Биомедицинская химия, 2017 том 63, вып. 4, с. 341-350.

ПРИМЕНЕНИЕ МЕТОДОВ *DE NOVO* СЕКВЕНИРОВАНИЯ ДЛЯ ИДЕНТИФИКАЦИИ БЕЛКОВ

В.С. Скворцов*, А.В. Микурова, А.В. Рыбина

Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; эл. почта: vladlen@ibmh.msk.su

С использованием трёх программ *de novo* секвенирования (Novor, PEAKS и PepNovo+) проведена идентификация белков из набора, включающего 48 белков человека (набор UPS2 компании “Sigma-Aldrich”, США). Экспериментальные данные получены при помощи тандемной масс-спектрометрии. В работе исследованы как набор данных белков, так и комбинированные пробы с добавлением экстракта *E. coli* и плазмы крови человека. При использовании критерия идентификации (два пептида длиной не менее девяти остатков или один пептид не короче 13 остатков) в “чистой” пробе обнаружено 13 (Novor), 20 (Peaks) и 11 (PepNovo+) белков из набора UPS2. Для комбинированных проб результат был или сопоставимым, или хуже. При использовании предсказаний, включающих высокодостоверный фрагмент (TAG) последовательности длиной не менее 7 остатков, и “остаточные” (неидентифицированные) массы с N- и C-концов (PepNovo+), результат существенно улучшается (~20%) и хорошо идентифицируются масс-спектрометрические артефакты и, вероятно, PTM. В работе зарегистрированы нестандартные C-концевые изменения масс (+23,02, +26,04 и +27,03), встречающиеся статистически достоверно. При использовании пептидов, содержащих эти модификации, а также пограничных критериев идентификации, в различных пробах удалось обнаружить 41 из 48 белков набора UPS2.

Ключевые слова: панорамная протеомика, *de novo* секвенирование, масс-спектрометрия, обработка данных

DOI: 10.18097/PBMC20176304341

ВВЕДЕНИЕ

Одна из первоочередных задач современной протеомики – каталогизация всех белков организма при различных физиологических (или патологических) состояниях, которая, несмотря на развитие современных технологий, до сих пор не решена. Большая часть белков, “вычисленная” из генома (до 17% [1]), так никогда и не была экспериментально определена в организме. Частично это связано с тем, что большая часть этих так называемых “отсутствующих” белков [2] (в англоязычной литературе обозначаемых как “missing proteins”) может присутствовать в организме в столь малых количествах, что они не доходят до “уровня детекции”. В настоящее время одним из наиболее широко применяемых способов для идентификации белков в низких и очень низких количествах является масс-спектрометрия.

Современная каталогизация белков основывается на двух основных подходах: таргетной протеомике и панорамной протеомике (shotgun proteomic technologies) [3]. Первый такой подход, например, как SRM (selected reaction monitoring), заключается в селективном мониторинге фрагментарных ионов, полученных при диссоциации в узком массовом диапазоне ионной пары “родительский ион – ион-фрагмент” [4]. За счёт этого достигается высокая чувствительность и специфичность в отношении анализируемых белков. Хотя метод SRM, как и другие методы таргетной протеомики, высокочувствителен, он времязатратен и дорог. Для его реализации необходимо синтезировать определённые (протеоспецифические) пептиды, которые должны учитывать возможные мутации (SAP; single amino-acid polymorphism) и пост-трансляционные модификации (PTM; Post-translational modification). Эти условия далеко не всегда выполнимы при анализе биологических объектов. На первый взгляд, в панорамной протеомике этой проблемы нет: верификация найденных пептидов осуществляется путём сравнения спектров (а чаще их фрагментов) с *a priory* заданной базой данных и со строго фиксированным набором возможных изменений. Однако и у этого метода есть существенный недостаток: все возможные варианты изменений пептидов (SAP, PTM и др.) должны быть заданы программе до проведения поиска. Этот недостаток может быть преодолён, если при идентификации белка на основе данных масс-спектрометрии проводится так называемое *de novo* секвенирование. Собственно, под *de novo* секвенированием следует понимать весь комплекс экспериментально-аналитических процедур, включающих подготовку проб, тандемную масс-спектрометрию (MS/MS) и анализ полученных результатов без привлечения данных о последовательностях белков, используя исключительно знание того, каким образом может происходить образование вторичных ионов (рис. 1) [5]. Исторически именно этот метод идентификации последовательности и был первым. Но из-за того, что точность приборов раньше была недостаточной, компьютеры небыстрыми, а спектры часто получаются неполными, более широкое распространение получил метод по известным БД последовательностей с использованием сравнения реальных спектров с теоретическими [6]. Название “*de novo* секвенирование” закрепилось в настоящее время за группой программ, которые продолжают расшифровывать последовательность

* - адресат для переписки

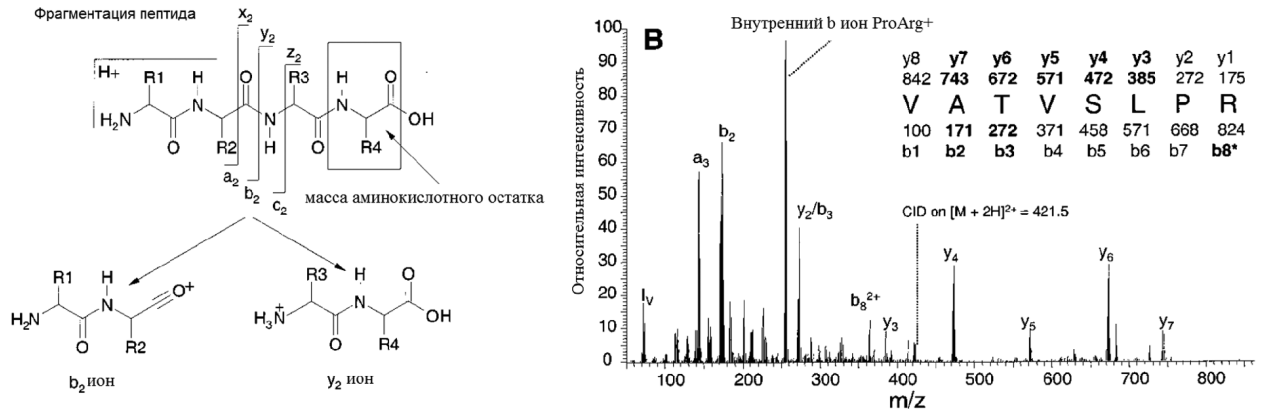


Рисунок 1. Восстановление последовательности пептида по масс-спектру первичного иона (адаптировано из [5]).

только по полученному спектру, не используя априорной информации о последовательности искомых белков. Однако точность приборов за последние годы значительно возросла, быстродействие компьютеров повысилось многократно, появились новые алгоритмы работы программ обработки MS/MS спектров.

Целью данной работы было оценить, как лучше использовать программы *de novo* секвенирования в задачах идентификации белков, насколько результат зависит от концентрации белков и состава проб, и какие есть ограничения для использования данного метода. Сравнение качества работы данных программ не было задачей работы (это не раз делалось и ранее [7]). Они примерно равны, хотя каждая имеет и свои особенности, определяющие возможность их применения для различных случаев. Например, в отличие от PEAKS [8] авторы Novor'а [9] ограничиваются всего одним вариантом предсказанного пептида для конкретного спектра, что не всегда удобно, так как в местах, где определяется высоко вероятное положение нескольких аминокислотных остатков (брутто), их порядок часто определяется неправильно (PEAKS их варьирует). Обе программы дают оценку вероятности правильности детекции по каждому аминокислотному остатку, это можно использовать. PepNovo+ [10] в первую очередь предсказывает так называемый TAG (часть аминокислотной последовательности, для которой программа в состоянии поставить в соответствие имеющиеся пики для каждого из предсказанных остатков), оставшаяся часть пептида представлена остаточными массами (масса группы остатков для последовательности неизвестной длины). В нашей работе использовались остаточные массы только с N- и C- концов, но возможно и использование вставок из остаточных масс [8].

МЕТОДИКА

В работе использованы данные масс-спектрометрического исследования, полученные на стандартизованном наборе из 48 белков человека (UPS2 компании "Sigma-Aldrich", США) (табл. 1), и доступные в хранилище протеомных данных PRIDE (https://www.ebi.ac.uk/pride, Submission Reference: PXD007643). По данным авторов: набор UPS2 состоит из 48 высокоочищенных белков человека без SAP (PTM возможны). Белки разделены на 5 групп по концентрации, присутствующей в пробе; концентрации различаются до 5 порядков (от 10^{-11} M до 10^{-6} M). Всего в работе были исследованы 5 проб: UPS2, не содержащая ничего кроме указанного набора (2 технических повтора); ISTD – плазма крови человека (2); MTRX-экстракт *E. coli* (3); UPS-ISTD – смесь набора UPS2 и плазмы крови человека (2); UPS-MTRX – смесь набора UPS2 и экстракта *E. coli* (3). Во всех пробах, содержащих набор UPS2, его количество было одинаковое. Пробоподготовка включала протеолиз с использованием трипсина и алкилирование йодацетамидом. MS анализ выполняли на масс-спектрометре Orbitrap Fusion ("Thermo Scientific", США). Диапазон сканирования от 400 до 1200 m/z. MS/MS сканирование выполнялось с разрешением 17500, окно изоляции 2,0 m/z, режим фрагментации HCD. Заряд первичного иона находился в диапазоне от +2 до +6.

В работе были использованы три наиболее популярные программы *de novo* секвенирования: PEAKS [8] (http://www.bioinfor.com/peaks/features/denovo.html), Novor [9] и PepNovo+ [10]. Две последние в составе общедоступного ПО DeNovo GUI версия 1.15.3 (http://compomics.github.io/projects/denovogui.html) [11]. Для работы всех трёх программ был использован сходный набор параметров: фермент гидролиза трипсин, точность детекции первичного иона 3 ppm, точность определения пиков вторичных ионов 0,01 Да. Для всех остатков цистеина принималось, что они присутствуют всегда в форме карбамидометил цистеина, допускалось вариативное окисление остатков метионина. В качестве факультативной замены допускалось спонтанное дезаминирование остатков глутамина и аспарагина; анализируемые масс-спектры содержали достаточно большое число таких замен. К сожалению, программа PepNovo+ в текущей версии поддерживает очеьограниченное число модификаций. Минимальный размер TAG (фрагмент пептида, разрешённый достоверно) для PepNovo+ был установлен на уровне 7. Программы PEAKS и PepNovo+ могли генерировать до 10 вариантов пептидов для каждого спектра (Novor только 1, см. выше).

Таблица 1. Описание набора UPS2 и результаты идентификации белков различными программами (пояснения в тексте)

N	UniProt ID	UniProt рекомендованное имя (краткое)	Chain	Источник белка	Potential PTM	Кол-во, фемтоМ	Novor	PEAKS					PepNovo+				
							UPS2	UPS2	ISTD	MTX	USPMTX	UPSISTD	UPS2	ISTD	MTX	USPMTX	UPSISTD
6	P00915	Carbonic anhydrase 1	2-261	Erythrocytes	Acetylation	50000	+	+			+	+	+	?		+	+
7	P00918	Carbonic anhydrase 2	2-260	Erythrocytes	Acetylation	50000		+			+	+	+	*			+
9	P01031	Complement C5/C5a anaphylatoxin	678-751	Recombinant		50000		+			+	+	+				+
40	P02768	Serum albumin	26-609	Recombinant		50000		+	+			+	+	+			+
27	P41159	Leptin	22-167	Recombinant		50000		+			+	+	*				+
46	P62988	Ubiquitin	1-76	Recombinant		50000	+										
21	P68871	Hemoglobin subunit beta	2-147	Erythrocytes	Acetylation Glycosylation Nitrosylation Phosphorylation	50000	+	+	+		+	+	+	+		+	+
20	P69905	Hemoglobin subunit alpha	2-142	Erythrocytes	Glycosylation Phosphorylation	50000	+	+	+		+	+	+	?		?	+
12	P00167	Cytochrome b5	2-134	Recombinant		5000	+	+				+	+				+
36	P01133	Pro-epidermal growth factor (EGF)/Epidermal growth factor	971-1023	Recombinant		5000					+	+	+				+
30	P02144	Myoglobin	2-154	Heart		5000	+	+			+	+	+				+
8	P04040	Catalase	2-527	Erythrocytes	Phosphorylation	5000	+	+			+	+	*				+
31	P15559	NAD(P)H dehydrogenase [quinone] 1	2-274	Recombinant		5000		+			+	+	+			#	+
33	P62937	Peptidyl-prolyl cis-trans isomerase A (PPIase A, Rotamase A)	1-165	Recombinant		5000		+				+	+			+	+
41	P63165	Small ubiquitin-related modifier 1 (SUMO-1)	35431	Recombinant		5000		+					+				+
34	Q06830	Peroxiredoxin 1	2-199	Recombinant		5000		+			+	+	+				+
1	P00709	Alpha-lactalbumin	20-142	Milk	Glycosylation	500		+					#	?		+	
37	P02753	Retinol-binding protein 4	19-201	Urine		500			+			+	#	+		+	+
11	P06732	Creatine kinase M-type	1-381	Heart		500		+					+				#
22	P12081	Histidyl-tRNA synthetase, cytoplasmic	1-509	Recombinant		500	+	+					+				+
38	P16083	Ribosyldihydronicotinamide dehydrogenase [quinone]	2-231	Recombinant		500	+	+					*	*		*	+
28	P61626	Lysozyme C	19-148	Milk		500	+						*				?
42	P63279	SUMO-conjugating enzyme UBC9	1-158	Recombinant		500	+						#	+		+	
32	Q15843	NEDD8	1-81	Recombinant		500											

Таблица 1. Описание набора UPS2 и результаты идентификации белков различными программами (продолжение)

15	O76070	Gamma-synuclein	1-127	Recombinant		50	+						*	+			*
3	P01008	Antithrombin-III (ATIII)	33-464	Plasma	Glycosylation	50			+			+				+	+
35	P01127	Platelet-derived growth factor subunit B (PDGF subunit B)	82-190	Recombinant		50	+						?				
23	P01344	Insulin-like growth factor II (IGF-II)	25-91	Recombinant		50		+					#			?	
17	P08263	Glutathione S-transferase A1 (GTH1, HA subunit 1)	2-222	Recombinant		50								?			
44	P10599	Thioredoxin (Trx)	2-105	Recombinant		50	+							?			
5	P55957	BH3-interacting domain death agonist	1-195	Recombinant		50	+									?	
4	P61769	Beta-2-microglobulin	21-119	Urine		50											
47	O00762	Ubiquitin-conjugating enzyme E2 C	1-179	Recombinant		5	+							?			
19	P01112	GTPase HRas	1-189	Recombinant		5	+									?	
24	P01579	Interferon gamma (IFN-gamma)	23-166	Recombinant		5							*			*	?
39	P02787	Serotransferrin (Transferrin)	20-698	Plasma	Phosphorylation Glycosylation	5			+			+	+				+
16	P06396	Gelsolin	28-782	Plasma	Phosphorylation	5											
18	P09211	Glutathione S-transferase P	2-210	Placenta		5	+						?			+	
48	P51965	Ubiquitin-conjugating enzyme E2 E1	1-193	Recombinant		5											
13	P99999	Cytochrome c	2-105	Recombinant		5	+									+	?
43	P00441	Superoxide dismutase [Cu-Zn]	2-154	Erythrocytes	Acetylation Phosphorylation	0,5	+										?
45	P01375	Tumor necrosis factor/Tumor necrosis factor, soluble form	77-233	Recombinant		0,5										?	
10	P02741	C-reactive protein	19-224	Plasma		0,5								+			
26	P02788	Lactotransferrin (Lactoferrin)	20-710	Milk	Glycosylation Phosphorylation	0,5										+	
14	P05413	Fatty acid-binding protein, heart	2-133	Heart	Acetylation Phosphorylation	0,5											
2	P08758	Annexin A5	2-320	Placenta	Acetylation Phosphorylation	0,5											
25	P10145	Interleukin-8 (IL-8)	28-99	Recombinant		0,5		+			+	+		?			
29	P10636	Microtubule-associated protein tau	2-441	Recombinant		0,5							*			?	

Примечания. 1. *E. coli* используется как рекомбинантная система для всех рекомбинантных белков кроме P02768 (*Pichia pastoris*). 2. Метки обнаружения белка: “+” - удовлетворяет условиям отбора полностью; “*” - удовлетворяет условиям отбора с учётом нестандартных добавочных масс на С-конце (23, 26, 27);”#” - условия отбора не выполняются, но пептиды, соответствующие белку, присутствуют (нестандартные добавочные массы игнорируются; “?” - аналогично “#”, но с учётом нестандартных добавочных масс.

Анализ идентификации и степени покрытия последовательности предсказанными процедурами *de novo* пептидами выполняли при помощи программы ProteoCat [12], которая выполняет “побуквенное сравнение аминокислотной последовательности” белков и пептидов, при этом остатки изолейцинов и лейцинов считались эквивалентными, допускалась 1 аминокислотная замена. Обычно последнее допущение служит для поиска SAP, однако, в настоящей работе все подобные замены могли быть объяснены спонтанными модификациями при пробоподготовке или особенностями работы алгоритмов программ *de novo* секвенирования. Для данных, полученных программой PepNovo+, была написана утилита, выполняющая сравнение по массам для конкретных положений остатков (это снимало часть артефактов предсказания, например, “перепутанные” остатки лизина и глутамина при неполном трипсинолизе белков) и суммам остаточных масс с последующим выравниванием пептид относительно последовательности целевого белка. Анализ выполнялся вручную, с помощью программы MS Excel. Для всех случаев, кроме специально оговорённых, признаком идентификации белка было обнаружение не менее двух пептидов длиной 9 и более аминокислотных остатков, или одного пептида длиной 13 и более остатков. Программа ProteoCat позволяет анализировать ложные срабатывания при наличии гомологов, но в данной работе эта функция не использовалась.

Для идентификации белков использовали базу данных (БД) аминокислотных последовательностей, содержащую все вычисленные из геномов белки человека и *E. coli*. Для оценки адекватности предсказания также использовали равноразмерную псевдослучайную выборку последовательностей (так называемую decoy БД), полученную путём инверсии целевой БД. Так как в случае PepNovo+ имелся элемент ручного анализа, то для этой группы предсказаний проводили сравнение только с 48 белками из набора UPS2. Для данных, полученных с помощью PEAKS и Novor, в работе рассчитывали показатель ложноположительной идентификации (FDR; false discovery rate), как отношение числа идентификаций, полученных на decoy БД к числу идентификаций, полученных на целевой БД [8].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Чтобы понять, какие белки могут быть обнаружены в сложной (биологической) смеси, проанализируем для начала состав выборки UPS2 [13], данные о котором представлены в таблице 1. Учитывая, что белки для набора получают из различных источников, то даже в чистой пробе, содержащей исключительно набор UPS2, можно ожидать присутствие и других белков человека (которые попадают в препарат целевого белка при очистке соответствующих тканей), а также белков *E. coli* (в случае для рекомбинантных белков). Даже при высокой степени очистки число примесей может быть достаточно велико. Например, если тот же гемоглобин содержит всего 1% белковых примесей, то они в 1000 раз по общему объёму белка превосходят интерлейкин-8, то есть отдельные белки в примеси вполне могут быть в сопоставимых количествах с белками, содержащимися в наборе UPS2 в низкой концентрации. Для части белков, полученных из биологического материала, возможно наличие PTM, осложняющих детекцию. И наконец, в то время как для сравнения используются последовательности белков “полной длины”, в наборе часть белков представлены в виде фрагментов или укороченные в результате пост-трансляционного процессинга до физиологического размера. Впрочем, это никак не влияло на качество детекции белков, хотя и вносило ошибку при расчёте степени покрытия белка пептидами при идентификации.

Несмотря на то, что точность измерения масс в MS/MS анализе определяется, в первую очередь, используемой приборной базой, тем не менее, было проверено, насколько влияют данные параметры на работу программ *de novo* секвенирования. Приведённые в таблице 2 значения оценочных функций (ALC >50% и Novor score >70) соответствуют тем, что обычно используются по умолчанию [8, 9]. Дополнительная проверка показала, что уменьшение порога не даёт увеличения количества идентифицированных белков из набора UPS2, а вот увеличение порога существенно уменьшает это число (табл. 2). Следует отметить, что для предсказаний PepNovo+ (табл. 2) были использованы только те пептиды,

Таблица 2. Сравнение результатов (число идентифицированных белков), полученных при разной точности идентификации спектров (проба UPS2)

ПО	Точность для первичного иона (ppm)	Белки UPS2 (точность вторичного иона, Да)				Другие белки человека (точность вторичного иона, Да)			
		0,1	0,05	0,02	0,01	0,1	0,05	0,02	0,01
PEAKS (ALC > 50%)	3	20	20	20	20	85	87	88	81
	5	14	14	20	18	91	94	88	68
Novor (ALC > 50%)	3			18	17			63	64
	5			18	13			64	42
PepNovo+ (pscore > 0, без остаточных масс)	3			11	10			23	23
	5			12	12			24	22

для которых программа предсказывала TAG, равный полной длине пептида. Выбор точности в 3 ppm и 0,01 Да оправдан.

Вторым важным фактором был выбор условий признания белка идентифицированным. Нами были проанализированы данные, полученные программой PEAKS (табл. 3). В таблице 3 для краткости представлены 5 групп, в каждой группе число “идентифицированных” белков по отдельному параметру зависит от того, был ли уже белок идентифицирован при более жёстких параметрах при одинаковом числе пептидов. Например, если белок идентифицирован при длине одиночного пептида в 15 аминокислотных остатков, то он уже не рассматривается при тестировании пептидов меньшей длины. Установленный *a priory* порог (два пептида по 9 остатков или один – от 13 остатков) даже несколько жёстче, чем мог бы быть. Использование decoy ДБ показало, что при длине пептидов (или TAG) более 7 случайных совпадений не наблюдается. При длине пептида в 7 остатков случайные совпадения встречаются, но не более чем по одному на белок (исключая случаи повторов участков последовательности в конкретном белке).

В случае применения всех трёх программ *de novo* секвенирования в варианте “работа по умолчанию”, они дают вполне предсказуемый результат (рис. 2, табл. 1). В случае UPS2 ожидаемо находятся белки, представленные в пробе в больших количествах. Больше половины идентифицированных белков имеют не менее трёх пептидов (рис. 2Б), в отдельных случаях до 15 пептидов и более 60% покрытия аминокислотной последовательности. Белки из набора UPS2 не обнаружены в экстракте *E. coli* (табл. 3, проба MTX), наличие же небольшого количества идентификаций белков человека связано с тем, что находятся гомологичные белки. Напомним, что в работе идентификация идёт по выявленным пептидам, а не так называемым “характеристическим”, которые специфичны исключительно для конкретных белков. Так как набор UPS2 содержит несколько “мажорных” белков плазмы крови и эритроцитов, неудивительно, что семь из них были обнаружены в пробе ISTD.

Давно известно, что метод панорамной протеомики очень сильно подвержен случайным флуктуациям, которые определяют высокий процент вариаций при технических повторах одной и той же пробы. Это хорошо видно при сравнении “чистых” проб плазмы крови и экстракта *E. coli* и тех же проб с добавленным UPS2 (рис. 3). В этих, по сути одних и тех же, пробах (выполненных в 2-3 технических повторах) были обнаружены значительные несовпадения. Основной же вывод при анализе проб UPSISTD и UPSMTX тривиален: внесение “белкового шума” ухудшает возможность детекции белков.

Но так ли это на самом деле? Основная проблема программ *de novo* секвенирования – это фрагментарность полученных спектров. Не вызывает сомнения, что для “идеальных” спектров

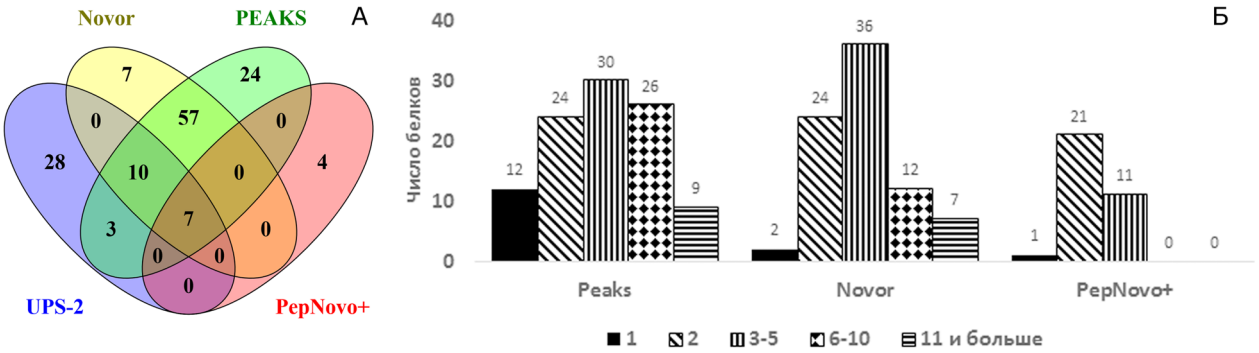


Рисунок 2. Сравнение результатов идентификации белков из набора UPS2 с пептидами, полученными разными программами *de novo* секвенирования. Проба UPS2. А. Диаграмма Венна. Б. Распределение по числу найденных пептидов (все белки, идентифицированные в пробе). Примечание. В случае PepNovo+ были использованы пептиды, входящие в “tag” целиком, без учёта остаточных масс на N- или C-концах.

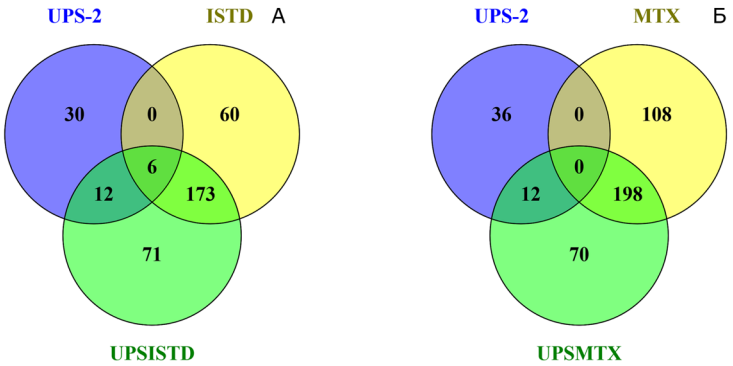


Рисунок 3. Сравнение результатов идентификации белков из набора UPS2 при анализе чистой пробы и с внесением шума при добавлении плазмы крови (А) и экстракта *E. coli* (Б). Идентификация программой PEAKS.

Таблица 3. Число идентифицированных пептидов при различных критериях достаточности для идентификации

	Число пептидов	Длина пептидов	Проба UPS2						Проба MTX						Проба UPSMTX						Проба ISTD						Проба UPSISTD					
			UPS2	HUMAN	E.COLI	Bcero	decoy	FDR %	UPS2	HUMAN	E.COLI	Bcero	decoy	FDR %	UPS2	HUMAN	E.COLI	Bcero	decoy	FDR %	UPS2	HUMAN	E.COLI	Bcero	decoy	FDR %	UPS2	HUMAN	E.COLI	Bcero	decoy	FDR %
1	1	7	23	115	1	139	0	0		15	480	495	4	0,81	16	105	362	483	6	1,23	8	300	1	309	5	1,62	21	326	1	348	4	1,14
	2	7	18	73		91	0	0			259	259	0	0	11	44	200	255	0	0	6	165		171	0	0	18	192		210	0	0
	3	7	17	59		76	0	0			159	159	0	0	8	37	115	160	0	0	4	113		117	0	0	14	154		168	0	0
	1	9	1	7		8	0	0		11	71	82	0	0	2	17	58	77	0	0	1	42		43	0	0		51		51	0	0
	1	11	1	16		17	0	0		2	70	72	0	0		1	43	44	0	0		5		5	0	0		10		10	0	0
	1	13	2	6		8	0	0		2	37	39	0	0	1	2	23	26	0	0		28		28	0	0	1	18		19	0	0
	1	15		4		4	0	0			24	24	0	0	1	3	10	14	0	0	1	36		37	0	0		26		26	0	0
2	1	9	22	106		128	0	0		15	460	475	0	0	14	65	329	408	0	0	7	274		281	0	0	19	296		315	0	0
	2	9	18	71		89	0	0			237	237	0	0	9	39	174	222	0	0	5	153		158	0	0	16	188		204	0	0
	3	9	13	52		65	0	0			146	146	0	0	6	35	98	139	0	0	4	104		108	0	0	13	145		158	0	0
	1	11	1	16		17	0	0		2	78	80	0	0		1	46	47	0	0		6		6	0	0	1	11		12	0	0
	1	13	2	6		8	0	0		2	40	42	0	0	2	5	26	33	0	0		31		31	0	0	1	18		19	0	0
	1	15		4		4	0	0			27	27	0	0	1	3	12	16	0	0	1	42		43	0	0	1	27		28	0	0
3	1	11	21	95		116	0	0		4	371	375	0	0	13	50	250	313	0	0	6	230		236	0	0	19	242		261	0	0
	2	11	12	55		67	0	0			167	167	0	0	8	30	122	160	0	0	4	135		139	0	0	16	161		177	0	0
	3	11	10	40		50	0	0			96	96	0	0	7	18	65	90	0	0	4	84		88	0	0	12	107		119	0	0
	1	13	4	11		15	0	0		2	52	54	0	0	2	10	42	54	0	0		33		33	0	0	1	27		28	0	0
	1	15		4		4	0	0				47	0	0	2	3	21	26	0	0	2	52		54	0	0	1	38		39	0	0
4	1	13	16	78		94	0	0		2	237	239	0	0	12	43	166	221	0	0	6	211		217	0	0	18	217		235	0	0
	2	13	10	49		59	0	0			92	92	0	0	7	27	63	97	0	0	4	100		104	0	0	13	125		138	0	0
	1	15		5		5	0	0			64	64	0	0	3	5	41	49	0	0	2	70		72	0	0	4	57		61	0	0
						0						0						0						0						0		
5	1	15	11	50		61	0	0			150	150	0	0	9	21	91	121	0	0	5	164		169	0	0	16	168		184	0	0
	1	17	7	28		35	0	0			82	82	0	0	6	15	52	73	0	0	4	117		121	0	0	15	126		141	0	0

они дают 100% результат. Собственно, программы поиска по готовой БД и решают эту проблему, так как позволяют идентифицировать спектр по его фрагменту, с учётом ограничения, что именно этот белок ожидается встретить в пробе. Вторая проблема – наличие SAP, PTM и всевозможных масс-спектрометрических артефактов. При настройке программы проблематично подключить весь набор возможных модификаций. С одной стороны, это существенно увеличивает время обработки данных (в случае поиска по теоретическим спектрам – катастрофически, при *de novo* секвенировании – не так уж и существенно). Но самая главная проблема при *de novo* секвенировании – заполнение участков, для которых не находится достоверных пиков на масс-спектре, неправдоподобным набором модифицированных аминокислотных остатков. Например, программа PEAKS может в таком случае выстроить подряд 9-10 остатков, модифицированных различными способами. Конечно, они будут иметь низкую поаминокислотную оценку достоверности, но, если остальная часть пептида будет высоко достоверна, средняя оценка может удовлетворять критериям отбора. Отсутствие вариативности в низкодостоверных частях пептидов – основная причина проигрыша программы Novor. Достаточно часто наблюдаются случаи, когда участок из двух или трёх остатков содержит правильные аминокислоты, но расположенные в другом порядке. Это не представляет серьёзной проблемы, если целевых белков немного, и можно обработать результат по частичному совпадению (программа ProteoCat позволяет выполнить эту процедуру в интерактивном режиме). Но если обрабатываются белки, закодированные всем геномом, то данная процедура становится проблематичной, даже если её автоматизировать. В таком случае удобнее, если недостоверные участки не будут на первом этапе “привязываться” к конкретным остаткам, а будут сравниваться по брутто массам.

Проиллюстрируем это на примере работы программы PepNovo+. Казалось бы, из рисунка 2 следует, что PepNovo+ работает не очень хорошо, но надо помнить, что в данном анализе были использованы только те решения, в которых TAG включал в себя весь пептид. Таких решений не очень много, около 10%. За счёт того, что программа могла сохранять 10 вариантов пептидов для каждого спектра, таких решений было несколько больше. Значение оценочной функции для этих вариантов часто меньше, чем для неполного решения, которое в общем случае можно представить, как последовательность: остаточная масса с N-конца, последовательность разрешённой части (TAG), остаточная масса с C-конца. Повторимся, что остаточная масса может быть и между двумя TAG, но в данной работе этот вариант не рассматривался. Так как анализ данных по неполной идентификации, полученным программой PepNovo+, проводился не полностью автоматически (автоматизировано было только выравнивание относительно искомых последовательностей и отбор записей), то в данном случае рассматривалось только одно решение для одного спектра и анализ проводился только для 48 белков из набора UPS2. В данном случае это ограничение оправдано. Конечно, бывают случаи, когда в пределах одного спектра программа PepNovo+ находит две независимые последовательности TAG, но чаще всего программа “гадает”, увеличивая длину TAG на 1-2 аминокислотных остатка с низкой достоверностью идентификации.

Результаты анализа представлены на рисунке 4. Полученные данные можно разделить на четыре основные группы:

1. Пептид достоверно совпадает с соответствующим пептидом в белке из набора UPS2 (“+”). Это, в первую очередь пептиды, у которых TAG совпадает по длине со всем пептидом, а также те пептиды, у которых массы N и C-концевых фрагментов совпадают полностью, либо с точностью до масс допустимых модификаций (окисление метионина, алкилирование цистеина и т.д.) Кроме того, выполняется условие идентификации белка (два пептида по 9 остатков и больше или 1, но не меньше 13 остатков).

Возможные модификации	TAG (длина)	N-term mass	C-term mass	N-term (число ост.)	N-term (дельта масс)	C-term (число ост.)	C-term (дельта масс)	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107
								S	K	M	D	Q	T	L	A	V	Y	Q	Q	I	L	T	S	M	P	S	R	N	V	I	Q	I	S	N	D	L	E	N	L	R	D
"O3H+"	7	0	759.40	0	0	6	19.02	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N	V	L	Q	L	S	N	-	-	-	-	-	-	-
Нет в БД UNIMOD	7	0	766.42	0	0	6	26.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N	V	L	Q	L	S	N	-	-	-	-	-	-	-
"O3H+"	10	0	402.25	0	0	3	19.02	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N	V	L	Q	L	S	N	D	L	E	-	-	-	-
	7	1306.64	0	11	16.01	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
oxy M	7	1306.62	0	11	15.99	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	7	1290.63	0	11	0.00	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
deam Q	7	1291.62	0	11	1.00	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
oxy M	8	1193.53	0	10	15.99	0	0	-	-	-	-	-	-	-	-	-	-	-	-	L	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	8	1177.56	0	10	0.01	0	0	-	-	-	-	-	-	-	-	-	-	-	-	L	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
deam Q -> cmm E	9	1107.55	0	9	58.06	0	0	-	-	-	-	-	-	-	-	-	-	-	Q	L	L	T	S	M	P	S	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-
oxy M + O 3H+	8	491.17	678.32	4	16.00	6	19.02	-	-	-	-	-	L	A	V	Y	Q	Q	L	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
"O3H+"	9	374.13	678.32	3	0.00	6	19.02	-	-	-	-	T	L	A	V	Y	Q	Q	L	L	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Рисунок 4. Пример анализа остаточных масс по данным (белок P41159), полученным программой PepNovo+.

2. Всё тоже самое, что и для группы 1, но условие достоверной идентификации не выполняется (рассматриваются пептиды длинной не меньше 8).

3. Группа пептидов, имеющая на С-конце добавочную массу вместо 19,02 (вода + два иона H^+) массы +23,02, +26,04 и +27,03 и удовлетворяющая условию достоверной идентификации. Эта группа была выявлена из анализа полученных данных. В БД UNIMOD (http://www.unimod.org, [14]) эти модификации найти не удалось. Такие модификации встречаются: для массы +23,02 в пяти пептидах у пяти белков (во всех зарегистрированных случаях на С-конце пептида остаток аргинина); +26,04 – девяти пептидов в девяти белках (на С-конце встречается как аргинин, так и лизин); +27,03 – 10 пептидов в девяти белках (на С-конце пептида остаток лизина). Учитывая, что проанализированы совпадения для 48 белков, эти изменения не случайны, а так как они встречаются как вариант и для пептидов, имеющих много вариантов (рис. 4), то вряд ли являются признаком неправильного решения для масс-спектра. Вероятнее всего, массы +23,02, +26,04 и +27,03 (или некоторые из них) указывают на образование основания Шиффа или вторичного амина [15]. Следует отметить, что при анализе выявляются и другие более известные артефакты (рис. 4), например, карбамидметилирование лизина и др.

4. Группа содержит пептиды, аналогичные группе 3, но условия идентификации не выполняются полностью, как это имеет место в группе 2.

Следует подчеркнуть, что ни в пробе с “чистым” экстрактом *E. coli*, ни в “decoy” совпадений с белками набора UPS2 нет. Случайные совпадения TAG из семи аминокислотных остатков встречаются, но “привязки” по концам пептидов относительно предсказанных для гидролиза при этом сделать нельзя. Именно поэтому мы и рассматривали пептиды длиной не менее восьми остатков. Легко видеть (рис. 5, табл. 1), что использование данного подхода позволяет идентифицировать в пробе больше белков, не менее чем на 20%, больше чем при использовании “готового” результата, выдаваемого программой PEAKS. А если суммировать результат по всем пробам, то с разной достоверностью был зарегистрирован 41 белок из 48. Суммировать в данном случае не совсем правильно, тем более что в пробе с добавлением плазмы часть искомых белков со 100% вероятностью присутствовала и в пробе без добавления UPS2. Следует напомнить о большой вариативности результатов панорамной протеомики, упомянутой ранее. К сожалению, вне зависимости от сложности приборов и математического аппарата в этом подходе всегда присутствует большой элемент случайности. Кроме того, из полученных результатов можно сделать и заключение о “чистоте пробы”. Это тривиальное наблюдение, но белки, полученные рекомбинантным способом, очищены лучше, в противном случае в пробе UPS2 было бы значительное число белков *E. coli* (как это имеет место для белков человека), а этого нет.

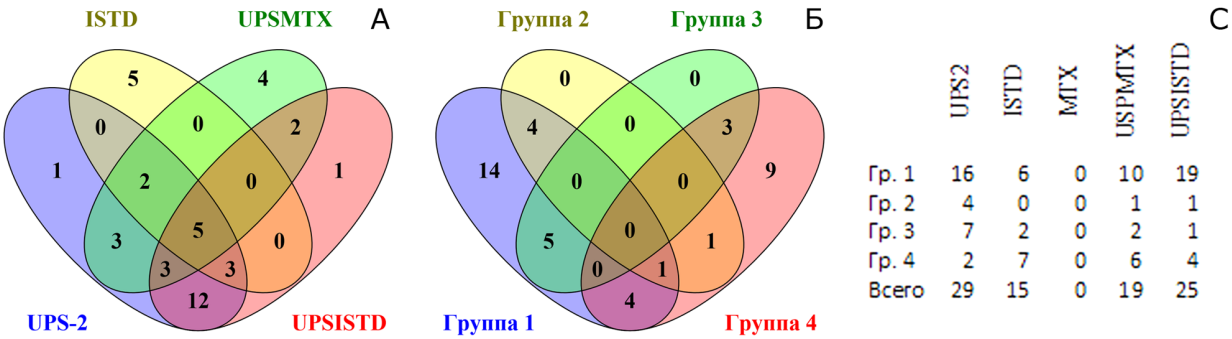
Таким образом, можно заключить, что имеющиеся в настоящее время методы *de novo* секвенирования дают очень хорошие результаты при идентификации в панорамной масс-спектрометрии. А если использовать данные с остаточными массами, то, вероятнее всего, будут давать результат лучше, чем поиск. Хотя, конечно, этот вопрос и требует дальнейшего исследования, а главное, адаптации существующего программного обеспечения.

БЛАГОДАРНОСТИ

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий наук на 2013-2020 годы.

ЛИТЕРАТУРА

1. *Reddy P.J., Ray S., Srivastava S.* (2015) OMICS: A Journal of Integrative Biology, **19**(5), 276-282.
2. *Lane L., Bairoch A., Beavis R.C., Deutsch E.W., Gaudet P., Lundberg E., Omenn G.S.* (2013) J. Proteome Res., **13**(1), 15-20.
3. *Nesvizhskii A.I., Aebersold R.* (2005) Molecular & Cellular Proteomics, **4**(10), 1419-1440.
4. *Завьялова М.Г., Згода В.Г., Харыбин О.Н., Николаев Е.Н.* (2014) Биомед. химия, **60**, 668-676. DOI: 10.18097/PBMC20146006668
5. *Aebersold R., Goodlett D.R.* (2001) Chemical Rev., **101**(2), 269-296.



	UPS2	ISTD	MTX	USPMTX	UPSISTD
Гр. 1	16	6	0	10	19
Гр. 2	4	0	0	1	1
Гр. 3	7	2	0	2	1
Гр. 4	2	7	0	6	4
Всего	29	15	0	19	25

Рисунок 5. Результаты обработки данных, полученных программой PepNovo+ с учётом варианта “TAG + остаточные массы”. А. Сравнение между пробами (проба MTX пропущена, так как в ней не идентифицированы белки из набора UPS2). Б. Распределение белков по группам надёжности. С. Распределение по пробам и группам надёжности.

6. *Keller A., Nesvizhskii A.I., Kolker E., Aebersold R.* (2002) Anal. Chem., **74**(20), 5383-5392.
7. *Gorshkov V., Hotta S.Y.K., Verano-Braga T., Kjeldsen F.* (2016) Proteomics, **16**(18), 2470-2479.
8. *Zhang J., Xin L., Shan B., Chen W., Xie M., Yuen D., Zhang W., Zhang Z., Lajoie G.A., Ma B.* (2012) Molecular & Cellular Proteomics, **11**(4), M111-010587.
9. *Ma B.* (2015) J. Am. Soc. Mass Spectrometry, **26**(11), 1885-1894.
10. *Frank A., Pevzner P.* (2005) Anal. Chem., **77**(4), 964-973.
11. *Muth T., Weilnböck L., Rapp E., Huber C.G., Martens L., Vaudel M., Barsnes H.* (2014) J. Proteome Res., **13**(2), 1143-1146.
12. *Скворцов В.С., Алексейчук Н.Н., Худяков Д.В., Микурова А.В., Рыбина А.В., Новикова С.Е., Тихонова О.В.* (2015) Биомед. химия, **61**, 770-776. DOI: 10.18097/PBMC20156106770
13. http://www.sigmaaldrich.com/life-science/proteomics/mass-spectrometry/ups1-and-ups2-proteomic.html (дата обращения 01.08.2017).
14. *Creasy D.M., Cottrell J.S.* (2004) Proteomics, **4**(6), 1534-1536.
15. *Bootorabi F., Jänis J., Valjakka J., Isoniemi S., Vainiotalo P., Vullo D., Supuran. C.T., Waheed A., Sly W.S., Niemelä O., Parkkila S.* (2008) BMC biochemistry, **9**(1), 32.

Поступила: 13. 06. 2017.
Принята к печати: 20. 07. 2017.

USE OF *DE NOVO* SEQUENCING FOR PROTEINS IDENTIFICATION

V.S. Skvortsov, A.V. Mikurova, A.V. Rybina

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: vladlen@ibmh.msk.su

Three *de novo* sequencing programs (Novor, PEAKS and PepNovo+) have been used for identification of 48 individual human proteins constituting the Universal Proteomics Standard Set 2 (UPS2) (“Sigma-Aldrich”, USA). Experimental data have been obtained by tandem mass spectrometry. The MS/MS was performed using pure UPS2 and UPS2 mixtures with *E. coli* extract and human plasma samples. Protein detection was based on identification of at least two peptides of 9 residues in length or one peptide containing at least 13 residues. Using these criteria 13 (Novor), 20 (PEAKS) and 11 (PepNovo+) proteins were detected in pure UPS2 sample. Protein identifications in mixed samples were comparable or worse. Better results (by ~20%) were obtained using prediction included high quality identified fragment (TAG) containing at least 7 residues and unidentified additional masses at N- and C-termini (PepNovo+). The latter approach confidently recognized mass-spectrometric artefacts (and probably PTM). Atypical mass changes missed in UNIMOD DB were found (PepNovo+) to be statistically significant at the C-terminus (+23.02, +26.04 and +27.03). Using peptides containing these modifications and milder detection threshold 41 of 48 UPS2 proteins were identified.

Key words: shotgun proteomics, *de novo* sequencing, mass spectrometry, data processing