УДК 543.51

МЕТОДЫ ОБРАБОТКИ МАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ ПРИ ИДЕНТИФИКАЦИИ ПЕПТИДОВ И БЕЛКОВ

Е.И. Беризовская[1], А.А. Ихалайнен[1], А.М. Антохин[1], В.Ф. Таранченко[1], В.М. Гончаров[1], Д.А. Митрофанов[1], А.В. Удинцев[1], А.В. Аксенов[1], О.А. Шевлякова[1], И.А. Родин[2], О.А. Шпигун[2]

([1]ФГУП НЦ «Сигнал»; [2]Московский государственный университет имени М.В. Ломоносова; e-mail: eiberizovskaya@rambler.ru)

Перед началом идентификации пептидов и белков методом масс-спектрометрии, как правило, проводят ферментативное расщепление, а затем записывают масс-спектрометрические данные полученных пептидов. Применяют различные алгоритмы и программы для идентифицирования пептидов и белков с помощью поиска по базам данных и *de novo* секвенирования. Рассмотрены основные используемые для этих целей программные продукты, кратко описаны базы данных.

Ключевые слова: масс-спектрометрия, ферментативное расщепление, системы поиска, базы данных, *de novo* секвенирование.

В настоящее время масс-спектрометрия (МС) является одним из самых распространенных методов в исследованиях пептидов и белков [1]. Быстрое развитие данного метода способствует увеличению скорости, чувствительности, качества данных и надежности результатов. Появляются новые инструменты, повышаются производительность и эффективность программного обеспечения, скорость получения данных и удобство использования [2]. При этом исследователи получают гигабайты данных в течение нескольких часов, что затрудняет их анализ вручную [3]. Форматы данных, получаемых в масс-спектрометрических исследованиях, зависят от фирмы-производителя масс-спектрометрического оборудования: MS («Agilent») [4], WIFF («ABI/Sciex») [5], FID/.YEP/.BAF («Shimadzu», «Bruker») [6, 7], RAW («Thermo Scientific») [8], MassLynx («Waters») [9]. Таким образом, обработка экспериментальных данных в исследованиях пептидов и белков представляет собой сложную задачу, для решения которой требуются обширные знания многих пакетов программ, имеющих различные алгоритмы, требования к формату данных и пользовательские интерфейсы.

Основные подходы

Для идентификации белков по масс-спектрам используют два метода: поиск по базам данных и *de novo* секвенирование. При масс-спектрометрическом анализе пептидов и белков различают три подхода [10]: «top-down», «middle-down» и «bottom-up».

«Top-down» – установление точной структуры белков исключительно возможностями масс-спектрометрии [11–13]. В данном методе информацию об аминокислотной последовательности пептида получают за счет выделения из узкого диапазона масс родительских ионов, их фрагментации и последующей записи масс-спектра фрагментных (дочерних) ионов. Схема эксперимента с использованием данной стратегии представлена на рис. 1 [10, 14].

Успешная идентификация пептидов с использованием вышеуказанного подхода продемонстрирована в работах [15–23]. «Middle-down» – недавно возникший подход, основанный на неполном протеолизе исходного белка с последующим масс-спектрометрическим анализом данных длинных пептидов методом «top-down» [24]. Таким образом, данный метод комбинирует главные особенности подходов «top-down» и «bottom up». Схема эксперимента с использованием данной стратегии представлена на рис. 2 [10].

Этот подход был предложен в работе [25]. При его реализации используют фермент, который «режет» пептид не так часто, как трипсин, что приводит к получению меньшего числа пептидов с большей молекулярной массой по сравнению с фрагментами, получаемыми при трипсинолизе. Примером такого фермента являются протеазы Lys-C (расщепляет связи Lys-X [26]), Asp-N (расщепляющая по остатку аспарагина с N-конца [27]), Arg-C и др. [28–31]. Для успешного применения данного метода необходима также приборная база, позволяющая получать масс-спектры высокого разрешения и масс-спектры высоких порядков [32, 33]. Показано применение данного подхода для получения информации о местах посттрансляционных

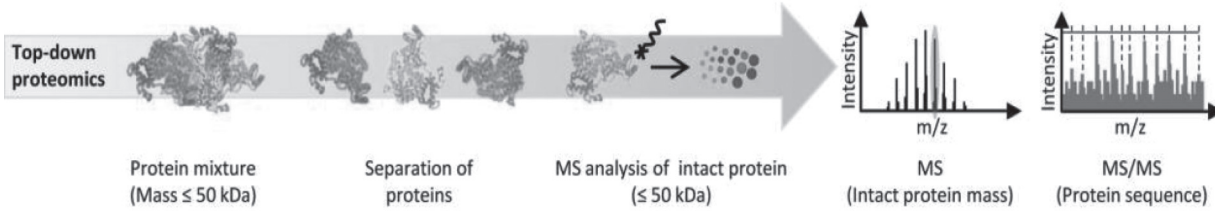

Рис. 1. Схема эксперимента с использованием стратегии «top-down» [10]

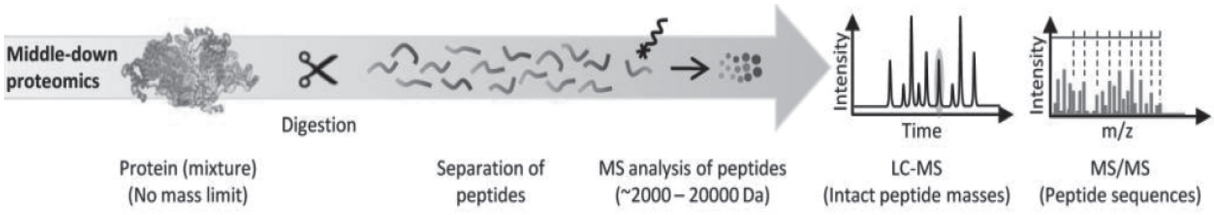

Рис. 2. Схема эксперимента с использованием стратегии «middle-down» [10]

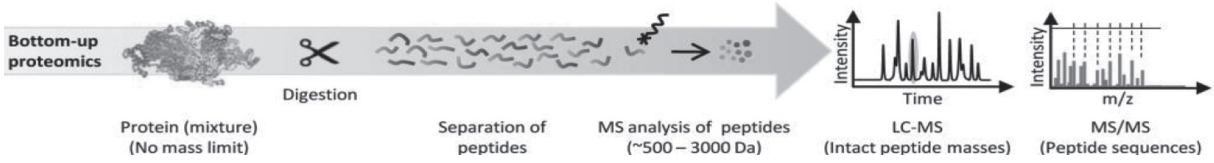

Рис. 3. Схема эксперимента с использованием стратегии «bottom-up» [10]

модификаций и идентифицирования конкретных изоформ [25, 34, 35]. Для анализа интактных белков, а также пептидов с большой молекулярной массой используют масс-спектрометрию с лазерной десорбционной ионизацией в присутствии матрицы (МАЛДИ).

«Bottom-up» – наиболее распространенный вариант анализа соединений пептидной структуры [1, 11, 36–42]. Эта методика была предложена одновременно пятью группами [37–41]. Образец расщепляется ферментом по заранее известным аминокислотам. В случае трипсинолиза в белке гидролизуются пептидные связи, образованные карбоксильными группами лизина и аргинина, в случае протеиназы V8 из *Staphylococcus aureus* гидролизуются связи Glu-X (X ≠ Pro) и т.д. [43]. Схема эксперимента с использованием данной стратегии представлена на рис. 3 [10].

После ферментативного расщепления получившуюся смесь пептидов очищают от низкомолекулярных примесей и анализируют на масс-спектрометрах, оснащенных ионным источником МАЛДИ или источником ионизации электрораспылением. По полученному спектру масс пептидов проводят поиск в базах данных с использованием различных поисковых систем, которые сопоставляют измеренные массы с массами пептидов, гидролизованных теоретически (если заранее задан известный фермент). Недавние исследования с использованием данного подхода представлены в работах [44–48].

Следует отметить, что подход «top-down» имеет несколько преимуществ перед «bottom-up», например, высокое покрытие последовательности для белка [49], способность определить посттрансляционные модификации и возможные мутации [50, 51]. Использование вышеуказанных ферментов возможно в комбинации, что позволяет увеличить число пептидов, регистрируемых масс спектрометрическим методом, и повышает эффективность протеолиза [52].

В табл. 1 приведены основные используемые ферменты и их характеристики [10, 43, 53]. Ферментативное расщепление белков осуществляется в геле или растворе, а также с помощью концентрирующих фильтров [54]. В первом случае протеолизу предшествует разделение белков по массе (1D-гель электрофорез) или по массе и изоэлектрической точке (2D-гель электрофорез). Ферментативное расщепление в геле дает возможность применять сильный детергент типа додецилсульфата натрия для экстракции белков из биологического материала. В данных условиях белки хорошо денатурированы и префракционированы перед воздействием фермента, однако данная методика плохо сопоставима с последующим масс-спектрометрическим количественным анализом, а также требует большого избытка фермента.

Сведения о специфичности используемых ферментов

Фермент	Оптимум pH	Основной тип гидролиза	Гидролиз не идет
Трипсин	8,0	–Lys–↓–X–; –Arg–↓–X–	–Lys–Pro–
Glu-C	4,0–7,8	–Glu–↓–X–; (–Asp–↓–X–)*	–Glu–Pro–; –Glu–Glu–
Asp-N	7,0–8,0	–X–↓–Asp–; –X–↓–cysteic acid	–X–↓–Cys–
Lys-C	8,5–8,8	–Lys–↓–X–	–
Arg-C	7,2–8,0	–Arg–↓–X–	–

*Скорость протеолиза в 3 000 раз медленнее основного.

Гидролиз в растворе, напротив, хорошо подходит для количественного анализа. Ему предшествуют дополнительные стадии пробоподготовки, включающие восстановление и алкилирование, направленные на денатурацию нативного белка и восстановление дисульфидных связей, что приводит к разворачиванию аминокислотной цепи и экспозиции аминокислотных сайтов для расщепления протеазами. Денатурирующие условия могут интерферировать с активностью ферментов, но в присутствии низких концентраций денатурирующих агентов типа мочевины большинство протеаз сохраняет активность. В качестве хаотропного агента используют детергенты, например дезоксихолат натрия, характеризующийся стабильностью и химической инертностью по отношению к белкам при повышенной температуре [55].

Гидролиз с помощью концентрирующих фильтров объединяет преимущества использования электрофоретического разделения в геле, высоких концентраций хаотропов (6–8 М мочевина) и возможности анализировать белки биологической пробы без предварительного разделения.

Полученные в результате протеолитического расщепления пептиды разделяют одномерной или многомерной жидкостной хроматографией (ЖХ) и вводят в масс-спектрометр. Отношение массы к заряду (*m/z*) и интенсивность регистрируются для всех ионов пептидов в масс-спектрах (MS^1), и один или более ионов пептидов отбирают для фрагментации. Полученные фрагментарные ионы анализируются и регистрируются в тандемном МС-спектре (МС/МС или MC^2). Спектры MS^2 используют для идентификации пептидов.

Фрагментация родительских ионов в столкновительной ячейке происходит путем разрыва одной из пептидных связей с образованием двух комплементарных друг другу серий ионов. Дочерние ионы обеих серий обозначаются в соответствии с общепринятой классификацией, предложенной в работе [56]: фрагменты, содержащие N-конец аминокислотной последовательности, в зависимости от разорванной связи обозначаются буквами *a*, *b* и *c*; фрагменты, содержащие С-конец аминокислотной последовательности – буквами *x*, *y* и *z* (рис. 4).

Разность масс между соседними пиками каждой из серий соответствует массе аминокислотного остатка, расположенного в соответствующем месте последовательности. Следует отметить, что в реальных масс-спектрах серии пиков представлены не полностью, что создает определенные трудности при интерпретации масс-спектра [57]. По полученному масс-спектру пептидов производят поиск в базах данных с использованием различных поисковых программ [58].

Наиболее распространенные поисковые системы

На практике пептиды идентифицируют, используя, как правило, системы поиска по базе данных белков, например: Mascot [59], SEQUEST [60], X!Tandem [61], Inspect [62], OMSSA [63], MassMatrix [64], Crux [65], MyriMatch [66], MS-GFDB [67] и др. Наиболее распространенными поисковыми системами являются первые три из перечисленных выше. Поисковая система Mascot основана на алгоритме MOWSE (Molecular Weight SEarch), предложенном в 1993 г. Ознакомиться с ресурсом можно на сайте http://www.matrixscience.com [40]. Данный алгоритм использует поиск по массовым «отпечаткам пальцев» пептидов. Вначале сравнивают массы пептидов из базы данных

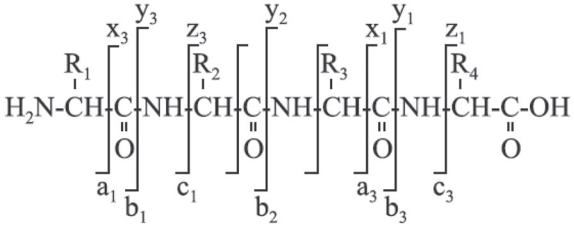

Рис. 4. Схематичное изображение места разрыва связи при образовании N-концевых (*a, b, c*) и С-концевых (*y, x, z*) фрагментов [24]

с экспериментальными данными масс пептидов с учетом заданной погрешности. Затем для каждого совпадения рассчитывают величину Score (величина уровня достоверности) в соответствии с (1):

$$\text{Score} = \frac{50000}{M_{prot} \times \Pi_n m_{i,j}}, \quad (1)$$

где M_{prot} – молекулярная масса каждого совпавшего белка, П – произведение, которое рассчитывается из Mowse-матрицы весов M для каждого совпадения экспериментальных данных и масс пептидов, рассчитанных из записей в геномной базе данных [68].

Данный алгоритм можно применять для поиска MC/MC. В этом случае в формуле для Score роль белка выполняет пептид, а роль пептида – фрагмент. Сумма Score пептидов дает Score для белка [59, 68].

Поисковая система Sequest основана на отдельной идентификации каждого масс-спектра [60]. Ознакомиться с ресурсом можно на сайтах http://proteomicsresource.washington.edu/protocols06/sequest.php и http://fields.scripps.edu/ sequest/index.html.

Вначале из белковой базы данных отбирают пептиды, соответствующие массе родительского иона исследуемого пептида. Для каждого кандидата генерируется теоретический масс-спектр фрагментации и сверяется с экспериментальными данными [69]. Затем проводят кросс-корреляционный анализ спектров, который сводится к вычислению целочисленной функции $R(\tau)$ по уравнению (2):

$$R(\tau) = \sum_{i=0}^{n-1} x[i]y[i+\tau], \quad (2)$$

где n – число каналов в масс-спектре; $x[i]$ и $y[i]$ – интенсивность сигналов масс-спектра на i-ом канале; τ – смещение рассчитанного спектра относительно экспериментального. Данная функция максимальна при $\tau = 0$ [70].

Поисковая система X!Tandem наиболее развита, так как является программным обеспечением с открытым исходным кодом [61]. Ознакомиться с ресурсом можно на сайте http://www.thegpm.org/tandem.

В данном алгоритме рассчитанный и экспериментальный масс-спектры приводятся к виду многомерного вектора из $n = m_{prt}/\Delta m$, где m_{prt} – масса родительского иона, а Δm – максимальная погрешность при определении массы дочернего иона. В рассчитанный масс-спектр включаются массы ионов серий и массы их ионов с нейтральными потерями (NH_3 и H_2O). Для оценки совпадения рассчитанного и экспериментального спектров используется рейтинг, вычисляемый по формуле (3):

$$x = n_b!n_y!\sum_{i=0}^{n} I_i P_i, \quad (3)$$

где n_b и n_y – число обнаруженных в экспериментальном масс-спектре ионов b- и y-серий соответственно; $\sum_{i=0}^{n} I_i P_i$ – скалярное произведение векторов экспериментального и рассчитанного масс-спектров.

Для оценки достоверности идентификации белка вычисляют рейтинг белка E_{pro} по формуле (4), основанный на достоверности e каждого спектра пептида этого белка:

$$E_{pro} = \sqrt{M}\prod_{i=1}^{n} e_i(x_i^*), \quad (4)$$

где N – общее число спектров; n – количество спектров, соотнесенных с белком [71].

X!Tandem выполняет также идентификацию пептидов с неполным или неспецифическим гидролизом или при наличии модификаций в них за относительно короткое время [72]. Краткое описание более 100 различных алгоритмов и пакетов программ обработки масс-спектрометрических данных по пептидам и белкам представлено на сайтах http://en.wikipedia.org/wiki/Mass_spectrometry_software и http://www.ms-utils.org.

Базы данных характеристик пептидов и белков

Использование баз данных для идентификации белков и пептидов позволяет расшифровывать масс-спектры сложных смесей за короткое время [73]. Почти все известные в настоящий момент аминокислотные последовательности белков и пептидов объединены в базы данных, которые находятся в открытом доступе в сети Интернет [24]. Каждая из них имеет свой формат хранения данных, разную степень избыточности, взаимосвязи с родственными или аналогичными базами данных. Все базы данных можно разделить на пять типов. Первый тип – архивные базы данных, в которых информация добавляется исследователями. К таким базам данных относятся (GenBank, EMBL, PDB). Второй тип – курируемые базы данных (содержание записей курируют специалисты), к ним относится, например, Swiss-Prot. Третий тип – автоматические базы данных (записи генерируются компьютерными программами), к ним относится, например, TrEMBL. Четвертый тип – производные

базы данных, которые пополняются за счет обработки данных из баз данных первых двух типов (SCOP, PFAM, GO и др.). Пятый тип – интегрированные базы данных, которые объединяют информацию из различных баз (ENTREZ) [74]. Ниже представлено краткое описание перечисленных баз данных.

GENBANK. База данных генетических последовательностей (ДНК, РНК и белков) Национального центра биотехнологической информации США GenBank была основана в 1982 г. Это аннотированная база данных всех общедоступных последовательностей, снабженная литературными ссылками и другой биологической информацией. Эта база является частью объединения International Nucleotide Sequence Database Collaboration (INSDC), которое включает три крупнейшие коллекции нуклеотидных последовательностей: DDBJ (DNA Data Bank of Japan), EMBL (European Molecular Biology Laboratory) и GenBank (National Center for Biotechnology Information). Постоянно совершенствуются и создаются новые средства для депонирования новых последовательностей в базу, средства эффективного поиска в базе. Содержимое банка удваивается в объеме каждые 18 месяцев [75–78].

Доступ к данным осуществляется либо через сеть интернет (http://www.ncbi.nlm.nih.gov/Genbank/), либо через файловый сервер FTP (http://ftp.ncbi.nlm.nih.gov/genbank).

EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL). База данных нуклеотидных последовательностей Европейской молекулярно-биологической лаборатории была основана в 1982 г., она содержит разнообразную информацию о каждом фрагменте последовательностей, включая литературные ссылки, перекрестные ссылки на документы других баз данных и др. [79–84]. На сегодняшний день EMBL состоит из 18 разделов. Доступ к данным может осуществляться либо через EBI веб интерфейс (http://www.ebi.ac.uk/embl/), либо с помощью CD-ROM.

PROTEIN DATA BANK (PDB). База данных Брукхейвенской национальной лаборатории США была основана в 1971 г., она содержит информацию о 3D-структурах биологических макромолекул. С 2002 г. в основном депозитарии хранятся структуры, определенные экспериментально с помощью методов ядерно-магнитного резонанса, рентгеноструктурного анализа и др. Теоретические структуры выделены в подбазу, доступную по FTP [85–88]. База данных обновляется еженедельно.

Доступ к данным может осуществляться через сеть интернет (http://www.rcsb.org/pdb/), через EBI веб интерфейс (http://www.ebi.ac.uk/pdbe/) и через EBI файловый сервер ftp (http://ftp.ebi.ac.uk/pub/databases/pdb/).

SWISS-PROT. Эта база данных была создана в 1986 г. в Европейском институте биоинформатики. На сегодняшний день эта база данных считается наиболее надежной, что обусловлено минимальной избыточностью вследствие высокого уровня аннотации, выполненной вручную. Туда включена информация о функциях белка, его структурных доменах, посттрансляционных модификациях, различных вариантах последовательности и др. База обладает высоким уровнем интеграции с другими базами данных [89–93].

Доступ к данным может осуществляться различными путями: через EBI веб интерфейс (http://www.ebi.ac.uk/), через EBI файловый сервер ftp (http://ftp.ebi.ac.uk/pub/databases/swissprot/) и через сеть интернет (www.expasy.org/sprot/).

EMBL PROTEIN-CODING DNA SEQUENCE FEATURES TRANSLATED INTO PEPTIDE SEQUENCES (TrEMBL). База данных была разработана в 1996 г. как приложение к Swiss-Prot, она является автоматически аннотируемой. База содержит белковые последовательности, полученные теоретически трансляцией нуклеотидных последовательностей [89, 91, 94–96].

Доступ к данным осуществляется через сеть интернет (http://www.uniprot.org/), и через EBI файловый сервер ftp (http://ftp.ebi.ac.uk/pub/databases/trembl).

PROTEIN INFORMATION RESOURSE (PIR). База данных была создана в 1984 г. в Национальном фонде биомедицинских исследований США на основе NBRF Protein Sequense Database, разрабатываемой в течение 20 лет Margaret O. Dayhoff. С 1988 г. является международной базой в результате сотрудничества между Национальным фондом биомедицинских исследований в США, институтом последовательностей белков, институтом биохимии им. Макса Планка в Германии, Международной базы данных по белкам в Японии [97].

База данных PIR разделена на 4 секции: PIR1, PIR2, PIR3 и PIR4. В настоящее время на PIR1 и PIR2 приходится около 99% от всех записей. PIR1 полностью классифицирована по суперсемействам и аннотирована. PIR2 является переходным разделом к PIR1 от PIR3. PIR3 служит временным хранилищем для новых записей и включает менее 1% от общей базы данных. В ней находятся неклассифицированные и неаннотированные записи. PIR4 включает последовательности, не встречающиеся в природе или неклассифицированные последовательности [98]. База данных PIR содержит следующую информацию: название белка и ор-

ганизма, из которого он был выделен, аминокислотная последовательность, основные характеристики белка, его функции в организме, ссылки на литературу [99–106].

Доступ к данным может осуществляться через интернет (http://pir.georgetown.edu/), а также с помощью CD-ROM.

С декабря 2003 г. начал свою работу проект UniProt [94], который объединил базы данных Swiss-Prot, TrEMBL и PIR-PSD [107]. Проект предоставляет четыре основных базы данных: UniProtKB (Swiss-Prot и TrEMBL), UniParc, UniRef и UniMes [108].

База знаний UniProtKB состоит из двух частей: UniProtKB/Swiss-Prot (содержит обзорные записи, аннотированные вручную) и UniProtKB/TrEMBL (содержит нерецензированные записи, аннотированные автоматически) [109]. По состоянию на 19 марта 2014 г. UniProtKB/Swiss-Prot содержал 542 782 последовательности, а UniProtKB/TrEMBL – 54 247 468 последовательностей [110, 111].

UniProtKB/Swiss-Prot – аннотированная вручную база данных белковых последовательностей. При составлении аннотации требуется подробный анализ последовательности белка и данных о нем из научной литературы [112]. Перед включением в UniProtKB/Swiss-Prot аннотированные записи проходят контроль качества. Все аннотации подвергаются регулярным проверкам и при появлении новых данных существующие записи обновляются [113]. Более подробную информацию о включаемых данных можно посмотреть в работе [94].

UniProtKB/TrEMBL – автоматически аннотированная база данных белковых последовательностей [114]. Она содержит аминокислотные последовательности из Protein Data Bank (PDB) [115], RefSeq [116], Complex Carbohydrate Structural Database (CCDS) [117] и трансляции аннотированных кодирующих последовательностей в базах данных последовательностей нуклеотидов EMBL-Bank [118], GenBank [119], DNA Data Bank of Japan (DDBJ) [120].

База знаний UniParc является архивом и содержит последовательности белков из основных общедоступных баз данных [121]. Так как один и тот же белок может находиться в нескольких базах данных или дублироваться в одной и той же базе данных, UniParc сохраняет каждую последовательность только один раз для минимизации избыточности. Одинаковые последовательности объединяются. Каждой из них присвоен уникальный код, который позволяет идентифицировать один и тот же белок из различных баз данных. UniParc содержит только белковые последовательности без аннотации. Если в исходных базах данных последовательности изменяются, эти изменения переносятся в UniParc, а история всех изменений сохраняется в архиве.

Кластер ссылок UniRef включает три базы данных (UniRef100, UniRef90 и UniRef50) и сформирован из наборов последовательностей из UniProtKB (в том числе изоформ) и выбранных записей UniParc [122]. База данных UniRef100 сочетает идентичные последовательности и фрагменты последовательности от 11 или более остатков (любого организма) в одной записи UniRef. UniRef90 построен путем объединения UniRef100 последовательностей с 11 или более остатками с помощью алгоритма CD-HIT, при этом каждый кластер состоит из последовательностей, которые имеют ~90% идентичности с самой длинной последовательностью. Аналогично UniRef50 построен путем объединения UniRef90 последовательностей, которые имеют ~50% идентичности с самой длинной последовательностью [123]. До 2013 г. не было порогового перекрытия, так как кластеры были более разнородными в длину.

Снижение избыточности увеличивает скорость поиска подобия и позволяет повысить надежность поиска далеких родственных белков. Так, UniRef90 и UniRef50 дают уменьшение размера базы данных примерно на 58 и 79% соответственно [124]. Записи UniRef содержат информацию о последовательности белка, а также регистрационные номера всех записей и ссылок на аннотации в UniProtKB. UniRef доступен с сайта UniREF FTP [125].

База знаний UniMES (The UniProt Metagenomic and Environmental Sequences database) – база данных метагеномных последовательностей и неизвестных последовательностей из окружающей среды [126]. UniMES в настоящее время содержит данные о белковых последовательностях организмов из мирового океана. Данные из этой базы отсутствуют в базе знаний UniProt или в кластерах ссылок UniRef, но интегрированы в UniParc [127]. UniMES доступен через UniProt сервер FTP файла в формате FASTA (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/unimes/).

Запросы для всех данных в UniProt осуществляются через веб-сайт (http://www.uniprot.org/), через MartView BioMart (http://www.ebi.ac.uk/uniprot/ biomart/martview) и через UniProt DAS-сервер (http://www.ebi.ac.uk/uniprot-das).

STRUCTURAL CLASSIFICATION OF PROTEINS (SCOP). Эта база данных создана в 1994 г. и непрерывно обновляется [128, 129]. Она устанавливает эволюционные и структурные взаимос-

вязи между всеми белками с известной структурой, включая белки из PDB. Классификация SCOP сделана вручную визуальным сравнением структур. База содержит следующую информацию: изображение структуры, данные об аминокислотной последовательности, литературные ссылки и др. [130–135]. О последних усовершенствованиях SCOP можно узнать в работах [136, 137].

Доступ осуществляется через интернет: http://scop.mrc-lmb.cam.ac.uk/scop/

PROTEIN FAMILIES DATABASE OF ALIGNMENTS AND HMMS (PFAM). Создана в 1995 г. [138, 139]. Это большая база данных семейств белков и доменов, состоящая из двух частей: PFAMA (содержит курируемые вручную аннотированные белковые семейства) и PFAMB (состоит из автоматически генерируемых из базы данных ProDom семей доменных белков) [140–146]. PFAM охватывает белковые последовательности, представленные в UniProtKB и NCBI GenPept [147, 148]. Об усовершенствованиях базы данных можно узнать в работах [149, 150].

Доступ осуществляется через интернет: http://pfam.sanger.ac.uk

GENE ONTOLOGY CONSORTIUM DATABASE (GO). База знаний была создана в 1998 г. как проект, целью которого было создание унифицированной терминологии для аннотации генов и генных продуктов всех биологических видов [151, 152]. Это позволило унифицировать описания в различных базах данных и облегчить поиск в них необходимого гена. GO является независимой базой данных, сотрудничающей с другими базами [153–156]. В течение нескольких последних лет GO внедрил ряд изменений для увеличения качества, количества и специфичности аннотаций [157–160].

Доступ осуществляется через сеть интернет: http://www.geneontology.org/

MOLECULAR BIOLOGY DATABASE AND RETRIEVAL SYSTEM (Entrez). Интегрированная база данных содержит нуклеотидные и белковые последовательности, геномные карты, семейства белков и доменов, структуры белков и др. Кроме того, Entrez позволяет проводить поиск литературы в данных через PubMed и OMIM. В состав данной интегрированной базы данных входят следующие ресурсы: GenBank, dbEST, dbSTS, SwissProt, PIR, PDB, PRF, GSDB и др. Данные из перечисленных ресурсов поступают в Entrez после присвоения уникального идентификатора последовательности, перевода документов в единое хранилище, проверки данных, проверки ссылок по базе данных MedLine и проверки названий организмов по классификации GenBank Taxonomy. [161–166].

Доступ осуществляется через интернет: http://www.ncbi.nlm.nih.gov/gquery/.

Более полный список баз данных разного назначения можно найти в работе [167].

Системы идентификации аминокислотной последовательности пептидов и белков методом de novo секвенирования. Определение аминокислотной последовательности пептидов и белков без использования поисковых программ и баз данных называют *de novo* секвенированием. Такой подход применяют для идентификации не описанных ранее белков, при наличии неисследованных мутаций, посттрансляционных модификаций и т.д. Применяемые алгоритмы *de novo* секвенирования основаны на различных математических методах. Первые алгоритмы определения аминокислотной последовательности [168, 169] представляли собой перебор всех возможных комбинаций аминокислот, составляющих массу родительского иона, фрагментацию которых сравнивали с экспериментальным масс-спектром. Очевидно, что погрешность измерения массы родительского иона влечет за собой увеличение числа соответствующих ему комбинаций.

Еще один подход представляет собой рассмотрение малой части последовательности (тэга), к которой с обеих сторон добавляются аминокислоты до тех пор, пока не будет достигнута соответствующая масса родительского иона [170–173]. При этом неполная фрагментация пептида может привести к потере кандидатных последовательностей.

В 1990 г. предложена теория графов [174]. Ее суть состоит в том, что каждый пик в масс-спектре сопоставляется с вершиной графа. Между двумя вершинами проводится ребро, если разница масс между соответствующими пиками в спектре равна массе одного или нескольких аминокислотных остатков. В граф также добавляют вершины N- и C-конца. Секвенирование *de novo* проводят за счет поиска пути в графе от N-конца к C-концу. Примеры использования теоретико-графового подхода описаны в работах [175–180] и используются в алгоритмах Lutefisk, Sherenga.

Динамическое программирование – это метод декомпозиции задач, который позволяет решить задачу об антисимметричных путях в графе спектра [181]. Более подробно он описан в работе [182]. Использование данного метода в алгоритмах PEAKS, PepNovo, AUDENS и др. приводится в работах [181, 183–186]. Для идентификации пептидов используют также скрытые Марковские

модели (СММ) [187]. Натренированная СММ определяет модель воспроизведения спектров, которую используют для оценки степени похожести теоретического спектра полученной последовательности-кандидата и экспериментального спектра. Более того, помимо предсказания хороших последовательностей СММ позволяет определить достоверность таких предсказаний [188]. Данный алгоритм применяется в программе NovoHMM [189].

Одним из новых алгоритмов, не основанных на теоретико-графовом подходе, является линейный алгоритм секвенирования [190] с высокой скоростью работы и эффективностью.

Краткое описание указанных алгоритмов приведено ниже.

Lutefisk. Данный алгоритм проводит обработку тандемных масс-спектров триптических пептидов, полученных с использованием низкоэнергетической столкновительной диссоциации и теоретико-графового подхода [176, 177]. Программа выполняет поиск *b*- и *y*-ионов в спектре, подтверждает их присутствие по наличию соответствующих пиков потери воды, аммиака и т.д., а затем по спектру от N-конца восстанавливается аминокислотная последовательность. Пептиды-кандидаты оцениваются в зависимости от длины цепочек пиков одного типа, по которым происходило восстановление последовательности и подходящие последовательности подвергают кросс-корреляционному анализу [60, 191]. Затем эти две оценки объединяют для получения результата.

Sherenga. Данный алгоритм [175] реализован в пакете программного обеспечения для обработки МС–МС-данных Spectrum Mill фирмы «Agilent Technologies» (www.agilent.com) и основан на поиске антисимметричных путей в графе спектра, т.е. в графе спектра происходит поиск сразу двух путей (для *y*- и *b*-ионов). Эти пути симметричны друг другу, направлены в противоположные стороны и не содержат более одной вершины из комплементарной пары $y \leftrightarrow b$. Оценка путей в спектре строится на основании предварительной статистической оценки вероятности появления и средней интенсивности сигналов ионов серий с различными нейтральными потерями для заданного типа инструмента. Детали поиска не раскрыты, видимо, вследствие его коммерческой реализации.

PEAKS. Алгоритм [184] использует измененную версию алгоритма *de novo* секвенирования, которая основана на динамическом программировании [192]. В отличие от теоретико-графового подхода PEAKS может идентифицировать аминокислотную последовательность при отсутствии некоторых пиков. При этом учитываются еще такие факторы, как интенсивность пиков, совпадение масс, наличие разных типов ионов и др. Также данный алгоритм проводит предварительную обработку спектров.

В 2013 г. выпущена версия 7.0 данного программного обеспечения. О внесенных изменениях можно узнать из работы [193].

PepNovo. Алгоритм основан на динамическом программировании и восстанавливает аминокислотную последовательность путем поиска антисимметричных путей в графе спектра. Алгоритм проверяет две гипотезы: первая состоит в том, что масса фрагмента образована при фрагментации пептида, которому соответствует исследуемый спектр; вторая гипотеза состоит в том, что все пики в спектре были получены в результате случайного процесса. В соответствии с первой гипотезой могут быть описаны правила фрагментации пептида. В результате каждой массе фрагмента ставится в соответствие величина, равная логарифму отношения правдоподобия этих двух гипотез. Для каждой вершины вычисляют несколько значений, которым соответствуют различные комбинации аминокислот. В итоге учитываются значения, рассчитанные для каждой величины [70, 188, 189, 194, 195].

AUDENS. Программа с открытым кодом использует автоматизированное *de novo* секвенирование, которое основано на динамическом программировании. Алгоритм предварительно обрабатывает спектры, чтобы отделить сигналы фрагментов пептидов от шума с использованием правил фрагментации при столкновительной диссоциации. Каждому пику в спектре соответствует «фактор значимости», с учетом которого восстанавливают аминокислотные последовательности [185, 196].

NovoHMM. Алгоритм использует СММ и основан на статистическом моделировании масс-спектров. Используемая Марковская модель состоит из двух основных частей: в первой вычисляются вероятности перехода из одного состояния в другие, во второй – вероятности образования пиков определенной интенсивности в каждом состоянии. Результат объединяет оба набора вероятностей [187, 197].

Таким образом, идентификация пептидов и белков с использованием поиска по базам данных является наиболее простым и распространенным методом интерпретации МС/МС-данных, однако он не лишен недостатков. Во-первых, эта стратегия применима только в случае известных белков, последовательности которых занесены в базы дан-

ных. Во-вторых, при наличии посттрансляционных модификаций время поиска может быть значительно увеличено, и при этом возрастает вероятность получения ложных результатов. Кроме того, типичной проблемой при анализе протеолитических смесей пептидов является высокая степень гомологии среди пептидов. В результате в списке последовательностей-кандидатов, выдаваемых поисковой программой, возникает множество последовательностей, которым присваиваются близкие значения индекса, причем программа в окончательном списке идентификаций оставляет только одну последовательность. На результатах интерпретации отрицательно сказывается наличие погрешности в аминокислотных последовательностях белков в базах данных. В целях устранения данного недостатка постоянно ведутся работы по корректировке и обновлению баз данных.

Секвенирование *de novo* незаменимо при работе с неизвестными пептидами и белками, но при этом предъявляются очень высокие требования к качеству получаемых фрагментных спектров. Так, необходимым условием является наличие полного набора фрагментных ионов основных серий. Наилучшие результаты данный метод показывает при использовании масс-спектрометров высокого разрешения.

СПИСОК ЛИТЕРАТУРЫ

1. *Aebersold R., Mann M.* // Nature. 2003. Vol. 422. P. 198.
2. *Weisser H., Nahnsen S., Grossmann J., Nilse L., Quandt A., Brauer H., Sturm M., Kenar E., Kohlbacher O., Aebersold R., Malmström L.* // J. Proteome Res. 2013. Vol. 12. P. 1628.
3. *Pedrioli P., Eng J., Hubley R., Vogelzang M., Deutsch E., Raught B., Pratt B., Nilsson E., Angeletti R., Apweiler R., Cheung K., Costello C., Hermjakob H., Huang S., Julian R., Kapp E., McComb M., Oliver S., Omenn G., Paton N., Simpson R., Smith R., Taylor C., Zhu W., Aebersold R.* // Nat Biotechnol. 2004. Vol. 22. P. 1459.
4. URL: www.agillent.com (дата обращения: 10.10.2014)
5. URL: http://www.appliedbiosystems.com (дата обращения: 10.10.2014)
6. URL: http://www.shimadzu.com (дата обращения: 10.10.2014)
7. URL: http://www.bruker.ru. (дата обращения: 10.10.2014)
8. URL: http://www.thermo.com (дата обращения: 10.10.2014)
9. URL: http://www.waters.com/waters/home.htm (дата обращения: 10.10.2014)
10. *Switzar L., Giera M., Niessen W.* // J. Proteome Res. 2013.12. P. 1067.
11. *Bogdanov B., Smith R.* // Mass Spectrom Rev. 2005. 24. P. 168.
12. *Kinter M., Sherman N.* Protein Sequencing and identification using tandem mass spectrometry. Wiley-Interscience Series on Mass Spectrometry. N.Y., 2000. 320 p.
13. *Liebler D.C.* Introduction to proteomics. Tools for the New Biology. Totowa, 2002. 210 p.
14. *Reid G., McLuckey S.* // J. Mass Spectrom. 2002. Vol. 37. P. 663.
15. *Ross P., Huang Y., Marchese J., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., Purkayastha S., Juhasz P., Martin S., Bartlet-Jones M., He F., Jacobson A., Pappin D.* // Mol. Cell. Proteomics. 2004. Vol. 3. P. 1154.
16. *Ryan C., Souda P., Halgand F., Wong D., Loo J., Faull K., Whitelegge J.* // Am. Soc. Mass Spectrom. 2010. Vol. 21. P. 908.
17. *Calligaris D., Villard C., Terras L., Braguer D., Verdier-Pinard P., Lafitte D.* // Anal. Chem. 2010. Vol. 82. P. 6176.
18. *Meyer B., Papasotiriou D., Karas M.* // Amino Acids. 2011. Vol. 41. P. 291.
19. *Théberge R, Infusini G, Tong W, McComb ME, Costello C.* // Int. J. Mass Spectrom. 2011. Vol. 300. P. 130.
20. *Xu F., Xu, Q., Dong X., Guy M., Guner H., Hacker T., Ge Y.* // Int. J. Mass Spectrom. 2011. Vol. 305. P. 95.
21. *Nicolardi S., Andreoni A., Tabares L., van der Burgt Y., Canters G., Deelder A., Hensbergen P.* // Anal. Chem. 2012. Vol. 84. P. 2512.
22. *Kellie J., Catherman A., Durbin K., Tran J., Tipton J., Norris J., Witkowski C. Thomas P., Kelleher N.* // Anal. Chem. 2012. Vol. 84. P. 209.
23. *Breuker K., Jin M., Han X., Jiang H., McLafferty F.* // J. Am. Soc. Mass Spectrom. 2008. Vol. 19. P. 1045.
24. *Лебедев А., Артеменко К., Самгина Т.* Основы масс-спектрометрии белков и пептидов. М., Vol. 2012. 180 с.
25. *Wu S., Kim J., Hancock W., Karger B.* // J Proteome Res. 2005. Vol. 4. P. 1155.
26. *Raijmakers R., Neerincx P., Mohammed S., Heck A.* // Chem. Commun. 2010. Vol. 46. P. 8827.
27. *Ahn J., Cao M., Yu Y., Engen J.* // Biochim. Biophys. Acta. 2013. Vol. 1834. P. 1222.
28. *Swatkoski S., Gutierrez P., Ginter J., Petrov A., Dinman J., Edwards N., Fenselau C.* // J. Proteome Res. 2007. Vol. 6. P. 4525.
29. *Smith B.* // Methods Mol. Biol. 2003. Vol. 211. P. 63.
30. *Crimmins D.; Mische S.; Denslow N.* // Curr. Protoc. Protein Sci. 2005. Vol. 11. P. 1.
31. *Wu C., Tran J., Zamdborg L., Durbin K., Li M., Ahlf D., Early B., Thomas P., Sweedler J., Kelleher N.* // Nat. Methods. 2012. Vol. 9. P. 822.
32. *Roepstorff P.* // EXS. 2000. Vol. 88. P. 81.
33. *Makarov A.* // Anal. Chem. 2000. Vol. 72. P. 1156.
34. *Wu S., Kim J., Bandle R., Liotta L., Petricoin E., Karger B.* // Mol. Cell Proteomics. 2006. Vol. 5. P. 1610.
35. *Kalli A., Sweredoski M., Hess S.* // Anal. Chem. 2013. Vol. 85. P. 3501.
36. *Cottrell J.* // Pept Res. 1994. Vol. 7. P. 115.
37. *Henzel W., Billeci T., Stults J., Wong S., Grimley C., Watanabe C.* // Proc. Natl. Acad. Sci. USA. 1993. Vol. 90. P. 5011.
38. *James P., Quadroni M., Carafoli E., Gonnet G.* // Biochem Biophys. Res. Commun. 1993. Vol. 195. P. 58.

39. *Mann M., Hojrup P., Roepstorff P.* // Biol. Mass Spectrom. 1993. Vol. 22. P. 338.
40. *Pappin D., Hojrup P., Bleasby A.* // Curr. Biol. 1993. Vol. 3. P. 327.
41. *Yates J. 3rd, Speicher S., Griffin P., Hunkapiller T.* // Anal. Biochem. 1993. Vol. 214. P. 397.
42. *Washburn M., Wolters D., Yates J.* // Nat. Biotechnol. 2001. Vol. 19. P. 242.
43. *Zhang Y., Fonslow B., Shan B., Baek M., Yates J.* // Chem. Rev. 2013. Vol. 113. P. 2343.
44. *Yang Z., Ke J., Hayes M., Bryant M., Tse F.* // J. Chromatogr. B: Analyt. Technol. Biomed. Life Sci. 2009. Vol. 877. P. 1737.
45. *Paulech J., Solis N., Cordwell S.* // Biochim. Biophys. Acta. 2013. Vol. 1834. P. 372.
46. *Cannon J., Nakasone M., Fushman D., Fenselau C.* // Anal. Chem. 2012. Vol. 84. P. 10121.
47. *Mo J., Tymiak A., Chen G.* // Drug Discovery Today. 2012. Vol. 17. P. 1323.
48. *Contrepois K., Ezan E., Mann C., Fenaille F.* // J. Proteome Res. 2010. Vol. 9. P. 5501.
49. *Forbes A., Patrie S., Taylor G., Kim Y., Jiang L., Kelleher N.* // Proc. Natl. Acad. Sci USA. 2004. Vol. 101. P. 2678.
50. *Boyne M. 2nd, Pesavento J., Mizzen C., Kelleher N.* // J. Proteome Res. 2006. Vol. 5. P. 248.
51. *Siuti N., Roth M., Mizzen C., Kelleher N., Pesavento J.* // J. Proteome Res. 2006. Vol. 5. P. 233.
52. Glatter T., Ludwig C., Ahrné E., Aebersold R., Heck A., Schmidt A. // J. Proteome Res. 2012. Vol. 11. P. 5145.
53. Дарбре А. Практическая химия белка. М., 1989. 623 с.
54. *Wiśniewski J., Zougman A., Nagaraj N., Mann M.* // Nature Methods. 2009. Vol. 6. P. 359.
55. *Proc J., Kuzyk M., Hardie D., Yang J., Smith D., Jackson A., Parker C., Borchers C.* // J. Proteome Res. 2010. Vol. 9. P. 5422.
56. *Roepstorff P., Fohlman J.* // Biomed. Mass Spectrom. 1984. Vol. 11. P. 601.
57. *Papayannopoulos L.* // Mass Spectrom. Rev. 1995. Vol. 14. P. 49.
58. *Helsens K., Martens L., Vandekerckhove J., Gevaert K.* // Proteomics. 2007. Vol. 7. P. 364.
59. *Perkins D., Pappin D., Creasy D., Cottrell J.* // Electrophoresis. 1999. Vol. 20. P. 3551.
60. *Eng J., McCormack A., Yates J.* // J Am Soc Mass Spectr. 1994. 5. P. 976.
61. *Craig R., Beavis R.* // Bioinform. 2004. Vol. 20. P. 1466.
62. *Tanner S., Shu H., Frank A., Wang L., Zandi E., Mumby M., Pevzner P., Bafna V.* // Anal. Chem. 2005. Vol. 77. P. 4626.
63. *Geer L., Markey S., Kowalak J., Wagner L., Xu M., Maynard D., Yang X., Shi W., Bryant S.* // J. Proteome Res. 2004. Vol. 3. P. 958.
64. *Xu H., Freitas M.* // Proteomics. 2009. 9. P. 1548.
65. *Park C., Klammer A., Kall L., MacCoss M., Noble W.* // J. Proteome Res. 2008. Vol. 7. P. 3022.
66. *Tabb D., Fernando C., Chambers M.* // J. Proteome Res. 2007. Vol. 6. P. 654.
67. *Kim S., Mischerikow N., Bandeira N,, Navarro J., Wich L., Mohammed S., Heck A., Pevzner P.* // Mol. Cell Proteomics. 2010. Vol. 9. P. 2840.
68. *Автономов Д., Агрон И., Кононихин А., Николаев Е.* // Труды МФТИ. 2009. Vol. 1. С. 24.
69. *Yates J., Eng J., McCormack A.* //Anal. Chem. 1995. Vol. 67. P. 3202.
70. *Лютвинский Я.* Метод распознавания аминокислотных последовательностей в масс-спектрах пептидов для задач протеомики. Дис. … канд. техн. наук. СПб., 2007.
71. *Fenyo D., Beavis R.* // Anal. Chem. 2003. Vol. 75. P. 768.
72. *Craig R., Beavis R.* // Rapid Commun. Mass Spectrom. 2003. Vol. 17. P. 2310.
73. *Sparkman D.* Informatics and Mass-Spectral Databases in the Evaluation of Enviromental Mass Spectral Data. Saint Albans, 2012. 528 p.
74. Объединенный центр вычислительной биологии и информатики [электронный ресурс] // сайт. URL: http://www.jcbi.ru/index.html (дата обращения: 10.10.2014).
75. *Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Rapp B., Wheeler D.* // Nucl. Acids Res. 2000. Vol. 28. P. 15.
76. *Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Rapp B., Wheeler D.* // Nucl. Acids Res. 2002. Vol. 30. P. 17.
77. *Benson D., Karsch-Mizrachi I., Lipman D., Ostell J., Wheeler D.* // Nucl. Acids Res. 2007. Vol. 35. P. D21.
78. *Benson D., Clark K., Karsch-Mizrachi I., Lipman D., Ostell J., Sayers E.* // Nucl. Acids Res. 2014. Vol. 42. P. D32.
79. *Baker W., van den Broek A., Camon E., Hingamp P., Sterk P., Stoesser G., Tuli M.* // Nucl. Acids Res. 2000. Vol. 28. P. 19.
80. *Stoesser G., Baker W., van den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Lombard V., Lopez R., Parkinson H., Redaschi N., Sterk P., Stoehr P., Tuli M.* // Nucl. Acids Res. 2001. Vol. 29. P. 17.
81. *Stoesser G., Baker W., van den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Leinonen R., Lin Q., Lombard V., Lopez R., Redaschi N., Stoehr P., Tuli M., Tzouvara K., Vaughan R.* // Nucl. Acids Res. 2002. Vol. 30. P. 21.
82. *Kulikova T., Aldebert P., Althorpe N. et al.* // Nucl. Acids Res. 2004. Vol. 32. P. D27.
83. *Sterk P., Kulikova T., Kersey P., Apweiler R.* // Methods Mol Biol. 2007. Vol. 406. P. 1.
84. *Cochrane G., Akhtar R., Aldebert P., Althorpe N., Baldwin A., Bates K., Bhattacharyya S., Bonfield J., Bower L., Browne P., Castro M., Cox T., Demiralp F., Eberhardt R., Faruque N., Hoad G., Jang M., Kulikova T., Labarga A., Leinonen R., Leonard S., Lin Q., Lopez R., Lorenc D., McWilliam H., Mukherjee G., Nardone F., Plaister S., Robinson S., Sobhany S., Vaughan R., Wu D., Zhu W., Apweiler R., Hubbard T., Birney E.* // Nucl. Acids Res. 2008. Vol. 36. P. D5.
85. *Berman H., Westbrook J., Feng Z., Gilliland G., Bhat T., Weissig H., Shindyalov I., Bourne P.* // Nucl. Acids Res. 2000. Vol. 28. P. 235.
86. *Berman H.* // Acta Crystallogr A. 2008. Vol. 64. P. 88.
87. *Berman H., Kleywegt G., Nakamura H., Markley J.* // Biopolymers. 2013. Vol. 99. P. 218.
88. *Gutmanas A., Alhroub Y., Battle G., Berrisford J., Bochet E., Conroy M., Dana J., Montecelo M., van Ginkel G., Gore S., Haslam P., Hatherley R., Hendrickx P., Hirshberg M., Lagerstedt I., Mir S., Mukhopadhyay A., Oldfield T., Patwardhan A., Rinaldi L., Sahni G., Sanz-García E., Sen S., Slowley R., Velankar S., Wainwright M., Kleywegt G.* // Nucl. Acids Res. 2014. Vol. 42. P. D285.
89. *Bairoch A., Apweiler R.* // Nucl. acids research. 1996. Vol. 24. P. 21.
90. *Bairoch A.* // Bioinform. 2000. Vol. 16. P. 48.
91. *Bairoch A, Apweiler R.* // Nucl. Acids Res. 2001. Vol. 28. P. 45.

92. *Boguski M.S., Ostell J., States D.J.* // Protein Engineering: a Practical Approach. Oxford, 1992. P. 57.
93. *Stockinger H., Altenhoff A., Arnold K., Bairoch A., Bastian F., Bergmann S., Bougueleret L., Bucher P., Delorenzi M., Lane L., Le Mercier P., Lisacek F., Michielin O., Palagi P., Rougemont J., Schwede T., von Mering C., van Nimwegen E., Walther D., Xenarios I., Zavolan M., Zdobnov E., Zoete V., Appel R.* // Nucleic Acids Res. 2014. Vol. 42. P. W436.
94. *O'Donovan C., Martin M., Gattiker A., Gasteiger E., Bairoch A., Apweiler R.* // Brief. Bioinform. 2002. Vol. 3. P. 275.
95. *Boeckmann B., Bairoch A., Apweiler R., Blatter M., Estreicher A., Gasteiger E., Martin M., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M.* // Nucl. acids res. 2003. Vol. 31. P. 365.
96. *Apweiler R., Bairoch A., Wu C.* // Curr Opin Chem Biol. 2004. Vol. 8. P. 76.
97. *George D., Barker W., Mewes H., Pfeiffer F., Tsugita A.* // Nucl. Acids Res. 1996. Vol. 24. P. 17.
98. *Barker W., Garavelli J., McGarvey P., Marzec C., Orcutt B., Srinivasarao G., Yeh L., Ledley R., Mewes H., Pfeiffer F., Tsugita A., Wu C.* // Nucl. Acids Res. 1999. Vol. 27. P. 39.
99. *Barker W., Garavelli J., Huang H., McGarvey P., Orcutt B., Srinivasarao G., Xiao C., Yeh L., Ledley R., Janda J., Pfeiffer F., Mewes H., Tsugita A., Wu C.* // Nucl. Acids Res. 2000. Vol. 28. P. 41.
100. *McGarvey P., Huang H., Barker W., Orcutt B., Garavelli J., Srinivasarao G., Yeh L., Xiao C., Wu C.* // Bioinform. 2000. Vol. 16. P. 290.
101. *Barker W., Garavelli J., Hou Z., Huang H., Ledley R., McGarvey P., Mewes H., Orcutt B., Pfeiffer F., Tsugita A., Vinayaka C., Xiao C., Yeh L., Wu C.* // Nucl. Acids Res. 2001. Vol. 29. P. 29.
102. *Wu C., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Ledley R., Lewis K., Mewes H., Orcutt B., Suzek B., Tsugita A., Vinayaka C., Yeh L., Zhang J., Barker W.* // Nucl. Acids Res. 2002. Vol. 30. P. 35.
103. *Wu C., Yeh L., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis P., Ledley R., Suzek B., Vinayaka C., Zhang J., Barker W.* // Nucl. Acids Res. 2003. Vol. 31. P. 345.
104. *Wu C., Nebert D.* // Hum. Genomics. 2004. Vol. 1. P. 229.
105. *Hinz U., UniProt Consortium* // Cell Mol. Life Sci. 2010. Vol. 67. P. 1049.
106. *Wu T., Shamsaddini A., Pan Y., Smith K., Crichton D., Simonyan V., Mazumder R.* // Database (Oxford). 2014. doi: 10.1093/database/bau022.
107. Funding for Global Protein Database Will Create One Reliable Resource [электронный ресурс] // сайт. URL: http://www.genome.gov/page.cfm?pageID=10005283 (дата обращения 10.10.2014 г.).
108. *UniProt Consortium* // Nucl. Acids Res. 2014. Vol. 42. P. D191.
109. *UniProt Consortium* // Nucleic acids research. 2010. Vol. 38. P. D142.
110. UniProtKB/SwissProt release statistics [электронный ресурс] // сайт. URL: http://www.expasy.org/sprot/relnotes/relstat.html (дата обращения 10.10.2014 г.).
111. UniProtKB/TrEMBL release statistics [электронный ресурс] // сайт. URL: http://www.ebi.ac.uk/uniprot/TrEMBLstats/ (дата обращения 10.10.2014 г.).
112. How do we manually annotate a UniProtKB entry электронный ресурс] // сайт. URL: http://www.uniprot.org/faq/45 (дата обращения 10.10.2014 г.).
113. *Apweiler R., Bairoch A., Wu C., Barker W., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M., Natale D., O'Donovan C., Redaschi N., Yeh L.* // Nucl. Acids Res. 2004. Vol. 32. P. D115.
114. Where do UniProtKB sequences come from [электронный ресурс] // сайт. URL: http://www.uniprot.org/faq/37 (дата обращения 10.10.2014 г.).
115. URL: http://www.wwpdb.org/ (дата обращения 10.10.2014 г.).
116. URL: http://www.ncbi.nlm.nih.gov/refseq/ (дата обращения 10.10.2014 г.).
117. URL: http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi (дата обращения 10.10.2014 г.).
118. URL: http://www.embl.org/ (дата обращения 10.10.2014 г.).
119. URL: http://www.ncbi.nlm.nih.gov/genbank/ (дата обращения 10.10.2014 г.).
120. URL: http://www.ddbj.nig.ac.jp/ (дата обращения 10.10.2014 г.).
121. *Leinonen R., Diez F., Binns D., Fleischmann W., Lopez R., Apweiler R.* // Bioinform. 2004. Vol. 20. P. 3236.
122. *Suzek B., Huang H., McGarvey P., Mazumder R., Wu C.* // Bioinform. 2007. Vol. 23. P. 1282.
123. *Li W., Jaroszewski L., Godzik A.* // Bioinform. 2001. Vol. 17. P. 282.
124. URL: http://www.uniprot.org/help/uniref (дата обращения 10.10.2014 г.).
125. UniREF FTP from [электронный ресурс] // сайт. ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/uniref/ (дата обращения 10.10.2014 г.).
126. *Yooseph S., Sutton G., Rusch D. et al.* // Public Library of Science Biology. 2007. Vol. 5. P. 432.
127. *UniProt Consortium* // Nucl. Acids Res. 2008. Vol. 36. P. D190.
128. *Murzin A., Brenner S., Hubbard T., Chothia C.* // J. Mol. Biol. 1995. 247. P. 536.
129. *Lo Conte L., Brenner S., Hubbard T., Chothia C., Murzin A.* // Nucl. Acids Res. 2002. Vol. 30. P. 264.
130. *Hubbard T., Ailey B., Brenner S., Murzin A., Chothia C.* // Nucl. Acids Res. 1999. Vol. 27. P. 254.
131. *Lo Conte L., Ailey B., Hubbard T., Brenner S., Murzin A., Chothia C.* // Nucl. Acids Res. 2000. Vol. 28. P. 257.
132. *Andreeva A., Howorth D., Brenner S., Hubbard T., Chothia C., Murzin A.* // Nucl. Acids Res. 2004. Vol. 32. P. D226.
133. *Andreeva A., Howorth D., Chandonia J., Brenner S., Hubbard T., Chothia C., Murzin A.* // Nucl. Acids Res. 2008. Vol. 36. D419.
134. *Jain P., Hirst J.* // Bioinform. 2010. Vol. 11. P. 364.
135. *Pethica R., Levitt M., Gough J.* // BMC Struct. Biol. 2012. Vol. 12. P. 27.
136. *Fox N., Brenner S., Chandonia J.* // Nucl. Acids Res. 2014. Vol. 42. P. D304.
137. *Asarnow D., Singh R.* // Bioinform. 2014. Vol. 15. P. S1.
138. *Sonnhammer E., Eddy S., Durbin R.* // Proteins. 1997. Vol. 28. P. 405.
139. *Sammut S., Finn R., Bateman A.* // Brief Bioinform. 2008. Vol. 9. P. 210.
140. *Sonnhammer E., Eddy S., Birney E., Bateman A., Durbin R.* // Nucl. Acids Res. 1998. Vol. 26. P. 320.
141. *Bateman A., Birney E., Durbin R., Eddy S., Finn R., Sonnhammer E.* // Nucl. Acids Res. 1999. Vol. 27. P. 260.

142. *Bru C., Courcelle E., Carrère S., Beausse Y., Dalmar S., Kahn D.* // Nucl. Acids Res. 2005. Vol. 33. P. D212.
143. *Bateman A., Coin L., Durbin R., Finn R., Hollich V., Griffiths-Jones S., Khanna A., Marshall M., Moxon S., Sonnhammer E., Studholme D., Yeats C., Eddy S.* // Nucl. Acids Res. 2004. Vol. 32. P. D138.
144. *Finn R., Mistry J., Schuster-Böckler B., Griffiths-Jones S., Hollich V., Lassmann T., Moxon S., Marshall M., Khanna A., Durbin R., Eddy S., Sonnhammer E., Bateman A.* // Nucl. Acids Res. 2006. Vol. 34. P. D247.
145. *Mistry J., Bateman A., Finn R.* // Bioinform. 2007. Vol. 8. P. 298.
146. *Finn R., Mistry J., Tate J., Coggill P., Heger A., Pollington J., Gavin O., Gunasekaran P., Ceric G., Forslund K., Holm L., Sonnhammer E., Eddy S., Bateman A.* // Nucl. Acids Res. 2010. Vol. 38. P. D211.
147. *Finn R., Tate J., Mistry J., Coggill P., Sammut S., Hotz H., Ceric G., Forslund K., Eddy S., Sonnhammer E., Bateman A.* // Nucl. Acids Res. 2008. Vol. 36. P. D281.
148. *Punta M., Coggill P., Eberhardt R., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E., Eddy S., Bateman A., Finn R.* // Nucl. Acids Res. 2012. Vol. 40. P. D290.
149. *Xu Q., Dunbrack R.* // Bioinform. 2012. Vol. 28. P. 2763.
150. *Finn R., Bateman A., Clements J., Coggill P., Eberhardt R., Eddy S., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E., Tate J., Punta M.* // Nucl. Acids Res. 2014. Vol. 42. P. D222.
151. *Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., Sherlock G.* // Nat. Genet. 2000. Vol. 25. P. 25.
152. *du Plessis L., Skunca N., Dessimoz C.* // Brief Bioinform. 2011. Vol. 12. P. 723.
153. *Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., Sherlock G.* // Nat. Genet. 2000. Vol. 25. P. 25.
154. *The Gene Ontology Consortium* // Genome Research. 2001. Vol. 11. P. 1425.
155. *The Gene Ontology Consortium* // Nucl. Acids Res. 2006. Vol. 34. P. D322.
156. *Roncaglia P., Martone M., Hill D., Berardini T., Foulger R., Imam F., Drabkin H., Mungall C., Lomax J.* // J. Biomed Semantics. 2013. Vol. 4. P. 20.
157. *The Gene Ontology Consortium* // Nucl. Acids Res. 2012. Vol. 40. P. D559.
158. The Gene Ontology Consortium // Nucl. Acids Res. 2013. Vol. 41. P. D530.
159. *Balakrishnan R., Harris M., Huntley R., Van Auken K., Cherry J.* // Database (Oxford). 2013. doi: 10.1093/database/bat054.
160. *Huntley R., Harris M., Alam-Faruque Y., Blake J., Carbon S., Dietze H., Dimmer E., Foulger R., Hill D., Khodiyar V., Lock A., Lomax J., Lovering R., Mutowo-Meullenet P., Sawford T., Van Auken K., Wood V., Mungall C.* // Bioinform. 2014. Vol. 15. P. 155.
161. *Sayers E., Barrett T., Benson D., Bryant S., Canese K., Chetvernin V., Church D., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer L., Helmberg W., Kapustin Y., Landsman D., Lipman D., Madden T., Maglott D., Miller V., Mizrachi I., Ostell J., Pruitt K., Schuler G., Sequeira E., Sherry S., Shumway M., Sirotkin K., Souvorov A., Starchenko G., Tatusova T., Wagner L., Yaschenko E., Ye J.* // Nucl. Acids Res. 2009. Vol. 37. P. D5.
162. *Sayers E., Barrett T., Benson D., Bolton E., Bryant S., Canese K., Chetvernin V., Church D., Dicuccio M., Federhen S., Feolo M., Geer L., Helmberg W., Kapustin Y., Landsman D., Lipman D., Lu Z., Madden T., Madej T., Maglott D., Marchler-Bauer A., Miller V., Mizrachi I., Ostell J., Panchenko A., Pruitt K., Schuler G., Sequeira E., Sherry S., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T., Wagner L., Wang Y., Wilbur J., Yaschenko E., Ye J.* // Nucl. Acids Res. 2010. Vol. 38. P. D5.
163. *Sayers E., Barrett T., Benson D., Bolton E., Bryant S., Canese K., Chetvernin V., Church D., DiCuccio M., Federhen S., Feolo M., Fingerman I., Geer L., Helmberg W., Kapustin Y., Landsman D., Lipman D., Lu Z., Madden T., Madej T., Maglott D., Marchler-Bauer A., Miller V., Mizrachi I., Ostell J., Panchenko A., Phan L., Pruitt K., Schuler G., Sequeira E., Sherry S., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T.A., Wagner L., Wang Y., Wilbur W., Yaschenko E., Ye J.* // Nucl. Acids Res. 2011. Vol. 39. P. D38.
164. *Sayers E., Barrett T., Benson D., Bolton E., Bryant S., Canese K., Chetvernin V., Church D., Dicuccio M., Federhen S., Feolo M., Fingerman I., Geer L., Helmberg W., Kapustin Y., Krasnov S., Landsman D., Lipman D., Lu Z., Madden T., Madej T., Maglott D., Marchler-Bauer A., Miller V., Karsch-Mizrachi I., Ostell J., Panchenko A., Phan L., Pruitt K., Schuler G., Sequeira E., Sherry S., Shumway M., Sirotkin K., Slotta D., Souvorov A., Starchenko G., Tatusova T., Wagner L., Wang Y., Wilbur W., Yaschenko E., Ye J.* // Nucl. Acids Res. 2012. Vol. 40. P. D13.
165. *NCBI Resource Coordinators* // Nucl. Acids Res. 2013. Vol. 41. P. D8.
166. *NCBI Resource Coordinators* // Nucl. Acids Res. 2014. Vol. 42. P. D7.
167. *Baxevanis A.* // Nucl. Acids Res. 2000. Vol. 28. P. 1.
168. *Sakurai T., Matsuo T., Matsuda H., Katakuse I.* // Biomed. Mass Spectrom. 1984. Vol. 11. P. 396.
169. *Hamm C., Wilson W., Harvan D.* // Comput. Appl. Biosci. 1986. Vol. 2. P. 115.
170. *Biemann K.* // Biomed. Environ. Mass Spectrom. 1988. Vol. 16. P. 99.
171. *Johnson R., Martin S., Biemann K., Stults J., Watson J.* // Anal. Chem. 1987. Vol. 59. P. 2621.
172. *Ishikawa K., Niva Y.* // Biomed. Environ. Mass Spectrom. 1986. Vol. 13. P. 373.
173. *Siegel M., Bauman N.* // Biomed. Environ. Mass Spectrom. 1988. Vol. 15. P. 333.
174. *Bartels C.* // Biomed. Environ. Mass Spectrom. 1990. Vol. 19. P. 363.
175. *Dancik V., Addona T., Clauser K., Vath J., Pevzner P.* // J. Comput. Biol. 1999. Vol. 6. P. 327.
176. *Taylor A., Johnson R.* // Rapid. Commun. Mass Spectrom. 1997. Vol. 11. P. 1067.
177. *Taylor A., Johnson R.* // Anal. Chem. 2001. Vol. 73. P. 2594.
178. *Fernandez-de-Cossio J., Gonzalez J., Betancourt L., Besada V., Padron G., Shimonishi Y., Takao T.* // Rapid Commun. Mass Spectrom.1998. Vol. 12. P. 1867.
179. *Scigelova M., Maroto F., Dufresne C., Vazquez J.* High-Throughput De Novo Sequencing // Proc. 50th

ASMS Conf. Mass Spectrom. and Allied Topics, Chicago, 2002.
180. *Zhong H., Li L.* // Rapid Commun. Mass Spectrom. 2005. Vol. 19. P. 1084.
181. *Chen T., Kao M., Tepel M., Rush J., Church G.* // J. Comput. Biol. 2001. Vol. 8. P. 325.
182. URL:http://msms.usc.edu/sub (дата обращения 10.10.2014 г.).
183. *Bafna V., Edwards N.* // Bioinform. 2001. Vol. 17. P. S13.
184. *Ma B., Zhang K., Hendrie C. et al.* // Rapid Commun. Mass Spectrom. 2003. Vol. 17. P. 2337.
185. *Grossmann J., Roos F., Cieliebak M., Lipta Z., Mathis L., Muller M., Gruissem W., Baginsky S.* // J. Proteome Res. 2005. Vol. 4. P. 1768.
186. *Fernandez-de-Cossio J., Gonzalez J., Besada V.* // Comput. Appl. Biosci. 1995. Vol. 11. P. 427.
187. *Fischer B., Roth V., Roos F., Grossmann J., Baginsky S., Widmayer P., Grulssem W., Buhmann J.* // Anal. Chem. 2005. Vol. 77. P. 7265.
188. *Певцов С.* // Масс-спектрометрия. 2006. Vol. 3. С. 255.
189. *Frank A., Pevzner P.* // Anal. Chem. 2005. Vol. 77. P. 964.
190. *Savitski M., Nielsen M., Kjeldsen F., Zubarev R.* // J. Proteome Res. 2005. Vol. 4. P. 2348.
191. *Owens K.* // Appl. Spectrosc. Rev. 1992. Vol. 27. P. 1.
192. *Ma B., Zhang K., Liang C.* // J. Comput. Sys. Sci. 2005. Vol. 70. P. 418.
193. Руководство пользователя PEAKS 7.0 [электронный ресурс] // сайт. URL: http://www.bioinfor.com/doc/peaks7/peaks7.pdf (дата обращения 10.10.2014 г.)
194. *Song Y.* // PLoS One. 2014. Vol. 9. P. e87476.
195. *Pan C., Park B., McDonald W., Carey P., Banfield J., VerBerkmoes N., Hettich R., Samatova N.* // Bioinform. 2010. Vol. 11. P. 118.
196. *Zhang S., Wang Y., Bu D., Zhang H., Sun S.* // Bioinform. 2011. Vol. 12. P. 346.
197. *Pevtsov S., Fedulova I., Mirzaei H., Buck C., Zhang X.* // J. Proteome Res. 2006. Vol. 5. P. 3018.

Поступила в редакцию 10.09.14

METHODS OF PROCESSING MASS SPECTROMETRY DATA FOR IDENTIFICATION OF THE PEPTIDES AND PROTEINS

E.I. Berizovskaya[1], A.A. Ichalaynen[1], A.M. Antochin[1], V.F. Taranchenko[1], V.M. Goncharov[1], D.A. Mitrofanov[1], A.V. Udintsev[1], A.V. Aksenov[1], O.A. Shevlyakova[1], I.A. Rodin[2], O.A. Shpigun[2]

([1]Federal State Unitary Enterprise Scientific Center "Signal"; [2]Chemistry Department, Lomonosov Moscow State University)

One of the most common method for peptide and protein identification is mass spectrometry. First step is an enzymatic digestion, then mass spectrometric data are received. Identification of this compounds is realized by using differents algorithms and programs for database search or de novo sequencing. The most popular software products and databases are briefly reviewed in the article.

Key words: peptides, protein, mass spectrometry, enzymatic digestion, database, *de novo* sequencing, programs.

Сведения об авторах: *Беризовская Елена Игоревна* – науч. сотр. ФГУП «НЦ «Сигнал» (eiberizovskaya@rambler.ru); *Ихалайнен Андрей Александрович* – вед. науч. сотр. ФГУП «НЦ «Сигнал», докт. биол. наук, доцент (an12321na@gmail.com); *Антохин Андрей Михайлович* – зам. директора ФГУП «НЦ «Сигнал» по научной работе, канд. техн. наук, доцент (antohin_08@mail.ru); *Таранченко Виктор Федорович* – начальник отдела ФГУП «НЦ «Сигнал», канд. хим. наук, доцент (victaran@rambler.ru); *Гончаров Валерий Михайлович* – вед. науч. сотр. ФГУП «НЦ «Сигнал», канд. хим. наук, доцент (GVM52005@ya.ru); *Митрофанов Дмитрий Александрович* – вед. науч. сотр. ФГУП «НЦ «Сигнал», канд. хим. наук, доцент (lab-tex@mail.ru); *Удинцев Андрей Валерьевич* – вед. науч. сотр. ФГУП «НЦ «Сигнал», канд. мед. наук (udintsevav@mail.ru); *Аксенов Алексей Вадимович* – зав. лаб. ФГУП «НЦ «Сигнал», канд. техн. наук, доцент (aksenov_av@list.ru); *Шевлякова Олеся Александровна* – науч. сотр. ФГУП НЦ «Сигнал» (olesya.shevlyakova@gmail.com); *Родин Игорь Александрович* – ст. науч. сотр. кафедры аналитической химии химического факультета МГУ, канд. хим. наук (igorrodin@ya.ru); *Шпигун Олег Алексеевич* – профессор кафедры аналитической химии химического факультета МГУ, чл.-корр. РАН, докт. хим. наук (shpigun@analyt.chem.msu.ru).