Биомедицинская химия, 2015 том 61, вып. 6, с. 770-776.

УДК 577.332

ПРОГРАММА ProteoCat КАК ИНСТРУМЕНТ ПЛАНИРОВАНИЯ ПРОТЕОМНОГО ЭКСПЕРИМЕНТА

В.С. Скворцов*, Н.Н. Алексейчук, Д.В. Худяков, А.В. Микурова, А.В. Рыбина, С.Е. Новикова, О.В. Тихонова

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

Программа ProteoCat предназначена для помощи исследователям при планировании широкомасштабных протеомных экспериментов. Центральной частью программы является модуль симуляции гидролиза, поддерживающий 4 протеазы (трипсин, лизин С, эндопротеиназы AspN и GluC). Для полученных в результате виртуального гидролиза или загруженных пептидов рассчитывается или предсказывается ряд важных для масс-спектрометрического эксперимента свойств, по которым данные могут быть проанализированы и отфильтрованы. Для предсказания pI и оценки вероятности детекции пептида в программе использованы новые улучшенные модификации описанных ранее собственных методов; pI также может быть предсказано по ряду популярных шкал pKa, предложенных другими исследователями. Предсказание времени удержания пептида реализовано в виде алгоритма аналогичного используемому в программе SSRCalc. Используя ProteoCat можно оценить степень покрытия аминокислотной последовательности анализируемых белков при заданных ограничениях на детекцию пептидов, а также возможность сборки протяжённых пептидов из фрагментов с установленными пользователем минимальными размерами “липких” концов. Программа имеет графический интерфейс, написана на языке JAVA и доступна по адресу http://www.ibmc.msk.ru/LPCIT/ProteoCat.

Ключевые слова: масс-спектрометрия, предсказание свойств пептидов, гидролиз, покрытие последовательности белка

DOI: 10.18097/PBMC20156106770

ВВЕДЕНИЕ

Количественный анализ экспрессии белков в клетках – одна из важнейших задач современной протеомики. Общепринятый подход для решения данной задачи – широко используемый в тандемной масс-спектрометрии метод направленного анализа (SRM – Selected Reaction Monitoring) [1]. Разделение сложных белковых или пептидных смесей на фракции в масс-спектрометрии используют: (а) жидкостную хроматографию (ЖХ), и тогда пептид (белок) характеризуется величиной “время удержания” (RT), (б) фракционирование с помощью изоэлектрического фокусирования, результат которого определяется величиной изоэлектрической точки пептида (pI). На масс-спектре пептид характеризуется соотношением массы к заряду (*m/z*) иона и интенсивностью пика. Важной задачей считается также выбор пептида, который должен быть, с одной стороны уникальным для исследуемого белка, а с другой, – обладать набором свойств, которые бы позволили масс-спектрометру определить (детектировать) данный пептид. Такие пептиды обозначают термином “протеотипические” пептиды. Следует подчеркнуть, что именно детектируемость пептида масс-спектрометром является решающим фактором для идентификации и определения количества белка [2]. Естественно, все эти свойства пептидов (белков) предсказываются большим количеством компьютерных программ, даже такое мало формализованное свойство как “протеотипичность” [3].

Другая актуальная задача протеомики – поиск существующих посттрансляционных модификаций (PTM), единичных аминокислотных замен и наличия альтернативного сплайсинга, что важно для выявления маркеров заболеваний и анализа причинно-следственных связей между модификациями белков и различными патологиями. В данном случае, идентификации нескольких небольших “протеотипичных” пептидов недостаточно, и требуется существенно большее покрытие аминокислотной последовательности белка пептидами. В какой-то

* - адресат для переписки

степени, может помочь использование нескольких протеаз с различной специфичностью, что расширяет возможности общепринятого подхода: гидролиз белковой смеси трипсином и последующая панорамная масс-спектрометрия [4]. Использование гидролиза множественными протеазами (MELD), а также варьирование условий гидролиза, увеличивает разнообразие пептидов в смеси. Кроме того, данный подход даёт возможность собирать более протяжённые фрагменты за счёт наличия пересекающихся частей достаточной длины (исключает или существенно снижает вероятность случайной "склейки" пептидов), увеличивая вероятность собрать белок полностью [5].

Для того чтобы обобщить различные методы предсказания свойств пептидов, включая наши собственные, а также дать возможность анализировать возможные варианты результатов эксперимента с целью более эффективного планирования, была создана программа ProteoCat, описанная ниже.

ОПИСАНИЕ ПРОГРАММЫ. МЕТОДЫ, АЛГОРИТМЫ И ПРИМЕРЫ ИСПОЛЬЗОВАНИЯ

Программа ProteoCat содержит несколько вычислительных блоков, часть из которых предсказывает различные свойства пептидов, а часть – имитирует экспериментальные процедуры и методы обработки результатов. Сначала остановимся на первой группе и опишем отличия применяемых методов от описанных ранее.

Предсказание величины pI пептидов

Данный модуль основан на созданной нами ранее программе pIPredict [6], позволяющей предсказать значение изоэлектрической точки набором некоторых методов. В основе большинства этих методов лежит уравнение Хендерсона-Хассельбаха и варьируются только шкалы pKa диссоциируемых групп. Используется также и нейросетевая модель, но она не включена в программу ProteoCat. Из имеющихся шкал следует отметить шкалу, взятую из работы Bjellkvist и соавторы [7] (используется в самом популярном сервисе предсказания pI портала биоинформатических ресурсов http://web.expasy.org/compute_pi), а также разработанную нами шкалу [6]. В ProteoCat используется уточнённая вторая версия шкалы, рассчитанная на увеличенной обучающей выборке пептидов (>38000), которая объединяла в себе обучающую выборку первой версии и обучающую выборку для немодифицированных пептидов программы PredpI [8]. Новая версия незначительно увеличивает R^2 обучения и тестирования на независимой выборке из 1700 пептидов [9], однако существенно уменьшает среднюю ошибку предсказания. Следует подчеркнуть, что при определении pI в обучающей и тестовой выборках имеет место априорная ошибка, связанная с особенностями метода фракционирования с помощью изоэлектрического фокусирования. Например, в выборке pIPredict она составляет 0,15 значений pH, в тестовой выборке – 0,19. В выборке PredpI ошибка кажется очень малой – сотые доли, однако выборка характеризует очень узкий диапазон pH (3,58 – 4,83). С этим, вероятно, и связан тот факт, что на нашей тестовой выборке программа PredpI показывает R^2 предсказания равный 0,858, против 0,944 (рис. 1), показанный нашей шкалой. Метод Bjellkvist на данной выборке даёт формально несколько худший результат (R^2 = 0,92), но качество предсказания метода Bjellkvist и шкалы pIPredict сопоставимы и отличаются только тем, что число "больших" ошибок у метода Bjellkvist больше (рис. 2). Кроме того, шкала pIPredict позволяет рассчитывать pI при наличии ряда PTM, например, фосфорилирования.

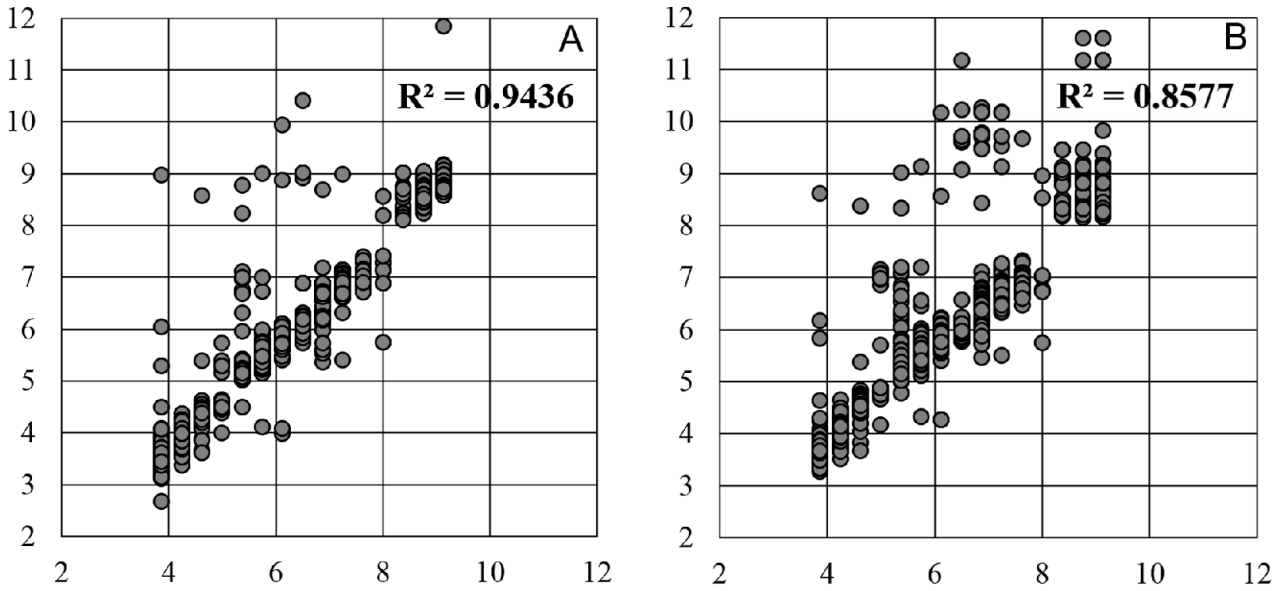

Рисунок 1. Сравнение предсказанных значений pI для тестовой выборки методами pIPredict версия 2 (A) и PredpI (B).

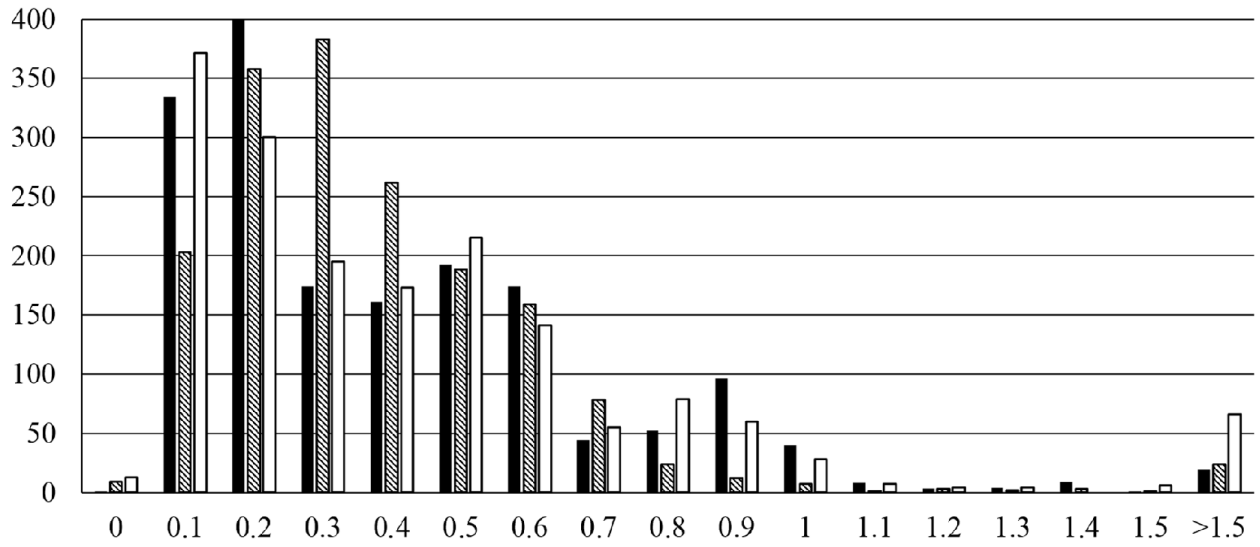

Рисунок 2. Распределение ошибки предсказания значений pI для тестовой выборки методами Bjellqvist и соавт., (чёрный цвет), pIPredict версия 2 (косой штрих) и PredpI (белый цвет).

Предсказание времени удержания пептида

Для предсказания величины RT был реализован алгоритм, описанный Krokhin в 2006 году [10]. Данный алгоритм использовался в программе SSRCalc (http://hs2.proteome.ca/SSRCalc/SSRCalcQ.html) в одной из предыдущих версий. К настоящему времени, методика расчёта SSRCalc усовершенствована и доступна в виде сервиса по вышеприведённому адресу. Однако, и в нашем варианте точность предсказания достаточно хорошая (рис. 3), а, учитывая, что в программе ProteoCat планируется делать предсказания для модифицированных пептидов, требовался собственный модуль расчёта, пригодный для модификаций. Метод рассматривает величину RT как линейно зависимую от величины "идеальной гидрофобности" (HI):

$$RT = a + b \cdot HI,$$

где a и b – коэффициенты фитирующей (выравнивающей) функции, определяемые особенностями оборудования, реактивов и условий жидкостной хроматографии. Расчёт HI проводится на основе последовательности пептида в 9 шагов. Первый – аддитивная схема, суммирующая коэффициенты удерживания отдельных аминокислот (учитывается также положение аминокислотного остатка в цепи), 8 последующих – корректирующие функции. К сожалению, не все корректирующие функции были описаны достаточно подробно, так что их пришлось изъять из алгоритма. Так, не используются коррекция при наличии гидрофобных факторов (п. 3 в [10]), коррекция по величины изоэлектрической точки (п. 5)

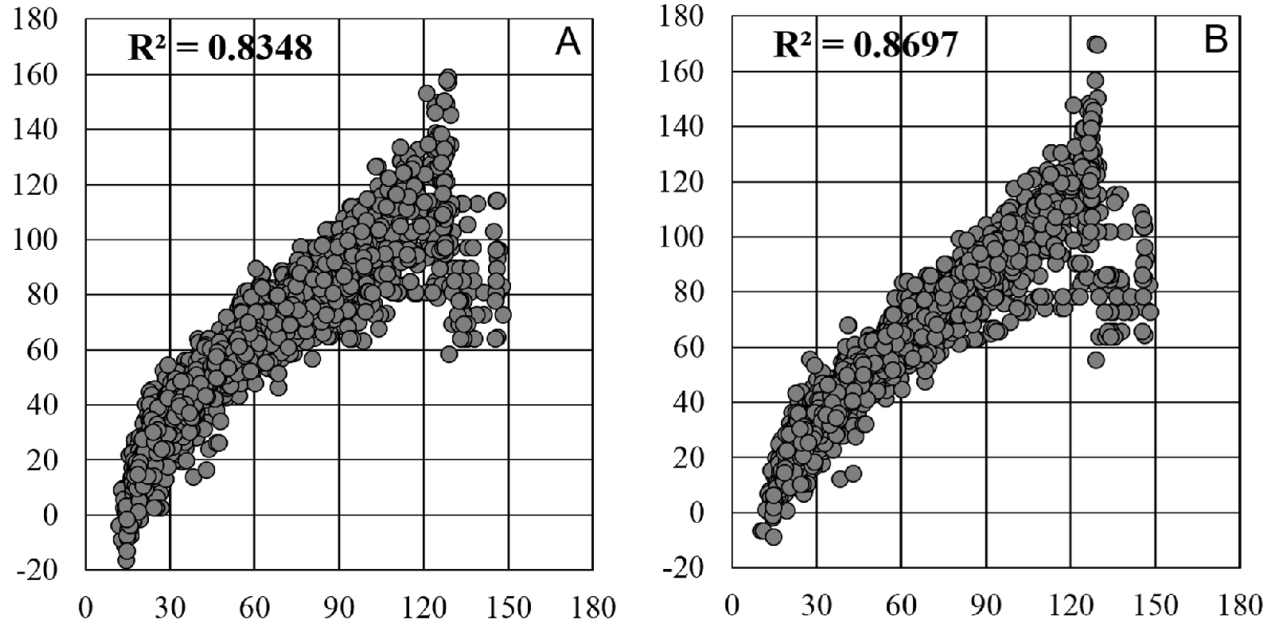

Рисунок 3. Сравнение предсказанных величин RT с экспериментально определёнными для 2750 пептидов. A. Программа ProteoCat. B. Программа SSRCalc.

и коррекция для пептидов склонных к образованию спиральных структур (п. 9). Используемая по умолчанию фитирующая функция была настроена по собственным данным идентификации 2750 пептидов из гидролизата печени мыши (трипсин). ЖХ проводили на аналитической колонке Zorbax 300SB-C18, диаметр частиц 3,5 мкм, 150 мм × 75 мкм ("Agilent Technologies", USA). Пользователь может заменить коэффициенты a и b на свои собственные в настройках программы, подстроив, таким образом, результаты предсказания под собственное оборудование.

Сравнение полученных результатов предсказания для 2750 пептидов с предсказаниями программы SSRCalc (вариант 2015 года) показало (рис. 3), что собственная реализация работает ненамного хуже современного варианта. И, хотя средняя ошибка на нашей тестовой выборке существенно меньше у программы SSRCalc (7,38 против 9,07 у ProteoCat), она всё равно слишком велика.

Оценка вероятности детекции пептида при масс-спектрометрии

В действительности, надёжная масс-спектрометрическая детекция пептида зависит от многих причин, связанных как со свойствами самого пептида, так и особенностями оборудования и условиями эксперимента. В программе ProteoCat для подобной оценки используется очень простой метод, предложенный нами ранее [11]. Его идея состоит в том, что при наличии известной зависимости интенсивности масс-спектрометрических пиков от концентрации пептидов, можно рассчитать ожидаемую интенсивность пика, а отклонение от этой величины в большую или меньшую сторону определяется аминокислотным составом пептида. Чем эта величина больше, тем больше вероятность того, что пептид будет уверенно детектироваться.

Следует подчеркнуть, что метод имеет следующие ограничения:

- Метод обучен на данных положительной электроспрейной ионизации и не применим к другим вариантам.
- Пептиды при pH=2,5 должны быть в основном представлены ионами 2^+ и 3^+.
- Метод (или, если сказать точнее, уравнение, коэффициенты которого приведены в таблице) работает только для триптических пептидов (на С-конце остатки аргинина или лизина).
- Длина пептидов в обучающей выборке была от 7 до 20 (медиана 11), так что при большей длине достоверность результата не определена. В программе длина пептидов, для которых выполняется предсказание, ограничена 30.

Несмотря на невысокую точность и наличие ограничений, модель уверенно отличает пептиды с высокой степенью ответа от пептидов с низкой степенью ответа и позволяет отобрать самые перспективные пептиды.

Для универсальности расчётов, в ProteoCat использован вариант линейной функции (таблица), отличный от варианта с наилучшими параметрами, описанного в работе [11]. Однако, параметры их близки: R^2 обучения 0,58, средняя ошибка обучения – 0,15 (диапазон значений ΔI при обучении от -1,43 до 1,26), Q^2 в процедуре скользящего контроля – 0,52, а средняя ошибка – 0,16, размер обучающей выборки – 378 пептидов. По понятным причинам остатки аргинина или лизина не влияют на качество предсказания (в обучающей выборке они имеются только на конце пептида, а различия между заканчивающимися на K или R пептидами несущественны). Учитывая, что в уравнении сохраняется концентрация пептида [11], при предсказании используется величина 0,1 пмоль/мкл. Следует также иметь в виду, что под цистеином в данном случае всегда подразумевается карбамидометил-цистеин. Впрочем, его вклад в предсказанное значение невелик.

Таблица. Коэффициенты линейного уравнения для предсказания величины ΔI, включающего аминокислотный спектр и концентрацию пептида.

Переменная	Коэффициент	Переменная	Коэффициент
Константа	нет	lg (концентрации, пмоль/мкл)	-0,329
Кол-во A	0,122	Кол-во M	0,054
Кол-во C	-0,056	Кол-во N	0,044
Кол-во D	-0,131	Кол-во P	0,163
Кол-во E	0,043	Кол-во Q	-0,09
Кол-во F	0,309	Кол-во R	нет
Кол-во G	-0,01	Кол-во S	0,002
Кол-во H	-0,018	Кол-во T	0,041
Кол-во I	0,275	Кол-во V	0,324
Кол-во K	нет	Кол-во W	-0,024
Кол-во L	0,504	Кол-во Y	0,072

Пост-трансляционные модификации и расчёт масс пептидов

Программа рассчитывает как среднеизотопные, так и моноизотопные (по умолчанию) массы пептидов и ионов. Программа понимает все основные PTM. Главное ограничение – наличие таких модификаций должно быть отражено в загружаемом файле с аминокислотными последовательностями. Гипотетических предсказаний PTM программа не делает. В отличие от PTM, химические модификации белков или пептидов, такие как, например, преобразование остатков цистеинов в карбамидометил-цистеин могут быть применены ко всем белкам и/или пептидам в списке по желанию пользователя. Расчёт pI для модифицированных пептидов (PTM) осуществляется, если соответствующая модификация полностью элиминирует возможность диссоциации у модифицированного аминокислотного остатка. Если используется шкала pIPredict, то pI также

рассчитывается в случае N-концевого метилирования и фосфорилирования. В существующей версии расчёты RT и ΔI для модифицированных пептидов не производятся.

Симуляция гидролиза различными протеазами

В качестве одной из центральных возможностей в ProteCat заложена функция симуляции процедуры гидролиза белков с использованием различных протеаз. В текущей версии – 4 протеазы: трипсин, лизин С (LysC), эндопротеиназа AspN и эндопротеиназа GluC. В общем случае трипсин расщепляет белки на пептиды с остатками лизина и аргинин на С-конце [12], лизин С – на пептиды с остатком лизина на С-конце, AspN – на пептиды с остатком аспарагина на N-конце, GluC – на пептиды с остатками глутамина и аспарагина на С-конце. Для каждого фермента имеются свои особенности в специфичности, они учтены в программе. Предусмотрена возможность как параллельного гидролиза (независимо каждым из ферментов), так и последовательного (комбинация ферментов, второй из которых гидролизует пептиды, полученные после работы первого). Полученные списки пептидов могут быть отфильтрованы по заданным параметрам: длина пептида, масса, pI, RT, ΔI (для трипсина; для лизина С, если в пептиде имеется аргинин результат не достоверен), величине *m/z* иона. Имеется также возможность отфильтровать пептиды случайным образом, чтобы имитировать реальный эксперимент (как пример, продемонстрировано на рисунке 5).

Возможности анализа результата.

Самый простой вариант применения программы ProteoCat – использование модуля симуляции гидролиза и последующую фильтрацию пептидов для анализа распределения пептидов по фракциям или специфическим группам (рис. 4), используя предсказанные и рассчитываемые характеристики пептидов.

Другой вариант связан с анализом возможного покрытия последовательности белков. В самом простом случае – это процент покрытия гидролизными пептидами после фильтрации по установленным пользователем правилам без учёта перекрытий. Более сложный – сборка из меньших фрагментов пептидов большего размера с использованием пересекающихся концов (минимум 5 остатков). Таким образом, можно предположить, насколько протяжённым будет непрерывный фрагмент белка при идеальной детекции пептидов в заданном окне масс и RT. Например, при анализе 277 белков, кодируемых генами 18 хромосомы, если предположить, что пептиды после параллельного гидролиза 4-мя протеазами идентифицируются идеально, но при этом их длина не более 15 остатков, можно восстановить полностью фрагменты длиной до 30-33 остатков, но это единичные случаи. Если же увеличить размер детектируемых пептидов до 20 остатков, то восстановленных фрагментов длиной 40-60 остатков достаточно много (максимальный вариант – 69). Фильтрация по ΔI или случайным образом может помочь спрогнозировать реальные цифры покрытия последовательности. В приведённом примере на рисунке 5 показаны изменения степени покрытия последовательности при условии, что возможно детектировать пептиды разной массы (А) и тот же вариант, но с добавленным условием, что произвольный пептид детектируется с вероятностью 0,6 (Б).

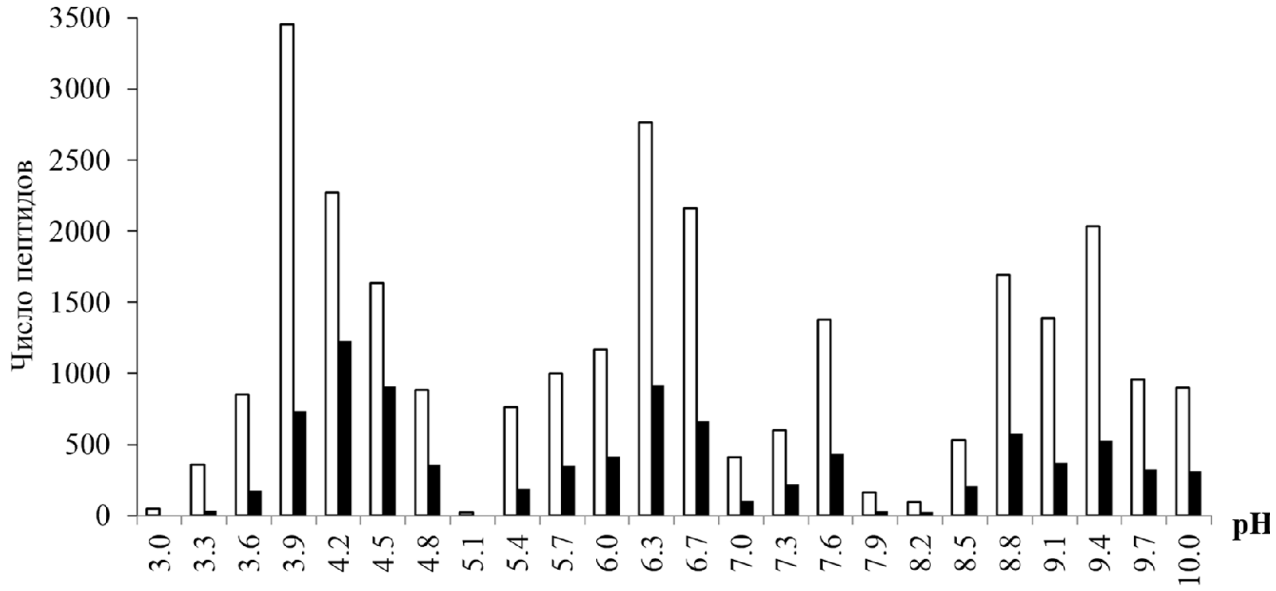

Рисунок 4. Распределение по фракциям пептидов после гидролиза. Симуляция фракционирования с помощью изоэлектрического фокусирования для 277 белков, кодируемых генами 18 хромосомы человека, pH 3-10, белые столбцы - параллельный гидролиз 4 протезами, чёрные - только трипсином.

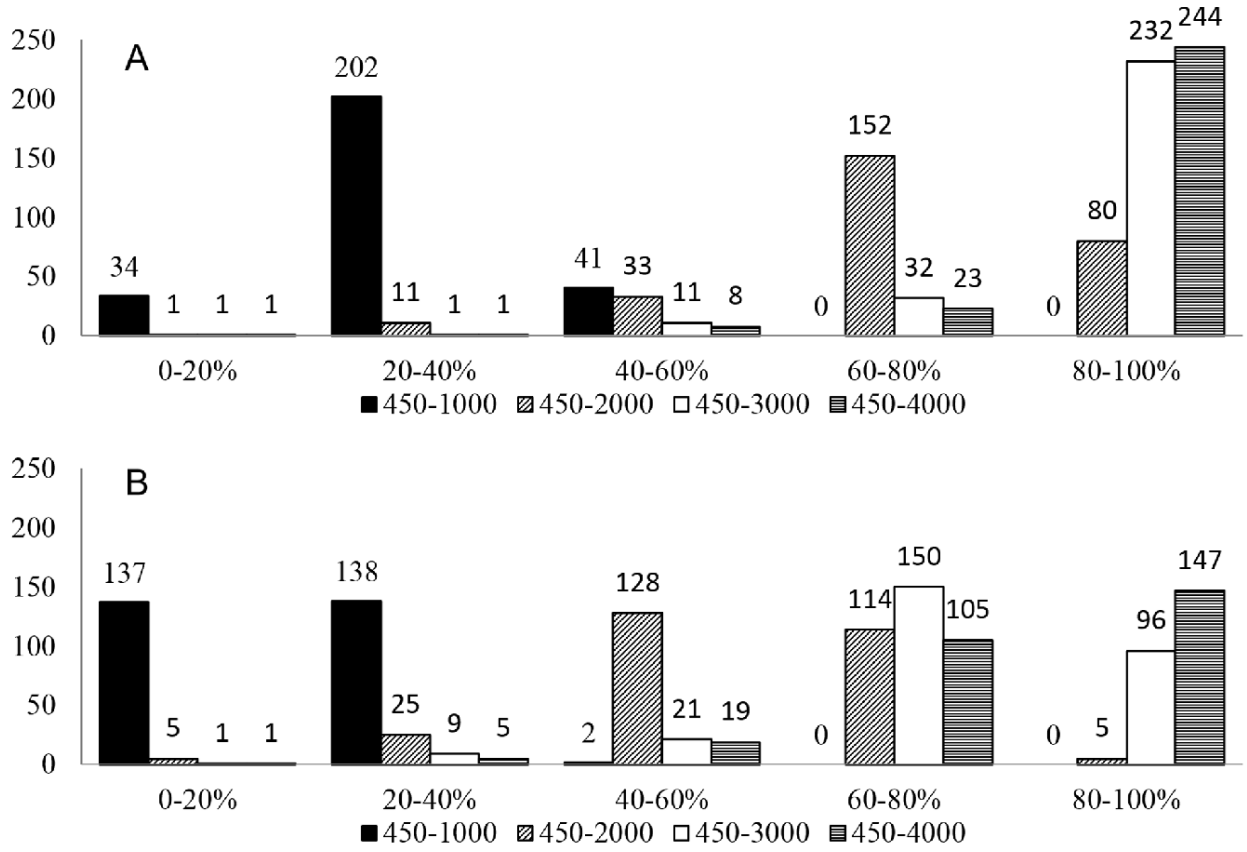

Рисунок 5. Распределение процента покрытия последовательностей 277 белков, кодируемых генами 18 хромосомы, при восстановлении последовательности для пептидов различных масс при идеальной детекции (А) и при случайном отборе пептидов с вероятностью 0,6 (В).

ЗАКЛЮЧЕНИЕ

Программа ProteoCat представляет собой мультиплатформенное приложение, реализованное на языке JAVA в виде исполняемого jar-архива и имеет интуитивно понятный графический интерфейс пользователя. Программа свободна для использования академическими пользователями и доступна по адресу http://www.ibmc.msk.ru/LPCIT/ProteoCat. Основное назначение программы – помощь при планировании широкомасштабных протеомных экспериментов.

Работа поддержана грантом РНФ№14-25-00132.

ЛИТЕРАТУРА

1. *Picotti P., Aebersold R.* (2012) Nature Methods, **9**(6), 555-566.
2. *Li Y.F., Arnold R.J., Tang H., Radivojac P.* (2010) J. Proteome Res., **9**(12), 6288-6297.
3. *Tang H., Arnold R.J., Alves P., Xun Z., Clemmer D.E., Novotny M.V., Reilly J.P., Radivojac P.* (2006) Bioinformatics, **22**(14), e481-e488.
4. *Swaney D.L., Wenger C.D., Coon J.J.* (2010) J. Proteome Res., **9**(3), 1323–1329.
5. *Mazzucchelli G., Zimmerman T., Smargiasso N., Baiwir D., Meuwis M.A., De Pauw E.* (2015) *De novo* sequencing using MELD proteolysis coupled to a “sequence assembly” algorithm In 63rd ASMS Conference on Mass Spectrometry & Allied Topics, http://hdl.handle.net/2268/182843
6. *Скворцов В.С., Алексейчук Н.Н., Худяков Д.В., Ромеро Рейес И.В.* (2015) Биомед. химия, **61** (1), 83-91.
7. *Bjellqvist B., Hughes G.J., Pasquali Ch., Paquet N., Ravier F., Sanchez J.-Ch., Frutiger S., Hochstrasser D.F.* (1993) Electrophoresis, **14**, 1023-1031.
8. *Branca R.M., Orre L.M., Johansson H.J., Granholm V., Huss M., Pérez-Bercoff A. et al.* (2014) Nature methods, **11**(1), 59-62.
9. *Heller M., Ye M., Michel P.E., Morier P., Stalder D., Jünger M.A. et al.* (2005) J. Proteome Res., **4**(6), 2273-2282.
10. *Krokhin O.V.* (2006) Analytical chemistry, **78**(22), 7785-7795.
11. *Рыбина А.В., Скворцов В.С., Копылов А.Т., Згода В.Г.* (2014) Биомед. химия, **60**(6), 707-712.
12. *Barrett A.J., Woessner J.F., Rawlings N.D. (eds.)* (2012) Handbook of proteolytic enzymes (vol. 1), Elsevier, 1009 c.

Поступила: 01. 11. 2015.

ProteoCat: A TOOL FOR PLANNING OF PROTEOMIC EXPERIMENTS

V.S. Skvortsov, N.N. Alekseychuk, D.V. Khudyakov, A.V. Mikurova, A.V. Rybina, S.E. Novikova, O.V. Tikhonova

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

ProteoCat is a computer program has been designed to help researchers in the planning of large-scale proteomic experiments. The central part of this program is the subprogram of hydrolysis simulation that supports 4 proteases (trypsin, lysine C, endoproteinases AspN and GluC). For the peptides obtained after virtual hydrolysis or loaded from data file a number of properties important in mass-spectrometric experiments can be calculated or predicted. The data can be analyzed or filtered to reduce a set of peptides. The program is using new and improved modification of our methods developed to predict pI and probability of peptide detection; pI can also be predicted for a number of popular pKa's scales, proposed by other investigators. The algorithm for prediction of peptide retention time was realized similar to the algorithm used in the program SSRCalc. ProteoCat can estimate the coverage of amino acid sequences of proteins under defined limitation on peptides detection, as well as the possibility of assembly of peptide fragments with user-defined size of “sticky” ends. The program has a graphical user interface, written on JAVA and available at http://www.ibmc.msk.ru/LPCIT/ProteoCat.

Key words: mass spectrometry, peptide properties prediction, protein hydrolysis, covering of protein sequence