ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2006, том 16, № 3, с. 122–131

МАСС-СПЕКТРОМЕТРИЯ ДЛЯ БИОТЕХНОЛОГИИ. ИНТЕРПРЕТАЦИЯ ДАННЫХ, МЕТОДОЛОГИЯ, ПРИМЕНЕНИЕ

УДК 621.384.668.8: 577.112.6

© Я. И. Лютвинский, В. В. Макаров, Н. В. Краснов, Е. П. Подольская, А. Н. Веренчиков

ЧАСТИЧНАЯ РАСШИФРОВКА АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ПЕПТИДА ПО ЕГО ФРАГМЕНТНОМУ МАСС-СПЕКТРУ: АЛГОРИТМ И РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ

Разработан новый алгоритм восстановления части аминокислотной последовательности пептида по фрагментному масс-спектру пептида. Алгоритм оптимизирован по числу проверяемых гипотез, что обеспечивает его высокое быстродействие. При оценке возможных гипотез используется байесов подход, основанный на фактической статистике пиков фрагментного масс-спектра. Поведена проверка алгоритма на данных тандемного времяпролетного масс-спектрометра Q-TOF и масс-спектрометра на основе ловушки Паули.

ВВЕДЕНИЕ

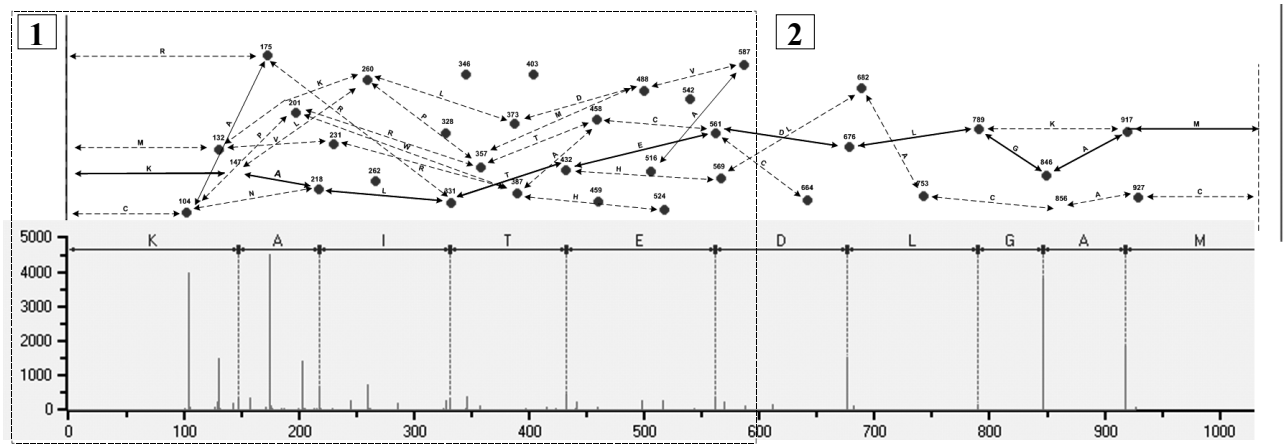
Восстановление аминокислотной последовательности пептида по его фрагментному масс-спектру — это один из основных приемов интерпретации масс-спектрометрических данных, используемых в протеомике. В настоящее время наиболее распространенным подходом к задаче восстановления аминокислотной последовательности является поиск наилучшего совпадения состава масс фрагментов в зарегистрированных экспериментальных фрагментных спектрах и теоретических спектрах, построенных на основании аминокислотных последовательностей известных белков, содержащихся в протеомных базах данных. Этот подход реализован в таких известных программных комплексах, как Mascot [1], Sequest [2], X!Tandem [3], и многих других.

Альтернативный подход подразумевает восстановление аминокислотной последовательности без обращения к базам данных на основании непосредственного анализа сигналов спектров. В рамках этой альтернативы можно выделить два метода.

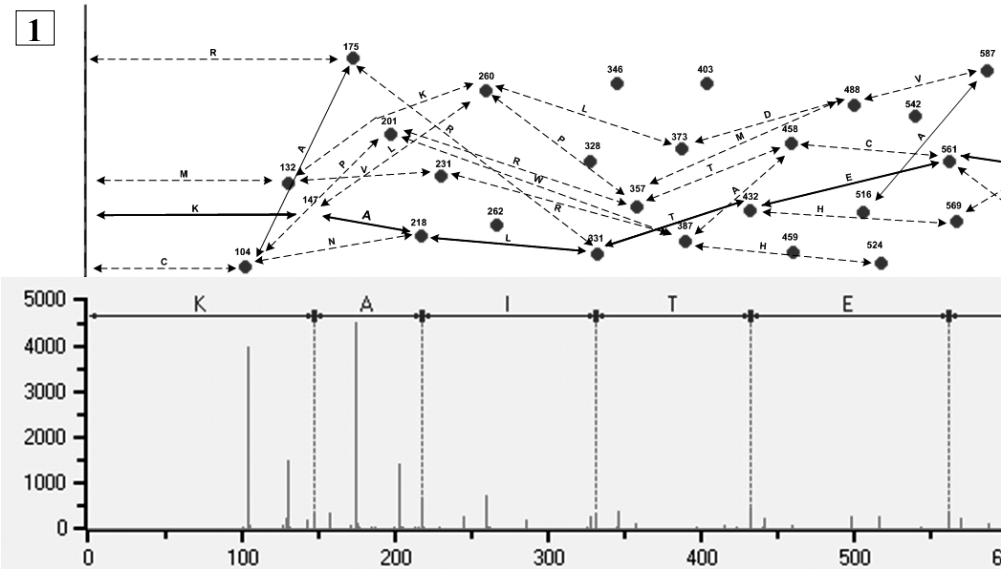
- Полное восстановление аминокислотной последовательности [4, 5] (de novo sequencing). Отсутствие во многих фрагментных масс-спектрах даже хорошего качества полных серий сигналов, соответствующих сериям фрагментных ионов, приводит к базовому недостатку этого метода — восстановленная последовательность содержит слабые предположения, основанные на неполной информации, и потому часто не соответствует действительной аминокислотной последовательности.
- Частичное восстановление аминокислотной последовательности. Для частичного восстановления используются наблюдаемые в спектре последовательности пиков, принадлежащих основным сериям фрагментов пептида, расстояние между которыми соответствует массам аминокислотных остатков пептида. Как правило, такая последовательность пиков покрывает не весь спектр и позволяет восстановить только часть аминокислотной последовательности пептида.

Данная работа посвящена выявлению наиболее вероятных последовательностей пиков, представляющих аминокислотную последовательность исходного пептида. Идея использовать эти последовательности пиков для поиска в базах данных белков впервые высказана в статье [6], и техника такого поиска в мировой литературе получила название Peptide Sequence Tag Search (поиск PST). В настоящее время ведутся интенсивные работы как в области разработки новых алгоритмов построения PST [7–9], так и использования PST для идентификации и характеризации белков в ходе белковых анализов [10, 11]. Интерес к алгоритмам построения и использования PST обусловлен следующими причинами.

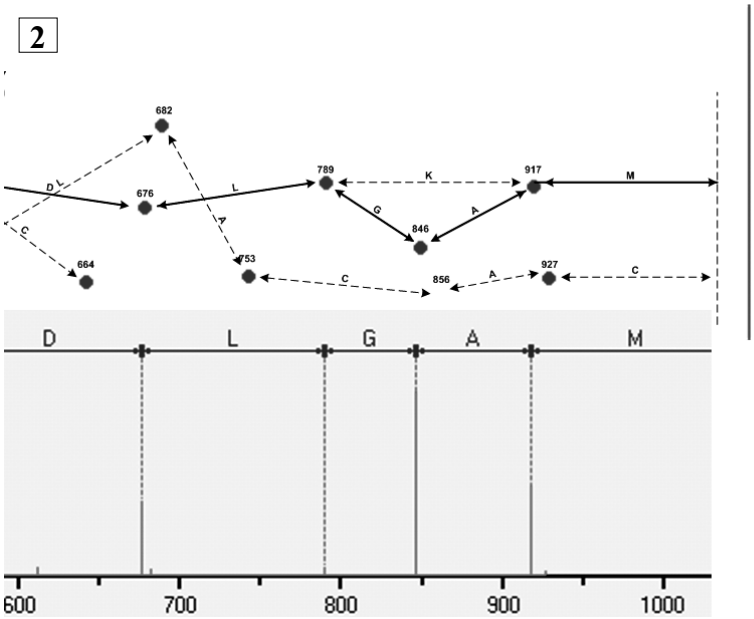
- PST, выделенные в результате анализа фрагментных масс-спектров пептидов, полученных в результате неполного или неспецифического гидролиза [10], а также содержащих посттрансляционные модификации [12], тем не менее позволяют использовать спектр для идентификации белков.
- PST пригодны для идентификации неизвестных белков у организмов с несеквенированным геномом [11].
- PST обладают меньшей информационной избыточностью по сравнению со списком пиков масс-спектра, что снижает время поиска в протеомных базах данных и позволяет использовать PST для создания систем быстрой обработки масс-спектрометрических данных белковых анализов [13].



а



б



в

Рис. 1. Представление масс-спектра как графа на примере масс-спектра пептида MAGLDETIAK (с упрощением).
а — общий вид; б, в — укрупненные изображения частей **1** и **2** общего вида

ФОРМУЛИРОВКА ЦЕЛЕЙ РАЗРАБОТКИ

Алгоритм распознавания PST формализуется как поиск частичного пути в взвешенном ориентированном ациклическом графе [14]. В таком графе вершины представлены сигналами масс-спектра, а ребра допустимых переходов — разницами масс, соответствующих массам аминокислотных остатков (см. рис. 1). На рисунке сплошными линиями выделен корректный путь через граф, а штриховыми линиями несколько вариантов ложнопозитивных результатов.

Классические алгоритмы поиска наилучшего пути в графе, такие как алгоритм Дейкстры [15] или алгоритм A-Star [16], оперируют понятием фиксированной начальной и конечной точек пути. В нашем случае, хотя начальные и конечные точки детерминированы (нулевая масса и масса родительского иона), в их достижении нет необходимости. Хуже того, правильного решения задачи нахождения полного пути для такого графа может и не существовать — далеко не все спектры содержат полные серии фрагментных ионов. Из-за этого подход классических алгоритмов поиска пути малоприменим.

Как правило, задачи поиска путей в графах связаны с рекурсивным обходом графа. Для решения задачи распознавания PST можно было бы применить такой метод. Для этого можно построить полный набор вариантов путей заданной длины для каждого узла графа и впоследствии оценить каждый из вариантов. Однако такая постановка задачи приводит к полному перебору вариантов возможных PST для данного спектра.

Последствия полного перебора гипотез могут быть легко продемонстрированы на примере спектра на рис. 1. Спектр получен с высоким качеством, содержит полную серию y-ионов и позволяет полностью восстановить последовательность пептида. При построении графа использовались только 33 наиболее интенсивных сигнала спектра из 141 имеющихся. Тем не менее даже этот граф демонстрирует множественные ложноположительные гипотезы PST, число которых быстро возрастает с введением в граф дополнительных пиков из масс-спектра. Предварительные оценки показывают, что для обеспечения достаточной избирательности при решении задачи идентификации белков PST должны состоять хотя бы из 4–5 аминокислот восстановленной последовательности. В спектре на рис. 1 возможно построить более 28 000 PST такой длины. На основании этого наблюдения приходим к двум следствиям.

1. Нахождение наилучшего пути потребует построения набора адекватных оценок для этих путей.
2. Методы, использующие полный перебор для нахождения наилучшего пути в данном случае неприемлемы, и алгоритм построения оптимального пути должен быть оптимизирован по отношению к числу проверяемых гипотез.

ОЦЕНКА ГРАФА МАСС-СПЕКТРА

Для оценки возможности вхождения пика в PST есть ряд эмпирических критериев, таких как:

- относительная интенсивность пика в его окрестности;
- зашумленность спектра вокруг пика в его окрестности;
- наличие в спектре пиков, парных данному по правилам построения серий ионов: y↔b, y↔a, x↔b и т. д.;
- наличие характерных нейтральных потерь: $-H_2O$, $-NH_2$ и т. д.

Заметим, что в некоторых исследованиях набор эмпирических критериев может меняться. Так, например, в работе Зубарева и др. [9] привлекаются такие критерии, как наличие сигналов фрагментных ионов, полученных при использовании различных методов фрагментации, наличие изотопных распределений для моноизотопных ионов и, возможно, другие критерии.

Значимость каждого из этих критериев A_i для конкретной экспериментальной установки можно установить с использованием референтного метода определения последовательности пептида. В данной работе в качестве референтного метода использовано программное обеспечение (ПО) для идентификации белков на основе МС-МС-спектров X!Tandem [3]. Результат распознавания X!Tandem, принимаемый за истину, позволяет численно оценить условные вероятности выполнения критериев для полного набора из двух гипотез:

1) H_1 — пик относится к серии фрагментных ионов b или y;
2) H_2 — пик не относится к серии фрагментных ионов b или y.

Решение о принадлежности или непринадлежности сигнала к серии фрагментных ионов принимается исходя из восстановления теоретической картины фрагментации для пептидов, предложенных в качестве результата интерпретации спектра системой X!Tandem. Описанные критерии можно разделить на две группы.

- Бинарные критерии, такие как наличие y иона пары, соответствующей потере группы H_2O. Оценка условных вероятностей выполнения критерия вычисляется для события критерия.
- Функциональные критерии, такие как относительная интенсивность. Условные вероятности

для таких критериев можно оценить отдельно для некоторых интервалов их значений. Например, для критерия относительной интенсивности удобно оценить условные вероятности соответствия сигнала масс-спектра ионам серий y и b для набора интервалов 1, (0.5:1], (0.25:0.5] и т. д.

Таким способом, мы получаем полный набор условных вероятностей для каждого из критериев $P(\mathrm{A}_i|\mathrm{H}_1)$, $P(\mathrm{A}_i|\mathrm{H}_2)$. Собранный набор оценок критериев позволяет оценить по Байесу вероятность гипотез H_1 для каждого пика спектра в том случае, если мы можем предполагать, что значимость критериев для этого спектра адекватна спектрам, использованным для накопления статистики. Оценку каждого пика мы получаем последовательным применением формулы Байеса для каждого из предварительно оцененных критериев A_i

$$P(\mathrm{H}_1 \mid \mathrm{A}_i) = \\ = \frac{P(\mathrm{H}_1)P(\mathrm{A}_i \mid \mathrm{H}_1)}{P(\mathrm{H}_1)P(\mathrm{A}_i \mid \mathrm{H}_1) + (1 - P(\mathrm{H}_1))P(\mathrm{A}_i \mid \mathrm{H}_2)}. \quad (1)$$

При первом применении теоремы Байеса в качестве априорной вероятности $P(\mathrm{H}_1)$ используется доля сигналов ионов серий y и b в спектрах пептидов. При последовательной оценке по ряду критериев в качестве априорной вероятности используется апостериорная вероятность, полученная на предыдущем шаге. В качестве итоговой оценки Q_j вершины j графа масс-спектра используется апостериорная вероятность, полученная после применения всех оцененных критериев.

Оценка ребер графа, построенного на основании масс-спектра, сводится к оценке допустимости предположения о том, что разность масс между двумя пиками является измерением массы аминокислотного остатка. В данный момент для оценки допустимости этого предположения используется нормальное распределение ошибки измерения разницы масс между фрагментными ионами y и b серий, полученное при статистическом анализе спектров

$$p(\delta) = \exp(-\delta^2 / 2\sigma^2), \quad (2)$$

где δ — наблюдаемая погрешность точной массы аминокислотного остатка для интервала между пиками; σ — численно оцениваемое среднеквадратичное отклонение измерения точной массы аминокислотного остатка для пиков, отнесенных к y и b сериям ионов на основании результатов распознавания X!Tandem.

Оценку Peptide Sequence Tag (PST) в целом будем строить как произведение оценок всех вершин и ребер графа масс-спектра, вошедших в PST. В терминах теории вероятностей это соответствует совпадению событий включения в путь PST вершин и ребер графа. На взгляд авторов, это соответствует нарастанию вероятности ошибки при увеличении длины PST. Таким образом, итоговая оценка построенного PST из n пиков будет

$$P_{tag} = \left(\prod_{j=1}^{n} Q_j \right) \left(\prod_{j=1}^{n-1} p(\delta_j^{j+1}) \right). \quad (3)$$

Заметим, что, хотя оценки выполнены в терминах теории вероятностей, процедура построения оценок содержит ряд нестрогих предположений, начиная с предположения об истинности результатов распознавания пептидов системой X!Tandem, которые позволяют говорить об оценках, но не о вероятностях в строгом значении этого слова.

ПОСТРОЕНИЕ PST

После описания и оценки графа масс-спектра переходим к собственно построению PST. Гипотеза об отнесении сигнала масс-спектра к PST включает предположения о том, что:

a) пик относится к одной из основных серий ионов b или y — оценка предположения Q_j;

b) в спектре есть пик, соответствующий следующему иону той же серии в направлении возрастания масс, — оценка предположения $p(\delta_j^{j+1})$;

c) в спектре есть пик, соответствующий следующему иону той же серии в направлении убывания масс, — оценка предположения $p(\delta_{j-1}^{j})$.

Если рассматривается только одно из двух последних условий, то данный пик — это конечный пик в PST.

Построим для каждого пика спектра полный набор структур данных, соответствующих всем гипотезам о включении пика в PST, в том числе и для завершения PST этим пиком. Оценим каждую из этих гипотез как произведение оценок пика и квадратных корней оценок интервалов

$$G_j = \sqrt{p(\delta_{j-1}^{j})} Q_j \sqrt{p(\delta_j^{j+1})}. \quad (4)$$

Упорядочим получившийся набор структур по убыванию оценок. Заметим, что любой PST может быть представлен как цепочка таких структур, замкнутая со стороны убывания масс гипотезой, для которой не рассматривается предположение c), а со стороны возрастания масс гипотезой для которой не рассматривается предположение b). Оценка PST будет равна произведению оценок структур его составляющих.

Далее извлекаем структуры из упорядоченного списка и для каждой структуры строим все варианты цепочек структур с участием предыдущих извлеченных структур. Те цепочки, длина которых соответствует заранее заданному требуемому числу аминокислот в PST и которые завершены

по концам односторонними структурами, рассматриваем как итоговые версии PST для данного спектра.

Благодаря монотонному убыванию оценок рассматриваемых гипотез в каждый момент времени мы располагаем полным списком PST, построенных из гипотез с наивысшей оценкой. Алгоритм останавливается, когда получено заданное количество PST или по исчерпании списка структур. Таким образом, мы получаем заданное число PST заданной длины и избегаем нахождения и оценки всех возможных вариантов PST для данного спектра.

РЕЗУЛЬТАТ РАЗРАБОТКИ

При реализации описанного алгоритма, получившего название CrystalTag, были выполнены следующие задачи.

- Разработана и наполнена информацией реляционная база данных, содержащая информацию о последовательностях белков, МС-МС-спектры и результаты их интерпретации.
- Разработано приложение MS/MS IAnI Viewer для визуализации протеомных данных, МС-МС-спектров и результатов их интерпретации.
- Разработана программа CrystalStat для статистического анализа спектров и результатов их интерпретации, поставляющая статистические данные для настройки алгоритма CrystalTag.
- Разработана программа, реализующая алгоритм CrystalTag и сопоставляющая полученные PST с протеомной базой данных.

В качестве среды разработки использована Visual Studio 7.0. MS/MS IAnI Viewer реализован на языке C# v.1.0, остальные программы реализованы на языке C++ в пределах стандарта ANSI. Все вызовы функций API Win32 и Microsoft COM изолированы в отдельных модулях. База данных размещена на Microsoft SQL Server 2000.

При разработке алгоритма в качестве тестового набора был использован набор спектров, полученных на приборе Q-TOF Ultima в Институте системной биологии (Сиэтл, США) для триптического гидролизата модельной смеси 17 известных белков, содержащий 1389 спектров. Набор данных находится в свободном доступе на сайте http://sashimi.sourceforge.net/repository.html. Файл данных предоставлен в открытом формате mzXML [17] и был приведен к текстовому формату .pkl при помощи утилиты mzXML2Other. Спектры набора данных были подвергнуты фильтрации с целью раскрытия изотопных распределений молекул на основе алгоритма структурной декомпозиции масс-спектров, разработанного в институте [18].

Впоследствии работоспособность алгоритма была показана на наборе из 6048 спектров, полученном в результате ряда (ЖХ-МС-МС)-экспериментов на масс-спектрометре Bruker Esquire 4000 в процессе белковых анализов препарата митохондрий клеток сердца быка, проведенных в Институте биоорганической химии РАН.

СТАТИСТИКА ЗНАЧИМЫХ ПИКОВ В (МС-МС)-СПЕКТРАХ

Для оценки значимости критериев, используемых при оценке PST, использована программа CrystalStat. Здесь приведены результаты работы этой программы для двух описанных наборов спектров. В табл. 1 приведены условные вероятности реализации критериев для гипотез H_1, H_2 для наборов спектров, полученных на приборах Q-TOF Ultima и Bruker Esquire 4000. Сравнение результатов анализа особенно интересно, поскольку эти два прибора относятся к архитектурно разным классам масс-спектрометров.

Табл. 1. Условные вероятности реализации критериев для гипотез H_1 и H_2

Критерии A_i	Q-TOF Ultima		Bruker Esquire 4000	
	$P(A_i \mid H_1)$	$P(A_i \mid H_2)$	$P(A_i \mid H_1)$	$P(A_i \mid H_2)$
y↔b	0.4938	0.03545	0.326	0.03671
y↔a	0.1573	0.05489	0.04309	0.03507
b↔x	0.04335	0.03174	0.026	0.03117
b↔a	0.1965	0.2819	0.04045	0.03982
y↔x	0.08348	0.2819	0.01934	0.03982
$-H_2O$	0.3514	0.247	0.1865	0.0511
$-NH_3$	0.319	0.202	0.0775	0.03586

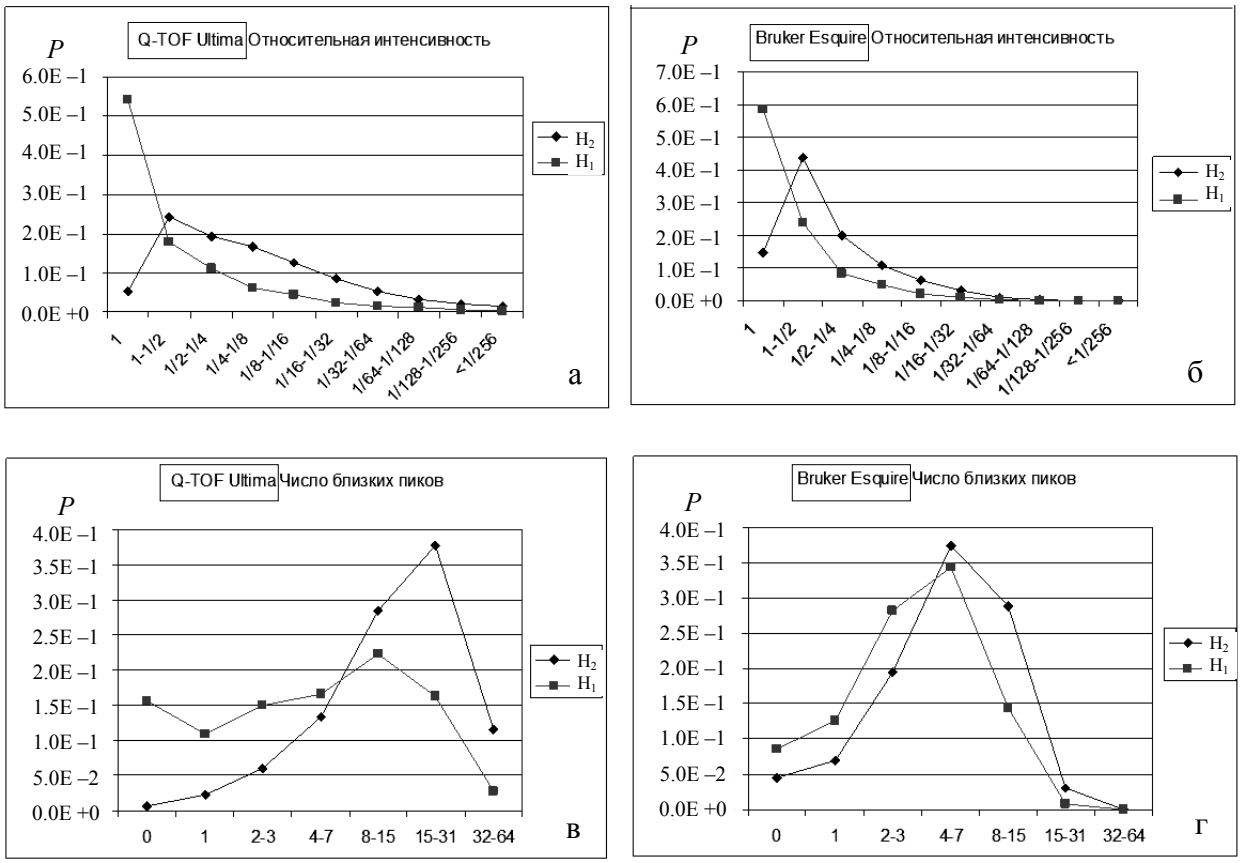


Рис. 2. Распределение условных вероятностей $P(\mathrm{A}_i \mid \mathrm{H}_1)$ и $P(\mathrm{A}_i \mid \mathrm{H}_2)$ для относительной интенсивности (а, б) и зашумленности (в, г) спектра для информативных и неинформативных ионов

Так, например, таблица показывает, что наличие нейтральных потерь $-H_2O$, $-NH_3$ является более значимым критерием для масс-спектрометра Bruker Esquire 4000, в то же время наличие пары сигналов типа y↔b — это сильный критерий для обоих приборов.

Рис. 2 показывает значимость критерия относительной интенсивности пиков и критерия числа соседствующих пиков в окрестности 50 Да от оцениваемого пика. Для обоих типов приборов эти два критерия оказываются существенно значимыми, т. е. интенсивные пики и пики, расположенные в незашумленных областях спектра, будут предпочтительны для построения PST. Тем не менее условная вероятность $P(\mathrm{A}_i \mid \mathrm{H}_1)$ для пиков, не доминирующих в своей окрестности, тоже может быть рассчитана, и эта вероятность отличается от нуля. Такие пики, возможно, получат меньшую общую оценку, однако не будут исключены из процедуры поиска PST.

Обращает на себя внимание тот факт, что в спектрах прибора Bruker Esquire информативные пики часто находятся среди множества неинформативных, поэтому критерий числа пиков в окрестности будет менее значимым для этих спектров, нежели для спектров, полученных на приборе Q-TOF Ultima.

Точность определения разностей масс, показанная на рис. 3, для исследованных наборов спектров соответствует ожидаемой для данных классов приборов. Средняя разница между пиками y и b серий для прибора Bruker Esquire будет несколько больше ожидаемой, что, видимо, связано с систематической погрешностью измерений при регистрации масс-спектров. Тем не менее полученная информация о дисперсии при измерении разницы масс между информативными ионами пригодна для использования при распознавании PST.

РЕЗУЛЬТАТЫ РАБОТЫ АЛГОРИТМА CRYSTALTAG

Полученные наборы оценок критериев были использованы при обработке указанных массивов спектров алгоритмом CrystalTag. Обработка проводилась на рабочей станции с процессором Intel Pentium M 1.73 GHz и 1 Gb RAM.

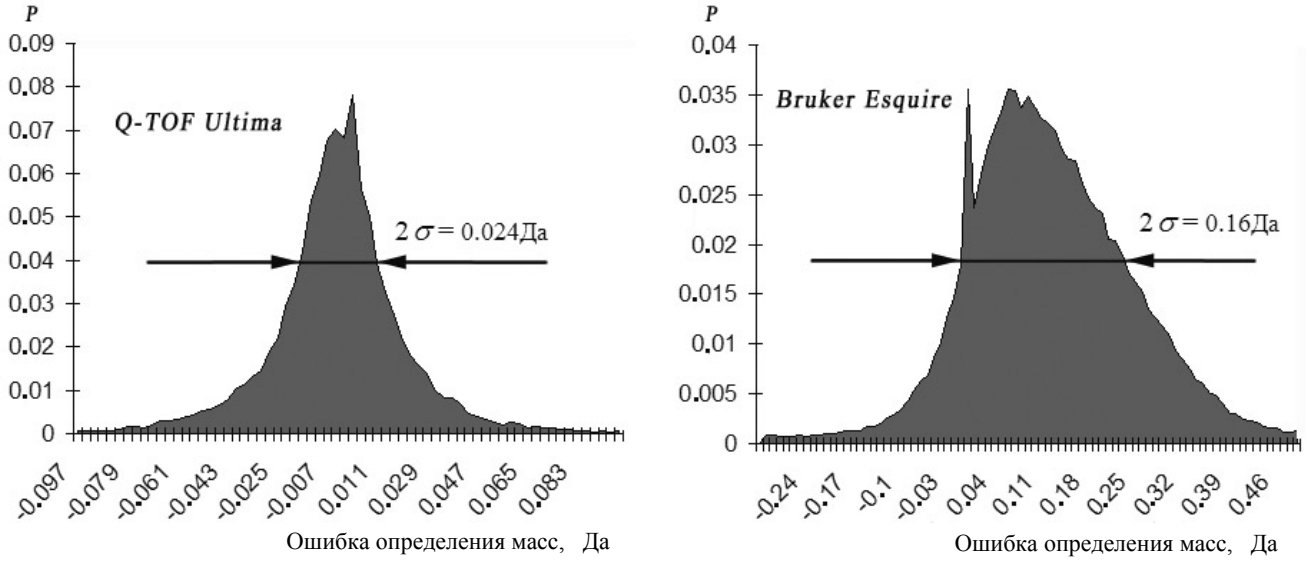


Рис. 3. Ошибка измерения разницы масс в масс-спектрах дочерних ионов

Табл. 2. Результаты тестирования алгоритма CrystalTag

№ п/п	Характеристика	Прибор Q-TOF Ultima, число спектров 1382, смесь: 18 модельных белков			Прибор Bruker Esquire, число спектров 6048, идентифицировано около 50 белков		
1.	Число PST / число аминокислот в PST	5/5	20/5	5/4	5/5	20/5	5/4
2.	Всего спектров, содержащих PST	627	627	773	3484	3484	4235
3.	Среднее время работы алгоритма CrystalTag, мс	8.05	17.27	5.61	5.16	14.32	3.26
4.	Опознанных спектров по X!Tandem	266	266	266	1392	1392	1392
5.	PST, совпадающих с X!Tandem	193	205	219	696	814	876
6.	Спектров, сопоставленных белкам из результата X!Tandem с использованием PST	280	313	385	1157	1314	1539
7.	Спектров из числа опознанных X!Tandem, допускающих построение верного PST	214	214	231	983	983	1186
8.	Процент пептидов, совпадающих с X!Tandem	72.6%	77.1%	78.5%	50%	58.5%	62.9%
9.	Процент корректно распознанных PST	90.2%	95.8%	94.8%	70.8%	82.8%	73.9%

Основные результаты тестирования алгоритма CrystalTag находятся в сводной табл. 2. Число гипотез PST и число аминокислот в каждой гипотезе заданы как параметры для каждого запуска алгоритма. В целом алгоритм показывает лучшие результаты для данных, полученных на приборе Q-TOF Ultima. Это ожидаемый результат — меньшая точность определения масс и меньшая избирательность критериев оценки пиков для Bruker Esquire приводят к меньшему числу верных гипотез PST. Для всех представленных вариантов поиска, алгоритм показывает быстродействие в единицы миллисекунд на спектр. Для тех спектров, которые содержат достаточно длинные последовательности пиков для построения правильного PST, набор гипотез PST, предложенный CrystalTag, содержит

правильную гипотезу в более чем 90 % случаев для Q-TOF Ultima и 70 % случаев для Bruker Esquire. Кроме того, CrystalTag позволяет выделить из общего набора спектров большое число дополнительных спектров, не опознанных X!Tandem, которые также могут быть соотнесены с белками, входящими в исходную смесь.

Этот эффект приводит к тому, что несмотря на то, что некоторые из спектров, опознанных X!Tandem, не были удачно распознаны при поиске PST, покрытие аминокислотной белковой последовательности для этих двух методов остается тем же или даже увеличивается, как это показано в табл. 3. В этой таблице приведены последовательности двух белков из результатов анализа проб; подчеркиванием обозначены пептиды в составе белка, сопоставленные спектрам.

Видно, что участки последовательности, сопоставленные спектрам, как минимум совпадают или расширяются для метода поиска PST по сравнению с X!Tandem, т. е. поиск PST воспроизводит результаты поиска X!Tandem и даже расширяет их.

ЗАКЛЮЧЕНИЕ

Разработанный алгоритм CrystalTag частичного восстановления последовательности пептида по его фрагментному масс-спектру обладает рядом достоинств. В их числе следующие.

- Быстродействие. Благодаря оптимизированному по числу проверяемых гипотез способу анализа графа масс-спектра время обработки спектра алгоритмом CrystalTag составляет единицы миллисекунд, что намного меньше времени регистрации спектра на существующих тандемных масс-спектрометрах.
- Качество распознавания. Алгоритм распознает до 90 % актуально существующих PST в составе масс-спектров. Данные о числе спектров сопоставленных белкам, присутствующим в смеси, хорошо согласуются с потенциальными возможностями метода поиска PST, представленными в работах [8, 13].
- Универсальность. Предложенная процедура оценки модели фрагментации, реализованная в программе CrystalStat, позволяет использовать алгоритм для масс-спектров, полученных на масс-спектрометрах различной конструкции, использующих разные физические принципы и имеющих различные аналитические характеристики.
- Расширяемость. Байесова модель формирования оценки пиков позволяет легко вводить новые критерии, значимые для восстановления исходной последовательности пептидов.

Для решения проблемы быстрой идентификации белков методом поиска PST следующая задача — это оценка значимости полученного набора PST при сопоставлении его белковой аминокислотной последовательности.

Табл. 3. Покрытие последовательности белков для алгоритма CrystalTag и программы X!Tandem

Алгоритм	Q-TOF Ultima. Прекурсор алкалин фосфатазы (swiss-prot ID — PPB_ECOLI)	Bruker Esquire. Изоцитрат дегидрогеназа бычья (swiss-prot ID — IDHP_BOVIN)
X!Tandem	1234567890123456789012345678901234567890123456789 0: MKQSTIALALLPLLFTPVTKARTPEMPVLENRAAQGDITAPGGARRLTGD 50: QTAALRDSLSDKPAKNIILLIGDGMGDSEITAARNYAEGAGGFFKGIDAL 100: PLTGQYTHYALNKKTGKPDYVTDSAASATAWSTGVKTYNGALGVDIHEKD 150: HPTILEMAKAAGLATGNVSTAELQDATPAALVAHVTSRKCYGPSATSEKC 200: PGNALEKGGKGSITEQLLNARADVTLGGGAKTFAETATAGEWQGKTLREQ 250: AQARGYQLVSDAASLNSVTEANQQKPLLGLFADGNMPVRWLGPKATYHGN 300: IDKPAVTCTPNPQRNDSVPTLAQMTDKAIELLSKNEKGFFLQVEGASIDK 350: QDHAANPCGQIGETVDLDEAVQRALEFAKKEGNTLVIVTADHAHASQIVA 400: PDTKAPGLTQALNTKDGAVMVMSYGNSEEDSQEHTGSQLRIAAYGPHAAN 450: VVGLTDQTDLFYTMKAALGLK Покрытие последовательности 19%	1234567890123456789012345678901234567890123456789 0: MAGYLRVVRSLCRASGSGSAWAPRALTAPNLQEQPRRHYADKRIKVAKPV 50: VEMDGDEMTRIIWQFIKEKLILPHVDVQLKYFDLGLPNRDQTNDQVTIDS 100: ALATQKYSVAVKCATITPDEARVEEFKLKKMWKSPNGTIRNILGGTVFRE 150: PIICKNIPRLVPGWTKPITIGRHAHGDQYKATDFVVDRAGTFKVVFTPKD 200: GSGPKEWEVYNFPAGGVGMGMYNTDESISGFAHSCFQYAIQKKWPLYMST 250: KNTILKAYDGRFKDIFQAIFEKHYKTEFDKHKIWYEHRLIDDMVAQVLKS 300: SGGFVWACKNYDGDVQSDILAQGFGSLGLMTSVLVCPDGKTIEAEAAHGT 350: VTRHYREHQKGRPTSTNPIASIFAWTRGLEHRGKLDGNQDLIRFAQTLEK 400: VCVETVESGAMTKDLAGCIHGLSNVKLNEHFLNTSDFLDTIKSNLDRALG 450: QQ Покрытие последовательности 39%
CrystalTag	1234567890123456789012345678901234567890123456789 0: MKQSTIALALLPLLFTPVTKARTPEMPVLENRAAQGDITAPGGARR**LTGD** 50: **QTAALR**DSLSDKPAKNIILLI**GDGMGDSEITAAR**NYAEGAGGFFKGIDAL 100: PLTGQYTHYALNKKTGKPDYVTDSAASATAWSTGVKTY**NGALGVD**IHEKD 150: H**PTILEMAK**AAGLATGNVSTA**ELQDAT**PAALVAHVTSRKCYGP**SATSEK**C 200: PGNALEKGGK**GSITEQLLNA**R**ADVTLG**GGAKTF**AETATA**GEWQGKTLREQ 250: AQARGYQLVSDAASLNSVTEANQQKPLLGLFADGNMPVRWLGPKATYHGN 300: IDKPAVTCTPNPQRNDSVPTLAQMTDKAIELLSKNEKGFFL**QVEGASID**K 350: QDHAANPCGQIGETVDLDEAVQRALEFAKKEGNTLVIVTADHAHASQIVA 400: PDTKA**PGLTQALNT**KDGAVMVMSYGNSEEDSQEHTGSQLRIAAYG**PHAAN** 450: VVGLTDQTDLFYTMKAALGLK Покрытие последовательности 32%	1234567890123456789012345678901234567890123456789 0: MAGYLRVVRSLCRASGSGSAWAPRALTAPNLQEQPRRHYADKRIKVA**KPV** 50: **VEMD**GDEMTRIIWQFIKEKLI**LPHVDVQL**KYFDLGLPNRDQTND**QVTIDS** 100: **ALATQK**YSVAVKCATITPDEARVEEFKLKKMWKSPNGTIRNI**LGGTV**FR**E** 150: **PIIC**KNIPRL**VPGWTKP**ITIGRHAHGDQYKATDFVVDRAGTFKVVFTPKD 200: GSGPKEWEVYNFPAGGVGMGMYNTDESISGFAHSCFQYAIQKK**WPLYM**ST 250: KNTILKAYDGRF**KDIFQAIFE**KHYKTEFDKHKIWYEHRLI**DDMVAQV**LKS 300: SGGFVWACKNYDGDVQSDILAQGFGSLGLMTSVLVCPDGKTI**EAEAAHGT** 350: VTRHYREHQKGRPTSTN**PIASIF**AWTRGLEHRGK**LDGNQD**LIRFAQTLEK 400: VCV**ETVESGA**MTKDLAGCIHGLSNVKLNEH**FLNTSDFLD**TIKSNLDRALG 450: QQ Покрытие последовательности 43%

СПИСОК ЛИТЕРАТУРЫ

1. *Perkins D.N., Pappin D.J., Creasy D.M., Cottrell J.S.* Probability-based protein identification by searching sequence databases using mass spectrometry data // Electrophoresis. 1999. V. 20, N 18. P. 3551–3567.
2. *Eng J.K., Ashley L.McCormack, Yates J.R.* An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database // JASMS. 1994. V. 5, N 11. P. 976–989.
3. *Craig R., Beavis R.C.* TANDEM: matching proteins with tandem mass spectra // Bioinformatics. 2004. V. 20, N 9. P. 1466–1467.
4. *Ma B. et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry // Rapid Commun. Mass Spectrom. 2003. V. 17, N 20. P. 2337–2342.
5. *Taylor J.A., Johnson R.S.* Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry // Anal. Chem. 2001. V. 73, N 11. P. 2594–2604.
6. *Mann M., Wilm M.* Error-tolerant identification of peptides in sequence databases by peptide sequence tags // Anal. Chem. 1994. V. 66, N 24. P. 4390–4399.
7. *Tabb D.L., Saraf A., Yates J.R.* GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model // Anal. Chem. 2003. V. 75, N 23. P. 6415–6421.
8. *Tanner S. et al.* InsPecT: identification of post-translationally modified peptides from tandem mass spectra // Anal. Chem. 2005. V. 77, N 14. P. 4626–4639.
9. *Savitski M.M., Nielsen M.L., Zubarev R.A.* New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques // Mol. Cell Proteomics. 2005. V. 4, N 8. P. 1180–1188.
10. *Sunyaev S., Liska A.J., Golod A., Shevchenko Anna, Shevchenko Andrej.* MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry // Anal. Chem. 2003. V. 75, N 6. P. 1307–1315.
11. *Shevchenko A., Sunyaev S., Loboda A. et al.* Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching // Anal. Chem. 2001. V. 73, N 9. P. 1917–1926.
12. *Tsur D., Tanner S., Zandi E., Bafna V., Pevzner P.A.* Identification of post-translational modifications by blind search of mass spectra // Nat. Biotechnol. 2005. V. 23, N 12. P. 1562–1567.
13. *Frank A., Tanner S., Bafna V., Pevzner P.* Peptide sequence tags for fast database search in mass-spectrometry // J. Proteome Res. 2005 V. 4, N 4. P. 1287–1295.
14. *Bartels C.* Fast algorithm for peptide sequencing by mass spectrometry // Biomedical and Environmental Mass. Spectrometry. 1990. V. 19. P. 363–368.
15. *Dijkstra E.W.* A note on two problems in connection with graphs // Numer. Math. 1959. V. 1. P. 269–271.
16. *Hart P., Nilsson N., Raphael B.* A formal basis for the heuristic determination of minimum cost paths' // IEEE Trans. on Systems Science and Cybernetics. 1968. V. 4, N 2. P. 100–107.
17. *Pedrioli P.G., Eng J.K., Hubley R. et al.* A common open representation of mass spectrometry data and its application to proteomics research // Nat. Biotechnol. 2004. V. 22, N 11. P. 1459–1466.
18. *Макаров В.В., Самокиш А.В., Лютвинский Я.И.* Метод извлечения значимой информации из масс-спектров пептидов // Научное приборостроение. 2004. Т. 14, № 2. С. 96–104.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 27.06.2006.

PARTIAL SEQUENCING OF PEPTIDES USING FRAGMENT MASS SPECTRA: ALGORITHM AND TESTING

Ya. I. Lutvinsky, V. V. Makarov, N. V. Krasnov, E. P. Podolskaya, A. N. Verentchikov

Institute for Analytical Instrumentation RAS, Saint-Petersburg

A new algorithm is developed for partial sequencing of peptides using MS/MS spectra for peptide sequence tag search. To accelerate the algorithm, it is optimized by reducing the number of tested sequence hypotheses. To evaluate sequence hypotheses a Bayesian approach is employed, based upon the actual statistics of mass specral peaks. The algorithm has been tested using data from a tandem Q-TOF MS as well as from Bruker Esquire mass spectrometer based on the Pauli ion trap.