ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ

ПРЕДСКАЗАНИЕ ЗНАЧЕНИЯ ИЗОЭЛЕКТРИЧЕСКОЙ ТОЧКИ ПЕПТИДОВ И БЕЛКОВ С ШИРОКИМ СПЕКТРОМ ХИМИЧЕСКИХ МОДИФИКАЦИЙ

В.С. Скворцов*, А.И. Воронина, Я.О. Иванова, А.В. Рыбина

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

Представлена шкала «виртуальных» значений pKa для расчёта изоэлектрической точки пептидов и белков, имеющих как химические, так и посттрансляционные модификации (PTM). Обучающая выборка для подбора значений pKa сформирована на основе данных из 25 экспериментов по изоэлектрическому фокусирования пептидов с последующей масс-спектрометрической идентификацией (ProteomeXchange accession codes: PXD000065, PXD005410, PXD006291, PXD010006 и PXD017201). Для всех наборов данных идентификация пептидов по «сырым» масс-спектрометрическим данным проведена заново с целью обогащения выборки пептидами с модификациями. В окончательную обучающую выборку включены пептиды, для которых выполнялись следующие условия: пептид встречался во фракции, для которой величина максимума оценочной функции при идентификации пептида совпадала с максимальным значением представленности («abundance»), пептид встречался более чем в одном эксперименте, причём величина pI между экспериментами не отличалась больше чем 0.15 значений единицы pH. Созданы два варианта шкал. В первом величина pKa зависела только от его положения относительно концов последовательности (N- или C-концевой остаток, либо внутри цепи). Во втором учитывали также влияние соседних остатков. Точность предсказания по второму варианту была выше. Проведено сравнение с другими методами предсказания pI. Несмотря на то, что шкала рассчитывалась по выборке, содержащей только пептиды, она применима и для предсказания pI белков как с наличием PTM, так и без. Создано программное обеспечение для предсказания pI с использованием полученных шкал pKa, доступное по адресу http://pIPredict3.ibmc.msk.ru.

Ключевые слова: пептид; изоэлектрическая точка; посттрансляционные модификации; химические модификации, предсказание свойств

DOI: 10.18097/BMCRM00161

ВВЕДЕНИЕ

Изоэлектрическая точка (pI) – важная физико-химическая характеристика пептидов и белков, которая широко используется в практике современного эксперимента [1-2]. Например, в области протеомных исследований наиболее часто используются такие методы, как двумерный (2D) гель-электрофорез для белков и фракционирование пептидов [3-6] с использованием изоэлектрического фокусирования (IEF) для последующего анализа масс-спектрометрическими методами. Кроме того, значение pI может быть использовано как дополнительный маркер при контроле правильности идентификации пептидов, реже – белков (так как белки в значительной степени подвергаются посттрансляционной модификации (PTM), то определение, какая конкретно протеоформа исследуется, – само по себе часто непростая задача [7]).

Для предсказания значения pI обычно используют два методических подхода. Самый распространённый – использование уравнения Хендерсона-Хассельбаха [8]. Как правило, в случае пептидов и белков используют шкалу значений констант диссоциации (pKa) для отдельных химических групп, рассчитанную заранее. Самой популярной является шкала из работы [9], на базе которой создан широко используемый калькулятор [10] pI белков на Swiss Bioinformatics Resource Portal. Несмотря на то, что авторы [9] прямо пишут о предсказании значения в заданной области pH (от 4 до 7), за давностью лет этот факт уже забылся. Как правило, не учитывают и зависимость величины pKa от температуры и ионной силы раствора (справедливости ради, Chemaxon Marvin Suite [11] учитывает температурную зависимость для предсказания pKa). В тоже время, для белков, в которых потенциально заряженных групп достаточно много, ошибка, вносимая выбранной шкалой, обычно, невелика и лежит в пределах до 0.5 единиц pH. Для широкого спектра задач такая точность может оказаться достаточной. Однако в случае пептидов с ограниченным числом диссоциирующих групп эта ошибка может быть существенно выше.

Существует несколько вариантов формирования шкалы значений pKa: квантово-химические расчёты (точность и время существенно зависят от сложности системы) и на основе эмпирических данных, полученных различными методами. Первый – проведение титрования модельных пептидов [12]. Это самый надёжный метод. Однако он долог, дорог и при использовании небольших выборок не даёт гарантии, что промоделированы все возможные комбинации. К тому же, число вариантов модельных структур растёт экспоненциально по мере увеличения длины пептида. Второй – решение обратной задачи: если для набора пептидов (белков) известны значения pI, то значения pKa подбираются таким образом, чтобы значение средней или среднеквадратичной ошибки было минимальным [13] (математический метод оптимизации

параметров может быть и другим, вплоть до простого перебора значений). Для последнего способа нужно иметь набор большого числа пептидов с известными значениями pI. Такие наборы данных доступны как «побочный продукт» «shotgun»-протеомики с использованием фракционирования при помощи изоэлектрического фокусирования и широко применяются [2, 14, 15]. В случае их использования необходимо учитывать следующее:

1. Большое число пептидов обнаруживается в нескольких фракциях, и нужно чётко понимать критерии, по которым пептиду приписывается конкретное значение pI.

2. Так как фракционирование обычно происходит в пределах определённого диапазона pH («ширина» фракции определяется экспериментатором произвольно в зависимости от задачи и возможностей), имеется априорная ошибка при определении pI для конкретного пептида.

3. Учитывая, что основным методом при приготовлении проб для протеомного анализа является трипсинолиз, существует вырожденность выборки относительно С-концевых аминокислотных остатков (в подавляющем большинстве это аргинин и лизин).

4. Как правило, авторы работ решают свои собственные задачи, и основной является определение по возможности большего числа пептидов, а не собственно значений pI. Точный диапазон значений pH для конкретной фракции не всегда можно выяснить из текста статьи, так как маркёров в пробе нет, а границы диапазона значений зависят от условий проведения эксперимента.

5. При формировании списка пептидов отмечаются только значимые для экспериментаторов PTM (например, фосфорилирование, дезаминирование, N-ацетилирование), при этом игнорируются метилирование, окисление и др. Описание облигатных химических модификаций (например, алкилирование цистеина) могжет быть опущено, хотя в методической части оно и приведено.

6. Пептиды, имеющие величину значений pI меньше или больше чем границы диапазона pH стрипа для IEF, остаются, как правило, в крайних фракциях, но могут встречаться и по всему диапазону.

7. И, наконец, ошибки идентификации пептидов не только возможны, они неизбежно имеются. Например, значение False Discovery Rate (FDR) для пептидов в 5% можно трактовать следующим образом, что априори имеется до 5% ложных идентификаций.

Из положительных сторон данного метода следует отметить, в первую очередь, большой размер выборок, а также наличие в выборках модифицированных пептидов. Это позволяет сформировать шкалу, учитывающую посттрансляционные модификации, что встречается в очень ограниченном числе случаев (например, фосфорилирование у [16] или химическая модификации с использованием TMT (Tandem mass tag) и iTRAQ (Isobaric tag for relative and absolute quantitation) [14]).

Кроме уравнения Хендерсона-Хассельбаха существует и другой подход для предсказания значения pI – применение методов машинного обучения с использованием генетических алгоритмов [17], искусственных нейронных сетей [13] и метода опорных векторов (SVM) [15, 18]. Однако, несмотря на кажущуюся привлекательность и несомненные достоинства этих методов, и у них также есть ряд ограничений и недостатков (переобучение выборки, невозможность предсказывать pI для полипептидов, существенно отличающихся по размеру от тех, что использовали в обучающей выборке, невозможность адаптации модели при введении новых вариантов PTM без полного пересчёта и др.). Что-то можно решить, используя отдельные модели для пептидов и белков, но для данных по белкам проблем ещё больше, чем для пептидов. Обычно значения pI для белков берут из данных, полученных при 2D электрофорезе. При этом часто неизвестно, о какой протеоформе идёт речь. Экспериментаторы не утруждают себя внесением маркёров pH, а привязку к шкале проводят либо по известным белкам, либо по предсказанным значениям, используя калькулятор pI с сайта Swiss Bioinformatics Resource Portal. Нередко публикуются данные, которые являются компиляцией из нескольких объединённых экспериментов, при этом в местах «склеек» могут быть ошибки. В любом случае, наборы данных существенно меньше по размеру, чем для пептидов.

Данная работа посвящена формированию шкалы значений pKa, с использованием которой можно предсказывать величину pI как для пептидов, так и белков с широким спектром PTM и химических модификаций, как преднамеренных, так и спонтанных, наблюдаемых при проведении масс-спектрометрического анализа.

МЕТОДИКА

Формирование обучающей выборки пептидов

Для формирования обучающей выборки были использованы данные экспериментов по масс-спектрометрической идентификации с примененим IEF, депонированные в БД ProteomeXchange (accession codes: PXD000065 [14], PXD005410 [19], PXD006291 [20], PXD010006 [21] и PXD017201 [22]). Всего было обработано 25 наборов данных (табл. 1), однако не все они были использованы для подбора обучающей выборки по причинам, указанным ниже. Так как для данной работы было важно получить максимально возможный спектр модификаций, что не было задачей авторов экспериментальных работ, то идентификацию пептидов провели заново с использованием программы Peaks Studio X Pro [23]. При этом, в качестве основных модификаций задавали химические модификации, специфичные для конкретного эксперимента (алкилирование цистеинов, TMT или iTRAQ метки), фосфорилирование, окисление метионина и дезаминирование. Остальные модификации находили, используя соответствующий модуль программы. Ограничение для идентификации прекурсора было установлено в 5 ppm, точность идентификации фрагментов – 0.01 Da. Были отобраны пептиды с уровнем FDR 5%. В ходе работы анализировали также варианты отбора с уровнями FDR 1% и 0.1%, но при резком сокращении числа вариантов существенного улучшения качества выборки после отбора не наблюдалось. Как правило, каждый идентифицированный пептид мог встречаться в нескольких пробах, ассоциированных с различными значениями pI. Это может быть связано как с ложной идентификаций, так и с тем фактом, что пептид встречается в нескольких соседних пробах с максимумом количества молекул в пробе с pH, наиболее близкой к pI пептида. Поэтому в результирующую выборку отбирали только те варианты, для которых максимум оценочной функции при идентификации пептида совпадал с максимумом площади под кривой

Таблица 1. Наборы «сырых» данных из Protein Exchange использованные в работе.

№ набора данных	Protein Exchange ID	Фрагмент имени файла RAW для идентификации набора данных	Диапазон pH стрипа для IEF	Число пептидов	Число пептидов, не вошедших в обучающую выборку	Число пептидов в обучающей выборке	Число пептидов в обучающей выборке с модификациями*
1	PXD000065	iTRAQ8_200ugIPG439-499	4.39-4.99	3002	2478	524	103
2	PXD000065	plain_200ugIPG440-465	4.4-4.65	9718	9679	68	29
3	PXD000065	plain_200ugIPG37-49_no-pImarkers	3.7-4.9	25018	24790	228	117
4	PXD000065	plain_200ugIPG37-49_with-pImarkers	3.7-4.9	22386	22385	319	2
5	PXD000065	plain_200ugIPG400-425	4.0-.25	14376	14308	81	54
6	PXD000065	plain_200ugIPG420-445	4.2-4.45	13848	13719	39	21
7	PXD000065	plain_200ugIPG370-405	3.7-4.05	15586	15505	129	83
8	PXD0006291	TMT10_set2_500ug_IEF_3-10	3.0-10.0	44509	16092	28417	8705
9	PXD0006291	TMT10_set2_500ug_IEF_37-49	3.7-4.9	52641	25201	27440	9522
10	PXD006291	TMT10_set1_500ug_IEF_3-10	3.0-10.0	53939	22502	31437	9098
11	PXD006291	TMT10_set1_500ug_IEF_37-49	3.7-4.9	43493	17714	25779	9743
12	PXD010006	IPG3-10-50m	3.0-10.0	15462	10179	5283	2738
13	PXD010006	IPG25_37_TMT10_5mg	2.5-3.7	8957	7500	1457	1070
14	PXD010006	non_phospho_500ug_IPG3-10	3.0-10.0	43937	28382	15555	4550
15	PXD005410	TiO2_TMT10_IPG3-10	3.0-10.0	19907	14118	5789	2450
16	PXD005410	TMT_nonphospho_300ugIPG3-10	3.0-10.0	72200	53424	18776	4255
17	PXD005410	TiO2_TMT10_IPG25-37	2.5-3.7	8951	7580	1371	1152
18	PXD000065	iTRAQ8_200ugIPG37-49_nuc-heavy	3.7-4.9	10990	10281	709	272
19	PXD017201	TMT_set1_1mgIPG3-10	3.0-10.0	7743	2968	4775	1588
20	PXD017201	TMT_set2_1mgIPG3-10	3.0-10.0	8591	3024	5567	1990
21	PXD017201	TMT_set3_1mgIPG3-10	3.0-10.0	8041	3104	4937	1744
22	PXD017201	TMT_set4_1mgIPG3-10	3.0-10.0	9632	3905	5727	2018
23	PXD017201	TMT_set5_1mgIPG3-10	3.0-10.0	8359	3098	5261	1918
24	PXD006291	TMT-ctrliii-2hii-6hii-24hiii_400ugIPG6-9	6.0-9.0	43237	29370	13867	3304
25	PXD006291	TMT-ctrliii-2hii-6hii-24hiii_400ugIPG11-6	6.0-11.0	48816	34519	14297	3302
		Всего уникальных пептидов:		372215	298352	73863	25159

Примечание. *- за вычетом облигатных химических модификаций (алкилирование цистеинов, TMT или iTRAQ метки).

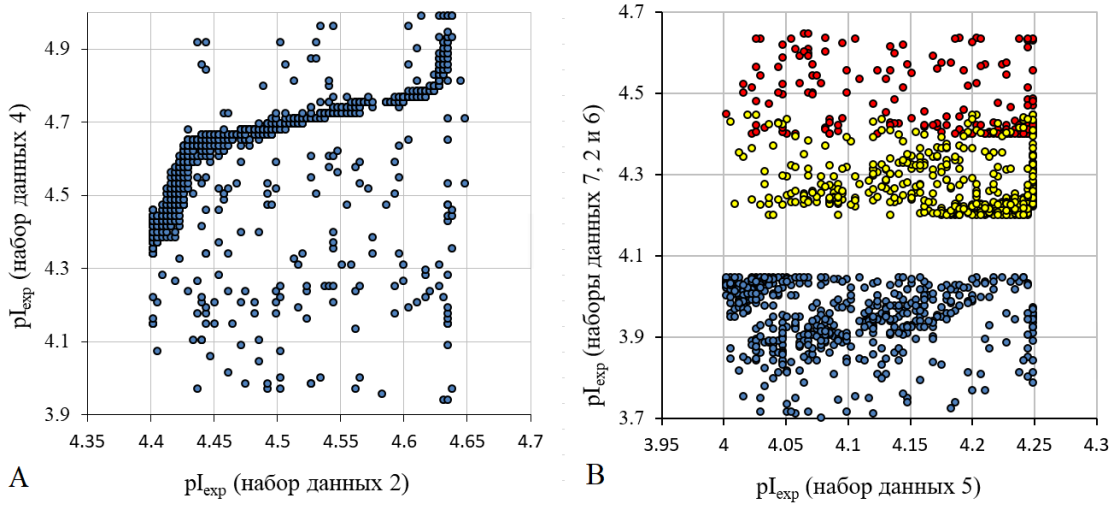

Рисунок 1. Попарное сравнение экспериментально определенных значений pI (pI$_{exp}$) для пептидов, имеющихся в каждой из выборок. А. Выборка **2** относительно выборки **4**. В. Выборки **2** (красный), **6** (жёлтый) и **7** (синий) относительно выборки **5**.

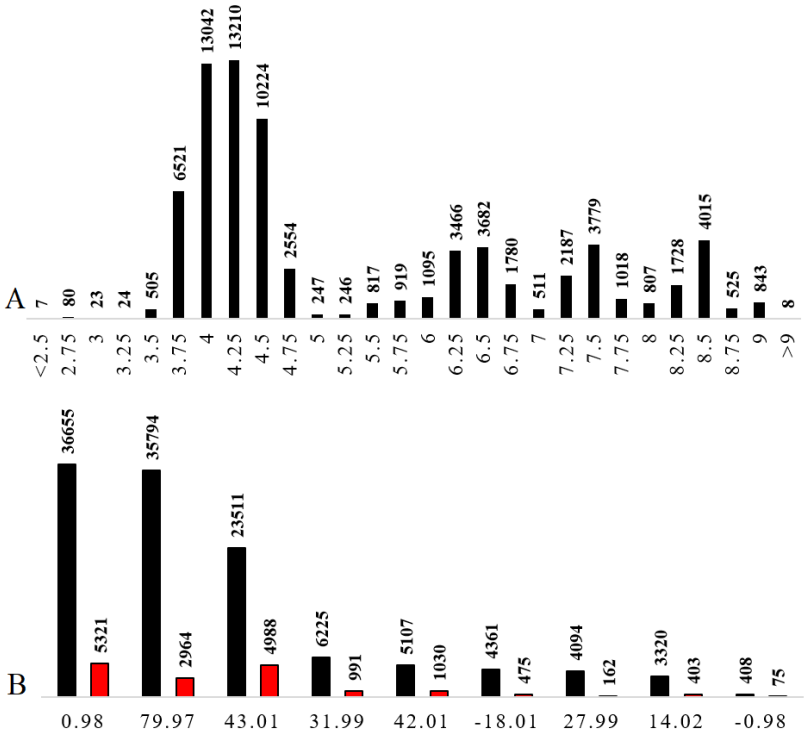

Рисунок 2. Характеристика обучающей выборки. А. Распределение количества пептидов пептидов по значению pI_{exp}. В. Сравнение исходной (черный) и отобранной в качестве обучающей (красный) выборок по общему числу самых распространённых модификаций без учёта алкилирования цистеинов, TMT или iTRAQ меток (указано изменение массы остатка, полная статистика приведена в дополнительных материалах).

распределения иона-прекурсора на хромотограме (данная величина может служить мерой количества пептида). Всего на первоначальном этапе было отобрано 613339 вариантов для 372216 неповторяющихся пептидов (табл. 1 и дополнительные материалы).

Для каждого пептида, идентифицированного более чем в одном наборе данных, разброс значений ΔpI колебался от 0 до 8 единиц pH. Частично это связано с тем, что реальные значения pH в каждом конкретном эксперименте могут отличаться от тех, что заявлены производителем стрипов для IEF, на которых происходит разделение пептидов по pH. В работе [14] один из наборов данных был получен с использованием pI маркеров. Таким образом, было возможно скорректировать значения pI, используя данные для пептидов, встречающихся и в данном наборе, и в том, для которого требуется коррекция. Исключение составил набор **15**, его удалось выровнять по другим наборам данных (использовали набор **16**, полученный теми же авторами). Наборы данных **13** и **17** (заявленный диапазон pH 2.5-3.7) из-за отсутствия значимых пересечений с другими наборами использовали без выравнивания. Следует отметить, что наличие одинаковых пептидов в наборах данных, полученных без облигатных химических модификаций, и с внесёнными метками (либо наборов с различными метками) скорее артефакт пробоподготовки. Ожидается, что все пептиды модифицированы, тем не менее в пробах всегда присутствует некоторая доля непрореагировавших пептидов, и если количество исходного пептида велико, то этой доли может хватить для его идентификации в неизменённом виде.

Конечный отбор для формирования обучающей выборки проводили с использованием, следующих правил: пептид должен был быть идентифицирован в двух и более наборах данных, при этом, если хотя бы один из наборов имел диапазон значений pH от 3 до 10, расхождение по значениям pI должно было быть не более чем 0.15 единицы pH, для диапазонов pH 6-9 и 6-11 – не более 0.1, в случае если совпадение было только в наборах данных с диапазонами pH 3.7-4.9 и уже – не более 0.05. Из сравнения были исключены данные наборов **2**, **5**, **6** и **7** с шириной диапазона значений pH менее 0.5. Причиной стали наличие нелинейных зависимостей при попарном сравнении наборов (рис. 1А), а также тот факт, что они содержали слишком большое количество пептидов, для которых значение pI, определённое в других наборах, не совпало с заявленным диапазоном даже приблизительно. Более того, несмотря на то, что сами по себе эти 4 набора закрывают последовательно 4 диапазона значений pH с небольшим пересечением по границе, значительная часть пептидов находится более чем в одном наборе и часто более чем в двух (рис. 1В). Для каждого конкретного пептида за значение pI принимали медианное значение по всем существующим вариантам. Сформированная обучающая выборка содержит 73863 пептида (см. дополнительные материалы); распределение значений pI и количество пептидов с наиболее представленными модификациями показаны на рисунке 2 (полностью статистика по всем модификациям приведена в дополнительных материалах).

Формирование шкалы значений pKa и тестирование

В ходе работы были исследованы два варианта шкалы значений pKa. Первый, аналогичный шкале, опубликованной нами ранее [13], в которой значения pKa различаются для остатков, расположенных на N- и C-концах пептида и остатков со второго по предпоследний (обозначим её как шкала 3). Второй вариант предполагает учёт влияния соседних остатков. В идеале следует проварьировать все возможные варианты соседних остатков, однако для выборки менее чем в 74 тысячи наблюдений это

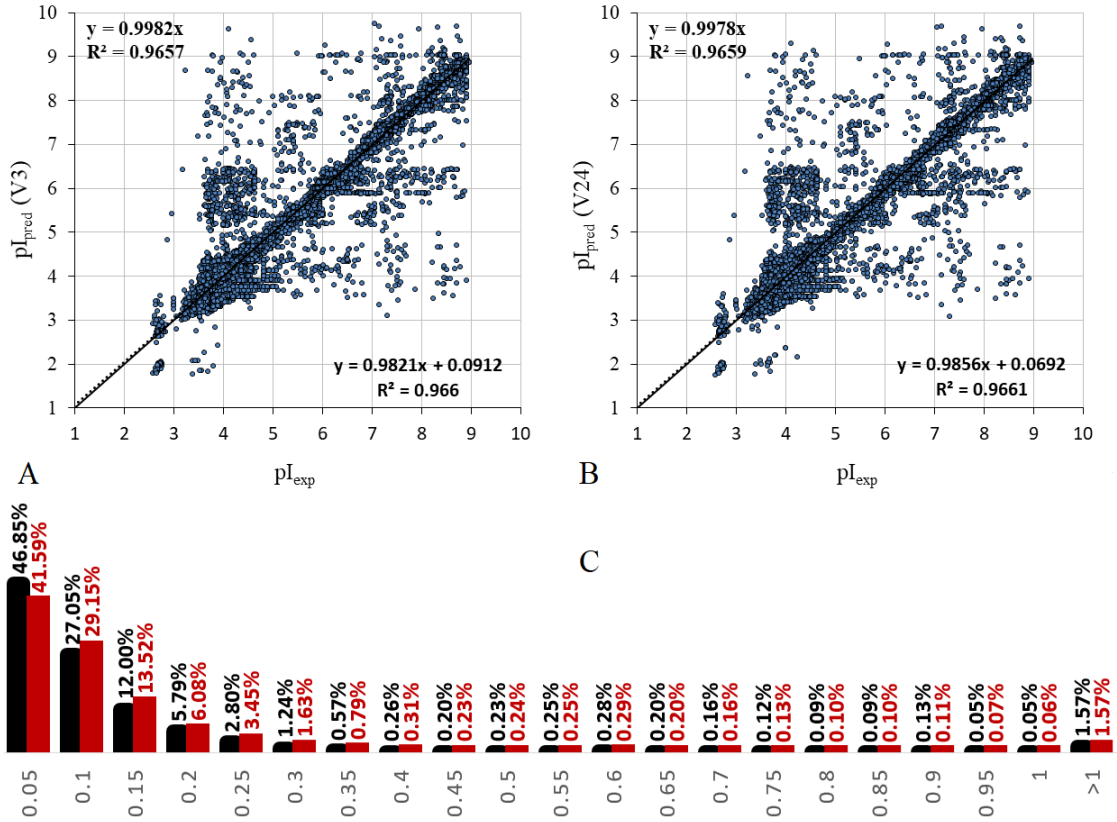

Рисунок 3. Сравнение результатов обучения без (A, красные столбцы на гистограмме) и с учётом (B, чёрные) соседних остатков. C. Распределение пептидов по абсолютной ошибке предсказания при обучении.

невозможно, число вариантов слишком велико (особенно с учётом PTM и химических модификаций). Чтобы сократить число вариантов для соседних остатков введено разделение на 4 группы: 1 – у остатка нет диссоциируемых групп; 2 - у остатка есть протонируемые группы; 3 - у остатка есть диссоциируемые группы; 4 - у остатка есть и протонируемые и диссоциируемые группы. Таким образом, возможно 24 варианта констант для каждого остатка (шкала 24), и хотя для каждого остатка может быть до 3 значений pKa и до 3 pKb, реальное количество переменных в шкале – 1740. Уменьшение числа переменных было получено и за счёт того, что в случае, когда модификация относится к N- или C-концевому остатку и имеет место образование амидной связи (например, TMT или iTRAQ), то такая модификация рассматривается как самостоятельный остаток, а аминокислотный остаток атрибутируется при этом как внутренний.

Обычно для тестирования выборку делят на две части – обучающую и тестовую, но выборка в 73863 наблюдений и так не слишком велика, особенно для шкалы 24. Поэтому в качестве тестовых рассматривали все пептиды по всем 25 наборам данных, не вошедшие в обучающую выборку. Хотя количество ошибочных идентификаций и неправильно определённых значений pI там больше, но в то же время подавляющая часть пептидов идентифицирована правильно, как и значения pI в существенном числе случаев. Мерой при этом может служить процент предсказаний, совпадающих с определёнными экспериментально. Кроме того, провели сравнение с данными, полученными различными методами, приведёнными в работе [15]. Тестирование качества предсказания для белков проводили на выборках из работ [15, 24]. Данные последней были загружены из World-2DPAGE Repository (https://world-2dpage.expasy.org/repository/) [25].

Оптимизацию значений pKa в шкалах проводили с использованием метода Монте-Карло (до 100000 итераций) с последующей процедурой покоординатного наискорейшего спуска (независимо по каждому из значений pKa). В качестве оптимизируемого параметра использовали среднюю абсолютную ошибку предсказания (MAE). Процедуру по всем значениям pKa повторяли циклически до 100 раз или до прекращения изменений оптимизируемой величины. Расчёт значений pI при оптимизации проводили численно (точность до 0.001 величины pH) с использованием уравнения Хендерсона-Хассельбаха:

$$cl = \sum_i \frac{Nb_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKb_i}} + \sum_i \frac{Na_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKa_i}} - Na_i$$

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В результате оптимизации минимальная величина MAE для шкал 3 и 24 составила 0.128 и 0.123 единицы pH. Качество предсказания значений pI на примере обучающей выборки можно оценить по рисунку 3. Значения R^2 обучения приведены на диаграммах в двух вариантах с учётом константы и без. По диаграммам сложно сказать, в чём преимущество шкалы 24, но это видно по распределению AE (рис. 3C). На качество предсказания для выбросов (за которые, следуя работе [15], приняты значения AE, превышающие 0.25) детализация шкалы особо не влияет, но для части

Таблица 2. Распределение (%) по ошибке предсказания pI пептидов, не входящих в обучающую выборку. Использована шкала с учётом соседних остатков.

Абс. ошибка	Набор 1	Набор 2	Набор 3	Набор 4	Набор 5	Набор 6	Набор 7	Набор 8	Набор 9	Набор 10	Набор 11	Набор 12	Набор 13	Набор 14	Набор 15	Набор 16	Набор 17	Набор 18	Набор 19	Набор 20	Набор 21	Набор 22	Набор 23	Набор 24	Набор 25	Медианное значение
0.05	33.2	9.2	33.0	0.7	11.7	12.9	10.8	23.3	30.0	28.5	26.8	27.2	8.3	27.3	30.2	0.1	6.0	44.8	19.1	23.2	18.9	22.9	23.6	13.6	22.8	25.7
0.10	25.8	14.0	23.2	1.3	14.5	17.2	14.7	19.6	22.6	23.7	21.9	24.1	8.1	21.9	28.5	0.2	8.4	29.2	16.9	19.0	18.9	20.2	22.1	14.1	23.0	21.4
0.15	12.2	11.3	13.0	2.6	18.4	18.5	19.1	14.7	13.1	16.1	14.2	16.1	10.0	17.6	13.5	0.2	12.3	11.7	14.9	13.8	14.1	15.5	12.9	14.3	13.2	14.6
0.20	6.7	7.9	6.0	6.2	15.2	11.3	16.7	10.0	6.8	8.4	7.8	10.8	12.9	8.7	7.4	0.2	13.3	5.1	9.8	9.4	8.9	10.6	9.2	13.9	10.0	9.4
0.25	2.7	7.4	2.1	12.1	8.8	6.0	13.1	5.7	4.0	5.1	4.2	6.1	12.8	4.4	4.3	0.3	18.1	1.6	8.0	6.8	7.5	5.4	5.8	10.8	4.1	5.7
0.30	1.3	4.3	1.5	18.7	4.1	2.6	4.8	3.0	2.9	3.2	3.2	3.6	10.7	2.2	2.5	0.2	11.3	0.7	5.5	4.6	5.4	4.0	2.8	5.7	4.5	3.3
0.35	0.9	2.6	1.2	17.8	2.0	1.5	1.9	3.8	1.4	2.0	1.5	2.4	6.0	1.7	1.5	0.3	6.3	0.3	3.7	2.7	3.9	2.4	3.3	2.4	1.9	2.0
0.40	0.7	1.7	1.0	12.1	1.4	1.0	1.0	1.7	1.0	1.1	0.9	1.4	2.4	0.9	0.9	0.3	2.5	0.3	3.0	2.1	2.0	1.5	2.4	1.2	2.3	1.2
0.45	0.8	2.0	0.9	5.2	1.3	1.5	1.0	1.1	0.8	0.8	0.6	0.8	1.4	0.6	0.4	0.2	1.2	0.4	2.4	1.7	1.4	1.3	1.8	0.5	1.6	0.9
0.50	0.6	1.8	1.2	2.1	1.1	1.6	1.2	1.0	0.8	0.6	0.8	0.5	0.9	0.7	0.3	0.2	0.6	0.4	2.0	1.4	1.4	0.9	1.2	0.3	1.4	0.8
0.55	0.9	1.6	1.9	1.1	1.0	1.6	1.4	0.7	0.8	0.5	0.7	0.4	1.3	0.6	0.3	0.1	0.6	0.4	1.1	0.9	1.2	0.7	0.9	0.4	1.0	0.8
0.60	1.3	1.8	2.2	0.9	1.0	1.0	2.1	0.7	0.7	0.5	0.6	0.3	1.4	0.6	0.2	0.1	0.9	0.5	1.1	0.6	0.5	0.4	0.3	0.6	0.7	0.8
0.65	1.6	1.6	1.8	0.9	1.9	1.2	2.9	0.4	0.7	0.4	0.7	0.3	1.0	0.8	0.2	0.1	0.6	0.5	0.3	0.7	0.5	0.5	0.6	0.7	0.6	0.9
0.70	1.3	3.2	1.5	0.8	2.8	1.5	3.0	0.5	0.5	0.3	0.5	0.3	1.6	0.6	0.2	0.1	0.9	0.2	0.7	0.5	0.6	0.5	0.6	1.1	0.4	0.9
0.75	1.5	3.3	1.5	0.9	3.6	2.5	2.0	0.6	0.3	0.5	0.5	0.2	2.1	0.5	0.2	0.1	0.7	0.3	0.6	0.5	0.7	0.6	0.5	1.3	0.5	1.0
0.80	1.1	1.5	1.4	1.0	3.6	4.5	0.7	0.7	0.4	0.4	0.4	0.2	2.3	0.4	0.2	0.0	1.5	0.2	0.6	0.3	0.6	0.5	0.4	1.6	0.2	0.9
0.85	0.8	1.8	0.5	1.4	1.8	3.9	0.2	0.8	0.4	0.4	0.2	0.3	1.7	0.6	0.2	0.1	2.7	0.2	0.5	0.3	0.3	0.3	0.3	1.4	0.4	0.8
0.90	0.7	2.2	0.3	1.8	0.8	2.3	0.1	0.8	0.3	0.4	0.3	0.2	1.4	0.5	0.2	0.1	1.8	0.1	0.4	0.5	0.3	0.5	0.4	1.1	0.4	0.6
0.95	0.3	3.9	0.4	1.9	0.2	1.1	0.1	0.5	0.4	0.4	0.2	0.1	0.5	0.4	0.3	0.1	1.4	0.0	0.1	0.4	0.3	0.2	0.4	0.7	0.2	0.5
1.00	0.5	4.2	0.4	1.3	0.1	0.6	0.0	0.6	0.1	0.3	0.1	0.2	0.4	0.4	0.1	0.1	1.0	0.0	0.2	0.3	0.3	0.2	0.3	0.5	0.2	0.4
>1	5.2	12.8	5.0	9.3	4.8	5.9	3.1	9.8	12.1	6.5	14.0	4.7	13.0	8.7	8.5	97.1	8.0	3.4	9.0	10.5	12.5	11.1	10.2	14.1	10.6	7.4
<=0.25	80.5	49.9	77.4	22.9	68.6	65.8	74.5	73.3	76.4	81.8	75.0	84.3	52.0	79.9	83.9	0.9	58.0	92.3	68.7	72.0	68.3	74.5	73.6	66.6	73.1	76.8
>0.25	19.5	50.1	22.6	77.1	31.4	34.2	25.5	26.7	23.6	18.2	25.1	15.8	48.0	20.1	16.1	99.0	42.0	7.7	31.3	28.0	31.7	25.5	26.4	33.4	26.9	23.2

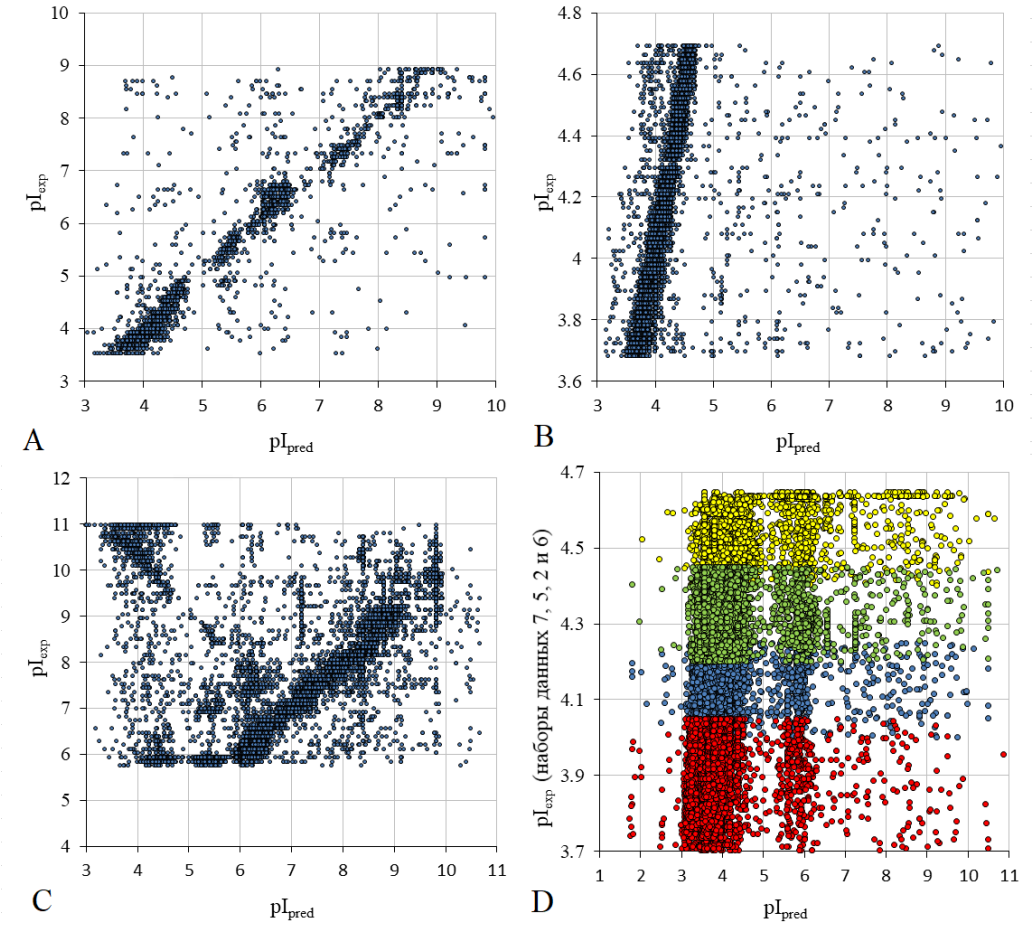

Рисунок 4. Сравнение предсказанных по шкале с учётом соседних остатков и экспериментальных значений pI для пептидов не входящих в обучающую выборку. A. Выборка **19**. B. Выборка **18**. C. Выборка **24.** C. Выборки **2** (жёлтый), **5** (синий), **6** (зелёный) и **7** (красный).

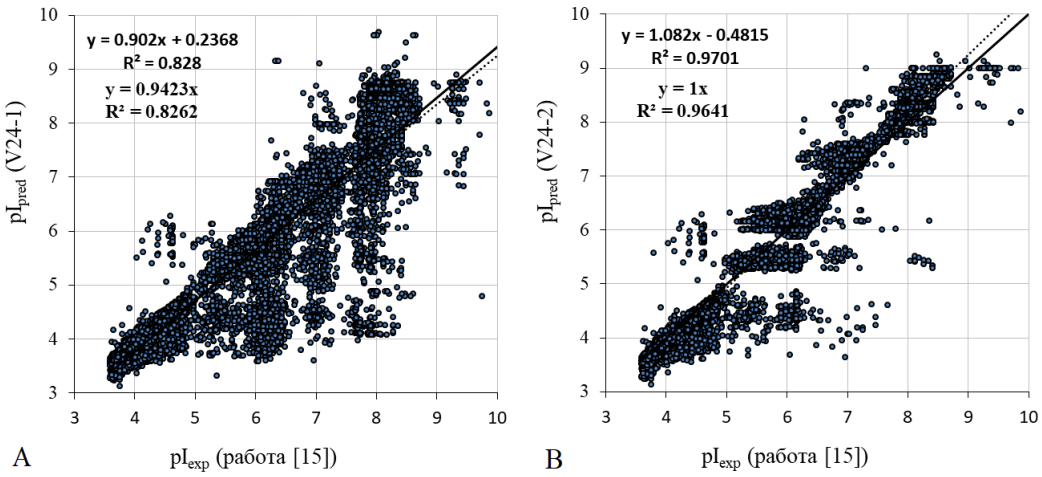

Рисунок 5. Сравнение предсказанных значений pI (pI_{pred}) с экспериментально определёнными (pI_{exp}) для пептидов из 25% тестовой выборки работы [15]. A. Шкала 24 без учёта наличия химических модификаций пептидов. B. Шкала 24 при внесении химических модификаций.

пептидов с AE в пределах 0.25 величина pI предсказывается точнее (в предел до 0.05 попадает на 10% больше). Больших различий между шкалами быть и не может, так как по сути отличие второй важно только для случаев, когда рядом расположены заряженные остатки, а большинство аминокислотных остатков не имеет заряженных групп. Не для всех из 24 вариантов было достаточно наблюдений по каждому из типов остатков. В таком случае недостающие значения дополняли, используя усреднённое значение для других типов остатков, схожих по спектру диссоциируемых и протонируемых групп. Это важно для предсказания пептидов, не входящих в обучающую выборку.

Результаты тестирования для всех 25 наборов данных представлены в таблице 2. При предсказании для наборов данных **2**, **5-7** от 40% до 60% наблюдений попадают в область с AE менее 0.25, а на диаграмме, объединяющей все 4 выборки (рис. 4D), можно видеть общую полосу, соответствующую зоне пептидов, позиционированных

Таблица 3. Распределение по ошибке предсказания pI пептидов из 25% тестовой выборки работы [15]. V24 – варианты шкалы с учётом соседних остатков, полученные в настоящей работе.

AE	IPC2.peptide. Conv2D	%	IPC2.peptide.svr.19	%	IPC2_peptide	%	Bjellqvist	%	Nozaki	%	Thurlkill	%	Sillero	%	Dawson	%	*V24-1*	%	*V24-2*	%	*V24-3*	%
0.05	11121	17.89	10431	17.89	8596	17.89	3374	17.89	3561	17.89	6241	17.89	6417	17.89	7400	17.89	5327	17.89	8800	29.56	13531	45.45
0.10	7654	16.76	7558	16.76	7386	16.76	3760	16.76	4121	16.76	5713	16.76	5908	16.76	6646	16.76	4991	16.76	7997	26.86	7758	26.06
0.15	4326	13.87	4689	13.87	5417	13.87	4332	13.87	4700	13.87	4976	13.87	5048	13.87	4724	13.87	4129	13.87	5743	19.29	3820	12.83
0.20	2313	10.58	2708	10.58	3486	10.58	3807	10.58	4419	10.58	3361	10.58	3076	10.58	2732	10.58	3151	10.58	3057	10.27	1552	5.21
0.25	1305	7.18	1408	7.18	1710	7.18	2862	7.18	3136	7.18	2301	7.18	1718	7.18	1574	7.18	2137	7.18	1269	4.26	708	2.38
0.30	778	5.53	803	5.53	817	5.53	2143	5.53	2138	5.53	1566	5.53	1144	5.53	988	5.53	1647	5.53	664	2.23	373	1.25
0.35	515	3.93	504	3.93	547	3.93	1678	3.93	1572	3.93	854	3.93	792	3.93	576	3.93	1170	3.93	325	1.09	263	0.88
0.40	358	3.12	255	3.12	307	3.12	1336	3.12	1336	3.12	463	3.12	551	3.12	308	3.12	930	3.12	239	0.80	219	0.74
0.45	261	2.58	240	2.58	243	2.58	1028	2.58	812	2.58	247	2.58	445	2.58	279	2.58	768	2.58	196	0.66	235	0.79
0.50	174	2.39	200	2.39	132	2.39	783	2.39	361	2.39	264	2.39	248	2.39	238	2.39	712	2.39	173	0.58	217	0.73
0.55	149	2.19	179	2.19	196	2.19	546	2.19	433	2.19	214	2.19	282	2.19	279	2.19	651	2.19	163	0.55	177	0.59
0.60	104	2.20	126	2.20	109	2.20	525	2.20	357	2.20	190	2.20	304	2.20	184	2.20	654	2.20	164	0.55	171	0.57
0.65	95	1.60	78	1.60	171	1.60	510	1.60	262	1.60	159	1.60	348	1.60	169	1.60	475	1.60	155	0.52	152	0.51
0.70	76	1.17	74	1.17	82	1.17	523	1.17	175	1.17	166	1.17	221	1.17	119	1.17	349	1.17	140	0.47	108	0.36
0.75	58	0.92	42	0.92	62	0.92	442	0.92	221	0.92	121	0.92	128	0.92	149	0.92	275	0.92	102	0.34	66	0.22
0.80	48	0.74	39	0.74	48	0.74	558	0.74	162	0.74	153	0.74	136	0.74	93	0.74	221	0.74	78	0.26	52	0.17
0.85	30	1.04	33	1.04	41	1.04	485	1.04	94	1.04	152	1.04	143	1.04	180	1.04	310	1.04	59	0.20	36	0.12
0.90	30	0.73	28	0.73	35	0.73	335	0.73	75	0.73	166	0.73	149	0.73	120	0.73	217	0.73	47	0.16	36	0.12
0.95	22	0.52	21	0.52	20	0.52	85	0.52	239	0.52	226	0.52	126	0.52	225	0.52	156	0.52	32	0.11	19	0.06
1.00	13	0.31	12	0.31	14	0.31	48	0.31	372	0.31	122	0.31	83	0.31	125	0.31	91	0.31	21	0.07	9	0.03
>1	344	4.75	346	4.75	355	4.75	614	4.75	1228	4.75	2119	4.75	2507	4.75	2666	4.75	1413	4.75	350	1.18	272	0.91
MAE	0.122		0.116		0.139		0.284		0.267		0.254		0.270		0.264		0.382		0.194		0.107	
R^2	0.976		0.974		0.970		0.920		0.919		0.903		0.891		0.883		0.828		0.970		0.977	
Outliers	3055	10.26	2980	10.01	3179	10.68	11639	39.09	9837	33.04	7182	24.12	7607	25.55	6698	22.5	10039	33.72	2908	9.77	2405	8.08

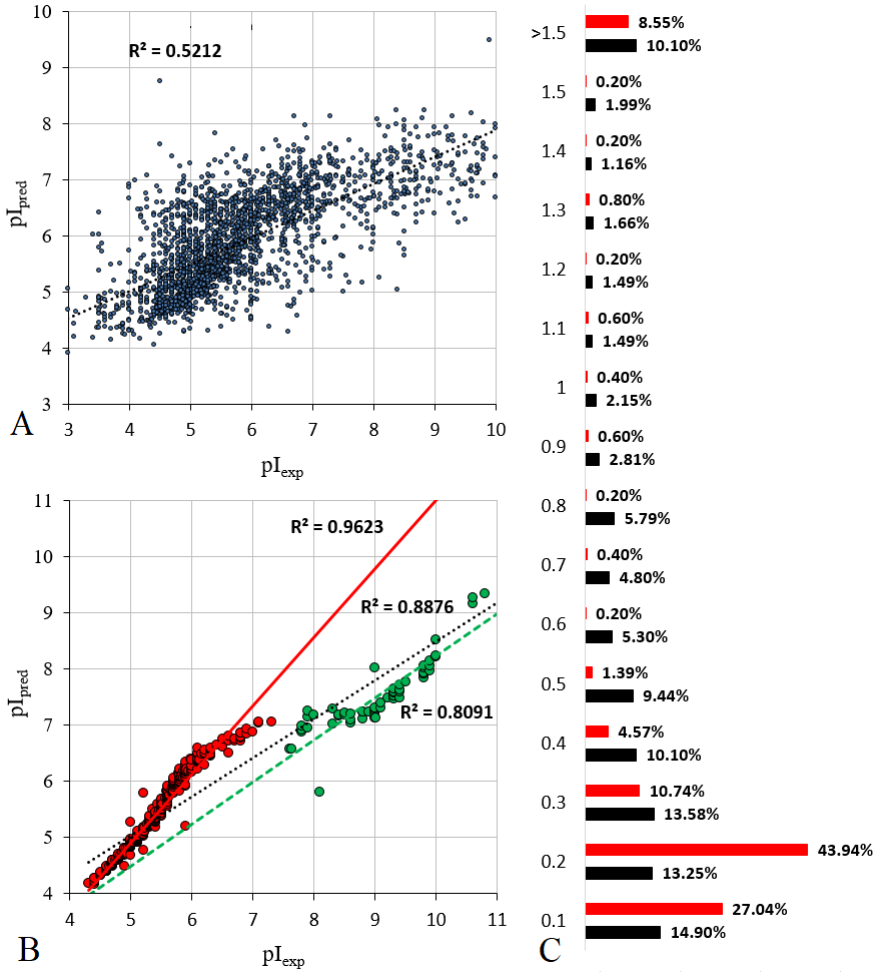

Рисунок 6. Сравнение предсказанных значений pI_{pred} с экспериментально определёнными (pI_{exp}) для выборки белков из работы [15] (A, в гистограмме черный) и белков из работы [27] (B, красный). C – гистограмма распределения абсолютной ошибки предсказания.

в соответствии с их pI, и шлейф из пептидов с более основным pI, которые не вышли за пределы стрипа для IEF. Для всех наборов данных с диапазоном от pH 3 до 10 (пример на рисунке 4A) количество выбросов не превышает 32% в худшем случае (напомним, что все они были обеднены «правильными» значениями). Наборы данных **13** и **17** не были выравнены с остальными и при этом имеют достаточно узкий диапазон значений pI. Так как для этих наборов имеется максимум на графике распределения MAE в районе 0.2-0.25, весьма вероятно, что имеет место системная ошибка по определению диапазона pH. Выборки 1 и 18 (рис. 4B) содержат пептиды с iTRAQ меткой, причём в обучающую выборку вошло менее 10% от общего числа пептидов, тем не менее они имеют лучший результат предсказания. Отдельно нужно отметить данные по наборам **24** (рис. 4C) и **25.** Мы не можем объяснить наличие в них больших групп пептидов, для которых pI предсказывается с точностью до наоборот под углом 90 градусов к основному тренду, ничем иным, кроме как наличием загрязнений. Имея немного представления о практике подобных экспериментов, наличие таких загрязнений можно предположить для набора **25**, данные для отдельных нарезок которого представлены в обратном порядке (от большего значения pI к меньшему). Но так как точный порядок масс-спектрометрического анализа проб неизвестен, то сделать конкретный вывод о возможности подобного загрязнения нельзя. В среднем, даже если учитывать выборки **13** и **17**, в пределах AE до 0.25 единицы pH находятся предсказания для 77% пептидов (если принимать за величину pI для конкретного пептида медианное значение по всем наборам данных).

Сравнение качества предсказания по шкале 24 с предсказаниями с использованием шкал других авторов можно провести по данным работы [15], в которой проводится сравнение собственного метода на основе SVM с большим числом вариантов шкал для расчётов по уравнению Хендерсона-Хассельбаха. Все данные предсказаний для выборки в 25% тестовых пептидов, кроме наших, взяты из файлов дополнительных материалов работы [15]. Распределение средней ошибки и R^2 представлены в таблице 3. В таблицу вошли только 5 лучших методов, использующих уравнение Хендерсона-Хассельбаха. На первый взгляд наша шкала (колонка V24-1 в таблице), несмотря на формально худшее значение R^2 по распределению абсолютной ошибки и количеству выбросов (согласно работе [15] за них принимаются значения, отличающиеся более чем на 0.25 значений pH), даёт результат, сравнимый с методами, использующими уравнение Хендерсона-Хассельбаха, но проигрывает при сравнении с предсказаниями, полученными с использованием SVM. Однако в работе [15] для формирования обучающей и тестовой выборок использовали те же самые наборы данных, которые входят и в число использованных в данной работе [14, 21]. В обеих работах указано, что пробы были обработаны йодацетамидом, а в работе [21] использовали и

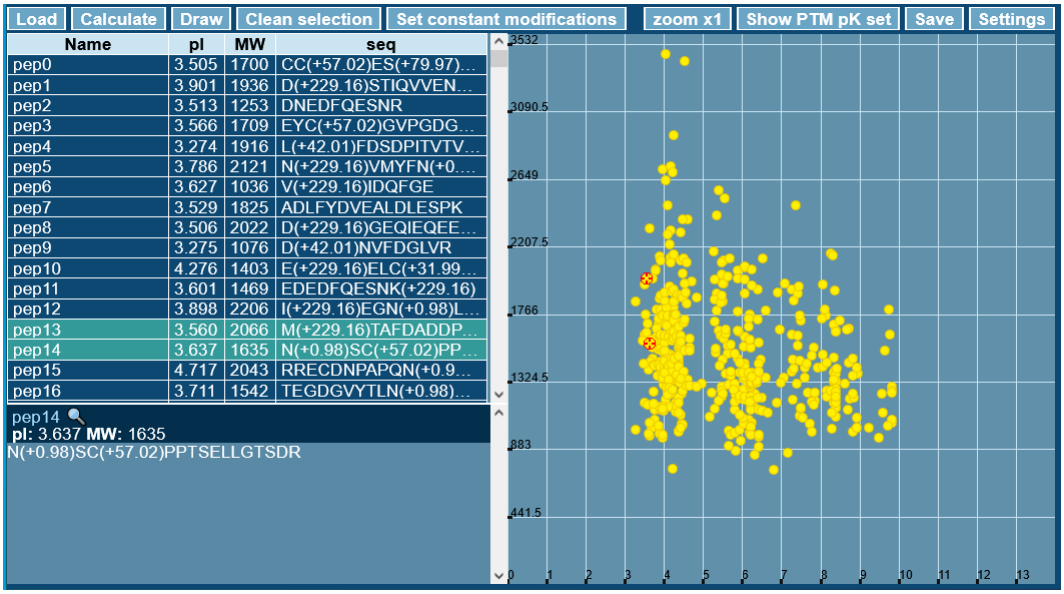

Рисунок 7. Графических интерфейс программы pIPredict 3 с примером предсказания значений pI для выборки из 500 пептидов с модификациями.

TMT-метки. Об этих фактах при описании выборки в работе [15] не упоминается. Если заменить все остатки цистеина на карбамидметилцистеин (+57.02 Da), а к пептидам с $pI_{эксп}$ больше 5 (данные из [14] и [21] имеют пересечение в диапазоне от 3 до 4.9, в котором пептиды могут быть как с TMT, так и без) добавить TMT (+229.16), то результат (колонка V24-2) становится сравним с результатами IPC2. Для наглядности изменения предсказываемых значениях pI с и без учёта PTM приведены на рисунке 5. Следует отметить, что используя вышеописанную методику подбора шкалы pKa и обучающую выборку из работы [15], можно получить более хороший результат (колонка V24-3) для предсказания значений pI тестовой выборки. Однако, использовать его для предсказания можно только для пептидов, имеющих облигатные модификации (карбамидметилцистеин и наличие TMT-меток).

Предсказание pI для пептидов – задача достаточно специфическая, значительно чаще необходимо предсказание pI для белков. Результаты предсказания с использование шкалы 24 для двух выборок белков представлены на рисунке 6. Так как, в отличие от работы [15], выборка с белками ни в какой её части не была использована в качестве обучающей, то для получения более полной картины для предсказания использовали всю выборку, а не 25%-ную часть (рис. 6A). В работе [15] для 25% выборки наилучший R^2=0.59 (для метода на основе уравнения Хендерсона-Хассельбаха и без обучения на выборке белков 0.52), при этом наименьшее число выбросов (абсолютная ошибка >0.5) – 247 (43%). При использовании нашей шкалы на вчетверо большей выборке R^2 = 0.52, а число выбросов –1052 (46%). Данная выборка представляет собой коллекцию данных, полученных разными исследователями; в работе [15] в случае наличия для одного идентифицированного белка нескольких значений pI брали среднее значение. Конкретная протеоформа заведомо не была известна. На рисунке 6B представлены результаты предсказания для белков *Staphylococcus aureus* [24], полученные одной группой (в двух экспериментах с разным диапазоном pH). В данном случае и вероятность того, что белки имеют PTM намного меньше, чем у эукариотических клеток. Соответственно, результат предсказания значительно лучше, хотя тот факт, что результаты получены при анализе двух IEFэлектрофорезов.

ЗАКЛЮЧЕНИЕ

Полученные шкалы могут быть использованы как для предсказания значения pI пептидов с модификациями или без, так и для предсказания pI белков с учётом некоторых PTM и химических модификаций. Для использования данных шкал созданы два варианта программы. Первый (написан на JavaScript) имеет графический интерфейс пользователя (рис. 7), работает как WEB-приложение и доступен по адресу http://pIPredict3.ibmc.msk.ru. Для данного варианта установлено ограничение по размеру загружаемой выборки – 1000 белков (пептидов). Если встречается запись с модификацией, не имеющейся в шкале, то такой остаток можно проигнорировать, либо считать не модифицированным. Второй вариант – исполняемая программа для win32, написанная на C++, работает из командной строки и не имеет ограничений по количеству белков (пептидов). Программа доступна в дополнительных материалах к статье. В случае «неизвестных» модификаций остаток игнорируется. Описание возможных модификаций и выборок данных, использованных в работе, также доступно в дополнительных материалах.

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).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

К данной статье приложены дополнительные материалы, свободно доступные (http://dx.doi.org/10.18097/BMCRM00161) на сайте журнала.

ЛИТЕРАТУРА

1. *Giglione, C., Boularot, A., Meinnel, T.* (2004) Protein N-terminal methionine excision. Cellular and Molecular Life Sciences CMLS, **61,** 1455–1474. DOI: 10.1007/s00018-004-3466-8
2. *Heller, M., Ye, M., Michel, P.E., Morier, P., Stalder, D., Jünger, M.A., Aebersold, R., Reymond, F., Rossier, J.* (2005) Journal of proteome research, **4**(6), 2273-2282. DOI: 10.1021/pr050193v
3. *Pernemalm, M., & Lehtiö, J.* (2013) A novel prefractionation method combining protein and peptide isoelectric focusing in immobilized pH gradient strips. Journal of proteome research, **12**(2), 1014–1019. DOI: 10.1021/pr300817y
4. *Zhu, M., Rodriguez,R., Wehr, T.* (1991) Optimizing separation parameters in capillary isoelectric focusing. Journal of chromatography, **559**, 479–488.
5. *Kirkwood, J., Hargreaves, D., O'Keefe, S., & Wilson, J.* (2015) Using isoelectric point to determine the pH for initial protein crystallization trials. Bioinformatics (Oxford, England), **31**(9), 1444–1451. DOI: 10.1093/bioinformatics/btv011
6. *Branca, R. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff,* Å., *Forshed, J., Käll, L., & Lehtiö, J.* (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nature methods, **11**(1), 59–62. DOI: 10.1038/nmeth.2732
7. *Naryzhny, S. N., Legina, O. K.* (2019) Structural-functional diversity of p53 proteoforms. Biomeditsinskaya khimiya, **65**(4), 263-276. DOI: 10.18097/PBMC20196504263
8. *Po, H. N., Senozan, N. M.* (2001) The Henderson-Hasselbalch Equation: Its History and Limitations. Journal of Chemical Education, **78**, 1499-1503. DOI: 10.1021/ed078p1499
9. *Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S., & Hochstrasser, D.* (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis, **14**(10), 1023–1031. DOI: 10.1002/elps.11501401163
10. *Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., Bairoch, A.* (2005) The Proteomics Protocols Handbook, pp. 571-607. DOI: 10.1385/1-59259-890-0:571
11. Chemaxon, Budapest, Hungary, http://www.chemaxon.com
12. *Patrickios, C. S.* (1995) Journal of Colloid and Interface Science, **175**, 256-256. DOI: 10.1006/jcis.1995.1454.
13. *Skvortsov, V. S., Alekseychuk, N. N., Khudyakov, D. V., Romero Reyes, I. V.* (2015) pIPredict: a computer tool for predicting isoelectric points of peptides and proteins. Biomeditsinskaya khimiya, **61**(1), 83-91. DOI: 10.18097/PBMC20156101083
14. *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
15. *Kozlowski, L. P.* (2021) IPC 2.0: prediction of isoelectric point and pKa dissociation constants. Nucleic Acids Research, **49**(W1, 2), W285–W292. DOI: 10.1093/nar/gkab295
16. *Halligan, B. D., Ruotti, V., Jin, W., Laffoon, S., Twigger, S. N., & Dratz, E. A.* (2004) ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels. Nucleic acids research, **32**(suppl_2), W638-W644. DOI: 10.1093/nar/gkh356
17. *Cargile, B. J., Sevinsky, J. R., Essader, A. S., Eu, J. P., & Stephenson, J. L., Jr* (2008) Calculation of the isoelectric point of tryptic peptides in the pH 3.5-4.5 range based on adjacent amino acid effects. Electrophoresis, **29**(13), 2768–2778. DOI: 10.1002/elps.200700701
18. *Perez-Riverol, Y., Audain, E., Millan, A., Ramos, Y., Sanchez, A., Vizcaíno, J. A., Wang, R., Müller, M., Machado, Y. J., Betancourt, L. H., González, L. J., Padrón, G., & Besada, V.* (2012) Isoelectric point optimization using peptide descriptors and support vector machines. Journal of proteomics, **75**(7), 2269–2274. DOI: 10.1016/j.jprot.2012.01.029
19. *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
20. *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
21. *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
22. *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
23. *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
24. *Plikat, U., Voshol, H., Dangendorf, Y., Wiedmann, B., Devay, P., Müller, D., Wirth, U., Szustakowski, J., Chirn, G. W., Inverardi, B., Puyang, X., Brown, K., Kamp, H., Hoving, S., Ruchti, A., Brendlen, N., Peterson, R., Buco, J., Oostrum, J. v., & Peitsch, M. C.* (2007) From proteomics to systems biology of bacterial pathogens: approaches, tools, and applications. Proteomics, **7**(6), 992–1003. DOI: 10.1002/pmic.200600925
25. *Hoogland, C., Mostaguir, K., Appel, R. D., & Lisacek, F.* (2008) The World-2DPAGE Constellation to promote and publish gel-base d proteomics data through the ExPASy server. Journal of proteomics, **71**(2), 245–248. DOI: 10.1016/j.jprot.2008.02.005

Поступила:	13.10.2021
После доработки:	21.11.2021
Принята к публикации:	23.11.2021

THE PREDICTION OF THE ISOELECTRIC POINT VALUE OF PEPTIDES AND PROTEINS WITH A WIDE RANGE OF CHEMICAL MODIFICATIONS

V.S. Skvortsov*, A.I. Voronina, Y.O. Ivanova, A.V. Rybina

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

The scale of virtual pKa values for calculating the isoelectric point of peptides and proteins with chemical and post-translational modifications (PTM) is presented. The learning set of pKa values is based on data from 25 experiments of isoelectric focusing of peptides with subsequent mass spectrometric identification (ProteomeXchange accession codes: PXD000065, PXD005410, PXD006291, PXD010006 and PXD017201). In order to enrich the resulting sets with peptides containing modifications the identification of peptides was repeated using raw mass spectrometry data of all datasets. In the final learning set have included peptides satisfying the following conditions: the peptide was found in the fraction with scoring function maximum and maximum peptide abundance; the peptide was found in more than one experiment, and differences of the pI value between experiments was less than 0.15 pH unit. Two variants of the scales were created. In the first variant, pKa values depended only on the residue position relative to the ends of the sequence (N- or C-terminal residue or inside the chain). In the second variant, the effect of neighboring residues was also taken into account. The prediction accuracy of the second variant was higher. The comparison with other methods of pI prediction was carried out. Although the scale was calculated from set containing only peptides, it would be applicable for pI prediction of proteins with and without PTM. The software for prediction of pI values using the resulting pKa scales is available at http://pIPredict3.ibmc.msk.ru.

Key words: peptide; isoelectric point; post-translational modifications; chemical modifications; property prediction

FUNDING

The work was done in the framework of the Russian Federation fundamental research program for the long-term period for 2021-2030.

Received: 13.10.2021, revised: 21.11.2021, accepted: 23.11.2021