УДК 543.51

Д.М. Автономов[1,2], *И.А. Агрон*[1,3], *А.С. Кононихин*[4,2], *Е.Н. Николаев*[4,2]

[1] Московский физико-технический институт (государственный университет)
[2] Институт биохимической физики им. Н.М. Эмануэля РАН
[3] Институт биомедицинской химии им. В.Н. Ореховича РАМН
[4] Институт энергетических проблем химической физики РАН

Создание базы данных точных массово-временных меток для качественного и количественного подхода в исследовании протеома мочи человека с использованием изотопного мечения

Актуальной задачей современной биологии и медицины (их физических методов исследования) является разработка методик быстрого и точного анализа человеческих жидкостей, в том числе их белкового состава. Уже были исследованы такие жидкости, как кровь, моча, слеза, на предмет белкового состава. Основные сейчас методики довольно медленны и плохо реализуются в плане высокопроизводительного количественного анализа. Поэтому мы предлагаем метод для быстрого и высокопроизводительного анализа, основанный на идеологии точной массово-временной метки и изотопного мечения.

Ключевые слова: протеом, человеческие жидкости, белковый состав, пептиды, геномная база, масс-спектрометр ионно-циклотронного резонанса, массово-временная метка, изотопное мечение.

I. Современные методики идентификации белков и пептидов

Сейчас наиболее эффективным способом идентификации белков является поиск по геномной базе данных с использованием спектров фрагментации (тандемной масс-спектрометрии) (рис. 1).

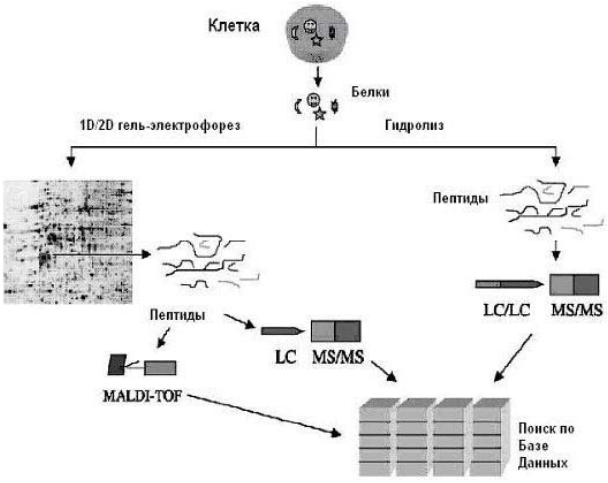

Рис. 1. Современные стандартные процедуры идентификации белков

Наиболее распространенными поисковыми движками являются Mascot (MatrixScience (Darryl Pappin)), Sequest (Jimmy Eng и John Yates), X! Tandem (проект — The Global Proteome Machine) и OMSSA (Национальный Институт Здоровья, США).

Все поисковые машины в общем рассчитаны под стандартную процедуру эксперимента: гидролиз белковой фракции с помощью сайт-специфичного фермента, обычно трипсина, разделение на хроматографической колонке с обращенной фазой (смывание пептидов по их гидрофобности в зависимости от концентрации ацетонитрила), масс-спектрометрический анализ с селективным измерением спектров фрагментации ионов пептидов [1, 2].

II. Поисковая машина Mascot

Идеология идентификации в поисковой машине Mascot основана на алгоритме MOWSE (MOlecular Weight SEarch), разработанном Дэррилом Паппиным (1993) [3].

Изначально алгоритм был рассчитан на так называемый поиск по массовым «отпечаткам пальцев» пептидов.

Первым этапом поиска с использованием алгоритма Mowse является сравнение вычисленных масс пептидов для всех

записей последовательностей в базе данных с набором экспериментальных данных. Каждая вычисленная величина, которая совпадает с заданной погрешностью с экспериментальной величиной, считается как совпавшая. Молекулярная масса интактного белка может быть использована как пре-фильтр.

Вместо простого счета числа совпавших пептидов Mowse-алгоритм использует эмпирически определённые факторы, чтобы приписывать статистический вес каждому совпавшему пептиду. Матрица весовых факторов рассчитывается в тот момент, когда происходит установка базы данных, следующим образом.

Матрица частотных факторов, F, создаётся, чтобы каждая строка соответствовала интервалу в 100 Да для масс пептидов, а каждая колонка — интервал в 10 кДа для масс интактных белков. Поскольку каждая запись последовательности в базе данных обрабатывается, то соответствующий матричный элемент $f_{i,j}$ добавляется так, чтобы собрать статистику распределения масс пептидов как функцию масс белков. Элементы F затем нормализуются посредством деления элементов каждого «10 кДа» — столбца на наибольшую величину в этом столбце, чтобы рассчитать Mowse-матрицу весов M в соответствии с формулой

$$m_{i,j} = \frac{f_{i,j}}{|f_{i,j}|_{\max incolumn(j)}}.$$

После сравнения экспериментальных масс с рассчитанными массами по базе данных скор (будем так называть некую величину уровня достоверности, Score на англ. счет) для каждого элемента рассчитывается в соответствии с формулой

$$Score = \frac{50\ 000}{M_{prot} * \prod_n m_{i,j}},$$

где M_{prot} — молекулярная масса для каждого совпавшего белка, а Π — произведение, которое рассчитывается из Mowse-матрицы весов M для каждого совпадения экспериментальных данных и масс пептидов, рассчитанных из записей в геномной базе данных.

В итоге скор есть абсолютная величина, которая соответствует тому, насколько наблюдаемый пептид — случайная величина.

Реализован поиск по «Peptide Mass Fingerprint» был раньше, чем поиск с использованием результатов тандемной масс-спектрометрии (MS/MS-поиск), но схема для «Peptide Mass Fingerprint» была позже также отлично применена и для MS/MS-поиска: только теперь роль белка в формуле для Score выполняет пептид, а фрагмент — роль пептида. Сумма же скоров пептидов суммируется, откуда получается суммарный скор для белка.

III. Поисковая машина Sequest

В случае же с программой Sequest поиск осуществляется иным способом [4].

Поисковая машина идентифицирует каждый тандемный масс-спектр отдельно.

Поисковой движок создаёт лист всех пептидов, которые могли бы получиться теоретически при гидролизе с помощью данного фермента. Далее происходит поиск с учётом заранее заданной точности измерения масс по этому листу совпадений с экспериментальными массами интактных пептидов (они известны из масс-спектра), в результате которого мы получаем новый лист пептидов-кандидатов. Для каждого кандидата строится теоретический спектр фрагментации, который потом опять же сверяется с экспериментальными данными (с MS/MS-спектрами). Результатом является значение взаимно корреляционной функции.

Среди плюсов методов, использующих тандемную масс-спектрометрию (MS/MS), можно указать:

1) высокую достоверность идентификации пептидов,

2) возможность deNovo [1] сиквенирования пептидов.

Среди недостатков таких подходов стоит указать:

1) из-за малого количества образца может быть недостаточно ионов для проведения MS/MS,

2) не всегда могут получаться хорошие спектры фрагментации,

3) дополнительное время, затрачиваемое на снятие спектров фрагментации,

4) увеличенное время хроматографии.

IV. Идеология точной массово-временной метки

В Pacific Nortwest Naional Laboratory (Richland, WA, US) группой Дика Смита была предложена методика точной массово-временной метки (AMT-tags) для идентификации белков и пептидов [5]. Преимущества данного метода заключаются в отказе от MS/MS, для проведения которого нужно больше образца, времени, а главное, что не всегда фрагментация в MS/MS хорошо получается и в использовании иного параметра, чем масса, не связанного с ней напрямую: времени удержания на хроматографической колонке с гидрофобно-обращенной фазой.

В обычном подходе при использовании тандемной масс-спектрометрии осуществляется поиск по геномной или протеомной базе данных, содержащей последовательности белков, с учётом гидролизующего фермента (то есть сайтов гидролиза) и способа фрагментации в масс-спектрометре; в случае же точных массово-временных меток предлагается использовать только точную массу пептида (продукта гидролиза) и времени его элюции с хроматографической колонки (точнее говоря, концентрации растворителя, обычно ацетонитрила). Во-первых, экономится образец и время, а во-вторых, этот метод с той же процедурой Bottom-Up становится более информативным, когда хроматограф перестает быть просто разделителем фракции пептидов.

В простейшем случае база данных состоит из таблицы, содержащей точную массу, аминокислотную последовательность пептида и время выхода (время начала и конца хроматографического пика). Хотя в реальной ситуации это несколько сложнее, поскольку этой информации мало, исследователю или пользователю БД нужно знать, к какому белку относится пептид и какая биологическая функция у белка, например. Конечно, все это могло быть в одной таблице, но управлять ею эффективно было бы невозможно при большом количестве пептидов в ней. Вследствие этого возникает проблема создания эффективной базы данных.

V. Количественный подход

В большинстве случаев сейчас масс-спектрометрия в протеомике является качественным методом анализа, особенно, когда речь идёт о такой высокопроизводительной системе, как сочетание жидкостной хроматографии и масс-спектрометрии с ионизацией электрораспылением. Принципиальным расширением такой системы анализа будет внедрение количественных методик [6]. Сейчас существует несколько вариаций, все они основаны на изотопном мечении: либо изотопами C^{12}/C^{13} углеродами, либо O^{16}/O^{18}, либо N^{14}/N^{15}. Основными являются ICAT (cICAT), iTRAQ, ICROS, гуанидирование и ферментативное C-концевое мечение.

1. ICAT (cICAT) — мечение происходит сайт-специфично либо протонированными/дейтерированными метками, либо метками на основе C^{12}/C^{13} углеродов.

2. ICROS — две пробы метятся по цистеину двумя реагентами: N-этилйодацетамид и N-D5-этилйодацетамид.

3. iTRAQ — метка (4 разновидности, состоит из рапортирующей и балансировочной группировок, которые в сумме дают одинаковую массу) связывается с пептидом, и при MS-анализе все пики совпадают, но при MS/MS происходит отделение пиков на масс-спектрах друг от друга, поскольку рапортирующие группировки все разные. То есть этот метод применим только вкупе с использованием тандемной масс-спектрометрии.

4. Гуандирование — это такой метод мечения, когда один из двух образцов подвергается модификации, модифицируется лизин в гомоаргинин.

5. Ферментативное C-концевое мечение O^{16}/O^{18} кислородами в составе воды — при трипсинолизе в силу двух реакций (энзиматической и кислородобменной в карбоксильной группе) присоединяется один или два кислорода, в одном случае изотоп O^{16}, во втором — O^{18}.

Каждый из методов имеет свои плюсы и недостатки, среди недостатков первых двух стоит особо отметить их специфичность по цистеину (или метионину) и проблемы ко-элюции изначально одинаковых пептидов (по-разному меченных), четвертый метод также специфичен, поскольку зависит от наличия лизина на C-кон-

це. Третий способ вообще может использоваться только в тандемной масс-спектрометрии. У пятого способа недостаток заключается в вариабельности включения одного или двух «тяжёлых» кислородов и малой разности масс, которая может перекрываться изотопными распределениями.

В нашем случае наиболее применим пятый способ мечения, поскольку:

1) он не зависит от аминокислотного состава, как первый, второй и четвёртый способы;

2) никак не связан с MS/MS-анализом;

3) при использовании ИЦР масс-спектрометрии с преобразованием Фурье мы вполне способны различить пики изотопного распределения самих пептидов (C^{12}/C^{13} углероды) и разницы в изотопах воды, которая использовалась при трипсинолизе.

Этот метод лишен проблемы неодновременной ко-элюции пептидов, что есть бич, например, двух первых способов.

VI. База точных массово-временных меток

За основу была взята система управления базами данных (СУБД) MySQL 5.0.45 с открытым кодом и свободно распространяемая под лицензией GNU GPL. Структура созданной базы (рис. 2) позволяет поддерживать одновременно несколько проектов, выявлять схожесть и различия экспериментов и, конечно же, производить поиск по точной массово-временной метке.

Рис. 2. Блок-схема внутреннего устройства базы данных и стандартных процедур заполнения и поиска

Исходными данными являются файлы результатов поиска в Mascot. Загрузив их в базу при помощи специальной программы-загрузчика, можно формировать отчёты об идентифицированных в экспериментах белках по разным критериям. Для использования в подходе точной массово-временной метки нам нужна база данных с высокой достоверностью идентифицированных пептидов — ведь любая ошибка приведёт к множественным неверным идентификациям в будущем. Поэтому лишь небольшая часть из всех идентифицированных пептидов попадает в итоговую базу, пептиды должны удовлетворять следующим двум критериям, чтобы пройти отбор: достоверность идентификации пептида больше 99,5% (пептидам присваивается рейтинг, который и определяет достоверность), в каждый идентифицированный белок должно входить минимум 2 пептида. На данный момент мы используем более строгий критерий: сначала каждому белку приписываются все из найденных пептидов, которые он мог содержать, но далее пептиды остаются только в белках, набравших наибольший суммарный пептидный рейтинг. Можно ещё дальше ужесточить критерии и потребовать уникальности пептидов, в этом случае допустима идентификация белка лишь по одному пептиду. Если более одного белка содержат одинаковый набор пептидов, то на данный момент в базу будут занесены все пептиды, но в ближайшем будущем будет добавлена проверка гомологичности идентифицированных белков. Если они не являются схожими, то подобные пептиды не могут считаться репрезентативными, следовательно, не должны заноситься в базу. Делаться это будет методом попарного выравнивания белковых последовательностей, например, с помощью пакетов clustalW, BLASTp, FASTA.

При заполнении базы записываются точно измеренное отношение массы к заряду, масса, полученная после деконволюции, нормализованные времена начала и конца выхода из колонки (рис. 2). Нормализация может производиться несколькими методами:

1) с использованием внешних калибрантов (в каждый исследуемый образец вводятся стандартные пептиды, смывающиеся с колонки в начале, середине и конце эксперимента, например, пептиды цитохрома C);

2) с использованием внутренних калибрантов (например, пептиды гидролизующего фермента — в нашем случае трипсина);

3) нормализация по теоретически рассчитанному времени.

Важно не перегрузить колонку во время эксперимента — это может сильно уширить хроматографичесике пики, что в свою очередь приведёт к потере уникальности массово-временной метки.

После заполнения базы референсными данными можно приступать к идентификации, используя массово-временные метки. Проведя эксперимент без использования тандемной масс-спектрометрии, полученные результаты обрабатываются программой Decon2LS (Pacific-Northwest National Laboratory) для проведения деконволюции и деизотопирования (может использоваться также и любая другая подобная программа). Далее нормализуются времена выхода обнаруженных ионов, и эти данные сравниваются с записями в базе данных массово-временных меток.

VII. Эксперименты, оборудование и пробоподготовка

Эксперименты проводились на масс-спектрометре ионно-циклотронного резонанса, совмещенного с линейной ионной ловушкой Thermo LTQ-FT и жидкостном хроматографе Agilent 1100. Для хроматографии использовались самодельные колонки внутренним диаметром 75 мкм, обращенная фаза Reprosil-Pur C18 3 мкм с порами 100 А.

Для тестовых экспериментов были собраны образцы мочи 10 человек, из которых выделялась белковая фракция и трипсинолизовалась в растворе без предварительного разделения белков при помощи гель-электрофореза.

Полученная пептидная смесь анализировалась в хромато-масс-спектрометрических экспериментах. Для хроматографии в качестве сольвентов использовались вода и ацетонитрил с линейным градиентом от 10% содержания ацетонитрила до 50% в электрораспылении. При масс-анализе измерялись масс-спектры высокой точности и разрешения в диапазоне 300–2000 m/z в ячейке ИЦР, по ним определялись массы родительских ионов и их зарядовые состояния. Для тех ионов, у которых удалось определить зарядовое состояние, измерялись спектры столкновительной диссоциации (CID) в линейной ионной ловушке.

VIII. Результаты и обсуждение

В проведённом эксперименте была отработана новая для практики цепочка по быстрой идентификации пептидов с возможным количественным анализом.

Во-первых, удалось наполнить базу данных экспериментальных величин, которые соответствуют массам и временам выхода с колонки для пептидов из белков протеома мочи, которую уже можно использовать для скрининга пациента (присутствует или нет данный белок). В базе содержится порядка 500 пептидов из 100 белков (ранее в медицине считалось, что белок в моче — плохой показатель), наличие или отсутствие (возможно и концентрация) которых потенциально может быть маркером заболеваний. Во-вторых, данная база разработана с прицелом под количественный анализ, схема которого уже предложена нами:

1) сбор мочи (с возможным ингибированием протеаз), выделение белковой фракции,

2) трипсинолиз обоих образцов («пациента» и контроля) и смешивание,

3) хроматографическое разделение на колонке, MS-анализ (без MS/MS),

4) поиск по базе данных точных массово-временных меток и одновременное сравнение интенсивностей пиков (приписанных пептиду из БД), отличающихся на 2 (или 4) Да.

На четвертом этапе пептиды, трипсинолизованные в обычной воде и в воде H_2O^{18}, будут появляться одновременно на спектре хроматограммы (рис. 3): пики, которые будут отличаться на 2 Да, будут выделяться, как первый вариант, программным обеспечением от Finnigan (Quan Browser) или open source ПО от BioGrid Illinois, но в тот же момент все пики будут искаться по БД. В конечном счете нас будет интересовать относительные интенсивности двух пиков, когда «более лёгкий» совпал по массе и времени выхода со значением в БД, но и ему соответствует пик, смещённый «вправо» на разницу между двумя изотопами кислорода.

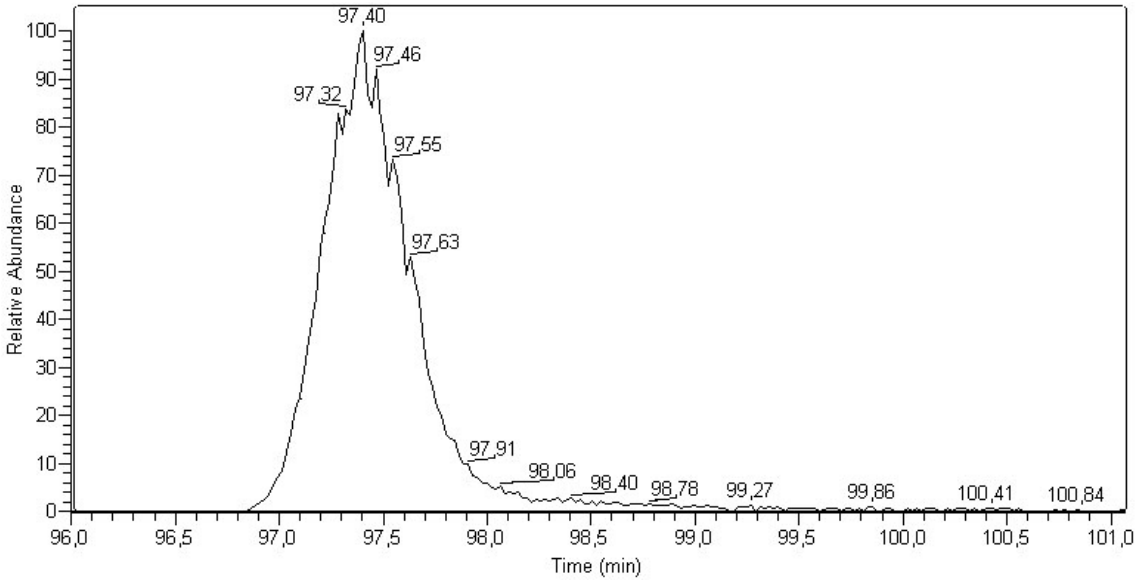

Рис. 3. Хроматограмма. В базу записываются времена начала и конца выхода

Как второй вариант, будут производиться два одновременных поиска по одной и той же БД, но со сдвигом примерно на 2 Да (разницу между кислородами $O^1$6 и $O^1$8) во втором поиске масс всех пептидов. Одновременное совпадение номера спектра, внутреннего номера пептида в обоих поисках будет гарантировать, что пептиды одинаковые, ко-элюировали одновременно.

В скором времени будут представлены результаты по количественному скринингу по нашей БД точных массово-временных меток с применением разработанной схемы.

СПИСОК ЛИТЕРАТУРЫ

1. *Bogdanov B.C., Smith R.D.* Proteomics by FTICR mass spectrometry: Top down and bottom up // Mass Spectrometry Reviews. — 2004. — V. 24, N. 2. — P. 168–200.

2. *Kelleher N.L., Hong Y.L., Valaskovic G.A., Aaserud D.J., Fridriksson E.K., McLafferty F.W.* Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry // J. Am. Chem. Soc. — 1999. — V. 10.1021/ja973655h, N. 121. — P. 806–812.

3. *Pappin D.J., Hojrup P.N., Bleasby A.J.* Rapid identification of proteins by peptide-mass fingerprinting // Curr Biol. — 1993. — N. 3(6). — 327–332.

4. *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.

5. *Pasa-Toliж L.F., Masselon C.E., Barry R.C., Shen Y.W., Smith R.D.* Proteomic analyses using an accurate mass and time tag strategy // Biotechniques. — 2004. — V. 37(4), N. 621-4. — P. 626–636.

6. *Mirgorodskaya E.L., Wanker E.D., Otto A.E., Lehrach H.H., Gobom J.O.* Method for Qualitative Comparisons of Protein Mixtures Based on Enzyme-Catalyzed Stable-Isotope Incorporation // J. Proteome Res. — 2005. — V. 10.1021/pr050219i, N. 4 (6). — P. 2109–2116.

Поступила в редакцию 22.01.2008.