УДК 577.112:004.021

ВОССТАНОВЛЕНИЕ АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ЦИКЛИЧЕСКИХ ПЕПТИДОВ ИЗ МАСС-СПЕКТРОВ

© 2014 г. **Э.С. Фомин**

Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: fomin@bionet.nsc.ru

Поступила в редакцию 4 сентября 2014 г. Принята к публикации 6 октября 2014 г.

Метод масс-спектрометрии − один из физических методов исследования протеомов различных организмов, позволяющий решать как задачи идентификации биологических макромолекул, так и секвенирования пептидных цепочек в случаях, когда нет информации о геномах либо эта информация крайне ограничена. В настоящее время существует множество компьютерных программ для поддержки исследований в этой области. Тем не менее, несмотря на высокую активность, имеется только незначительный прогресс в создании программ, позволяющих решать задачи *de novo* секвенирования для циклических пептидов, к которым относятся наиболее эффективные антибиотики, противоопухолевые агенты, иммунодепрессанты, токсины и множество пептидов с неизвестными функциями, синтезируемые в клетке по нерибосомальному пути. Предложен эффективный алгоритм для решения задачи секвенирования циклических пептидов, который позволяет восстанавливать последовательности большой (до 160 аминокислотных остатков) длины.

Ключевые слова: масс-спектрометрия, секвенирование циклических пептидов, проблема beltway.

ВВЕДЕНИЕ

Протеомика – наука о протеомах – долгое время развивалась благодаря методу электрофореза, который позволяет разделить макромолекулы, различающиеся по молекулярной массе, пространственной конфигурации и электрическому заряду за счет разной скорости их диффундирования в буферном растворе под действием электрического тока (Остерман, 1981). Этот метод используют почти в каждой биохимической лаборатории, но для больших протеомных проектов в настоящее время определяющую роль играют другие, более мощные физические методы исследования, такие как ядерный магнитный резонанс (ЯМР) и метод масс-спектрометрии (МС).

Метод ЯМР основан на поглощении электромагнитной энергии ЯМР чувствительными ядрами (такими как, например, ^{1}H или ^{13}C) в сильном магнитном поле. Поскольку разные атомы в ближайшем окружении любого ядра по-разному экранируют внешнее магнитное поле, то положение резонансных линий одних и тех же ядер в зависимости от локального окружения различается, и по величине сдвига резонансных линий можно судить о том, какие атомы и на каком расстоянии от ЯМР-чувствительных ядер они находятся. Метод позволяет решать не только задачи идентификации, как метод электрофореза, но и задачи восстановления первичной нуклеотидной и аминокислотных последовательностей, определения пространственной структуры белков (Wuthrich, 1986), и, более того, он позволяет исследовать динамические свойства молекул − константы скорости химических реакций и величины энергетических барьеров внутримолекулярного вращения (Lambert *et al.*, 2000). Тем не менее, несмотря на широкие возможности, метод ЯМР еще не стал рабочим инструментом для каждой лаборатории из-за высокой цены спектрометров и ряда присущих ему недостатков, которые включают

требование обогащения образцов ЯМР-чувствительными радиоактивными ядрами ^{13}C, ^{15}N и ^{17}O, необходимость выполнения экспериментов в растворителе, где нет водородов (D_2H, CCl_4 и др.), и низкую чувствительность. Низкая чувствительность, как следствие, приводит к необходимости иметь большой объем очищенного исследуемого вещества (не менее миллиграмм) и к длительному времени проведения эксперимента для накопления статистики.

Метод МС в отличие от ЯМР для исследования протеомов различных организмов оптимален с точки зрения соотношения «затраты/результат». Он позволяет выполнить идентификацию белков и получить количественные соотношения между различными белками (Aebersold, Mann, 2003), выявить аминокислотный состав и их последовательность в пептидных цепочках (Hubbard, Jones, 2010). Белки идентифицируются обычно в одностадийном МС-эксперименте, когда массы белков, выявленные из масс-спектра («отпечатки пальцев»), сравнивают с полученными выборкой из компьютерных баз данными последовательностей (Pappin *et al.*, 1993). Для задач секвенирования используют техники тандемной МС, в которых индивидуальные пептиды, отселектированные и накопленные после первого этапа МС, на втором этапе подвергают дальнейшей фрагментации и анализируют полученные фрагменты. Количество этапов селектирования и фрагментации обозначают надстрочным индексом *n*, например, для тандемной масс-спектрометрии принято обозначение $МС^2$ или МС/МС, для спектрометрии с большим числом этапов − обозначение $МС^n$. Одним из существенных преимуществ МС/МС над методом ЯМР являются более низкие требования к количеству исследуемого вещества (достаточно пикограмм), что весьма важно для биологических экспериментов, где стоимость получения образцов высока.

Анализ большого количества масс-спектрометрических данных, получаемых в области протеомики, – узкое место многих проектов. Современные установки могут генерировать до десятков тысяч ионных фрагментов в час. Совокупность вычислительных задач, связанных с сопоставлением этих фрагментов с пептидными последовательностями, удалением шума, идентификацией пептидных цепочек и восстановлением последовательности пептидов, представляет собой серьезный вызов для биоинформатики. В силу статистической природы экспериментальных спектров, наличия в них пропусков или ложных выбросов решение вышеупомянутых задач не является строго однозначным и может приводить к ошибочным результатам, накапливающимся в литературе и базах данных, запуская процесс их деградации и затрудняя дальнейший анализ новых данных. По этой причине актуальность разработки современных подходов, валидации баз данных, развития существующих программ и увеличение их эффективности с течением времени только увеличиваются.

Биоинформационные подходы для анализа масс-спектров

В настоящее время существующие биоинформационные подходы для анализа масс-спектрометрических данных разбиты на две категории:

- идентификация макромолекул с использованием баз данных (поиск сходства между спектром ионного фрагмента и теоретическими и/или экспериментальными спектрами пептидных цепочек, сохраненными в библиотеках);
- *de novo* секвенирование (восстановление пептидных цепочек прямым образом из МС/МС спектров).

Для больших протеомных проектов использование баз данных является основным способом идентификации образцов, другие же стратегии дают привлекательную альтернативу в особых ситуациях, например, когда исследуемый образец ранее не был зафиксирован в базах данных. Множество различных программ, поддерживающих стратегию идентификации пептидных цепочек через использование баз данных, разработано к настоящему времени (Eng *et al.*, 1994; Clauser *et al.*, 1999; Perkins *et al.*, 1999; Zhang *et al.*, 2002; Colinge *et al.*, 2003; Craig, Beavis, 2004).

Программы загружают спектр фрагмента пептида и ранжируют его относительно теоретических спектров, конструируемых для пептидов из баз данных. Количество возможных

решений ограничено согласно критериям, задаваемым пользователем, таким как точность совпадения масс фрагментов, типы разрешенных посттрансляционных модификаций и прочее. Лучшие найденные решения подвергают дальнейшему контролю методами статистического анализа (Benjamini, Hochberg, 1995; Keller *et al.*, 2002; Storey, Tibshirani, 2003; Elias, Gygi, 2007). Ряд схем ранжирования решений, описанных в литературе, включает использование спектральных корреляционных функций (Eng *et al.*, 1994), количество сходных фрагментов (Perkins *et al.*, 1999; Craig, Beavis, 2004) и статистически вычисленные частоты их встречаемости (Colinge *et al.*, 2003) или использует эмпирически подобранные правила (Dančík *et al.*, 1999), полученные с помощью технологий машинного обучения.

К настоящему времени из-за большого объема систематических исследований протеомов огромного числа организмов и большого сходства между протеомами различных организмов велика вероятность того, что исследуемый образец либо его родственные формы уже когда-либо экспериментально были изучены и занесены в ту или иную базу данных. Это позволило использовать стратегии поиска, основанные на сравнении с ранее сделанными экспериментальными данными, и разработать соответствующие программы (Craig *et al.*, 2005, 2006; Frewen *et al.*, 2006). Описанные подходы существенно более быстры, чем генерация для задач сравнения образцов теоретических спектров, и могут стать первым эффективным фильтром в задачах идентификации.

Стратегии *de novo* секвенирования (прямое восстановление первичной последовательности пептидных цепочек из масс-спектров) используют с начала 2000-х гг. (Johnson, Taylor, 2002; Ma *et al.*, 2003; Frank, Pevzner, 2005). Их главное преимущество состоит в том, что они не заменимы в тех случаях, когда либо нет информации о геномах, либо эта информация существенно ограничена, либо если подходы, основанные на поиске аналогов в базах данных, в чем-то не сработали.

Таким образом, *de novo* секвенирование может применяться к белкам, которые имеют полиморфизмы, либо к искусственно модифицированным пептидным цепочкам. В отличие от задач идентификации с поиском по базам данных, стратегии *de novo* секвенирования пептидов чрезвычайно затратны в вычислительном плане и требуют масс-спектры высочайшего качества с минимальным количеством шума.

De novo секвенирование циклических цепочек

В настоящее время наблюдается большой прогресс в создании компьютерных программ для анализа и валидации спектров МС/МС (Nesvizhskii *et al.*, 2007; Allmer, 2011). Следует обратить внимание на то, что большинство программ для решения задачи *de novo* секвенирования пептидных последовательностей работает с линейными незамкнутыми цепочками. Это связано с преобладанием подобного рода биологических макромолекул в природе. Тем не менее замкнутые в кольцо пептиды также существуют. Например, такие пептиды представляют антибиотики, синтезируемые по нерибосомальному пути, − ванкомицин, даптомицил, тироцидин и прочие. Эти антибиотики необычны тем, что их синтез в почвенных микроорганизмах не следует основному постулату молекулярной биологии «от ДНК к матричной РНК и далее к пептидной макромолекуле», они вообще не закодированы в ДНК. Вместо этого в ДНК закодированы некоторые белки (синтетазы), которые и собирают эти антибиотики (Marahiel *et al.*, 1993; Sieber, Marahiel, 2005). Подобным способом закодирована широкая область циклических пептидов, которая включает не только антибиотики, но и противоопухолевые агенты, иммунодепрессанты, токсины и множество пептидов с неизвестными функциями. Интересно также и то, что большинство пептидов, синтезируемых по нерибосомальному пути, включает нестандартные остатки, например, тироцидин включает орницин (Orn), ванкомицин включает гликозилированные остатки, причем общее число таких строительных блоков потенциально может доходить до нескольких сотен.

De novo секвенирование циклических пептидов сталкивается с новой, неожиданной на первый взгляд, дополнительной сложностью по сравнению с *de novo* секвенированием линейных пептидов − вероятность успешного восстановления последовательности компью-

терными подходами с увеличением длины последовательности становится экспоненциально малой (Jaganathan, Hassibi, 2013). Сущность этой неожиданности разъясняется только с использованием математических подходов.

В математической постановке задача секвенирования последовательности сводится к давно известной среди математиков задаче восстановления множества целых чисел из мультимножества парных расстояний между ними, причем в этой задаче различают два четко выделенных случая:

- все точки исходного множества расположены на отрезке ограниченной длины (в англоязычной литературе на этот случай ссылаются как на проблему turnpike);
- все точки исходного множества расположены на кольце ограниченной длины (проблема beltway).

Следует заметить, что к данной математической постановке сводится множество задач из кристаллографии (Millane, 1990), астрономии (Dainty, Fienup, 1987), оптики (Walther, 1963; Jaganathan *et al.*, 2013), обработки сигналов (Rabiner, Juang, 1993) и даже теории музыки (Rahn, 1994). В биоинформатике задачи, сводимые к проблемам turnpike и beltway, обнаружены в задачах картирования сайтов рестрикции ДНК (Stefik, 1978; Allison, Yee, 1988; Pandurangan, Ramesh, 2002) и *de novo* секвенирования пептидов (Chen *et al.*, 2000; Mohimani *et al.*, 2011). Ввиду высокой практической значимости в 1977 г. эти проблемы были перечислены в списке основных проблем вычислительной геометрии (Shamos, 1977). Следует отметить, что до сих пор, несмотря на более полувековой интерес математиков к решению этих проблем, нет доказательства, принадлежат ли эти задачи к классу nondeterministic polinomial (NP), т. е. к классу задач, для которых не существует решения с полиномиальной вычислительной сложностью и построение эффективного алгоритма для произвольных данных невозможно, либо все же эти задачи не относятся к классу NP и построение эффективного алгоритма возможно.

Исчерпывающее решение обозначенных проблем дано в работах Skiena с соавт. (1990) и Lemke с соавт. (2003). Было показано, что число уникальных решений для одномерной проблемы turnpike находится в диапазоне

$$1/2n^{0,8107144} \le H_1(n) \le 1/2n^{1,2324827}$$

и для проблемы beltway в диапазоне

$$\exp\left(2^{\frac{ln\, n}{ln\, ln\, n} + o(1)}\right) \le S_1(n) \le 1/2n^{n-2} \quad , \qquad (1)$$

где n – число элементов в исходной последовательности точек. Полученные формулы могут быть интерпретированы следующим образом:

- проблема turnpike скорее всего не принадлежит к строгому NP-классу; вероятность получения решения в полиномиальное время в произвольном случае высока; число случаев, для которых решение не может быть получено в полиномиальное время, экспоненциально мало;
- проблема beltway скорее всего принадлежит к строгому NP-классу; вероятность получения решения в полиномиальное время в произвольном случае экспоненциально мала; число случаев, для которых решение не может быть получено в полиномиальное время, велико.

По сути, работа Lemke с соавт. (2003) обосновала известные из практики факты: для проблемы turnpike возможно построение алгоритма с вычислительной сложностью $O(N^2)$ (Dakic, 2000) для большинства данных, для проблемы beltway такого алгоритма построить не удалось и вычислительная сложность предложенных в литературе алгоритмов для большинства данных равна $O(N^N \log N)$ (Lemke a *et al.*, 2003). Именно по этой причине для линейных пептидов (задача *de novo* секвенирования сводится к проблеме turnpike) сделан большой прогресс в области разработки программ, а для циклических пептидов (сводится к проблеме beltway) – такие программы имеются в единичных экземплярах, и их результаты не всегда приводят к убедительным выводам. Например, неудачей завершилась попытка восстановления циклических пептидов микроорганизма *Oscillatoria* sp. из МС-спектров (Ng *et al.*, 2009).

В данной работе рассмотрено узкое, но наименее разработанное в литературе подмножество задач, связанных с восстановлением циклических пептидных цепочек из масс-спектров (задачи, сводимые к проблеме beltway). Вычислительные ресурсы для решения подобных задач с ростом длины последовательности

возрастают экспоненциально. Например, если допустить, что последовательность длиной равной одному элементу восстанавливается за один машинный такт (10^{-9} с), то последовательность длиной в 10 элементов будет восстановлена за 10^{10} тактов, что равно 10 с; последовательность длиной в 15 элементов будет восстановлена за 15^{15} тактов, или за 13,8 лет; последовательность длиной в 20 элементов будет восстановлена за 20^{20} тактов, или за 3,3 млрд лет. Поскольку математически доказано, что для проблемы beltway избежать экспоненциального роста вычислительных затрат невозможно, то целью исследований в этой области биоинформатики может быть только разработка подходов, которые ограничивают скорость этого роста. Как будет показано далее, возможно существенно ограничить скорость роста и получить эффективный алгоритм, который позволяет при доступных вычислительных ресурсах восстанавливать циклические последовательности длиной до 160 элементов.

Предлагаемый путь решения проблемы beltway

Получение эффективного алгоритма для решения проблемы может быть основано на ряде ограничений:

- Ограничимся восстановлением последовательностей, состоящих из 18 заранее известных элементов из множества $\Omega = \{57^{Gly}, 71^{Ala}, 87^{Ser}, 97^{Pro}, 99^{Val}, 101^{Thr}, 103^{Cys}, 113^{Ile, Leu}, 114^{Asn}, 115^{Asp}, 128^{Gln, Lys}, 129^{Glu}, 131^{Met}, 137^{His}, 147^{Phe}, 156^{Arg}, 163^{Tyr}, 186^{Trp}\}$, которые представляют собой веса стандартных аминокислотных остатков (без учета H_2O) (Lide, 1991). Заметим, что веса аминокислотных остатков {Ile, Leu} и {Gln, Lys} совпадают, и по этой причине число элементов во множестве Ω меньше на 2, чем общее число стандартных аминокислотных остатков.
- Предположим, что используемый для восстановления последовательности масс-спектр идеален, т. е. не содержит пропусков, дубликатов и лишних элементов, характерных для реального экспериментального спектра.

Идеализация задачи является основным приближением. Реальные спектры могут содержать пропуски (недостаток чувствительности аппаратуры), лишние элементы (загрязнения образца) и дубликаты (различная вероятность разрыва пептида в том или ином месте). Все эти эффекты в настоящей работе не учтены. Основная цель работы состоит в обеспечении ограничения экспоненциального роста вычислительных затрат и достижении максимальной длины последовательности, которая может быть восстановлена при современных вычислительных ресурсах.

Для решения задачи использован следующий подход. Назовем идеальным масс-спектром S полное множество масс всех подпоследовательностей различной длины от 1 до N, образованных одно- и двукратными разрывами некоторой циклической последовательности $\{m_1, m_2, .., m_N\}$, состоящей из N элементов $m_i \in \Omega$. К примеру, такой масс-спектр включает массы подпоследовательностей длины 3: $\{m_2, m_3, m_4\}$ (образована двумя разрывами исходной циклической цепочки в позициях $m_1 \downarrow m_2$ и $m_4 \downarrow m_5$) и $\{m_N, m_1, m_2\}$ (образована разрывами в позициях $m_{N-1} \downarrow m_N$ и $m_2 \downarrow m_3$). Также спектр Ω включает N одинаковых масс $M = \Sigma_i m_i$ для подпоследовательностей $\{m_1, m_2, .., m_N\}$, $\{m_2, m_3, .., m_N, m_1\}$, .., $\{m_N, m_1, m_2, .., m_{N-1}\}$, образованных однократными разрывами циклической цепочки в позициях $m_N \downarrow m_1$, $m_1 \downarrow m_2$, .., $m_{N-1} \downarrow m_N$ соответственно.

Назовем любую n-параметрическую циклическую последовательность $\{m_1, m_2, .., m_n, M - \Sigma_i^n m_i\}$, где $n < N$ и $m_i \in \Omega$, частичным решением, если ее масс-спектр S_n является подмножеством полного масс-спектра S, $S_n \in S$. Совокупность всех частичных решений длиной $n = N$ очевидно образует полное решение исходной задачи. Следует заметить, что, в силу требования цикличности, одно и то же частичное решение любой длины $2 < n \leq N$ может быть записано в $2n$ вариантах, различающихся циклическим сдвигом элементов и/или их инверсией. Например, последовательности: $\{m_1, m_2, m_3\}$, $\{m_2, m_3, m_1\}$, $\{m_3, m_1, m_2\}$, $\{m_3, m_2, m_1\}$, $\{m_2, m_1, m_3\}$ и $\{m_1, m_3, m_2\}$ представляют собой одну и ту же циклическую последовательность.

Единственное частичное решение с нулевым числом параметров (ранга 0) тривиально. Им является последовательность {M}, которая имеет один элемент с массой, равной массе всей искомой последовательности. Частичные решения

ранга 1 строятся на основе частичного решения ранга 0 делением элемента M на две части в виде $\{m_1, M - m_1\}$, где $m_1 \in \Omega$, и сохранением в полученном множестве тех последовательностей, чей спектр S_1 является подмножеством полного спектра задачи, $S_1 \in S$. Полное число N_1 частичных решений, получаемых таким способом, ограничено числом элементов $|\Omega|$ во множестве Ω, $N_1 \le |\Omega|$. Процесс дробления последнего элемента последовательности продолжается далее, образуя частичные решения ранга 2: $\{m_1, m_2, M - m_1 - m_2\}$, ранга 3: $\{m_1, m_2, m_3, M - m_1 - m_2 - m_3\}$ и т. д. Так продолжается до тех пор, пока не будут построены частичные решения ранга N, полное множество которых и является решением задачи. Следует заметить, что полное число возможных частичных решений N_n ранга n с увеличением значения n растет экспоненциально, $N_n \le |\Omega|^n$. Некоторые существенные детали алгоритма, позволяющие снизить скорость экспоненциального роста, приведены в приложении.

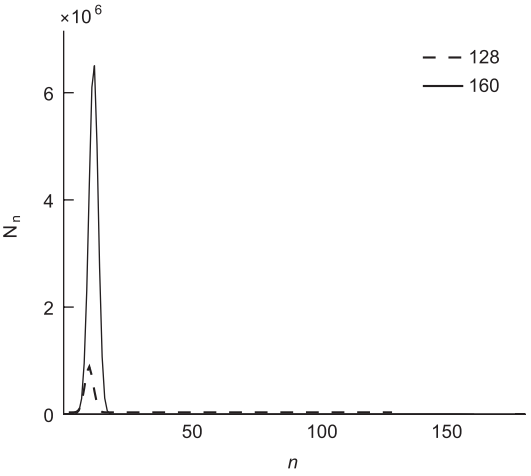

Рис. 1. Зависимость числа частичных решений разного ранга n для случайных последовательностей длиной в 128 (штриховая линия) и 160 (сплошная линия) элементов.

РЕЗУЛЬТАТЫ

На рис. 1 показано, как велико и как сильно меняется число частичных решений разного ранга на примере восстановления двух случайных последовательностей в 128 и 160 элементов. Как можно видеть, поначалу число частичных решений при увеличении длины n резко возрастает, а затем так же резко падает, образуя резкий пик при малых значениях $n \sim 10$. Величина этого пика составляет $\sim 0{,}9 \times 10^6$ для последовательности в 128 элементов и $\sim 6{,}6 \times 10^6$ для последовательности в 160 элементов. Подобный пик вполне ожидаем, так как он образован двумя противодействующими факторами: (1) экспоненциальным ростом числа решений $N_n \le |\Omega|^n$ и (2) резким падением вероятности того, что произвольная сгенерированная последовательность длины n имеет спектр, который является подмножеством заданного спектра.

На рис. 2 показано, как зависит время получения решения от числа элементов в искомой последовательности. Для получения этих данных было сгенерировано более 300 случайных последовательностей с элементами $m_i \in \Omega$ в диапазоне $n \in [30, 160]$. Ось абсцисс дана в логарифмическом масштабе. Как видно из

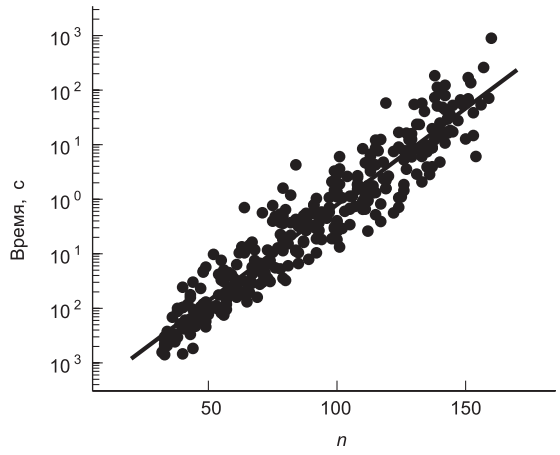

Рис. 2. Время восстановления циклической последовательности в зависимости от ее длины n.

графика, точки хорошо ложатся на прямую линию. Регрессионная прямая, дающая лучшее согласие с данными, может быть записана в виде уравнения, комбинирующего параметры $\log_{10} T$ и n: $\log_{10} T = 0{,}353n - 3{,}63$. Уравнение позволяет сделать оценку времени восстановления последовательности в зависимости от ее длины. Например, оценка для последовательности длиной в 250 элементов дает величину порядка 3 дней, а для последовательностей в 500 элементов − порядка 9 млн лет.

ВЫВОДЫ

Результаты демонстрируют, что разработанный нами алгоритм, детали которого описаны в приложении, достаточно эффективен. Он позволяет решить проблему beltway для последовательностей длиной до 160 элементов в пределах нескольких минут на персональном компьютере, что является очень хорошим результатом по ограничению экспоненциального роста для подобного рода задач. Для сравнения можно упомянуть работы (Ng *et al.*, 2009; Mohimani *et al.*, 2011), в которых решались задачи секвенирования циклических последовательностей существенно меньшей длины (~10), хотя и для реальных масс-спектров. Таким образом, полученные в данной работе результаты с существенным запасом превышают требования, возникающие в сегодняшних задачах *de novo* секвенирования. Это, в свою очередь, позволяет двигаться в направлении решения задач, интересных практически, т. е. к задачам восстановления последовательностей из реальных масс-спектров, содержащих пропуски, дубликаты и лишние данные.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке междисциплинарными интеграционными проектами СО РАН № 130, 39, 47, а также проектом фундаментальных исследований СО РАН VI.61.1.2.

ЛИТЕРАТУРА

Остерман Л.А. Методы исследования белков и нуклеиновых кислот: электрофорез и ультрацентрифугирование. М.: Наука, 1981. 286 с.

Aebersold R., Mann M. Mass spectrometry-based proteomics // Nature. 2003. V. 422. P. 198−207.

Allison L., Yee C.N. Restriction site mapping is in separation theory // Comput. Appl. Biol. Sci. 1988. V. 4. P. 97–101.

Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra // Expert Review of Proteomics. 2011. V. 8. No. 5. P. 645–657.

Benjamini Y., Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing // J. R. Stat. Soc. Ser. B. Methodol. 1995. V. 57. P. 289–300.

Chen T., Kao M., Tepel M., Rush J., Church G.M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry // Proc. of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). San Francisco. CA. 2000. P. 389–398.

Clauser K.R., Baker P., Burlingame A.L. Role of accurate mass measurement (+/-10 ppm) in protein identification strategies employing MS or MS/MS and database searching // Anal. Chem. 1999. V. 71. P. 2871–2882.

Colinge J., Masselot A., Giron M. *et al.* OLAV: Towards high-throughput tandem mass spectrometry data identification // Proteomics. 2003. V. 3. P. 1454–1463.

Craig R., Beavis R.C. TANDEM: matching proteins with tandem mass spectra // Bioinformatics. 2004. V. 20. P. 1466–1467.

Craig R., Cortens J.P., Beavis R.C. The use of proteotypic peptide libraries for protein identification // Rapid Commun. Mass Spectrom. 2005. V. 19. P. 1844–1850.

Craig R., Cortens J.C., Fenyo D., Beavis R.C. Using annotated peptide mass spectrum libraries for protein identification // J. Proteome Res. 2006. V. 5. P. 1843–1849.

Dainty J.C., Fienup J.R. Phase Retrieval and Image Reconstruction for Astronomy. Image Recovery: Theory and Application, 1987. P. 231–275.

Dakic T. On the Turnpike Problem. PhD Thesis. Simon Fraser University, 2000.

Dančík V., Addona T.A., Clauser K.R., Vath J.E., Pevzner P.A. De Novo Peptide Sequencing via Tandem Mass Spectrometry // J. Computational Biology. 1999. V. 6. No. 3-4. P. 327–342.

Elias J.E., Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry // Nat. Methods. 2007. V. 4. P. 207–214.

Eng J.K., McCormack A.L., Yates J.R. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database // J. Am. Soc. Mass Spectrom. 1994. V. 5. P. 976–989.

Frank A., Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling // Anal. Chem. 2005. V. 77. P. 964–973.

Frewen B.E., Merrihew G.E., Wu C. *et al.* Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries // Anal. Chem. 2006. V. 78. P. 5678–5684.

Hubbard S.J., Jones A.R. Proteome Bioinformatics. Humana Press, 2010.

Jaganathan K., Hassibi B. Reconstruction of Integers from Pairwise Distances // Acoustics, Speech and Signal Processing (ICASSP). IEEE International Conference. 2013. P. 5974–5978.

Jaganathan K., Oymak S., Hassibi B. Sparse phase retrieval: Uniqueness guarantees and recovery algorithms. arXiv:1311.2745 [cs, math], Nov. 2013. [Online]. Available: http://arxiv.org/abs/1311.2745.

Johnson R.S., Taylor J.A. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry // Mol. Biotechnol. 2002. V. 22. P. 301–315.

Keller A., Nesvizhskii A.I., Kolker E., Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search // Anal. Chem. 2002. V. 74. P. 5383–5392.

Lambert B., Jacques V., Shivanyuk A. *et al.* Calix[4]arenes as selective extracting agents. An NMR dynamic and conformational investigation of the lanthanide (III) and thorium (IV) complexes // Inorg. Chem. 2000. V. 39. No. 10. P. 2033−2041.

Lemke P., Skiena S.S., Smith W.D. Reconstructing Sets From Interpoint Distances // Discrete Computational Geometry Algorithms Combinatorics. 2003. V. 25. P. 597–631.
Lide D.R. Handbook of Chemistry and Physics. 72nd Ed. CRC Press. Boca Raton. FL., 1991.
Ma B., Zhang K., Hendrie C. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry // Rapid Commun. Mass Spectrom. 2003. V. 17. P. 2337–2342.
Marahiel M.A., Nakano M.M., Zuber P. Regulation of peptide antibiotic production in Bacillus // Mol Microbiol. 1993. V. 7. No. 5. P. 631–636.
Millane R.P. Phase retrieval in crystallography and optics // J. Opt. Soc. Am. A. 1990. V. 7. No. 3. P. 394–411.
Mohimani H., Liu W.T., Yang Y.L. *et al.* Multiplex De Novo Sequen-cing of Peptide Antibiotics // J. Comp. Biol. 2011. V. 18. No. 11. P. 1371–1381.
Nesvizhskii A.I., Vitek O., Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry // Nature methods. 2007. V. 4. No. 10. P. 787–797.
Ng J., Bandeira N., Liu W. *et al.* Dereplication and de novo sequencing of nonribosomal peptides // Nat. Methods. 2009. V. 6. P. 596–599.
Pandurangan G., Ramesh H. The restriction mapping problem revisited // J. Computer System Sciences. 2002. V. 65. P. 526–544.
Pappin D.J.C., Hojrup P., Bleasby A.J. Rapid identification of proteins by peptide-mass fingerprinting Transportable // Current Biology. 1993. V. 3. P. 327−332.
Perkins D.N., Pappin D.J.C., Creasy D.M., Cottrell J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data // Electrophoresis. 1999. V. 20. P. 3551–3567.
Rabiner L., Juang B.H. Fundamentals of Speech Recognition. Signal Processing Series. Prentice Hall. 1993.
Rahn J. Possible and impossible melodies: Some formal aspects of contour // Journal Music Theory. 1994. V. 36. No. 2. P. 259–279.
Shamos M.I. Problems in computational geometry. CMU. Pittsburgh. PA, 1977.
Sieber S., Marahiel M. Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics // Chem. Rev. 2005. V. 105. P. 715–738.
Skiena S.S., Smith W.D., Lemke P. Reconstructing sets from interpoint distances // Proc. Sixth ACM Symposium Computational Geometry. Berkeley. CA, 1990. P. 332–339.
Stefik M. Inferring DNA structures from segmentation data // Artif. Intell. 1978. V. 11. P. 85–114.
Storey J.D., Tibshirani R. Statistical significance for genomewide studies // Proc. Natl. Acad. Sci. USA. 2003. V. 100. P. 9440–9445.
Walther A. The question of phase retrieval in optics // Opt. Acta. 1963. V. 10. P. 41–49.
Wuthrich K. NMR of Proteins and Nucleic Acids. John Wiley and Sons. N. Y., 1986.
Zhang N., Aebersold R., Schwilkowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data // Proteomics. 2002. V. 2. P. 1406–1412.

ПРИЛОЖЕНИЕ

Детали алгоритмов

Графы частичных решений ранга n и сеть частичных решений

Построим ненаправленный граф всех частичных решений ранга 1 G_1 следующим образом. Вершинами графа являются все уникальные частичные решения $\{w_1, w_2, .., w_n\}$ ранга 1, где n – полное число решений ранга 1. Пометим вершины кортежами с тремя элементами вида $\langle w_i, \varnothing, \varnothing\rangle$, где w_i – вес соответствующей вершины. Две произвольные вершины графа $\langle w_i, \varnothing, \varnothing\rangle$ и $\langle w_j, \varnothing, \varnothing\rangle$ соединим ребром, если последовательность $\{w_i, w_j\}$ является частичным решением ранга 2. Пометим такое ребро кортежом с тремя элементами $\langle w_i + w_j, i, j\rangle$, где $w_i + w_j$ – вес данного ребра. Построенный таким образом граф может включать петли, например ребро, соединяющее вершину w_k с собой и соответствующее частичному решению $\{w_k, w_k\}$ ранга 2, но не может включать кратные ребра.

На следующем шаге мы строим ненаправленный граф G_2 всех частичных решений ранга 2. Вершинами данного графа объявляем все ребра графа G_1. Соединяем ребром любые две вершины графа G_2, если выполняются следующие условия: (1) соответствующие ребра графа G_1 являются соседями, т. е. они имеют общую вершину, (2) объединение двух частичных решений ранга 2, связанных с данными вершинами, представляет собой решение ранга 3. Любое новое ребро графа G_2 помечается кортежом с тремя элементами $\langle w_i + w_j - w_{i\cap j}, i, j\rangle$, где w_i и w_j – веса соответствующих ребер графа G_1, $w_{i\cap j}$ – вес их общей вершины $i \cap j$. Граф G_2 также может включать петли, но не может включать кратные ребра. Более того, петли в графе G_2 могут быть кратными. В самом деле, ребро $\langle w, k, m\rangle \in G_1$ может быть связано с самим собой через вершину $k \in G_1$, порождая ребро $\langle w + w - w_k, k, k\rangle$ в G_2, и через вершину $m \in G_1$, порождая ребро $\langle w + w - w_m, m, m\rangle$.

Подобным образом мы определим ненаправленный граф G_n всех частичных решений ранга n. Все вершины графа G_n образуются из ребер графа G_{n-1}. Любые две вершины $i, j \in G_n$ связываются ребром, если выполняются условия: (1) соответствующие ребра графа G_{n-1} являются

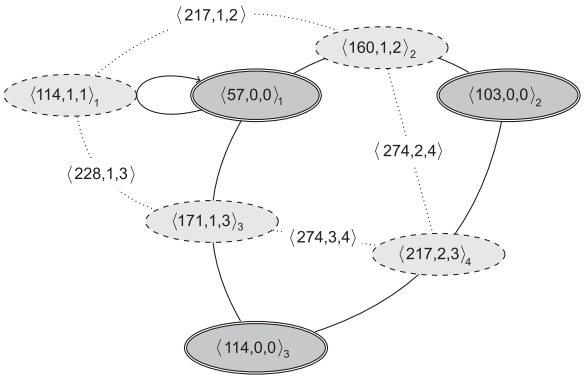

Рис. 3. Пример двух ненаправленных графов G_1 (сплошная и штриховая линии) и G_2 (штриховая и пунктирная линии) для последовательности {57, 114, 103, 57}.

соседями, т. е. имеют общую вершину, (2) объединение двух частичных решений ранга $n - 1$, связанных с данными вершинами, представляет собой решение ранга n. Вес такого ребра рассчитан по формуле $w_{i \cup j} = w_i + w_j - w_{i \cap j}$.

Заметим, что поскольку граф G_n любого ранга n строится на графе G_{n-1} предыдущего ранга $n - 1$, то совокупность всех графов образует сеть графов. На рис. 3 представлен пример объединения в сеть двух графов G_1 и G_2 для последовательности {57, 114, 103, 57}. Вершины графа G_1 выделены сплошными линиями, ребра, которые одновременно являются вершинами графа G_2, – штриховыми линиями, ребра графа G_2 – пунктирными линиями.

Оценка полного числа операций при поиске решения

Алгоритм нахождения решений основывается на последовательном построении сети из графов G_1, G_2, ... G_N. Совокупность вершин графа G_N дает полное решение поставленной задачи. Сделаем оценку полного числа операций при получении полного решения. Данное число равно $N(G_1 \cup G_2 \cup .. \cup G_N) = \Sigma N(G_k)$, где $N(G_k)$ – число операций для построения графа G_k.

Для естественного алгоритма, в котором ребра графа строятся между всеми парами вершин с дальнейшей их проверкой на то, являются ли они частичными решениями или нет, число операций $N(G_k)$ для построения ребер графа вычисляется как $N(G_k) = N_k^2\, k^2$, где N_k – число вершин в графе G_k. В этой формуле первый сомножитель N_k^2 обусловлен необходимостью проверки всех пар вершин, а второй сомножитель отражает число операций на построение спектра потенциального частичного решения длины k (аналогичное построение всех циклических пар расстояний между элементами последовательности длиной k). Общее число полученных ребер графа G_k будет равно $N(E) \le N_k^2$. Эти полученные ребра являются вершинами следующего графа G_{k+1}, что позволяет построить рекуррентную зависимость для вычисления полного числа операций построения сети. Общее число операций для «естественного» алгоритма равно $N(G_1 \cup G_2 \cup .. \cup G_N) \le \le \Sigma\, N_1^{2k}\, k^2$, где N_1 – число вершин в начальном графе G_1. Значение N_1 известно, $N_1 \le 18$, где 18 есть число элементов в последовательности весов стандартных аминокислотных остатков.

В алгоритме, который представлен в этой работе, число шагов существенно меньше, чем в «естественном» алгоритме. Это обусловлено тем, что ребра графа G_k строятся только между теми вершинами, которые являются соседями в графе меньшего ранга G_{k-1}, т. е. в графе G_{k-1} они связаны общей вершиной. Вычислительные эксперименты на большом числе случайных последовательностей показывают, что среднее число подобных соседей $<m>_k$ у любой вершины графа ранга k много меньше, чем число вершин, т. е. $<m>_k << N_k$ и обычно находится в диапазоне от 3 до 10. Это приводит к тому, что рост числа вершин при переходе от графа ранга k к графу ранга $k + 1$ перестает быть квадратичным. Другим способом ограничения числа операций является то, что спектр любого частного решения в алгоритме строится на базе уже рассчитанного спектра соседней вершины, т. е. требует не k^2 операций, а всего лишь k операций. Построив аналогичную рекуррентную зависимость для расчета полного числа операций, получим, что алгоритм требует $N(G_1 \cup G_2 \cup .. \cup G_N) \le N_1^{\,2}\, \Sigma <m>_k^{k-1}$.

Таким образом, построение сети графов частичных решений, когда каждый элемент сети строится на базе уже построенных соседних элементов, позволяет существенно снизить вычислительную сложность алгоритма от $\Sigma\, N_1^{2k}\, k^2$ до $N_1^{\,2}\, \Sigma <m>_k^{k-1}$.

RECONSTRUCTION OF AMINO ACID SEQUENCES OF CYCLIC PEPTIDES FROM THEIR MASS SPECTRA

E.S. Fomin

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: fomin@bionet.nsc.ru

Summary

Mass spectrometry is a physical method, which can be applied to the investigation of proteomes of different organisms. It allows us both to solve the problem of identification of biological macromolecules and to sequence peptide chains in cases where information on the genomes is scarce or absent. Currently, there are many software programs to support research in this area. Nevertheless, in spite of all efforts, there is little progress in the development of programs able to solve the problem for de novo sequencing of cyclic peptides, which are most effective antibiotics, antitumor agents, immunosuppressants, toxins, and a vast number of nonribosomal peptides with unknown functions. In this paper, an effective algorithm for solving the problem of de novo sequencing cyclic peptides is proposed. The algorithm allows us to reconstruct sequences of lengths up to 160 amino acid residues.

Key words: mass spectrometry, sequencing of cyclic peptides, beltway problem.