СТЕПЕНЬ ПОКРЫТИЯ АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ПРИ ИСПОЛЬЗОВАНИИ РАЗЛИЧНЫХ МЕТОДОВ АНАЛИЗА МАСС-СПЕКТРОМЕТРИЧЕСКИХ ДАННЫХ, ПОЛУЧЕННЫХ НА МОДЕЛЬНЫХ БЕЛКАХ

А.В. Микурова*, С.Е. Новикова, В.С. Скворцов, Н.Н. Алексейчук, А.В. Рыбина, Ю.В. Мирошниченко

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; эл. почта: a.mikurova@ibmc.msk.ru

В работе проведена оценка на 5 модельных белках (CYB5A, SMAD4, RAB27B, FECH и CXXC1) степени покрытия аминокислотной последовательности в панорамной протеомике при использовании различных методов обработки данных, включающих поисковые системы (Mascot и X!Tandem) и программы *de novo* секвенирования (PEAKS, Novor и PepNovo+). Для достижения максимального результата в работе применялся мультипротеазный гидролиз (ферменты Трипсин, Lys-C, AspN и GluC) в растворе и по методике FASP. Масс-спектрометрический анализ высокого разрешения проводили с помощью гибридного масс-спектрометра Q Exactive HF в режиме положительной ионизации, родительские ионы с наибольшей интенсивностью и зарядом в диапазоне от +2 до +6, фрагментировали в режиме HCD. Всего было проведено 27 экспериментов (гидролиз каждым из 5 ферментов в растворе; 4 для FASP-протокола, по три технических повтора). При использовании параметров, ограничивающих ложные идентификации пептидов, поисковые системы и программы *de novo* секвенирования дают сходные результаты. Степень покрытия аминокислотной последовательности не меньше 40%, а в лучших случаях 80-90%. Использование программ *de novo* секвенирования позволило легко обнаружить аминокислотную замену Y12H в одном из целевых белков (CYB5A).

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

DOI: 10.18097/PBMC20176305397

ВВЕДЕНИЕ

Внедрение протеомных исследований в область персонализированной медицины требует изменения основных подходов при анализе масс-спектрометрических данных. В данном случае уже недостаточно провести только идентификацию конкретных белков по отдельным фрагментам, необходимо максимально точно представлять какие протеоформы выявляются у конкретного индивидуума, какой вариант белка обнаруживается, какие варианты аминокислотного полиморфизма (SAP; single amino-acid polymorphism) и посттрансляционных модификаций (PTM; Post-translational modification) существуют. К сожалению, таргетная протеомика, имеющая высокую чувствительность и специфичность в отношении анализируемых белков [1] не всегда может быть применима, так как часто все возможные варианты просто неизвестны. В то же время панорамная протеомика (shotgun proteomic technologies) в своём общепринятом виде, как правило, имеет дело лишь с небольшими фрагментами белков, покрывающими не очень большую часть аминокислотной последовательности. Кроме того, верификация пептидов осуществляется путём сравнения спектров (а чаще их фрагментов) с *a priory* заданной базой данных и со строго фиксированным набором возможных изменений, когда все возможные варианты изменений пептидов (SAP, PTM и др.) должны быть заданы программе до проведения поиска, что катастрофически увеличивает время обработки данных. Мы полагаем, что, с учётом существенного увеличения точности современных приборов, можно существенно улучшить результаты анализа масс-спектрометрических данных, если использовать методы *de novo* секвенирования [2]. По сути *de novo* секвенирование – это весь комплекс экспериментально-аналитических процедур, включающих подготовку проб, тандемную масс-спектрометрию (MS/MS) и анализ полученных результатов исключительно на основе знаний о том, каким образом может происходить образование вторичных ионов [3].

Целью данной работы была оценка на нескольких модельных белках, какой степени покрытия аминокислотной последовательности можно достичь, используя различные методы обработки данных (как поисковые системы, так и программы *de novo* секвенирования). Кроме того, для достижения максимального результата в работе применили мультипротеазный гидролиз [4].

МЕТОДИКА

Состав анализируемой пробы, пробоподготовка и масс-спектрометрический анализ

В работе использовали пять рекомбинантных белков, полученных из института биоорганической химии НАН Беларуси, степень чистоты и другие характеристики приведены в [5, 6]:

1. P00167 (UNIPROT ID), CYB5A (название гена), цитохром *b5*, длина аминокислотной последовательности 134 остатка [5].
2. Q13485, SMAD4, рецептор для различных SMAD-белков; регулятор транскрипции, 552 остатка [6].

* - адресат для переписки

3. O00194, RAB27B, мембраносвязанный белок, предположительно, связанный с ориентацией уроплакинов, 218 остатков [6].

4. P22830, FECH, белок катализирующий присоединения гема к протопорфирину IX, 423 остатка [6].

5. Q9P0U4, CXXC1, регулятор генной экспрессии, 656 остатков [6].

Общая масса пробы составила 100 нг, по 20 нг каждого белка. Гидролиз проводился как в растворе, так и по методике FASP (Filter Aided Sample Preparation) [7] с использованием фильтра Microcon 10 (“Millipore”, США). В обоих случаях условия эксперимента при восстановлении и алкилировании совпадали. Для восстановления проводили инкубацию с 10 мМ DTT в 100 мМ Tris HCl, pH 8,5, в течение 40 мин при температуре 60°C. Для алкилирования: инкубацию с 40 мМ хлорацетамида (CAA) в темноте при комнатной температуре в течение 60 мин. Для гидролиза в растворе после восстановления и алкилирования каждый образец был разведён в 5 раз гидролитическим буфером (табл. 1). После этого добавляли протеазы. Для каждой протеазы инкубация проводилась в течение 12 ч при температуре 37°C (AspN, Lys-C, Трипсин) и 20°C (GluC).

Для FASP протокола использовали то же соотношение фермент/белок (см. табл. 1) и те же гидролитические буферы. Протеазы использовались в следующей последовательности: AspN/GluC/Lys-C/Трипсин. Инкубация с каждой протеазой проводилась в течение 2 ч. После инкубации с определенным ферментом пептиды смывались с фильтров центрифугированием (11000 g, 15 мин, 20°C), фильтр промывался гидролитическим буфером, и добавлялась следующая протеаза.

Хроматографический анализ осуществляли с помощью системы UltiMate 3000 (“Dionex”, США). Смесь пептидов разделяли на реактивированной колонке обращённой фазы Zorbax C18-300SB 150 мм × 75 мкм, диаметр частиц 3,5 мкм (0,1% водный раствор муравьиной кислоты, раствор A; “Agilent Technologies”, США), в градиенте подвижной фазы (80% раствор ацетонитрила в 0,1% муравьиной кислоте, раствор B) при скорости потока 0,4 мкл/мин. Использовали следующий градиент раствора B: 5%-15% B за 23 мин, 15%-60% B за 25 мин и 60%-99% B за 5 мин.

Масс-спектрометрический анализ высокого разрешения проводили с помощью гибридного масс-спектрометра Q Exactive HF (“Thermo Scientific”, США) в режиме положительной ионизации, в диапазоне сканирования родительских ионов от 400 до 1500 *m/z* с разрешением 60000. AGC показатель для MC-скана составил 10^6 родительских ионов при максимальном времени накопления в 50 мс. 20 наиболее интенсивных родительских ионов с зарядом в диапазоне от +2 до +6 с интенсивностью, превышающей 20000 условных единиц, фрагментировали в режиме HCD с окном изоляции 1,0 *m/z* и нормализованной энергией соударения 25. Дочерние ионы (MC/MC) сканировали с разрешением 15000. AGC показатель для MC/MC-скана составил 10^5 родительских ионов при максимальном времени накопления в 100 мс. Внутренняя калибровка осуществлялась относительно фиксированной массы *m/z* = 445,12008.

Таким образом, с учётом выполнения трёх технических повторов было проведено 27 экспериментов (гидролиз каждым из 5 ферментов в растворе; 4 для FASP-протокола, так как пробы анализировалась после каждого шага).

Анализ масс-спектрометрических данных

Анализ масс-спектрометрических данных был проведён как с использованием поисковых машин, так и программ *de novo* секвенирования.

Поиск белков в программах Mascot (“Matrix Science”, Великобритания, version 2.4.1; www.matrixscience.com) и X!Tandem (GUI версия 3.2.5 [8]) проводили по базе данных (БД), включающей последовательности белков из базы данных Swiss-Prot (версия 13.02.2017), для видов: *Homo sapiens*, *Arabidopsis thaliana*, *Escherichia coli*, *Lysobacter enzymogenes* (Q7M135), *Pseudomonas fragi* (Q9R4J4), *Staphylococcus aureus* (P0C1U8, Q2FZL2, Q99V45, Q5HH35, Q7A6A6, Q8NX98, Q6GI34, Q6GAG4), *Sus scrofa* (P00761). Состав базы данных для поиска определялся несколькими факторами. Целевые белки принадлежат человеку. Кроме того, было известно, что до реактивации хроматографическая колонка использовалась для исследования плазмы крови человека, таким образом было нужно также оценить наличие примесей. С этой же целью в БД были добавлены белки *Escherichia coli* (так как все целевые белки рекомбинантные, и заявленная степень очистки >95% [6] предполагает наличие примесей). В БД были добавлены последовательности ферментов, использованных в работе. В свою очередь, белки *Arabidopsis thaliana* были включены для оценки “степени фантазии” используемых программ.

В программах поиска устанавливали следующие параметры: расщепляющий фермент (в зависимости от протокола пробоподготовки) – Трипсин/Lys-C/Asp-N/Glu-C-V8E/Glu-C-V8DE (для X!Tandem только Трипсин и Glu-C), либо “фермент не определён”

Таблица 1. Состав гидролитического буфера

Фермент:	Трипсин	Lys-C	AspN	GluC (V8E)	GluC (V8DE)
Соотношение Фермент/Белок:	1/10	1/10	1/10	1/10	1/10
Гидролитический буфер:	40 мМ тетраэтиламмония бикарбонат (ТЭАБ), pH=8,5	40 мМ тетраэтиламмония бикарбонат (ТЭАБ), pH=8,5	40 мМ тетраэтиламмония бикарбонат (ТЭАБ), pH=8,5	40 мМ тетраэтиламмония бикарбонат (ТЭАБ), pH=8,5	фосфатный буфер (PBS), pH=7,4

(в последнем случае результат рассматривался независимо от первого варианта); точность совпадения теоретической и экспериментальной массы пептида (peptide tolerance) – 1 м.д., 5 м.д. и 10 м.д. (ppm); точность совпадения теоретической и экспериментальной масс фрагментарных ионов (MS/MS-tolerance) – 0,02 Да, 0,1 Да и 0,5 Да; количество возможных пропущенных участков расщепления ферментом – 2; фиксированная модификация (fixed modifications) – карбамидометилирование цистеина (cysteine carbamidomethylation); вариабельная модификация (variable modifications) – окисленный метионин (methionine oxidation). Параллельно проводили поиск по базе данных реверсных последовательностей аминокислот (decoy) с максимальным порогом FDR 1%. Варьирование точности совпадения было проведено для подбора оптимальных параметров. Данные по результатам, кроме особо оговоренных случаев, приводятся для значений "peptide tolerance" 5 м.д. и "MS/MS-tolerance" 0,02 Да.

Программы *de novo* секвенирования, использованные в работе: PEAKS [9], Novor [10] и PepNovo+ [11]. Две последние в составе общедоступного ПО DeNovo GUI версия 1.15.3 [12]. Для работы всех трёх программ был использован набор параметров, сходный с теми, что применяли при поиске: фермент гидролиза, соответствующий эксперименту (если имелась возможность) и "фермент не определён", точность детекции первичного иона 5 м.д. (ppm), точность определения пиков вторичных ионов 0,02 Да. Как и в случае поиска для подбора оптимальных параметров последние два параметра варьировались: 1 м.д., 5 м.д. и 10 м.д. и 0,02 Да, 0,1 Да и 0,5 Да, соответственно.

Анализ идентификации и степени покрытия последовательности предсказанными или найденными пептидами выполняли при помощи программы ProteoCat [13]. Использовали 2 варианта расчёта покрытия. Первый, "как есть", то есть накладывался целиком идентифицированный пептид, ограничение – длина пептида не менее девяти аминокислотных остатков. Второй предусматривал пересечение двух и более пептидов разной длины, при этом фрагмент, не входящий в пересекающуюся часть, отбрасывался, минимальная длина общей части также равнялась девяти остаткам. Второй вариант, по сути, неинформативен для результатов поиска (найденные пептиды заведомо совпадают с фрагментами последовательности целевых белков), но для результатов *de novo* секвенирования служит дополнительным критерием достоверности.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В результате масс-спектрометрического эксперимента в общей сложности было получено 425000 спектров МС/МС, которые были проанализированы с применением поисковых систем и программ *de novo* секвенирования.

Первым результатом, который необходимо обозначить, было фактическое отсутствие пептидов для белка CXXC1. Единственно, что было показано, одиночный пептид, выявленный поиском X!Tandem, длиной в 43 аминокислотных остатка (при установке "фермент не определён). Так как в достоверности этой идентификации имеются огромные сомнения (о чём будет написано ниже), то при дальнейшем изложении будем считать, что либо белок "испортился" при хранении, либо имела место наша ошибка при подготовке пробы.

Второй важный вопрос – подбор оптимальных параметров поиска и/или процедур *de novo* секвенирования. Несомненно, основным определяющим фактором должны быть номинальные характеристики точности используемых приборов (или анализ исходных данных). Выборочные данные представлены в таблицах 2 и 3, полностью таблицы приведены в дополнительных материалах (см. приложение). Основной вывод достаточно тривиален, но, тем не менее, важен – если для программ поиска загрубение параметров точности увеличивает степень покрытия и, таким образом, как бы "улучшает" результат,

Таблица 2. Сравнение покрытия аминокислотной последовательности при различных установках "точности совпадения" (фрагмент для результатов, полученных программой PEAKS)

		O00194	P00167	P22830	Q13485
Общая длина последовательности (по Uniprot)		218	134	423	552
Вариант «с пересечением»					
«MS/MS-tolerance», Да («peptide tolerance» 5 м.д.)	0,5	99	44	103	60
	0,1	108	44	105	69
	0,02	128	50	114	107
«peptide tolerance», м.д. («MS/MS-tolerance», 0,02 Да)	10	112	47	116	84
	5	128	50	114	107
	1	92	46	112	62
Вариант «как есть»					
«MS/MS-tolerance», Да («peptide tolerance» 5 м.д.)	0,5	130	57	141	119
	0,1	128	58	146	97
	0,02	134	56	145	120
«peptide tolerance», м.д. («MS/MS-tolerance», 0,02 Да)	10	131	56	148	112
	5	134	56	145	120
	1	137	56	150	119

Таблица 3. Сравнение покрытия последовательности при различных установках “фермента гидролиза” (параметры точности: 5 м.д., 0,02 Да)

		Вариант «с пересечением»				Вариант «как есть»			
		O00194	P00167	P22830	Q13485	O00194	P00167	P22830	Q13485
Длина последовательности		218	134	423	552	218	134	423	552
NOVOR	Фермент указан точно	108	46	81	72	111	48	105	96
	Указан трипсин	86	48	86	68	119	57	126	112
	Фермент не определён	128	50	114	107	134	56	145	120
PEAKS	Фермент указан точно	152	57	81	68	186	54	188	184
	Указан трипсин	129	57	125	74	165	70	193	184
	Фермент не определён	174	86	126	123	194	80	206	240
MASCOT	Фермент указан точно	113	68	174	96	190	86	307	208
	Указан трипсин	120	52	147	140	162	70	292	319
	Фермент не определён	159	116	256	508	210	125	384	522
X!TANDEM	Фермент указан точно	197	83	281	325	210	123	307	473
	Указан трипсин	190	74	266	330	210	101	316	404
	Фермент не определён	197	92	277	440	207	123	319	524

то для программ *de novo* секвенирования это либо не приводит к существенным улучшениям, либо наоборот приводит к тому, что программы начинают фантазировать и результат ухудшается. Последнее особенно заметно на варианте с пересечениями, когда оценивается только покрытие более достоверным фрагментом, имеющим дополнительное подтверждение. Выбор параметра отсечения по величине оценочных функций программ *de novo* секвенирования был сделан в пользу ALC >=50% для PEAKS, “Novor score” >=70 для Novor (см. приложение) и pscore >0 для PepNovo+. Уменьшение порогов не даёт значительного выигрыша для целевых белков, в то же время существенно увеличивает число обнаружения примесей и ложных идентификаций [2].

Собственно, результаты анализа степени покрытия последовательности для целевых белков неоднозначны (рис. 1). Казалось бы, при использовании процедур поиска в ряде вариантов степень покрытия больше 90%, в то время как программы *de novo* секвенирования за одним исключением (O00194, PEAKS) не показывают результат больше 60%. Но так ли это на самом деле? Рассмотрим полученные результаты на примере цитохрома *b5* (рис. 2). Первый вывод очевиден: при использовании поисковых машин с установкой “фермент не определён” выявленные пептиды в ряде случаев недостоверны, так как их появление не может быть объяснено имеющимися условиями гидролиза белка (даже если предположить наличие примеси химотрипсина в трипсине, на что, например, указывает гидролиз между F и L), а спонтанный гидролиз вряд ли бы привёл к накоплению достаточного для детекции количества конкретного пептида. Таким образом, имеется только 1 вариант, когда предположительно был найден пептид, соответствующий мембран-связанной части цитохрома *b5*. Данный пептид (DSSSSWWTNWVIPAISAVAVALMYR) состоит из 25 аминокислотных остатков и соотношение *m/z* для двухзарядного иона (единственного, что могли зарегистрировать) находится пусть и близко к верхней границе, но в пределах диапазона сканирования. Остальные пептиды в этой части последовательности (начиная с 90 остатка) вероятнее всего следует признать ложными идентификациями. Так как если бы имел место неспецифический гидролиз, то они бы выявлялись и при поиске, и при *de novo* секвенировании. Существует вариант анализа, при котором часть данного фрагмента может быть выявлена. Это использование программы PepNovo+ в варианте, когда рассматривается только достоверный фрагмент и остаточные массы с N- и/или C-конца [2]. В таком случае можно выделить фрагмент “AVALMYR” и N-концевую массу, соответствующую либо паре остатков Val-Ala (в произвольном порядке), либо Gly-Leu. Однако, наличие данного пептида также нельзя объяснить имеющимися условиями гидролиза. В то же время, если сам по себе фрагмент “AVAVALMYR” (по поиску на https://blast.ncbi.nlm.nih.gov) встречается фактически только у цитохрома *b5*, то мотивы близкие к нему, а особенно к мотивам “GLAVALMYR” и “LGAVALMYR”, нередко встречаются у *Escherichia*. Строго говоря, результаты поиска с использованием X!Tandem мало различаются при различных установках для фермента гидролиза, в то время как при явном определении фермента пептиды, найденные поисковой системой Mascot, соответствуют ожидаемым по протоколу гидролиза. Программы *de novo* секвенирования также выявляют пептиды,

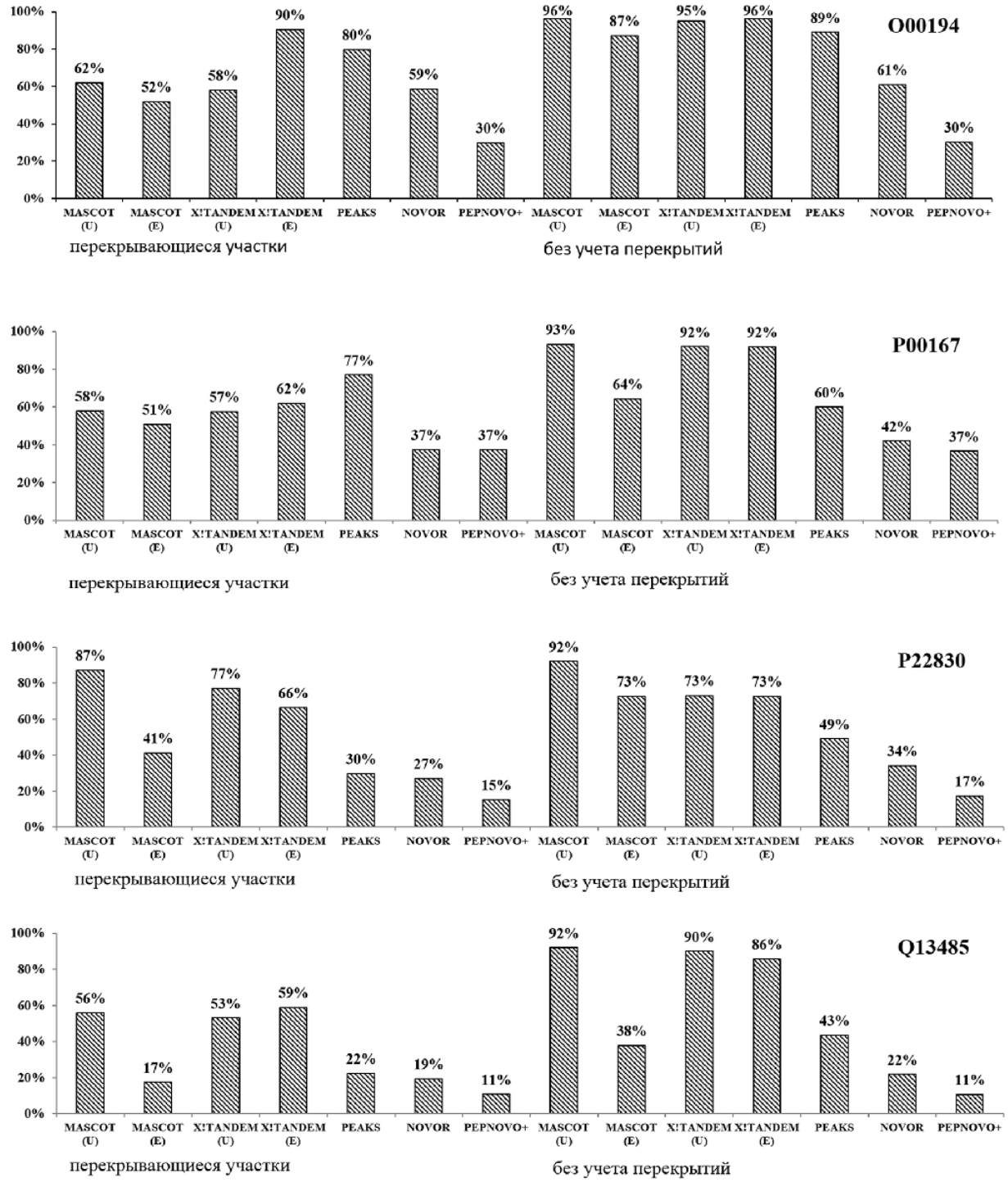

Рисунок 1. Сравнение покрытия последовательностей при идентификации спектров различными методами.

соответствующие протоколу гидролиза, и, в общем случае, дают результаты по степени покрытия последовательности сопоставимые (рис. 1) с идентификациями Mascot при заданном ферменте гидролиза (то есть при отсутствии нереальных пептидов). К тому же у программ *de novo* секвенирования есть одно преимущество. Несомненно, поисковая машина Mascot обнаружила бы найденную мутацию в цитохроме *b5* (Y12H, рис. 1), если бы в установках был задан поиск SAP. Однако, это бы многократно увеличило время обработки результатов. В то же время, для программ *de novo* секвенирования это не увеличивает время работы, необходимо только иметь возможность отслеживать это событие при конечном сравнении аминокислотных последовательностей идентифицированных пептидов и целевых белков. Данная возможность имеется в программе ProteoCat. Мы не ставили себе задачи найти возможные мутации, и обнаружили её фактически случайно при анализе результатов идентификации. Авторы [5], у которых был получен цитохром *b5*, подтвердили наличие мутации.

```
          1.......10.#.....20........30........40........50........60........70

P00167    MAEQSDEAVKYHTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
X!Ta-E    -AEQSDEAVK----EEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
X!Ta-U    -AEQSDEAVK----EEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
Masc-E    -AEQSDEAVK--------------STWLILHHK--DLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
Masc-U    -AEQSDEAVK--TLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
PAEKS1    --------------EELQKHNHSKSTWLLLHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
PAEKS2    -----DEAVKYHTLEELQK-----STWLLLHHK--DLTKFLEEHPGGEEVLREQAGGDATENFEDVGHST
PepN+1    ------------------------------------FLEEHPGGEEVLR-----DATENFEDVGHST
PepN+2    -----DEAVKYHTLEELQKHNHSK----------***KFLEEHPGGEEVLR*****DATENFEDVGHST

          ........80........90.......100.......110.......120.......130....
P00167    DAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISAVAVALMYRLYMAED
X!Ta-E    DAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISAVAVALMYR------
X!Ta-U    DAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISAVAVALMYR------
Masc-E    DAREMSKTFIIGELHPDDRPKLNKPPETLITTI--------------------------------
Masc-U    DAREMSKTFIIGELHPDDRPKLNKPPETLITTIDSSSSWWTNWVIPAISAVAVALMYR------
PAEKS1    DAREMSKTFLLGELHPDDRPK---PPETLLTTL--------------------------------
PAEKS2    DAREMSKTFLLGELHPDDRPKLNKPPE----------------------AVAVALMYR------
PepN+1    DAR-----------------------------------------------AVAVALMYR------
PepN+2    *******-----**HPDDRPK-------------------------------------------
```

Рисунок 2. Покрытие аминокислотной последовательности цитохрома *b5* пептидами, полученными в результате различных методов обработки данных масс-спектрометрического анализа. Знаком # помечена выявленная замена Y12H. Жирным шрифтом выделены остатки с N-конца идентифицированных пептидов. Подчерк - выделены остатки с C-конца идентифицированных пептидов. * - помечены позиции, для которых совпадают остаточные массы, но точно аминокислотные остатки не разрешены.

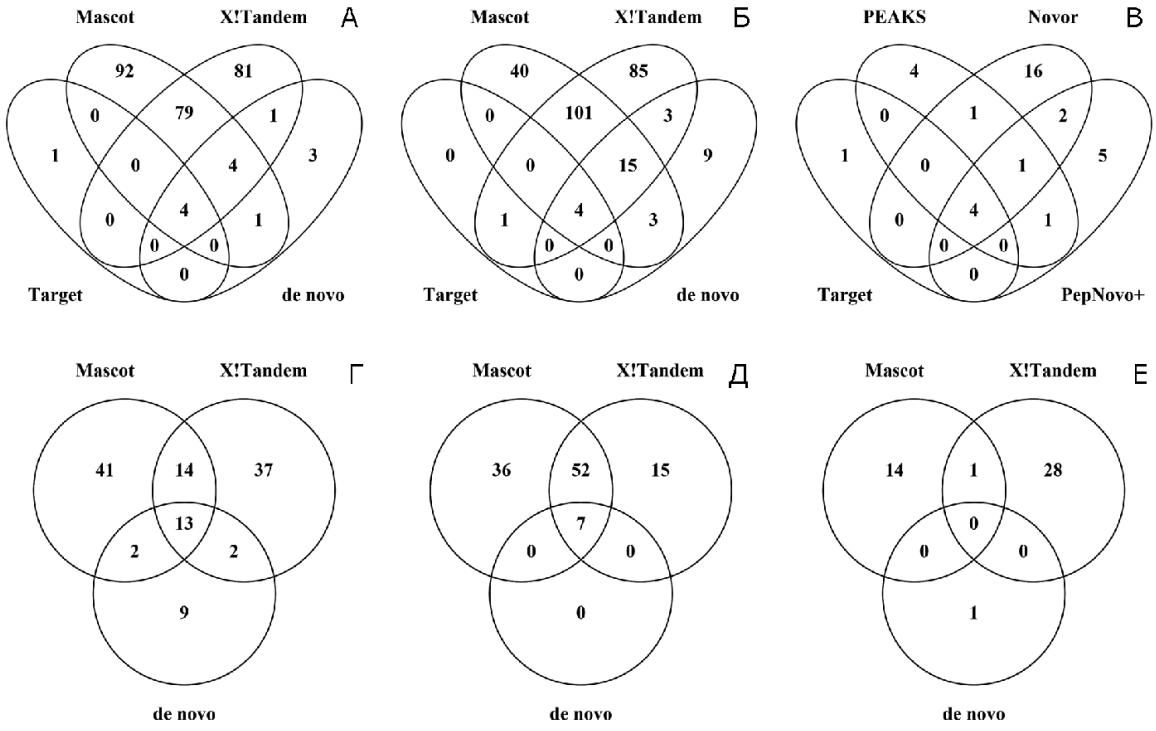

Рисунок 3. Сравнение количества белков, идентифицированных при различных методах анализа масс-спектрометрических данных. А. Все идентификации, при условии, что фермент задан явно, результаты работы программ PEAKS и Novor объединены (PepNovo+ не поддерживает данную опцию). Target - целевые белки. Б. Все идентификации, при условии, что “фермент не определён”, результаты работы программ PEAKS, Novor и PepNovo+ объединены. В. Белки, идентифицированные программами *de novo* секвенирования (“фермент не определён”). Г, Д, Е. Сравнение идентифицированных белков по видовой принадлежности (для программ поиска “фермент задан явно”, данные *de novo* секвенирования объединены и “фермент не определён”).

И, наконец, несколько слов об обнаружении примесей (рис. 3, приложение). И программы поиска, и программы *de novo* секвенирования нашли в пробе определённое количество примесей. В ряде случаев, например, как в случае белка Rab-27A (P51159) или цитохрома *b5* из *Arabidopsis thaliana* (O48845), это, несомненно, не реальные идентификации, а паразитное выявление гомологов. При пофрагментном анализе средствами программы ProteoCat легко обнаружить, что все те же фрагменты входят

и в последовательность целевых белков. Так как имел место последовательный гидролиз, то наличие фрагментов протеаз также ожидаемо. Странно, что только детектирование AspN подтверждается как поиском, так и *de novo* секвенированием. Учитывая, что данный фермент первый в последовательности FASP, его обнаружение наиболее вероятно, но и второй в последовательности эксперимента фермент (GluC) должен бы быть обнаружен, да и трипсин тоже [14] (при поиске они идентифицируется). Общая тенденция, что белки *Homo sapiens*, *Escherichia coli* и *Arabidopsis thaliana* идентифицируются в большей степени при поиске, чем при *de novo* секвенировании (уточним, что все значимые идентификации для *Arabidopsis thaliana* фактически могут быть объяснены наличием гомологов). С одной стороны, это можно бы было объяснить большей чувствительностью поисковых машин при заданных параметрах. С другой, наличие в списке цитохромов P450 или дермцидина, которые вряд ли могли бы оказаться в крови в достаточном количестве, позволяет предположить или ошибку поисковых машин, или попадание в пробу загрязнений из внешних источников. Учитывая, что при режиме "фермент не определён" поисковые машины почти наверняка ошибаются, то можно сделать вывод об их излишней оптимистичности. В то же время, среди белков, идентифицированных программами *de novo* секвенирования, нет таких, наличие которых нельзя бы было объяснить.

ЗАКЛЮЧЕНИЕ

В конечном итоге можно заключить, что если не рассматривать завышенный результат, полученный поисковой системой X!Tandem (в случае которой имеются определённые подозрения, что обнаруженные ею пептиды нереальные), то лучшая из программ *de novo* секвенирования PEAKS даёт вполне сравнимые результаты по покрытию аминокислотной последовательности с поисковой системой Mascot. При этом выявление SAP (и вероятнее всего PTM, которых просто не было в целевых белках) происходит при анализе без существенного увеличения временных затрат, имеющих место при аналогичных ситуациях в поисковых системах. Ещё одним недостатком поисковых машин является их излишняя "оптимистичность", конечно, примеси в пробе (и колонке) присутствовали, но явно не в таких количествах, чтобы идентифицировать более 200 белков (X!Tandem). Но и программы *de novo* секвенирования не лишены недостатков, и главный из них (призванный сделать их более удобными): они начинают фантазировать, варьируя аминокислотные остатки, если сумма масс совпадает, но порядок неизвестен. Конечно, это находит отражение в величине оценочной функции, но требует сохранения большого числа вариантов и волей-неволей может быть источником "подгонки" результата. Выходом может быть использование только достоверно разрешённых фрагментов пептида и остаточных масс для остальной его части [2].

БЛАГОДАРНОСТИ

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий наук на 2013-2020 годы. Экспериментальная часть и поиск Mascot выполнены с использованием оборудования ЦКП "Протеом человека", поддержанного Минобрнауки России в рамках выполнения соглашения №14.621.21.0017 (уникальный идентификатор проекта RFMEFI62117X0017).

ЛИТЕРАТУРА

1. *Nesvizhskii A.I., Aebersold R.* (2005) Mol. Cell. Proteomics, **4**(10), 1419-1440.
2. *Скворцов В.С., Микурова А.В., Рыбина А.В.* (2017) Биомед. химия, **63**, 341-350. DOI: 10.18097/PBMC20176304341
3. *Aebersold R., Goodlett D.R.* (2001) Chem. Rev., **101**(2), 269-296.
4. *Ni W., Lin M., Salinas P., Savickas P., Wu S.L., Karger B.L.* (2013) J. Am. Soc. Mass Spectrometry, **24**(1), 125-133.
5. *Ershov P., Mezentsev Y., Gnedenko O., Mukha D., Yantsevich A., Britikov V., Kaluzhskiy L., Yablokov E., Molnar A., Ivanov A., Lisitsa A., Gilep A., Usanov S.* (2012) Proteomics, **12**(22), 3295-3298.
6. *Иванов А.С., Ершов П.В., Мольнар А.А., Мезенцев Ю.В., Калужский Л.А., Яблоков Е.О., Флоринская А.В., Гнеденко О.В., Медведев А.Е., Козин С.А., Митькевич В.А., Макаров А.А., Гилеп А.А., Лущик А.Я., Гайдукевич И.В., Усанов С.А.* (2016) Биоорган. химия, **42** (1), 18-27.
7. https://www.biochem.mpg.de/226356/FASP
8. *Vaudel M., Barsnes H., Berven F.S., Sickmann A., Martens L.* (2011) Proteomics, **11**(5), 996-999.
9. *Zhang J., Xin L., Shan B., Chen W., Xie M., Yuen D., Zhang W., Zhang Z., Lajoie G.A., Ma B.* (2012) Mol. Cell. Proteomics, **11**(4), M111-010587. DOI: 10.1074/mcp.M111.010587
10. *Ma B.* (2015) J. Am. Soc. Mass Spectrometry, **26**(11), 1885-1894.
11. *Frank A., Pevzner P.* (2005) Anal. Chem., **77**(4), 964-973.
12. *Muth T., Weilnböck L., Rapp E., Huber C.G., Martens L., Vaudel M., Barsnes H.* (2014) J. Proteome Res., **13**(2), 1143-1146.
13. *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**(6), 770-776. DOI: 10.18097/PBMC20156106770
14. *Karty J.A., Ireland M.M., Brun Y.V., Reilly J.P.* (2002) J. Chromatogr. B, **782**(1), 363-383.

Поступила: 31. 08. 2017.
Принята к печати: 29. 09. 2017.

THE SEQUENCE COVERAGE IN DIFFERENT METHODS OF MASS SPECTROMETRY DATA ANALYSIS OBTAINED ON MODEL PROTEINS

A.V. Mikurova, S.E. Novikova, V.S. Skvortsov, N.N. Alekseychuk, A.V. Rybina, Yu.V. Miroshnichenko

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: a.mikurova@ibmc.msk.ru

The aim of this study was to evaluate sequence coverage of five model proteins (CYB5A, SMAD4, RAB27B, FECH, and CXXC1) by means of shotgun proteomic data analysis employing different methods of data treatment including database-dependent search engines (MASCOT and X!Tandem) and de novo sequencing software ((PEAKS, Novor, and PepNovo+). In order to achieve maximal results, multiprotease hydrolysis including enzymes trypsin, LYS-C, ASPN and GluC was performed in solution and using the FASP method. High resolution mass spectrometry was carried out with a Q EXACTIVE HF hybrid mass spectrometer in the positive ionization mode; parent ions with the highest intensity and a charge range from +2 to +6 were fragmented in the HCD mode. 27 experiments were carried out (hydrolysis with each of 5 enzymes in solution, 4 for the FASP protocol, three technical repeats). Using parameters limiting false identification of peptides, the search engines and *de novo* sequencing software gave similar results. The degree of sequence coverage was not at least 40%, and in the best cases it reached 80-90%. The use of *de novo* sequencing software resulted in identification of the Y12H amino acid substitution in one model protein (CYB5A).

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