Задачи анализа спектров тандемной масс-спектрометрии

Иванова Елизавета Владимировна, гр. 15.M03-мм

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

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

Санкт-Петербург
2017 г.

Базовые определения

- **Масс-спектрометрия** — техника, которую используют для определения химического состава веществ.
- **Масс-спектр** — сигнал, представляющий зависимость количества ионов вещества (интенсивность) от m/z.

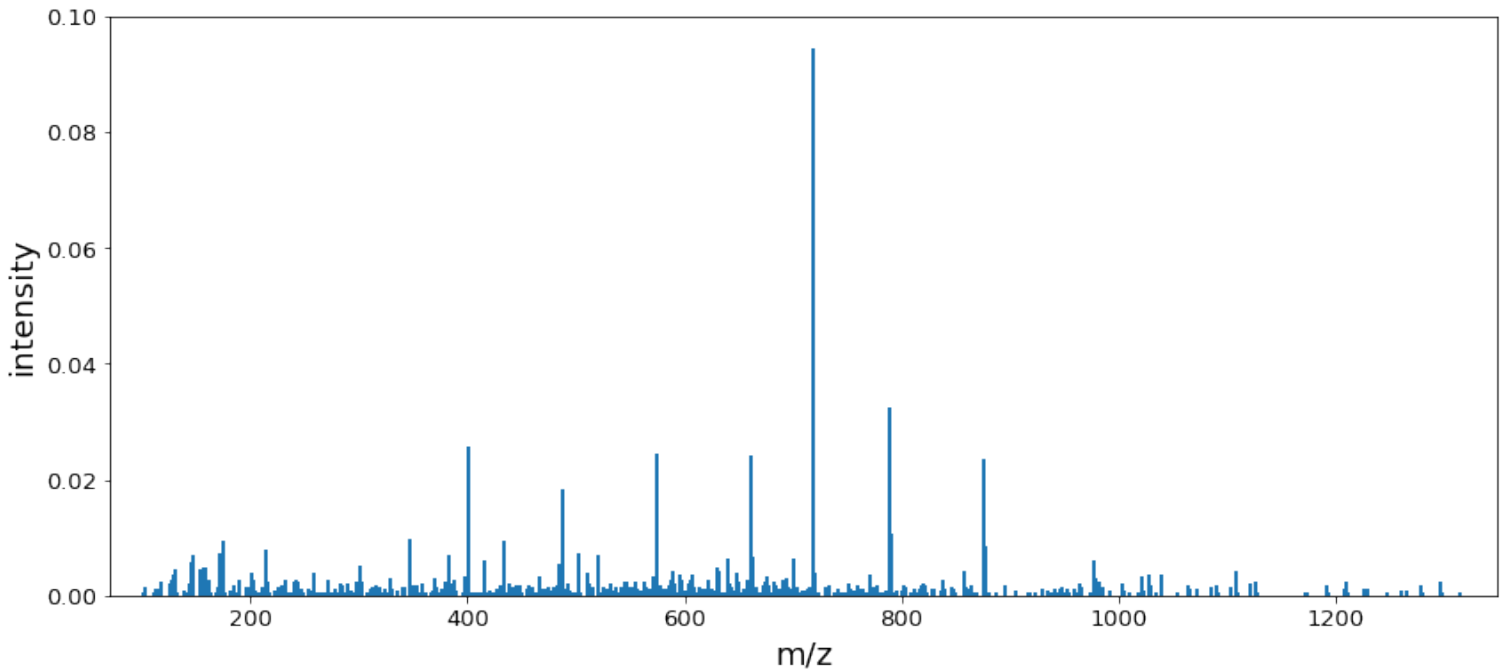


Рис. : Пример масс-спектра.

Задача фильтрации спектров

Свойства эмпирических спектров

- Имеются лишние (шумовые) пики,
  а интенсивности могут быть искажены трендом.
- Некоторые пики могут сливаться или разделяться на несколько пиков.

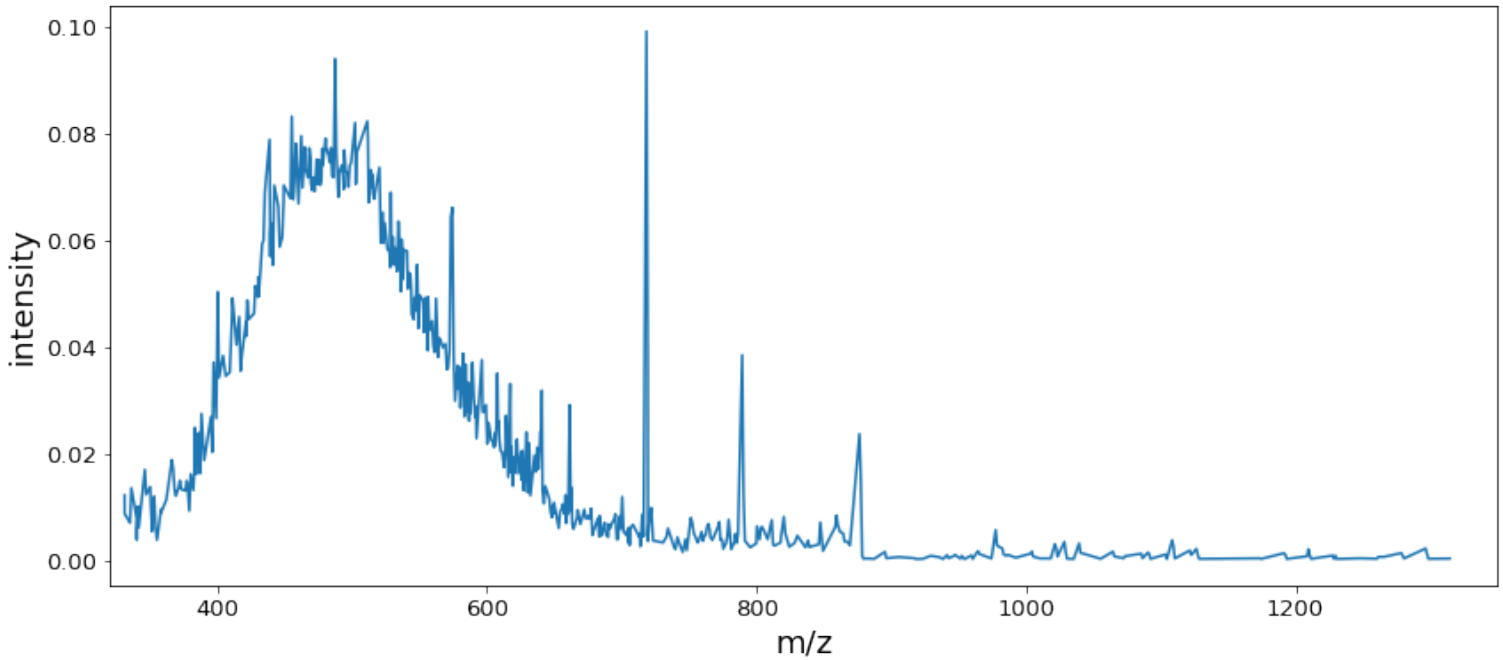


Рис. : Пример масс-спектра с выраженным трендом.

Фильтрация пиков

- Фильтрация является важной процедурой, так как лишние пики и артефакты влияют на результаты дальнейшего анализа.
- Задача фильтрации хорошо изучена, так как масс-спектры используются в различных областях знаний.

Нас будет интересовать задача фильтрации в контексте потоковой обработки большого количества спектров.

Фильтрация масс-спектров в Дерепликаторе

В работе был рассмотрен случай **Дерепликатора** (H. Mohimani *et al*, 2016), одного из алгоритмов идентификации пептидов.

Его процедура фильтрации имеет ряд недостатков, самый важный из них — добавление новых пиков.

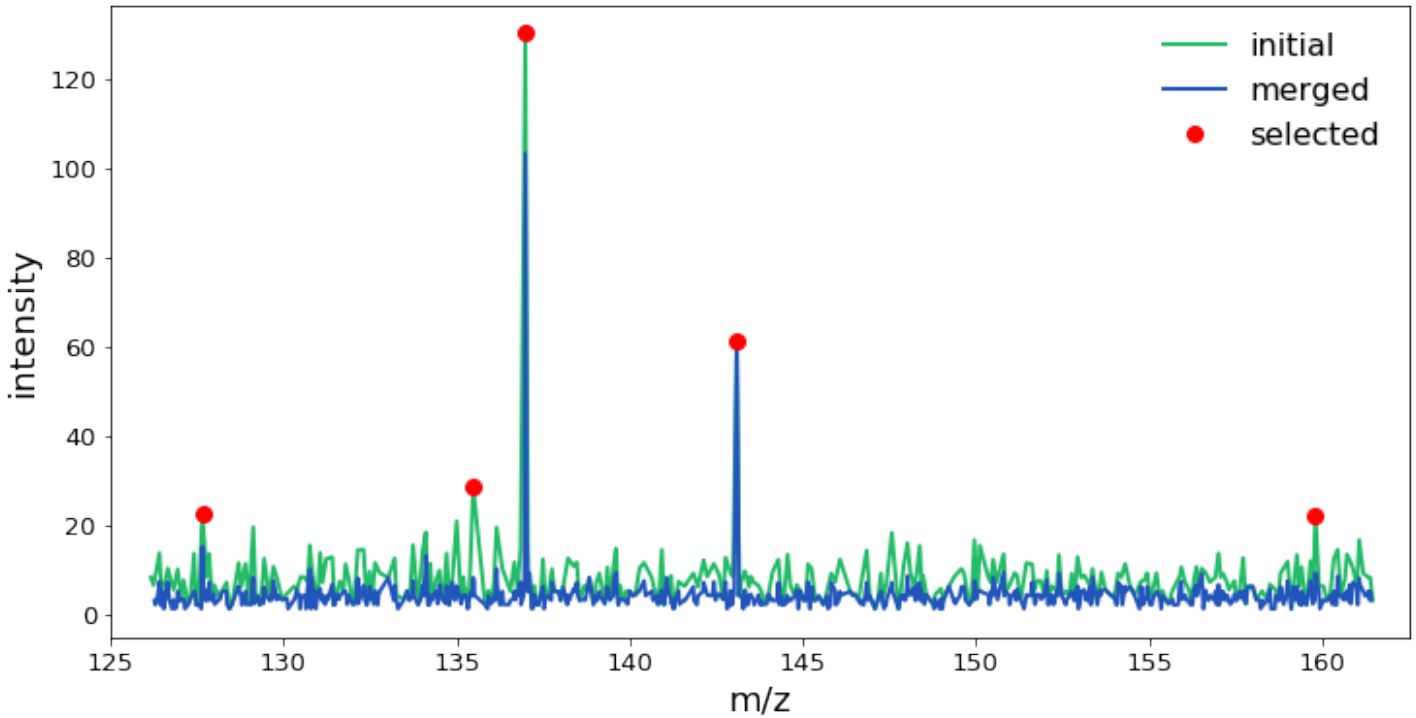


Рис. : Пример обработки спектра с помощью фильтрации в Дерепликатор.

Постановка задачи

Задача:

- Изучить методы фильтрации пиков.
- Выбрать наиболее подходящий согласно критериям:
  - Наличие небольшого количества настраиваемых параметров;
  - Параметры легко интерпретировать.
- Интегрировать решение в Дерепликатор.

Методы фильтрации пиков

- Большинство методов состоят из трех шагов — сглаживание, выделение тренда и определение пиков.
- Были рассмотрены два алгоритма:
  - PROcess (X. Li *et al.*, 2005) из Bioconductor;
  - MassSpecWavelet (P. Du *et al.*, 2006) из Bioconductor.
- Оба алгоритма оценивают сигнал и шум в каждой точке спектра, а затем отбираются пики, для которых частное оценок больше некоторого порога.

- $\psi(t)$ — материнский вейвлет,
  $\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$ — вейвлеты.
- Модель эмпирического масс-спектра:
  $Y(t) = S(t) + B(t) + E(t)$, где $S(t)$ — искомый спектр,
  $B(t)$ — тренд, $E(t)$ — шум с нулевым средним;
- Коэффициенты вейвлет-преобразования для $Y(t)$:

$$C(a,b) = \int_{\mathbb{R}} \psi_{a,b}(t)Y(t)dt$$

$$= \int_{\mathbb{R}} \psi_{a,b}(t)S(t)dt + \int_{\mathbb{R}} \psi_{a,b}(t)B(t)dt + \int_{\mathbb{R}} \psi_{a,b}(t)E(t)dt.$$

- Идея метода основана на предположении

$$\int_{\mathbb{R}} \psi_{a,b}Y\,dt \approx \int_{\mathbb{R}} \psi_{a,b}S dt.$$

Замечания

- MassSpecWavelet неявно использует предположение о равноотстоящих отсчетах в спектре.
- Существует модификация MassSpecWavelet (French *et al*, 2015), подходящая для неравноотстоящих m/z.
- В этой работе модификация была улучшена:
  - Новый алгоритм корректно обрабатывает короткие спектры;
  - Параметры адаптированы под обработку спектров любой длины.

Сравнение алгоритмов

1) Эксперимент с длинными масс-спектрами, для которых неизвестны пептиды. Длины спектров от 8000 до 15000.

- Назовем пик **сомнительным**, если в его окрестности находится пик с большей интенсивностью.
- Дерепликатор: 38% сомнительных пиков.
  Предлагаемый алгоритм: 21% сомнительных пиков.

2) Эксперимент с масс-спектрами, для которых известны пептиды. Длины спектров от 600 до 700.

	Точность*	Полнота*
Дерепликатор	0.31	0.58
Предлагаемый алгоритм	0.37	0.62

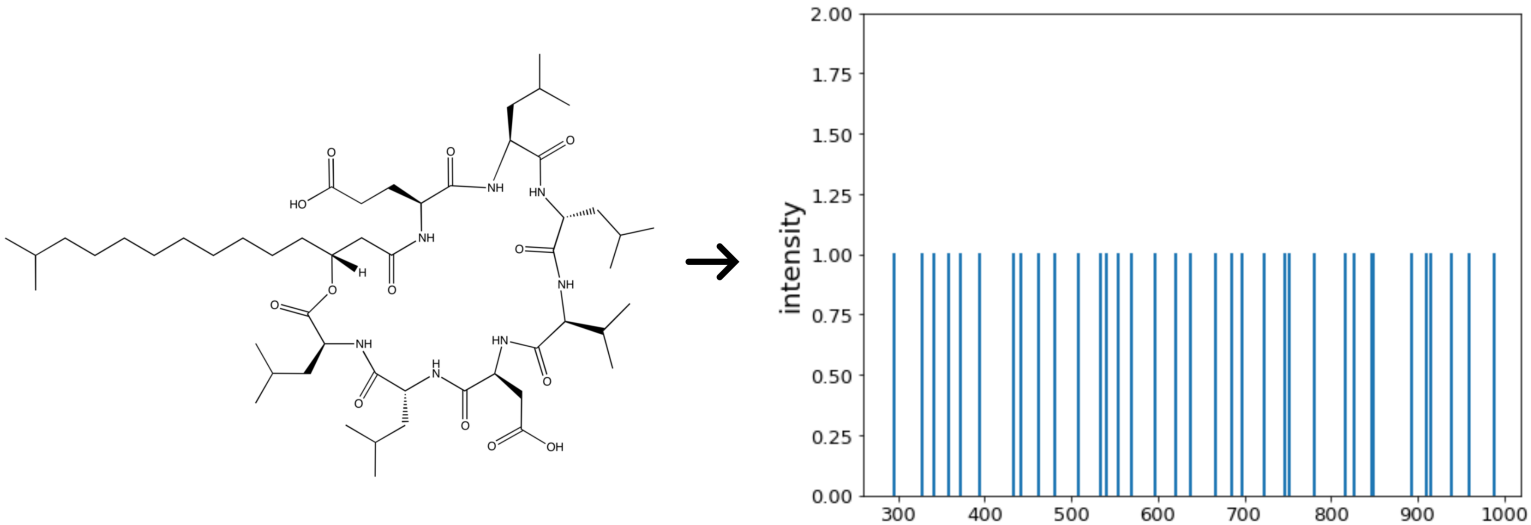
*Усреднено по спектрам.

Задача кластеризации спектров

Идентификация пептидов

- Масс-спектрометрия является основным инструментом определения пептидов.

- Большинство методов идентификации пептидов по масс-спектру используют базу данных пептидов.
  Для пептида из базы данных строится ожидаемый спектр и сравнивается с наблюдаемым.



Ожидаемый спектр

- $\mathcal{A}$ — алфавит аминокислот;
- $m(p) : \mathcal{A} \to (0, +\infty)$ — масса аминокислоты;
- $P = p_0 p_1 \ldots, p_{n-1}$ — строка в алфавите $\mathcal{A}$, пептид.

Ожидаемый спектр — это набор масс

$$b(P,k) = \sum_{i=0}^{k-1} m(p_i), \;\; y(P,k) = \sum_{i=k}^{n-1} m(p_i), \;\; k = 1, \ldots, n.$$

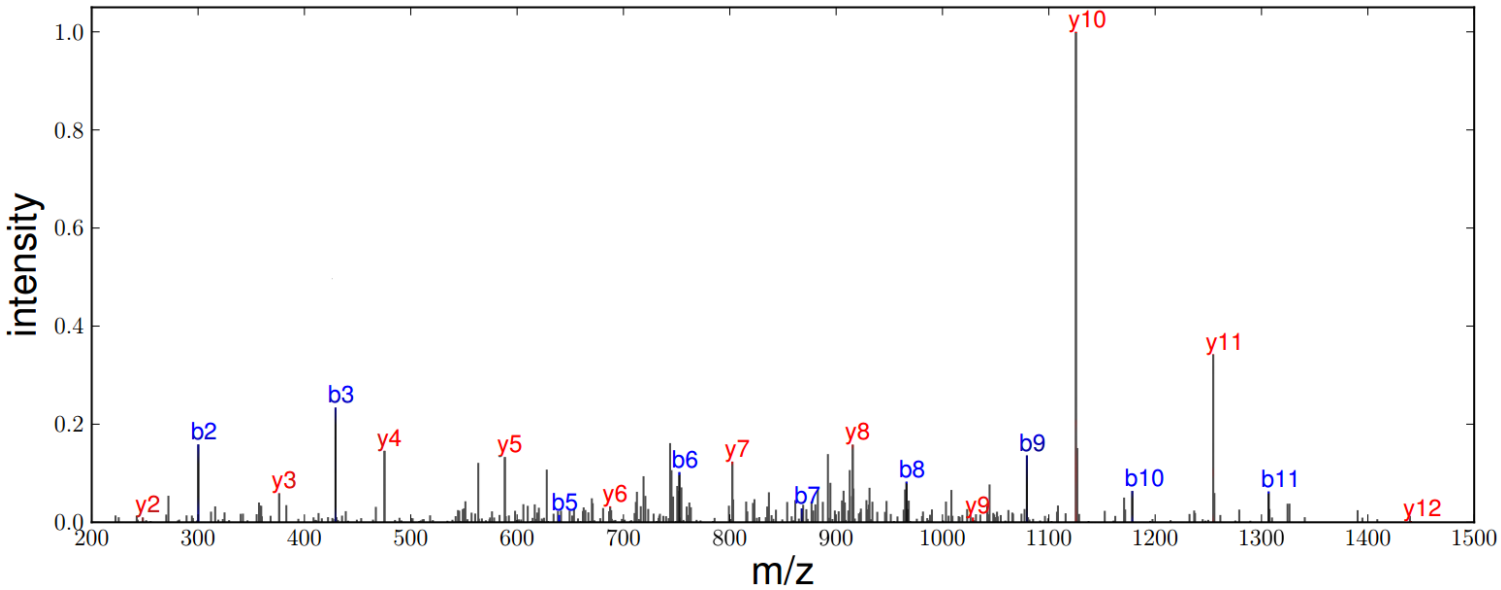


Рис. : Пример эмпирического масс-спектра. Синим и красным цветами отмечены пики, относящиеся к ожидаемому спектру.

Ускорение процедуры идентификации пептидов

- Для идентификации пептида по масс-спектру требуется перебрать все ожидаемые спектры, построенные по базе данных.
- Если N — размер базы данных, а M — количество эмпирических спектров, то общее время работы — $O(MN)$.

Идея ускорения процедуры — кластеризовать ожидаемые спектры.

Постановка задачи

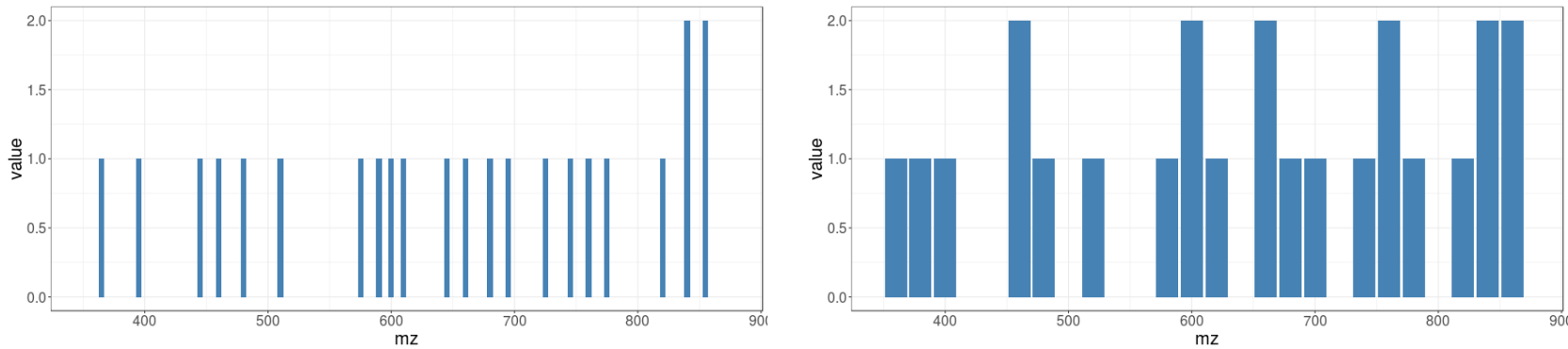
Задача:

Научиться кластеризовать ожидаемые спектры, в том числе:

- Выбрать расстояние между пептидами как строками.
- Подобрать подходящее векторное представление спектра.
- Выбрать меру близости между спектрами, которая вычислялась бы за разумное время.

Мера близости биологических последовательностей

- Будем измерять близость пептидов как оценку выравнивания с матрицей замены BLOSUM62 (Henikoff *et al*, 1992).
- Векторное представление спектров — гистограмма с определенной шириной столбцов.



(a) bin length = 5

(b) bin length = 20

Рис. : Пример гистограмм с шириной столбца bin length.

Расстояние между спектрами

Пусть $R = (r_0, \ldots, r_{n-1})$ и $Q = (q_0, \ldots, q_{n-1})$ — гистограммы, m — суммарное число их ненулевых столбцов.

Рассмотренные расстояния:

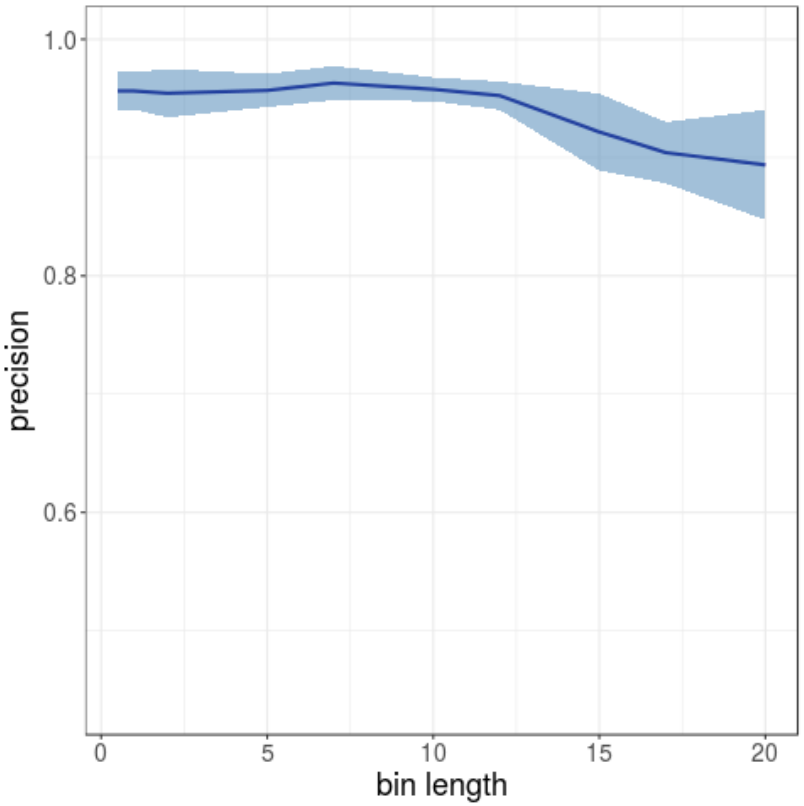
1) Синус угла между векторами, временная сложность — $O(n)$;
2) Модификация Earth Mover's Distance (Pele and Werman, 2008), временная сложность — $O(m^2)$;
3) Quadratic-Chi Histogram Distance Family (Pele and Werman, 2010), временная сложность — $O(m^2)$.

Вычислительные эксперименты

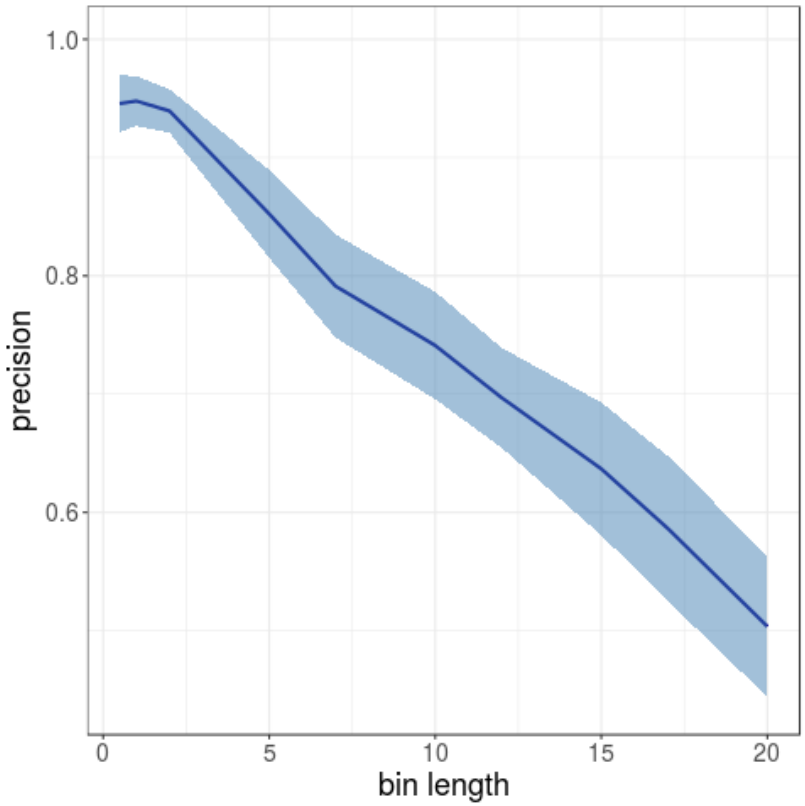
- Данные — 17000 пептидов.
  Из них построено K непересекающихся подвыборок.
- Для каждой подвыборки
  - вычислялись матрицы расстояний;
  - запускалась иерархическая кластеризация
    с динамическим обрезанием дерева (Langfelder, 2007).
- Подсчитывались различные статистики полученного разбиения и усреднялись по выборкам.

Пример

Параметры: число выборок $K = 8$.



(a) SIN



(b) QC

Рис. : Зависимость доли правильно кластеризованных пептидов от ширины столбца bin length у гистограмм.

Результаты

Предложен алгоритм кластеризации ожидаемых спектров, в том числе:

- Выбрано векторное представление ожидаемых спектров в виде гистограммы с регулируемой шириной столбца;
- Изучены три варианта расстояний между спектрами, выбран наилучший, исходя из вычислительных экспериментов;
- Проведены вычислительные эксперименты, подтверждающие корректность предложенной процедуры.