Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Абрамова Анастасия Николаевна

Метод Монте-Карло по схеме марковской цепи для оценки вероятности редких событий в задачах биоинформатики

Выпускная квалификационная работа

Научный руководитель:
к. ф.-м. н., доцент А. И. Коробейников
Рецензент:
Разработчик ПО А. Л. Тарасов

Санкт-Петербург
2017

Saint Petersburg State University
Applied Mathematics and Computer Science
Computational Stochastics and Statistical Models

Abramova Anastasiia

Monte Carlo method by Markov chain for rare event probability estimation in bioinformatics problems

Graduation Project

Scientific Supervisor:
Associate Professor A. I. Korobeynikov

Reviewer:
Software Developer A. L. Tarasov

Saint Petersburg
2017

Оглавление

Введение

Пептидные соединения (пептиды) — это вещества, молекулы которых содержат два и более остатков аминокислот, соединенных в цепь пептидными связями. Существуют живые организмы, способные продуцировать природные пептидные соединения, оказывающие сильное подавляющее действие на рост и размножение бактерий – натуральные антибиотики. В связи с проблемой резистентности существующих антибиотиков к грамм-положительным бактериям и недавними успехами, связанными с открытием теиксобактина – антибиотика, активного в отношении данных бактерий, задача идентификации природных пептидных соединений вновь становится задачей высокой важности в сфере протеомики. Она заключается в следующем: для исследуемого образца в базе необходимо найти наиболее близкое по структуре к нему пептидное соединение. В связи с тем, что схожесть структур во многих случаях влечет за собой схожесть свойств соединений, решение данной задачи помогает в исследовании новых образцов.

Самым распространенным инструментом для идентификации пептидных соединений является масс-спектрометрия: исследуемый образец режется на части при помощи химических реакций, после чего измеряется масса каждого полученного фрагмента, и в дальнейшем исследуется полученный массив масс, который называют *спектром* [1]. Тогда задача идентификации пептидного соединения сводится к тому, чтобы найти наиболее похожий спектр теоретического пептида из базы на полученный по образцу экспериментальный спектр и оценить эту похожесть.

Для случая пептидных соединений линейной структуры, существует метод MS-GF+ [2], решающий задачу, используя комбинаторные методы. Однако в случае природных пептидных соединений, зачастую имеющих сложную нелинейную структуру, метод MS-GF+ не может быть применен, и задача решается только вероятностными подходами.

В [3] был предложен новаторский подход под названием MS-DPR (Mass Spectrometry Direct Probality Distribution), работающий при определенных ограничениях, наложенных на структуру химических соединений, и основанный на методе Монте-Карло на марковских цепях. Тем не менее, данный метод не дает никакой информации о точности полученных оценок, а также сильно зависит от длины марковской цепи, строящейся в процессе работы метода.

В данной работе был предложен алгоритм, который также основывается методе Монте-Карло на марковских цепях: доказано что оценки по данному методу являются асимптотически несмещенными. В добавление к этому предложен критерий остановки построения марковской цепи, который позволяет получить длину траектории, достаточную для получения оценки заданной точности.

Глава 1

Постановка задачи

Будем считать, что мера схожести экспериментального спектра $Spectrum$ на теоретический спектр известного пептида P из базы вычисляется при помощи некоторой функции $Scoring(Spectrum, P)$. Ясно, что спектр, вычисленный по пептидному соединению P зависит от его фрагментации в масс-спектрометре, а значит, при различных схемах фрагментации P значения $Scoring(Spectrum, P)$ будут различны. Опишем модель фрагментации пептидных соединений, предложенную в [3].

1.1. Модель фрагментации пептидных соединений

Фиксированную химическую структуру пептидного соединения P будем представлять в виде слабо-связного ориентированного графа $G = (V, E)$, вершинами которого будут аминокислоты, а ребрами — химические связи между ними. Множество вершин графа G будем обозначать за $V(G)$, а множество ребер за $E(G)$. Для подграфа $G' \subset G$ будем определять его массу как

$$Mass(G') = \sum_{v \in V(G')} Mass(v), \tag{1.1}$$

где $Mass(v)$ — масса вершины v. Соответственно, масса всего графа $Mass(G)$ будет равняться сумме масс всех аминокислот, входящих в данное соединение.

Теперь, ребро в графе будем называть *мостом*, если при его удалении граф становится несвязным. Пару ребер будем называть *tc-ребрами* (от two-cut), если оба ребра не являются *мостами*, но при этом их совместное удаление делает граф несвязным.

Обозначим $\mathcal{C}_b$ — множество всех мостов графа G, $\mathcal{C}_{tc}$ — множество всех его *tc-ребер*. Их объединение $\mathcal{C} = \mathcal{C}_b \cup \mathcal{C}_{tc}$ будем называть *множеством разрезов*. Все одноэлементные подмножества множества $\mathcal{C}$ (то есть мосты и пары *tc-ребер*) будем называть соответственно *разрезами*.

Теперь каждому разрезу $C' \in \mathcal{C}$ можно сопоставить такой подграф $G' = G(C')$, что $E(G') = E(G) \setminus C'$, а $V(G') = V(G)$. По определению разрезов, полученный подграф G' распадается на две компоненты связности $G_1(C')$ и $G_2(C')$, которым по формуле (1.1) можно сопоставить их массы $m_b = Mass(G_1(C'))$ и $m_y = Mass(G_2(C'))$.

Таким образом по графу G и множеству $\mathcal{C}$ можно сформировать вектора $\overline{m}_b = (m_b^{(1)}, \ldots, m_b^{(|\mathcal{C}|)})$ и $\overline{m}_y = (m_y^{(1)}, \ldots, m_y^{(|\mathcal{C}|)})$, где $m_b^{(i)} = m(G_1(C^{(i)}))$, $m_y^{(i)} = m(G_2(C^{(i)}))$, а $C^{(i)} \in \mathcal{C}$.

Вектор $\overline{m}_b$ в дальнейшем будем называть *теоретическим спектром* и положим $S = \overline{m}_b$. Заметим, что по $Mass(G)$ и вектору $\overline{m}_b$ можно вычислить вектор $\overline{m}_y$.

Теоретический спектр может быть представлен иным образом. Сформируем матрицу $H = \{h_{ij}\}$ размера $|\mathcal{C}| \times |V(G)|$ по следующему правилу:

$$h_{ij} = \begin{cases} 1, & \text{если } j \in V(G_1(C^{(i)})) \\ 0, & \text{иначе} \end{cases}.$$

Тогда, теоретический спектр имеет вид $S = H\mu$, где μ — вектор масс аминокислот (вершин графа G).

1.2. Вероятностная модель спектра пептида

Пусть $Spectrum$ — спектр исследуемого образца, вычисленный экспериментально, P — пептидное соединение с известной структурой. Граф G — граф химической структуры пептида P.

Обозначим за $\mathcal{M}$ множество векторов, удовлетворяющих следующему условию:

$$\mathcal{M} = \left\{ \left(\mu_1, \ldots, \mu_{|V(G)|}\right), \mid \mu_i > 0, \sum_{i=1}^{|V(G|} \mu_i = Mass(G) \right\}. \tag{1.2}$$

Данное множество представляет собой всевозможные пептидные соединения, имеющие одинаковую массу и химическую структуру. Отсюда, нашей задачей является оценка вероятности

$$p = \mathbb{P}(Scoring(Spectrum, H\mu) \geq t), \tag{1.3}$$

где μ — случайная величина, равномерно распределенная на множестве $\mathcal{M}$, t — некоторый заранее фиксированный порог. В дальнейшем будем обозначать:

$$Score(\mu) := Scoring(Spectrum, H\mu).$$

Глава 2

Метод Монте-Карло на марковских цепях

2.1. Метод Монте-Карло и метод существенной выборки

Стандартным методом для вычисления оценок вероятности вида (1.3) является метод Монте-Карло. Рассмотрим случайную величину ξ, определенную на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$ с распределением $\mathcal{P}$ и соответствующей плотностью f относительно меры ν. Обозначим $p = \mathbb{P}(\xi \in A)$.

Определение 1. *Пусть $x_1, \ldots, x_N$ — независимые одинаково распределенные случайные величины с распределением $\mathcal{P}$. Оценкой величины p по методу Монте-Карло называется*

$$\widehat{p}_{MC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{x_i \in A\}.$$

Дисперсия такой оценки $\mathbb{D}(\widehat{p}_{MC}) = \frac{p(1-p)}{N}$ сходится к 0 при $N \to \infty$, однако относительная ошибка

$$\mathrm{RE}(\widehat{p}_{\mathrm{MC}}) = \frac{\mathbb{D}(\widehat{p}_{\mathrm{MC}})}{p^2} = \frac{p(1-p)}{Np^2} = \frac{1}{Np} - \frac{1}{N} \longrightarrow \infty, \tag{2.1}$$

при $p \to 0$. Из формулы (2.1) следует, например, что при $\widehat{p}_{\mathrm{MC}} \approx 10^{-10}$ для получения $\mathrm{RE}(\widehat{p}_{\mathrm{MC}}) \approx 10^{-2}$ нужно иметь последовательность x_i длиной в 10^{14}. Поэтому обычный метод Монте-Карло является очень трудоемким для задачи оценки редких событий.

Одним из способов уменьшения относительной ошибки является метод существенной выборки. В предложенных выше обозначениях положим $\mathcal{Q}$ — некоторое распределение, определенное на $(\Omega, \mathcal{F})$ с соответствующей функцией распределения Q и плотностью q относительно меры ν. Предположим также, что существует производная Радона-Никодима $d\mathcal{P}/d\mathcal{Q}$, которая в случае существования плотностей f и q имеет вид

$$d\mathcal{P}/d\mathcal{Q} = \begin{cases} \frac{f(x)}{q(x)}, & q(x) \neq 0 \\ 0, & q(x) = 0 \end{cases}.$$

Определение 2. *Пусть $x_1, \ldots, x_N$ — одинаково распределенные случайные величины с распределением $\mathcal{Q}$. Оценкой Монте-Карло по методу существенной выборки для ве-*

роятности p будем называть величину

$$\widehat{p}_{IS} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(x_i)}{q(x_i)}\mathbb{1}_{\{x_i\in A\}}.$$

Замечание 1. *Оценка является состоятельной, так как по закону больших чисел,*

$$\widehat{p}_{IS} = \frac{1}{m}\sum_{i=1}^{m}\frac{f(x_i)}{q(x_i)}\mathbb{1}_{\{x_i\in A\}} \xrightarrow[m\to\infty]{\mathbb{P}} \int\limits_A \frac{f(x)}{q(x)}\,dQ = \int\limits_A f(x)\,d\nu = \mathbb{E}_F\mathbb{1}_{\{X_n\in A\}} = p.$$

Предположим теперь, что плотность q имеет конкретный вид, а именно зависит от плотности f и весовой функции w:

$$q(x) = cw(x)f(x), \tag{2.2}$$

где $c > 0$ — нормирующая константа. Тогда оценка по методу существенной выборки имеет следующий вид:

$$\widehat{p}_{IS} = \frac{\sum\limits_{n=1}^{N}\mathbb{1}_{\{x_n\in A\}}/w(x_n)}{\sum\limits_{n=1}^{N}1/w(x_n)}. \tag{2.3}$$

То есть при таком виде моделирующей плотности q получаем, что оценка по методу существенной выборки не будет зависеть от вида плотностей f и q, но будет зависеть от вида весовой функции w.

2.2. Метод Метрополиса-Гастингса

Чтобы построить оценку по методу существенной выборки с моделирующей плотностью вида (2.2), нужно уметь моделировать случайные величины с распределением $\mathcal{Q}$. Данная задача является нетривиальной из-за того, что нормирующая константа c неизвестна и веса могут быть произвольными. Вместо этого мы будем строить марковскую цепь $(x_n)_{n\geq 1}$ со стационарным распределением, равным $\mathcal{Q}$, при помощи метода Метрополиса-Гастингса [4]: он удобен тем, что не требует вычисления константы нор-

мировки c. Ниже представлено формальное описание алгоритма.

Algorithm 1: Алгоритм Метрополиса-Гастингса

Input: Текущее состояние x_i, переходная плотность $\gamma(\cdot \mid x)$

Output: Следующее состояние x_{i+1}

1 Моделируется случайная величина y с условным распределением $\gamma(\cdot|x = x_i)$

2 Моделируется случайная величина r, равномерно распределенная на промежутке $[0; 1]$

3 Вычисляется *доверительная вероятность* $\alpha = \min\left(\frac{q(y)\gamma(x_i|y)}{q(x_i)\gamma(y|x_i)}, 1\right)$

4 Если $r < \alpha$, то $x_{i+1} \leftarrow y$, иначе $x_{i+1} \leftarrow x_i$

Замечание 2. *Данный алгоритм удовлетворяет уравнению детального баланса [5]. Стационарным распределением полученной в результате работы алгоритма марковской цепи является* $\mathcal{Q}$.

2.3. Описание алгоритма построения $\widehat{p}_{IS}$

Таким образом, общая идея алгоритма сводится к следующему:

1. Определенным образом оцениваются значения весовой функции w.
2. При помощи метода Метрополиса-Гастингса моделируется марковская цепь со стационарным распределением $\mathcal{Q}$ с плотностью q вида (2.2).
3. Вычисляется оценка интересующей нас вероятности по формуле (2.3).

В следующей главе описаны особенности данного алгоритма для задачи оценки вероятности (1.3): представлен способ оценки значений весовой функции w, предложен вид переходной плотности для метода Метрополиса-Гастингса.

Глава 3

Особенности реализации построения $\widehat{p}_{IS}$

Итак, используем описанный подход для оценки вероятности (1.3).

1. В качестве плотности f будем рассматривать плотность равномерного распределения на множестве $\mathcal{M}$.

2. Положим $q(\mu) = cw(Score(\mu))f(\mu)$, то есть весовая функция w зависит от значений функции $Score$.

3. Заметим, что если задать значения весовой функции как

$$w(S) \approx \frac{1}{\mathbb{P}(Score(\mu) = S)}, \tag{3.1}$$

в предположении, что функция $Score$ принимает значение на дискретном множестве, то при моделировании векторов масс μ с такой плотностью q, распределение значений $Score(\mu)$ будет близко к равномерному на множестве $[S_{min}, S_{max}]$ всевозможных значений $Score$.

То есть выбор весов таким образом уменьшает относительную ошибку оценки $\mathrm{RE}(\widehat{p}_{\mathrm{IS}})$.

3.1. Оценка весовой функции по методу Ванга-Ландау

Для построения оценки такой весовой функции $\widehat{w}$ будем использовать алгоритм Ванга-Ландау [6]. Данный алгоритм является адаптивной модификацией метода Метрополиса-Гастингса и позволяет оценить требуемые веса одновременно с моделированием траектории цепи. Однако получающаяся в результате работы алгоритма цепь, вообще говоря, не является марковской, так как зависит от всей «истории» цепи. В связи с этим данный алгоритм используется только для оценки значений весовой функции w.

Ниже приведено формальное описание этого алгоритма.

Algorithm 2: Алгоритм Ванга-Ландау

Input: Количество значений функции $Score$, максимальное количество шагов для метода Метрополиса-Гастингса на каждой итерации M_{max}, диапазон $[LC_{min}; LC_{max}]$, S_{min}, S_{max}

Output: Оценки $\ln(w)$

$LGW[S_{min}, \ldots, S_{max}] \leftarrow 0$, $Hist[S_{min}, \ldots, S_{max}] \leftarrow 0$

$LC \leftarrow LC_{max}$, $M = 0$ – счетчик шагов

while $LC > LC_{min}$ **do**

  Моделируется текущее значение $\mu \in \mathcal{M}$, вычисляется $Score(\mu)$.

  while $M < M_{max}$ *или* $Hist$ *не соответствует «критерию равномерности»* **do**

    Шаг метода Метрополиса-Гастингса с $q(\mu) = cw(Score(\mu))f(\mu)$, где $w(Score(\mu)) = LGW[Score(\mu) - S_{min} + 1]$. В результате этого шага получается новое состояние цепи $\tilde{\mu}$.

    $LGW[Score(\tilde{\mu})] \leftarrow LGW[Score(\tilde{\mu})] - LC$, $Hist[Score(\tilde{\mu})] \leftarrow Hist[Score(\tilde{\mu})] + 1$.

  end

  $LC \leftarrow LC/2$ и обнулить $Hist$, M.

end

Замечание 3. *Будем считать, что гистограмма удовлетворяет «критерию равномерности», если для всех* $i \in \{S_{min}, \ldots, S_{max}\}$ *выполняется* $Hist[i] > 20$ *и* $Hist[i]$ *больше, чем* 70% *и меньше, чем* 130% *от среднего значения по гистограмме.*

3.2. Выбор переходной плотности γ

В методе Метрополиса-Гастингса кандидата на следующее значение марковской цепи будем моделировать следующим образом:
пусть $\mu = (\mu_1, \ldots, \mu_{|V(G)|}) \in \mathcal{M}$ — текущее состояние. Пронумеруем ребра графа G и смоделируем следующее состояние $\tilde{\mu}$ таким образом:

1. Промоделируем равномерно распределенный индекс i на $\{1, \ldots, |E(G)|\}$ и выберем ребро $e_i \in E(G)$. Положим $v_1 = beg(e_i)$, $v_2 = end(e_i)$ как начало и конец ребра e_i.

2. Положим $\mu_{beg} \leftarrow m(v_1)$, $\mu_{end} \leftarrow m(v_2)$.

3. Смоделируем равномерно случайную величину δ на $[-\mu_{beg}; \mu_{end}]$.

4. Положим $\tilde{\mu}_{beg} \leftarrow \mu_{beg} + \delta$, $\tilde{\mu}_{end} \leftarrow \mu_{end} - \delta$, и $\tilde{\mu}_i \leftarrow \mu_i$ для остальных i.

Замечание 4. *Можно проверить, что для данной переходной плотности выполняется уравнение детального баланса [5].*

Замечание 5. *При таком выборе переходной плотности γ доверительная вероятность α в методое Метрополиса-Гастингса для моделирования марковской цепи со стационарным распределением $\mathcal{Q}$ и плотностью $q(\mu) = cw(Score(\mu))f(\mu)$ будет иметь вид*

$$\alpha = \min\left(\frac{w(Score(\mu))}{w(Score(\tilde{\mu}))}, 1\right).$$

Иными словами, на каждом шаге мы отдаем предпочтение тому вектору масс, чей весовой коэффициент значения Score больше.

Глава 4

Оценка дисперсии и критерий остановки

4.1. Способы вычисления дисперсии

Теперь по описанному выше алгоритму мы можем получить оценку $\widehat{p}_{IS}$ для вероятности (1.3). Однако необходимо уметь вычислять точность $\widehat{p}_{IS}$.

В предположении, что выполняется центральная предельная теорема [7] для марковской цепи $\mu_1, \ldots, \mu_N$

$$\sqrt{N}(\widehat{p}_{IS} - p) \xrightarrow[N\to+\infty]{\mathcal{D}} \mathcal{N}(0, \sigma_p^2), \tag{4.1}$$

для некоторого $\sigma_p^2 > 0$, и зная некоторую строго состоятельную оценку $\widehat{\sigma}_N^2$ величины σ_p^2 ($\widehat{\sigma}_N^2 \xrightarrow[N\to+\infty]{\text{п.н.}} \sigma_p^2$), мы можем построить доверительный интервал:

$$C_N = (\widehat{p}_{IS} - z_{\delta/2}\widehat{\sigma}_N/\sqrt{N}; \widehat{p}_{IS} + z_{\delta/2}\widehat{\sigma}_N/\sqrt{N}), \tag{4.2}$$

где $z_{\delta/2}$ — квантиль нормального распределения $\mathcal{N}(0,1)$ уровня $\delta/2$. Тогда точность оценки $\widehat{p}_{IS}$ будет охарактеризована шириной этого доверительного интервала

$$w_\delta = 2z_{\delta/2}\widehat{\sigma}_N/\sqrt{N}.$$

Существуют различные способы вычисления $\widehat{\sigma}_N^2$ вдоль траектории марковской цепи [8]. Одним из таких методов них является метод overlapping batch means (OBM), принцип которого основан на разбиении траектории $\overline{\mu}_1, \ldots, \overline{\mu}_N$ на $N - b_N + 1$ перекрывающихся отрезков длины b_N и вычислении интересующей оценки на каждом из отрезков

$$\widehat{p}_{IS}^{(j)} = \frac{\sum\limits_{i=1}^{b_N} \mathbb{1}_{\{\mu_{i+j}\in A\}}/w(Score(\mu_{i+j}))}{\sum\limits_{i=1}^{b_N} 1/w(Score(\mu_{i+j}))}. \tag{4.3}$$

Тогда оценка дисперсии будет иметь вид

$$\widehat{\sigma}_N^2 = \frac{Nb_N}{(N-b_N)(N-b_N+1)} \sum_{j=0}^{N-b_N} \left(\widehat{p}_{IS} - \widehat{p}_{IS}^{(j)}\right)^2.$$

В [8] показано, что такая оценка является сильно состоятельной оценкой дисперсии.

Тем не менее, этот метод не может оценивать дисперсию рекурсивно, и следовательно, если траектория цепи увеличивается последовательно (N, $2N$,...), то при каждом увеличении длины траектории необходимо полностью пересчитывать оценку и поэтому хранить всю траекторию. Поэтому в случае, когда длина траектории становится очень большой, данный подход становится трудоемким по времени и требует хранения массива большой длины.

Однако появились методы, позволяющие рекурсивно пересчитывать оценку дисперсии σ_p^2 за $O(1)$ по времени и памяти. Метод [9], появившийся исторически раньше остальных, использует следующий алгоритм: пусть $(a_k)_{k\in\mathbb{N}}$ — строго возрастающая последовательность целых чисел, $a_1 = 1$ и $a_{k+1} - a_k \longrightarrow \infty$ при $a_k \to \infty$. Определим последовательность $(t_i)_{i\in\mathbb{N}}$: $t_i = a_k$ при $a_k \le i < a_{k+1}$. Тогда оценка дисперсии имеет вид

$$\widehat{\sigma}_N^2 = \sum_{i=1}^{N} \left(\widehat{p}_{IS} - \widehat{p}_{IS}^{(j)}\right)^2 / \sum_{i=1}^{N} (i - t_i + 1),$$

где

$$\widehat{p}_{IS}^{(i)} = \frac{\sum\limits_{k=t_i}^{i} \mathbb{1}_{\{\mu_k \in A\}} / w(Score(\mu_k))}{\sum\limits_{k=t_i}^{i} 1/w(Score(\mu_k))}.$$

В данной работе используется разработанный позднее подход [10] базирующийся на идее, описанной в [9], с тем дополнением, что размер отрезков корректируется при помощи второй последовательности $\{b_k\}$, $a_k - b_k > 0$, определяющей количество элементов в $k-$ом отрезке.

В [10] показано, что оценки являются состоятельными и имеют среднеквадратическое отклонение такого же порядка, как и оценки, полученные при помощи ОВМ.

4.2. Варианты критерия остановки

В практических задачах необходимо уметь вычислять вероятности (1.3) для коллекции пар $(Spectrum, P)$. Поэтому нет возможности находить достаточное для построения оценки заданной точности N для каждой конкретной траектории.

В [11] был предложен автоматический критерий остановки, основанный на том, что моделирование траектории прекращается, когда отношение ширины доверительного интервала C_N к оценке дисперсии вдоль траектории меньше заданного порога.

Формально, обозначим теоретическую дисперсию вдоль траектории за λ_p. Заметим, что $\lambda_p \neq \sigma_p^2$ из-за того, что случайные величины, входящие в марковскую цепь, коррелированные. Положим $\widehat{\lambda}_N \xrightarrow[N\to+\infty]{\text{п.н.}} \lambda_p$.

Дополнительно введем обозначение

$$N_\varepsilon = \inf\{n \geq 0 : w_{\delta/2}\sqrt{N} \leq \varepsilon\widehat{\lambda}_N\}.$$

Тогда при $N \to \infty$ и $\varepsilon \to 0$ моделирование прекратится п.н. и при $N \to \infty$ имеет место сходимость

$$\mathbb{P}(p \in C_{N_\varepsilon}) \to 1 - \delta,$$

где C_N определено в (4.2).

Полезная модификация этого подхода основывается на том, что часто нужно не вычислить p максимально точно, а понять, выполняется ли неравенство

$$p < p_0, \tag{4.4}$$

где p_0 — фиксированный порог, который часто значительно больше оцениваемых вероятностей p (например $p_0 \approx 10^{-7}$ при $p < 10^{-10}$). Таким образом, если для какого-то ε_0 выполняется $p_0 \notin C_{N_{\varepsilon_0}}$, то из этого автоматически вытекает (4.4) или обратное неравенство с вероятностью $1 - \delta$ при $N \to \infty$.

Поэтому в данной постановке задачи моделирование будем прекращать, когда p_0 в первый раз выйдет за пределы доверительного интервала $C_{N_{\varepsilon_i}}$, где $\{\varepsilon_i\}$ — убывающая последовательность. Такая модификация позволяет существенно уменьшить длину моделируемой траектории, так как она будет зависеть от $p_0 \gg p$.

4.3. Теоретические аспекты в рамках исследуемой задачи

Для того, чтобы пользоваться оценкой дисперсии и критерием остановки, необходимо, чтобы марковская цепь $\overline{\mu}_1, \ldots, \overline{\mu}_N$ удовлетворяла определенным свойствам: стационарным распределением должно являться распределение $\mathcal{Q}$; должен выполняться закон больших чисел, чтобы оценка $\widehat{p}_{IS}$ была состоятельной оценкой вероятности (1.3); а также должна выполняться центральная предельная теорема (4.1) для возможности вычисления дисперсии и доверительных интервалов.

4.3.1. Дискретный случай

Сначала рассмотрим следующую марковскую цепь. Предположим, что пространство состояний — это $\mathcal{M}_{\mathrm{d}} = \{(m_1, \ldots, m_{|V(G|}), \mid m_i > 0, \sum_{i=1}^{|V(G)|} \mu_i = M\}$, где все $m_i \in \mathbb{N}$, M фиксировано. Переходная плотность γ в методе Метрополиса-Гастингса имеет следующий вид (текущее состояние $(m_1, \ldots, m_{|V(G)|})$, следующее состояние $(\tilde{m}_1, \ldots, \tilde{m}_{|V(G)|})$).

1. Равномерно моделируется индекс i на множестве $\{1, \ldots, |V(G)|\}$.

2. Равномерно моделируется случайная величина δ на множестве $\{-m_i + 1, \ldots, m_{(i+1) \bmod |V(G)|} - 1\}$.

3. Положим

$$\tilde{m}_i \leftarrow m_i + \delta, \tag{4.5}$$

$$\tilde{m}_{i+1} \leftarrow m_{i+1} - \delta, \tag{4.6}$$

$$\tilde{m}_k \leftarrow m_k, \ k \neq i, i+1. \tag{4.7}$$

Такая цепь является дискретным аналогом марковской цепи, полученной в предложенном алгоритме. Ясно, что пространство состояний данной цепи конечно и дискретно (количество целых точек в ограниченной области в $\mathbb{R}^{|V(G)|}$).

Теорема 1. *Данная марковская цепь является эргодической.*

Доказательство. Апериодичность цепи очевидна по структуре алгоритма. Теперь докажем неприводимость данной цепи. Покажем, что вероятность перейти из произвольного состояния μ в некоторое произвольное состояние $\tilde{\mu}$ положительная. В силу того, что на каждом шаге метода Метрополиса-Гастингса мы с ненулевой вероятностью переходим в новое состояние, то достаточно предложить последовательность $m^{(1)}, \ldots, m^{(k)}$, такую что $m^{(1)} \sim \gamma(\cdot \mid m), \ldots, \tilde{m} \sim \gamma(\cdot \mid m^{(k)})$, иными словами, предъявить последовательность действий вида (4.7).

Во-первых, заметим, что (4.7) реализует всевозможные транспозиции вектора (а значит из любой перестановки вектора можно получить исходный за конечное число шагов).

Докажем наше утверждение по индукции ($n = |V(G)|$):

1. База $n = 2$. Очевидно, что из вектора $m = (m_1, m_2)$ можно получить вектор $(\tilde{m_1}, \tilde{m_2})$, взяв $i = 1$ и $\delta = \tilde{m}_1 - m_1$. Условие $\tilde{m}_1 - m_1 \in \{-m_1 + 1, \ldots, m_2 - 1\}$ переписывается в виде $\tilde{m}_1 \in \{1, \ldots, M - 1\}$, что всегда верно.

2. Теперь покажем, что если вектора $m = (m_1, \ldots, m_n)$ и $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_n)$ не имеют общих элементов, то за конечное число шагов из вектора m можно получить вектор, имеющий один общий элемент с $\tilde{m}$. Рассмотрим некоторую пару (m_i, m_j). В силу того, что нашим преобразованием реализуются транспозиции, можем считать не умаляя общности $j = i + 1, m_i < m_j$. После преобразования (4.7) каждый из этих элементов может перейти в диапазон $\{1, \ldots, m_i + m_j - 1\}$. Значит, если если в какой-то паре $(\tilde{m}_k, \tilde{m}_\ell)$ хотя бы один из элементов лежит в данном промежутке, то из (m_i, m_j) можно получить пару, имеющую общий элемент с $(\tilde{m}_k, \tilde{m}_\ell)$. А такая пара обязательно существует, в противном случае все элементы $\tilde{m}$ больше m, что противоречит равенству сумм $\sum \tilde{m}_i$ и $\sum m_i$.

3. Теперь можно последовательно применить описанный выше шаг, уменьшая количество не совпадающих элементов на единицу, пока их не станет два.

Таким образом, марковская цепь является неприводимой и апериодичной, а следовательно является эргодической. □

Следствие. *Для представленной марковской цепи выполяется центральная предельная теорема и закон больших чисел.*

В случае, когда пространство состояний становится непрерывным, ситуация усложняется. Сначала введем несколько определений, для того, чтобы предъявить необходимые условия для ЗБЧ и ЦПТ в данном контексте [12].

4.3.2. Сведения из теории марковских цепей

Пусть $(X_n)_{n \geq 0}$ — марковская цепь, определенная на множестве $\mathbf{X}$, с переходным распределением P. Иными словами

$$P(x, A) = \mathbb{P}(X_n \in A \mid X_{n-1} = x),$$
$$P^n(x, A) = \mathbb{P}(X_n \in A \mid X_0 = x),$$

для некоторого борелевского множества A.

Определение 3. *Марковская цепь* $(X_n)_{n\geq 0}$ *называется* ***φ-неприводимой****, если для меры* φ *выполняется*

$$\varphi(A) > 0 \Longrightarrow \sum_n P^n(x, A) > 0, x \in \mathbf{X}.$$

Определение 4. *Множество* $C \subset \mathbf{X}$ *называется* ***small-множеством****, если существуют* $\delta > 0$, $n > 0$ *и вероятностная мера* ν, *сосредоточенная на множестве* C, *такие что*

$$P^n(x, \cdot) \geq \delta\nu(\cdot), x \in C. \tag{4.8}$$

Определение 5. *Марковская цепь* $(X_n)_{n\geq 0}$ *называется* ***апериодичной****, если для всех small-множеств* C, *таких что* $\varphi(C) > 0$, *наибольший делитель констант* n *из* (4.8) *равен единице.*

Определение 6. *Марковская цепь* $(X_n)_{n\geq 0}$ *называется* ***Харрис-рекуррентной****, если существует* $A \subset \mathbf{X}$, $\beta > 0$, $m \geq 1$ *и вероятностная мера* ψ *на* $\mathbf{X}$, *такие что*

1. $\mathbb{P}_x(T_A < \infty) = 1, x \in \mathbf{X}$, *где* $T_A = \inf\{n \geq 0 \mid X_n \in A\}$
2. $\mathbb{P}_x(X_m \in \cdot) \geq \beta\psi(\cdot)$, $x \in A$.

Если марковская цепь является φ-неприводимой, апериодичной и Харрис-рекуррентной, то говорят, что она является **Харрис-эргодической**.

Предположим что марковская цепь является Харрис-эргодической со стационарным распределением π. В таком предположении выполняется сильная сходимость [12]

$$||P^n(\mu^{\text{leb}}, \cdot) - \pi(\cdot)|| \longrightarrow 0, \ n \to \infty,$$

где $P^n(\mu^{\text{leb}}, \cdot) = \int\limits_{\mathbf{X}} P^n(x, A)\mu^{\text{leb}}(dx)$ и $||\cdot||$ — норма полной вариации.

Определение 7. *Говорят, что марковская цепь сходится с геометрической скоростью сходимости, если*

$$||P^n(x, \cdot) - \pi(\cdot)|| \leq M(x)t^n,$$

где $M(x)$ *— некоторая неотрицательная функция, а* $0 < t < 1$.

Имеет место следующая теорема [12]:

Теорема 2. *Рассмотрим марковскую цепь* $(X_n)_{n\geq 0}$ *со стационарным распределением* π *и функцию* $f : \mathbf{X} \longrightarrow \mathbb{R}$. *Пусть* $V : \mathbf{X} \longrightarrow [1;\infty)$. *Если существуют константы* $d > 0$, $b < \infty$ *и* $0 \leq \tau < 1$, *такие что*

$$\int V(y)P(x,dy) - V(x) \leq -d[V(x)]^T + b\mathbb{1}_{\{x\in C\}},\ x \in \mathbf{X}, \tag{4.9}$$

где C *— small-множество, и при этом* $f^2(x) \leq V(x)$ *для всех* $x \in \mathbf{X}$, *то* $\sigma_f^2 \in [0;\infty)$, *и если* $\sigma_f^2 > 0$, *то для любого начального распределения*

$$\sqrt{N}(\frac{1}{N}\sum_i = 1^N f(X_i) - \mathbb{E}_\pi f) \xrightarrow{D} \mathcal{N}(0,\sigma_f^2).$$

4.3.3. Непрерывный случай

Утверждение 1. *Марковская цепь, полученная в предложенном алгоритме является* q*-неприводимой.*

Доказательство. Для каждого вектора $\mu_i \in \mathcal{M}$ существует открытое множество A_i такое, что $\gamma(\mu \mid \mu_i) > 0$, $\mu \in A_i$. Ясно, что $\bigcup_i A_i$ покрывает все множество $\mathcal{M}$. В силу того, что $\mathcal{M}$ является компактным, то из такого покрытия можно выбрать конечное подпокрытие. А значит, $P^n(x,A) > 0$ для любого $x \in \mathcal{M}$, $A \in \mathcal{B}(\mathcal{M})$. □

Имеет место следующее утверждение [13].

Теорема 3. *Любая* φ*-неприводимая марковская цепь, для которой выполняется* $P(x,\{x\}) > 0$, *является апериодичной.*

Так как доверительная вероятность α положительная, то условие теоремы выполняется.

С учетом того, что марковская цепь является апериодичной, из доказательства неприводимости следует, что вероятность достичь некоторого множества A из начального состояния x равна единице, а значит, предложенная марковская цепь также является Харрис-рекуррентной, а значит, и Харрис-эргодической.

Замечание 6. *Таким образом, для полученной марковской цепи выполняется закон больших чисел. Для того чтобы получить центральную предельную теорему, необходимо проверить условие геометрической сходимости и условие* (4.9).

Глава 5

Численные результаты

5.1. Одиночные идентификации

Корректность работы алгоритма проверялась на нескольких пептидах линейной и циклической структуры. Линейные пептиды были выбраны таким образом, чтобы оценки вероятности (1.3) можно было получить при помощи метода Монте-Карло за разумное количество времени.

В качестве соединений циклической структуры были выбраны несколько примеров из [3], а именно $(10, 20, 40)$, $(10, 20, 40, 80)$, $(10, 20, 40, 80, 160)$, $(10, 20, 40, 80, 160, 320)$, и $(10, 20, 40, 80, 160, 320, 640)$, а также белок *Surfactin*.

Введем следующие обозначения:

1. $\widehat{p}_{MC}$ — оценки, полученные при помощи стандартного метода Монте-Карло.
2. $\widehat{p}_{IS}$ — оценки, полученные по представленному методу.
3. $\widehat{p}_{DPR}$ — оценки, полученные в методе MS-DPR, описанном в [3]. Заметим, что данный метод не позволяет сосчитать дисперсию оценки, поэтому доверительные интервалы в таблицах не представлены.

Верификация алгоритма осуществлялась следующим образом:

1. Были сосчитаны оценки $\widehat{p}_{MC}$ с $N = 50 \cdot 10^6$; $\widehat{p}_{IS}$ при помощи критерия остановки с $\epsilon = 0.02$; $\widehat{p}_{DPR}$ были получены с настройками MS-DPR по умолчанию.
2. Для $\widehat{p}_{MC}$ и $\widehat{p}_{IS}$ были получены доверительные интервалы, которые можно видеть в таблице 5.2.

Результаты представлены в таблицах 5.1 и 5.2. Из них можно видеть, что доверительные интервалы для $\widehat{p}_{IS}$ лежат внутри доверительных интервалов по методу Монте-Карло и часто имеют меньшую ширину. Также можно заметить, что $\widehat{p}_{DPR}$, напротив, лежат вне левой границы доверительных интервалов. Такое смещение оценок влечет за собой ложные «идентификации».

Длины $N = 50{\cdot}10^6$ оказалось недостаточно, чтобы оценить $\widehat{p}_{MC}$ для $(10, 20, 40, 80, 160)$, $(10, 20, 40, 80, 160, 320)$, and $(10, 20, 40, 80, 160, 320, 640)$, а также метод MS-DPR не смог оценить $\widehat{p}_{DPR}$ для последнего пептида.

Таблица 5.1. Сравнение методов Монте-Карло, MCMC и MS-DPR: оценки

Пептид	$\widehat{p}_{IS}$	$\widehat{p}_{MC}$	$\widehat{p}_{DPR}$
PPAEDSQK	$4.87 \cdot 10^{-7}$	$4.20 \cdot 10^{-7}$	$6.6 \cdot 10^{-7}$
PPAEDSQK	$1.40 \cdot 10^{-6}$	$1.52 \cdot 10^{-6}$	$6.5 \cdot 10^{-7}$
GQGDPGSNPNK	$4.70 \cdot 10^{-7}$	$6.40 \cdot 10^{-7}$	$1.5 \cdot 10^{-8}$
HSNAAQTQTGEANR	$2.39 \cdot 10^{-6}$	$2.22 \cdot 10^{-6}$	$4.9 \cdot 10^{-8}$
GEEEPSQGQK	$1.03 \cdot 10^{-6}$	$1.04 \cdot 10^{-6}$	$3.6 \cdot 10^{-7}$
$(10, 20, 40)$	0.00184	0.00184	0.00197
$(10, 20, 40, 80)$	$7.35 \cdot 10^{-6}$	$7.34 \cdot 10^{-6}$	$9.36 \cdot 10^{-6}$
$(10, 20, 40, 80, 160)$	$6.76 \cdot 10^{-9}$	N/A	$4.49 \cdot 10^{-9}$
$(10, 20, 40, 80, 160, 320)$	$1.74 \cdot 10^{-12}$	N/A	$1.56 \cdot 10^{-12}$
$(10, 20, 40, 80, 160, 320, 640)$	$4.08 \cdot 10^{-16}$	N/A	N/A
Surfactin	$1.18 \cdot 10^{-5}$	$1.13 \cdot 10^{-5}$	$1.01 \cdot 10^{-5}$

Теперь, с N, полученным по критерию остановки, мы получили оценки по $\widehat{p}_{MC}$. Учитывая, что теперь размер выборки одинаковый, мы сравнили дисперсии $\widehat{p}_{IS}$ и $\widehat{p}_{MC}$. Заметим, что величины N, полученной из критерия остановки, было недостаточно для линейного пептида GQGDPGSNPNK, поэтому размер выборки пришлось увеличить. В таблице 5.3 представлены результаты сравнения дисперсий. Результаты показывают, что чем меньше оцениваемая вероятность, тем больше отношение дисперсии $\widehat{\sigma}^2_{MC}$ к $\widehat{\sigma}^2_{MC}$.

Таблица 5.2. Сравнение методов Монте-Карло, MCMC и MS-DPR: 95% доверительные интервалы

Пептид	Доверительный интервал, $\widehat{p}_{IS}$		Доверительный интервал, $\widehat{p}_{MC}$	
PPAEDSQK	$4.74 \cdot 10^{-7}$	$4.99 \cdot 10^{-7}$	$2.40 \cdot 10^{-7}$	$6.00 \cdot 10^{-7}$
PPAEDSQK	$1.36 \cdot 10^{-6}$	$1.43 \cdot 10^{-6}$	$1.18 \cdot 10^{-6}$	$1.86 \cdot 10^{-6}$
GQGDPGSNPNK	$4.53 \cdot 10^{-7}$	$4.87 \cdot 10^{-7}$	$4.18 \cdot 10^{-7}$	$8.62 \cdot 10^{-7}$
HSNAAQTQTGEANR	$2.30 \cdot 10^{-6}$	$2.48 \cdot 10^{-6}$	$1.81 \cdot 10^{-6}$	$2.63 \cdot 10^{-6}$
GEEEPSQGQK	$9.96 \cdot 10^{-7}$	$1.07 \cdot 10^{-6}$	$7.57 \cdot 10^{-7}$	$1.32 \cdot 10^{-6}$
$(10, 20, 40)$	$1.80 \cdot 10^{-3}$	$1.88 \cdot 10^{-3}$	$1.82 \cdot 10^{-3}$	$1.85 \cdot 10^{-3}$
$(10, 20, 40, 80)$	$7.12 \cdot 10^{-6}$	$7.58 \cdot 10^{-6}$	$6.60 \cdot 10^{-6}$	$8.10 \cdot 10^{-6}$
$(10, 20, 40, 80, 160)$	$6.4 \cdot 10^{-9}$	$7.10 \cdot 10^{-9}$	N/A	N/A
$(10, 20, 40, 80, 160, 320)$	$1.51 \cdot 10^{-12}$	$1.97 \cdot 10^{-12}$	N/A	N/A
$(10, 20, 40, 80, 160, 320, 640)$	$3.60 \cdot 10^{-16}$	$4.55 \cdot 10^{-16}$	N/A	N/A
Surfactin	$1.14 \cdot 10^{-5}$	$1.22 \cdot 10^{-5}$	$1.03 \cdot 10^{-5}$	$1.23 \cdot 10^{-5}$

Таблица 5.3. Сравнение методов Монте-Карло и MCMC: дисперсии

Пептид	$\hat{\sigma}^2_{IS}$	$\hat{\sigma}^2_{MC}$	$\hat{\sigma}^2_{MC}/\hat{\sigma}^2_{IS}$	N
PPAEDSQK	$2.09 \cdot 10^{-10}$	$4.94 \cdot 10^{-7}$	2358.98	5000000
PPAEDSQK	$1.25 \cdot 10^{-9}$	$1.11 \cdot 10^{-6}$	890.33	3400000
GQGDPGSNPNK	$2.33 \cdot 10^{-10}$	$1.49 \cdot 10^{-7}$	639.49	20000000*
HSNAAQTQTGEANR	$5.56 \cdot 10^{-9}$	$2.24 \cdot 10^{-6}$	403.23	2800000
GEEEPSQGQK	$1.23 \cdot 10^{-9}$	$7.89 \cdot 10^{-7}$	642.19	3800000
$(10, 20, 40)$	$5.47 \cdot 10^{-4}$	$1.88 \cdot 10^{-3}$	3.43	1500000
$(10, 20, 40, 80)$	$9.93 \cdot 10^{-8}$	$7.60 \cdot 10^{-6}$	76.53	7500000
Surfactin	$1.15 \cdot 10^{-7}$	$1.00 \cdot 10^{-5}$	86.96	2000000

5.2. База данных GNPS

Предложенный алгоритм позволяет оценить указанную вероятность для некоторого экспериментального спектра $Spectrum_i$ и пептида с известной структурой P_j. Однако на практике мы имеем множество экспериментальных спектров $\{Spectrum_i\}_i$ и базу данных пептидных соединений различной структуры и массы $\{P_j\}_j$.

Таким образом, в случае поиска спектра по базе данных, необходимо уметь оценивать вероятность

$$\mathbb{P}(Scoring(Spectrum, P) > t),$$

где P из множества всех возможных спектров, зная лишь оценки вероятностей (5.2). Данная проблема аналогична проблеме множественных сравнений (при проверке статистических гипотез), и стандартной техникой, которая используется в задачах такого рода, является оценка так называемого False Discovery Rate (FDR). Зафиксировав некоторый порог α, определим

$$\mathrm{FDR}(\alpha) = \frac{\mathbb{E}V_\alpha}{R_\alpha},$$

где R_α — количество спектров S_j из базы, таких что $\mathbb{P}(Score(Spectrum_i, S_j) > t) < \alpha$, где $Spectrum_i$ — некоторый экспериментальный спектр, V_α — количество тех спектров из S_j, которые действительно имеют «схожую» структуру с исследуемым образцом $Spectrum_i$.

Оценка $\widehat{\mathrm{FDR}}$ величины $\mathrm{FDR}(\alpha)$ строится следующим образом [14]: база делится на две части — target и decoy. При этом часть target состоит из теоретических масс-спектров известных пептидов, а decoy заполнена (некоторыми специально построенными, см. [14]) случайными векторами. Тогда оценка $\widehat{\mathrm{FDR}}$ вычисляется следующим образом:

$$\widehat{\mathrm{FDR}}(\alpha) = \frac{V_\alpha^{\mathrm{target}}}{V_\alpha^{\mathrm{target}} + V_\alpha^{\mathrm{decoy}}},$$

где $V_\alpha^{\mathrm{target}}$, $V_\alpha^{\mathrm{decoy}}$ — количество спектров S_j из target и decoy баз соответственно, таких что $\mathbb{P}(Score(Spectrum_i, S_j) > t) < \alpha$.

В таблице 5.4 представлены результаты оценки $\widehat{\mathrm{FDR}}$ для разных α, полученные по базе Global Natural Products Social Molecular Networking (GNPS) [15]. Это база открытого доступа, предназначенная для совместного использования необработанных, обработанных и идентифицированных тандемных масс-спектрометрических данных, позволяющая поддерживать улучшенные аннотации.

Таблица 5.4. Сравнение методов MSDPR и MCMC на данных GNPS

	MSDPR			MCMC		
$-\log_{10}\alpha$	target	decoy	$\widehat{FDR}$ %	target	decoy	$\widehat{FDR}$ %
7	762	188	19.78	744	179	19.39
8	619	110	15.08	610	104	14.56
9	505	52	9.33	473	51	9.73
10	443	33	6.93	415	30	6.74
11	393	21	5.07	354	20	5.34
12	354	15	4.06	312	12	3.70
13	322	11	3.30	271	7	2.51
14	293	11	3.61	238	2	0.83
15	264	7	2.58	201	1	0.49
16	238	5	2.05	169	0	0.0
17	211	2	0.93	138	0	0.0
18	188	0	0.0	104	0	0.0
19	157	0	0.0	87	0	0.0
20	139	0	0.0	76	0	0.0

В практических задачах интерес представляют «идентификации» с очень маленьким уровнем значимости α. Заметим, что в таблице 5.4 для $\alpha \in [10^{-12}; 10^{-17}]$ оценка $\widehat{\text{FDR}}$ полученная по методу MCMC меньше, чем полученная в MS-DPR.

Заключение

Таким образом, в данной работе был предложен алгоритм оценки вероятности (1.3) для пары («Экспериментальный спектр», «Теоретический пептид»). Был продемонстрирован способ оценки ее дисперсии и предложен критерий остановки моделирования траектории марковской цепи, чтобы ее длины было достаточно для построения оценки заданной точности.

Корректность работы полученного алгоритма была показана на наборе пептидов различной структуры и проведено сравнение с существующим методом MS-DPR. Согласно результатам, полученный метод имеет меньший FDR для низких порогов значимости, что доказывает применимость метода в задачах биоинформатики.

Список литературы

1. Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases / H. Mohimani, W. Liu, J. Mylne et al. // Journal of Proteome Research. — 2011. — Vol. 10.
2. Kim S., Pevzner P. MS-GF+ makes progress towards a universal database search tool for proteomics // Nature Communications. — 2014. — Vol. 5.
3. Mohimani H., Kim S., Pevzner P. A. A new approach to evaluating statistical significance of spectral identifications // J. Proteome Res. — 2013. — Vol. 12, no. 4. — P. 1560–1568.
4. David D., Minh L., Minh D. Understanding the Hastings algorithm // Communications in Statistics - Simulation and Computation. — 2014. — Vol. 44. — P. 332–349.
5. Harris T. E. The existence of stationary measures for certain Markov processes. // In Proc. 3rd Berkeley Symp. Math. Statist. Probab. — Vol. 2. — California Press, Berkeley, 1956. — P. 113 – 124.
6. Iba Y., Saito N. D., Kitajima A. Multicanonical MCMC for sampling rare events: An illustrative review. // Annals of the Institute of Statistical Mathematics. — 2014. — Vol. 66. — P. 611–645.
7. Tierney L. Markov chains for exploring posterior distributions // Ann. Statist. — 1994. — Vol. 22, no. 4. — P. 1701–1728.
8. Flegal J., Jones G. Batch means and spectral variance estimators in Markov chain Monte Carlo // Annals of Statistics. — 2010. — Vol. 38. — P. 1034–1070.
9. Wu W. Recursive estimation of time-average variance constants // The Annals of Applied Probability. — 2009. — Vol. 19. — P. 1529–1552.
10. Chan K., Chun Y. New recursive estimators of the time-average variance constants // Statistics and Computing. — 2016. — Vol. 26. — P. 609–627.
11. Flegal J. M., Gong L. Relative fixed-width stopping rules for Markov chain Monte Carlo simulations // Statistica Sinica. — 2015. — Vol. 25. — P. 655–676.
12. Jones G. L. On the Markov chain central limit theorem // Probability Surveys. — 2004. — Vol. 1. — P. 299–320.
13. Mengersen K. L., Tweedie R. L. Rates of convergence of the Hastings and Metropolis algorithms // Ann. Stat. — 1996. — Vol. 24. — P. 101–121.
14. Elias J. E., Gygi S. P. Target-decoy search strategy for increased confidence in large-

scale protein identifications by mass spectrometry // Nat. Methods. — 2007. — Vol. 4, no. 3. — P. 207–214.

15. Wang M. et al. Sharing and community curation of mass spectrometry data with GNPS // Nature Biotechnology. — 2016. — Vol. 34. — P. 828–837.