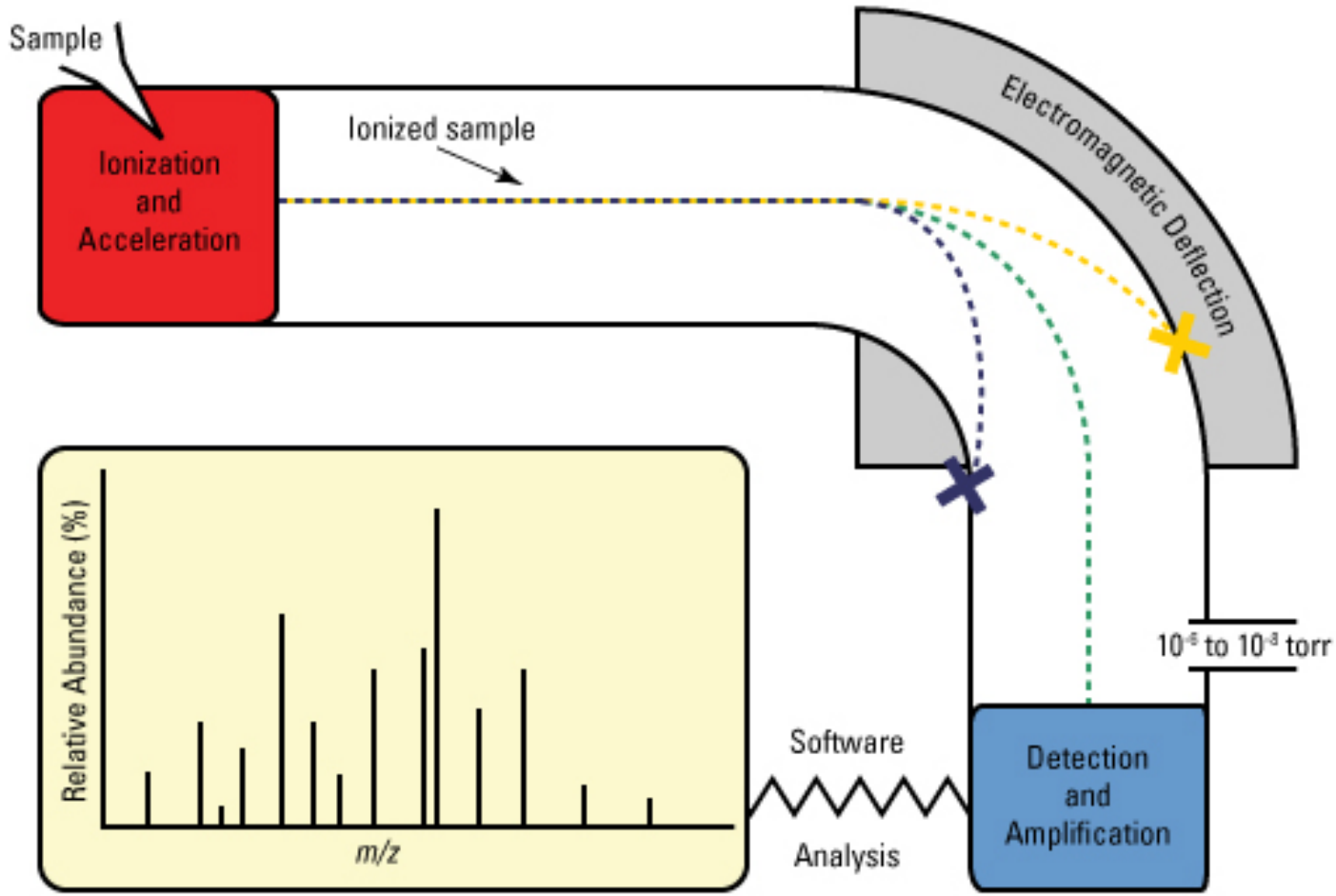
Метод Монте-Карло по схеме марковской цепи для оценки вероятности редких событий в задачах биоинформатики

Абрамова Анастасия Николаевна, гр. 15.М03-мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

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

Методы масс-спектрометрии



Модель масс-спектра пептида (Mohimani *et al.*, 2013)

Пептид P с массой M, состоящий из k аминокислот

- вектором масс аминокислот $m = (m_1, \ldots, m_k)$
- матрицей структуры $\mathbb{H}$ размера $k \times r$.

Ожидаемый масс-спектр пептида P имеет вид $S = \mathbb{H}m$.

Определим множество $\mathcal{M}$:

$$\mathcal{M} = \{m = (m_1, \ldots, m_k) \mid m_i > 0, \sum_{i=1}^{k} m_i = M\}.$$

Тогда $\mathcal{M}$ и фиксированная матрица $\mathbb{H}$ описывают множество пептидов одинаковой массы и химической структуры.

Оценка статистической значимости

Рассмотрим экспериментальный спектр:

$$\tilde{S} = (\tilde{s}_1, \ldots, \tilde{s}_\ell), \quad \tilde{s}_i > 0.$$

Мерой похожести экспериментального спектра $\tilde{S}$ и масс-спектра $\mathbb{H}m^*$ назовем значение $t = \overline{\text{Score}}(\tilde{S}, \mathbb{H}m^*)$.

Определение

Статистической значимостью значения t будем называть вероятность

$$p = \mathbb{P}(\overline{\text{Score}}(\tilde{S}, \mathbb{H}m) \geq t) = \mathbb{P}(\text{Score}(m) \geq t) = \mathbb{P}(m \in \mathcal{S}),$$

предполагая что m имеет равномерное распределение на $\mathcal{M}$.

В приложениях $p \approx 10^{-20}$.

Алгоритм MS-DPR

В работе (Mohimani *et al.*, 2013) предложен алгоритм MS-DPR, вычисляющий оценку статистической значимости.

Недостатки данного алгоритма состоят в том, что

- его точность неизвестна,
- размер выборки, используемой для построения оценки, задается заранее,
- оценки являются смещенными вниз.

Цели работы

Таким образом, целью работы является:

- построение оценки $\widehat{p}$,
- вычисление ее дисперсии $\sigma^2_{\widehat{p}}$ и построение доверительного интервала для $\widehat{p}$,
- оценка достаточного размера выборки для получения $\widehat{p}$ заданной точности.

Оценка по методу существенной выборки

Рассмотрим выборку $m_1, \ldots, m_N \sim f$, где f — плотность $U(\mathcal{M})$. Пусть g — плотность некоторого распределения $\mathcal{G}$.

Оценка по методу существенной выборки для вероятности $p = \mathbb{P}(m \in \mathcal{S})$:

$$\widehat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(m_i)}{g(m_i)} \mathbb{I}_{\mathcal{S}}(m_i).$$

Пусть $g(m) \propto w(\mathrm{Score}(m)) f(m)$, тогда:

$$\widehat{p}_{IS} = \frac{\sum\limits_{i=1}^{N} \mathbb{I}_{\mathcal{S}}(m_i) / w(\mathrm{Score}(m_i))}{\sum\limits_{i=1}^{N} 1 / w(\mathrm{Score}(m_i))}.$$

Для построения $\widehat{p}_{IS}$ необходимо уметь моделировать случайные величины из распределения $\mathcal{G}$.

Алгоритм Метрополиса–Гастингса позволяет построить марковскую цепь $m_1, \ldots, m_N$ со стационарным распределением $\mathcal{G}$.

Предложение

Если для марковской цепи выполняется закон больших чисел, то оценка $\widehat{p}_{IS}$ является состоятельной оценкой p.

Выбор весов w

Когда $g(m) \propto w(\mathrm{Score}(m))f(m)$, выбор плотности g сводится к выбору весов w.
Если значения функции Score дискретны, то выбор

$$w(x) \propto \frac{1}{\mathbb{P}(\mathrm{Score}(m) = x)}$$

уменьшит дисперсию $\widehat{p}$.

Алгоритм Ванга-Ландау (Iba *et al.*, 2014) является модификацией метода Метрополиса-Гастингса: в процессе моделирования цепи строится оценка весов.

Построение оценки $\widehat{p}$

1. Построение оценки $\widehat{w}$ алгоритмом Ванга–Ландау.
2. Построение марковской цепи со стационарным распределением $g(m) \propto \widehat{w}(\mathrm{Score}(m))f(m)$ алгоритмом Метрополиса–Гастингса.
3. Построение оценки

$$\widehat{p}_{IS} = \frac{\sum\limits_{i=1}^{N} \mathbb{I}_{\mathcal{S}}(m_i)/\widehat{w}(\mathrm{Score}(m_i))}{\sum\limits_{i=1}^{N} 1/\widehat{w}(\mathrm{Score}(m_i))}.$$

Теоретические результаты

На практике, вместо множества $\mathcal{M}$ можно рассматривать его дискретное подмножество: зафиксируем некоторое $m \in \mathcal{M}$ и рассмотрим вектора

$$\mathcal{M}_{d,m} = \{\tilde{m} = m + dv \mid \tilde{m} \in \mathcal{M}, \sum_{i=1}^{k} v_i = 0, v_i \in \mathbb{Z}\}.$$

Теорема

1. *Марковская цепь, построенная методом Метрополиса-Гастингса на множестве $\mathcal{M}_{d,m}$, является эргодической.*
2. *Марковская цепь, построенная методом Метрополиса-Гастингса на множестве $\mathcal{M}$, является Харрис-эргодической.*

Критерий остановки

Теорема (Flegal *et al.*, 2013)

Пусть $\widehat{\lambda}_N$ *– оценка дисперсии вдоль траектории* λ_p*:* $\widehat{\lambda}_N \xrightarrow{\text{п. н.}} \lambda_p$, $\widehat{\sigma}^2_N \xrightarrow{\text{п. н.}} \sigma^2_p$. *Предположим, что* $\sqrt{N}(\widehat{p}_N - p) \xrightarrow{d} \mathcal{N}(0, \sigma^2_p)$, $N \to \infty$.

Обозначим

- $N_\epsilon = \inf\left\{N > 0 : 2z_{\delta/2}\widehat{\sigma}_N/\sqrt{N} \le \epsilon\widehat{\lambda}_N\right\}$
- $C_N = (\widehat{p}_N - z_{\delta/2}\widehat{\sigma}^2_p/\sqrt{N}; \widehat{p}_N + z_{\delta/2}\widehat{\sigma}^2_p/\sqrt{N})$

Тогда при $N \to \infty$ *и* $\epsilon \to 0$ *моделирование прекратится с вероятностью* 1 *и* $\mathbb{P}(p \in C_{N_\epsilon}) \to 1 - \delta$ *при* $N \to \infty$.

Оценка дисперсии: методы

Оценка $\widehat{p}_{IS}$, полученная по траектории марковской цепи $m_1, \ldots, m_N$, имеет вид

$$\widehat{p}_{IS} = \frac{1}{N} \sum_{i=1}^{N} h(m_i).$$

Дисперсия такой оценки вычисляется как

$$\sigma_p^2 = \frac{1}{N} \sum_{k=-(N-1)}^{N-1} \left(1 - \frac{|k|}{N}\right) \mathrm{cov}(h(m_i), h(m_{i+k})).$$

Методы оценки дисперсии: метод перекрывающихся средних (Jones *et al.*, 2006), спектральные оценки (Hobert *et al.*, 2002).

Оценка дисперсии: рекурсивные методы

В предложенном алгоритме длина траектории N марковской цепи увеличивается последовательно, поэтому при использовании классических методов возникают проблемы

- хранения траектории,
- высокая трудоемкость.

Используется рекурсивная оценка по «правилу выбора трапеции» (Chan and Yau, 2014).

Свойства:

- вычислительная сложность рекурсивного пересчета $O(1)$,
- затрачивает $O(1)$ памяти.

Схема проведения экспериментов

Были построены:

- полученные оценки $\widehat{p}_{IS}$,
- оценки по методу Монте-Карло $\widehat{p}_{MC}$,
- оценки по алгоритму MS-DPR $\widehat{p}_{DPR}$,

для пептидов:

- простой структуры: линейных и циклических (Mohimani *et al.*, 2013),
- сложной, циклической с разветвлениями структуры (*Surfactin*).

Также для оценок $\widehat{p}_{IS}$ и $\widehat{p}_{MC}$ были сосчитаны оценки дисперсий $\widehat{\sigma}^2_{IS}$, $\widehat{\sigma}^2_{MC}$ и построены 95% доверительные интервалы.

Численные результаты: точность оценок

Для оценок по методу Монте-Карло, построенных с достаточно большим N:

1. оценки $\widehat{p}_{IS}$ лежат в доверительных границах оценок по методу Монте-Карло.
2. оценки $\widehat{p}_{DPR}$ в большинстве случаев выходят за границы доверительного интервала по методу Монте-Карло.

Численные результаты: сравнение дисперсий

Для оценок, построенных по одинаковому размеру выборки:

Пептид	$\widehat{p}_{IS}$	$\hat{\sigma}^2_{IS}$	$\hat{\sigma}^2_{MC}$	$\hat{\sigma}^2_{MC}/\hat{\sigma}^2_{IS}$
PPAEDSQK	$4.87 \cdot 10^{-7}$	$2.09 \cdot 10^{-10}$	$4.94 \cdot 10^{-7}$	2358.98
GQGDPGSNPNK	$4.70 \cdot 10^{-7}$	$2.33 \cdot 10^{-10}$	$1.49 \cdot 10^{-7}$	639.49
GEEEPSQGQK	$1.03 \cdot 10^{-6}$	$1.23 \cdot 10^{-9}$	$7.89 \cdot 10^{-7}$	642.19
Surfactin	$1.18 \cdot 10^{-5}$	$1.15 \cdot 10^{-7}$	$1.00 \cdot 10^{-5}$	86.96
$(10, 20, 40)$	$1.84 \cdot 10^{-3}$	$5.47 \cdot 10^{-4}$	$1.88 \cdot 10^{-3}$	3.43

Результаты демонстрируют, что отношение $\hat{\sigma}^2_{MC}/\hat{\sigma}^2_{IS}$ увеличивается с уменьшением значения оцениваемой вероятности p.

Выводы

1. Был предложен способ оценки статистической значимости меры похожести двух спектров.
2. Проведено сравнение полученного алгоритма с MS-DPR.
3. Его код был написан на C++ и интегрирован в Dereplicator — метод идентификации пептидных спектров.
4. Написана статья и подана на конференцию по алгоритмической биоинформатике (WABI, 2017).