Метод зацепления в задачах Метода Монте-Карло на Марковских Цепях

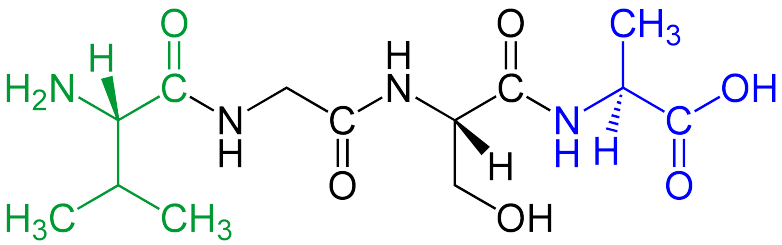
Мехнин Павел Владимирович, гр. 21.М03-мм

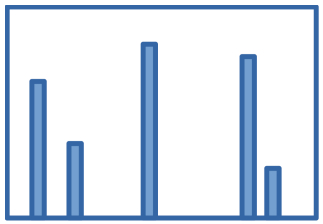
Санкт-Петербургский государственный университет
Прикладная математика и информатика

Научный руководитель: к.ф.-м.н. Шпилёв П. В.
Консультант: к.ф.-м.н. Коробейников А. И.
Рецензент: к.ф.-м.н. Гуревич А. А.

Санкт-Петербург, 2023

Идентификация пептидов





Выявление пептидов
с похожими свойствами

⇓

Обнаружение соединений
с аналогичной структурой

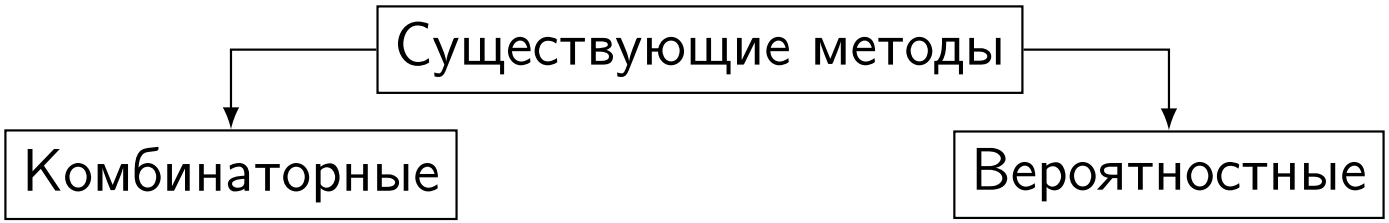
⇓

Получение экспериментального
спектра исследуемого образца
методом масс-спектрометрии

⇓

Поиск в базе данных спектра,
наиболее схожего
с экспериментальным, и
оценка этого сходства

Актуальность



- MS-GF+ (Kim et al., 2014)
- только для пептидов линейной структуры

- MS-DPR (Mohimani et al., 2013)
- оценки потенциально смещены

Предлагаемое решение: обобщение метода зацепления марковских цепей (Jacob et. al., 2020).

Цель работы: разработка алгоритма вычисления несмещённых оценок значимости совпадений пептидного спектра.

Задачи:

- эффективное построение марковских цепей;
- валидация работы для пептидов различной структуры.

Вероятностная модель спектра (Abramova et al., 2017)

Пусть P — пептид из k аминокислот общей массой M,
$\mu = (\mu_1, \ldots, \mu_k)$ — вектор масс аминокислот,
$\mathbb{H}$ — матрица фрагментации пептида,
$Score(\mu) = Score(S, \mathbb{H}\mu)$ — функция оценки сходства экспериментального S и ожидаемого $\mathbb{H}\mu$ спектров.

Предполагая, что μ равномерно распределён на множестве $\mathcal{M} = \{(\mu_1, \ldots, \mu_k) | \, \mu_i > 0, \sum_{i=1}^{k} \mu_i = M\}$, определение значимости совпадений спектра сводится к оценке вероятности

$$p = \mathbb{P}(Score(\mu) \geq r),$$

где r — заранее фиксированный порог.

Оценка по методу существенной выборки

Обозначим множество $\mathcal{A} = \{\nu \in \mathcal{M} : Score(\nu) \geq r\}$,
f — плотность равномерного распределения на множестве $\mathcal{M}$,
плотность $q(\nu) \propto w(Score(\nu))\, f(\nu)$.

Рассмотрим выборку $\nu_1, \ldots, \nu_N \sim q$.

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

$$\hat{p}_{IS} = \frac{\sum_{n=1}^{N} \mathbb{1}_{\{\nu_n \in A\}} / w(Score(\nu_n))}{\sum_{n=1}^{N} 1 / w(Score(\nu_n))}.$$

Идея метода зацепления (Jacob et al., 2020)

Определение

Парой сцепленных марковских цепей с пространством состояний $\mathcal{X}$ и стохастическим ядром $P(\cdot,\cdot)$ называется марковская цепь $Z_t = (X_t, Y_t)$ с пространством состояний $\mathcal{X} \times \mathcal{X}$, такая что:

$$\mathbb{P}\{X_{t+1} = x' | Z_t = (x,y)\} = \mathbb{P}\{X_{t+1} = x' | X_t = x\} = P(x,x'),$$
$$\mathbb{P}\{Y_t = y' | Z_{t-1} = (x,y)\} = \mathbb{P}\{Y_t = y' | Y_{t-1} = y\} = P(y,y').$$

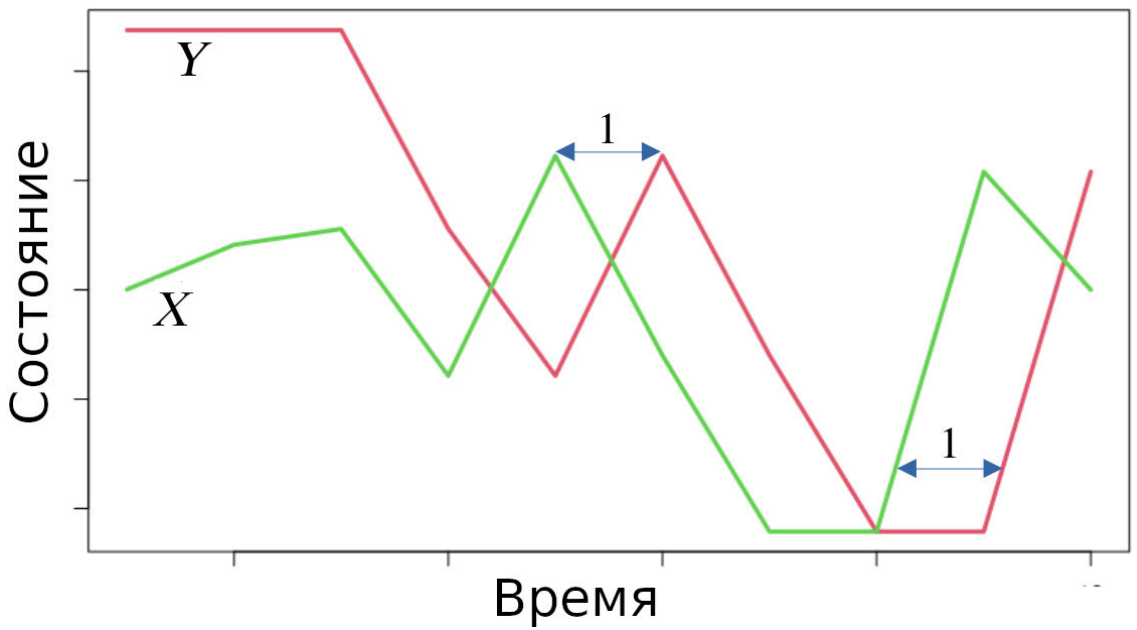


Рис. 1: Траектории сцепленных марковских цепей

Несмещённая оценка

С помощью выборки $\{(X_t, Y_{t-1})|t = 1, 2, ...\}$ из сцепленных цепей Маркова можем вычислить несмещённую оценку

$$H_i = h(X_i) + \sum_{t=i+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\},$$

где h — индикатор множества $\mathcal{A}$,
$\tau = \inf\{t \geq 1 : X_t = Y_{t-1}\}$ — момент зацепления цепей.

Алгоритм моделирования сцепленных цепей

Алгоритм 1 Построение сцепленных марковских цепей

Вход: число итераций m, оценки весов w
Выход: выборка $\{(X_t, Y_{t-1}) | t = 1, 2, \dots\}$
$X_1 \leftarrow \nu_x$, $Y_0 \leftarrow \nu_y$, $t \leftarrow 1$
while $t < \max(m, \tau)$, где $\tau = \inf\{t \geq 1 : X_t = Y_{t-1}\}$ **do**
    $\nu^* \sim \gamma(\cdot|\nu)$
    $\alpha_x \leftarrow \min\left[1, \frac{w(Score(\nu^*))}{w(Score(\nu_x))}\right]$, $\alpha_y \leftarrow \min\left[1, \frac{w(Score(\nu^*))}{w(Score(\nu_y))}\right]$
    $u \sim U[0, 1]$
    if $u \leq \alpha_x$ **then**
        $\nu_x \leftarrow \nu^*$, $X_{t+1} \leftarrow \nu^*$
    else
        $X_{t+1} \leftarrow X_t$
    if $u \leq \alpha_y$ **then**
        $\nu_y \leftarrow \nu^*$, $Y_t \leftarrow \nu^*$
    else
        $Y_t \leftarrow Y_{t-1}$
    $t \leftarrow t + 1$

Несмещённая оценка

Схема вычисления несмещённой оценки $\hat{p}_C$:

1. Выбор весов $\hat{w}(s)$ алгоритмом Ванга–Ландау (Iba et al., 2014).

2. Построение **сцепленных** марковских цепей со стационарным распределением $q(\nu) \propto \hat{w}(Score(\nu))\, f(\nu)$ **модифицированным** алгоритмом Метрополиса-Гастингса.

3. Вычисление **несмещённой** оценки

$$\hat{p}_C = \frac{\sum_{i=1}^{n} H_i \,/\, w(Score(\nu_i))}{\sum_{i=1}^{n} 1 \,/\, w(Score(\nu_i))}$$

Уменьшение дисперсии

Несмещённые оценки можно усреднить,
чтобы уменьшить дисперсию оценок

$\Downarrow$

Промоделируем параллельно множество цепей и
объединим результаты независимых вычислений

$\Downarrow$

Для уменьшения дисперсии оценки
отбросим первые k элементов цепи
(0.99-квантиль распределения времени зацепления τ)

$\Downarrow$

Для уменьшения «бесполезных» вычислений
ограничим число итераций m как кратное k

Схема проведения экспериментов

- Для пептидов различной структуры были вычислены:
  1. оценки по методу существенной выборки $\hat{p}_{IS}$
  2. оценки по методу зацепления $\hat{p}_C$

- Для оценок построены 95% доверительные интервалы. Оценки дисперсий вычислены по рекурсивному методу TSR (Yau et al., 2016).

- Выполнено сравнение смещения оценок $b = \hat{p} - p$ от ожидаемого значения.

Оценки и их доверительные интервалы

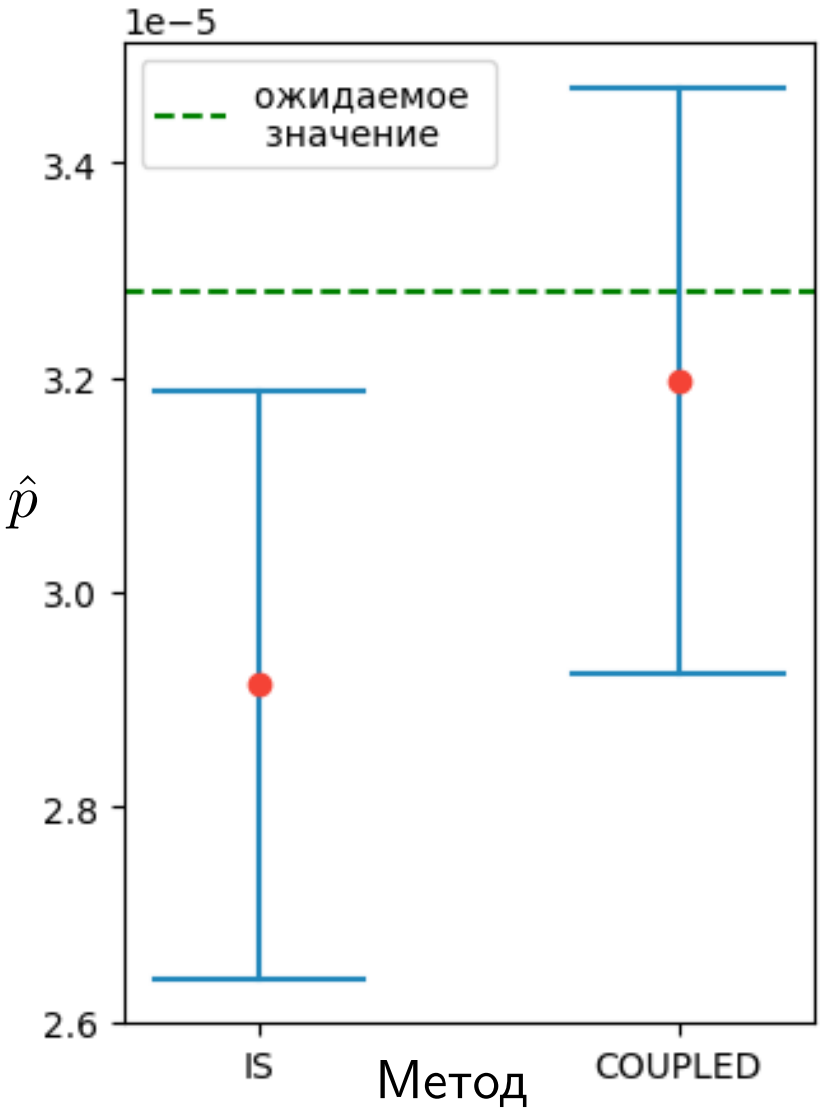


Рис. 2: GPDGPEEK

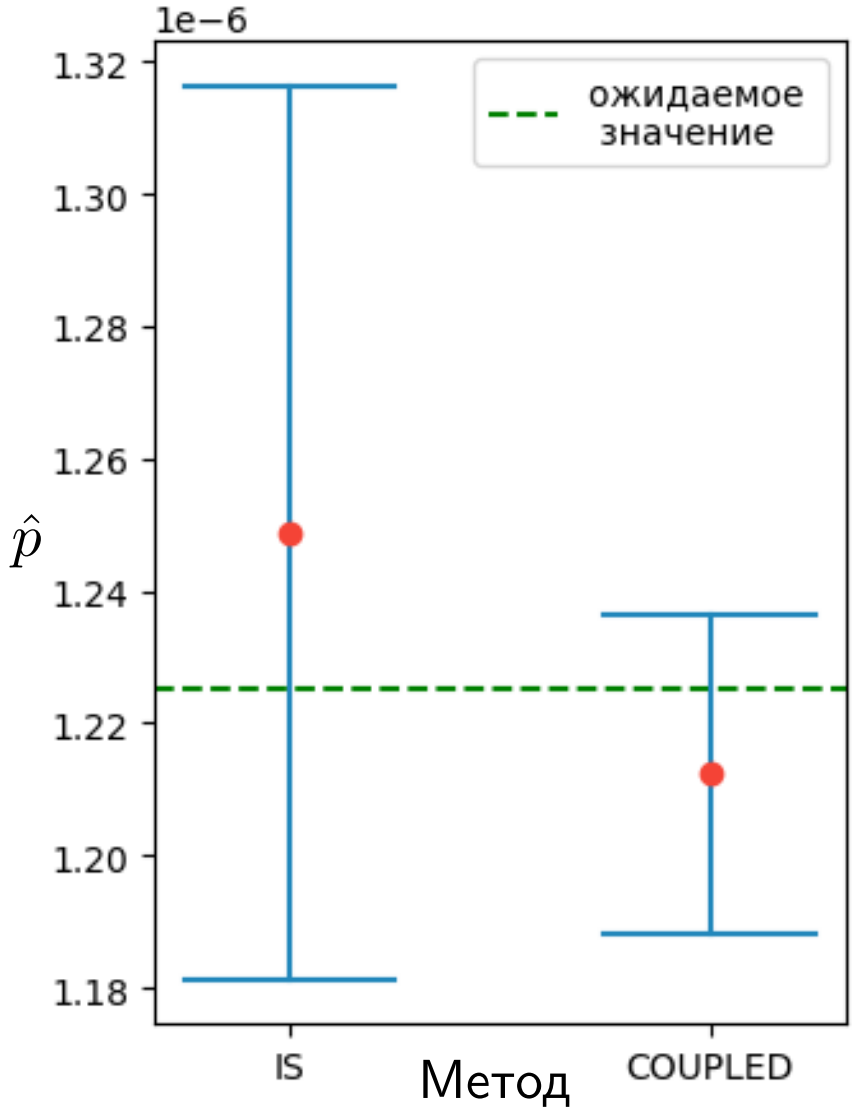


Рис. 3: PPAEDSQK

Оценки и их доверительные интервалы

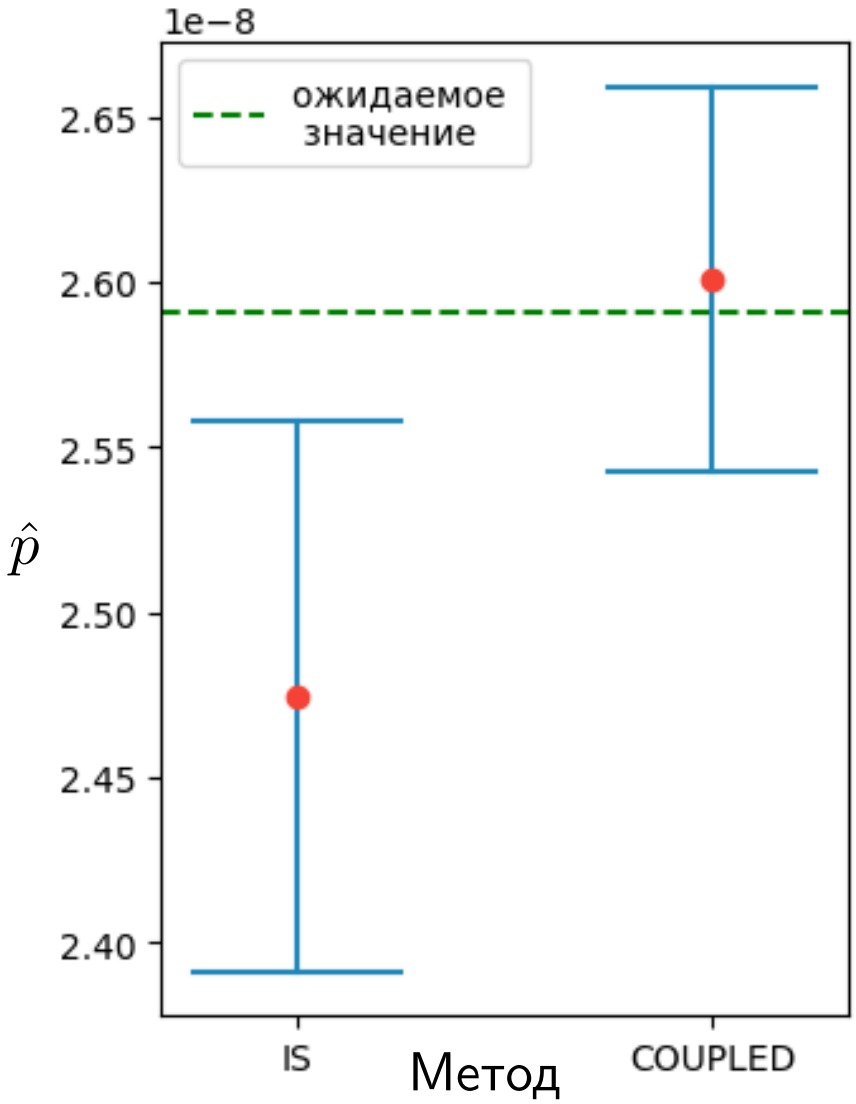


Рис. 4: (10,20,40,80,160)

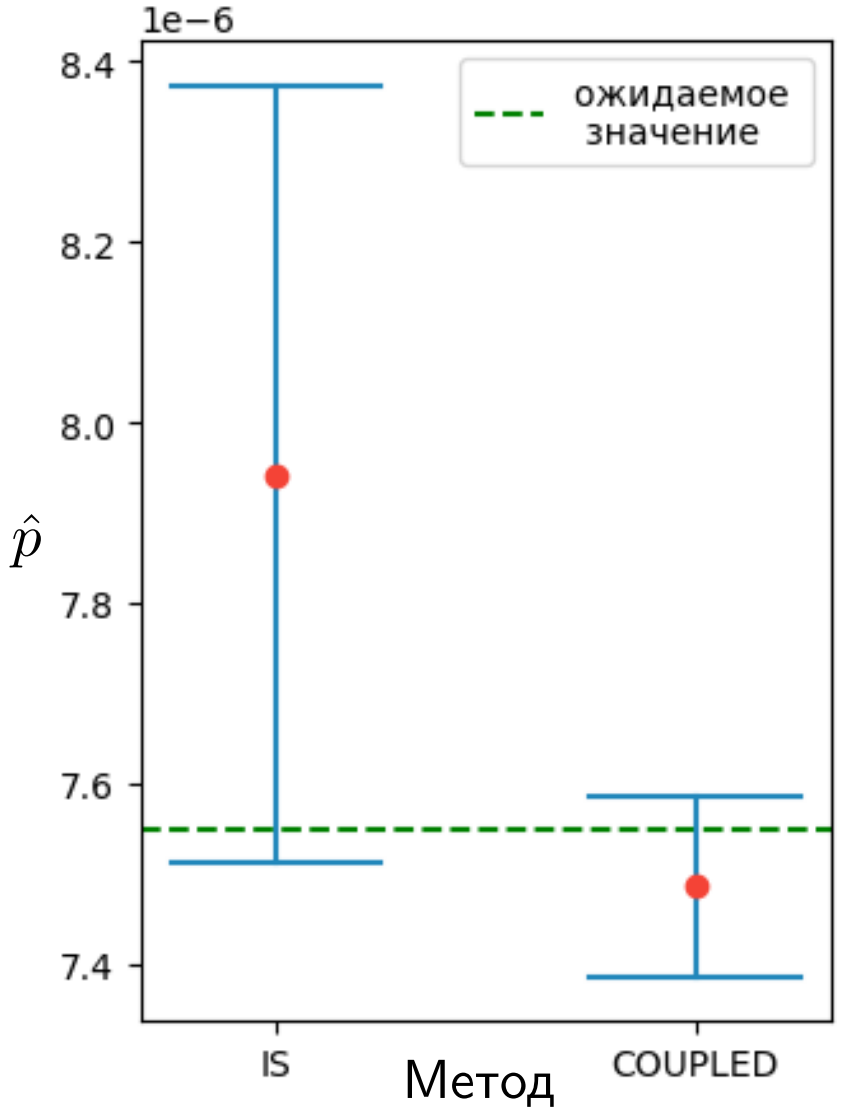


Рис. 5: *Surfactin*

Численные результаты

Таблица 1: Сравнение смещения оценок

Пептид	$\lvert b_{IS}\rvert$	$\lvert b_C\rvert$	$\frac{\lvert b_{IS}\rvert}{\lvert b_C\rvert}$
GPDGPEEK	11.2%	2.55%	4.38
GEEEPSQGQK	8.07%	0.70%	11.5
PPAEDSQK	1.91%	1.07%	1.79
(10,20,40,80)	4.48%	0.40%	11.2
(10,20,40,80,160)	5.43%	0.47%	11.6
Surfactin	5.22%	0.82%	6.36

Выводы

- В работе исследован метод зацепления, позволяющий уменьшить смещение в оценках, полученных с помощью алгоритмов MCMC.
- Разработан алгоритм вычисления несмещённых оценок статистической значимости совпадений спектра пептидов с использованием сцепленных марковских цепей.
- Проведено сравнение полученного алгоритма с методом существенной выборки.
- Эмпирически показано, что подход довольно общий и потенциально может применяться к пептидам различной структуры.
- Метод открывает простор для масштабирования посредством многопоточных приложений или облачной контейнеризации.