ПРОТОКОЛЫ ЭКСПЕРИМЕНТОВ, ПОЛЕЗНЫЕ МОДЕЛИ, ПРОГРАММЫ И СЕРВИСЫ

ПРОГРАММА ПРЕДСКАЗАНИЯ ВРЕМЕНИ УДЕРЖАНИЯ ПЕПТИДОВ С УЧЁТОМ ПОСТТРАНСЛЯЦИОННЫХ МОДИФИКАЦИЙ

А.И. Воронина*, А.В. Рыбина

Научно-исследовательский институт Биомедицинской химии им. В.Н. Ореховича,
119121 Москва, Погодинская ул., 10; *e-mail: an.voronina@list.ru

В работе представлена программа и web-сервис Retention Time Predictor (RTP), предназначенные для предсказания времени удержания пептидов на хроматографической колонке в экспериментах по масс-спектрометрии и учитывающая посттрансляционные модификации аминокислотных остатков (а.о.). Программа представляет собой модификацию известной программы SSRCalc версии 3 (Krokhin, Anal. Chem., 2006, 78(22), 7785–7795). В нее добавлены значения коэффициентов удержания для модифицированных а.о. и алгоритм расчёта величины изоэлектрической точки из программы pIPredict (Skvortsov et al., Biomed. Chem. Res. Meth., 2021, 4(4), e00161). Модификации, описанные в программе, включают: (i) Tandem Mass Tag (TMT) и Isobaric Tags for Relative and Absolute Quantification (iTRAQ) метки; (ii) ацетилирование, формилирование и метилирование N-концевого остатка и/или бокового радикала лизина; (iii) карбамидометилирование остатков цистеина, аспарагиновой и глутаминовой кислот; (iv) окисление и двойное окисление остатков метионина и пролина; (v) фосфорилирование остатков серина, треонина и тирозина; (vi) С-концевое амидирование остатков лизина и аргинина; (vii) образование пропионамида с остатком цистеина. Подбор коэффициентов удержания проведён с использованием данных 25 масс-спектрометрических экспериментов, для которых идентификация была выполнена заново по исходным (RAW) данным, депонированным в БД ProteomeXchange. Программа RTP и web-сервис свободно доступны по адресу http://lpcit.ibmc.msk.ru/RTP.

Ключевые слова: время удержания пептида; изоэлектрическая точка; посттрансляционные модификации; web-сервис

DOI: 10.18097/BMCRM00196

ВВЕДЕНИЕ

В экспериментах по протеомике время удерживания пептидов на хроматографической колонке (retention time, RT) – важный показатель, который определяется только свойствами самого пептида и особенностями оборудования для жидкостной хроматографии и не зависит от метода фрагментации и способов последующей идентификации [1]. Это даёт возможность использовать данный параметр для оценки достоверности идентификации пептидов [2]. Кроме того, предсказание RT можно использовать при планировании эксперимента для более прицельного отбора пептидов для последующего анализа.

Величина RT, наблюдаемая в каждом конкретном эксперименте, зависит от множества факторов даже при работе на одном и том же оборудовании: различий в настройках прибора и характеристик колонки, температурного режима, состава смеси [2]. Однако это препятствие можно обойти, если использовать не непосредственное предсказание величины RT, а предсказание более фундаментальной характеристики пептида, от которой в свою очередь зависит RT. Например, Krokhin и соавторы предложили использовать величину «гидрофобности» (Hydrophobicity Index, HI) и создали программу Sequence-Specific Retention Calculator (SSRCalc) для предсказания данной величины [3], которая использует аддитивную схему с коррекцией. Первоначальная сумма величин, названных коэффициентами удержания, индивидуальных для каждого из аминокислотных остатков в зависимости от его положения в полипептидной цепи, корректируется набором коэффициентов. Эти коэффициенты связаны с размером пептида, зарядом, величиной изоэлектрической точки, наличием более одного остатка пролина рядом, способностью образовывать гидрофобные кластеры или устойчивые спирали.

Существуют и другие алгоритмы предсказания RT от простых линейных регрессионных моделей, базирующихся на свойствах аминокислот до сложных биофизических моделей, либо использующие методы глубокого обучения (deep learning). Так, авторы SSRCalc использовали методы глубокого обучения в своей новой программе Chronologer [4]. При этом, Chronologer может предсказывать RT и для пептидов с некоторыми посттрансляционными и химическими модификациями. В данной работе тот факт, вносятся ли эти модификации искусственно или в процессе жизнедеятельности, не существенен, и оба варианта будут обозначены, как ПТМ. Тем не менее, несмотря на видимый выигрыш в точности предсказания, методы глубокого обучения продолжают оставаться «чёрным ящиком», так как нет возможности интерпретировать полученную модель с точки зрения физико-химических свойств объектов. В свою очередь, такая величина как гидрофобность может быть использована сама по себе для предсказания каких-либо других свойств, связанных с пептидами [5], а, например, коэффициенты удержания для отдельных остатков могут быть дескрипторами в QSAR-моделях.

Таким образом, представляется полезным добавить в программу SSRCalc возможность предсказания величины HI для пептидов с ПТМ, которой в исходном варианте нет. В качестве базовой была выбрана оригинальная программа SSRCalc версии 3.0, написанная на языке Perl и

Таблица 1. Описание 25 наборов данных масс-спектрометрических экспериментов, использованных в работе.

ID в работе [7]	Всего уникальных пептидов	UP	FR	ПТМ-1	Идентификатор PRIDE	Химическая метка	Диапозон pH
S2	13223	3022	69	6496	PXD000065	-	4.4-4.65
S3	28079	4364	72	12113	PXD000065	-	3.7-4.9
S4	26417	4799	72	10973	PXD000065	-	3.7-4.9
S5	17849	3299	72	8228	PXD000065	-	4.0-4.25
S6	17114	3427	72	8314	PXD000065	-	4.2-4.45
S7	18154	2754	72	7867	PXD000065	-	3.7-4.05
S8	23594	1456	54	9887	PXD0006291	TMT10	3.0-10.0
S9	32843	2448	72	12521	PXD0006291	TMT10	3.7-4.9
S10	28428	2073	49	12088	PXD006291	TMT10	3.0-10.0
S11	18042	1626	72	7779	PXD006291	TMT10	3.7-4.9
S12	10922	415	24	4227	PXD010006	TMT10	3.0-10.0
S13	720	86	3	238	PXD010006	TMT10	2.5-3.7
S14	38161	2935	46	17293	PXD010006	TMT10	3.0-10.0
S15	12473	367	22	5520	PXD005410	TMT10	3.0-10.0
S16	61651	3784	65	29439	PXD005410	TMT10	3.0-10.0
S17	3821	74	9	1565	PXD005410	TMT10	2.5-3.7
S18	11238	83	72*	4925	PXD000065	iTRAQ8	3.7-4.9
S19	2679	172	18	996	PXD017201	TMT10	3.0-10.0
S20	2692	157	16	897	PXD017201	TMT10	3.0-10.0
S21	2787	186	18	973	PXD017201	TMT10	3.0-10.0
S22	2847	185	19	929	PXD017201	TMT10	3.0-10.0
S23	2686	159	16	933	PXD017201	TMT10	3.0-10.0
S24	24129	1637	31	10957	PXD006291	TMT10	6.0-9.0
S25	30125	2317	42	13612	PXD006291	TMT10	6.0-11.0
[13]	26473	16419	18	7595	PXD018450	-	-

Примечание. UP – число немодифицированных пептидов, по которым возможен пересчёт RT в virtual HI; FR – число фракций в выборке, для которых было достаточно данных для пересчёта, * – все фракции рассматривались как единственная; ПТМ-1 – число пептидов с одной единственной ПТМ.

доступная в рамках Artistic License. Подробности алгоритма программы и используемые коэффициенты опубликованы в работе [3]. Добавочные параметры для остатков с ПТМ могут быть рассчитаны на основании наборов данных масс-спектрометрического анализа.

МЕТОДИКА

Набор целевых значений для пептидов, идентифицированных при анализе масс-спектрометрических данных

Для формирования выборки пептидов для анализа и тестирования были использованы полученные ранее (с использованием программы Peaks Studio X Pro [6]) в работе [7] результаты по идентификации пептидов для 25 наборов исходных (RAW) данных масс-спектрометрических экспериментов, выполненных с применением метода изоэлектрического фокусирования (IEF) и депонированных в БД ProteomeXchange (PXD000065 [8], PXD005410 [9], PXD006291 [10], PXD010006 [11] и PXD017201 [12]). В части из этих экспериментов авторы работ использовали TMT (Tandem Mass Tag) или iTRAQ (Isobaric Tags for Relative and Absolute Quantification) метки (только одна выборка с iTRAQ из двух была использована в работе, причина описана далее), в 6 случаях метки не использовали. В каждом наборе данных были представлены 72 фракции, отобранные в узком поддиапазоне pH. В отличие от работы [7], из общего массива идентифицированных пептидов отбирали только те, которые удовлетворяли двум условиям: уровень FDR 0.1% и значение ASCORE при наличии модификации не менее 500. В работе [7] ограничение для идентификации прекурсорного иона было установлено в 5 ppm, точность идентификации фрагментов – 0.01 Da. Все модификации, включая алкилирование остатков цистеина и наличие меток TMT или iTRAQ, считали вариабельными. Это позволило ожидать, что часть непрореагировавших с советующими реагентами пептидов также будут зарегистрированы. Кроме того, были наложены дополнительные ограничения. Пептид не должен был быть идентифицирован менее чем в 2-х соседних фракциях (отличаются по pH), но и не более чем в числе фракций, соответствующих диапазону pH шириной 0.3 для экспериментов с широким диапазоном pH и 0.1 с узким. Кроме того, был использован дополнительный набор данных из работы [13], в которой каждая фракция была получена из фрагмента 2D геля. Это позволило получить данные для описание дополнительной модификации –

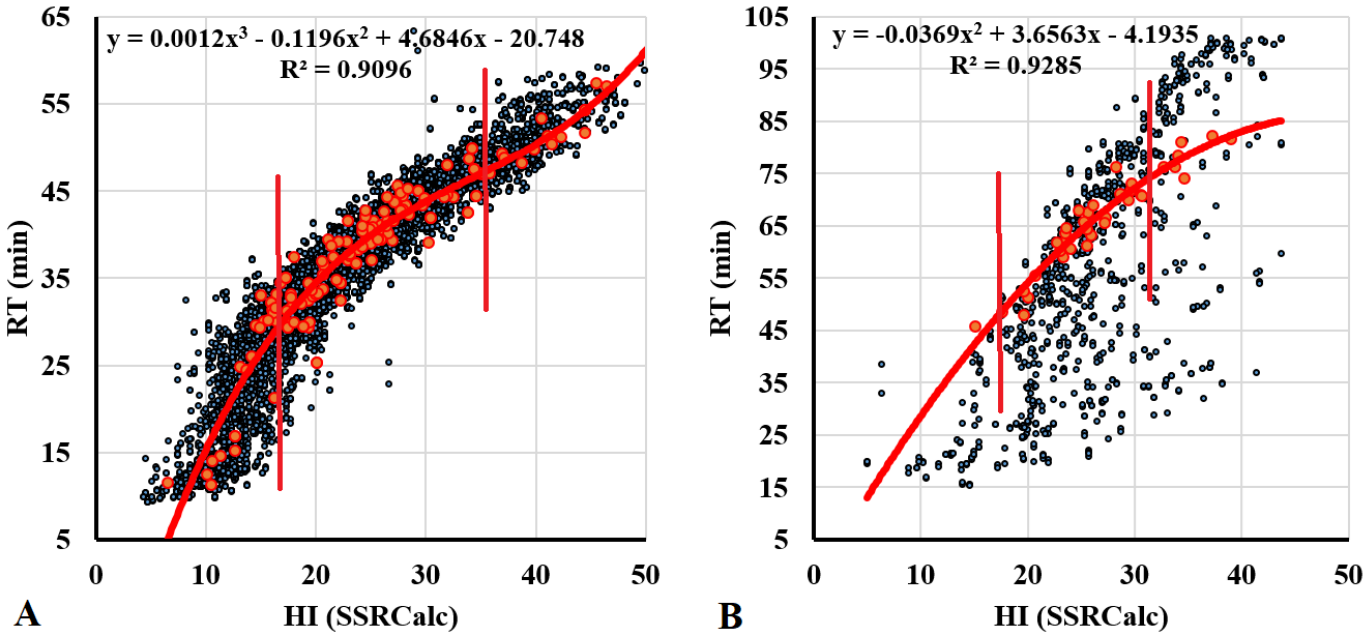

Рисунок 1. Примеры корреляции наблюдаемых величин RT и величин HI, рассчитанных программой SSRCalc для выборок S5 (A) и S22 (B). Красным цветом выделена одна из фракций, для которой приведены параметры линии тренда. Вертикальные линии ограничивают участок линейной зависимости.

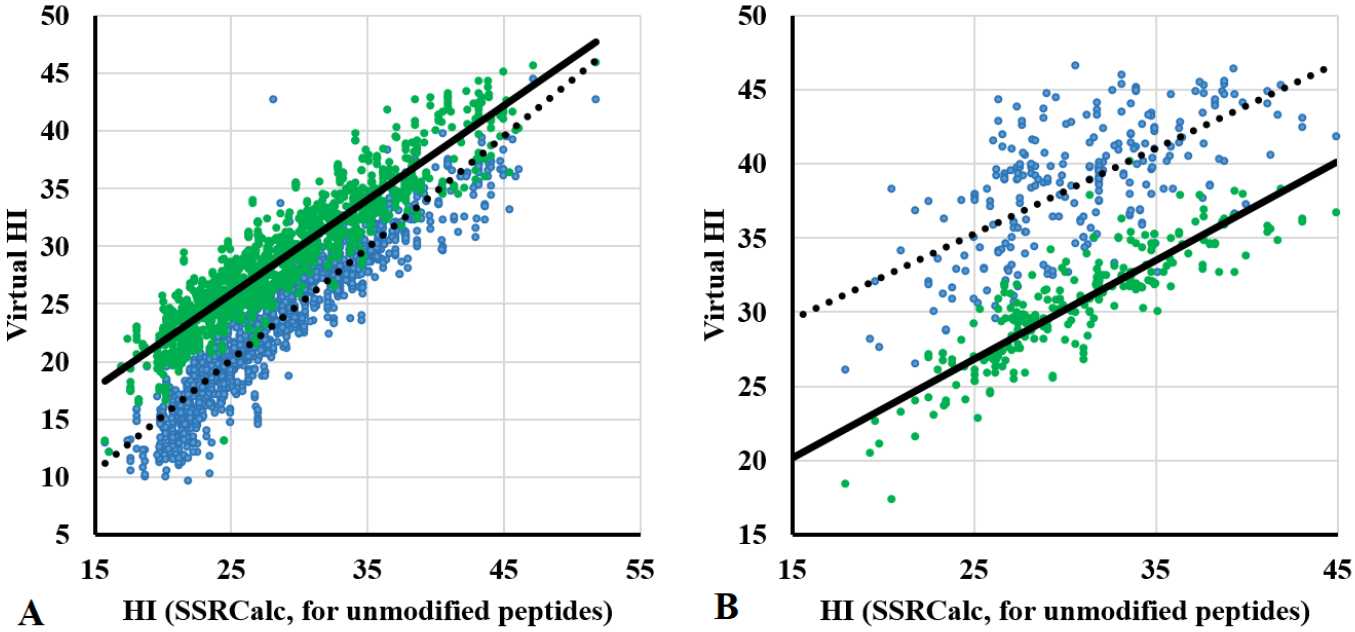

Рисунок 2. Сравнение величины virtual HI для пептидов с одинаковой последовательностью с ПТМ (синий) и без (зелёный) с величиной HI, рассчитанной программой SSRCalc для пептидов без модификаций. (A) Модификация M(+15.99) в положении M; (B) N-концевая модификация [+42.01].

образование пропионамида, а также позволило добрать данные по пептидам с оксипролином. Фильтрацию по диапазону pH в последнем случае не использовали. Описание каждого из наборов данных представлено в таблице 1.

Как и можно было ожидать, число непрореагировавших пептидов в экспериментах с использованием меток очень мало. В то же время число немодифицированных пептидов в конкретной фракции имеет определяющее значение для второго этапа подготовки данных – пересчёта значений RT в величину гипотетической гидрофобности (virtual HI) на основе корреляции наблюдаемых величин RT и рассчитанных программой SSRCalc для каждой отдельной фракции (рис. 1). Для построения корреляционного уравнения было установлено ограничение: использовали только данные для немодифицированных пептидов в пределах одной фракции, не менее 7 уникальных последовательностей с разницей между минимальным и максимальным значением RT не менее 15 минут. Для создания корреляционных уравнений с модификацией TMT отбирали только непрореагировавшие пептиды без модификаций, имеющие аналоги с TMT в той же фракции, исключая этим часть возможных ошибок

идентификации. Учитывая, что при сравнении наблюдаемых величин RT и значений HI только в части диапазона значений RT наблюдалась линейная зависимость (рис. 1), то наблюдения, выходящие за пределы этого диапазона, в работе не рассматривали.

После получения уравнений корреляции для каждой фракции в отдельности по данным наблюдаемого значения RT было рассчитано значение virtual HI. В выборку для подбора коэффициентов удержания отбирали пептиды, имеющие только одну требуемую модификацию в соответствующей позиции (N-концевой остаток, 2-й остаток, «средний» – от 3-го до 2-го с С-конца, предпоследний и С-концевой остаток). Для условно «длинных (от 10 остатков) и условно «коротких» (от 5 до 10 остатков) пептидов выборки формировались отдельно. Учитывали только наборы уникальных пептидов от 8 штук. Подбор коэффициентов удержания проводили в диапазоне от -10 до +20 с шагом в 0.05. Наиболее подходящее значение выбирали по минимальной сумме квадратов отклонений вновь рассчитанных значений Hi от значений virtual HI по всей выборке.

В работе также были ограничены варианты модификаций для которых отбирали данные (в скобках указана моноизотопная дельта масс): TMT (+229.16) и iTRAQ (+304.21) метки, а также ацетилирование (+42.01), формилирование (+27.99) и метилирование (+14.02) N-концевого остатка и/или бокового радикала лизина; карбамидометилирование (+57.02) остатков цистеина, аспарагиновой и глутаминовой кислот; окисление (+15.99) и двойное окисление (+31.99) остатков метионина и пролина, фосфорилирование (+79.97) остатков серина, треонина и тирозина; С-концевое амидирование (0.98) остатков лизина и аргинина; образование пропионамида (+71.04) с остатком цистеина. Нет нужды отдельно рассматривать в работе дезамидирование (+0.98) остатков глутамина и аспарагина, так как по сути это иммитирует аминокислотную замену на остаток аспарагиновой и глутаминовой кислот, параметры для которых в программе SSRCalc есть. К сожалению, некоторые из востребованных модификаций, например, фосфорилирование остатка тирозина, при заданных ограничениях не были представлены в выборке в достаточном количестве (8 и более аминокислотных остатков). В таком случае подбор проводили по всему массиву данных, включая пептиды с двумя и более модификациями, например, наличие TMT метки и фосфорилирование остатка тирозина в одном пептиде.

Следует отметить, что в случае выборки с iTRAQ меткой, пары пептидов с меткой и без не были обнаружены. Вероятно, процент непрореагировавших пептидов, в отличие от процедуры внесения TMT метки или обработкой йодацетамидом (+57.02), ничтожно мал. Однако в данном случае при вычислении virtual HI можно опереться на загрязнение лабораторного оборудования, т.е. использовать пептиды, не обработанные в пробе, но уверенно определяемые при масс-спектрометрии (рис. 2). Только для одной из двух выборок с iTRAQ меткой из работы [7] таких данных было достаточно для подбора уравнения пересчёта RT в virtual HI. В данном случае все фракции рассматривали как единое целое, так как данных по примесям было не так много. Однако если сравнивать между собой значения RT для одинаковых пептидов из разных фракций, то для данной выборки отклонения в среднем не превышали одной минуты.

Программа Retention Time Predictor (RTP) – модификация программы SSRCalc для работы с учётом ПТМ

Оригинальный код программы SSRCalc v3.0 на языке Perl был транслирован на язык Python (версия 3.6.8). Воспроизводимость результатов для немодифицированных пептидов была проверена на выборке из более чем 215 тысяч уникальных пептидных последовательностей, собранных из всех 25 выборок, рассматриваемых в работе. В случае округления результатов до второго знака после запятой при записи в файл максимальное отклонение в 0.01 единицу HI наблюдали у 239 пептидов, что составило 0,11% от общей выборки. Остальные значения были идентичны. Наличие небольшой ошибки округления вероятнее всего является следствием различий в округлении дробных чисел в языках Perl и Python.

В соответствии с алгоритмом оригинальной программы SSRCalc, вычисление параметра HI происходит в два этапа. Первоначально суммируются табличные коэффициенты удержания для каждого из аминокислотных остатков пептида в зависимости от их расположения в пептиде. Различаются позиции N-концевая (N1), вторая с N-конца (N2), любая в середине цепи (M), предпоследняя с С-конца (C2), С-концевая (C1). На втором этапе последовательно рассчитывается ряд поправочных коэффициентов, в зависимости от аминокислотного состава и свойств пептида, его длины и т.д. Итоговая величина является гидрофобностью пептида. Разработчики SSRCalc [3] описали 2 шкалы коэффициентов удержания для колонок с размером пор 10 нм и 30 нм, в настоящей работе использовали последнюю. Способ учёта модифицированных остатков при вычислении поправочных коэффициентов определяли из общих соображений, как могут измениться свойства остатков при внесении данной модификаций. Например, остаток метионина при расчёте поправки «кластерность» определён как сильно гидрофобный, следовательно, его окисленная форма не учитывается при расчёте данной поправки. В тоже время остаток оксипролина своей геометрии не меняет и при расчёте поправки на наличие полипролинового фрагмента продолжает учитываться. При расчёте поправок, связанных с наличием заряда на остатке, учитывали изменения в числе протонируемых или диссоциируемых групп для данной модификации. Так как табличных данных по значениям pKa для расчёта величины pI модифицированных пептидов в программе SSRCalc нет, то для её расчёта программа RTP использует шкалу программы pIPredict версии 3 [7] (вариант без учёта соседних остатков). При тестировании влияния замены алгоритма расчёта pI на той же выборке из более чем 215 тысяч немодифицированных пептидов среднее отклонение составило 0.04 единицы HI, максимальное – 0.82, при этом среднее отклонение величины pI было 0.32 единицы pH, максимальное – 4.25.

Между программами имеются также различия в описании остатков, связанных с модификациями N-концевой аминогруппы. Они трактуются как добавочный остаток, для которого имеются собственные табличные данные. Например, пептид A(+229.16)GGSTR в программе преобразуется в форму [+229.16]AGGSTR, где [+229.16] – самостоятельный остаток с массой 229.16 и собственными табличными значениями коэффициентов удержания, а остаток аланина становится вторым. Это позволяет сократить размеры шкалы. В свою очередь пептид K(+229.16)(+229.16)GGSTR

Таблица 2. Количество наблюдений для пептидов с различными модификациями для наборов, по которым проводили подбор значения коэффициентов удержания.

ПТМ	Длина пептида 10 и более остатков					Длина пептида меньше 10				
	N1	N2	M	C2	C1	N1	N2	M	C2	C1
[+304.21]	3029 (8463)	NA	NA	NA	NA	1103 (877)	NA	NA	NA	NA
K(+304.21)	NA	55 (57)	279 (460)	57 (99)	1906 (3093)	NA	0 (0)	27 (29)	0 (0)	569 (596)
[+229.16]	61 (150958)	NA	NA	NA	NA	22 (23346)	NA	NA	NA	NA
K(+229.16)	NA	529 (3033)	1896 (6587)	293 (1475)	28691 (80264)	0 (0)	181 (351)	237 (274)	63 (90)	15306 (15984)
T(+79.97)	0 (0)	0 (113)	43 (2187)	0 (19)	0 (0)	0 (0)	0 (9)	0 (155)	0 (0)	0 (0)
S(+79.97)	0 (32)	9 (842)	363 (12755)	0 (261)	0 (10)	0 (17)	0 (82)	31 (887)	0 (24)	0 (0)
Y(+79.97)	0 (0)	0 (69)	0 (534)	0 (32)	0 (0)	0 (0)	0 (16)	0 (59)	0 (10)	0 (0)
C(+71.04)	0 (0)	0 (0)	27 (100)	0 (9)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
C(+57.02)	254 (555)	277 (3017)	2330 (24527)	273 (2805)	0 (209)	49 (104)	47 (490)	159 (1852)	55 (502)	0 (18)
E(+57.02)	191 (254)	58 (128)	170 (438)	19 (37)	0 (0)	29 (43)	0 (0)	0 (16)	0 (0)	0 (0)
D(+57.02)	425 (644)	38 (83)	126 (285)	12 (20)	0 (0)	202 (283)	0 (0)	0 (0)	0 (0)	0 (0)
K(+42.01)	NA	0 (0)	11 (48)	0 (10)	0 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
[+42.01]	355 (2267)	NA	NA	NA	NA	61 (305)	NA	NA	NA	NA
P(+31.99)	0 (0)	0 (0)	28 (163)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
M(+31.99)	0 (12)	11 (26)	92 (290)	17 (37)	0 (0)	0 (0)	0 (0)	11 (17)	0 (0)	0 (0)
K(+27.99)	NA	84 (109)	243 (332)	65 (91)	268 (377)	20 (21)	13 (19)	13 (14)	10 (11)	40 (59)
[+27.99]	1283 (2065)	NA	NA	NA	NA	143 (87)	NA	NA	NA	NA
M(+15.99)	406 (848)	444 (901)	2776 (23639)	340 (2712)	0 (177)	95 (188)	70 (120)	187 (922)	36 (234)	0 (31)
P(+15.99)	0 (0)	0 (0)	100 (102)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
K(+14.02)	NA	0 (10)	19 (55)	0 (0)	61 (119)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
[+14.02]	24 (34)	NA	NA	NA	NA	0 (0)	NA	NA	NA	NA
K(-.98)	NA	NA	NA	NA	16 (0)	NA	NA	NA	NA	0 (0)
R(-.98)	NA	NA	NA	NA	19 (0)	NA	NA	NA	NA	0 (0)

Примечание. Здесь и далее в описании ПТМ: [] – соответствует N-концевой модификации основного хода полипептидной цепи; () – модификация бокового радикала аминокислотного остатка или С-концевой COOH группы; NA – данная позиция для модификации невозможна из-за особенностей алгоритма программы. Указано число пептидов с одиночными ПТМ (значения меньше 8 обнулялись), в скобках указано число пептидов с данной ПТМ, имеющих и другие модификации.

преобразуется к виду [+229.16]K(+229.16)GGSTR, где K(+229.16) – остаток с массой 357.26 (с модифицированным боковым радикалом), также имеющий собственные табличные значения коэффициентов удержания. Если программа встречает пептид K(+229.16)GGSTR, то трактует его как пептид с модификацией N-концевой аминогруппы, а не бокового радикала. Строго говоря, последнее правило может не соответствовать действительности, однако, способа различить на каком из атомов азота первого остатка находятся модификации [+304.21], [+229.16] или [+27.99] нет. Для модификации [+42.01] и [+14.02], если пептид не соответствует N-концевому пептиду белка, можно предположить, что модифицирован боковой радикал, но всё равно применяется данное правило.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Подбор значений коэффициентов удержания для пептидов с модификациями и тестирование результатов

В таблице 2 указано количество отобранных наблюдений в подвыборках для пептидов с различными модификациями и по каждой из возможных позиций. Первая величина – количество наблюдений, когда в каждом пептиде модифицирован только 1 остаток, величина в скобках – количество пептидов, содержащих данную ПТМ. В случае одиночной ПТМ хорошо видно согласованное изменение гидрофобности пептидов (рис. 2). При этом направление изменений хорошо согласуется с ожидаемым результатом,

Таблица 3. Значения коэффициентов удержания, рассчитанных в данной работе и использующихся в программе RTP.

ПТМ	Длина пептида 10 и более остатков					Длина пептида меньше 10				
	N1	N2	M	C2	C1	N1	N2	M	C2	C1
[+304.21]	3.35	NA	NA	NA	NA	5.5	NA	NA	NA	NA
K(+304.21)	NA	-1.3	-0.3	-1.4	0.05	NA	1.6**	1.35	1.6**	1.85
K(+229.16)	NA	-1.65	-0.95	-1.25	1.55	NA	1.83**	0.35	1.83**	3.3
[+229.16]	3.6	NA	NA	NA	NA	4.6	NA	NA	NA	NA
S(+79.97)	1.4*	1.95	-0.05	2*	1.7*	1.3*	1.6*	2.7	1.9*	1.61*
T(+79.97)	2.37**	3.9*	1.05	2.2*	2.37**	2.37**	1.6*	3.1*	2.37**	2.37**
Y(+79.97)	4.65**	4.6*	4.3*	6.2*	4.65**	4.65**	4.8*	4.1*	3.9*	4.65**
C(+71.04)	3.37**	3.37**	2.25	4.5*	3.37**	3.37**	3.37**	3.37**	3.37**	3.37**
C(+57.02)	0.75	1	0	0.05	3.1*	2.15	2.65	1.7	1.3	3.6*
D(+57.02)	1.95	0.45	-2.25	-0.2	0.71**	3.6	0.71**	0.71**	0.71**	0.71**
E(+57.02)	0.5	0.45	-2.4	0.65	0.41**	2.85	0.41**	0.41**	0.41**	0.41**
[+42.01]	8.55	NA	NA	NA	NA	7.85	NA	NA	NA	NA
K(+42.01)	NA	2.8**	3	2.00*	3.4*	NA	2.8**	2.8**	2.8**	2.8**
M(+31.99)	0.9*	0.45	0.65	1	1.01**	1.01**	1.01**	2.05	1.01**	1.01**
P(+31.99)	0.3**	0.3**	0.3	0.3**	0.3**	0.3**	0.3**	0.3**	0.3**	0.3**
[+27.99]	7.45	NA	NA	NA	NA	6.85	NA	NA	NA	NA
K(+27.99)	0.8	1.85	1.55	0.75	2.05	2.1	1.8	4.4	2.55	3.25
M(+15.99)	-0.2*	-0.65	-1.4	-1.25	3.3*	1.15	0.75	0.8	0.35	3.1*
P(+15.99)	1.4**	1.4**	1.40	1.4**	1.4**	1.4**	1.4**	1.4**	1.4**	1.4**
[+14.02]	3.7	NA	NA	NA	NA	3.7**	NA	NA	NA	NA
K(+14.02)	NA	-0.3*	0.1	0.55**	1.85	NA	0.55**	0.55**	0.55**	0.55**
K(-.98)	NA	NA	NA	NA	-2.4	NA	NA	NA	NA	-2.4**
R(-.98)	NA	NA	NA	NA	-0.8	NA	NA	NA	NA	-0.8**

Примечание. * – значение вычислено по набору пептидов с более чем одной ПТМ; ** – недостаточно данных для расчёта, значение взято как среднее по уже рассчитанным для данной ПТМ по другим сайтам; NA – данная позиция для модификации невозможна.

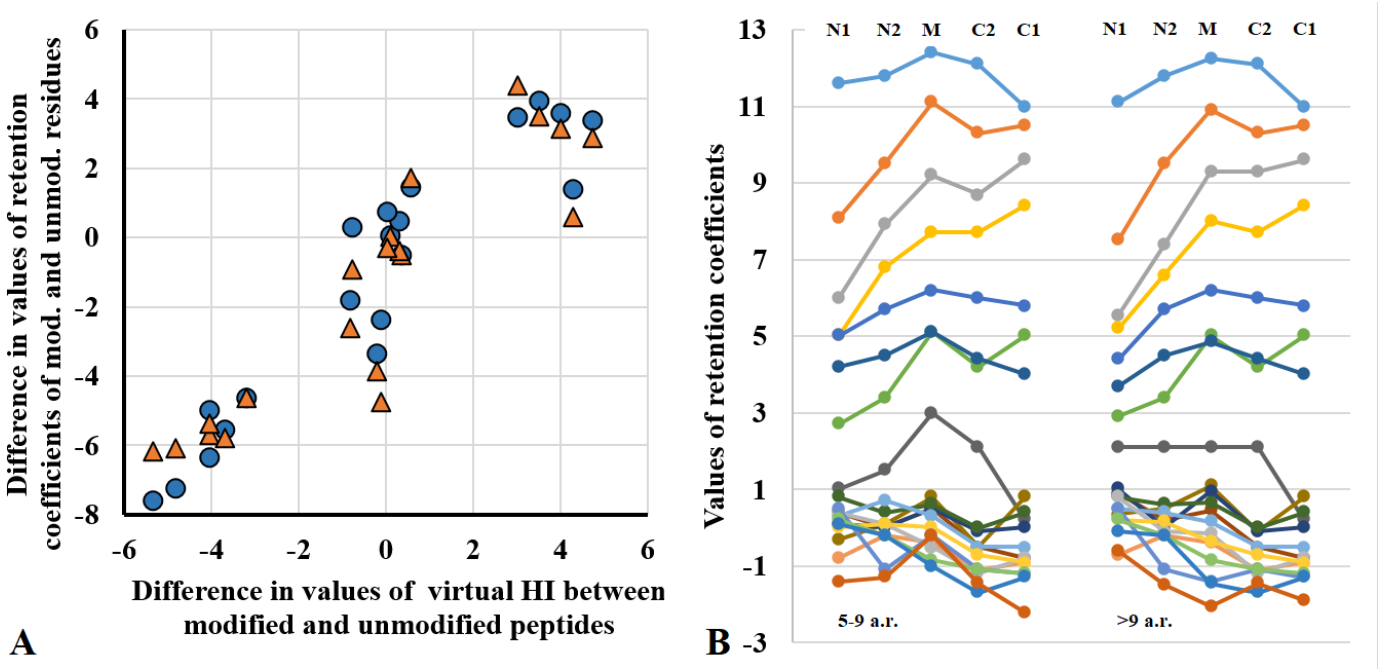

Рисунок 3. (A) Сравнение усреднённых величин смещения величины virtual HI для пептидов с одинаковой аминокислотной последовательностью после внесения ПТМ с величиной коэффициента удержания, определённого в работе. Синим цветом обозначены величины, рассчитанные для наборов пептидов с одиночной ПТМ, оранжевым – для пептидов с более чем одной ПТМ. (B) Сравнение значений коэффициентов удержания для 20 канонических аминокислотных остатков, установленных в программе SSRCalc для разных позиций в пептиде.

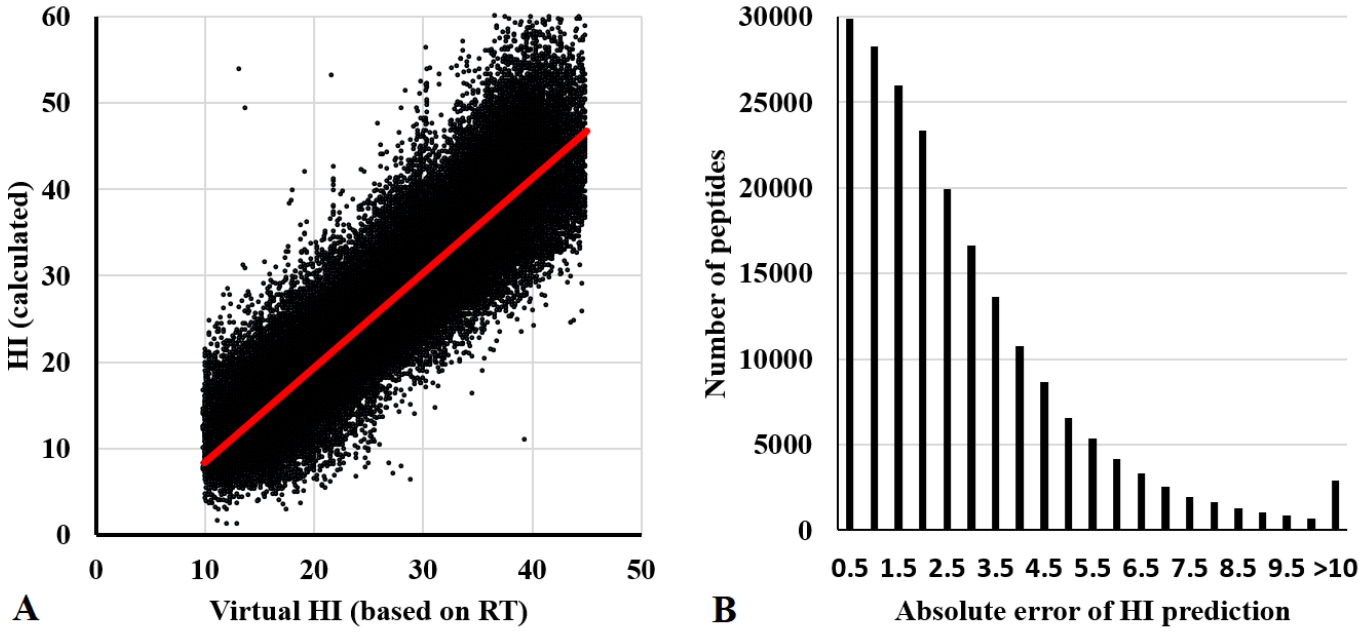

Рисунок 4. Сравнение расчётной величины HI и величины virtual HI (A) и распределение абсолютной ошибки (B) для полного набора данных, включая пептиды с более чем одной ПТМ.

что и должно быть для аддитивной схемы расчёта при изменении единственного параметра. В общем случае можно просто вычислить среднюю величину изменений (или лучше величину смещения линии тренда) и добавить её к значению коэффициента удержания немодифицированного остатка; результат будет близким к величине, подобранной перебором на всех пептидах с данной ПТМ (рис. 3A). Расчёт тех же величин на выборке пептидов, включающих более одной ПТМ, даёт близкие величины. Таким образом, можно ожидать, что и для случая, когда нет достаточного набора данных с одиночными ПТМ, результат будет адекватный.

Подобранные значения коэффициентов удержания приведены в таблице 3. К сожалению, некоторые из модификаций не имеют достаточного числа наблюдений по отдельным позициям ни в одном из вариантов (одиночные или множественные модификации). Но если проанализировать таблицу значений для немодифицированных остатков (рис. 3B), то в среднем минимальное и максимальное значение коэффициентов удержания отличаются на 1.7 единиц, так что в некотором приближении можно взять среднюю величину по имеющимся значениям и заполнить таблицу полностью. На анализируемых в этой работе данных ошибка в 1 единицу величины HI приводила к ошибке при определении величины RT в 1-3 минуты. При этом величина RT по сути дискретная, так как определяется с некоторым шагом, а минимальное среднее отклонение от линии тренда при сравнении величин HI и RT для немодифицированных пептидов для отдельных фракций составило около 3 минут.

Рисунок 4 демонстрирует зависимость величины virtual HI, вычисленной из RT, и величины HI, рассчитанной с учётом имеющихся ПТМ, на всём массиве пептидов с ПТМ (около 200 тысяч), использованных в работе. Для 90% наблюдений абсолютная ошибка не превышает 5. Следует напомнить, что расчёт величины virtual HI с использованием корреляционных уравнений может вносить определённую ошибку. Часть значений соответствуют пептидам с возможной ошибочной идентификацией (с учётом выбранного уровня FDR их не меньше 1%), что соответствует количеству тех пептидов, для которых абсолютная ошибка больше 10. Это можно использовать для фильтрации пептидов с ложной идентификацией.

Программная реализация, доступность кода и ограничения при работе программы

Для расчёта величины HI с использованием полученной шкалы значений коэффициентов удержания можно использовать программу RTP, свободно доступную для пользователей. В функции программы RTP входит также предсказание величины RT по коэффициентам уравнения корреляции в виде многочлена до пятой степени (вводятся пользователем). В процессе выполнения программы для немодифицированных пептидов возможен выбор метода предсказания pI (с помощью шкалы pIPredict 3 или оригинальной шкалы SSRCalc). Так как для немодифицированных пептидов изменения в шкалу значений коэффициентов удержания не вносили, то для таких пептидов программа повторяет вычисления SSRCalc версии 3.

Программа RTP доступна через веб-интерфейс по адресу http://lpcit.ibmc.msk.ru/RTP, а также в виде исполняемого файла для операционной системы MS Windows 10 (по ссылке на той же странице). Код программы может быть предоставлен по запросу. Программа работает с входными файлами в формате FASTA или TXT, в котором пептиды располагаются построчно без заголовков. Описание ПТМ соответствует выходным файлам программы Peaks Studio X Pro. Для веб-интерфейса установлено ограничение при загрузке в 10000 пептидов, размер файла не более 900 кбайт. В случае превышения лимита, данные будут загружены не полностью, и обработаны будут только первые 10000 пептидов. Результаты представлены в виде таблицы, содержащей последовательность пептидов, предсказанные значения HI, pI и RT (последнее при условии ввода

коэффициентов уравнения). Результаты также доступны в виде текстового файла с разделителем «табуляция». Формируется гистограмма распределения значений предсказанной величины Hi или RT.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Данная работа не содержит каких-либо исследований с использованием людей и животных в качестве объектов исследования.

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021 - 2030 годы) (№ 122030100170-5).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. Bączek, T., Kaliszan, R. (2009) Predictions of peptides' retention times in reversed-phase liquid chromatography as a new supportive tool to improve protein identification in proteomics, Proteomics, **9**(4), 835-847. DOI: 10.1002/pmic.200800544

2. Krokhin, O. (2012) Peptide retention prediction in reversed-phase chromatography: proteomic applications. Expert Review of Proteomics, **9**(1), 1-4. DOI: 10.1586/epr.11.79

3. Krokhin, O.V. (2006) Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300-and 100-Å pore size C18 sorbents. Analytical Chemistry, **78**(22), 7785-7795. DOI: 10.1021/ac060777w

4. Wilburn, D.B., Shannon, A.E., Spicer, V., Richards, A.L., Yeung, D., Swaney, D.L., Krokhin, O.V., Searle, B.C. (2023) Deep learning from harmonized peptide libraries enables retention time prediction of diverse post translational modifications, bioRxiv **5**(30), 542978. DOI: 10.1101/2023.05.30.542978

5. Hemshekhar, M., Faiyaz, S., Choi, K. Y. G., Krokhin, O. V., Mookherjee, N. (2019) Immunomodulatory functions of the human cathelicidin LL-37 (aa 13–31)-derived peptides are associated with predicted α-helical propensity and hydrophobic index. Biomolecules, **9**(9), 501. DOI: 10.3390/biom9090501

6. 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 Communications in Mass Spectrometry: RCM, **17**(20), 2337–2342. DOI: 10.1002/rcm.1196

7. Skvortsov, V.S., Voronina, A.I., Ivanova, Y.O., Rybina, A.V. (2021) The prediction of the isoelectric point value of peptides and proteins with a wide range of chemical modifications. Biomedical Chemistry: Research and Methods, **4**(4), e00161. DOI: 10.18097/BMCRM00161

8. Branca, R., Orre, L., Johansson, H., Granholm, V., Huss, M., Pérez-Bercoff, A., Forshed, J., Käll, L., Lehtiö, J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods, **11,** 59–62. DOI: 10.1038/nmeth.2732

9. Panizza, E., Branca, R. M. M., Oliviusson, P. et al. (2017) Isoelectric point-based fractionation by HiRIEF coupled to LC-MS allows for in-depth quantitative analysis of the phosphoproteome. Scientific Reports, **7**, 4513. DOI: 10.1038/s41598-017-04798-z

10. Zhu, Y., Orre, L. M., Johansson, H. J. et al. (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun, **9**, 903. DOI: 10.1038/s41467-018-03311-y

11. Panizza, E., Zhang, L., Fontana, J. M., Hamada, K., Svensson, D., Akkuratov, E. E., Scott, L., Mikoshiba, K., Brismar, H., Lehtiö, J., & Aperia, A. (2019) Ouabain-regulated phosphoproteome reveals molecular mechanisms for Na^+, K^+-ATPase control of cell adhesion, proliferation, and survival. FASEB journal: official publication of the Federation of American Societies for Experimental Biology, **33**(9), 10193–10206. DOI: 10.1096/fj.201900445R

12. Babačić, H., Lehtiö, J., Pico de Coaña, Y., Pernemalm, M., & Eriksson, H. (2020) In-depth plasma proteomics reveals increase in circulating PD-1 during anti-PD-1 immunotherapy in patients with metastatic cutaneous melanoma. Journal for immunotherapy of cancer, **8**(1), e000204. DOI: 10.1136/jitc-2019-000204

13. Kiseleva, O, Zgoda, V, Naryzhny, S, Poverennaya, E. (2020) Empowering Shotgun Mass Spectrometry with 2DE: A HepG2 Study. International Journal of Molecular Sciences, **21**(11), 3813. DOI: 10.3390/ijms21113813

Поступила: 18.06.2023
После доработки: 02.07.2023
Принята к публикации: 14.07.2023

A PROGRAM FOR PREDICTING THE RETENTION TIME OF PEPTIDES WITH POST-TRANSLATIONAL MODIFICATIONS

A.I. Voronina*, A.V. Rybina

Institute of Biomedical Chemistry, 10 Pogodinskaya st., Moscow, 119121 Russia, *e-mail: an.voronina@list.ru

This paper describes the Retention Time Predictor (RTP) program and web service for predicting the retention time of peptides on a chromatographic column in mass spectrometry experiments. Taking into account post-translational modifications of peptides the program represents a modification of the well-known SSRCalc version 3 (Krokhin, Anal. Chem. 2006, 78(22), 7785-7795). The values of retention coefficients for modified amino acid residues and the algorithm for calculating the isoelectric point value were from the pIPredict program (Skvortsov et al., Biomed. Chem. Res. Meth. 2021, 4(4), e00161). Modifications described in the program include (i) Tandem Mass Tag (TMT) and Isobaric Tags for Relative and Absolute Quantification (iTRAQ) labels; (ii) acetylation, formylation, and methylation of the N-terminal residue and/or lysine side chain; (iii) carbamidomethylation of cysteine, asparagine, and glutamic acid residues; (iv) oxidation and double oxidation of methionine and proline residues; (v) phosphorylation of serine, threonine, and tyrosine residues; (vi) C-terminal amidation of lysine and arginine residues; (vii) formation of propionamide with a cysteine residue. Retention coefficient estimation was based on data from 25 mass spectrometry experiments for which identification was performed from the raw data deposited in the ProteomeXchange database. The RTP program and web service are freely available at http://lpcit.ibmc.msk.ru/RTP.

Key words: peptide retention time; isoelectric point; post-translational modifications; web service

FUNDING

The work was performed within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021-2030) (№ 122030100170-5).

Received: 18.06.2023, revised: 02.07.2023, accepted: 14.07.2023