

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
Математико-механический факультет

Кафедра Системного Программирования

Атаманова Анна Михайловна

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

Бакалаврская работа

Допущена к защите.  
Зав. кафедрой:  
д. ф.-м. н., профессор Терехов А. Н.

Научный руководитель:

Рецензент:  
Лебедев С. А.

Санкт-Петербург  
2015

SAINT-PETERSBURG STATE UNIVERSITY  
Mathematics & Mechanics Faculty

Chair of Software Engineering

Anna Atamanova

# Variable-length hidden Markov models for ChIP-seq data analysis

Graduation Thesis

Admitted for defence.

Head of the chair:  
professor Andrey Terekhov

Scientific supervisor:

Reviewer:  
Sergei Lebedev

Saint-Petersburg  
2015

# Оглавление

Введение	4
1. Постановка задачи	6
2. Обзор существующих решений	7
3. Скрытые марковские модели переменного порядка	8
3.1. Марковские модели . . . . .	8
3.2. Скрытые марковские модели . . . . .	8
3.3. Скрытые Марковские модели переменного порядка . . . . .	9
3.4. Обучение модели VLHMM . . . . .	9
3.5. Обучение на нескольких выборках . . . . .	11
4. Simulation	12
5. Chip-seq, реальные данные	13
6. Оценка модели	14
Заключение	15
Список литературы	16

# Введение

## Предметная область

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 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.

## Формулировка проблемы

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.

По этому для дальнейшего анализа результатов эксперимента требуется построение

некоторой вероятностной модели, способной отделять ошибки, а также выявлять зависимости и кратко описывать структуру данных.

# 1. Постановка задачи

Целью данной работы является:

1. Изучение скрытых марковских моделей переменного порядка
2. Реализация скрытой марковской модели переменного порядка
3. Применение модели к результатам биологического эксперимента ChIP-seq

## 2. Обзор существующих решений

Большинство существующих моделей (TODO: ref) для данных хроматин-иммунопреципитата 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 искомых моделей, активное применение находит НММ (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.

### 3. Скрытые марковские модели переменного порядка

#### 3.1. Марковские модели

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

$$P(X_t = x | \text{VOHMM(VariableOrderHiddenMarkovModel)}[3] | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2} \dots X_{-\infty} = x_{-\infty}) = P(X_t = x | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2} \dots X_{t-m+1} = x_{t-m+1})$$

**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  $a(q; x^m) = P(q | x^m)$ .

#### 3.2. Скрытые марковские модели

**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.



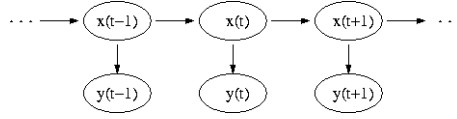


Рис. 2: HMM order 2

### 3.3. Скрытые Марковские модели переменного порядка

**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-Length Hidden Markov Models (VLHMM))*- 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  $b(y, x) = P(y|x)$ .

### 3.4. Обучение модели VLHMM

Задача:

По цепи наблюдений  $Y = (y_1, \dots, y_T)$  найти модель VLHMM с параметрами  $\Lambda$ , с минимальными по длине контекстами и максимальным правдоподобием. <sup>1</sup>

Другими словами, найти

$$\Lambda = \operatorname{argmin}_{\Lambda} \{|\Lambda(C)| \mid \Lambda \in \operatorname{argmax}_{\Lambda} \{P(Y|\Lambda)\}\}$$

Алгоритм:

Параметры алгоритма:  $m$  - максимальная длина контекста,  $\epsilon_{EM}$  - барьер для остановки ЕМ,  $\epsilon_{prune}$  - барьер для обрезания дерева

1. Инициализация контекстов.

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

Начальное распределение переходов произвольное. <sup>2</sup>

<sup>1</sup>Параметр алгоритма  $\epsilon$  определяет допустимое отклонение распределений

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

## 2. EM (Expectation–Maximization algorithm).

Пересчет производится подобно алгоритму Baum-Welch для HMM

### (a) Expectation

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

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

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

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

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

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

### (b) Maximization

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

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

Пересчет  $B$  зависит от принятого семейства моделей испусканий и производится с помощью  $\gamma$  в точности также как и в алгоритме Baum-Welch.

В случае распределения Пуассона  $b(.|c)$   $Poisson(\lambda_c)$

$$\lambda_c = \frac{\sum_t \gamma_t(c)y_t}{\sum_t \gamma_t(c)}$$

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

## 3. Обрезание дерева.

Если существует внутренний лист  $s$  такой, что  $\forall q \in S P(sq)kl(sq, s) < \epsilon_{prune}$  (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:

$c[0]$  - состояние, являющееся началом контекста  $c$ <sup>3</sup>

$c(x_t)$  - 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.** ЕМ может застревать в локальных максимумах функции правдоподобия.

### 3.5. Обучение на нескольких выборках

Пусть дано  $N$  выборок  $\{Y^1 \dots Y^N\}$

ЕМ

#### 1. Expectation

Считаем для каждой выборки  $\alpha^d, \beta^d, \gamma^d, \xi^d$

Общая  $\gamma$  - конкатенация гамм на выборках  $\gamma = [\gamma^1, \dots, \gamma^N]$

$$p = \prod_d p^d$$

#### 2. Maximization

$$a(q; c) = \frac{\sum_d \sum_t \xi_t^d(q; c)}{\sum_t \gamma_t(c)}$$

$$\text{и нормировка } a(q; c) = \frac{a(q; c)}{\sum_q a(q; c)}$$

Параметры испускания считаются по общей  $\gamma$  так же, как в обычной модели.

---

<sup>3</sup>Контекст  $c$  представляем как последовательность состояний  $c[0]c[1]\dots c[l-1]$ , где  $l$  - длина контекста.

## 4. Simulation

План проверки работы VLHMM.

1. Генерация параметров  $\Lambda$  начальной модели VLHMM.
2. Сэмплирование выборки  $Y$  из заданной модели.
3. Обучение новой модели на  $Y$ .  
Получение предсказанных параметров  $\hat{\Lambda}$ .
4. Сравнение параметров  $\Lambda$  и  $\hat{\Lambda}$ .

Ниже приведены два примера теста.

- Смесь.

1. Параметры начальной модели:  $\Lambda = (C, A, B)$   
множество контекстов  $C = \{""\}$ , множество переходов  $A = \{[0.4, 0.6]\}$ , множество распределений испусканий  $B = \{Pois(2.61), Pois(15.34)\}$ .
2. Выборка длиной  $T = 1000$
3. Обучение проходило начиная с полного дерева глубиной  $m = 3$ , случайным распределением переходов.  
Остальные параметры алгоритма: барьер для обрезания  $\epsilon_{prune} = 0.01$ , барьер для остановки ЕМ  $\epsilon_{EM} = 0.15$
4. На рисунке 3 представлен график логарифма правдоподобия по всем итерациям, реальное и предсказанное деревья переходов.  
Параметры предсказанной модели  $\hat{\Lambda} = (\hat{C}, \hat{A}, \hat{B})$   
 $\hat{C} = \{""\}$ ,  $\hat{A} = \{[0.41, .59]\}$ ,  $\hat{B} = \{Pois(2.7), Pois(15.3)\}$ .  
Параметры исходной и предсказанной модели схожи.

- Более интересный случай.

1. Параметры начальной модели:  
множество контекстов и множество переходов представлены на рисунке 4, распределения испусканий -  $Pois(13.1), Pois(36.1)$ .
2. Выборка длиной  $T = 10000$
3. Обучение начиналось с полного дерева глубиной  $m = 4$ , остальные параметры алгоритма те же, что и у примера выше.
4. Предсказанное алгоритмом дерево переходов и график логарифма правдоподобия представлены на рисунках 5,6  
Распределения предсказанных испусканий -  $Pois(13), Pois(35.9)$ .

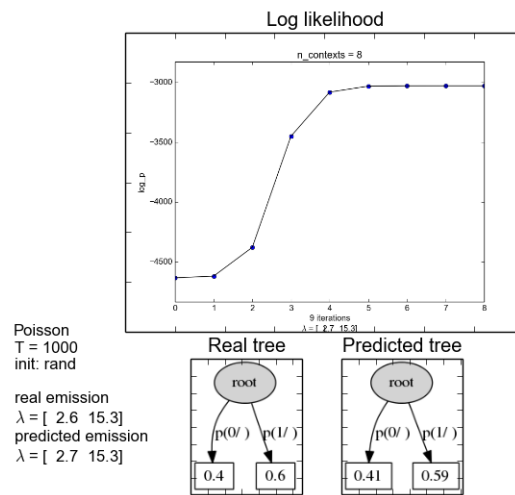


Рис. 3: Результат работы алгоритма VLHMM на смеси Пуассонов

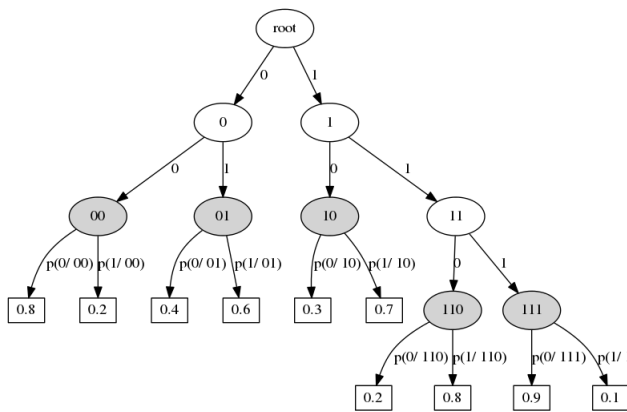


Рис. 4: Реальное дерево

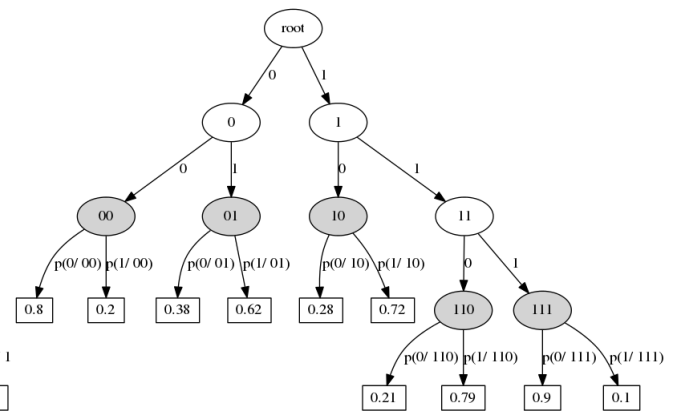


Рис. 5: Предсказанное дерево

## 5. Chip-seq, реальные данные

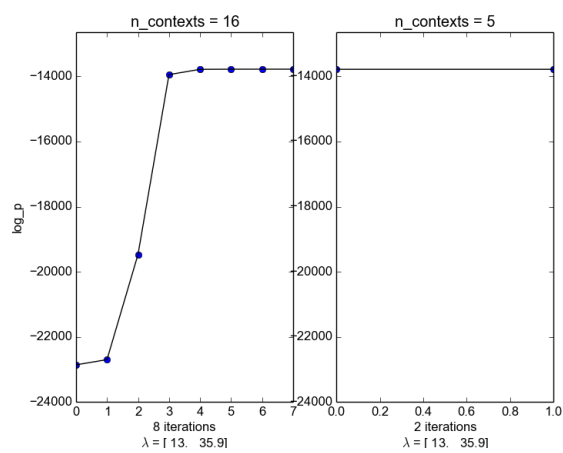


Рис. 6: График роста логарифма правдоподобия

## 6. Оценка модели

## 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.