Оценка статистической значимости «схожести» пептидных спектров

Небожатко Екатерина Павловна

Институт биоинформатики

Научный руководитель — к.ф.-м.н. **А.И. Коробейников**

Санкт-Петербург
2018г.

Постановка задачи

- Есть база данных спектров пептидов (теоретических спектров) и экспериментальный спектр, полученный в результате применения метода масс-спектрометрии
- Задаем функцию Score, вычисляющую меру «схожести» двух спектров

Нашей задачей является вычисление вероятности

$$p^* = \mathbb{P}(\text{Score}(Spectrum, peptide) \geq s)) = \mathbb{P}(peptide \in \mathcal{S}).$$

Здесь peptide:

$$pepdie = \{\vec{\mu} = (\vec{\mu}_1, \ldots, \vec{\mu}_k), |\vec{\mu}_i > 0, \sum_{i=1}^{k} \vec{\mu}_i = M\} \tag{1}$$

Метод Монте-Карло

Пусть $(x_1, \dots x_n)$ — равномерно распределенная выборка пептидов с распределением $\mathcal{P}$.

Определение

Оценкой по методу Монте-Карло будем называть оценку вида

$$\widehat{p}_{MC} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(x_i).$$

Дисперсия такой оценки $\mathbb{D}(\hat{p}_{MC}) = \frac{p(1-p)}{n}$ стремится к 0 при n $\to \infty$, однако относительная ошибка возрастает с уменьшением p

$$\mathrm{re}(\hat{p}_{MC}) = \frac{\mathbb{D}(\hat{p}_{MC})}{p^2} = \frac{p(1-p)}{np^2} = \frac{1}{np} - \frac{1}{n} \to \infty, \; p \to 0$$

Метод существенной выборки

Рассмотрим некоторое распределение $\mathcal{Q}$ с плотностью $q(x)$. Пусть $p(x)$ — плотность распределения $\mathcal{P}$.

Пусть $y_1, \ldots, y_n$ независимые одинаково распределенные случайные величины с распределением $\mathcal{Q}$.

Определение

Оценкой по методу существенной выборки будем называть оценку

$$\widehat{p}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \frac{p(y_i)}{q(y_i)} \mathbb{I}_{\mathcal{S}}(y_i).$$

Метод существенной выборки

Как выбрать $q(x)$?

Если $q(x) \propto w(x)p(x)$, тогда:

$$\widehat{p}_{IS} = \frac{\sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(y_i)/w(y_i)}{\sum_{i=1}^{n} 1/w(y_i)}$$

При таком выборе $q(x)$ оценка зависит только от весов w, и не зависит от, вообще говоря, неизвестной плотности $p(x)$.

Алгоритм Ванга–Ландау (Landau, Wang, 2004)

За счет выбора весов w можно уменьшить дисперсию оценки $\widehat{p}_{IS}$.

Как выбрать веса?

Если истинные значения вероятностей p известны и функция Score дискретна, оптимальные веса находятся из соотношения:

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

Описание алгоритма

1. Строится оценка $\widehat{w}$ методом Ванга–Ландау
2. Моделируется марковская цепь со стационарным распределением $q(x) \propto \widehat{w}(\mathrm{Score}(x))p(x)$

Определение

Оценкой по методу Ванга–Ландау будем называть оценку

$$\widehat{p}_{WL} = \frac{\sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(x_i)/\widehat{w}(\mathrm{Score}(x_i))}{\sum_{i=1}^{n} 1/\widehat{w}(\mathrm{Score}(x_i))}.$$

Replica Exchange (Geyer, 1991)

Replica exchange выбирает веса $w = e^{\beta \cdot \mathrm{Score}(x)}$ из сетки значений.

Описание алгоритма

1. Параллельно моделируются k цепей с различными значениями параметров $(\beta_1 \dots \beta_k)$ и распределением $q_t(x) \propto e^{\beta_t \cdot \mathrm{Score}(x)} p(x)$;
2. Через r итераций выбирается пара цепей i и j для обмена состояниями. Обмен происходит с некоторой вероятностью
3. Для каждой цепи строится оценка

$$\widehat{p}_{RE}^{(j)} = \frac{\sum_{i=1}^{n} \mathbb{I}_{\mathcal{S}}(x_i) \exp(-\beta_j \cdot \mathrm{Score}(x_i))}{\sum_{i=1}^{n} \exp(-\beta_j \cdot \mathrm{Score}(x_i))}.$$

Определение

Оценкой по методу replica exchange будем называть

$$\widehat{p}_{RE} = \frac{1}{k} \sum_{j=1}^{k} \widehat{p}_{RE}^{(j)}.$$

Stochastic Approximation Monte Carlo (Liang et al., 2007)

Разобьем выборочное пространство на l областей:
$E_1 = \{x : \text{Score}(x) \leq s_1\}, E_2 = \{x : s_1 < \text{Score}(x) \leq s_2\}, \ldots,$
$E_l = \{x : \text{Score}(x) > s_l\}$. Пусть $\psi(x)$ — некоторая неотрицательная функция и $g_i = \int_{E_i} \psi(x)dx$.

Пусть $\theta_t^{(i)}$ обозначает оценку $\log(g_i/\pi_i)$, полученную на итерации t.

Теорема

Если положить $\psi(x) \propto 1$, тогда g_i — количество элементов выборки, принадлежащих E_i, и оценка p определяется как

$$\widehat{p}_{SAMC_t} = \mathbb{P}_t(\text{Score}(\xi) > s_k) = \frac{\sum_{i=k+1}^{l} \exp(\theta_t^{(i)})(\pi_i + 1)}{\sum_{j=1}^{l} \exp(\theta_t^{(j)})(\pi_j + 1)},$$

и сходится к $\mathbb{P}(\text{Score}(\xi) > s_k)$ для достаточно больших t. Здесь $s_k = s^*$.

Схема проведения экспериментов

Были построены:

- оценки по методу Монте-Карло $\widehat{p}_{MC}$,
- оценки по методу Ванга–Ланду $\widehat{p}_{WL}$,
- оценки по методу replica exchange $\widehat{p}_{RE}$,
- оценки по методу stochastic approximation monte carlo $\widehat{p}_{SAMC}$.

Все оценки были получены четырех выбранных пептидов. $n = 10^7$. Также для всех оценок были сосчитаны оценки дисперсий по методу batch means и построены 95% доверительные интервалы.

Оценки и их доверительные интервалы

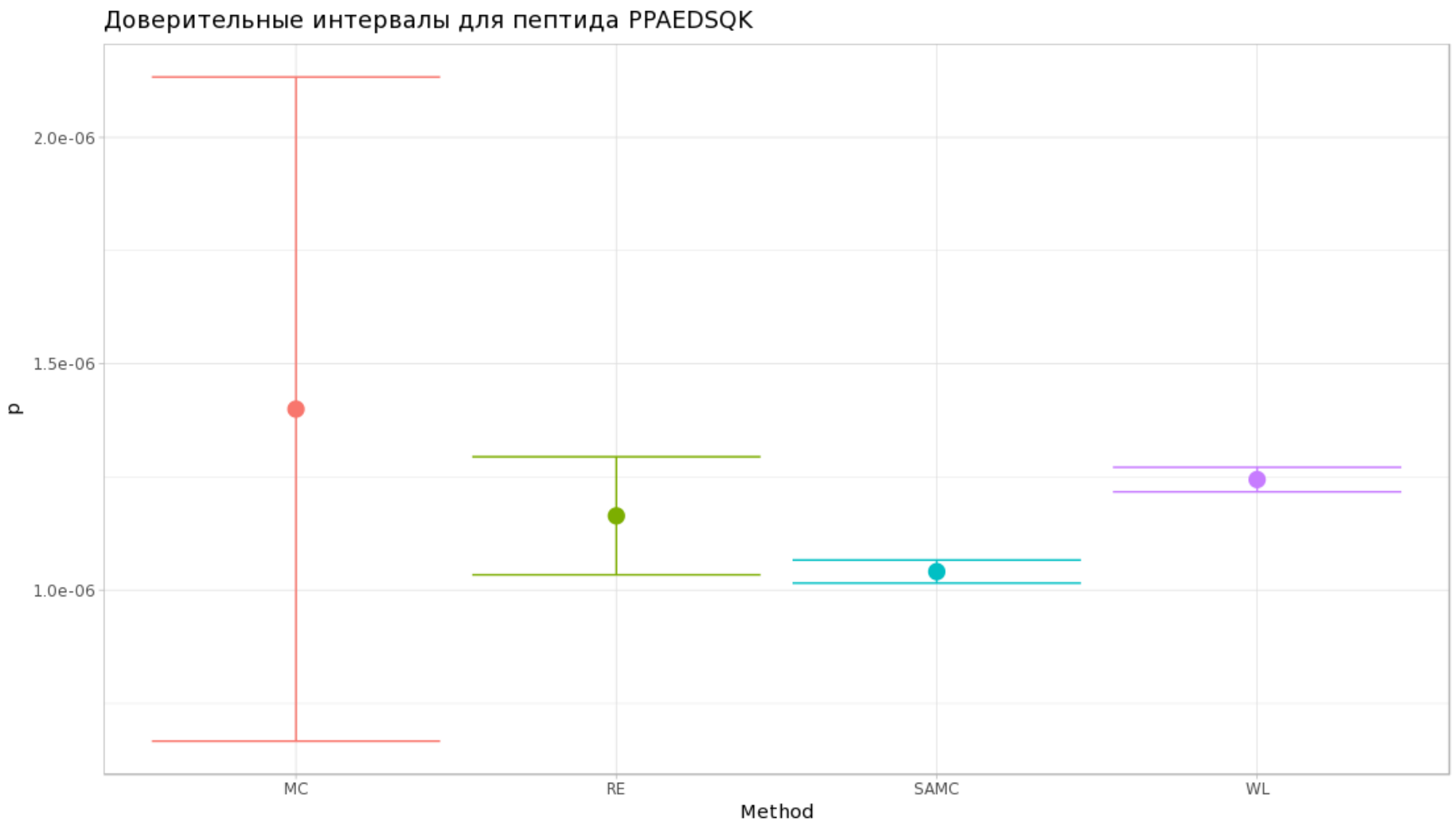

Рис. : PPAEDSQK

Оценки и их доверительные интервалы

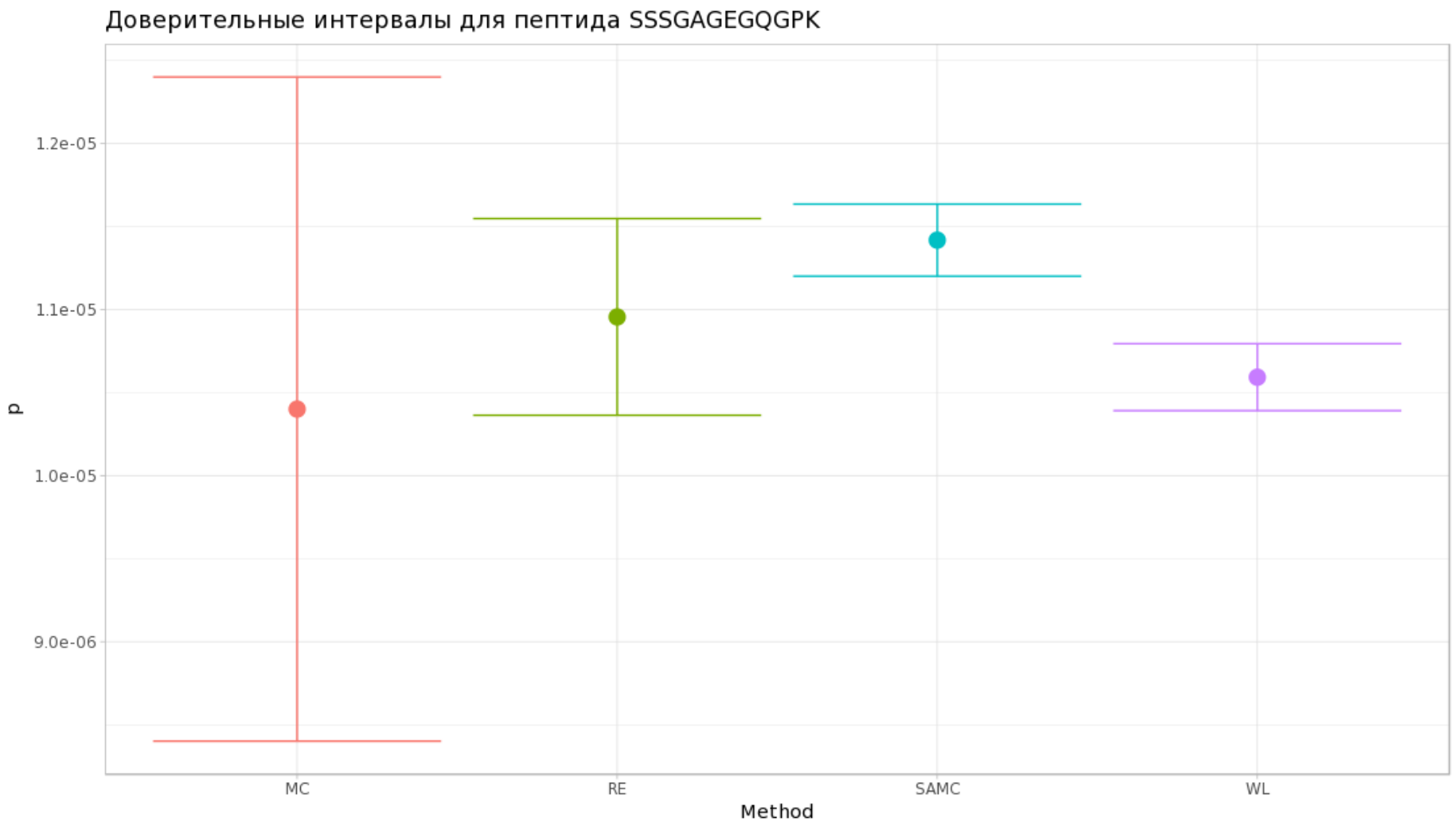

Рис. : SSSGAEGQGPK

Оценки и их доверительные интервалы

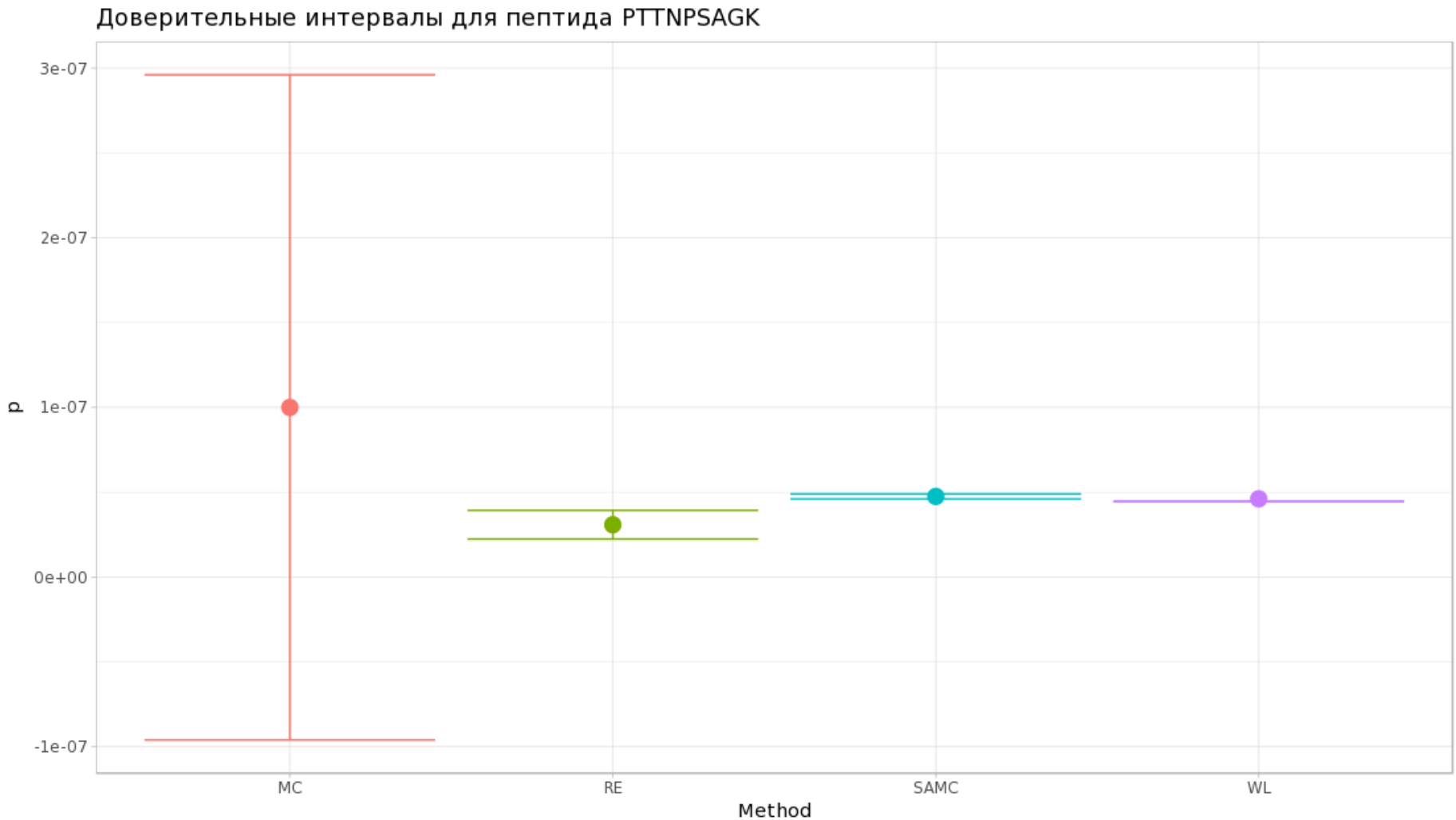

Рис. : PTTNPSAGK

Оценки и их доверительные интервалы

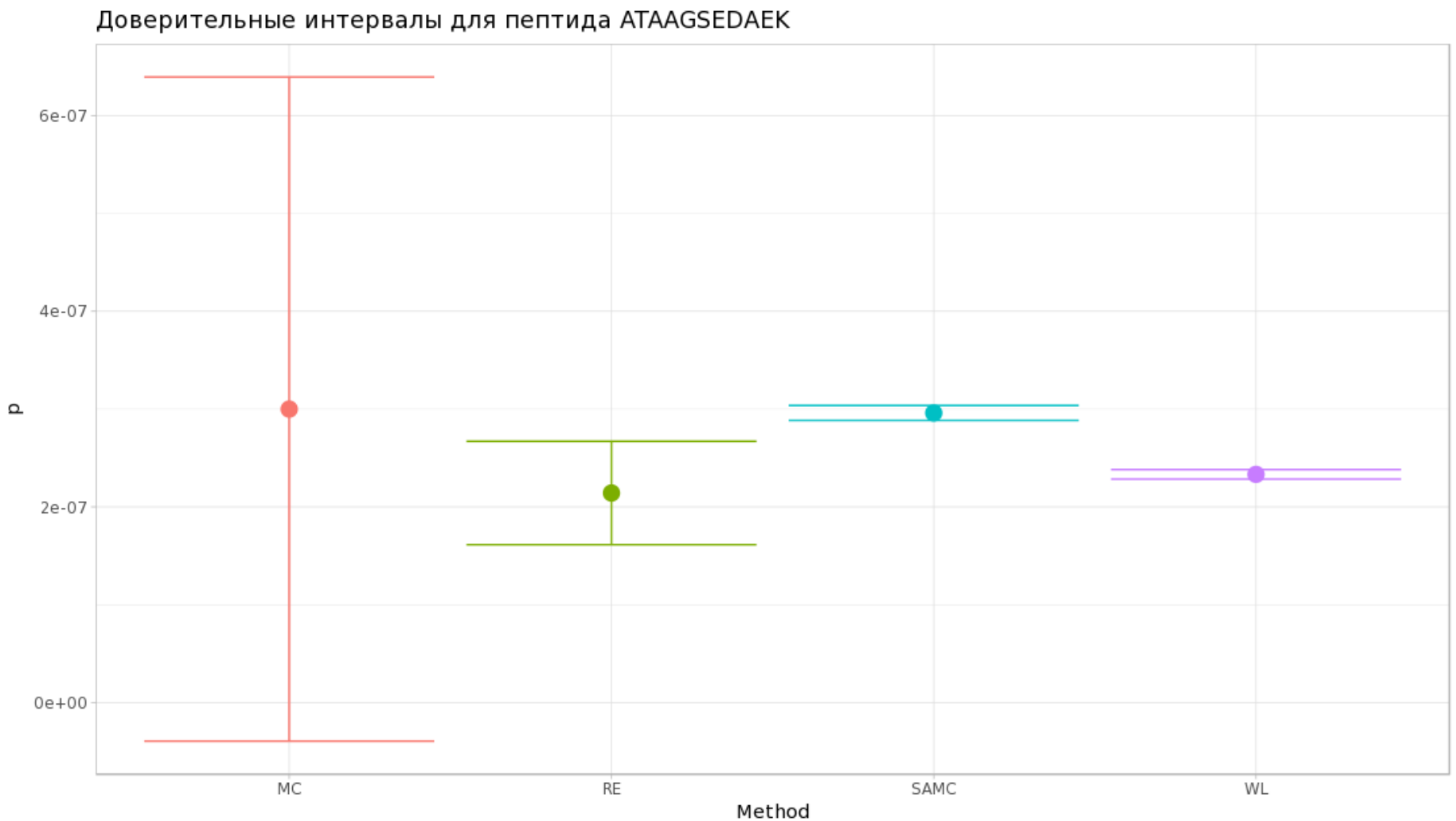

Рис. : ATAAGSEDAEK

Численные результаты

Таблица : Сравнение относительных ошибок (re) оценок

Пептид	$\mathrm{re}(\hat{p}_{MC})$	$\mathrm{re}(\hat{p}_{WL})$	$\mathrm{re}(\hat{p}_{RE})$	$\mathrm{re}(\hat{p}_{SAMC})$
SSSGAGEGQGPK	30.4	0.29	2.4	0.29
PPAEDSQK	225.87	0.39	10.32	0.49
ATAAGSEDAEK	1054.09	0.35	50.08	0.55
PTTNPSAGK	3162.27	0.95	61.98	0.81

Относительная ошибка всех трех методов значительно меньше, чем у Монте-Карло. И эта разница тем больше, чем меньше значение оцениваемой вероятности.

Выводы

- Эмпирически показано, что оценки, полученные с помощью данных алгоритмов, лежат в границах доверительных интервалов оценок Монте-Карло и имеют меньшую дисперсию.
- Наименьшая дисперсия была достигнута для оценок вероятностей по методам Ванга-Ландау и stochastic approximation Monte Carlo.

Github https://github.com/KateNebo/BioInf-Project18