

ОБЗОР

DE NOVO СЕКВЕНИРОВАНИЕ БЕЛКОВ И ПЕПТИДОВ: АЛГОРИТМЫ, ПРИЛОЖЕНИЯ, ПЕРСПЕКТИВЫ

К.В. Вяткина^{1,2}¹Санкт-Петербургский национальный исследовательский Академический университет Российской Академии Наук; 194021, Санкт-Петербург, ул. Хлопина, д. 8, к. 3; эл. почта: vyatkina@spbau.ru²Санкт-Петербургский государственный университет; 199034, Санкт-Петербург, Университетская наб., д. 7-9

Установление первичной структуры белков и пептидов является важным этапом при изучении их свойств. В настоящее время для решения данной задачи наиболее часто используется масс-спектрометрия. Результаты масс-спектрометрических измерений могут быть интерпретированы посредством поиска в базе данных или методами *de novo* секвенирования. Привлекательность последних обусловлена возможностью их применения для исследования неизвестных белков, а также тех, которые не могут быть проанализированы методами геномики или транскриптомики. В данной статье предлагается краткий обзор существующих подходов к *de novo* секвенированию белков и пептидов и решаемых с их помощью задач, а в завершение обозначены направления и перспективы их дальнейшего развития.

Ключевые слова: протеомика; масс-спектрометрия; *de novo* секвенирование; белки; пептиды; аминокислотная последовательность

DOI: 10.18097/BMCRM00005

ВВЕДЕНИЕ

Необходимость в исследовании структуры белков и пептидов возникает при решении самых разнообразных задач современной биологии и медицины. Ключевым его этапом является установление их аминокислотной последовательности, или секвенирование. Классическим методом, предложенным для этой цели, стала деградация по Эдману [1, 2], суть которой заключается в циклически повторяющемся отщеплении меченого N-концевого аминокислотного остатка и его идентификации при помощи хроматографии. К основным факторам, ограничивающим её применимость, относятся высокая стоимость реагентов и низкая скорость анализа. В частности, по этим причинам в настоящее время предпочтение нередко отдаётся более быстрому и менее дорогостоящему масс-спектрометрическому секвенированию.

Установление первичной структуры белков или пептидов с использованием масс-спектрометрического метода сопровождается либо идентификацией масс-спектров посредством поиска в базе данных, либо их интерпретацией *de novo*. В первом случае необходимо наличие базы данных, предположительно содержащей аминокислотные последовательности изучаемых белков или пептидов. Второй же способ позволяет анализировать ранее неизвестные белки и пептиды, а также предоставляет дополнительные возможности для исследования аминокислотных последовательностей, содержащих неизвестные или многочисленные модификации.

Алгоритмы идентификации масс-спектров путём поиска в базе данных разрабатываются на протяжении более двух десятилетий. В девяностых годах прошлого века были предложены, в частности, первые версии алгоритмов, ставших основой для коммерческих программных систем SEQUEST [3] и Mascot [4], широко используемых

для идентификации пептидов и по сей день. К популярным программам, решающим ту же задачу и находящимся в свободном доступе, относятся, например, MS-GF+ [5, 6] и Andromeda [7]. В течение последнего десятилетия были также разработаны и реализованы в программных инструментах подходы к идентификации белков по данным масс-спектрометрии “сверху вниз” (top-down), анализирующие белковые молекулы целиком; наиболее широкое распространение получила коммерческая реализация ProSightPC, сопровождаемая бесплатной Интернет-версией ProSight PTM [8, 9], а также свободно распространяемые MS-Align+ [10] и TopPIC [11].

Следует отметить, что процесс идентификации масс-спектров по базе данных нередко включает в себя элементы *de novo* секвенирования: для ускорения поиска на предшествующем ему этапе может применяться фильтрация базы данных с использованием *тегов пептидных последовательностей* (peptide sequence tags) – коротких последовательностей аминокислот, представляющих собой результаты локальной интерпретации обрабатываемых масс-спектров. При этом в рассмотрении остаются лишь те аминокислотные последовательности из базы, которые согласуются с найденными тегами. Данный подход был впервые предложен в 1994 г. М. Манном и М. Вилмом [12], а впоследствии были опубликованы десятки статей, в которых обсуждались эффективные способы генерации тегов и их использования в качестве фильтров (см., например, [13-25] и [8]).

Параллельно с методами идентификации по базе данных разрабатывались и алгоритмы *de novo* секвенирования. Первым алгоритмом, решающим данную задачу, стал Lutefisk97 [26], опубликованный в 1997 г. В основе этого метода, как и многих последующих, лежит концепция спектрального графа (spectrum graph), предложенная в 1990 г. Бартельсом [27]: такой граф сопоставляется



масс-спектру, причем его вершины порождаются пиками, а рёбра соединяют пары вершин, “отличающихся” друг от друга на массу остатка какой-либо аминокислоты (рис. 1). Каждое ребро спектрального графа помещено соответствующей аминокислотой (с точностью до замены I/L), а любой путь в этом графе определяет аминокислотную последовательность, образованную метками составляющих его рёбер. Таким образом, полная или частичная интерпретация масс-спектра сводится к нахождению в его спектральном графе оптимального или нескольких лучших (с точки зрения используемой оценочной функции) путей.

На сегодняшний день наиболее мощной и часто используемой коммерческой программой, несомненно, является PEAKS [28]; из бесплатных программных инструментов уже более десяти лет пользуется популярностью PepNovo [29]. К недавним разработкам относится метод Twister, изначально предназначенный для *de novo* секвенирования пептидов по наборам tandemных масс-спектров “сверху вниз” (top-down) [30-32], а впоследствии адаптированный к случаю данных “снизу вверх” (bottom-up) высокого разрешения [33], и Novor [34], позволяющий обрабатывать масс-спектры триптических пептидов.

Идентификация масс-спектров путем поиска в базе данных традиционно считается более надежным методом определения аминокислотной последовательности, нежели *de novo* секвенирование. Действительно, количество потенциально возможных интерпретаций масс-спектра, которые могут быть получены из базы данных, заведомо окажется существенно меньше числа всех возможных его интерпретаций*, а, следовательно, значительно сократится и число неверных интерпретаций, что, в свою очередь, должно уменьшить риск ошибки при попытке выбрать единственный правильный вариант. Ещё одна причина заключается в недостатке методов контроля качества результатов *de novo* секвенирования, сопоставимых с методами оценки уровня ложноположительных результатов (False Discovery Rate, FDR), используемых при поиске в базе

данных [35]. Однако технологические достижения и алгоритмические разработки последних лет обеспечили повышение надежности методов *de novo* секвенирования, что открывает новые перспективы для их применения и позволяет эффективнее использовать их в тех случаях, когда идентификация по базе данных не представляется возможной.

Физико-химические основы масс-спектрометрического метода анализа белков и пептидов детально изложены, например, в обзоре [36] и монографиях [37, 38]. В данной же статье будет приведен краткий обзор существующих на сегодняшний день алгоритмов *de novo* секвенирования, а также задач, при решении которых они могут быть успешно использованы. В заключение будут обозначены дальнейшие пути их совершенствования и некоторые дополнительные области их применения.

1. АЛГОРИТМЫ

За последние двадцать лет были предложены десятки алгоритмических подходов к решению задачи *de novo* секвенирования белков и пептидов. Подавляющее их большинство было разработано для установления первичной структуры пептидов по данным масс-спектрометрии “снизу вверх” (bottom-up). К тем из них, что оказались наиболее привлекательными для специалистов, могут быть отнесены Lutefisk [26, 39], SHERENGA [40], PEAKS [28], PepNovo [29] и усовершенствованная его версия PepNovo+ [41-43], NovoHMM [44], Vonode [23], pNovo [45], UniNovo [46] и Novor [34].

PEAKS представляет собой коммерческую программную систему с дорогостоящей пользовательской лицензией; прочие перечисленные выше алгоритмы реализованы в программных продуктах свободного доступа. Суть метода, лежащего в основе PEAKS, заключается в генерации для каждого масс-спектра большого числа аминокислотных последовательностей – потенциальных его интерпретаций (в соответствии

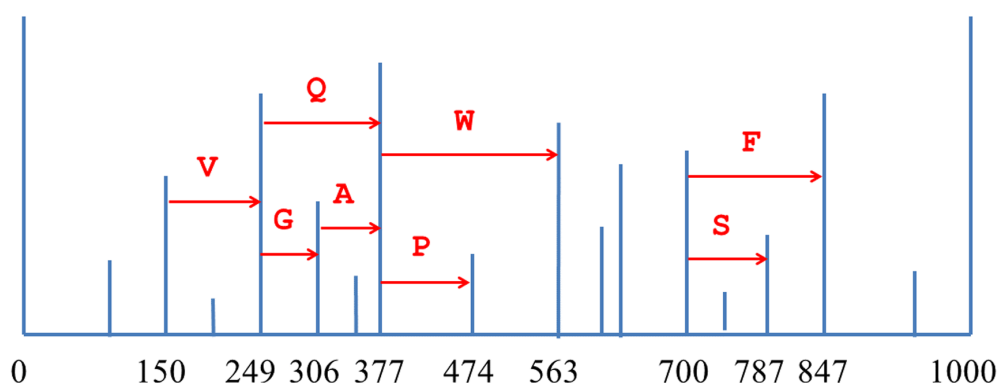


Рисунок 1. Спектральный граф для “игрушечного” MS/MS-спектра с нейтральными моноизотопными массами. Его вершины соответствуют пикам, а ребра определяются парами вершин, отстоящих друг от друга на массу остатка некоторой аминокислоты. Каждое ребро направлено в сторону вершины с большей массой.

* Поясним данное утверждение на «игрушечном» примере. Пусть база данных состоит из трёх последовательностей AN, TI и GEN массой 203, 232 и 318 Да. Соответственно. Допустим, что масса родительского иона для обрабатываемого масс-спектра составляет 203 Да. Тогда на основе базы данных для него может быть предложена лишь одна потенциально возможная интерпретация – AN, в то время как при *de novo* секвенировании *a priori* их будет пять: AN, NA, AGG, GAG и GGA. Какие именно из них будут проанализированы явно, и как будет выглядеть окончательный ответ, определяется деталями алгоритма и используемой функцией оценки надежности решения.

с исходной версией подхода их количество составляло 10000), и последующем выборе той из них, что наилучшим образом объясняет данный масс-спектр. NovoHMM и Novog опираются на методы машинного обучения и используют, соответственно, скрытые марковские модели (Hidden Markov Models, HMMs) и деревья решений (decision trees). В рамках остальных шести подходов нашла применение концепция спектрального графа [27].

Каждый из этих методов *de novo* секвенирования включает в себя функцию оценки надежности предлагаемой интерпретации масс-спектра. При определении такой функции или же при вычислении значений её параметров явно или неявно принимаются во внимание свойства масс-спектрометрических данных, для обработки которых, в первую очередь, предназначен соответствующий алгоритм, что необходимо учитывать при выборе метода анализа данных конкретного эксперимента (рис. 2).

Так, ранние подходы используют свойства фрагментации молекулярных ионов при помощи диссоциации пептидных связей, индуцированной соударениями (ДИС; collision-induced dissociation, CID), или диссоциации, активированной соударениями (ДАС; collision activated dissociation, CAD), в то время как pNovo был разработан для анализа масс-спектров, полученных с использованием ДАС при повышенной энергии (ДАСПЭ; higher-energy C-trap dissociation, HCD) [47], предложенной в 2007 г. Однако более поздние версии алгоритмов интерпретации ДИС- и ДАС-МС/МС спектров, как правило, позволяют получать приемлемые результаты и для ДАСПЭ-МС/МС спектров. Последние версии PEAKS, наряду с ДИС- и ДАС-МС/МС спектрами, позволяют анализировать масс-спектры, полученные с помощью диссоциации при переносе электрона (ДПЭ; electron transfer dissociation, ETD) [48] или диссоциации при захвате электрона

(ДЗЭ; electron capture dissociation, ECD) [49], а также комбинированного метода активации диссоциации EThcD (electron-transfer/higher-energy collision dissociation) [50] и ультрафиолетовой фотодиссоциации (УФФД; ultraviolet photodissociation, UVPD) [51], однако детали соответствующих алгоритмов не разглашаются. Прочие перечисленные выше алгоритмы не предназначены для обработки ДПЭ/ДЗЭ-, EThcD- или УФФД-МС/МС спектров. Отдельного упоминания в данном контексте заслуживает недавно предложенный подход UVNovo [52] к *de novo* секвенированию меченых хромофором триптических пептидов по наборам УФФД-МС/МС спектров; алгоритмическая его составляющая базируется на машинном обучении и включает в себя использование метода случайного леса и скрытых марковских моделей.

При выборе подходящего метода *de novo* секвенирования существенное значение имеет и то, с низким или высоким разрешением были сняты МС/МС-спектры. Из обсуждаемых девяти алгоритмов лишь Vnode был изначально предназначен для обработки МС/МС-спектров высокого разрешения. Кроме того, к этому случаю адаптирована версия PepNovo+ алгоритма PepNovo, а также достаточно поздние версии программной системы PEAKS. В то же время, качество результатов применения к таким данным алгоритмов, ориентированных на обработку масс-спектров низкого разрешения, порой оказывается абсолютно неприемлемым. Так, в [33] приведены крайне бедные результаты обработки наборов масс-спектров, снятых с высоким разрешением на приборах Q-Exactive plus MS и Q-Exactive Orbitrap HF (“Thermo Fisher Scientific”, Бремен, Германия) при помощи метода Novog, способного весьма эффективно обрабатывать МС/МС-спектры низкого разрешения в режиме реального времени. И в противоположной ситуации, очевидно, не будет оснований рассчитывать на хорошие результаты.

Наконец, следует иметь в виду, что практически все методы *de novo* секвенирования пептидов рассчитаны, прежде всего, на случай *триптических* пептидов. Из упомянутых выше программ лишь PEAKS предоставляет возможность явного указания других ферментов, использованных для гидролиза исследуемых белков, при выборе параметров эксперимента**. Правда, многие алгоритмы предоставляют опции работы с “полутриптическими” или “нетриптическими” пептидами – однако при их выборе результаты интерпретации нередко теряют в аккуратности и надежности.

Алгоритмы *de novo* секвенирования, реализованные в программах pNovo+ [53], ADEPTS [54] и UniNovo [46], а также методы, изложенные в работах [55-60], используют информацию из масс-спектров, снятых с одного и того же пептида с использованием различных методов инициирования фрагментации молекулярных ионов, что позволяет

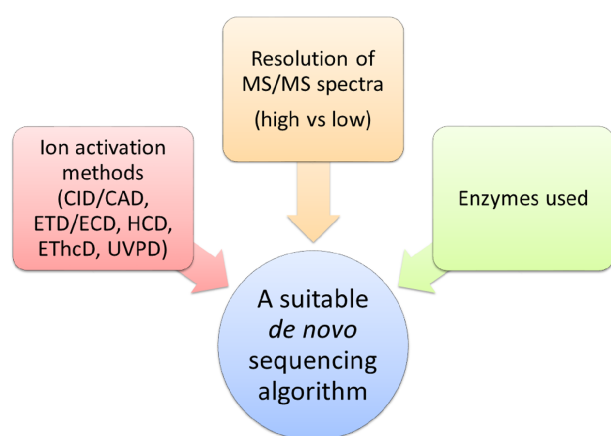


Рисунок 2. Условия пробоподготовки и масс-спектрометрического эксперимента, которые необходимо принимать во внимание при выборе подходящего алгоритма *de novo* секвенирования.

** Заметим, что скриншоты, иллюстрирующие возможности сервиса по *de novo* секвенированию белков, предоставляемого компанией Rapid Novor – разработчиком алгоритма Novog, доступные на сайте <https://www.rapidnovor.com/demo/coverageview>, содержат результаты для пептидов, полученных гидролизом шестью различными ферментами (трипсином, пепсином, химотрипсином, GluC, AspN и протеиназой K). Очевидно, закрытая версия Novog обладает соответствующей функциональностью, однако последняя его версия (1.05), находящаяся в свободном доступе, по-прежнему позволяет указать в качестве использованного фермента лишь трипсин.

существенно улучшить покрытие последовательности, а, следовательно, и качество результатов. (Следует отметить, что UniNovo позволяет обрабатывать и наборы масс-спектров, полученные с использованием лишь одного из методов ДИС, ДАСПЭ или ДПЭ.)

Также был разработан ряд подходов к *de novo* секвенированию белков по данным масс-спектрометрии “снизу вверх” для перекрывающихся пептидов, полученных гидролизом с применением нескольких ферментов различной специфичности или одного фермента без выраженной специфичности. Их суть заключается либо в выравнивании и объединении масс-спектров, предположительно относящихся к перекрывающимся пептидам, с последующим секвенированием полученного “суперспектра”, либо в объединении результатов *de novo* секвенирования отдельных масс-спектров. Методы первого типа на этапе выравнивания масс-спектров могут не использовать [61, 62] или использовать [63, 64] гомологичные белковые последовательности. Алгоритм Champs, изложенный в [65], применяет PEAKS для интерпретации исходных масс-спектров и нахождения в базе данных последовательности белка, гомологичного исследуемому, выравнивает по отношению к этой последовательности полученные *de novo* теги, а затем предсказывает аминокислотную последовательность целевого белка путём её уточнения.

Наконец, повышение надежности результатов *de novo* секвенирования может быть обеспечено за счёт сопоставления и комбинирования результатов, сгенерированных различными алгоритмами [66].

В завершение обсуждения подходов к *de novo* секвенированию пептидов следует упомянуть недавно предложенный разработчиками pNovo и pNovo+ алгоритм Open-pNovo [67], позволяющий учитывать потенциальное присутствие в последовательности любых из посттрансляционных модификаций, включенных в базу данных UniProt [68], а также выполненный в статье [69] анализ эффективности ряда методов в применении к смешанным (mixed) MC/MC-спектрам.

Первый метод определения аминокислотной последовательности белка по данным масс-спектрометрии “сверху вниз” (top-down) был изложен в работе [70]; входными данными для него служат пары ДАС- и ДЗЭ-MC/MC спектров. В 2014 г. был предложен комбинированный метод TVNovo для *de novo* секвенирования белков по наборам масс-спектров “сверху вниз” и “снизу вверх” [71], в соответствии с которым вначале выполнялось *de novo* секвенирование пептидов по данным “снизу вверх” с применением программы PEAKS, а затем полученные пептидные последовательности объединялись в белковую с использованием “каркаса” (scaffold), сформированного на основе данных “сверху вниз”. Первым же алгоритмом *de novo* секвенирования белков лишь по масс-спектрам “сверху вниз”, реализованным в виде свободно распространяемого программного инструмента, стал Twister [30-32]. Основная его идея заключается в том, чтобы дать отдельным масс-спектрам частичную, но максимально

надёжную интерпретацию, а затем скомбинировать полученные таким образом теги пептидных последовательностей. Впоследствии этот подход был адаптирован к случаю данных масс-спектрометрии “снизу вверх” высокого разрешения [33]. Для того чтобы обеспечить возможность его применения к масс-спектрам “снизу вверх”, необходимо предварительно их обработать при помощи метода деконволюции (удаления изотопных пиков и перехода от отношения массы к заряду к нейтральным массам), предназначенного для данных “сверху вниз”. Результаты анализа, приведённого в [33], свидетельствуют о том, что предпочтительным является использование алгоритма MS-Deconv [72]. После этого в большей части MC/MC-спектров “снизу вверх” пиков останется совсем немного (не более 10-20) или не останется вовсе. Как следствие, индивидуальная интерпретация большинства из них окажется невозможной. Однако оставшиеся пики будут хорошо подтверждены и позволят сгенерировать аккуратные теги пептидных последовательностей, за счёт комбинирования которых далее могут быть получены длинные фрагменты последовательностей анализируемых пептидов, а порой и полные их последовательности.

Результаты предварительных экспериментов подтверждают возможность использования метода Twister и для обработки данных масс-спектрометрии “с середины вниз” (middle-down) высокого разрешения: в данном случае, речь идет о *de novo* секвенировании длинных пептидов, полученных путём ферментативного или химического гидролиза. Twister позволяет обрабатывать ДИС/ДАС-, ДПЭ/ДЗЭ- и ДАСПЭ-MC/MC спектры и может быть также применен к EThcD-MC/MC спектрам. При обработке масс-спектров “снизу вверх” и “с середины вниз” Twister не использует никаких предположений относительно того, какие ферменты или химические соединения использовались для гидролиза, что делает его более универсальным по сравнению с другими методами.

2. ПРИЛОЖЕНИЯ

Алгоритмы *de novo* секвенирования востребованы, прежде всего, при изучении белков, которые не могут быть проанализированы на уровне генома или транскриптома, но они также находят применение и при исследовании сложных белковых смесей (рис. 3).

Наиболее распространенным применением методов *de novo* секвенирования белков и пептидов, несомненно, является их использование для установления первичной структуры антител. Необходимость в этом возникает, в частности, при разработке лекарственных препаратов на основе моноклональных антител, которые, согласно оценке экспертов, в последние годы занимают наибольшую долю фармацевтического рынка среди биопрепаратов (см., например, [73]). Поэтому неудивительно, что многие алгоритмы были либо целенаправленно разработаны для решения этой задачи [63, 64, 74], либо протестированы на наборах масс-спектрометрических

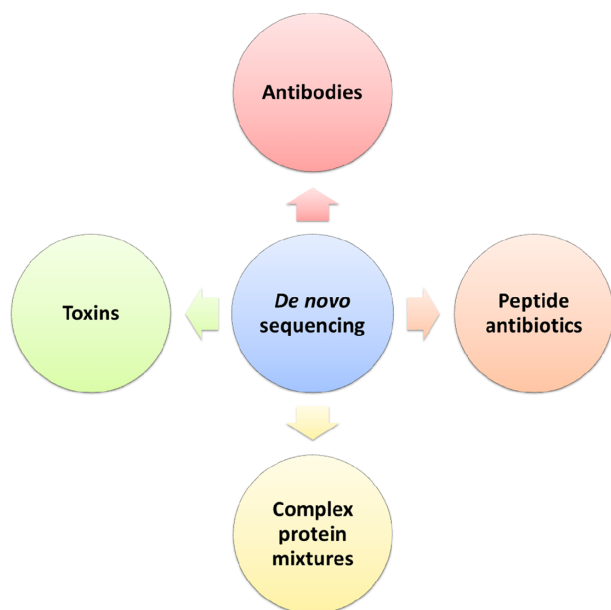


Рисунок 3. Белковые субстанции, при анализе которых применяются методы *de novo* секвенирования.

данных для антител [30-32, 68], а компания Bioinformatics Solutions – разработчик системы PEAKS – летом 2017 г. выпустила на рынок программный пакет PEAKS AB [74], предназначенный для *de novo* секвенирования антител и анализа их аминокислотных последовательностей. Все эти алгоритмы предназначены для анализа либо отдельных моноклональных антител, либо простых их смесей, называемых “коктейлями” (antibody cocktails), однако недавно была предпринята попытка разработать алгоритм, позволяющий анализировать поликлональные антитела [75].

Ещё одним примером белков, при изучении которых возникает потребность в использовании методов *de novo* секвенирования, являются токсины. Понимание их структуры и функций также необходимо для разработки ряда лекарственных средств, включая противоядия. Полноценное исследование токсинов методами геномики или транскриптомики не представляется возможным, во-первых, в силу недостатка необходимой для этого информации (в частности, единственной змеёй, для которой полностью секвенирован геном и транскриптом ядовитой железы, остаётся королевская кобра [76, 77]), а во-вторых, ввиду их подверженности посттрансляционным модификациям, оказывающим существенное влияние на их биологическую активность. В работе [62] приведены результаты анализа с применением изложенного в ней метода “скорострельного секвенирования белка” (shotgun protein sequencing) состава яда техасского гремучника (*Crotalus atrox*), а в статье [78] предложен метод *de novo* секвенирования “с ограничениями”, предназначенный для анализа токсинов улиток-конусов (результаты исследования приведены для видов *Conus textile* и *Conus stercusmuscarum*). Ограничения формулируются на основе априорных знаний о последовательности

токсина: это может быть, например, количество входящих в её состав аминокислотных остатков цистеина или какие-либо “мотивы” (motifs), которые должны в ней присутствовать; введение их в рассмотрение позволяет существенно уменьшить число последовательностей-кандидатов, из которых выбирается окончательный ответ.

Не менее интересным объектом исследований методами *de novo* секвенирования являются пептиды из кожного секрета амфибий, обладающие антимикробными, противоопухолевыми, противогрибковыми и другими свойствами [79] – в частности, кожные пептиды-антибиотики ранидных лягушек, различные виды которых широко представлены в России и Европе [80, 81]. Структурной их особенностью является наличие С-концевого цикла, который возникает за счёт образования дисульфидной связи между двумя цистеиновыми остатками; он носит название Rana box [81]. Для определения аминокислотной последовательности внутри такого цикла его необходимо предварительно раскрыть. Полученный в результате пептид будет иметь линейную структуру, однако, очевидно, не будет походить на триптический, что существенно снизит эффективность применения для его анализа практически всех существующих алгоритмов *de novo* секвенирования (напомним, что из обсуждавшихся выше методов возможности для анализа нетриптических пептидов предоставляют лишь поздние версии PEAKS и недавно предложенный алгоритм Twister, а также потенциально Novor). Как следствие, ранее для установления первичной структуры таких пептидов либо применялась деградация по Эдману [82], либо *de novo* секвенирование выполнялось вручную [80].

Кроме того, алгоритмы *de novo* секвенирования белков и пептидов с успехом применялись для решения самых разнообразных специализированных задач – например, при изучении абиогенных пептидов [83], микробных сообществ [84], нейропептидома морского ежа [85], особенностей адаптации личинок к зимовке [86], а также в контексте других исследований.

3. ПЕРСПЕКТИВЫ

К основным направлениям дальнейшего совершенствования подходов к *de novo* секвенированию белков и пептидов с целью расширения их применимости на практике относится повышение аккуратности выдаваемых ими результатов, а также разработка универсальных методов оценки их надёжности. Для достижения первой цели при улучшении существующих и разработке новых алгоритмов следует в максимально возможной степени использовать современные достижения масс-спектрометрии высокого разрешения и учитывать особенности получаемых с её помощью данных. В частности, представляет интерес разработка специальных алгоритмов для интерпретации весьма информативных, но сложных с точки зрения обработки УФФД- и EThcD-MC/MC спектров. Кроме того, найдут применение алгоритмы

de novo секвенирования пептидов, полученных гидролизом ферментами без выраженной специфичности к конкретным аминокислотным остаткам (например, секретируемой аспарагиновой протеазой 9 [87, 88]), а также эндогенных пептидов, включая пептиды-антибиотики амфибий [80, 81].

Для оценки надежности результатов идентификации белков или пептидов посредством поиска в базе данных, как правило, оценивается уровень ложноположительных результатов (FDR), определяемый как отношение числа ложноположительных идентификаций к общему их количеству. Наиболее распространенной является следующая схема: на основе “целевой” (target) базы данных путем реверсирования каждой из входящих в неё последовательностей генерируется “ложная” (decoy), а затем выполняется поиск в объединенной базе данных [35]. При этом предполагается, что число ложноположительных идентификаций в целевой базе данных совпадет с числом (заведомо ошибочных) идентификаций в ложной. Далее критерии, на основе которых предложенные идентификации разделяются на “принятые” и “отклонённые”, могут быть выбраны таким образом, чтобы обеспечить желаемый уровень FDR (например, 1%). В англоязычной литературе данный подход носит название “target-decoy approach”.

Очевидно, этот метод не может быть непосредственно применён в случае *de novo* секвенирования. Однако в недавних работах [66, 89] была продемонстрирована возможность получения аккуратной оценки FDR для результатов *de novo* секвенирования с использованием результатов идентификации в базе данных. Также представляет интерес опубликованный в январе 2018 г. метод оценки уровня “ложных аминокислот” (false amino-acid rate, FAR) [90], определяемого как отношение числа неправильно предсказанных аминокислот к общему количеству аминокислот в последовательностях, сгенерированных алгоритмом *de novo* секвенирования. В то же время не вызывает сомнений, что исследования в данном направлении будут продолжаться.

ЗАКЛЮЧЕНИЕ

Методы *de novo* секвенирования представляют собой незаменимый инструмент анализа белков из организмов, геном которых неизвестен, а также тех белков, которые напрямую не закодированы в геноме – например, антител. Достижения современной масс-спектрометрии высокого разрешения открывают новые возможности для их совершенствования, а, вместе с тем, расширяется и круг задач, для решения которых они могут быть с успехом использованы. Неудивительно, что в ведущих научных журналах регулярно появляются новые публикации, посвящённые разработке и применению алгоритмов *de novo* секвенирования, всё более востребованных как при изучении отдельных белков и пептидов, так и в комплексных протеогеномных исследованиях.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант №16-54-21006).

ЛИТЕРАТУРА

1. Edman P. (1949) A method for the determination of amino acid sequence in peptides. Arch. Biochem., 22(3):475-476.
2. Edman P. (1950) Method for determination of the amino acid sequence in peptides. Acta Chem. Scand., 4:283-293.
3. Eng J.K., McCormack A.L., Yates J.R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom., 5(11):976-989. DOI: 10.1016/1044-0305(94)80016-2
4. Perkins D.N., Pappin D.J.C., Creasy D.M., Cottrell J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18):3551-3567. DOI: 10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2
5. Kim S., Gupta N., Pevzner P.A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J. Proteome Res., 7(8):3354-3363. DOI: 10.1021/pr8001244
6. Kim S., Pevzner P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun., 5:5277. DOI: 10.1038/ncomms6277
7. Cox J., Neuhauser N., Michalski A., Scheltema R.A., Olsen J.V., Mann M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. J. Proteome Res., 10(4):1794-1805. DOI: 10.1021/pr101065j
8. LeDuc R.D., Taylor G.K., Kim Y.B., Janusz T.E., Bynum L.H., Sola J.V., Garavelli J.S., Kelleher N.L. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res., 32(Web Server issue):W340-W345. DOI: 10.1093/nar/gkh447
9. Zamdborg L., LeDuc R.D., Glowacz K.J., Kim Y.B., Viswanathan V., Spaulding I.T., Early B.P., Bluhm E.J., Babai S., Kelleher N.L. (2007) ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res., 35(Web Server issue):W701-W706. DOI: 10.1093/nar/gkm371
10. Liu X., Sirotkin Y., Shen Y., Anderson G., Tsai Y.S., Ting Y.S., Goodlett D.R., Smith R.D., Bafna V., Pevzner P.A. (2012) Protein identification using top-down spectra. Mol. Cell Proteomics, 11(6):M111.008524. DOI: 10.1074/mcp.M111.008524
11. Kou Q., Xun L., Liu X. (2016) TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics, 22(22):3495-3497. DOI: 10.1093/bioinformatics/btw398
12. Mann M., Wilm M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem., 66(24):4390-4399. DOI: 10.1021/ac00096a002
13. Taylor J.A., Johnson R.S. (2011) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. Anal. Chem., 73(11):2594-2604. DOI: 10.1021/ac001196o
14. Tabb D.L., Saraf A., Yates J.R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem., 75(23):6415-6421. DOI: 10.1021/ac0347462

15. Sunyaev S., Liska A.J., Golod A., Shevchenko A., Shevchenko A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.*, 75(6):1307-1315. DOI: 10.1021/ac026199a
16. Searle B.C., Dasari S., Turner M., Reddy A.P., Choi D., Wilmarth P.A., McCormack A.L., David L.L., Nagalla S.R. (2004) High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results *Anal. Chem.*, 76(8):2220-2230. DOI: 10.1021/ac035258x
17. Savitski M.M., Nielsen M.L., Zubarev R.A. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol. Cell. Proteomics*, 4(8):1180-1188. DOI: 10.1074/mcp.T500009-MCP200
18. Frank A., Tanner S., Bafna V., Pevzner P. (2005) Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.*, 4(4):1287-1295. DOI: 10.1021/pr050011x
19. Cao X., Nesvizhskii A.I. (2008) Improved sequence tag generation method for peptide identification in tandem mass spectrometry. *J. Proteome Res.*, 7(10):4422-4434. DOI: 10.1021/pr800400q
20. Na S., Jeong J., Park H., Lee K.J., Paek E. (2008) Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. *Mol. Cell. Proteomics*, 7(12):2452-2463. DOI: 10.1074/mcp.M800101-MCP200
21. Shen Y., Tolic N., Hixson K.K., Purvine S.O., Anderson G.A., Smith R.D. (2008) *De novo* sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Anal. Chem.*, 80(20):7742-7754. DOI: 10.1021/ac801123p
22. Tabb D.L., Ma Z.-Q., Martin D.B., Ham A.-J.L., Chambers M.C. (2008) DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.*, 7(9):3838-3846. DOI: 10.1021/pr800154p
23. Pan C., Park B.H., McDonald W.H., Carey P.A., Banfield J.F., VerBerkmoes N.C., Hettich R.L., Samatova N.F. (2010) A high-throughput *de novo* sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics*, 11:118. DOI: 10.1186/1471-2105-11-118
24. Liu W.T., Kersten R.D., Yang Y.L., Moore B.S., Dorrestein P.C. (2011) Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective agent arylomycin in *Streptomyces roseosporus*. *J. Am. Chem. Soc.*, 133(45):18010-18013. DOI: 10.1021/ja2040877
25. Kersten R.D., Yang Y.L., Xu Y., Cimermancic P., Nam S.J., Fenical W., Fischbach M.A., Moore B.S., Dorrestein P.C. (2011) Natural product peptidogenomics: A mass spectrometry-guided genome mining approach. *Nat. Chem. Biol.* 7(11):794-802. DOI: 10.1038/nchembio.684
26. Taylor J.A., Johnson R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 11(9):1067-75. DOI: 10.1002/(SICI)1097-0231(19970615)11:9<1067::AID-RCM953>3.0.CO;2-L
27. Bartels C. (1990) Fast algorithm for peptide sequencing by mass spectroscopy. *Biol. Mass Spectrom.*, 19:363-368. DOI: 10.1002/bms.1200190607
28. Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 17(20):2337-2342. DOI: 10.1002/rcm.1196
29. Frank A., Pevzner P. (2005) PepNovo: *De novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 77(4):964-73. DOI: 10.1021/ac048788h
30. Vyatkina K., Wu S., Dekker L.J.M., VanDuijn M.M., Liu X., Tolic N., Dvorkin M., Alexandrova S., Luider T.M., Pasa-Tolic L., Pevzner P. A. (2015) *De novo* sequencing of peptides from top-down tandem mass spectra. *J. Proteome Res.* 14(11):4450-62. DOI: 10.1021/pr501244v
31. Vyatkina K., Wu S., Dekker L.J.M., VanDuijn M.M., Liu X., Tolic N., Luider T.M., Pasa-Tolic L., Pevzner P.A. (2016) Top-down analysis of protein samples by *de novo* sequencing techniques. *Bioinformatics*, 32(18):2753-2759. DOI: 10.1093/bioinformatics/btw307
32. Vyatkina K. (2017) *De novo* sequencing of top-down tandem mass spectra: A next step towards retrieving a complete protein sequence. *Proteomes*, 5(1): 6. DOI: 10.3390/proteomes5010006
33. Vyatkina K., Dekker L.J.M., Wu S., VanDuijn M.M., Liu X., Tolic N., Luider T.M., Pasa-Tolic L. (2017) *De novo* sequencing of peptides from high-resolution bottom-up tandem mass spectra using top-down intended methods. *Proteomics*, 17(23-24). DOI: 10.1002/pmic.201600321
34. Ma B. (2015) Novor: Real-time peptide *de novo* sequencing software. *J. Am. Soc. Mass Spectrom.* 26(11):1885-1894. DOI: 10.1007/s13361-015-1204-0
35. Elias J.E., Gygi S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, 4(3):207-214. DOI: 10.1038/nmeth1019
36. Artemenko K.A., Samgina T.YU., Lebedev A.T. (2006) Mass-spektrometricheskoe *de novo* sekvenirovanie peptidov. *Mass-spektrometriya*, 3(4):225-254.
37. Lebedev A.T., Artemenko K.A., Samgina T.Yu. (2012) Osnovy mass-spektrometrii belkov i peptidov, M.: Tekhnosfera, 176 s.
38. Lebedev A.T., Artemenko K.A., Samgina T. (2015) Mass-spektrometriya v organicheskoy himii (2-e izd.), M.: Tekhnosfera, 704 s.
39. Taylor J.A., Johnson R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem.*, 73(11):2594-2604. DOI: 10.1021/ac001196o
40. Dancik V., Addona T.A., Clauser K.R., Vath J.E., Pevzner P.A. (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 6(3-4):327-42. DOI: 10.1089/106652799318300
41. Frank A.M., Savitski M.M., Nielsen M.L., Zubarev R.A., Pevzner P.A. (2007) *De novo* peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.*, 6(1):114-123. DOI: 10.1021/pr060271u
42. Frank A.M. (2009) A ranking-based scoring function for peptide-spectrum matches. *J. Proteome Res.*, 8(5):2241-2252. DOI: 10.1021/pr800678b
43. Frank A.M. (2009) Predicting intensity ranks of peptide fragment ions. *J. Proteome Res.*, 8(5):2226-2240. DOI: 10.1021/pr800677f
44. Fischer B., Roth V., Roos F., Grossmann J., Baginsky S., Widmayer P., Gruissem W., Buhmann J.M. (2005) NovoHMM: a hidden Markov model for *de novo* peptide sequencing. *Anal. Chem.*, 77(22):7265-7273. DOI: 10.1021/ac0508853
45. Chi H., Sun R.X., Yang B., Song C.Q., Wang L.H., Liu C., Fu Y., Yuan Z.F., Wang H.P., He S.M., Dong M.Q. (2010) pNovo: *De novo* peptide sequencing and identification using HCD spectra. *J. Proteome Res.*, 9(5):2713-2724.

DOI: 10.1021/pr100182k

46. Jeong K., Kim S., Pevzner P.A. (2013) UniNovo: a universal tool for *de novo* peptide sequencing. *Bioinformatics*, 29(16):1953-1962. DOI: 10.1093/bioinformatics/btt338
47. Olsen J.V., Macek B., Lange O., Makarov A., Horning S., Mann M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods*, 4(9):709-712. DOI: 10.1038/nmeth1060
48. Syka J.E., Coon J.J., Schroeder M.J., Shabanowitz J., Hunt D.F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 101(26):9528-33. DOI: 10.1073/pnas.0402700101
49. Zubarev R.A., Kelleher N.L., McLafferty F.W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.*, 120(13):3265-3266. DOI: 10.1021/ja973478k
50. Frese C.K., Altelaar A.F., van den Toorn H., Nolting D., Griep-Raming J., Heck A.J., Mohammed S. (2012) Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal. Chem.*, 84(22):9668-9673. DOI: 10.1021/ac3025366
51. Madsen J.A., Boutz D.R., Brodbelt J.S. (2010) Ultrafast ultraviolet photodissociation at 193 nm and its applicability to proteomic workflows. *J. Proteome Res.*, 9(8):4205-4214. DOI: 10.1021/pr100515x
52. Robotham S.A., Horton A.P., Cannon J.R., Cotham V.C., Marcotte E.M., Brodbelt J.S. (2016) UVnovo: A *de novo* sequencing algorithm using single series of fragment ions via chromophore tagging and 351 nm ultraviolet photodissociation mass spectrometry. *Anal. Chem.*, 88(7):3990-3997. DOI: 10.1021/acs.analchem.6b00261
53. Chi H., Chen H., He K., Wu L., Yang B., Sun R.X., Liu J., Zeng W.F., Song C. Q., He S.M., Dong M.Q. (2013) pNovo+: *De novo* peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res.*, 12(2):615-625. DOI: 10.1021/pr3006843
54. He L., Ma B. (2010) ADEPTS: advanced peptide *de novo* sequencing with a pair of tandem mass spectra. *J. Bioinform. Comput. Biol.*, 8(6):981-994. DOI: 10.1142/S0219720010005099
55. Savitski M.M., Nielsen M.L., Zubarev R.A. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol. Cell Proteomics*, 4(8):1180-1188. DOI: 10.1074/mcp.T500009-MCP200
56. Savitski M.M., Nielsen M.L., Kjeldsen F., Zubarev R.A. (2005) Proteomics-grade *de novo* sequencing approach. *J. Proteome Res.*, 4(6):2348-2354. DOI: 10.1021/pr050288x
57. Bertsch A., Leinenbach A., Pervukhin A., Lubeck M., Hartmer R., Baessmann C., Elnakady Y.A., Muller R., Bocker S., Huber C.G., Kohlbacher O. (2009) *De novo* peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis*, 30(21):3736-47. DOI: 10.1002/elps.200900332
58. Datta R., Bern M. (2009) Spectrum fusion: using multiple mass spectra for *de novo* peptide sequencing. *J. Comput. Biol.*, 16(8):1169-1182. DOI: 10.1089/cmb.2009.0122
59. Guthals A., Clauser K.R., Frank A.M., Bandeira N. (2013) Sequencing-grade *de novo* analysis of MS/MS Triplets (CID/HCD/ETD) from overlapping peptides. *J. Proteome Res.*, 12(6):2846-2857. DOI: 10.1021/pr400173d
60. Horton A.P., Robotham S.A., Cannon J.R., Holden D.D., Marcotte E.M., Brodbelt J.S. (2017) Comprehensive *de novo* peptide sequencing from MS/MS pairs generated through complementary collision induced dissociation and 351 nm ultraviolet photodissociation. *Anal. Chem.*, 89(6):3747-3753. DOI: 10.1021/acs.analchem.7b00130
61. Bandeira N., Tang H., Bafna V., Pevzner P. (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Anal. Chem.*, 76(24):7221-7233. DOI: 10.1021/ac0489162
62. Bandeira N., Clauser K.R., Pevzner P.A. (2007) Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell Proteomics*, 6(7):1123-1134. DOI: 10.1074/mcp.M700001-MCP200
63. Bandeira N., Pham V., Pevzner P., Arnott D., Lill J.R. (2008) Automated *de novo* protein sequencing of monoclonal antibodies. *Nat. Biotechnol.*, 26(12):1336-1338. DOI: 10.1038/nbt1208-1336
64. Castellana N.E., Pham V., Arnott D., Lill J.R., Bafna V. (2010) Template proteogenomics: sequencing whole proteins using an imperfect database. *Mol. Cell Proteomics*, 9(6):1260-1270. DOI: 10.1074/mcp.M900504-MCP200
65. Liu X., Han Y., Yuen D., Ma B. (2009) Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics*, 25(17):2174-80. DOI: 10.1093/bioinformatics/btp366
66. Blank-Landeshammer B., Kollipara L., Bi K., Pfenninger M., Malchow S., Shuvaev K., Zahedi R. P., Sickmann A. (2017) Combining *de novo* peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. *J. Proteome Res.*, 16(9):3209-3218. DOI: 10.1021/acs.jproteome.7b00198
67. Yang H., Chi H., Zhou W.-J., Zeng W.-F., He K., Liu C., Sun R.-X., He S.-M. (2017) Open-pNovo: *De novo* peptide sequencing with thousands of protein modifications. *J. Proteome Res.*, 16(2):645-654. DOI: 10.1021/acs.jproteome.6b00716
68. Creasy D.M., Cottrell J.S. (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics*, 4(6):1534-1536. DOI: 10.1002/pmic.200300744
69. Gorshkov V., Hotta S.Y.K., Verano-Braga T., Kjeldsen F. (2016) Peptide *de novo* sequencing of mixture tandem mass spectra. *Proteomics*, 16(18):2470-2479. DOI: 10.1002/pmic.201500549
70. Horn D.M., Zubarev R.A., McLafferty F.W. (2000) Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 97(19):10313-10317. DOI: 10.1073/pnas.97.19.10313
71. Liu X., Dekker L.J.M., Wu S., VanDuijn M.M., Luider T.M., Tolic N., Kou Q., Dvorkin M., Alexandrova S., Vyatkina K., Pasa-Tolic L., Pevzner P. A. (2014) *De novo* protein sequencing by combining top-down and bottom-up tandem mass spectra. *J. Proteome Res.*, 13(7):3241-3248. DOI: 10.1021/pr401300m
72. Liu X., Inbar Y., Dorrestein P.C., Wynne C., Edwards N., Souda P., Whitelegge J.P., Bafna V., Pevzner P.A. (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell Proteomics*, 9(12):2772-2782. DOI: 10.1074/mcp.M110.002766
73. Ecker D.M., Jones S.D., Levine H.L. (2015) The therapeutic monoclonal antibody market. *MAbs*, 7(1):9-14. DOI: 10.4161/19420862.2015.989042
74. Tran N.H., Rahman M.Z., He L., Xin L., Shan B., Li M. (2016) Complete *de novo* assembly of monoclonal antibody sequences. *Sci. Rep.*, 6:31730. DOI: 10.1038/srep31730
75. Guthals A., Gan Y., Murray L., Chen Y., Stinson J., Nakamura G., Lill J.R., Sandova W., Bandeira N. (2017) *De novo* MS/MS sequencing of native human antibodies. *J. Proteome Res.*, 16(1):45-54. DOI: 10.1021/acs.jproteome.6b00608
76. Vonk F.J., Casewell N.R., Henkel C.V., Heimberg A.M., Jansen H.J., McCleary R.J., Kerkkamp H.M., Vos R.A.,

- Guerreiro I., Calvete J.J., Wuster W., Woods A.E., Logan J.M., Harrison R.A., Castoe T.A., de Koning A.P., Pollock D.D., Yandell M., Calderon D., Renjifo C., Currier R. B., Salgado D., Pla D., Sanz L., Hyder A.S., Ribeiro J.M., Arntzen J.W., van den Thillart G.E., Boetzer M., Pirovano W., Dirks R.-P., Spaink H.P., Duboule D., McGlinn E., Kini R.M., Richardson M.K. (2013) The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. USA*, 110:20651-20656. DOI: 10.1073/pnas.1314702110
77. Petras D., Heiss P., Harrison R.A., Sussmuth R.D., Calvete J.J. (2016) Top-down venomomics of the East African green mamba, *Dendroaspis angusticeps*, and the black mamba, *Dendroaspis polylepis*, highlight the complexity of their toxin arsenals. *J. Proteomics*, 46:148-164. DOI: 10.1016/j.jprot.2016.06.018
78. Bhatia S., Kil Y.J., Ueberheide B., Chait B.T., Tayo L., Cruz L., Lu B., Yates III J. R., Bern M. (2012) Constrained *de novo* sequencing of conotoxins. *J. Proteome Res.*, 11(8): 4191-4200. DOI: 10.1021/pr300312h
79. Pukala T.L., Bowie J.H., Maselli V.M., Musgrave I.F., Tyler M.J. (2006) Host-defence peptides from the glandular secretions of amphibians: structure and activity. *Nat. Prod. Rep.*, 23(3):368-393. DOI: 10.1039/b512118n
80. Samgina T.Yu., Artemenko K.A., Gorshkov V.A., Ogourtsov S.V., Zubarev R.A., Lebedev A.T. (2008) *De novo* sequencing of peptides secreted by the skin glands of the Caucasian Green Frog *Rana ridibunda*. *Rapid Commun Mass Spectrom.*, 22(22):3517-3525. DOI: 10.1002/rcm.3759
81. Lebedev A., Samgina T. (2013) O chem mogut rasskazat' lyagushki? Izuchenie peptidnogo sostava kozhnogo sekreta amfibij. *Analitika*, 5(12):38-47.
82. Simmaco M., Mignogna G., Barra D., Bossa F. (1994) Antimicrobial peptides from skin secretions of *Rana esculenta*. Molecular cloning of cDNAs encoding esculentin and brevinins and isolation of new active peptides. *J. Biol. Chem.*, 269(16):11956-11961.
83. Terterov I., Vyatkina K., Kononikhin A.S., Boitsov V., Vyazmin S., Popov I. A., Nikolaev E.N., Pevzner P., Dubina M. (2014) Application of *de novo* sequencing tools to study abiogenic peptide formations by tandem mass spectrometry. The case of homo-peptides from glutamic acid complicated by substitutions of hydrogen by sodium or potassium atoms. *Rapid Commun Mass Spectrom.*, 28(1):33-41. DOI: 10.1002/rcm.6757
84. Robidart J., Callister S.J., Song P., Nicora C.D., Wheat C.G., Girguis P.R. (2013) Characterizing microbial community and geochemical dynamics at hydrothermal vents using osmotically driven continuous fluid samplers. *Environ. Sci. Technol.*, 47(9):4399-4407. DOI: 10.1021/es3037302
85. Menschaert G., Vandekerckhove T.T., Baggerman G., Landuyt B., Sweedler J.V., Schoofs L., Luyten W., Van Criekeing W. (2010) A hybrid, *de novo* based, genome-wide database search approach applied to the sea urchin neuropeptidome. *J. Proteome Res.*, 9(2):990-996. DOI: 10.1021/pr900885k
86. Carrasco M.A., Buechler S.A., Arnold R.J., Sformo T., Barnes B.M., Duman J.G. (2011) Elucidating the biochemical overwintering adaptations of larval *Cucujus clavipes puniceus*, a nonmodel organism, via high throughput proteomics. *J. Proteome Res.*, 10(10):4634-4646. DOI: 10.1021/pr200518y
87. Laskay U.A., Srzentic K., Monod M., Tsybin Y.O. (2014) Extended bottom-up proteomics with secreted asparatic protease Sap9. *J. Proteomics*, 110:20-31. DOI: 10.1016/j.jprot.2014.07.035
88. Srzentic K., Fornelli L., Laskay U.A., Monod M., Beck A., Ayoub D., Tsybin Y.O. (2014) Advantages of extended bottom-up proteomics using Sap9 for analysis of monoclonal antibodies. *Anal. Chem.*, 86(19):9945-9953. DOI: 10.1021/ac502766n
89. Devabhaktuni A., Elias J.E. (2016) Application of *de novo* sequencing to large-scale complex proteomics data sets. *J. Proteome Res.*, 15(3):732-742. DOI: 10.1021/acs.jproteome.5b00861
90. Yang H., Chi H., Zhou W.-J., Zeng W.-F., Liu C., Wang R.-M., Wang Z.-W., Niu X.-N., Chen Z.-L., He S.-M. (2018) pSite: Amino acid confidence evaluation for quality control of *de novo* peptide sequencing and modification site localization. *J. Proteome Res.*, 17(1):119-128. DOI: 10.1021/acs.jproteome.7b00428

Поступила: 15. 03. 2018.
Принята к публикации: 20. 03. 2018.

DE NOVO SEQUENCING OF PROTEINS AND PEPTIDES: ALGORITHMS, APPLICATIONS, PERSPECTIVES

K.V. Vyatkina

¹Saint Petersburg National Research Academic University of the Russian Academy of Sciences;
8/3 Khlopina str., St Petersburg, 194021 Russia; e-mail: vyatkina@spbau.ru

²Saint Petersburg State University; 7-9 Universitetskaya nab., St Petersburg, 199034 Russia.

Determination of the primary structure of proteins and peptides constitutes an important step in studying their properties. Currently, mass spectrometry is commonly applied to this end. The results of mass spectrometric measurements can be interpreted by means of either database search or *de novo* sequencing methods. The appeal of the latter is due to their applicability to investigating unknown proteins, as well as the ones that cannot be analyzed with genomics or transcriptomics methods. In this paper we briefly review the existing approaches to *de novo* sequencing of proteins and peptides, along with the problems that can be solved using those, and indicate directions and perspectives for their further development.

Key words: proteomics; mass spectrometry; de novo sequencing; proteins; peptides; amino acid sequence