ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2008, том 18, № 4, с. 114–119

МАТЕМАТИЧЕСКИЕ МОДЕЛИ

УДК 621.384.668.8; 577.112.6

© С. В. Фиронов, Я. И. Лютвинский, Н. В. Краснов

ИСПОЛЬЗОВАНИЕ НЕДЕТЕРМИНИРОВАННЫХ КОНЕЧНЫХ АВТОМАТОВ И ИНДЕКСА МАСС ПЕПТИДОВ ДЛЯ БЫСТРОГО СОПОСТАВЛЕНИЯ ФРАГМЕНТНЫХ МАСС-СПЕКТРОВ ПЕПТИДОВ БАЗАМ ДАННЫХ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Разработаны алгоритмы, позволяющие находить пептиды в базе данных аминокислотных последовательностей белков по частично интерпретированным фрагментным масс-спектрам пептидов. Показано, что использование индекса масс пептидов позволяет производить поиск одиночного образца за интерактивное время, а использование идеологии недетерминированных конечных автоматов позволяет производить поиск большого числа образцов за время чтения базы данных. Впервые разработан алгоритм, обладающий устойчивостью к ошибкам типа замены одной аминокислоты другой, либо одной из концевых масс другой.

ВВЕДЕНИЕ

Интерпретация данных тандемной масс-спектрометрии в белковых анализах — одно из наиболее востребованных математических приложений в стремительно развивающейся новой отрасли биологических знаний — протеомике. Данная работа развивает подход к интерпретации фрагментных масс-спектров пептидов, подразумевающий восстановление аминокислотных последовательностей, представленных в масс-спектре сериями фрагментов (рис. 1), и последующее сопоставление восстановленных последовательностей с базами данных аминокислотных последовательностей белков. Впервые этот подход предложен в статье [1] и получил в мировой научной литературе название Peptide Sequence Tag Search.

В процессе анализа масс-спектра с целью установить аминокислотную последовательность пептида, пики масс-спектра сопоставляются ионам фрагментов какой-либо серии. Если бы все ионы фрагментов были бы представлены пиками масс-спектра, то образовалась бы последовательность пиков такая, что разность масс соседних пиков последовательности была бы равна массе какого-либо аминокислотного остатка в нативной или модифицированной форме. Однако часто возникает ситуация, представленная на рис. 2, когда удается обнаружить только часть фрагментов, что позволяет восстановить лишь частичную последовательность пептида [2].

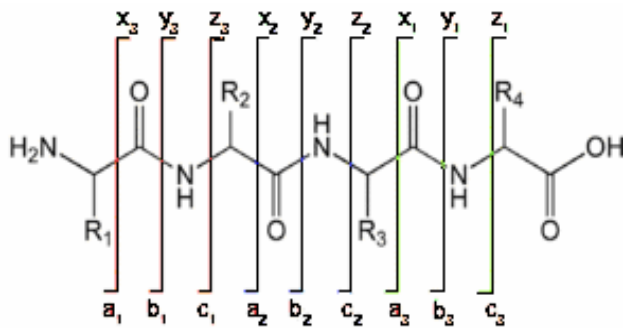

Рис. 1. Серии фрагментов a,b,c,x,y,z

Такую последовательность пиков масс-спектра можно записать в виде строки, содержащей:

1) массу первого пика в последовательности;

2) последовательность аминокислот, соответствующих расстоянию между пиками последовательности;

3) разницу между массой родительского иона и массой последнего пика последовательности.

Например, для последовательности, представленной на рис. 1, образец записывается как [372.240]AAGII[234.079]. Записанная таким образом строка будет использована как образец для поиска полной аминокислотной последовательности пептида в базе данных белков.

Для того чтобы принять во внимание возможные модификации исходного пептида, следует сформулировать задачу поиска соответствия следующим образом: в базе данных аминокислотных последовательностей белков найти пептиды, соответствующие образцу с ошибкой таких типов, как замена одной аминокислоты другой, либо замена одной из концевых масс на другую. В случае если обе концевые массы больше нуля, то поиск должен быть произведен как по прямой, так и по обратной последовательностям.

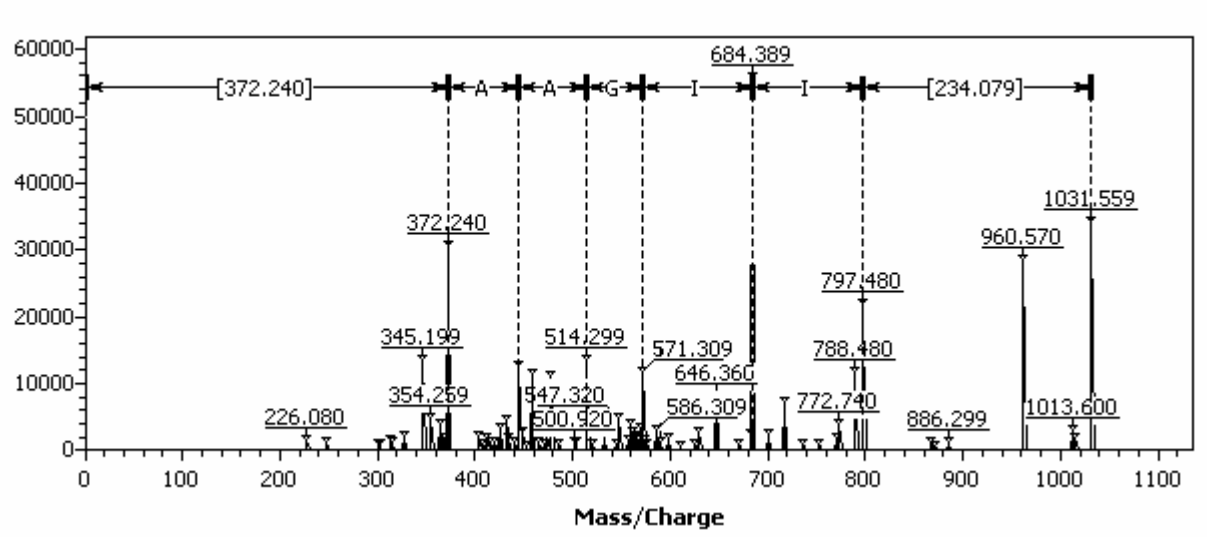

Рис. 2. Частично интерпретированный масс-спектр. Разницы масс между пиками, сопоставленными ионам неизвестной серии, соответствуют массам аминокислот

Образец для поиска сопоставляется со всеми пептидами, содержащимися в базе данных аминокислотных последовательностей белков. В зависимости от предположения о том, к какой серии ионов принадлежат обнаруженные пики масс-спектра, критерии соответствия пептида образцу будут разными. В частности, для a- , b- и c-серий следует сравнивать образец с прямой аминокислотной последовательностью пептида, тогда как для x- , y- и z-серий — с обратной. Отличие массы первого фрагмента от суммарной массы составляющих его аминокислот также зависит от серии.

Для примера дадим определение соответствия пептида образцу для b-серии, для остальных же серий с учетом предыдущих замечаний определения строятся аналогично.

Пептид соответствует образцу, если существует такое разбиение пептида на три части, удовлетворяющее следующим условиям:

1) суммарная масса аминокислот первой части разбиения равна первой массе образца за вычетом массы протона H^+;

2) вторая часть эквивалентна частичной последовательности, содержащейся в образце, с точностью до замены друг на друга масс-спектрометрически неразличимых аминокислот, таких как например лейцин и изолейцин;

3) суммарная масса аминокислот, входящих в третью часть разбиения, равна последней массе образца за вычетом массы H_2O и протона H^+.

Например, образцу [372.240]AAGII[234.079] из вышеприведенного примера соответствует пептид DQQAAGIIES с разбиением (DQQ)—(AAGII)—(ES). Учитывая возможность замены одной аминокислоты другой, пептид DQQAAKIIES также соответствует образцу [372.240]AAGII[234.079] с ошибкой: G в образце заменено на K в пептиде.

Базу данных аминокислотных последовательностей белков, записанных в однобуквенной нотации можно рассматривать как текст, к которому применимы алгоритмы текстового поиска. Задача поиска в текстовой базе данных с ошибками — классическая задача биоинформатики. Один из алгоритмов, решающих эту задачу, BLAST [3] позволяет находить в базе данных аминокислотные последовательности, наиболее близкие к заданной. Классические алгоритмы текстового поиска удалось бы эффективно применить только в том случае, если бы все массы, входящие в образец для поиска, могли бы быть интерпретированы с использованием небольшого числа вариантов аминокислотной последовательности, но в подавляющем большинстве случаев это не так.

Современные подходы к задаче поиска частично-восстановленных аминокислотных последовательностей пептида в белковой базе данных представлены в работах [4–7]. Однако в этих работах присутствует ряд ограничений. Так, работа Тэйлора и Джонсона [4] описывает поиск только по текстовой информации. Алгоритм, реализованный в SPIDER [5], учитывает только ошибку типа замены одной последовательности аминокислот другой, имеющей ту же массу. В работе [6] частичные аминокислотные последовательности используются только для фильтрации базы данных. Алгоритм Paragon [7] предназначен для работы с данными, полученными на приборе класса QqTOF.

Задача поиска образца в базе данных аминокислотных последовательностей белков имеет две модификации в зависимости от применения: это задача поиска одного образца, используемая при обработке масс-спектра специалистом в "ручном" режиме, и задача поиска большого количества образцов, используемая в автоматической обработке набора масс-спектров для идентификации белков в смеси.

Испытания алгоритмов проводились на компьютере Intel Pentium Core™2 Quad 2.67 ГГц с 2 ГБ оперативной памяти. В качестве базы данных аминокислотных последовательностей были взяты базы NRdb и SwissProt [8] по состоянию на май 2008 года. Скорость чтения базы 21 МБ / с, время произвольного доступа 0.031 с.

ЗАДАЧА ПОИСКА ОДНОГО ТЭГА В БАЗЕ ДАННЫХ

Скорость чтения базы данных не позволяет искать образец в базе данных за интерактивное время, просматривая всю базу данных целиком, поэтому для ускорения поиска применяется предварительная индексация базы данных.

Индексирование

В продуктах гидролиза, наиболее часто используемых для масс-спектрометрического анализа белков, массы пептидов, как правило, не превышают 3000 Да. Поэтому при индексировании ограничимся аминокислотными последовательностями, имеющими суммарную массу 3000 Да. Хотя это и не существенно для скорости работы алгоритма, но с увеличением максимальной индексируемой массы растет размер индекса. Составим индекс всевозможных пептидов массой меньше 3000 Да, имеющихся в базе данных, согласно следующему принципу. В файле базы данных позиция, с которой начинается пептид, имеет набор индексов, равных:

1) суммарной массе пептида;

2) суммарной массе части пептида, предшествующей выделенной аминокислоте, умноженной на 3000, и суммарной массе части пептида после выделенной аминокислоты, где в качестве выделенной аминокислоты перебираются все аминокислоты последовательности пептида исключая первую и последнюю.

Так, например, последовательность MAKVDIDIVDFEY попадает в индексы: 1538 (суммарная масса), 394336 (при выделенной аминокислоте A), 607208 (при выделенной аминокислоте K) и так далее.

Размер такого индекса для базы данных SwissProt составляет 40 ГБ и 240 ГБ для NRdb.

Для ускорения работы алгоритма поиска в случае, когда требуется вести поиск только по пептидам, принадлежащих белкам определенных организмов, база данных белков сортируется в соответствии с таксономическим индексом. В этом случае позиции в файле базы данных, попадающие в один индекс, должны быть отсортированы в возрастающем порядке, что позволяет быстро найти часть, в которой необходимо вести поиск, и прочитать с жесткого диска в оперативную память только эту часть, а не весь индекс, что существенно увеличивает скорость работы.

Поиск

Рассмотрим задачу поиска образца вида

[масса 1][последовательность аминокислот][масса 2]

в базе данных аминокислотных последовательностей белков. В том случае, если первая и последняя массы не равны нулю, неизвестно в прямой или обратной аминокислотной последовательности белка надо искать подходящий пептид.

Стоит заметить, что если взять любую аминокислоту образца, то все пептиды базы данных, которые содержат данный образец с ошибкой в этой аминокислоте, имеют индекс, равный сумме масс части образца до этой аминокислоты, умноженной на 3000, с массой части образца после выделенной аминокислоты. Таким образом, перебирая подряд все элементы образца, получаем список всех возможных позиций в файле базы данных аминокислотных последовательностей белков, с которых могут начинаться искомые пептиды.

Для каждой из позиций, найденных с помощью индекса, производится сравнение частичной аминокислотной последовательности белка, начинающейся с этой позиции, с образцом. В том случае, если образец совпадает с данной частичной последовательностью с точностью до одной аминокислоты или массы, то пептид, начинающийся с этой позиции в базе данных, включается в список кандидатов идентифицированных пептидов.

Достоинством такого подхода является высокая скорость работы за счет относительно небольших затрат на чтение с жесткого диска. Для базы данных SwissProt суммарный размер индекса, который необходимо прочитать с носителя данных при поиске образца [239]YAGFL[382] составляет 600 кБ, что в свою очередь меньше 1 % объема самой базы данных.

ЗАДАЧА МНОЖЕСТВЕННОГО ПОИСКА ТЭГОВ В БЕЛКОВОЙ БАЗЕ ДАННЫХ

При автоматической обработке набора масс-спектров, число гипотез вида

[масса 1][последовательность аминокислот][масса 2]

обычно составляет от 100 до 10^7. Разработанный алгоритм множественного поиска позволяет искать пептиды, удовлетворяющие этим образцам с ошибкой за один цикл чтения базы данных.

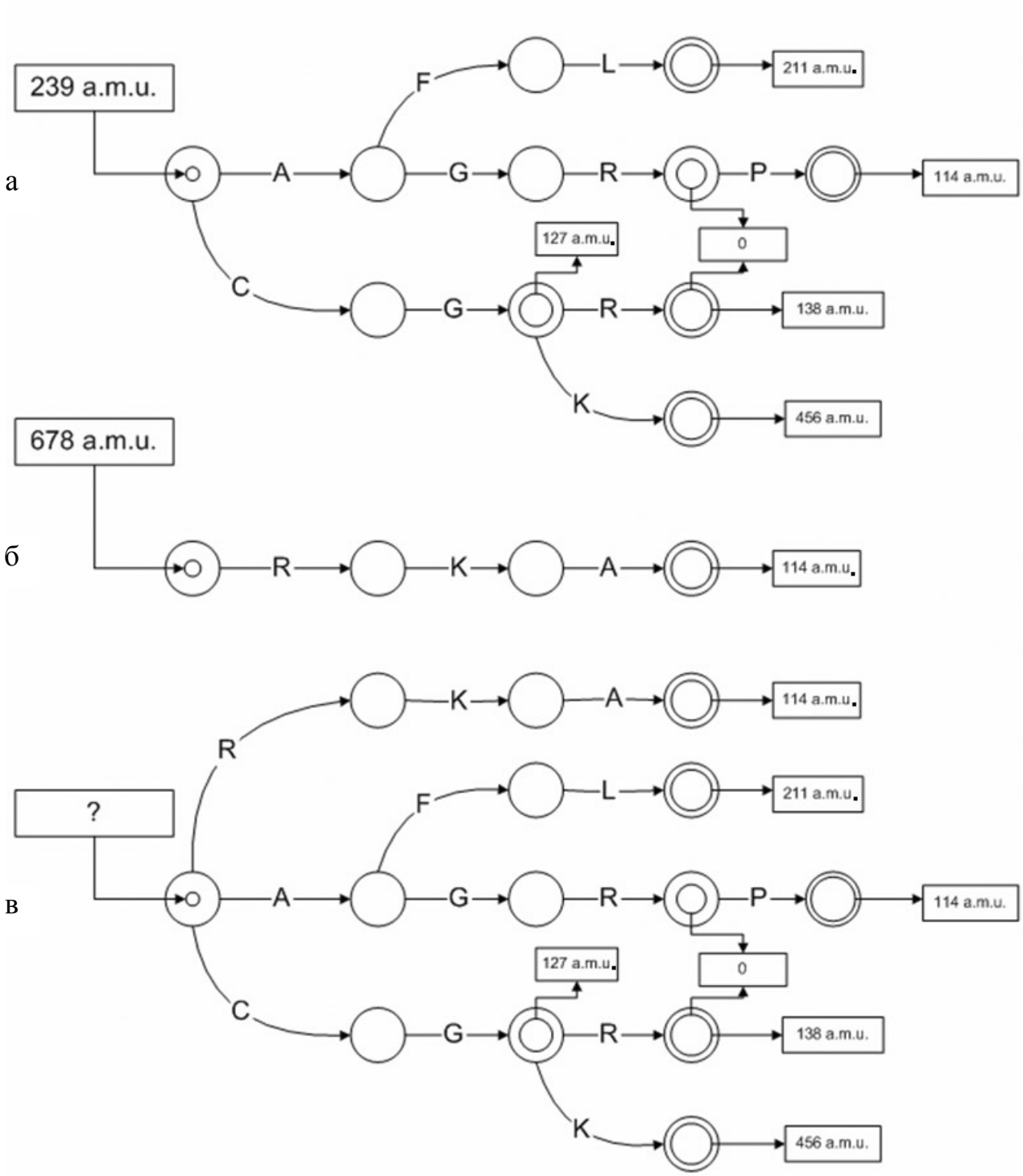

Рис. 3. Конечный автомат, построенный по набору образцов.
Графы переходов из начальных состояний, идентифицированных массой образца:
а — для массы 239 а.е.м.; б — 678 а.е.м.; в — для произвольной массы

Задача поиска формулируется следующим образом: допустим, имеется белковая база данных, содержащая N букв и M ($<< N$) пронумерованных образцов со средней длиной $<L>$ и максимальной — L_M. Необходимо найти все триптические пептиды из базы данных аминокислотных последовательностей белков, удовлетворяющих этим образцам.

Для решения задачи построим по образцам конечный недетерминированный автомат и пропустим через него текст базы данных. Алгоритм построения автомата аналогичен алгоритму построения поискового автомата, использующемуся в алгоритме Ахо—Корасик [9], с той лишь разницей, что автомат имеет несколько начальных состояний, идентифицированных начальными массами образцов.

Рассмотрим более подробно структуру автомата. Каждое состояние автомата содержит следующую информацию:

1) порядковый номер состояния;

2) признак конечного состояния (конечное/не конечное);

3) список образцов, заканчивающихся в данном состоянии.

Переходы автомата помечены буквами — обозначениями основных аминокислот. На рис. 3 показан автомат, построенный для тэгов [239]AFL[211], [239]AGR[0], [239]AGRP[114], [239]CG[127], [239]CGR[0], [239]CGR[138], [239]CGK[456], [678]RKA[114]. Одно начальное состояние (знак вопроса на рис. 3) предназначено для поиска с ошибкой в первой массе.

Построенный автомат позволяет искать точное совпадение подстрок белков с образцами, начинающимися на некоторую массу и далее состоящими только из букв, в то время как следует учитывать ошибку в любом элементе образца. Поэтому состояние всей системы должно описываться не только списком состояний автомата, но и накопленными в каждом из этих состояний ошибками. Кроме того, если в некотором состоянии ошибки нет, то при обработке текста базы данных вне зависимости от обрабатываемой буквы должны происходить все существующие переходы из данного состояния с накоплением ошибки, если метка перехода не совпадает с обрабатываемой буквой.

В случае неспецифического частичного гидролиза белков неизвестна позиция конца образца, поэтому необходим препроцессинг белка, предоставляющий для каждой позиции информацию о возможных массах подстрок, начинающихся вслед за текущей позицией. Тогда если автомат пришел в конечное состояние с накопленной ошибкой, то для каждого номера образца, соответствующего данному конечному состоянию, необходимо проверить, есть ли для данной позиции концевая масса образца.

ЗАКЛЮЧЕНИЕ

При использовании алгоритма поиск 100 000 образцов со средней длинной 6 в базе данных SwissProt был произведен за 5.61 с.

Разработанные алгоритмы позволяют находить пептиды в базе данных аминокислотных последовательностей белков по частично интерпретированным фрагментным масс-спектрам пептидов. Использование индекса масс пептидов позволяет производить поиск одиночного образца за интерактивное время, не превышающее 3 с (для базы данных SwissProt), а использование недетерминированных конечных автоматов позволяет производить поиск большого числа образцов за время чтения базы данных. Ключевая особенность данных алгоритмов — устойчивость к ошибкам типа замены одной аминокислоты другой, либо одной из концевых масс другой. Результатом работы представленных алгоритмов является список пептидов — кандидатов для последующей более подробной идентификации по фрагментным масс-спектрам.

СПИСОК ЛИТЕРАТУРЫ

1. *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.
2. *Лютвинский Я.И., Макаров В.В., Краснов Н.В., Подольская Е.П., Веренчиков А.Н.* Частичная расшифровка аминокислотной последовательности пептида по его фрагментному масс-спектру: алгоритм и результаты применения // Научное приборостроение. 2006. Т. 16, № 3. С. 122–131.
3. *Myers E.* Algorithmic Advances for Searching Biosequence Databases // Computational Methods in Genome Research. Plenum Press, 1994. P. 121–135.
4. *Taylor J.A., Johnson R.S.* Sequence Database Searches via de Novo Peptide Sequencing by Tandem Mass Spectrometry // Rapid Commun. Mass Spectrom. 1997. V. 11, N 9. P. 1067–1075.
5. *Han Y., Ma B., Zhang K.* SPIDER: Software for Protein Identification from Sequence Tags Containing De Novo Sequencing Error // Journal of Bioinformatics and Computational Bioliogy. 2005. V. 3, N 3. P. 697–716.
6. *Frank A., Tanner S., Bafna V., Pevzner P.* Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry // Proteome Res. 2005. V. 4, N 4. P. 1287–1295.
7. *Shilov I.V., Seymour S.L., Patel A.A., et al.* The Paragon Algorithm: A Next Generation Search

Engine that Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. Mass Spectrometry Informatics R&D. Applied Biosystems|MDS Sciex, Foster City. CA 94404.
8. *Wu C.H, Apweiler R., Bairoch A., et al.* The Universal Protein Resource (UniProt): an Expanding Universe of Protein Information // Nucleic Acid. Res. 2006. Jan. 1. D187–191.
9. *Aho A.* Algorithms for Finding Patterns in Strings // Handbook of Theoretical Computer Science. MIT Press., 1990. V. A. P. 257–300.

Институт аналитического приборостроения РАН, Санкт-Петербург

Материал поступил в редакцию 13.10.2008.

FAST PEPTIDE SEQUENCE TAG SEARCH IN PROTEIN DATABASES USING FINITE STATE AUTOMATON AND PEPTIDE MASSES INDEXING

S. V. Fironov, Ya. I. Lutvinsky, N. V. Krasnov

Institute for Analytical Instrumentation RAS, Saint-Petersburg

Algorithms to find peptides in protein databases for partially interpreted mass spectra of peptides have been developed. It has been shown that the use of mass index peptides allows users to search for online time a single sample, while the use of nondeterministic finite automata allows users to search a large number of samples during the reading of the database. For the first time algorithm tolerant to errors caused by replacing one amino acid by another or by one final mass by another was developed.