоподобие модели).

Для хранения контекстов используется дерево (бор), в котором вершины являются строками, ребра буквами. Корень - это пустой контекст, а каждый ребенок вершины является уточнением контекста родителя на состояние ребра. Итого, каждая внутренняя вершина имеет ровно n потомков. Листья - главные контексты.

Заметим, что, если дети какой-то вершины имеют одинаковые вероятности переходов, то они контекст вершины не уточняют, и их можно обрезать.

Так же заметим, что достаточно хранить только листья и распределение переходов на них, и, в случае необходимости, пересчитывать его на внутренние вершины

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

где $q$ - состояние, $C(s)$ - все листья (контексты), являющися потомками $s$

Изначально берется полное дерево некоторой фиксированной глубины $m$.

Обучение модели по цепочки наблюдений происходит с помощью чередования EM-алгоритма и обрезания дерева.

EM-часть максимизирует правдоподобие модели алгоритмом Баума-Велха

Обрезание дерева производится нахождением родителей лиистьев, все потомки которых имеют близкие распределения переходов. Все дети такой вершины ликвидируются, а сама вершина становится новым контекстом. Критерием сравнения распределений переходов служит расстояние Кульбака-Лейблера.

Инициализация каждого следующего EM ведется с помощью пересчета распределения переходов на новые контексты.

Изначальная инициализация производится путем построения цепочки состояний алгоритмом $k$-means по наблюдениям (где $k = n$) и частотной оценкой вероятности переходов на ней.

## 3 Оценка модели

## 4 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.