ОЦЕНОЧНЫЕ МЕТОДЫ В ПРОТЕОМИКЕ[1]

Гришин Е. М.[2]
(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Современные математические методы исследования белка, такие как database search и de novo, имеют свои недостатки. При помощи database search невозможно определить белок, который отсутствует в базах данных. Методы de novo позволяют идентифицировать новые белки, но при этом являются очень ресурсоемкими (требуется использование суперкомпьютера). В рамках данного проекта был разработан комплексный подход приближенного анализа исследуемого белка, проводимый на персональном компьютере. Задача качественного и количественного определения исходной последовательности (белка) состоит из трех подзадач. Первая – устранение шумов и выделение пиков по данным масс-спектрометрии. Был разработан алгоритм, сочетающий метод скользящего среднего и технологию вычислительной фотографии HDR. Вторая подзадача – идентификация пиков. Она была сведена к задаче о рюкзаке и решена при помощи метода ветвей и границ. Последняя подзадача – восстановление исходной последовательности по набору фрагментов (пики и их интенсивности). Данная подзадача была решена при помощи построения двоичных деревьев и поиска пути максимальной длины. Все вычисления проводились на ПК с применением технологии параллельных вычислений CUDA.

Ключевые слова: протеомика, задача о рюкзаке, метод ветвей и границ, параллельные вычисления.

1. Введение

В настоящее время бурно развиваются методы геномики и протеомики. Эти науки достаточно молодые (появились в конце XX века) и используют комбинацию современных компьютерных технологий и методы исследования и идентификации вещества, прежде всего масс-спектрометрические с высоким разрешением. Протеомика (от слов протеин и геномика) – наука, занимающаяся анализом аминокислотных последовательностей в белках [15]. Наиболее часто встречается анализ при помощи

[1] *Работа выполнена при частичной поддержке фонда Базис, грант №20-2-9-12-1.*
[2] *Егор Максимович Гришин, инженер* (*grishin.em16@physics.msu.ru*).

масс-спектрометрии (в том числе тандемной) с последующей обработкой.

Существуют два подхода к исследованию белков: протеомика «снизу–вверх» [2] и протеомика «сверху–вниз» [10]. В первом методе все интересующие исследователя белки объединяются в сложную смесь пептидов, которая затем анализируются, чтобы определить, какие белки присутствовали в образце. В подходе «сверху–вниз» из образца выделяются интересующие белки, которые уже независимо друг от друга исследуются. Данная работа посвящена второму подходу и анализу одного белка.

В протеомике для качественного и количественного анализа в основном применяются два метода анализа данных, полученных при помощи масс-спектрометрии. Первый из них – поиск по базам данных. Основной идеей является нахождение спектра, наиболее похожего на полученный, в одной из соответствующих баз. После чего производится сравнение спектров, снятого масс-спектрометром, теоретического и найденного в базе данных. Очевидным недостатком такого подхода является неполнота баз данных. Если ранее не был идентифицирован некоторый спектр, то при помощи поиска по базам данных его определить не удастся. Однако подобный инструмент анализа применим как к протеомике «снизу–вверх», так и «сверху–вниз».

Чтобы избавиться от этого недостатка, были разработаны методы de novo [5], заключающиеся в анализе входных данных без использования референсных значений. В отличие от поиска по базам данных им не требуется дополнительная информация для идентификации аминокислотой последовательности. Они позволяют достаточно точно определить исследуемый образец при помощи активно развивающихся математических методов. Однако для их применения нужно использовать суперкомпьютер, что не всегда возможно и требует продолжительного времени вычислений.

Согласно публикации в журнале Science [6] данная задача является одной из 125 важнейших современных научных проблем. В данной работе предложены оценочные методы для ана-

лиза белков на персональном компьютере. Они не дают точное решение, но позволяют получить представления об образце и выстроить при необходимости дальнейшие исследования более рационально.

1.1. ЛИТЕРАТУРНЫЙ ОБЗОР

Анализ белковых структур является актуальной темой уже долгое время. Одним из первых методов, который был применен, является деградация по Эдману [4]. Было предложено поочередно отщеплять концевые аминокислотные остатки. Имея множество всех возможных подпоследовательностей, можно восстановить исходную структуру. Для идентификации *n*-концевых остатков использовалась хроматография. Все операции проводились вручную с использованием дорогостоящих реактивов. Помимо этого было невозможно исследовать длинные последовательности.

С распространением персональных компьютеров и адаптации масс-спектроскопии были разработаны методы database search [16] и de novo. В настоящее время чаще используется подход database search, в котором полученный спектр сравнивается со спектрами уже известных белков, так как подходы de novo более чувствительны к шумам и ошибкам в масс-спектрах. Алгоритмы поиска по базам данных состоят из нескольких шагов. Первоначально необходимо экспериментально получить спектр исследуемого белка при помощи масс-спектрометрии. Далее производится обработка полученного спектра. На последнем шаге происходит идентификация исследуемых. В дополнение может производиться сравнение теоретического спектра и экспериментального. При сравнении оценивается вероятность совпадения исследуемого и теоретически полученного белков.

В [8] подробно разбираются в преимущества и недостатки современных подходов de novo. Часто применяется метод overlap layout consensus [13]. Среди всех последовательности аминокислот одинаковой длины производится поиск пересечений и на основе этого строится полная последовательность пептидов. Основной идеей является построение графа де Брюина возмож-

ных перестановок на основе данных масс-спектрометрии. Пики спектрограммы соответствуют вершинам графа, а расстояния между пиками соответствуют ребрам. В [12] предложен алгоритм поиска Евклидова пути в подобном графе, который и является искомой последовательность. Однако размеры построенных графов велики, и для поиска приходится использовать суперкомпьютеры. В [14] проведен подробный анализ использования графов для сборки белковых последовательностей, построенных по спектрам, а также представлены основные задач, решение которых необходимо для более эффективного применения подхода de novo.

Методы масс-спектрометрии активно развиваются, что позволяет получить более четкие спектры с меньшим количеством шумов, что важно для таких чувствительных методов как de novo. Новые разработанные алгоритмы также позволяют сократить время обработки данных и более качественно идентифицировать неизвестные образцы. В [11] сделан обзор последних достижений в области подходов de novo.

В данной работе предложен метод для оценочного анализа белков, состоящий из трех шагов. Прежде всего надо обработать спектр и устранить шумы. Для этого был разработан алгоритм с применением метода скользящего среднего (его схожее применение можно найти в [1]) и технологии вычислительной фотографии HDR [7]. Это позволяет на основе нескольких спектров для исследуемого белка накопить больше информации и точнее ее предобработать, избавившись от шумов и неточностей. Данная часть алгоритма выполняется при помощи параллельных вычислений с использованием ресурсов видеокарты – GPU-вычисления.

Далее необходимо определить, каким аминокислотам соответствует каждый локальный пик. Эта задача сведена к задаче о рюкзаке, которая была решена методом ветвей и границ при помощи параллельных вычислений. В [3] приведена реализация метода ветвей и границ с применением параллельных вычислений на GPU.

Последний шаг – восстановление исходной последовательности. Для этого был использован, предложенный в [9]. В его

основе лежит построение двоичных деревьев для различных начальных элементов. По пересечению фрагментов белка, входящих в построенные графы, осуществляется поиск оптимального пути в них, что и является искомой последовательностью. Построение двоичных деревьев и поиск пути в них так же выполнялся параллельно.

2. Методы в протеомике

Задача протеомики формируется следующим образом. Белок кодируется последовательностью аминокислот. Аминокислоты, в свою очередь, являются триплетами нуклеотидов. Нуклеотиды – неделимые элементы, существует четыре разновидности которых (A, T, G, C). При исследовании одновременно анализируется большое количество копий одного и того же белка. Каждая копия разрезается случайным образом на фрагменты (аминокислоты остаются неделимыми). Все полученные фрагменты (от всех копий) одновременно анализируются в масс-спектрометре. В результате получается спектр интенсивностей некоторых неопределенных фрагментов, которые называются ридами. Характеристики всех аминокислот и некоторых пептидов (недлинная последовательность аминокислот) известны и их можно идентифицировать по спектру без дополнительных вычислений. Оставшиеся пики необходимо обработать и определить набор аминокислот и пептидов, которым они соответствуют.

Данную задачу можно переформулировать. Пусть задан алфавит из четырех букв (A, T, G, C). Словом будет являться последовательность слогов, состоящих из трех букв (аминокислоты). Известны количественные характеристики неопределенных фрагментов слов (некоторые известны). Изначальное количество копий исследуемого слова точно неизвестно. Каждому слогу соответствует некоторый весовой коэффициент. Если имеется несколько неизвестных фрагментов с одинаковой весовой характеристикой (сумма весовых коэффициентов составляющих слогов), то с большой долей вероятности можно утверждать, что все такие фрагменты одинаковы. Хотя могут встре-

чаться фрагменты с равными весовыми характеристиками, но вероятность такого события мала, поэтому принято такое допущение. После идентификации всех фрагментов (составляющих их слогов) необходимо восстановить слово, длина и весовая характеристика которого лежит на некотором заданном интервале. В копиях слова и фрагментах могут быть ошибки, которые необходимо исправлять.

2.1. ОБРАБОТКА СПЕКТРА

Первоначально необходимо обработать «сырой» спектр, полученный при помощи масс-спектрометрии. Необходимо удалить шумы, которые могли появиться из-за примесей, и выделить пики, которые будут использованы для дальнейшего анализа.

Для повышения точности были использованы несколько спектров (полученные при повторных экспериментах). Сначала в каждом спектре выделялись пики. Для этого был использован метод скользящего среднего. Для каждой выбранной точки и ее соседей вычислялось среднее с учетом весовых коэффициентов. Далее из значения в выбранной (центральной) точки вычиталось значения взвешенного среднего. Если получалось отрицательное значение, то оно приравнивалось к нулю. Это позволяет увеличить значение максимумов и избавиться от фонового тренда. Было проведено исследование по эффективности выделения пиков от величины используемого интервала. Оказалось, что наиболее подходящим является использование пяти соседних значений помимо центрального.

После того как выделены необходимые пики на каждом отдельном спектре, эти наборы данных требуется объединить. Это необходимо, потому что чем выше точность идентификации пиков, тем более точно можно определить исходную последовательность белков. В силу технической погрешности пики смещаются относительно их истинного положения. Чтобы уменьшить эту погрешность, был применен аналог технологии вычислительной фотографии HDR для набора спектров. Это позволяет не только более точно установить положение пиков, но и дополнительно уменьшить воздействие шумов на итоговый ре-

зультат. Результирующее значение на основе предобработанных спектров получается следующим образом. На каждом спектре выделяются локальные максимумы (пики). Для каждого локального максимума выбранного спектра осуществляется поиск локальных максимумов на остальных спектрах на заданном интервале. Если локальные максимумы найдены, то их положение и их количество заносится в память. Если локальные максимумы не найдены, то принимаем данное значение за ошибочное. Далее, зная количество локальных максимумов и их положение на спектрах, необходимо определить координату точки, расположенную к ним ближе остальных по сумме расстояний. Варьируя радиус окрестности локального максимума, можно настраивать алгоритм для разных задач с учетом точности эксперимента.

2.2. ИДЕНТИФИКАЦИЯ ПИКОВ

Теперь, когда по набору спектров устранены шумы и погрешности, а пики однозначно выделены, последние нужно идентифицировать. Все белки состоят из аминокислот, которые, в свою очередь, являются триплетами нуклеотидов (A, C, G, T). Аминокислоты (характеристики всех известны) образуют пептиды, характеристики некоторых из них известны. Положение пиков соответствуют массам фрагментов белка. По интенсивностям пиков можно определить количество данного фрагмента в исследуемом образце. Одному пику могут соответствовать разные фрагменты (с очень близкими массами), но в угоду быстроте работы алгоритма это очень маловероятное событие не учитывается.

Расстояние между двумя пиками на масс-спектре задает разницу масс двух фрагментов. Если эта разность равна массе некоторого фрагмента, присутствующего в спектре, известного пептида или аминокислоты, то можно с большой долей вероятности утверждать, что тяжелый фрагмент (и соответствующий пик) состоит из более легких фрагментов. Таким образом, задачу идентификации пиков можно свести к известной задаче о рюкзаке.

В переформулированной задаче пики поочередно являются как рюкзаками, так и предметами. Алгоритм начинает свою работу с самого тяжелого фрагмента, представленного в спектре. Этот фрагмент является рюкзаком. Остальные же более легкие фрагменты (пики которых присутствуют в спектрограмме) являются предметами. При этом количество предметов ограничено (определяется по интенсивности пиков). Далее для таких входных данных ищутся несколько наиболее точных решений при помощи метода ветвей и границ. Аналогичные процедуры повторяются для всех пиков (фрагментов).

Самые легкие фрагменты состоят из отдельных аминокислот, поэтому их можно идентифицировать достаточно точно (так как массовые характеристики коротких пептидов и всех аминокислот известны с высокой точностью). Тогда, двигаясь в обратном направлении (от легких фрагментов к тяжелым) можно идентифицировать все пики спектра (определить их аминокислотный состав).

При решении методом ветвей и границ запоминаются все допустимые решения. В конце, когда наборы возможных решений для всех пиков известны, необходимо получить общее решение с учетом общего количества всех фрагментов. Для этого решается задача о рюкзаке повторно. На этот раз для каждого пика надо выбрать ровно одно решение, которое учитывает как массовые характеристики фрагментов, так и их общее количество в исходном масс-спектре. Причем общее количество всех различных фрагментов по отдельности, входящих в состав каждого пика, должно быть как можно ближе к количеству таких фрагментов, представленных в изначальном спектре.

2.3. ВОССТАНОВЛЕНИЕ ПОСЛЕДОВАТЕЛЬНОСТИ

Наконец, зная составляющие каждого из пиков и их количество (интенсивности), можно восстановить исходную последовательность. При получении масс-спектра множество копий исходного слова (белка) произвольно разрезается на некоторые фрагменты. Фрагменты разных копий слова могут быть пересекающимися. Может случиться так, что на одном спектре будут присутствовать определенные пики, а на другом (при повторном

анализе этого же белка) эти пики будут отсутствовать. Для того чтобы этого избежать, проводится несколько повторных экспериментов. Благодаря этому можно понизить ошибку и накопить больше информации об исходном белке (аналогично методу рядов).

Основой решения данной задачи является построение двоичных деревьев. Идеи, представленные в [9], были несколько упрощены, чтобы процесс вычисления занимал меньшее время. Чтобы восстановить исходную последовательность, строятся двоичные деревья пересечений фрагментов. При этом в качестве начальной вершины выбираются наиболее длинные фрагменты, полученные на предыдущем шаге. Так как спектры при повторных экспериментах могут различаться, то поиск подходящих фрагментов осуществляется для набора фрагментов, соответствующих только одному спектров.

Совпадение пересечений фрагментов не должно быть абсолютным: суффикс и префикс двух фрагментов могут иметь небольшие различия, количество совпадений было выбрано на уровне 85%, что позволяет сочетать сравнительно высокую скорость работы и точность решений. В каждом двоичном дереве ищутся пути максимальной длины (от листьев в центр), с длинами и весовыми характеристиками, лежащими на интервалах, определенных для исследуемого слова изначально. Последовательности, полученные в результате поиска наилучшего пути, объединяются в исходную последовательность аналогично предыдущему шагу при помощи решения задачи о рюкзаке.

3. Результаты

Таким образом, общая задача была разбита на три основные подзадачи: обработка спектров, идентификация пиков и восстановление исходной последовательности. Все задачи решаются последовательно, выходные данный одной задачи являются входными данными для следующей.

В первой подзадаче параллельно выполнялся пересчет значений в каждой точке по методу скользящего среднего с весовыми коэффициентами и объединение полученных данных

с нескольких масс-спектров. На рис. 1 представлен пример работы алгоритма на тестовом примере. Даны семь наборов входных данных (полученных при повторных экспериментах), один из которых (синий, седьмой) является преднамеренно ошибочным. В средней части данного рисунка представлены наборы данных без шумового. Видно, что положение пиков и их интенсивности различаются. В верхней части рисунка голубыми точками обозначены идентифицированные пики, при этом шумовые данные были исключены из рассмотрения алгоритмом, и соответствующие пики не были идентифицированы за истинные. Несмотря на большие шумы и неточности в одном из спектров, пики идентифицированы абсолютно точно с учетом фонового тренда. Для каждой центральной точки каждого спектра вычисления проводились параллельно при помощи графического процессора.

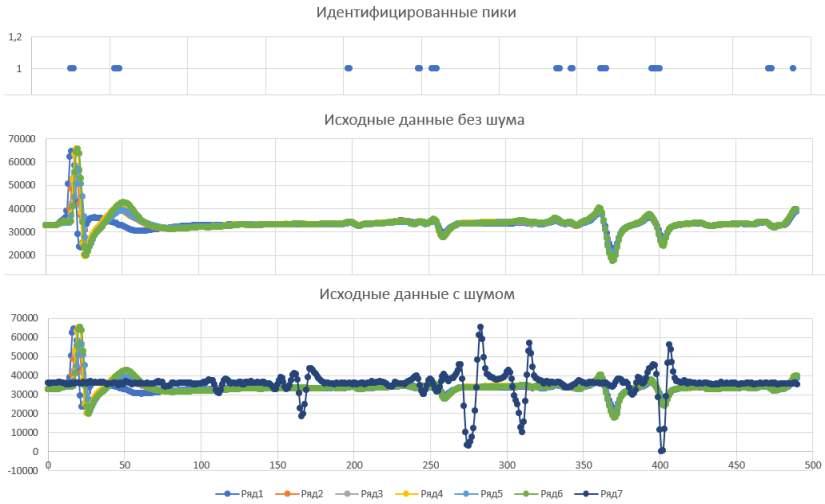

Рис. 1. Идентифицированные пики

На рис. 2 схематически показана формулировка задачи идентификации пиков в виде задачи о рюкзаке. Наиболее тяжелый пик из рассматриваемых выше соответствует рюкзаку, а более легкие – предметам. Так, в рюкзак вместимости 21

необходимо положить предметы (кирпичи с весом 9 в количестве не более 10 и книги с весом 3 в количестве не более 5) с суммарной массой, наиболее близкой к вместимости рюкзака. В дереве решений при достижении определенного количества вершин поиск оптимального решения по ветвям графа может происходить параллельно и независимо, что сокращает время работы алгоритма. Задача о рюкзаке была решена при помощи метода ветвей и границ на графическом процессоре.

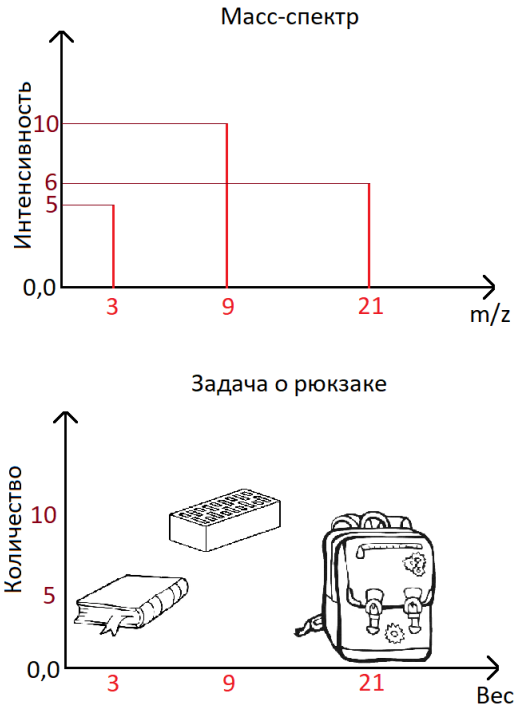

Рис. 2. Сведение решаемой проблемы к задаче о рюкзаке

На рис. 3 схематично показано построение двоичного дерева. Пусть есть фрагмент TATCG (выделено желтым). Среди других фрагментов ищутся совпадения суффиксов и префиксов (зеленые буквы). Так как совпадение не обязательно должно

быть полным (регулируется исследователем), то могут быть допущены противоречия (красные буквы).

В построенном дереве можно найти путь максимальной длины GCTTCTATCG(C/T)FAGTA (выделено голубым). При этом допущена неточность, так как по данному дереву невозможно определить значение одной из букв. Далее аналогичным образом восстанавливается исходная последовательность. Для различных начальных фрагментов построение таких деревьев и поиск пути в них проводился параллельно.

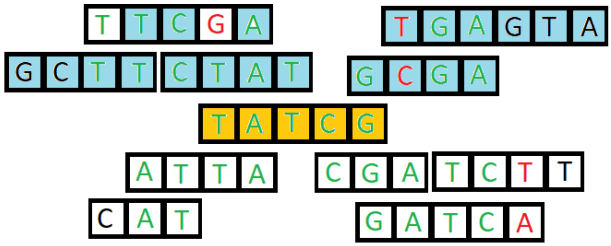

Рис. 3. Пример двоичного дерева

4. Заключение

Все вычисления проводились на персональном компьютере с процессором Intel Core i9-10980HK и видеокартой Nvidia RTX 2080 Max-Q. Алгоритмы были имплементированы на языке C++ с использованием технологи параллельного программирования CUDA. Количество пиков на масс-спектрах достигало миллионов, до 10 масс-спектров было получено (проведено повторных экспериментов) для исследуемого образца. Время работы предложенного комплексного подхода достигало нескольких часов работы. Наиболее ресурсоемкими задачами были решение задачи о рюкзаке (идентификация пиков – до 30% общего времени работы) и построение двоичных деревьев с поиском пути в них (восстановление исходной последовательности – до 70% общего времени). Предложенный подход позволяет значительно сокра-

тить время для приближенного анализа исследуемого белка с последующим применением более точных и время затратных методов при необходимости.

Литература

1. DARMAWAN R.A.S., SUTOMO A.S., LEGOWO B. *Determination of optimal window in spectrum analysis process with moving average method gravity data measurement* // Journal of Physics: Conf. Series. – 2019. – Vol. 1153. – P. 1–6.
2. DI SILVESTRE D., BRAMBILLA F., AGNETTI G., MAURI P. *Bottom-Up Proteomics* // In: Agnetti G., Lindsey M., Foster D. (eds) Manual of Cardiovascular Proteomics. – Springer, 2016.
3. DIDIER E.B., LALAMI M.E. *GPU Implementation of the Branch and Bound method for knapsack problems* // IEEE 26th Int. Parallel and Distributed Processing Symposium Workshops & PhD Forum. – 2012. – P. 1763–1771.
4. EDMAN P. *Method for determination of the amino acid sequence in peptides* // Acta Chemica Scandinavica. – 1950. – Vol. 4. – P. 283–293.
5. HUBBARD S.J., JONES A.R. *Proteome Bioinformatics* // Springer, New-York, 2010. – P. 397.
6. KENNEDY D., NORMAN C., SIEGFRIED T. *125 Science Questions: What Don't We Know?* // Science. – 2005. – Vol. 309, No. 5731. – P. 75–102.
7. MANN S., PICARD R.W. *On being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures* // Proc. of IS&T 48th Annual Conference Society for Imaging Science and Technology Annual Conference. – 1995. – P. 422–428.
8. MUTH T., RENARD B.Y. *Evaluating* de novo *sequencing in proteomics: already an accurate alternative to database-driven peptide identification?* // Briefings in Bioinformatics. – 2018. – Vol. 19, Iss. 5. – P. 954–970.
9. NARZISI G., MISHRA B. *Scoring-and-unfolding trimmed tree assembler: concepts, constructs and comparisons* // Bioinformatics. – 2011. – Vol. 27, No. 2. – P. 153–160.

10. PAMREDDY A., AGENDER R.P. *Top-down proteomics: applications, recent developments and perspectives* // Journal of Applied Bioanalysis. – 2016. – Vol. 2, Iss. 2. – P. 52–75.
11. PAN X., KORTEMME T. *Recent advances in de novo protein design: Principles, methods, and applications //* Journal of Biological Chemistry. – 2021. – Vol. 19, Iss. 296. – P. 1–16.
12. PEVZNER P.A., HAIXU T., WATERMAN M.S. *An Eulerian path approach to DNA fragment assembly* // Proc. of the National Academy of Sciences of the United States of America. – 2001. – Vol. 98, Iss. 17. – P. 9748–9753.
13. RIZZI R., BERETTA S., PATTERSON M., PIROLA Y., PREVITALI M., VEDOVA G.D., BONIZZONI P. *Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era* // Quantitative Biology. – 2019. – Vol. 7. – P. 278–292.
14. SUTTON G., WHITE O., ADAMS M., KERLAVAGE A. *TIGR assembler: A new tool for assembling large shotgun sequencing projects* // Genome Sci Technol. – 1995. – Vol. 1. – P. 9–19.
15. WU Q., GORSHKOV M.V., PAŠA-TOLIĆ L. *Towards increasing the performance of FTICR-MS with signal detection at frequency multiples: Signal theory and numerical study* // Int. Journal of Mass Spectrometry. – 2021. – Vol. 469. – P. 116669.
16. XU H., FREITAS A.F. *A high mass accuracy sens tive probability based scoring algorithm for database searching of tandem mass spectrometry data* // BMC Bioinformatics. – 2007. – Vol. 8. – P. 133.

APPROXIMATE APROACH IN PROTEOMICS

Egor Grishin, engineer, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow (grishin.em16@physics.msu.ru).

Abstract: Modern mathematical methods for protein analysis, such as database search and de novo methods, have their own drawbacks. It is not possible to identify proteins that are not included in databases using database search. The de novo methods allow us to identify new proteins but they are very computationally demanding (requiring the use of a supercomputer). In this project a complex approach of approximate protein analysis conducted on a personal computer was developed. A problem of qualitative and quantitative determination of initial sequence (protein)

consists of three subproblems. The first one is noise cancellation and peak identification using mass spectrometry data. An algorithm combining a sliding average method and computational photography HDR technology was developed. The second subproblem is peak identification. It was reduced to a knapsack problem and solved using the branch and bound method. The last subproblem is initial sequence reconstruction using a set of fragments (peaks and their intensities). This subproblem was solved by constructing double trees and searching for a path of maximum length. All calculations were performed on a PC using CUDA parallel computing technology.

Keywords: proteomics, the knapsack problem, the branch and bound method, parallel computing.

УДК 519.85
ББК 22.176
DOI: 10.25728/ubs.2022.95.3

Статья представлена к публикации
членом редакционной коллегии Э.Ю. Калимулиной.

Поступила в редакцию 09.11.2021.
Опубликована 31.01.2022.