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

Кафедра информационных систем

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

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

Пусть $\vec{x} = (x_1, .., x_N)$ - вектор прочтений, построенный из какого-то BAM файла. Сопоставим каждому наблюдению некоторую метку-состояние s_n из множества базовых состояний $s = \{1, .., S\}$, истинные значения которых не знаем.

Вероятностная модель позволяет найти наиболее правдоподобную последовательность состояний.

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

1.2. Смесь многомерных распределений Пуассона

Будем моделировать количество прочтений вдоль генома с помощью смеси многомерных распределений Пуасона.

Пусть $\pi = (\pi_1, .., \pi_S)$ - априорные вероятности компонент. $\lambda = (\lambda_1, .., \lambda_S)$ - параметры пуасоновских испусканий для каждой компоненты смеси.

Тогда правдоподобие неполных данных записывается так:

$$p(x; \pi, \lambda) = \prod_{n=1}^{N} \sum_{i=1}^{S} \pi_i \mathcal{P}(x_n; \lambda_i)$$

Интерпретация в качестве порождающей модели следующая: сначала случайным образом выбираем скрытое состояние, применяя распределение π, а затем используем выбранное скрытое состояние для порождения наблюдения.

Чтобы найти параметры модели с помощью оценки максимамльного правдоподобия, следует обратить в максимум совместное правдоподобие наблюдаемых и скрытых переменных. Поэтому Методом максимального 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. Присвоение ожидаемых значений скрытым переменным при условии текущих оценок параметров и нахождение математического ожидания правдоподобия.

$$\mathrm{Q}(\theta|\theta^{\mathrm{old}}) = \mathrm{E}[\log p(x, s; \theta)] \leq \log p(x; \theta)$$

3. Переоценка параметров с учетом обновленных ожидаемых значений скрытых переменных

$$\theta^{\text{new}} = \arg\max_{\theta \in \Theta} \mathrm{Q}(\theta|\theta^{\text{old}})$$

4. Вычислить логарифм правдоподобия и проверить на сходимость

Выведем EM-алгоритм для смеси многомерных Пуасоновских испусканий.
Запишем логарифм функции правдоподобия:

$$\ln p(x; \pi, \lambda) = \sum_{n=1}^{N} \ln \sum_{i=1}^{S} \pi_i \mathcal{P}(x_n; \lambda_i) \tag{1}$$

Шаг E

$$\gamma(s_{ni}) = \frac{\pi_i \mathcal{P}(x_n; \lambda_i)}{\sum_{j=1}^{S} \pi_j \mathcal{P}(x_n; \lambda_j)} \tag{2}$$

Шаг M

Чтобы максимизировать логарифмическое правдоподобие относительно параметров, необходимо взять частные производные и приравнять их к нулю.

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_i} \mathrm{E}[\log p(x; s, \theta)] \\
&\qquad = \frac{\partial}{\partial \lambda_i} \sum_{n=1}^{N} \sum_{s=1}^{S} \mathrm{E}[s_{ni}] \{\log \pi_i + \log \mathcal{P}(x_n|\theta)\} \\
&\qquad = \sum_{n=1}^{N} \mathrm{E}[s_{ni}] \frac{\partial}{\partial \lambda_i} \log \mathcal{P}(x_n|\theta) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{nk}) \frac{\partial}{\partial \lambda_i} \log \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{nk}) \frac{\partial}{\partial \lambda_i} (\log {\lambda_i}^{x_n} + \log e^{-\lambda_k} - \log x_n!) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{ni}) \frac{\partial}{\partial \lambda_i} (x_n \log \lambda_i - \lambda_i - \log x_n!) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{ni}) (\frac{x_n}{\lambda_S} - 1) \\
&\qquad = 0
\end{aligned}$$

получаем

$$\lambda_i^* = \frac{1}{N_i} \sum_{n=1}^{N} \gamma(s_{ni}) x_n \tag{3}$$

где

$$N_i = \sum_{n=1}^{N} \gamma(s_n i) \tag{4}$$

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

$$\begin{aligned}
&\frac{\partial}{\partial \pi_i}(\mathrm{E}[\log p(x; s, \theta)] + \lambda(\sum_{j=1}^{S} \pi_j - 1)) \\
&\qquad = \frac{\partial}{\partial \pi_i} \sum_{n=1}^{N} \sum_{i=1}^{S} E[s_{ni}]\{\log \pi_i + \log \mathcal{P}(x_n; \theta)\} + \frac{\partial}{\partial \pi_i} \lambda(\sum_{j=1}^{S} \pi_j - 1) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{ni}) \frac{\partial}{\partial \pi_i}\{\log \pi_i + \log \mathcal{P}(x_n|\theta)\} + \frac{\partial}{\partial \pi_i} \lambda(\sum_{j=1}^{S} \pi_j - 1) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{ni}) \frac{\partial}{\partial \pi_i}\{\log \pi_i\} + \frac{\partial}{\partial \pi_i} \lambda(\sum_{j=1}^{S} \pi_j - 1) \\
&\qquad = \sum_{n=1}^{N} \gamma(s_{ni}) \frac{1}{\pi_i} + \lambda \\
&\qquad = 0
\end{aligned}$$

Домножив на π_k, получаем

$$\sum_{n=1}^{N} \gamma(s_{ni}) + \pi_i \lambda = 0 \tag{5}$$

Просуммируем вдоль i

$$\sum_{n=1}^{N} \sum_{i=1}^{S} \gamma(s_{ni}) + \sum_{i=1}^{S} \pi_i \lambda = 0 \tag{6}$$

Применяя

$$\sum_{i=1}^{S} \gamma(s_{ni}) = 1$$

и

$$\sum_{i=1}^{S} \pi_i = 1$$

Из выражения (7) получаем

$$N + \lambda = 0$$

$$\lambda = -N$$

Далее, подставляя найденное λ в выражение (6), получаем

$$\pi_i = -\frac{\sum_{n=1}^{N} \gamma(s_{ni})}{\lambda} = \frac{N_i}{N}$$

Наконец,

$$\pi_i^* = \frac{N_i}{N}$$

Для нахождения наиболее правдоподобной последовательности скрытых состояний достаточно выбрать состояния с наибольшей апостериорной вероятностью для каждого наблюдения:

$$s_n = \arg\max_{i \in 1,..,S^N} \gamma(s_{ni})$$

Примечание. Чтобы успешно обучить наши данные, следует правильно проинициализировать начальные значения входных параметров модели. Для этой задачи достаточно использовать алгоритм кластеризации KMeans++[2], 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}. В результате, начальные значений параметров Пуасоновского распределения можно принять за значения кластеров.

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.

$z_{(n+1),i}$ зависит от $z_{n,i}$

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.

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

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

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

Состояние S_n с одинаковым базовыми состояниями $(+,+), (-,-), (\text{null}, \text{null})$ означает, что данные сравниваемых образцов наблюдения x_n похожи между собой.

Чтобы задать модель, нужно определить распределения испусканий, то есть $p(x_n|s_{ni} = 1)$, где i — индекс состояния из S^2, а n — индекс наблюдения. Будем считать, что x_n — это наблюдение из многомерного распределения Пуассона с независимыми компонентами, то есть:

$$p(x_n|s_{ni}) = \prod_{d=1}^{2} p(x_{nd}|\lambda_{id})$$

На данном этапе для каждого состояния и каждого образца есть свой параметр распределения Пуассона, что является неверной параметизацией. Рассмотрим два состояния $i = (+,+)$ и $j = (+,-)$ и выпишем для них функцию вероятности распределения Пуассона:

$$p(x_n|s_{ni}) = \prod_{d=1}^{2} p(x_{nd}|\lambda_{id}) = p(x_{n1}|\lambda_{i1})p(x_{n2}|\lambda_{i2})$$

$$p(x_n|s_{nj}) = \prod_{d=1}^{2} p(x_{nd}|\lambda_{jd}) = p(x_{n1}|\lambda_{j1})p(x_{n2}|\lambda_{j2})$$

Первый множитель в обоих выражениях соответствует наблюдению, порождённому базовым состоянием (+) в первом образце. Логично положить, что $\lambda_{i1} = \lambda_{j1}$, потому что в обратном случае испусканиям для одного и того же базового состояния будут соответствовать разные параметры распределения Пуассона. Таким образом, различных лямбд у нас не $2 * |S^2| = 8$, а $2 * |S| = 4$.

Для реализации удобно представлять $\vec{\lambda}$ в виде вектора размерности $2 * |S|$, а для состояния i из S^2 использовать матрицу трансляции D.

Матрица трансляции - это двухмерная матрица размерности $2 \times |S^2|$, где D_{di} - индекс лямбды из вектора:

$$\vec{\lambda} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4 \lambda_5, \lambda_6\}$$

		(null, null)	(+, +)	(−, −)	(null, +)	(null, −)	(−, null)	(+, null)	(+, −)	(−, +)
D	d = 1	1	3	2	1	1	2	3	3	2
	d = 2	4	6	5	6	5	4	4	5	6

Рис. 1: Матрица трансляции

Перепишем функцию вероятности распределения Пуассона в терминах D.

$$p(x_n; s_{ni}) = \prod_{d=1}^{2} \prod_{s=1}^{|S|} p(x_{nd}; \lambda_s)^{I[D_{di}=s]} \tag{7}$$

Функция правдоподобия определяется как

$$p(x, s; \theta) = p(s_1; \pi) \left[\prod_{n=2}^{N} p(s_n; s_{n-1}, A) \right] \prod_{m=1}^{N} p(x_n; s_m, \theta) \tag{8}$$

Подставив определение функции вероятности распределения (7) в функцию правдоподобия для СММ (8) можно убедиться, что М-шаг для вектора лямбд:

$$\lambda_s = \frac{\sum_{d=1}^{2} \sum_{i=1}^{|S|} I[D_{di} = s] \sum_{n=1}^{N} \gamma_{ni} x_{nd}}{\sum_{d=1}^{2} \sum_{i=1}^{|S|} I[D_{di} = s] \sum_{n=1}^{N} \gamma_{ni}} \tag{9}$$

Пусть

$\pi = (\pi_1, .., \pi_S^2)$ - априорные вероятности состояний.
$\mathcal{A}$ - матрица вероятностей перехода между состояниями.
$\lambda = (\lambda_1, .., \lambda_S)$ - параметры многомерного распределения Пуасона.

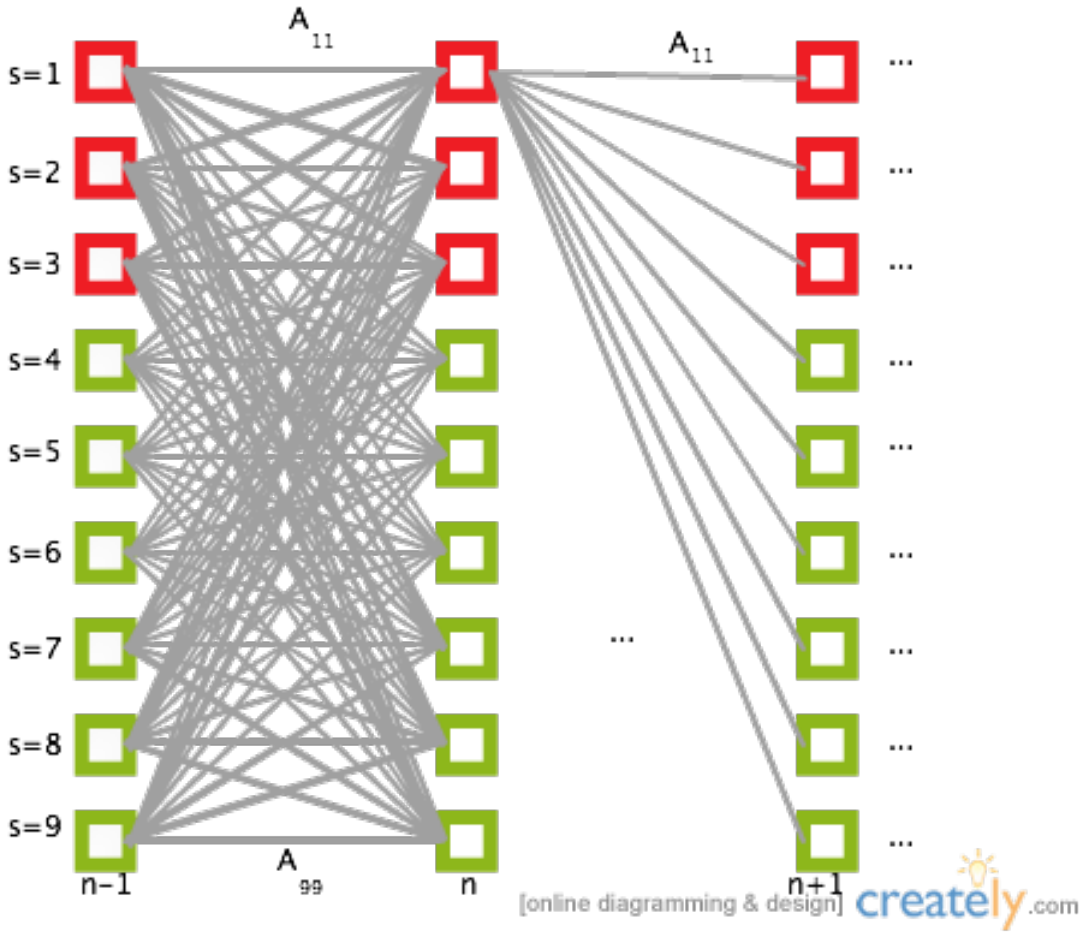

Рис. 2: решетка состояний

Решетка представляет собой диаграмму переходов между скрытыми состояниями модели. Красным светом выделены состояния, в котором базовые состояния похожи, то есть $(+,+),(-,-),(\text{null},\text{null})$

Шаг E

$$\gamma_{ni} = \frac{\alpha_{ni}\beta_{ni}}{\sum_{j=1}^{S}\alpha_{nj}\beta_{nj}} \tag{10}$$

$$\xi_{nij} = \frac{\alpha_{(n-1),i}\mathcal{A}_{ij}\mathcal{P}(x_n;\lambda_j)\beta_{nj}}{\sum_{i'=1}^{S}\sum_{j'=1}^{S}\alpha_{(n-1),i'}A_{i'j'}\mathcal{P}(x_n;\lambda_{j'})\beta_{nj'}} \tag{11}$$

Где,

$\alpha_{ni} = p(s_{ni} = 1, x_1, x_2, ..x_n; \theta)$

$\beta_{ni} = p(x_{n+1}, .., x_N; s_{ni} = 1, \theta)$

Вычисления α и β производится с помощью алгоритма прямого-обратного хода:

$$\alpha_{1i} = \pi_i \mathcal{P}$$

$$\beta_{Ni} = 1$$

$$\alpha_{ni} = \mathcal{P}(x_n; \lambda_i) \sum_{j=1}^{S} \alpha_{(n-1),j} A_{ji}$$

$$\beta_{ni} = \sum_{j=1}^{S} A_{ij} \mathcal{P}(x_{n+1}; \lambda_j) \beta_{(n+1),j}$$

Шаг M

Новые значения параметров модели вычисляются так:

$$\pi_i^* = \gamma(s_{1i}) \tag{12}$$

$$A_{ij}^* = \frac{\sum_{n=2}^{N} \xi_{nij}}{\sum_{j'=1}^{S} \sum_{n=2}^{N} \xi_{nij} \xi_{nij'}} \tag{13}$$

$$\lambda_i^* = \frac{\sum_{n=1}^{N} \gamma(s_{ni}) x_n}{\sum_{n=1}^{N} \gamma(s_{ni})} \tag{14}$$

Замечание. В алгоритме прямого-обратного хода может быть underflow - это значит, что не хватает точности чисел с плавающей точкой. Поэтому удобно проводить вычисления в логарифмах.

1.4. Выбор модели

Для задачи сравнения двух образцов была выбрана Скрытая Марковская Модель. В частном случае, для анализа одного биологического образца СММ более правдоподобнее, чем смесь Пуасоновских испусканий. Это значит, что модель, в которой предполагается зависимость между состояниями соседних наблюдений, более правдоподобная.

1.5. Актуальности разработки

Существуют инструменты для анализа одного эксперимента с использованием FAIRE-seq:

- ChromHMM
- ZINBA
- Fseq
- ChIPOTle Peak Finder
- и другие ...

Примечание. Косвенный аналог ChromHMM моделирует многомерную последовательность из $\{0, 1\}$. Данный инструмент, как и другие аналоги, используется для анализа одного FAIRE-seq образца. Стало быть, для задачи сравнения она не подходит.

2. Оценка модели

2.1. Оценка качества работы модели

При оценке качетсва работы модели мы хотели бы получать число неверных предсказаний FDR[3] среди всех предсказаний модели.

Для того чтобы ввести FDR, сформулируем гипотезы, которые будем проверять с помощью модели.

Для данных FAIRE-seq возможны две гипотезы:

- H_0 - разницы между состояниями эксперементальных данных в одном наблюдении нет
- H_1 - разница есть

Рассмотрим некоторое наблюдение с индексом i в нашей выборке. Как понять отвергаем ли мы или принимаем для него нулевую гипотезу?

Воспользуемся для этого апостериорными вероятностями:

- $P(\text{отличий в i нет}; x, \theta) := p_0$
- $P(\text{отличия в i есть}; x, \theta) := p_1$

Нулевая гипотеза отвергается если $p_0 < p_1$ и не отвергается в обратном случае.

Примечание. Вспомним, что $p_0 + p_1 = 1$, поэтому наш критерий можно записать как:

$$p_0 <= 0.5$$

Таким образом, применив сформулированный выше критерий ко всем бинам $i = 1, ..., N$ мы получим N результатов. FDR $= a \in [0, 1]$ означает, что среди N результатов $a * N$ — не верны.

В общем виде FDR записывается так:

$\text{FDR} = \text{E}[\text{FP} / (\text{TP} + \text{FP})]$,

где FP — количество неверно отвергнутых нулевых гипотез, TP — количество верно отвергнутых нулевых гипотез.

Для удобства будем использовать следующую разновидность FDR:

$$\text{mFDR} = \text{E}[\text{FP}] / \text{E}[\text{TP} + \text{FP}] \tag{15}$$

2.2. Контроль FDR

Также мы хотим выдавать вектор предсказаний, в котором гарантированно не более чем фиксированное число ошибок.

Заключение

Список литературы

[1] Bishop Christopher. Pattern Recognition and Machine Learning (Information Science and Statistics). — Springer, 2007. — Ozon Books : http://www.ozon.ru/context/detail/id/2978313/.

[2] Murphy Kevin P. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning series. — MIT Press, 2012. — Amazon Books : http://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020.

[3] Wikipedia. False discovery rate // Wikipedia, the free encyclopedia. — 2012. — URL: `http://en.wikipedia.org/wiki/False_discovery_rate` (online; accessed: 15.05.2015).