Биомедицинская химия, 2015 том 61, вып. 1, с. 83-91.

БИОИНФОРМАТИКА

УДК 577.332

pIPREDICT – КОМПЬЮТЕРНАЯ ПРОГРАММА ДЛЯ ПРЕДСКАЗАНИЯ ИЗОЭЛЕКТРИЧЕСКОЙ ТОЧКИ ПЕПТИДОВ И БЕЛКОВ

В.С. Скворцов*, Н.Н. Алексейчук, Д.В. Худяков, И.В. Ромеро Рейес

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

Данные о приблизительном значении pI, полученные в ходе фракционирования пептидов методом изоэлектрического фокусирования, могут быть с успехом использованы для расчёта шкалы значений pKa аминокислотных остатков, на основании которой можно предсказать pI для любого произвольного пептида. Такого рода данные содержат информацию о различных посттрансляционных модификациях (PTM), благодаря чему предсказания pI могут быть осуществлены для широкого спектра белковых форм. В настоящей работе расчёт шкалы значений pKa проведён на основании выборки в 13448 пептидов, 300 из которых имели PTM значимые для расчёта pI. Константы для N-концевых и C-концевых аминокислотных остатков учитывались самостоятельно. Сравнительный анализ показывает, что использование нашей шкалы повышает точность предсказания pI как для пептидов, так и для белков, с успехом конкурируя как с традиционными шкалами, так и с узкоспециализированными методами предсказания, такими как методы опорных векторов и искусственных нейронных сетей. Использовать данную шкалу для предсказания pI можно с помощью разработанной нами программы с графическим интерфейсом, написанной на языке JAVA в виде исполняемого jar-архива. Программа свободна для использования академическими пользователями и доступна по адресу http://www.ibmc.msk.ru/LPCIT/pIPredict. В программе реализованы также возможности предсказания по ряду других широко распространённых шкал, учёта некоторых PTM и добавления собственных вариантов шкалы.

Ключевые слова: пептид, изоэлектрическая точка, посттрансляционные модификации, предсказание свойств.

ВВЕДЕНИЕ

Белки и пептиды – амфотерные электролиты, которые являются удобными объектами для фракционирования с помощью изоэлектрического фокусирования. Это свойство активно используется в практике современного эксперимента [1-2]. В качестве основных примеров наиболее часто используемых методов можно назвать двумерный (2D) гель-электрофорез для белков и фракционирование триптических пептидов для последующего анализа масс-спектрометрическими методами в протеомике. Кроме того, значение изоэлектрической точки может быть использовано как дополнительный маркер при контроле правильности идентификации пептидов, реже – белков (белки в значительной степени подвергаются посттрансляционной модификации, и, как правило, исследователи не знают точно, с какой формой они работают в данный момент времени).

Теории буферных растворов и уравнению Хендерсона-Хассельбаха более 100 лет [3]. Эти знания входят в базовый набор для каждого химика и биолога, и в определённой степени приобрели признаки монументальности. Многие исследователи не обращают внимания, что и само это уравнение является аппроксимацией, и что значения констант диссоциации, использующиеся в расчётах, являются эмпирическими, а для пептидов и белков часто усреднёнными. За скобками, как правило, остаются зависимости от температуры и ионной силы раствора. Всё это известные вещи, однако если специализация авторов не связана с данной областью, то они не обращают на это внимания. Например, авторы работы [4], предложившие

* - адресат для переписки

шкалу констант диссоциации (pKa) для белков, на базе которой создан самый популярный калькулятор [5] изоэлектрической точки (pI), прямо пишут о "предсказании" значения в заданной области pH (от 4 до 7). Исследователи же в области протеомики чаще употребляют термин "теоретический расчёт", используют предсказанные значения для калибровки по pH, и по всему диапазону значений pH [6]. К счастью, для белков, в которых потенциально заряженных групп достаточно много, ошибка, вносимая выбранной шкалой, как правило, невелика. По различным оценкам, для различных шкал средняя ошибка составляет от 0,15 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.

Существуют 2 принципиальные возможности получения шкалы значений pKa для расчётов, в основе которых эмпирические данные (квантово-химические расчёты мы рассматривать в данной работе не будем). Первая – проведение титрования модельных пептидов [7]. 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. Вторая – решение обратной задачи: если для набора пептидов (белков) известны значения pI, то значения pKa подбираются таким образом, чтобы значение среднеквадратичной ошибки было минимальным (математический метод оптимизации параметров может быть и другим). Для этого необходим анализ большого числа пептидов с известными значениями pI. Такие данные в настоящее время доступны как "побочный продукт" "shotgun"-протеомики с использованием фракционирования при помощи изоэлектрического фокусирования.

Однако, эти данные эти могут быть неоднозначными в силу разных причин:

1. Использование данных для белков проблематично, так как неизвестно какую форму белка исследовали авторы (масс-спектрометрические методы, как правило, не позволяют различать изоформы и белки с пост-трансляционными модификациями (PTM) и отдельными аминокислотными заменами). На уровне пептида такой проблемы нет, если методика эксперимента описана адекватно.

2. Ряд пептидов обнаруживаются в нескольких фракциях, так что придать пептиду конкретное значение pI затруднительно.

3. Так как фракционирование обычно происходит в пределах определённого диапазона pH, имеется априорная ошибка при определении pI для конкретного пептида. Причём в разных работах она разная.

4. 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).

5. Так как авторы работ решают свои собственные задачи, и основной является определение по возможности большего числа пептидов, точный диапазон значений pH для конкретной фракции по тексту статьи не всегда можно установить.

6. И, наконец, возможны также ошибки идентификации пептидов и технические ошибки при публикации данных.

В то же время, из положительных сторон следует отметить, в первую очередь, большой размер выборок, а также наличие в выборках модифицированных пептидов, что позволяет формировать шкалу, учитывающую посттрансляционные модификации.

В ходе данной работы такая шкала была подобрана и вместе с рядом других методов предсказания pI используется в разработанной авторами программе pIPredict.

МЕТОДИКА

В качестве основных в работе использованы 2 выборки пептидов. Данные выборки имеют общее свойство: они уже были использованы в работах по разработке методов предсказания pI для пептидов ранее другими авторами. Первая, содержащая данные о 7390 немодифицированых пептидов была использована в работе Perez-Riverol с соавторами [8], для предсказания pI при помощи метода опорных векторов (MOB). Вторая [9], посвящённая созданию программы калькулятора pI с использованием уравнения Хендерсона-Хассельбаха, содержит данные о 5758 немодифицированных пептидах, а также о 300 пептидах с PTM, включая такие как N-концевое ацетилирование, фосфорилирование и окисление остатков метионина. Методы предсказания, описанные в данных работах, наряду с рядом других, реализованных в программе pIPredict, использованы для сравнения с полученными результатами.

Несколько слов об атрибутировании экспериментальных значений pI для фракций. Далеко не все авторы приводят диапазон значений pH для конкретной фракции;

как правило, указывается только данные по маркировке стрипа на котором проводилось фракционирование. Это закономерно, так как большинство авторов не интересуют конкретные значения pI. Они решают другие задачи. Однако, значительное несоответствие предсказанной величины pI для пептида фракции, в которой его обнаружили, может быть индикатором возможной ошибки при идентификации. Тем не менее, зная параметры эксперимента, можно каждой фракции приписать конкретное значение pI с ошибкой меньшей чем ошибка вносимая за счёт величины диапазона значений pH для конкретной фракции.

Построение нейросетевой модели

При построении искусственной нейронной сети (ИНС) в качестве независимых переменных был использован спектр аминокислотного состава каждого пептида, причём учитывалась локализация аминокислотного остатка на N-конце, на C-конце и "внутри" пептида. Рассматривались только остатки, содержащие диссоциируемую (протонируемую) группу. Следовательно, для внутренних аминокислотных остатков учитывались только D, E, Y, C, R, K и H. Таким образом, для каждого пептида был получен вектор из 47 значений. Методика построения модели и собственная программная реализация ИНС подробно описаны ранее [10-11]. В качестве обучающей была использована выборка из 7390 немодифицированных пептидов [8].

Оптимизация значений pKa для виртуальной выборки пептидов

При подборе значений pKa для уравнения Хендерсона-Хассельбаха в качестве процедуры оптимизации использовали метод наименьших квадратов и процедуру наискорейшего спуска. Для проверки адекватности метода подбора значений была случайным образом создана виртуальная выборка объёмом 4800 пептидов от 15 до 30 аминокислотных остатков длиной. Для каждого из пептидов были предсказаны значения pI методом Bjellkvist и соавторы [4] и с использованием метода реализованного в вычислительном модуле Marvin calculation plugin [12]. Первый метод в основе своей имеет конечную шкалу значений pKa для аминокислотных остатков. В основе второго лежит предсказание pKa для каждой диссоциируемой (протонируемой) группы. В идеале, для первого варианта процедура подбора должна найти табличные значения. Второй вариант имитирует реальную ситуацию, когда фактически каждая группа уникальна и подбирается усреднённое значение.

Как и в случае ИНС, учитывалось положение аминокислотного остатка на N-конце, на C-конце или "внутри" пептида, однако в данном случае, если N- или C-концевой остаток имел 2 заряженные группы, они учитывались независимо, и, таким образом, рассчитывалось 61 значение pKa.

Построение шкалы значений pKa по выборке экспериментальных значений

Расчёт шкалы для выборок пептидов с известными экспериментальными значениями проводили так же как и для виртуальных выборок. Однако, для модифицированных аминокислотных остатков были введены собственные типы, позволяющие отличать модифицированные остатки от немодифицированных. Кроме того, для остатка фосфорной кислоты каждая из OH-групп была самостоятельной. Таким образом, число табличных значений pKa было увеличено до 86. Следует отметить, что в выборках представлены не все теоретически возможные варианты положения аминокислотных остатков, это следует учитывать при использовании рассчитанной шкалы для предсказания pI.

Оценка возможности использования шкалы для предсказания изоэлектрической точки белков

Априорно можно утверждать, что методы опорных векторов [8] и ИНС не могут быть использованы для предсказания pI белков. В то же время рассчитанная на основе пептидов шкала по сути своей ничем не отличается от любой другой использующейся для предсказания pI другими авторами. Для того чтобы понять насколько хорошо она будет работать на белках, был проведён тест на выборке из 1100 пептидных форм (441 белков) протеома *Fusarium graminearum* идентифицированных в работе Pasquali и соавторы [13]. Следует отметить, что многие белки в выборке представлены большим числом вариантов, и какой из них кокой форме соответствует неизвестно. Всего в выборке 252 белка представлены в виде единственной формы, 155 имеют до 5 форм, 44 от 5 до 30 форм.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Нейросетевая модель предсказания pI пептидов

Конечная нейросетевая модель была отобрана по наименьшему значению среднеквадратичного отклонения (RMSE) для контрольной выборки (10% отобранных случайным образом наблюдений, R^2=0,98;

RMSE=0,16). Она содержит 7 нейронов в скрытом слое. Примечательно, что если поделить выборку пополам и перенастроить веса нейронной сети по половинам выборки, сохранив её архитектуру, то R^2 предсказания для второй половины составит 0,981 и 0,983 соответственно.

На рисунке 1 хорошо видно, что нейросетевая модель при сравнении с методами, описанными в работе X005, даёт предсказания лучше согласующиеся с экспериментом, выигрывая по RMSE, в том числе, и у метода опорных векторов. Кроме того, и метод опорных векторов, и ИНС не имеют систематической ошибки, наблюдаемой в диапазоне значений от 3 до 5,5 единиц pH (эту ошибку можно увидеть уже в оригинальной работе Bjellqvist и соавт., выполненной на белках [4]). Также нейросетевая модель, в отличие от 3-х остальных, не даёт “ступеньки” по средним значениям в области значений pH от 8 до 9.

Рассмотрим теперь предсказание с использованием ИНС для независимой выборки из 5758 немодифицированных пептидов (рис. 2). При сравнении с предсказанием по шкале Bjellqvist с соавт., нейронная сеть даёт несколько лучший результат по R2, и не даёт системной ошибки в области pH от 3 до 5. В то же время, нейронная сеть демонстрирует системную ошибку в области pH>5 и большую дисперсию при предсказании для каждой конкретной фракции. В среднем по 10 фракциям стандартное отклонение составляет 0,31 для нейронной сети и 0,23 для метода Bjellqvist.

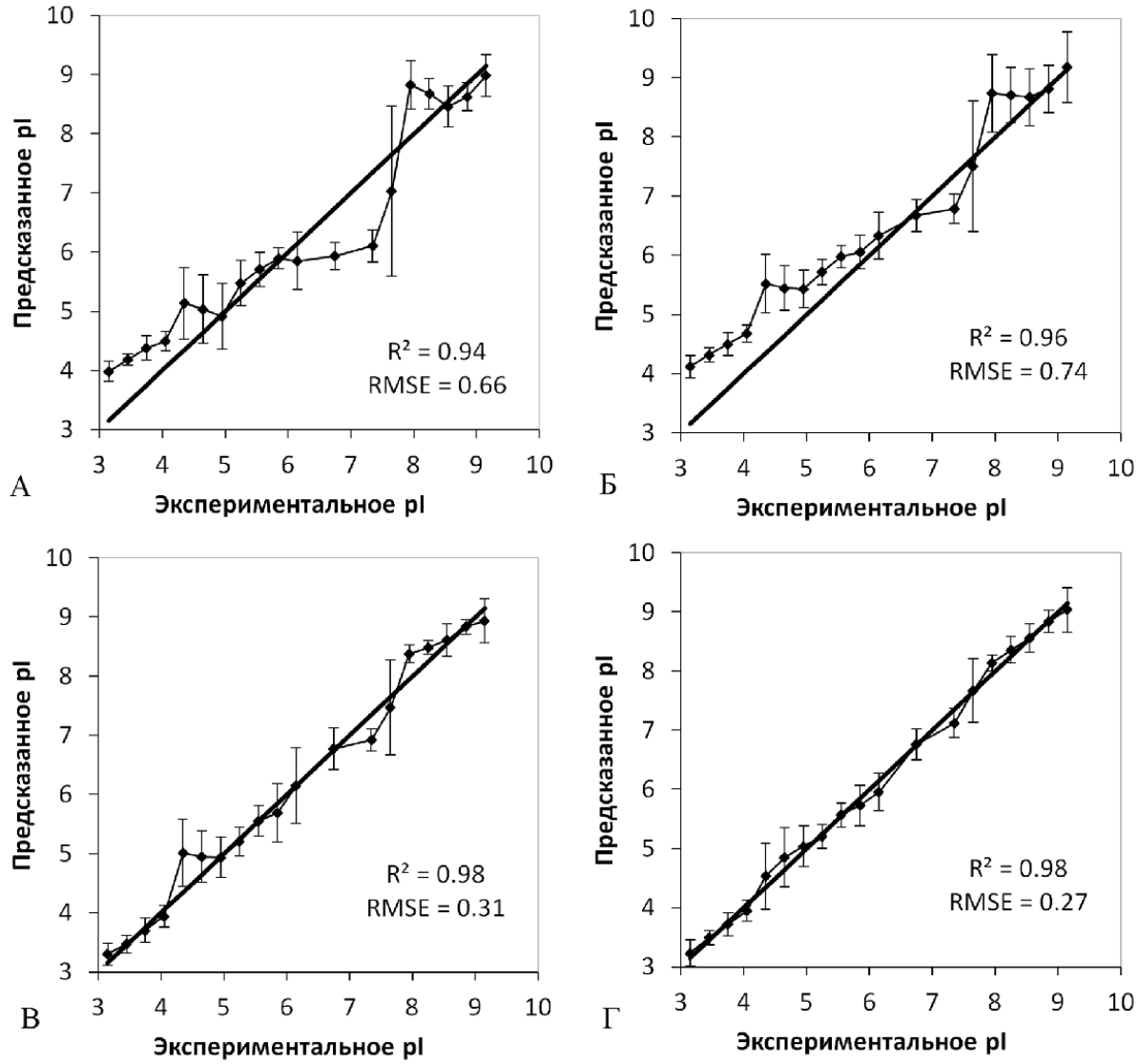

Рисунок 1. Сравнение экспериментальных и предсказанных значений pI для выборки из 7390 пептидов показанных в работе Perez-Riverol и соавт. и в данной работе (ИНС).Приведены средние значения и величина дисперсии по отдельным фракциям. RMSE - среднеквадратичное отклонение. А - Bjellqvist и соавт., Б - Cargile и соавт., В – МОВ, Г – ИНС.

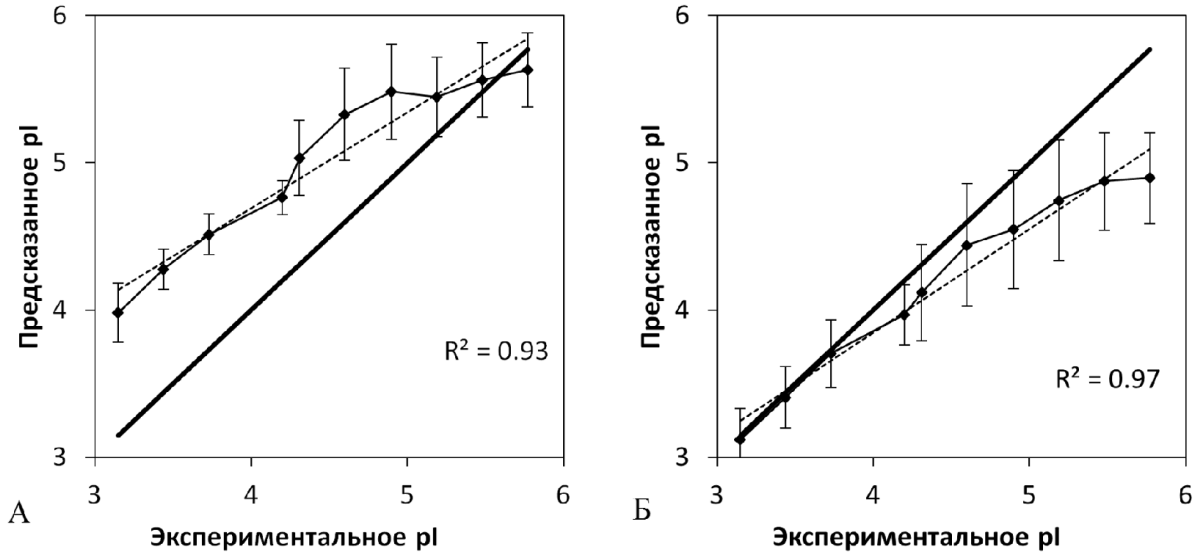

Рисунок 2. Сравнение экспериментальных и предсказанных значений pI для выборки из 5758 пептидов полученных методами Bjellqvist и соавт. и ИНС. Приведены средние значения и величина дисперсии по отдельным фракциям. А - Bjellqvist и соавт., Б - ИНС.

Следует отметить также тот факт, что нейросетевые модели обычно очень чувствительны к входным данным, и если они выпадают из диапазона значений, использовавшихся при обучении сети, то ошибка может быть существенной. Нейронные сети хороши в области интерполирования, но как правило, не работают за пределами области приложения. Таким образом, в данном случае нейросетевую модель скорее стоит рассматривать как имитатор процедуры фракционирования, так как пептиды ведут себя в них схожим образом. На большом объёме данных результат в целом хороший, в то время как для единичного пептида нет критерия достоверности предсказания.

Подбор значений pKa для создания шкалы на основе выборки пептидов с известными pI

Результаты оптимизации процедуры подбора параметров для настраиваемой шкалы pKa представлены на рисунке 3. При подборе табличных значений pKa для предсказаний, сделанных методом Bjellqvist с соавт. процедура восстанавливает значения pI со средней точностью 0,04. Наличие небольшой ошибки связано, в первую очередь, с ошибками

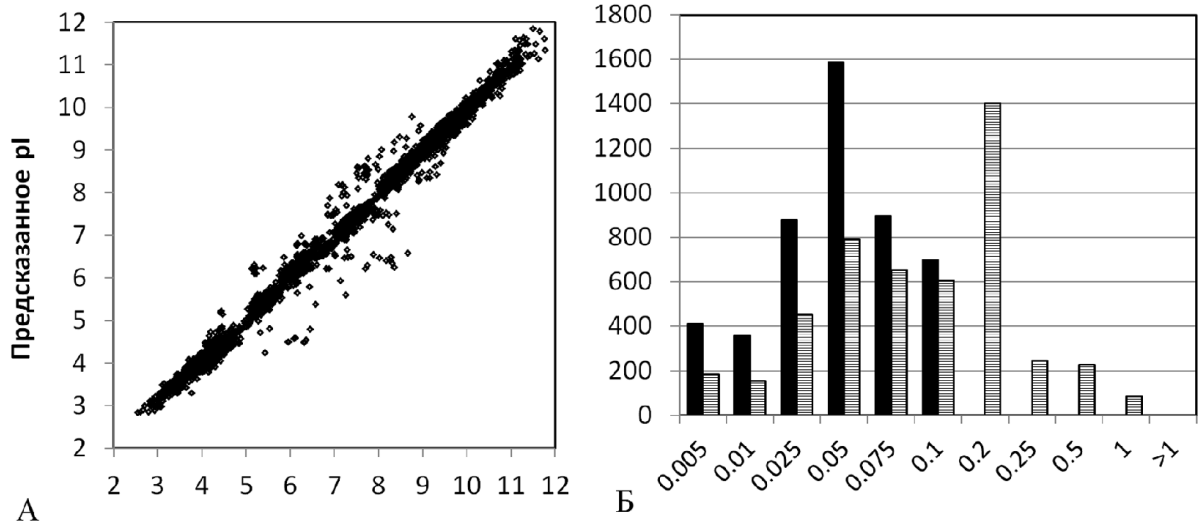

Рисунок 3. Сравнение предсказанных по восстановленным шкалам pI для выборки из 4800 модельных пептидов. А - рассчитанные Marvin plugin и предсказанные имитационной шкалой значения pI. Б - распределение ошибки предсказания для восстановленных значений заданных методами Bjellqvist и соавт. и Marvin plugin.

округления при расчёте pI, так как вычисляется величина не аналитически, а с использование дихотомической процедуры конечной точности (в работе достаточной точность при расчёте pI считалась величина 0,01). Параметры, полученные для предсказаний с использованием Marvin plugin, дают высокую корреляцию с собственно марвиновскими значениями, несмотря на то, что число вариантов pKa по сравнению с числом рассчитываемых в Marvin plugin сильно ограничено. Следует отметить, что сами по себе предсказания, сделанные Marvin plugin на выборке в 7390 пептидов, показывают результат, уступающий только ИНС и МОВ (рис. 3), и лучший, чем любая из шкал, использованных в программе pIPredict.

Так как процедура подбора доказала свою состоятельность, то следующим шагом стала подборка шкалы pKa на основе выборки из 7390 пептидов (рис. 4) для предсказания изоэлектрической точки немодифицированных пептидов. В качестве контроля приведены данные по предсказанию на основе полученной шкалы pI для 5300 немодифицированных пептидов из второй выборки. Результат предсказания, полученный с использованием данной шкалы, даёт наилучший результат по сравнению с имеющимися в программе pIPredict шкалами (табл. 1), и проигрывает по формальным показателям только методу ИНС. Однако, учитывая ограничения, описанные выше, метод ИНС лучше использовать как дополнительный фильтр достоверности результата, а не как самостоятельный метод предсказания.

Следует ещё раз подчеркнуть, что использованная выборка представляет собой набор триптических пептидов, которые специфическим образом обработали при подготовке к фракционированию. В ней практически отсутствуют пептиды, содержащие цистеин (есть всего один содержащий 2 остатка цистеина, но с точки зрения статистики этого явно недостаточно). Вероятнее всего это связано с тем, что после химической обработки пробы,

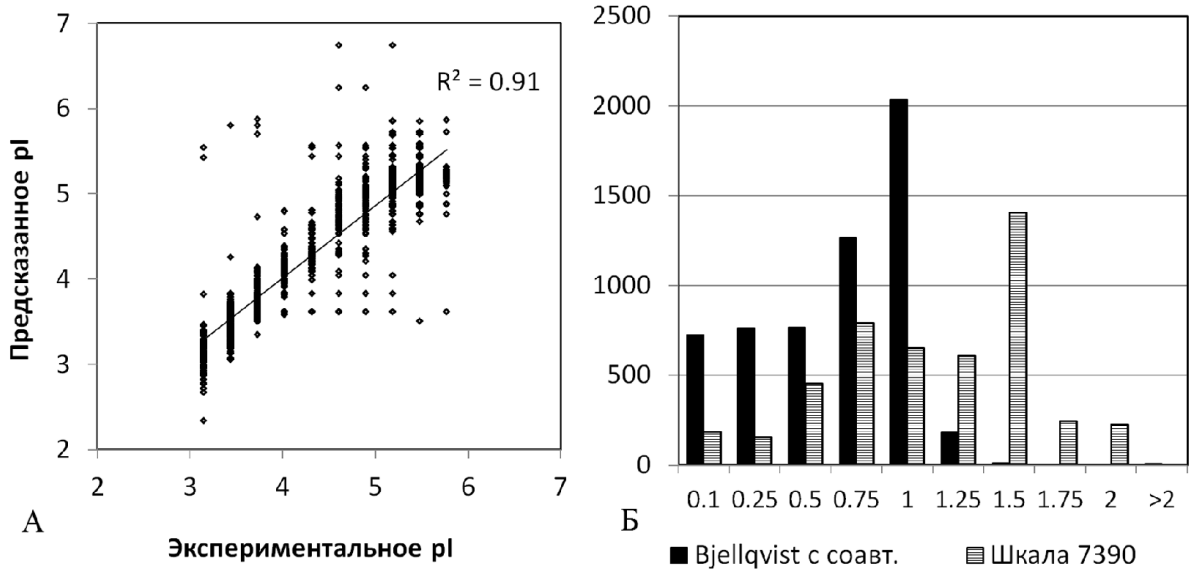

Рисунок 4. Предсказание pI для выборки из 5758 немодифицированных пептидов по шкале, рассчитанной по выборке в 7390 пептидов. А - сравнение предсказанных и экспериментальных значений. Б - распределение ошибки предсказания в сравнении с методом Bjellqvist и соавт.

Таблица 1. Предсказания pI для выборки из 5758 немодифицированных пептидов различными методами.

Метод предсказания	R^2	Средняя ошибка (MSE)	R^2 для средних значений по фракциям	Средняя дисперсия по фракциям
Bjellqvist и соавт.	0,839	0,558	0,933	0,226
ИНС	0,905	0,393	0,972	0,305
По выборке Marvin plugin	0,848	0,515	0,948	0,314
По выборке 7390	0,9	0,243	0,958	0,214
По объединённой выборке	0,928	0,196	0,972	0,204

все остатки цистеина были модифицированы до карбоксиметилцистеина [2, 8]. Кроме того, только 100 из 7390 пептидов в выборке не содержат на С-конце остатки аргинина или лизина, что также ухудшает результат предсказания. Последнее можно избежать, если использовать данные по фракционированию нетриптических пептидов. Для предсказания pI для пептидов, полученных в результате трипсинолиза белков, подвергнувшихся пострансляционным модификациям, мы объединили выборки 1 и 2. В результате точность предсказания для немодифицированых пептидов увеличилась (табл. 1, рис. 5) и стало возможным делать предсказания для пептидов, ацетилированных по N-концу и/или имеющих фосфорилированные остатки серина и треонина (R2 для пептидов с PTM равно 0,88).

Использование расчётной шкалы для предсказания pI белков

На имеющемся наборе белков ни один из методов не даёт таких же хороших результатов, как на выборках пептидов (табл. 2). Отчасти это связано с тем, что неизвестно, какая точно форма белка была в эксперименте детектирована. Кроме того, как правило, при определении экспериментальных значений pI, первичная привязка изображения геля проводится по реперным значениям. Эти значения являются предсказаниями, выполненными с помощью какого-либо метода на основе последовательности, взятой из базы данных (FGDB в данном случае [14]). Чаще всего, используется калькулятор с сервера expasy.org [5]. Последнее обстоятельство может вносить искажения при сравнении. Например, из таблицы 2 видно, что, несмотря на то, что коэффициент детерминации для предсказания методом Bjellqvist и соавт. самый низкий, средняя ошибка (MSE) минимальная. Авторы, вероятно, использовали в своей работе сервер expasy.org базирующийся на данном методе, и в данных для экспериментальных pI содержится систематическая ошибка, внесённая ошибкой предсказания для реперных точек.

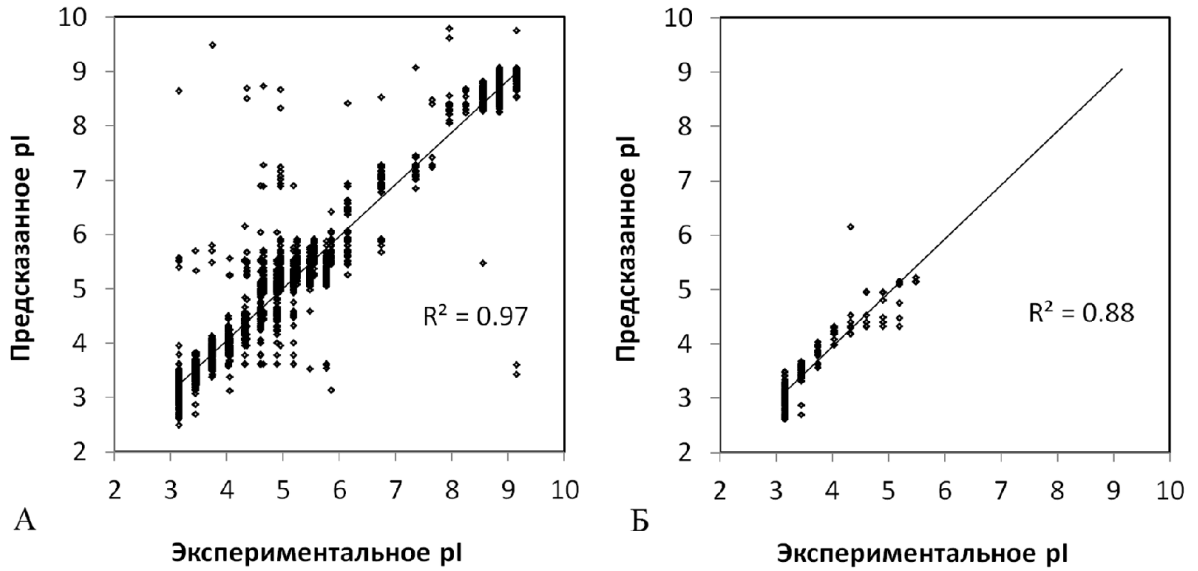

Рисунок 5. Предсказание pI по шкале, рассчитанной по объединённой выборке. А - сравнение предсказанных и экспериментальных значений для полной выборки. Б - для 300 пептидов с PTM.

Таблица 2. Предсказания pI для выборки белков (1100 протеоформ) различными методами.

Метод предсказания	R^2	Средняя ошибка (MSE)	R^2 (удалённый Met)	MSE (удалённый Met)
Bjelqvist и соавт.	0,619	0,529	0,623	0,538
Turklwill и соавт.	0,677	0,597	0,677	0,597
НВСР	0,645	0,613	0,644	0,616
По выборке Marvin plugin	0,706	0,683	0,704	0,682
По выборке 7390	0,66	0,855	0,664	0,875
По объединённой выборке	0,717	0,703	0,718	0,716

Результат можно немного улучшить, если внести в расчёт имитацию отщепления N-концевого остатка метионина. Об этой посттрансляционной модификации часто забывают при состыковке программ расчётов с получением последовательностей из публичных баз данных, созданных на основе геномной информации. В то же время, эта модификация очень распространена [1] и для отдельных белков разность в предсказываемых значениях pI может доходить до 0,5 единиц pH. Разумеется, это актуально только для тех шкал, для которых значения pKa для различных аминокислотных остатков зависят от того находятся ли они на N-конце или нет.

В конечном итоге можно констатировать, что использование расчётной шкалы даёт улучшение качества предсказания (рис. 6).

ЗАКЛЮЧЕНИЕ

Использовать, полученную в результате работы шкалу значений pKa, для предсказания pI можно с помощью разработанной авторами программы pIPredict, написанной на языке JAVA в виде исполняемого jar-архива. Программа свободна для использования академическими пользователями и доступна по адресу http://www.ibmc.msk.ru/LPCIT/pIPredict.

Программа имеет графический интерфейс пользователя и следующий набор возможностей:

1. Загрузка последовательностей белков и пептидов из файлов в формате FASTA или простом текстовом формате с указанием отдельных модификаций аминокислотных остатков (ацетилирование, фосфорилирование).
2. Предсказание pI с использованием нескольких шкал значений pKa, включая расчётную шкалу, описанную в данной работе, и шкалу имитирующую расчёт Marvin plugin.
3. Предсказание pI по нейросетевой модели для пептидов.
4. Преобразование по единому правилу выборки последовательностей (удаление N-концевого метионина и преобразование цистеина в карбоксиметилцистеин).
5. Расчёт зависимости усреднённого заряда белков (пептидов) от pH для выбранных белков и значения усреднённого заряда при заданном pH для полной выборки аминокислотных последовательностей.

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий. наук на 2013-2020 годы. Авторы благодарят С.Н. Нарыжного за консультации по вопросам анализа 2D-электрофореграмм.

ЛИТЕРАТУРА

1. *Giglione C., Boularot A., Meinnel T.* (2004) 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. at al.* (2005) J. proteome res., **4**(6), 2273-2282. DOI: 10.1021/pr050193v
3. *Po H.N., Senozan N.M.* (2001) J. Chemical Education, **78**, 1499-1503. DOI: 10.1021/ed078p1499
4. *Bjellqvist B., Hughes G.J., Pasquali Ch., Paquet N., Ravier F., Sanchez J.-Ch., Frutiger S., Hochstrasser D.F.* (1993) Electrophoresis, **14**, 1023-1031. DOI: 10.1002/elps.11501401163

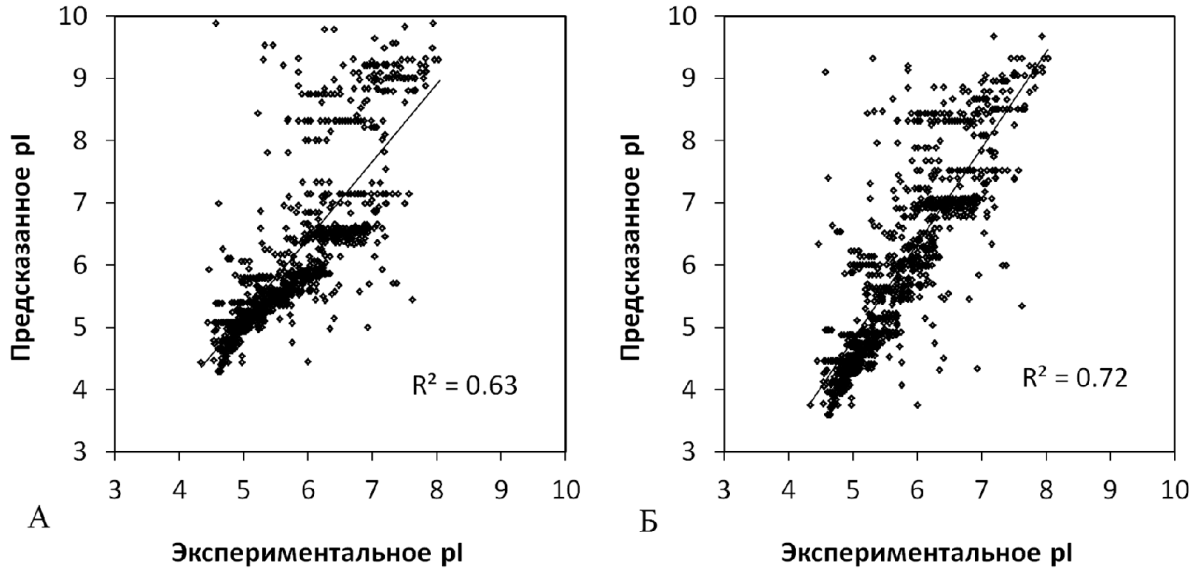

Рисунок 6. Сравнение предсказанных pI для выборки белков (1100 протоформ) методом Bjellqvist и соавт. (А) и по шкале, рассчитанной по объединённой выборке (Б).

5. *Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.* (2005) The Proteomics Protocols Handbook, pp 571-6071. DOI: 10.1385/1-59259-890-0:571
6. MELANIE version 7.0, GeneBio, Switzerland.
7. *Patrickios C.S.* (1995) J. Colloid and Interface Sci., **175**, 256-256. doi:10.1006/jcis.1995.1454.
8. *Perez-Riverol Y., Audain E., Millan A., Ramos Y., Sanchez A., Vizcaíno J. A., Wang R., Muller M., Machado Y.J., Betancourt L.H., Padrón G., Besada V.* (2012) J. Proteom., **75**, 2269-2274. DOI:10.1016/j.jprot.2012.01.029
9. *Gauci S., Van Breukelen B., Lemeer S.M., Krijgsveld J., Heck A.J.* (2008) Proteomics, **8**, 4898-4906. DOI: 10.1002/pmic.200800295
10. *Romero Reyes I.V., Fedyushkina I.V., Skvortsov V.S., Filimonov D.A.* (2013) Int. J. Math. Model. Meth. Appl. Sci., **7**, 303-310.
11. *Ромеро Рейес И.В.* (2014) Оценка аффинности комплексов белок-лиганд с применением нейронных сетей, Диссертация на соискание учёной степени к.ф-м.н., Москва, 107 с.
12. Chemaxon, Budapest, Hungary, http://www.chemaxon.com
13. *Pasquali M., Serchi T., Renaut J., Hoffmann L., Bohn T.* (2013) Electrophoresis, **34**, 505-509. DOI: 10.1002/elps.201200256
14. *Wong P., Walter M., Lee W., Mannhaupt G., Münsterkötter M., Mewes H.W., Adam G., Güldener U.* (2011) Nucleic Acids Research, **39**(suppl. 1): D637-D639. DOI: 10.1093/nar/gkq1016

Поступила: 05. 12. 2014.

pIPREDICT: A COMPUTER TOOL FOR PREDICTING ISOELECTRIC POINTS OF PEPTIDES AND PROTEINS

V.S. Skvortsov, N.N. Alekseychuk, D.V. Khudyakov, I.V. Romero Reyes

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

The data on approximate values of isoelectric point (pI) of peptides obtained during their fractionation by isoelectric focusing can be successfully used for the calculation of the pKa's scale for amino acid residues. This scale can be used for pI prediction. The data of peptide fractionation also provides information about various posttranslational modifications (PTM), so that the prediction of pI may be performed for a wide range of protein forms. In this study, pKa values were calculated using a set of 13448 peptides (including 300 peptides with PTMs significant for pI calculation). The pKa constants were calculated for N-terminal, internal and C-terminal amino acid residues separately. The comparative analysis has shown that our scale increases the accuracy of pI prediction for peptides and proteins and successfully competes with traditional scales and such methods as support vector machines and artificial neural networks. The prediction performed by this scale, can be made in our program pIPredict with GUI written in JAVA as executable jar-archive. The program is freely available for academic users at http://www.ibmc.msk.ru/LPCIT/pIPredict. The software has also the possibility of pI predicting by some other scales; it recognizes some PTM and has the ability to use a custom scale.

Key words: peptide, isoelectric point, posttranslational modifications, property prediction.