Оценка малых вероятностей при помощи метода Монте-Карло по схеме марковской цепи

Абрамова А. Н., студент кафедры статистического моделирования СПбГУ, abramova.asya93@gmail.com

Коробейников А.И., кандидат физико-математических наук, доцент кафедры статистического моделирования СПбГУ, a.korobeynikov@spbu.ru

Аннотация

Рассматривается задача поиска пептида по базе данных и оценивания вероятности близости данного пептида к некоторому пептиду из базы. Сформулирована и формализована на вероятностном языке более простая задача сравнения данной нуклеотидной строки с некоторым множеством нуклеотидных строк. В работе предложено решение этой задачи при помощи метода Метрополиса-Гастингса и алгоритма Ванга-Ландау.

Введение

Пептиды — это семейство веществ, молекулы которых построены из двух и более остатков аминокислот, соединённых в цепь пептидными связями. Существуют бактерии способные продуцировать пептидные соединения, подавляющие рост определённых микроорганизмов или вызывающие их гибель, то есть антибиотики.

Важным вопросом, связанным с исследованием пептидов, является их идентификация. Другими словами, необходимо понять, насколько исследуемый пептид P близок по структуре к некоторому известному пептиду P^* (близость структур пептидов влечет за собой близость их свойств). Наиболее распространенным инструментом для решения данной задачи является масс-спектрометрия. Методы масс-спектрометрии заключаются в том, что по данному пептиду экспериментально строится его так называемая «*фрагментация*», после чего измеряется масса каждого фрагмента, и в дальнейшем исследуется полученный массив масс, именуемый *спектром* [7].

Таким образом, существует задача идентификации пептида по его масс-спектру: исследование схожести двух пептидов сводится к исследованию схожести их спектров, а именно к оценке их близости на статистическом языке.

Существуют определенные методы, в том или ином роде решающие данную задачу (например, [6]). Тем не менее, остается неизвестной точность полученных оценок, а также возникает потребность увеличения скорости рабо-

ты существующих алгоритмов в связи с большими объемами данных, которые характерны для поставленной задачи.

На начальном этапе была рассмотрена упрощенная задача — задача идентификации нуклеотидных строк, которую формально опишем в следующем разделе.

Постановка задачи

Обозначим $\sigma = \{A, C, G, T\}$, зафиксируем $k \in \mathbb{N}$. Пусть σ_k — множество всех строк длины k над алфавитом σ.

Определение 1 *Рассмотрим строки $x, y \in \sigma_k$. Локальным выравниванием длины m строк x и y будем называть набор индексов $B = B(x, y) = \{(i_1, j_1), \ldots, (i_m, j_m)\}$, где:*

1. *При $u < v$ выполняется $i_u < i_v$, $j_u < j_v$,*
2. *Для любых u, v: если $i_u < j_u$, то $i_v < j_v$.*

Дополнительно, если для любого $u \in \{1, \ldots, m\}$ выполняется $i_{u+1} = i_u + 1$ и $j_{u+1} = j_u + 1$, то выравнивание называется выравниванием без пропусков.

Множество всевозможных выравниваний будем обозначать $\mathbb{B}$. На множестве пар строк определим функцию «расстояния» s, такую что для строк $x, y \in \sigma_k$:

$$s(x,y) = \max_{B,m} \left(c \sum_{\ell=1}^{m} \mathbb{I}_{\{x_{i_\ell} = y_{j_\ell}\}} + d \sum_{\ell=1}^{m} \mathbb{I}_{\{x_{i_\ell} \neq y_{j_\ell}\}} + h \sum_{\ell=1}^{m-1} \max\{i_{\ell+1} - i_\ell, j_{\ell+1} - j_\ell\} \right),$$

где $c, d, h \in \mathbb{R}$ — некоторые фиксированные константы.

Замечание 1 *Значения функции s зависят от параметров c, d и h, поэтому само значение $s(x, y)$ для некоторых $x, y \in \sigma_k$ не несет никакой информации относительно их «близости». Поэтому можно рассматривать следующие задачи:*

1. *Вычислить $\#\{(x, y) : s(x, y) > S^*\}$, где S^* — заранее заданный порог.*
2. *Вычислить $\#\{y : s(x_0, y) > S^*\}$, где S^* — заранее заданный порог, $x_0 \in \sigma_k$ — фиксированная строка.*

Вторая задача соответствует поиску образца в базе, где базой является множество σ_k, и поэтому далее мы будем рассматривать ее.

Статистическая постановка задачи

Теперь формализуем задачу с вероятностной точки зрения. Для этого будем считать, что строки — случайные величины.

Пусть $\mathbf{2}^{\sigma_k}$ — множество всех подмножеств σ_k. $\mathcal{P}_k$ — распределение на σ_k, такое что $\mathcal{P}_k(A) = \#A/\#\sigma_k$ для любого $A \in \mathbf{2}^{\sigma_k}$. На некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$ определим случайную величину $\xi : (\Omega, \mathcal{F}) \longrightarrow (\sigma_k, \mathbf{2}^{\sigma_k})$, такую что $\mathbb{P}(\xi \in A) = \mathcal{P}_k(A)$ для любого $A \in \mathbf{2}^{\sigma_k}$. Таким образом, все буквы строки будут равновероятны.

Зафиксируем строку $x_0 \in \sigma_k$. Тогда в статистическом смысле задача, описанная в замечании 1, эквивалентна оценке вероятности

$$p = \mathbb{P}(s(x_0, \xi) > S^*). \tag{1}$$

Известно, что для локального выравнивания без пропусков имеется следующий асимптотический результат [1]: распределение значений функции s сходится к распределению Гумбеля с определенными параметрами при $k \to \infty$. В дальнейшем для сравнения оценок величины (1), полученных в результате реализованного алгоритма, мы будем использовать оценки, полученные при помощи программы BLAST, которая использует данную асимптотическую аппроксимацию (подробнее см. [1]).

Замечание 2 *Далее будем обозначать* $s(y) := s(x_0, y)$.

Проблема оценки редких событий

Стандартным методом оценивания вероятности (1) является метод Монте-Карло. Рассмотрим случайную величину ξ, определенную на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$ с распределением $\mathcal{P}$. Обозначим $p = \mathbb{P}(\xi \in A)$, A — некоторое борелевского множество.

Определение 2 *Пусть* $x_1, \ldots, x_N$ *— независимые одинаково распределенные случайные величины с распределением* $\mathcal{P}$. *Оценкой по методу Монте-Карло называется*

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

Заметим, что дисперсия такой оценки $\mathbb{D}(\widehat{p}_{MC}) = \frac{p(1-p)}{N}$. В задаче оцениваемая вероятность p крайне мала, поэтому оценки, полученные при помощи

метода Монте-Карло, будут обладать очень большой дисперсией по отношению к p_{MC}. Вследствие этого задача получения оценок с заданной точностью становится очень трудоемкой, поэтому будем использовать метод MCMC (Markov Chain Monte-Carlo).

Оценка по методу MCMC

Метод существенной выборки В предложенных выше обозначениях рассмотрим следующий способ построения оценки $\widehat{p}$. Пусть $\mathcal{Q}$ — некоторое распределение, определенное на $(\Omega, \mathcal{F})$ с соответствующей функцией распределения Q и плотностью q относительно некоторой меры ν. Предположим также, что существует производная Радона-Никодима $d\mathcal{P}/d\mathcal{Q}$ (в данном случае $(d\mathcal{P}/d\mathcal{Q})(x) = f(x)/q(x)$ в точках, где $q(x) \neq 0$ и равна нулю в точках, где $q(x) = 0$).

Определение 3 *Пусть $x_1, \ldots, x_N$ — одинаково распределенные случайные величины с распределением $\mathcal{Q}$. Оценкой Монте-Карло по методу существенной выборки для вероятности p будем называть*

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

Такая оценка будет несмещенной оценкой p.

Выбор моделирующего распределения $\mathcal{Q}$ Далее, будем считать, что плотность q имеет конкретный вид, а именно

$$q(x) = cw(s(x))f(x), \tag{2}$$

где w — некоторая положительная функция, $c > 0$ — нормирующая константа, s — функция, определенная в замечании (2).

Утверждение 1 *Если плотность q имеет вид* (2)*, то оценка*

$$\widehat{p} = \frac{\sum\limits_{n=1}^{N} \mathbb{I}_{\{X_n \in A\}} / w(s(X_n))}{\sum\limits_{n=1}^{N} 1/w(s(X_n))}, \tag{3}$$

является несмещенной оценкой p.

Выбор весовой функции w Напомним, что нашей задачей является оценка вероятности (3) и необходимо минимизировать относительную ошибку $(\mathbb{D}\widehat{p}_{MC}/p_{MC})$ этой оценки, поэтому будем искать функцию q вида (2) так, чтобы для случайной величины $\xi \sim \mathcal{Q}$ распределение $s(\xi)$ было близко к равномерному распределению на $[S_{min}, S_{max}]$. Для этого достаточно выбирать весовую функцию w так, чтобы $w(S) \approx 1/\mathbb{P}(s(\xi) = S)$.

Построение марковской цепи Будем моделировать марковскую цепь $(X_n)_{n\geq 1}$ со стационарным распределением $\mathcal{Q}$ с помощью метода Метрополиса-Гастингса [2]. Для этого достаточно задать переходную плотность марковской цепи, удовлетворяющей уравнению детального баланса [3].

Теперь опишем, как выглядит переходная плотность γ в поставленной задаче. Пусть x — текущее состояние. Тогда новое состояние y получим следующим образом: смоделируем случайную величину j равномерно на множестве индексов $\{1, \ldots, k\}$ и букву ch равномерно на множестве $\sigma = \{A, C, G, T\}$.

- С вероятностью 1/2 положим $y[z] \leftarrow x[z]$ для $z = 1, \ldots, k$, $y[j] \leftarrow ch$,
- С вероятностью 1/8 положим $y[z] \leftarrow x[z+1]$ для $z = j, \ldots, k-1$, $y[k] \leftarrow ch$
- С вероятностью 1/8 положим $y[z] \leftarrow x[z-1]$ для $z = 2, \ldots, j$, $y[1] \leftarrow ch$
- С вероятностью 1/8 положим $y[z+1] \leftarrow x[z]$, $z = j, \ldots k-1$, $y[j] \leftarrow ch$.
- С вероятностью 1/8 положим $y[z-1] \leftarrow x[z]$, $z = 2, \ldots j$, $y[j] \leftarrow ch$.

Замечание 3 *Для данной переходной плотности и целевой плотности q, вероятность перехода в новое состояние равна* $\min\left(\frac{w(s(x))}{w(s(y))}, 1\right)$.

Весовую функцию w будем оценивать алгоритмом Ванга-Ландау, который в общем виде описан в [4]. Данный алгоритм является адаптивной модификацией метода Метрополиса-Гастингса и позволяет оценить требуемые веса одновременно с моделированием траектории марковской цепи.

Однако получающаяся в результате работы алгоритма марковская цепь нестационарна и, вообще говоря, не обязательно является эргодической. В связи с этим, построение оценок проводится в два этапа: сначала при помощи алгоритма Ванга-Ландау оцениваются значения функции $w(s)$, затем полученные веса используются для моделирования марковской цепи с соответствующей целевой плотностью q методом Метрополиса-Гастингса.

Численные результаты

Результатом описанного алгоритма является марковская цепь $\tilde{X}_1, \ldots, \tilde{X}_{M_{max}}$ со стационарным распределением $\mathcal{Q}$.

Чтобы строить доверительные интервалы для оценки (3) необходимо знать ее дисперсию, а значит нужно получить набор слабо коррелированных величин $X_1, \ldots X_N$. При помощи пакета «coda» языка программирования **R** [5] изменим полученную марковскую цепь следующим образом:

1. Найдем такой индекс k в траектории, начиная с которого наступает момент «стабилизации», и рассмотрим $\tilde{X}_k, \ldots, \tilde{X}_{M_{max}}$ (функция cumuplot).

2. С помощью функции autocorr.diag найдем такое значение l, что корреляция между X_k и $\tilde{X}_{k+l}$ достаточно мала и положим

$$X_i = \tilde{X}_{k+l(i-1)}, \text{для } i = 1, \ldots, \left[\frac{M_{max} - k}{l}\right].$$

Полученные оценки будем сравнивать следующим образом:

1. Для строк длины k при $S^* = S_{max}$ в (1) оцениваемая вероятность равна $1/|\sigma|^k$.

2. Сравним результаты описанного алгоритма с методом Монте-Карло.

3. Проведем сравнение с BLAST (с оценками на основе асимптотических результатов).

Результаты описанных выше сравнений представлены в таблице 1. Вычисление доверительных интервалов для оценок по полученному методу является нетривиальной задачей, которая будет рассмотрена при дальнейшей работе.

Таблица 1: Сравнение методов, $k = 10$, теоретическое значение $p-value$ для $S^* = 10$ равняется $9.5 \cdot 10^{-7}$

	$S^* = 10$	$S^* = 8$
	$\widehat{p}$	$\widehat{p}$
MC	$5 \cdot 10^{-7}$	0.00014
MCMC	$1.3 \cdot 10^{-6}$	0.00013
BLAST	$5 \cdot 10^{-5}$	$8 \cdot 10^{-4}$

Заключение

В работе был рассмотрен вопрос сравнения двух случайных строк заданной длины над фиксированным алфавитом.Также был реализован алгоритм, вычисляющий оценки для вероятности (1) при помощи метода Метрополиса-Гастингса и Ванга-Ландау. Было проведено сравнение данного метода со стандартным методом Монте-Карло и численно была продемонстрирована адекватность полученных оценок. В дальнейшем планируется модифицировать данный алгоритм для сравнения близости пептидных спектров и сравнить его с уже существующими методами.

Литература

[1] Altschul S. F., Gish W., W. M. Basic local alignment search tool // Journal of Molecular Biology. — 1990. — Vol. 215. — P. 403–410.

[2] David D. L. M., Minh D. L. P. Understanding the Hastings algorithm // Communications in Statistics - Simulation and Computation. — 2014. — Vol. 44. — P. 332–349.

[3] 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.

[4] 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.

[5] Plummer M., Best N., Cowles K. et al. coda: Output analysis and diagnostics for MCMC. — 2015.

[6] Mohimani H., Kim S., Pevzner P. A. A New Approach to Evaluating Statistical Significance of Spectral Identifications // Journal of Proteome Research. — Department of Electrical and Computer Engineering and Department of Computer Science and Engineering, University of California — San Diego, San Diego, California 92093

[7] Mohimani H., Liu, W., Mylne, J., Poth, A., Colgrave M., Tran D., Selsted M., Dorrestein P., Pevzner P. Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases // Journal of Proteome Research. — 2011 — Vol. 10 — P. 4505–4512.