

УДК 621.384.668.8: 577.112.6

© Я. И. Лютвинский, Н. В. Краснов

РАЗРАБОТКА CRYSTALTAG — АЛГОРИТМА ЧАСТИЧНОГО РАСПОЗНАВАНИЯ ФРАГМЕНТНЫХ МАСС-СПЕКТРОВ ПЕПТИДОВ

Статья представляет созданный в Институте аналитического приборостроения РАН оригинальный алгоритм частичного распознавания последовательности пептидов по масс-спектру продуктов фрагментации. Алгоритм предназначен для решения задач белкового и/или пептидного анализа посредством масс-спектрометрии и является одним из ключевых алгоритмов биоинформационного обеспечения МС-эксперимента. Первичное тестирование алгоритма показало приемлемое качество распознавания фрагментных масс-спектров пептидов. Алгоритм носит универсальный характер и может быть использован как для приборов, разработанных в ИАиП РАН, так и для импортных масс-спектрометров, особенно тандемов МС-МС.

ВВЕДЕНИЕ

Одной из наиболее динамично развивающихся областей современной молекулярной биологии является протеомика — исследование белкового пула организма (протеома) как единого целого. К числу ведущих методологий в протеомных исследованиях относится тандемная масс-спектрометрия высокого разрешения с мягкими методами ионизации [1]. На нужды протеомики ориентирована в значительной степени разработка новых современных тандемных масс-спектрометров [2–4]. Появление новых приборов вызывает необходимость в разработке новых методов обработки информации, получаемой на этих приборах.

Как правило, масс-спектрометры в протеомике используются для анализа смесей белков, представленных продуктами избирательного ферментативного гидролиза. Получаемые масс-спектры представляют собой масс-спектры пептидов — продуктов гидролиза. Важнейшая задача при обработке данных — это восстановление последовательности пептида по его фрагментному спектру. При решении этой задачи используется ряд подходов, получивших в мировой литературе специфические названия.

— *Fragment Map*. Сопоставление приборных масс-спектров с теоретическими спектрами продуктов фрагментации пептидов, построенными на основе протеомных баз данных. Подход реализован в таких продуктах, как Mascot [5], X!Tandem [6], и многих других.

— *de Novo Sequencing*. Полная интерпретация последовательности пептида по его фрагментному спектру. Этот способ используют такие программы, как Peaks Studio [7], Lutefisk [8].

— *Sequence Tag*. Частичная интерпретация последовательности пептида по его масс-спектру [9].

Несмотря на то что все эти алгоритмы решают

сходную задачу, методы их реализации сильно отличаются друг от друга. Алгоритм CrystalTag относится к алгоритмам распознавания Sequence Tag, поэтому подробнее опишем происхождение этого термина.

Впервые этот термин был введен в работе М. Манна (*Mann M.*) [10]. Группа ученых обнаружила в масс-спектрах фрагментов пептидов, полученных при контролируемом столкновении пептидов с молекулами газа, устойчивые последовательности пиков, расстояния между которыми в массовой шкале соответствовали массе аминокислотных остатков. Такая последовательность пиков соответствовала последовательности аминокислот в пептиде. Как правило, последовательность пиков покрывает не весь спектр и соответственно охватывает не всю последовательность аминокислот.

Sequence Tag принято записывать как массу первого пика в последовательности, последовательность аминокислот, соответствующую расстояниям между пиками последовательности, и разницу между последним пиком последовательности и массой родительского иона. Например, Sequence Tag на рис. 1 можно обозначить как [611.30]LGAD[242.05].

Таким образом, Sequence Tag — это последовательность пиков во фрагментном масс-спектре пептида, трактуемая как часть последовательности аминокислот за счет отнесения пиков последовательности к одной серии фрагментных ионов пептида. Когда говорят о поиске и распознавании Sequence Tag, тем самым говорят о решении задачи частичной интерпретации фрагментного масс-спектра.

Надо отметить, что по последовательности пиков в общем случае не удастся установить направление последовательности — от N-конца к C-концу или наоборот, поскольку неизвестно, ионы

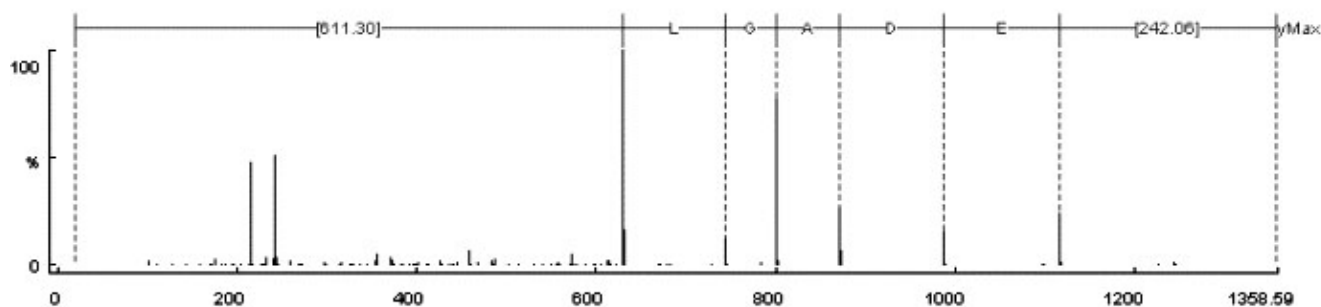


Рис. 1. Спектр пептида IEEDAGLGNGGLGR с выделенным Sequence Tag [611.30]LGADE[242.05]

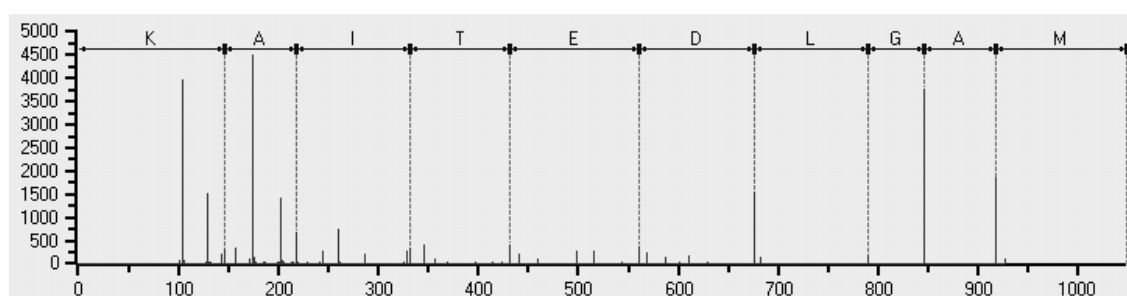


Рис. 2. Спектр пептида MAGLDETIK с выделенной у-серией ионов

какой серии фрагментов (a, b, или y в нотации, впервые предложенной П. Роепсторффом (*Roepstorff P.*) в 1984 году [11]) представлены пиками. Также на основании Sequence Tag не удастся различить аминокислоты лейцин и изолейцин с одинаковой молекулярной массой и часто (при недостаточной массовой точности прибора) не удастся различить аминокислоты глутамин и лизин с близкими молекулярными массами.

Техника распознавания Sequence Tags широко вошла в практику протеомных исследований. Некоторые стратегии использования масс-спектрометрической информации построены на основе данного подхода [12, 13, 14]. Особенно полезны Sequence Tag при исследовании организмов с несеквенированным геномом, поскольку метод интерпретации Sequence Tag не требует поиска в протеомных либо геномных базах данных. Однако до сих пор в большинстве случаев исследователи прибегают к ручному распознаванию Sequence Tags, хотя это и приводит к значительному количеству монотонной работы. Объясняется это как тем, что биологи более доверяют собственному экспертному опыту, нежели данным компьютерной обработки, так и отсутствием адекватных алгоритмов для распознавания Sequence Tag.

ПОДХОД К ПРОБЛЕМЕ

На первый взгляд расшифровка Sequence Tag представляется тривиальной задачей: нужно лишь найти серию интенсивных пиков, расстояние между которыми в массовой шкале соответствует массе аминокислотных остатков, — это и будет искомым Sequence Tag. На деле же кроме сигналов, составляющих истинный Sequence Tag, в спектре есть множество сигналов, расстояния между которыми также соответствуют массам аминокислот. Хорошо, если есть какие-то неоднозначности: казалось бы, можно составить весь список возможных последовательностей и среди них выбирать лучшие. Однако масс-спектр на рис. 2 показывает, что это не так. Это довольно качественный спектр, в котором у-серия ионов представлена полностью, и, казалось бы, его расшифровка не должна составлять никакого труда. Тем не менее из-за зашумленности областей средних и низких масс каждому пику в этих областях соответствует 3–8 пиков в направлении как низких, так и высоких масс с разницей, соответствующей какому-либо аминокислотному остатку. По этой причине в спектре можно обнаружить более 400 000 различных вариантов расшифровки последовательности.

стей, в том числе, например, более 28 000 только Sequence Tag длиной 5 аминокислот.

На генерацию и оценку всех этих вариантов тратится процессорное время и ресурс памяти. Таким образом, алгоритм с разумными требованиями по времени выполнения и используемой памяти должен быть построен так, чтобы избежать перебора всех возможных вариантов и в результате своей работы выдать только набор наиболее вероятных последовательностей для данного спектра.

РАЗРАБОТКА АЛГОРИТМА CRYSTALTAG

В рамках представленной работы был создан алгоритм частичного распознавания последовательности аминокислот в пептиде, получивший название CrystalTag. В этом алгоритме Sequence Tag строятся, начиная с наиболее высокооцениваемых пиков. Таким образом, они как бы кристаллизуются вокруг пиков, чья вероятность войти в тэг наиболее высока, — это и послужило причиной назвать алгоритм CrystalTag.

Использование только наиболее высоко оцениваемых пиков и совпадений по разнице масс с точностью, адекватной массовой точности прибора, дало возможность избежать задачи перебора всех возможных вариантов.

Алгоритм основан на методологическом базисе теории вероятностей и теории графов. Основными входными параметрами алгоритма являются длина последовательности аминокислот и максимальное количество тэгов как результат работы алгоритма. Тэги, найденные алгоритмом, — это варианты Sequence Tag, составленные из наиболее высокоранговых пиков по данному спектру.

Алгоритм реализован на Visual C++ 7.0. При реализации алгоритма использовались только стандартные библиотеки C++, такие как STL, так что алгоритм может быть легко перенесен на любую другую платформу, для которой существует компилятор C++.

Спектральная информация, как и информация по распознаванию Sequence Tag, размещены в реляционной базе данных MSSQL.

Для просмотра спектров и результатов распознавания на языке C# реализована утилита IAnI MS/MS Browser, обладающая графическим пользовательским интерфейсом, снимки экранов которой использованы в этой статье.

Для оценки каждого из пиков были использованы такие критерии, как:

- относительная интенсивность пика в его окрестности в интервале ± 50 Да;
- зашумленность спектра вокруг пика в его окрестности в интервале ± 50 Да;
- наличие в спектре пиков, парных данному, по правилам построения серий ионов — $y \leftrightarrow b$, $x \leftrightarrow a$ и т. д.;

— наличие характерных нейтральных потерь для $-H_2O$, $-NH_3$, $-CO$ и т. д.

Кроме того, использовались оценки интервалов между пиками — точность совпадения измеренной разницы между пиками с точной массой аминокислотных остатков. При этом учитывались:

- приборная точность определения массы фрагментных ионов;
- абсолютная величина разницы между пиками;
- приборная точность определения массы родительских ионов.

ПЕРВИЧНОЕ ТЕСТИРОВАНИЕ АЛГОРИТМА CRYSTALTAG

Алгоритм был протестирован на наборе данных, полученных на приборе Q-TOF Ultima в Институте системной биологии (Сиэтл, США) для триптического гидролизата смеси 17 известных белков, содержащем 1389 спектров. Набор данных находится в свободном доступе на сайте [15]. Файл данных предоставлен в открытом формате mzXML [16], и был приведен к текстовому формату .pkl при помощи утилиты mzXML2Other.

Спектры набора данных были подвергнуты фильтрации на основе алгоритма структурной декомпозиции масс-спектров [17].

В качестве референтного метода для проверки результатов поиска Sequence Tags использовалась программа X!Tandem (Университет Манитобы, Канада). В качестве протеомной базы данных для распознавания белков использовалась локальная копия базы данных Swiss-prot [18], находящейся в открытом доступе. Все 17 белков исходной смеси описаны в этой базе данных.

Полученные Sequence Tags были сопоставлены с результатами распознавания спектров программой X!Tandem.

В таблице приведены результаты сравнения, полученные для трех запусков алгоритма Sequence Tag с различными параметрами по числу генерируемых Sequence Tags и по количеству аминокислотных остатков, включаемых в результат распознавания.

Эксперимент проводился на рабочей станции на основе процессора Intel Celeron с тактовой частотой 2.4 ГГц с 256 мегабайтами оперативной памяти. Среднее время работы алгоритма составило 10–15 миллисекунд на спектр.

В первую очередь, обращает на себя внимание, что число спектров с распознанными Sequence Tags намного больше, чем число спектров, распознанных X!Tandem. Это может быть объяснено двумя причинами.

1. Некоторая часть спектров соответствует пептидам с нетривиальными сайтами гидролиза либо

Проверка алгоритма CrystalTag

Параметр оценки	5 лучших из 5 аминокислот	20 лучших из 5 аминокислот	5 лучших из 4 аминокислот
Всего спектров с Sequence Tags	627	627	773
Опознанных спектров по X!Tandem	266	266	266
Sequence Tags, совпадающих с X!Tandem	186	203	209
Sequence Tags, не совпадающих с X!Tandem	54	37	47
Спектров по X!Tandem без обнаруженных Sequence Tag	26	26	10
Спектров без серий ионов для Sequence Tag	52	52	35
Процент правильно опознанных пептидов в сравнении с X!Tandem	69.9%	76.3%	78.5%
Процент правильно опознанных Sequence Tags	86.2%	94.9%	90.5%

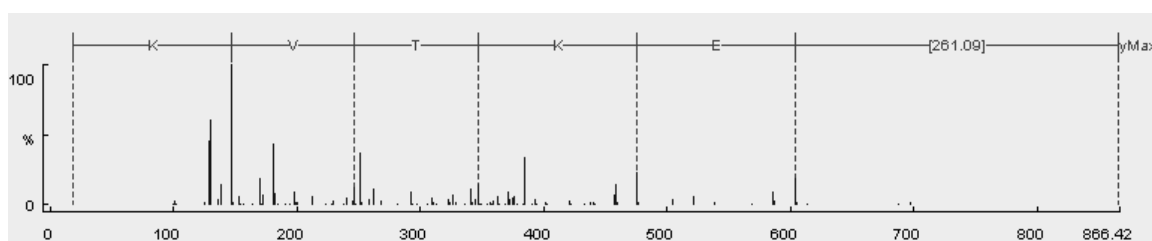


Рис. 3. Спектр пептида NMEQTVK+16 Да с Sequence Tag KVTKE[260.07]

модифицированным. Спектры таких пептидов могут быть сопоставлены с белками исходной смеси при помощи поиска в базах данных по тэгу, допускающему при составлении одну ошибку. Например, спектр на рис. 3, нераспознанный при помощи X!Tandem, имеет Sequence Tag KVTKE[260.07], что соответствует триптическому пептиду NMEQTVK из белка MYH4_RABIT с массой, измененной на +16 Да, что, скорее всего, соответствует окисленному метионину во второй позиции этого пептида. Такой белок действительно есть в исходной смеси, и, видимо, исходный пептид действительно был окислен.

2. Некоторая часть спектров не является спектрами пептидов, и все обнаруженные Sequence Tags являются ложными. CrystalTag обнаруживает любые последовательности пиков, которые можно трактовать как Sequence Tag. Возможно, число таких спектров можно будет уменьшить, введя порог по оценке Sequence Tags, чтобы исключить тэги с пренебрежимо малыми вероятностями.

Тем не менее визуальный контроль спектров показывает, что, как правило, тэги обнаруживаются в спектрах, которые выглядят информативными, имеют достаточно большое количество сигналов в большом массовом и динамическом диапазонах. Хочется надеяться, что ответить на вопрос о природе и полезности спектров могут дальнейшие исследования.

17 дополнительных верных Sequence Tags, обнаруженных при увеличении числа распознанных тэгов до 20, показывают неидеальность модели оценки Sequence Tags алгоритма. В этих случаях правильный Sequence Tag имеет ранг между 6-м и 20-м.

Заметим, что в некоторых спектрах Sequence Tags действительно нет. Так, 26 спектров, опознанных X!Tandem, в которых CrystalTag не нашел Sequence Tags, действительно не имеют серий ионов, которые позволили бы образовать Sequence Tag, включающий 5 аминокислот. В целом же, CrystalTag обнаружил около 90% действительно существовавших Sequence Tags, что охватывает 70–79% пептидов, распознанных X!Tandem.

ЗАКЛЮЧЕНИЕ

В этой статье приведены первые данные по характеристике алгоритма CrystalTag. Алгоритм все еще находится в процессе разработки и совершенствования.

Так, например, на момент написания статьи значимость критериев оценки Sequence Tag выставлена методом экспертной оценки. При помощи референтных методов, таких как X!Tandem, можно построить статистическую модель для автоматического определения параметров алгоритма

в соответствии с конкретным используемым масс-спектрометром. В принципе, в качестве референтного метода для настройки алгоритма может быть использована также и проверка получаемых Sequence Tag по протеомным базам данных. В этом случае точная настройка алгоритма будет происходить уже в процессе его работы.

Тем не менее существующий алгоритм работоспособен и может быть использован в практических протеомных исследованиях тремя основными способами.

1. Использование точного Sequence Tag для индексирования баз данных может дать хороший инструмент для быстрой идентификации смесей белков методами масс-спектрометрии.

2. Использование алгоритмов, устойчивых к однократным ошибкам в Sequence Tag, может дать способ идентификации спектров модифицированных и нетипично гидролизированных пептидов, что увеличит процент использования информации, полученной в ходе масс-спектрометрического эксперимента.

3. Скорость распознавания Sequence Tag и небольшой процент потерь информативных спектров позволяют использовать CrystalTag как фильтр перед передачей массивов спектров программам библиотечного поиска, таким как Mascot.

Впрочем разработка методологического и алгоритмического обеспечения использования Sequence Tag для поиска в протеомных базах данных при характеристике белковых смесей — это тема отдельного исследования.

СПИСОК ЛИТЕРАТУРЫ

1. Mann M., Hendrickson R.C., Pandey A. Analysis of proteins and proteomes by mass spectrometry // *Annu. Rev. Biochem.* 2001. V. 70. P. 437–473.
2. Веренчиков А.Н. Параллельный (МС-МС)-анализ во времяпролетном тандеме... // *Научное приборостроение*. 2004. Т. 14, № 2. С. 24–37.
3. Horning S., Malek R., Wieghaus A. et al. A Hybrid two-dimensional quadrupole ion trap-Fourier transform ion cyclotron mass spectrometer: Accurate mass and high resolution at a chromatography timescale // *Extended abstract on ASMS conference 2003*. (www.asms.org).
4. Hu Qizhi, Noll Robert J., Li Hongyan, Makarov Alexander, Hardman Mark, Cooks R. Graham The orbitrap: a new mass spectrometer // *Journal of Mass Spectrometry*. 2005. V. 40, N 4. P. 430–443.
5. 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.
6. Craig R., Beavis R.C. TANDEM: matching proteins with tandem mass spectra // *Bioinformatics*. 2004, Jun. 12. V. 20, N 9. P. 1466–1467.
7. Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry // *Rapid Commun. Mass Spectrom.* 2003. V. 17, N 20. P. 2337–2342.
8. Taylor J.A., Johnson R.S. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry // *Anal Chem.* 2001, Jun. 1. V. 73, N 11. P. 2594–2604.
9. Tabb D.L., Saraf A., Yates J.R. 3rd GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model // *Anal Chem.* 2003, Dec. 1. V. 75, N 23. P. 6415–6421.
10. Mann M., Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags // *Anal Chem.* 1994, Dec. 15. V. 66, N 24. P. 4390–4399.
11. Roepstorff P., Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides // *Biomed Mass Spectrom.* 1984, Nov. V. 11, N 11. P. 601.
12. Sunyaev S., Liska A.J., Golod A., Shevchenko A., Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry // *Anal Chem.* 2003, Mar. 15. V. 75, N 6. P. 1307–1315.
13. Liska A.J., Shevchenko A., Pick U., Katz A. Enhanced photosynthesis and redox energy production contribute to salinity tolerance in *Dunaliella* as revealed by homology-based proteomics // *Plant Physiol.* 2004, Sep. V. 136, N 1. P. 2806–2817.
14. Shevchenko A., Sunyaev S., Loboda A., Shevchenko A., Bork P., Ens W., Standing K.G. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching // *Anal Chem.* 2001, May. 1. V. 73, N 9. P. 1917–1926.
15. (<http://sashimi.sourceforge.net/repository.html>).
16. 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, Nov. V. 22, N 11. P. 1459–1466.
17. Макаров В.В., Самокиш А.В., Лютовинский Я.И. Метод извлечения значимой информации из масс-спектров пептидов // *Научное приборостроение*. 2004. Т. 14, №2 С. 96–104.
18. (<http://www.expasy.org>).

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 8.07.2005.

DEVELOPMENT OF THE CRYSTALTAG ALGORITHM FOR PARTIAL RECOGNITION OF PEPTIDE FRAGMENT MASS SPECTRA

Ya. I. Lyutvinsky, N. V. Krasnov

Institute for Analytical Instrumentation RAS, Saint-Petersburg

This article describes an original sequence tag search algorithm developed at the Institute for Analytical Instrumentation, RAS. This algorithm is intended for solving the tasks of protein and/or peptide analysis by means of mass spectrometry. It is one of the fundamental bioinformatic algorithms for mass spectrometry data analysis and interpretation. Initial testing of the algorithm has shown good quality of peptide fragment mass spectrum recognition. The algorithm is of a universal nature and can be applied both to instruments developed by IAI and third-party mass spectrometers, especially MS/MS tandems.