Санкт-Петербургский государственный университет

Клименко Полина Игоревна

Выпускная квалификационная работа

«Разработка компьютерного приложения для анализа данных UPVD-масс спектров»

Бакалавриат
Направление 01.03.01 «Математика»
Основная образовательная программа СВ.5000.2018 «Математика»

Научный руководитель:
Кандидат физико-математических наук,
Зав. кафедрой биоинформатики и математической биологии
СПбАУ РАН им. Ж.И. Алфёрова
Вяткина Кира Вадимовна

Рецензент:
Кандидат физико-математических наук,
Научный сотрудник,
Федеральное государственное бюджетное учреждение науки
Институт химической физики им. Н.Н. Семенова Российской академии наук
Иванов Марк Витальевич

Санкт-Петербург
2022

Содержание

Введение

Необходимость в исследовании белков и пептидов возникает при решении самых разных задач современной биологии, химии и медицины. Наиболее широко используемым методом их анализа является масс-спектрометрия.

На вход прибора, называемого масс-спектрометром, поступают ионизированные молекулы исследуемого белка или пептида. Далее они разделяются по отношению m/z массы к заряду, для определенного значения m/z обладающие им ионы изолируются и фрагментируются. Информация о фрагментных ионах и их интенсивности записывается в так называемый тандемный, или MC/MC, масс-спектр.

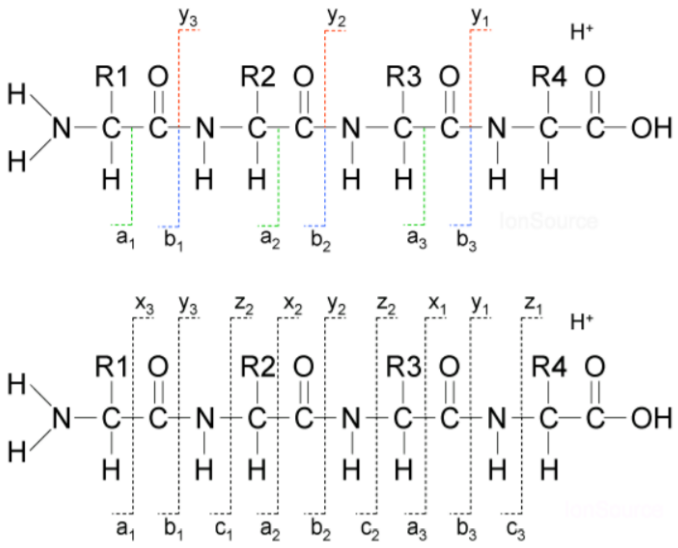

Рис. 1

Типы получаемых фрагментных ионов зависят от выбора метода активации фрагментации. Так, для CID-масс-спектров (collision-induced dissociation – диссоциация, индуцированная столкновениями) характерно присутствие b- и y-ионов, в то время как UVPD-спектры (ultraviolet photodissociation, ультрафиолетовая фотодиссоциация) могут одновременно содержать фрагментные ионы девяти различных типов

(a,x,b,y,c,z,a+, x+,y-). Такие масс-спектры весьма информативны, однако сложны с точки зрения интерпретации. Этим обусловлена потребность в программном инструменте, облегчающем данную задачу.

Результатом масс-спектрометрии обычно является набор тандемных масс-спектров, также называемых сканами. В данной работе мы будем использовать их представление в формате mzXML – одним из наиболее распространенных форматов, разработанных для этих целей.

В результате работы масс спектрометра делается некоторое количество сканов, данные с которых записываются в файл определенного формата.

В рамках проекта, на текущий момент, мы работаем с mzXML файлами.

По сути это стандартный XML файл, элементами которого являются $< scan > \ldots < scan/ >$.

```
<scan num="19"
          scanEvent="3"
          scanType="FULL"
          centroided="1"
          msLevel="1"
          peaksCount="15"
          polarity="+"
          retentionTime="PT353.43S"
          lowMz="400.38999999999999"
          highMz="1795.5599999999999"
          basePeakMz="445.34699999999998"
          basePeakIntensity="120053"
          totIonCurrent="16675500"
          msInstrumentID="LCQDeca">
      <peaks precision="64"
             byteOrder="network"
             pairOrder="m/z-int">
                AAAAAAAAAABALgAAAAAAAD/wAAAAAAAAQCwAAAAAAABAAA
                AAAAAAAEAqAAAAAAAAQAgAAAAAAABAKAAAAAAAAEAQAAAA
                AAAAQCYAAAAAAABAFAAAAAAAAEAkAAAAAAAAQBgAAAAAAA
                BAIgAAAAAAAEAcAAAAAAAAQCAAAAAAAABAIAAAAAAAEAc
                AAAAAAAAQCIAAAAAAABAGAAAAAAAAEAkAAAAAAAAQBQAAA
                AAAABAJgAAAAAAEAQAAAAAAAAQCgAAAAAABACAAAAAAA
                AEAqAAAAAAAAQAAAAAAAAABALAAAAAAAAD/wAAAAAAAA
      </peaks>
</scan>
```

Рис. 2

Внутри элемента $< peaks >$ хранится информация о всех пиках – парах ((m+z)/z; intensity) - в формате base64. На данный момент это

самая важная строчка для интерпретации данных.

Можно воспользоваться уже существующими инструментами визуализации и анализа данных масс спектров.

Например использовать питоновскую библиотеку *spectrum_utils*.

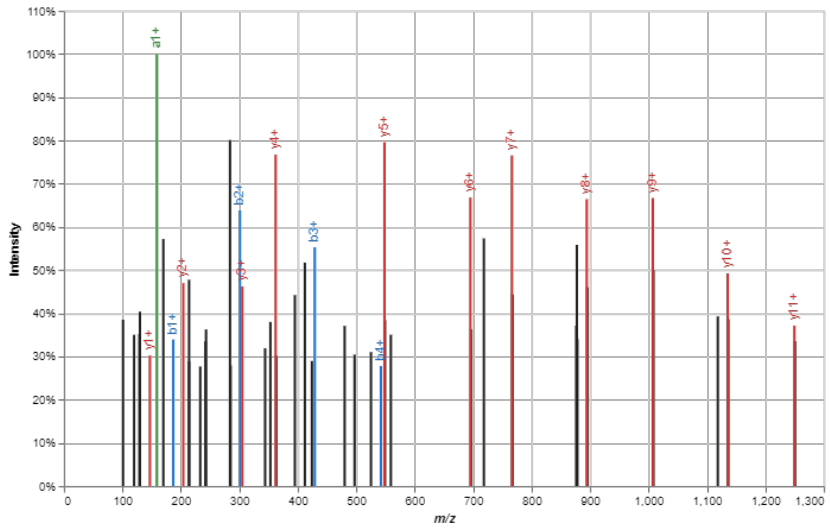

Рис. 3

Она неплохо справляется с тем чтобы аннотировать пики соответствующими ионами, однако качество визуализации страдает: пики слипаются и непонятно расстояние между ними. К тому же инструмент анализа и визуализации нужен биологам, не все из которых умеют пользоваться Jupyter Notebook.

Другой инструмент - SeeMS - программа с пользовательским интерфейсом и написанная под нее библиотека Proteowizard.

Рис. 4

SeeMS справляется с визуализацией сканов, хотя функция зума очень неудобная. Однако, возможности для анализа очень скудные: можно сделать предположение о том, какой пептид может быть в данной области и программа попытается этот пептид найти (при этом не проверяется ни существование такого пептида, ни возможность его существования в других участках спектра).

Оба инструмента нам не подходят. Нужен гибкий визуализатор и возможность получать первичный анализ сразу, не строя никаких конкретных предположений.

Для этого было решено реализовать собственное приложение.

Постановка задачи

Передо мной была поставлена задача создать инструмент для визуализации и анализа данных UVPD масс спектрометров.

Для этого было необходимо реализовать:

- парсер файлов формата .mzXML
- получение визуализации спектров по каждому скану
- алгоритмы проверки по каждой биологической гипотезе
- проверку выделенных пиков реализованными алгоритмами

- визуализацию результатов проверки выделенных пиков

- решение проблемы близко стоящих пиков

Основные результаты

После длительного изучения разных инструментов, в том числе ознакомления с открытым кодом библиотеки Proteowizard, было принято решение писать на языке *C#*, используя платформу WPF.

На данный момент реализован парсер mzXML файлов, который считывает внутренность элементов scan, выбирает только те сканы, msLevel которых равен 2 (это свидетельствует о фрагментации), и записывает данные в класс MassSpectrum. Информация о массе всего пептида содержится внутри элемента precursorMz.

Поля класса MassSpectrum соответствует атрибутам элемента scan:

```
spect.Add(new MassSpectrum()
{
    Num = Convert.ToInt32(num.Value),
    MsLevel = Convert.ToInt32(ms_level.Value),
    PeaksCount = Convert.ToInt32(peaks_count.Value),
    Polarity = polarity.Value,
    ScanType = scan_type.Value,
    LowMz = Convert.ToDouble(low_mz.Value.Replace(".", ",")),
    HighMz = Convert.ToDouble(high_mz.Value.Replace(".", ",")),
    BasePeakMz = Convert.ToDouble(base_peak_mz.Value.Replace(".", ",")),
    BasePeakIntensity = Convert.ToDouble(base_peak_intensity.Value.Replace(".", ","))
    PeptideMass = whole_mass_peptide,
    ByteString = bytes_array,
    MzList = MZ_Array(bytes_array),
    IntensityList = Intensity_Array(bytes_array)
});
```

Реализована визуализация каждого скана с удобным зумом. Для визулизации было перепробовано несколько библиотек, после чего выбор пал на ScottPlot.WPF

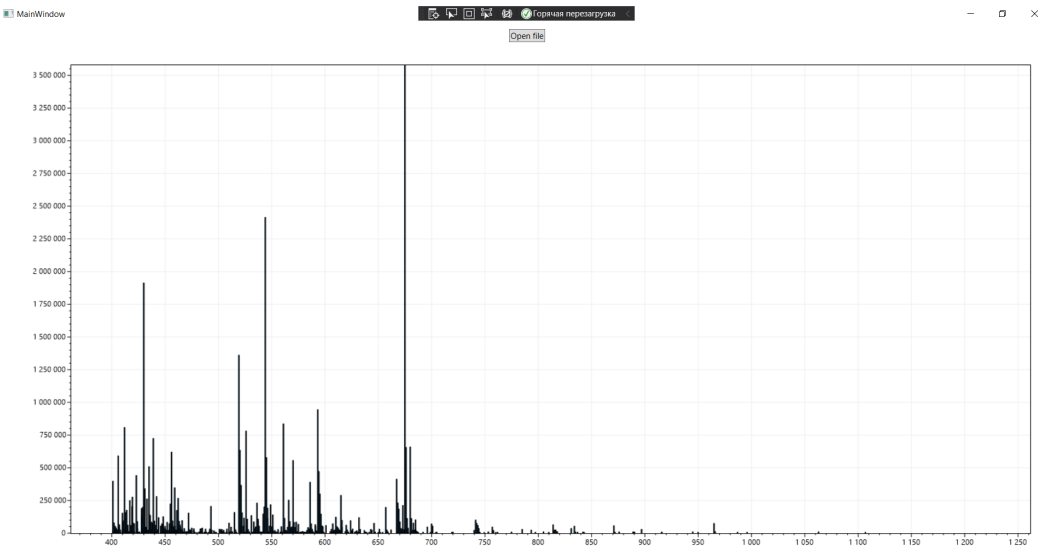

Рис. 5

В рамках работы были реализованы алгоритмы проверки на соответствие выбранного пика b/y- иону.

Для этого надо массовое число выделенного пика - (mz, intensity) - перевести в массу фрагмента и проверить существование пиков на расстоянии масс аминокислот от выделенного пика (величины заряда в масс спектрометре - целочисленные значения от 2 до 5):

```
CheckNeighbours(mz, peaks):
    # mz - массовое число пика
    # peaks - список пиков, подтверждающих гипотезу
    for z in {2,3,4,5}:
        mass_of_frag = (mz - 1) * z
        for all AA in list_of_amino_acids:
           peaks += FindPeaks(mass_of_frag + M(AA))
           peaks += FindPeaks(mass_of_frag - M(AA))
    if peaks not empty:
        return "Probably, it is b-ion!"
```

Функция FindPeaks ищет пики на расстоянии массы аминокислоты от пика:

```
FindPeaks(M):
```

```
peaks = []
for z in {2,3,4,5}:
    MassCount = (M + z) / z
    if MassCount is in mz_list:
        peaks.append((M + z) / z)
return peaks
```

При этом проверка принадлежности MassCount списку mz_list учитывает неточное совпадение пиков. В рамках проекта допускается погрешность в $\frac{1}{100}Da$.

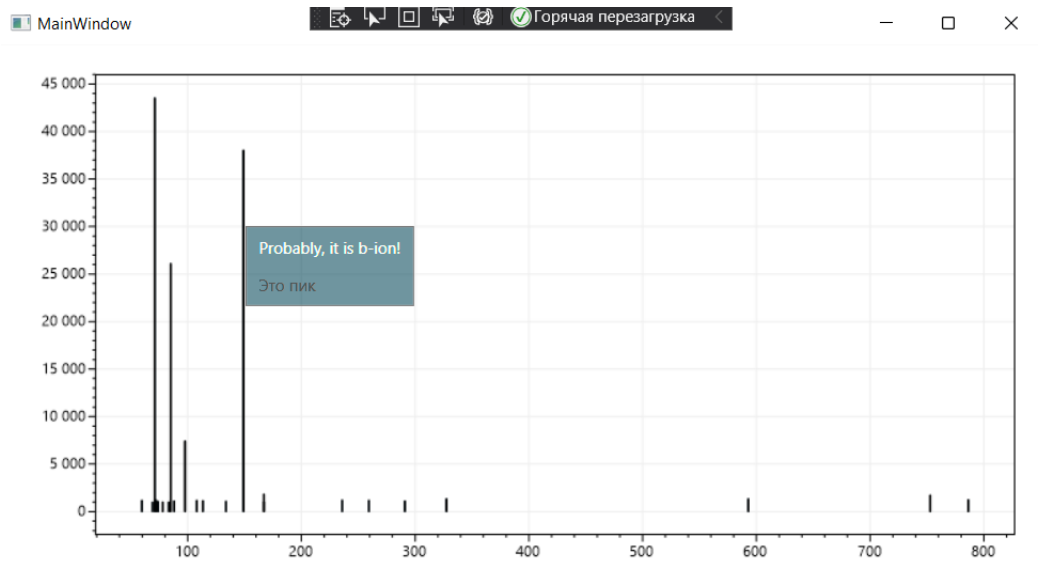

Рис. 6

Для большей точности, была реализована проверка существования комплиментарного пика (поиск пика, соответствующего второму фрагменту пептида после разбиения):

```
FindComplimetPeak(mz):
    for z in {2,3,4,5}:
        mass_of_frag = (mz - 1) * z
        mass_of_comp_peak = PeptideMass - mass_of_frag
        MassCount = (mass_of_comp_peak + z) / z
        if MassCount is in mz_list:
            return MassCount
```

Итоги

В ходе проделанной работы был разработан прототип программной системы, позволяющей находить подтверждения гипотезам о типах фрагментных ионов и наглядно визуализировать используемые ими пики. Далее планируется усовершенствовать предложенные методы и, тем самым, расширить возможности системы и обеспечить удобство ее применения при решении практических задач.

Репозиторий проекта

`https://github.com/PolinaKlimenko/MassSpectrumApp`

Список литературы

[1] К. В. Вяткина. *De novo секвенирование белков и пептидов: алгоритмы, приложения, перспективы.*

[2] Kira Vyatkina, Si Wu, Lennard J. M. Dekker, Martijn M. VanDuijn, Xiaowen Liu, Nikola Tolic, Mikhail Dvorkin, Sonya Alexandrova,Theo M. Luider, Ljiljana Pasa-Tolic and Pavel A. Pevzner. *De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra*

[3] Xiaowen Liu, Yuval Inbar, Pieter C. Dorrestein, Colin Wynne, Nathan Edwards, Puneet Souda, Julian P. Whitelegge, Vineet Bafna and Pavel A. Pevzner *Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins*

[4] Huseyin Guner, Patrick L. Close, Wenxuan Cai, Han Zhang, Ying Peng, Zachery R. Gregorich and Ying Ge *MASH Suite: A User-Friendly and Versatile Software Interface for High-Resolution Mass Spectrometry Data Interpretation and Visualization.*

[5] Chambers, M.C., MacLean, B., Burke, R., Amode, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L. Mallick, P. Nature Biotechnology 30, 918-920 (2012).*A cross-platform toolkit for mass spectrometry and proteomics.*

Приложение

```
public class DialogService : IDialogService
{
    Ссылок: 3
    public string FilePath { get; set; }
    Ссылок: 2
    public bool OpenFileDialog()
    {
        OpenFileDialog openFileDialog = new OpenFileDialog
        {
            Filter = "mzXML (*.mzXML)|*.mzXML|mzML (*.mzML)|*.mzML"
        };
        if (openFileDialog.ShowDialog() == true)
        {
            FilePath = openFileDialog.FileName;
            return true;
        }
        return false;
    }
```

Рис. 7

```
68      // есть ли пик с таким массовым числом в спектре;
69      // на вход: номер пика, который проверяем и массовое число нового пика, список mz_list;
70      // возвращает номер пика
        Ссылок: 2
71      public int PeakContains(int num_pos, double MassCount, double[] mz_list)
72      {
73          double delta = 0.00000000000000000166 / 100; // 1/100Da
74          double MassCount_old = mz_list[num_pos];
75          int l, r;
76          if (MassCount_old < MassCount)
77          {
78              l = num_pos - 1;
79              r = mz_list.Length;
80          }
81          else
82          {
83              l = -1;
84              r = num_pos;
85          }
86          while (l<=r)
87          {
88              int mid = (l+r) / 1;
89              double midVal = mz_list[mid];
90
91              if (midVal < MassCount)
92                  l = mid + 1;
93              else if (midVal > MassCount)
94                  r = mid - 1;
95              else if (Math.Abs(midVal - MassCount) < delta)
96                  return mid; // peak found
97          }
98          return -(l + 1);  // peak not found.
99      }
```

Рис. 8

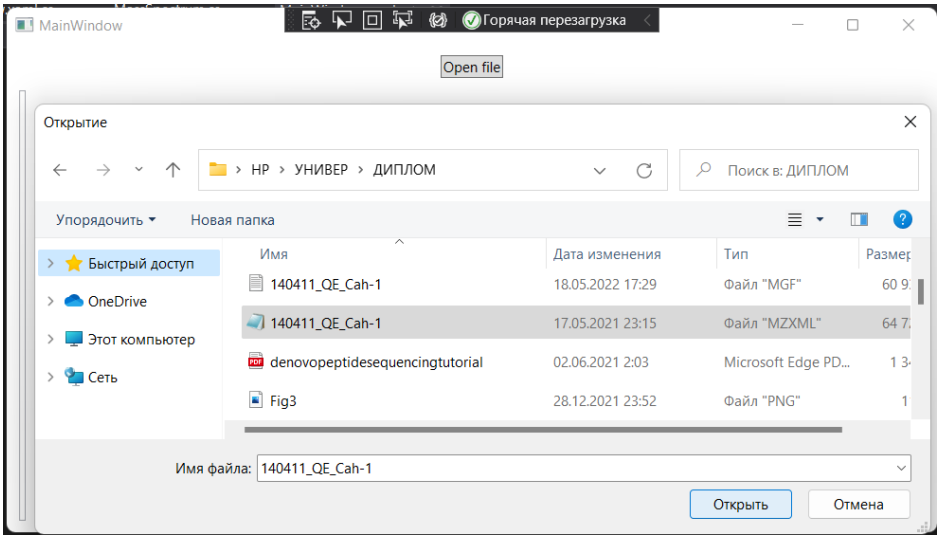
MainWindow
Горячая перезагрузка
Open file
Открытие
НР › УНИВЕР › ДИПЛОМ
Поиск в: ДИПЛОМ
Упорядочить
Новая папка
Быстрый доступ
OneDrive
Этот компьютер
Сеть
Имя
Дата изменения
Тип
Размер
140411_QE_Cah-1
18.05.2022 17:29
Файл "MGF"
140411_QE_Cah-1
17.05.2021 23:15
Файл "MZXML"
denovopeptidesequencingtutorial
02.06.2021 2:03
Microsoft Edge PD...
Fig3
28.12.2021 23:52
Файл "PNG"
Имя файла:
140411_QE_Cah-1
Открыть
Отмена

Рис. 9