

ПРОТОКОЛЫ ЭКСПЕРИМЕНТОВ, ПОЛЕЗНЫЕ МОДЕЛИ, ПРОГРАММЫ И СЕРВИСЫ

ФИЛЬТРАЦИЯ ДАННЫХ 2D ЭЛЕКТРОФОРЕЗА ПРИ СОЗДАНИИ ВЫБОРКИ ДЛЯ ПРЕДСКАЗАНИЯ ЗНАЧЕНИЯ ИЗОЭЛЕКТРИЧЕСКОЙ ТОЧКИ БЕЛКОВ

В.С. Скворцов*, А.В. Рыбина

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

Рассмотрен ряд простых фильтров, сформулированных из общих соображений и учитывающих особенности проведения экспериментов, а также полученных в ходе выполнения двумерного (2D) электрофореза результатов. Данные фильтры могут быть использованы для автоматизированного формирования и верификации выборок для обучения систем предсказания значений изоэлектрической точки белков. Среди них: (i) фильтрация явных ошибок, внесённых при формировании исходной базы данных; (ii) отбор заведомо достоверного диапазона значений; (iii) отбор единственного варианта среди различных протеоформ; (iv) отбор в пределах заданной величины отклонения электрофоретического сдвига и другие. На примере выборки, объединяющей данные из 8 карт *Homo sapiens*, *Mus musculus* и *Rattus norvegicus*, применение данного набора фильтров позволило улучшить R^2 предсказания от 0.44 до 0.67.

Ключевые слова: изоэлектрическая точка; 2D электрофорез; подбор данных**DOI:** 10.18097/BMCRM00162

ВВЕДЕНИЕ

Ранее мы представили программу предсказания изоэлектрической точки (pI) пептидов и белков [1], основанную на использовании уравнения Хендерсона-Хассельбаха [2]. При этом табличные значения pK_a были рассчитаны на основании данных, полученных для большой выборки пептидов, в том числе и с различными химическими и посттрансляционными модификациями (PTM), в экспериментах по изоэлектрофокусированию пептидов с последующей масс-спектрометрической идентификацией. В свою очередь, эта выборка пептидов была получена в результате тщательной фильтрации полного набора данных многократно большего размера, проведённой с целью увеличения процента достоверных значений в выборке. В то же время тестирование качества предсказания pI для белков показало, что качественного большого набора данных для решения этой задачи нет. Отдельные выборки, полученные в экспериментах по двумерному (2D) геле-электрофорезу, имеют, на первый взгляд, хорошее качество и предсказываются хорошо. Однако в случае смешанных выборок, например из работы [3], около 60% наблюдений имеют ошибку предсказания не хуже 0.5 значений pH, в пределах ошибки до 0.1 значений pH было всего 15% [1]. Для решения некоторых задач такой точности могло бы хватить, но для идентификации конкретной протеоформы [4] этого оказывается недостаточно. Возникает вопрос, является ли невысокая точность предсказания проблемой самого метода предсказания? На выборках, полученных в рамках одного эксперимента, результаты были значительно лучше, как в работе [1], так и при использовании других методов, основанных на применении уравнения Хендерсона-Хассельбаха, в том числе и использованного далее в настоящей работе [5].

Почему возникают сомнения в качестве подобных сборных выборок? В первую очередь, этот результат отражает стремительный рост всевозможных омик-наук и направления «big data». Данные становятся вторичными и теряют часть информации, содержащейся в первоисточниках. Попадая в различные базы данных, эти вторичные данные становятся основой для формирования новых выборок и это может быть не одна итерация. В области предсказания значения pI для белков примером такой вторичной базы данных может быть Proteome-pI [6]. При этом, на каждом этапе исследователи решают свои собственные задачи, и часто для них часть информации несущественна, но в случае, когда идёт речь о «big data», достоверность данных имеет ключевое значение [7]. В настоящей работе показаны некоторые «подводные камни», которые могут свести на нет достоверность выборок при автоматической обработке исходных данных.

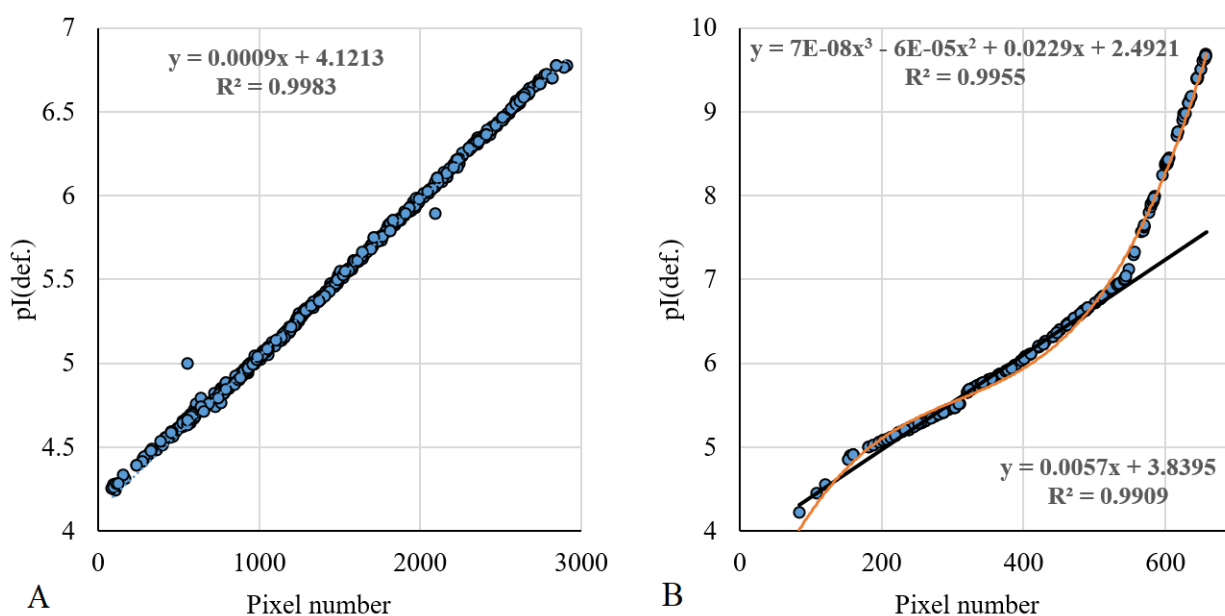
МЕТОДИКА

Для анализа данных были использованы 2 различных варианта предсказания значения pI для белков, базирующиеся на использовании уравнения Хендерсона-Хассельбаха. Первый из них, разработанный Bjellqvist и соавт. [5] – самый часто применяемый в настоящее время, так как именно он используется в программе предсказания pI с сервера exrasy.org. Второй, разработанный нами [1], кроме собственной шкалы pK_a , охватывающей не только основные аминокислотные остатки, но и остатки с модификациями, различает также местоположение аминокислотного остатка в полипептидной цепи. Для немодифицированных аминокислотных остатков в целом на больших выборках оба метода дают близкие результаты [1], но в каждом конкретном случае могут быть некоторые различия. В связи с этим для надёжности и внутреннего контроля лучше использовать оба.



Таблица 1. Карты 2D электрофореза, данные из которых использованы в работе

Вид	Код в World-2DPAGE Repository	Источник биологического материала	Ссылка
<i>Mus musculus</i>	BAT_MOUSE	Brown adipose tissue	[10]
<i>Homo sapiens</i>	CSF_HUMAN	Cerebrospinal FLUID	[11]
<i>Homo sapiens</i>	DLD1_HUMAN	Colorectal adenocarcinoma cell line (DL-1)	[12]
<i>Homo sapiens</i>	HEPG2_HUMAN	HepG2 cell line	[11]
<i>Mus musculus</i>	MUSCLE_MOUSE	Gastrocnemius muscle	[10]
<i>Homo sapiens</i>	PLASMA_HUMAN	Plasma	[13]
<i>Rattus norvegicus</i>	RAT_INS1E_4-7	INS-1E cell line, a model rat cell line for pancreatic beta cells	[14]
<i>Homo sapiens</i>	RBC_HUMAN	Red blood cells	[13]
<i>Staphylococcus aureus</i>	STAPHY4-7	<i>Staphylococcus aureus</i> (strain Mu50 / ATCC 700699)	[15]
<i>Staphylococcus aureus</i>	STAPHY6-11	<i>Staphylococcus aureus</i> (strain Mu50 / ATCC 700699)	[15]
<i>Marthasterias glacialis</i> (Spiny starfish)	RADIAL_NERVE_CORD_3-11	2-DE of the sea star radial nerve cord and the synaptosomal membranes subproteome	[16]

**Рисунок 1.** Сравнение приписанных значений pI и координат по оси абсцисс на 2D электрофоретической карте. А – данные RAT_INS1E_4-7; имеются 2 точки с ошибками в данных. В – данные RBC_HUMAN; приведено как уравнение кубического сплайна, описывающего весь набор данных, так и уравнение описывающее линейный участок. Если данные получены в одном эксперименте и без сбоев, то «ступеньки» около 310 пикселя быть не должно.

В качестве демонстрационных выборок данных для белков с «известными» значениями pI были использованы несколько наборов данных, полученных в экспериментах по 2D гель-электрофорезу (табл. 1), депонированных в World-2DPAGE Repository [8]. Данные были собраны собственной программой, которая также автоматически подгружала последовательности идентифицированных белков из БД Uniprot [9]. Если автоматизированная процедура не смогла найти аминокислотную последовательность в БД, данное наблюдение отбрасывали. Как правило, это связано с тем, что в различных редакциях БД идентификаторы были изменены, объединены или удалены (нельзя исключить также и факт ошибок и опечаток при формировании страницы в World-2DPAGE Repository.) Используя архивные данные, можно отыскать вручную все измененные последовательности, но так как задачей нашей работы было показать именно проблемы автоматизированных процедур, этого не

делали. Для анализа данных использовали координаты (номера пикселей по горизонтали), поставленные в соответствие точке идентификации, приписанное значение pI и предсказанные значения pI.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

При анализе результатов из любых глобальных БД всегда следует помнить, что они вторичны и сбор информации или её ввод авторами осуществляется либо вручную, либо с использованием внешних программ. И в том, и другом случае возможны ошибки ввода. Автоматизированные процедуры не предполагают работы оператора со статьями, где были опубликованы первичные данные, а в случае работы с большими данными («big data») это физически невозможно. Если информация, содержащаяся в изначальной статье, не соответствует структуре изначальной публикации, то она

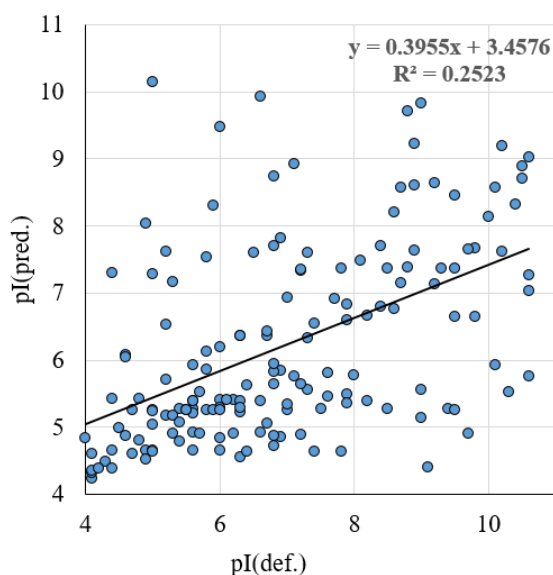


Рисунок 2. Сравнение приписанных значений pI и величин, предсказанных методом Bjellqvist и соавт. для набора данных из 2D электрофоретической карты RADIAL_NERVE_CORD_3-11. Все идентифицированные белки – гомологи белков *Marthasterias glacialis* из других организмов.

теряется или, что ещё хуже, искажается. Тем не менее, возможность проанализировать данные на ошибки или выявить отдельные особенности, которые следует учесть при отборе информации, все-таки существует. Рассмотрим два набора данных RAT_INS1E_4-7 и RBC_HUMAN (рис. 1). Если проанализировать значение pI и координаты по оси абсцисс на опубликованной карте (проще говоря, номер пикселя по горизонтали), то из общих соображений ясно, что зависимость должна быть монотонной. В зависимости же от типа «стрипа» (полоска с иммобилизованным градиентом pH), использованного в эксперименте, эта зависимость может быть либо линейной, либо соответствующей заявленной производителем (а их не так много). Данные RAT_INS1E_47 (рис. 1A) явным образом указывают на линейную зависимость, однако имеется несколько точек (2 из них ярко выражены), отклонение которых нельзя описать ошибками округления. С 100% вероятностью эти точки можно считать ошибками. Второй случай (рис. 2 B), на первый взгляд, соответствует нелинейной зависимости pI от координат, которая может быть описана кубическим сплайном и не имеет видимых ошибок. Характер кривой напоминает ожидаемую зависимость, такую как, например, у Nonlinear pH 3-10 ReadyStrip IPG strip («Bio-Rad Laboratories, Inc.», США). В тоже время при автоматизированной обработке данных сложно учесть все варианты, если этих данных нет в БД. Тем более, что предсказанные величины pI (например, методом Bjellqvist, см. дополнительные материалы) часто не совпадают с наблюдаемыми, сохраняя линейность и в этой части. Хотя шкала, полученная Bjellqvist и соавт. ещё в 1983 году, основана на данных для диапазона pI от 4 до 7, другие методы также дают близкие результаты. При первичной фильтрации данных пользоваться результатами каких-либо предсказаний неправильно. Существующий выход – вычленивать данные для линейного участка (рис. 1 B) и ограничиться ими. Для 2D электрофоретических карт редко, но встречается и ещё один вариант, когда авторы просто приписывают конкретным белкам «теоретические» значения pI. Поскольку при этом для расчёта используется

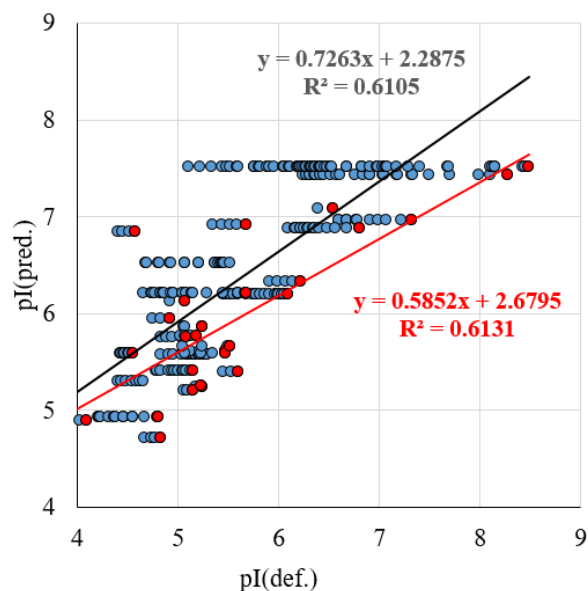


Рисунок 3. Сравнение приписанных значений pI и величин, предсказанных методом Bjellqvist и соавт. для набора данных из 2D электрофоретической карты CSF_HUMAN (отброшена одна точка без модификаций с $\Delta pI > 3$). Синий цвет - все варианты, красный только один самый основной вариант для каждого из белков.

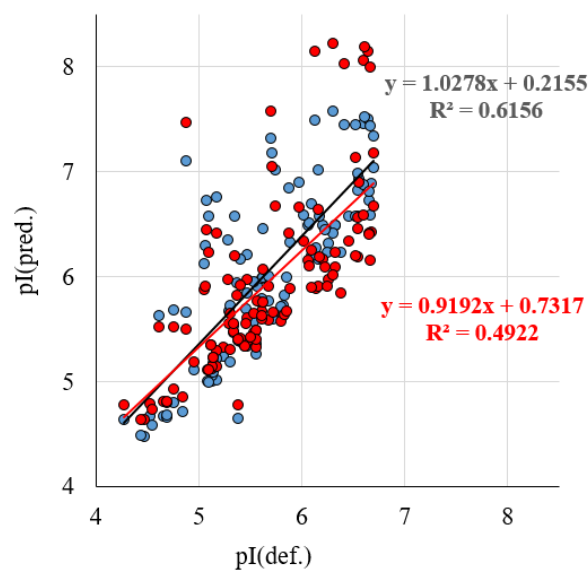


Рисунок 4. Сравнение приписанных значений pI и величин, предсказанных методами Bjellqvist и соавт. (красный) и pIPredict 3 (синий) для набора данных из 2D электрофоретической карты RAT_INS1E_4-7.

программа с сервера expasy.org, то простое сравнение данных с результатами предсказания Bjellqvist и соавт. легко выявляет такой случай, обычно это видно и при анализе зависимости pI от координат на карте.

Отдельно необходимо рассмотреть случай, когда исследуются белки организмов, для которых неизвестен, либо частично известен геном, например, белки морской звезды *Marthasterias glacialis* [16]. В такой ситуации при масс-спектрометрической идентификации используют последовательности близких или нет видов. Ожидать, что в этом случае совпадут значения pI, полученные из 2D электрофоретической карты, и предсказанные величины не приходится (рис. 2). В рассматриваемом примере в списке нет ни одного белка *Marthasterias glacialis*. Если добавить

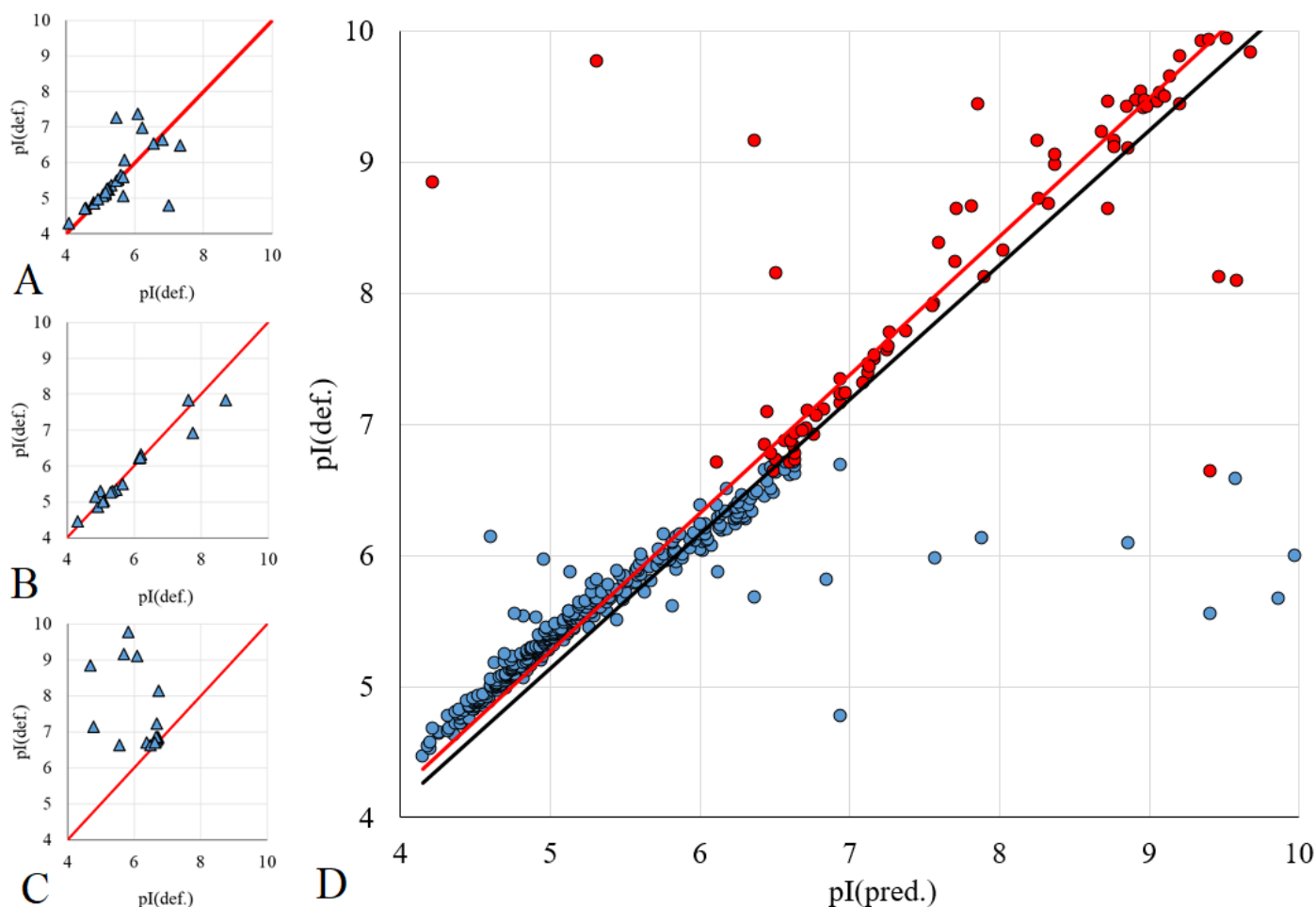


Рисунок 5. А, В, С – попарное сравнение приписанных величин pI из разных экспериментов (*Homo sapiens*, *Mus musculus* и *Staphylococcus aureus*). D – сравнение приписанных и предсказанных значений pI для белков *Staphylococcus aureus*, полученных из разных 2D карт (синий - STAPHY4-7, красный – STAPHY6-11).

в процедуру сбора данных простое сравнение по видовой принадлежности, то это решит проблему. К сожалению, в случае внутривидовых различий, особенно выраженных у одноклеточных организмов, это не поможет. В данном случае можно только надеяться, что один и тот же функционально белок имеет и схожие физико-химические свойства.

Ещё одним фактором, вносящим существенные искажения в значения pI , является наличие PTM. В 2D электрофорезе это выглядит как идентификация одного белка для нескольких «пятен». Широко распространённой практикой среди исследователей, формирующих выборки для создания методов предсказания значений pI , является усреднение значений [3]. В ряде случаев, если варианты мало различаются по величине pI , это не вносит больших искажений. Но пример, представленный на рисунке 3, демонстрирует обратное. Статистически усреднённые данные схожи с данными, полученными при отборе конкретного значения, но полученные зависимости существенно отличаются. Возникает вопрос, каким образом выбрать конкретное значение? Универсального решения нет. Из общих соображений это может быть самое большое по площади пятно на 2D карте, но данной информации в БД может и не быть. Более приемлемый вариант – выбирать самое основное значение pI , так как большинство имеющихся физиологических PTM смещают значение pI в кислую сторону.

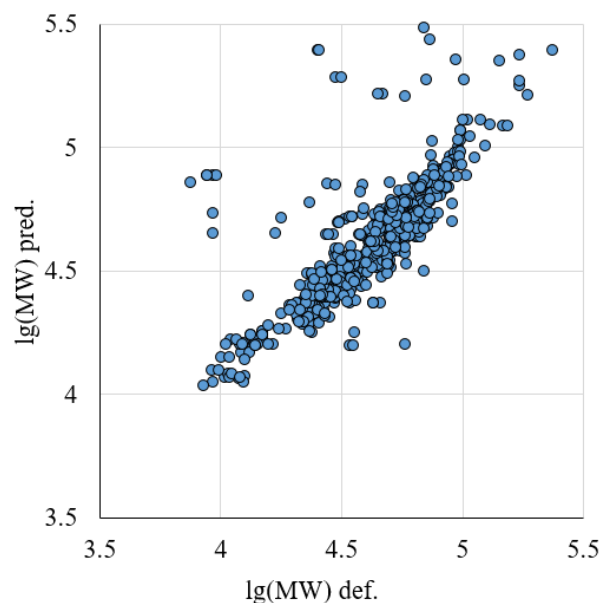


Рисунок 6. Сравнение расчётной и наблюдаемой величин $\lg(MW)$ для выборки, объединяющей данные из карт *Homo sapiens*, *Mus musculus* и *Rattus norvegicus*.

Одна PTM всё же может быть учтена. Речь идёт об удалении N-концевого остатка метионина. В зависимости от объекта исследований решение о сохранении его при расчётах или удалении может быть принято заранее. И если

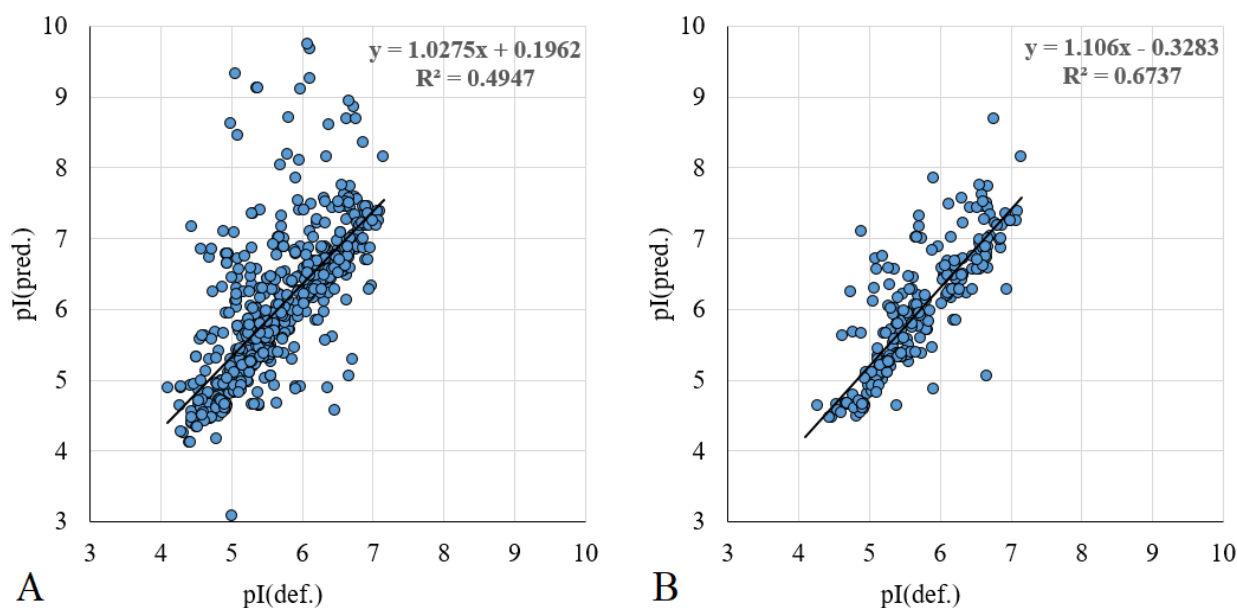


Рисунок 7. Сравнение приписанных и предсказанных значений pI для выборки белков, объединяющей данные из карт *Homo sapiens*, *Mus musculus* и *Rattus norvegicus*. А – фильтрация только на явные ошибки, линейный участок зависимости от координат пикселя, самое основное значение при наличии вариантов. В – добавлена фильтрация по значению $\Delta \lg(MW) \leq 0.05$.

для такого метода предсказания, как Bjellqvist и соавт., эта процедура не даст никаких изменений, то предложенный нами метод предсказания pIPredict [1] и ряд других учтут это изменение. Все предсказанные величины до этого момента были получены методом Bjellqvist и соавт. – не самым лучшим, но самым распространённым и известным среди исследователей в области 2D электрофореза. Однако предложенный нами метод предсказания pIPredict дает несколько лучший результат (рис. 4), поэтому в дальнейшем все предсказания были выполнены этим методом.

Хорошим вариантом при формировании выборки значений pI было бы подтверждение значения как минимум в двух экспериментах. Однако выборка при этом может быть очень небольшой. В нашем случае (рис. 5 А,В) для всех 5 карт *Homo sapiens* как минимум 2 значения было всего у 31 белка, в двух картах *Mus musculus* имеется 17 совпадений. И даже в этом случае не все приписанные значения pI совпадают (при наличии вариантов использовали самое основное значение). На рисунке 5 С приведены данные для пересечения двух карт *Staphylococcus aureus*, выполненных одними и теми же авторами в двух различных диапазонах pH, имеющих пересечение в области значений от 6 до 7. Для 8 из 19 пересекающихся белков приписанные значения отличаются кардинально. Кроме того, если сравнить приписанные значения с предсказанными (рис. 5 D), то хорошо видно, что для правой части имеет место небольшое смещение вверх. Если ориентироваться на приписанные величины pI для пересекающихся белков, то это смещение примерно в 0.1-0.15 значений pH. Величина небольшая, но вполне значимая.

Ну и наконец, до сих пор никак не обсуждался второй параметр из 2D карт – электрофоретический сдвиг, пропорциональный молекулярному весу белков (MW). Этот параметр также зависит от множества факторов: тех же PTM, гликозилирования (на величину pI практически не влияет), наличия в пробах редуцированных форм или фрагментов белков, наличия в составе белков аминокислотных паттернов, характерных для устойчивых элементов вторичной структуры

и не денатурирующих полностью [17], различиями в способности конкретных белков связывать молекулы SDS (т.е. в конечном итоге аминокислотный состав и те же паттерны) и др. К сожалению, в большинстве случаев в настоящее время нет возможности без знания всех особенностей исследуемого белка определить, почему в электрофоретическом сдвиге наблюдаются отклонения от ожидаемого. Для каких-то конкретных модельных белков с заранее точно известными аминокислотной последовательностью и имеющимися модификациями можно найти данные по электрофоретическому сдвигу, но в подавляющей части случаев при анализе природных проб точных данных нет. Тем не менее, можно предположить, что если величина отклонения от ожидаемого велика, то мы имеем дело с какой-то формой, отличной от простой полной аминокислотной последовательности из БД Uniprot. Как видно из рисунка 6, в основном различия между расчётной величиной MW и наблюдаемой в большинстве случаев находятся в пределах 0.2 единиц $\lg(MW)$. Большее отклонение можно заведомо считать ошибкой. Предел отклонения по MW может быть и более жёстким. При этом число наблюдений в выборке будет уменьшаться, но не катастрофично, а качество выборки будет расти (рис. 7). Если при первоначальной фильтрации (на явные ошибки, выбору линейного участка зависимости pH от координат по оси абсцисс, выбору только одной протеоформы для каждого белка) размер выборки составлял 614 наблюдений, и R^2 при сравнении приписанных и предсказанных значений pI был равен 0.49 (без фильтрации протеоформ – 0.44), то после добавления фильтрации по значению $\Delta \lg(MW) \leq 0.05$ число наблюдений было уменьшено до 235, а R^2 предсказания составил 0.67 (без фильтрации протеоформ – 0.48).

ЗАКЛЮЧЕНИЕ

Таким образом, при применении очень простых фильтров, установленных из общих соображений, учитывающих особенности проведения экспериментов и

природу полученных результатов, качество автоматически собранных выборок для обучения систем предсказания значений pI белков может быть существенно улучшено. Число таких фильтров не исчерпывается теми, что описаны в данной статье.

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.

К данной статье приложены дополнительные материалы, свободно доступные (<http://dx.doi.org/10.18097/BMCRM00162>) на сайте журнала.

ЛИТЕРАТУРА

1. 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
2. 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
3. Kozłowski, 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
4. Naryzhny, S.N., Legina, O.K. (2019) Structural-functional diversity of p53 proteoforms. *Biomeditsinskaya khimiya*, **65**(4), 263-276. DOI: 10.18097/PBMC20196504263
5. 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
6. Kozłowski, L. P. (2022) Proteome-pI 2.0: proteome isoelectric point database update. *Nucleic acids research*, **50**(D1), D1535-D1540. DOI: 10.1093/nar/gkab944
7. Kitchin, R. (2014) Big Data, new epistemologies and paradigm shifts. *Big data & society*, **1**(1), 2053951714528481. DOI: 10.1177/2053951714528481
8. Hoogland, C., Mostaguir, K., Appel, R.D., Lisacek, F. (2008) The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. *Journal of proteomics*, **71**(2), 245–248. DOI: 10.1016/j.jprot.2008.02.005
9. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, **49**(D1), D480–D489. DOI: 10.1093/nar/gkaa1100
10. Sanchez, J. C., Chiappe, D., Converset, V., Hoogland, C., Binz, P.A., Paesano, S., Appel, R.D., Wang, S., Sennitt, M., Nolan, A., Cawthorne, M.A., Hochstrasser, D.F. (2001) The mouse SWISS-2D PAGE database: a tool for proteomics study of diabetes and obesity. *Proteomics*, **1**(1), 136–163. DOI: 10.1002/1615-9861(200101)1:1<136::AID-PROT136>3.0.CO;2-1
11. Sanchez, J.C., Appel, R.D., Golaz, O., Pasquali, C., Ravier, F., Bairoch, A., Hochstrasser, D.F. (1995) Inside SWISS-2DPAGE database. *Electrophoresis*, **16**(7), 1131–1151. DOI: 10.1002/elps.11501601190
12. Demalte-Annessi, L., Sanchez, J.-C., Hoogland, C., Rouge, V., Binz, P.-A., Appel, R.D., Hochstrasser D.F. (1999) Submitted JAN-1999 to SWISS-2DPAGE. Retrieved from: https://world-2dpagexpasy.org/swiss-2dpagemap=dld1_human
13. Golaz, O., Hughes, G.J., Frutiger, S., Paquet, N., Bairoch, A., Pasquali, C., Sanchez, J. C., Tissot, J. D., Appel, R.D., Walzer, C. (1993) Plasma and red blood cell protein maps: update 1993. *Electrophoresis*, **14**(11), 1223–1231. DOI: 10.1002/elps.11501401183
14. D'Hertog, W., Maris, M., Thorrez, L., Waelkens, E., Overbergh, L., Mathieu, C. (2011) Two-dimensional gel proteome reference map of INS-1E cells. *Proteomics*, **11**(7), 1365–1369. DOI: 10.1002/pmic.201000006
15. 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
16. Franco, C.F., Santos, R., Coelho, A.V. (2011) Exploring the proteome of an echinoderm nervous system: 2-DE of the sea star radial nerve cord and the synaptosomal membranes subproteome. *Proteomics*, **11**(7), 1359–1364. DOI: 10.1002/pmic.201000541
17. Rath, A., Glibowicka, M., Nadeau, V. G., Chen, G., Deber, C. M. (2009) Detergent binding explains anomalous SDS-PAGE migration of membrane proteins. *Proceedings of the National Academy of Sciences*, **106**(6), 1760-1765. DOI: 10.1073/pnas.0813167106

Поступила: 24.02.2022
После доработки: 11.03.2022
Принята к публикации: 14.03.2022

THE FILTRATION OF 2D ELECTROPHORESIS DATA DURING CREATION OF A LEARNING SET FOR PREDICTION OF THE VALUE OF THE ISOELECTRIC POINT OF PROTEINS

V.S. Skvortsov, A.V. Rybina*

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

A number of simple filters formulated from general considerations that take into account the peculiarities of the experiments as well as results obtained in 2D electrophoresis experiments are considered. These filters can be used for automated dataset formation and verification of learning of system for predicting protein isoelectric point values. These include: (i) filtering obvious errors introduced during initial database formation; (ii) selection of a known plausible range of values; (iii) selection of a single variant among various proteoforms; (iv) selection within a preset value of electrophoretic shift deviation, etc. Using a dataset combining data from 8 maps of *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*, the application of this set of filters improved the R^2 value of predictions from 0.44 to 0.67.

Key words: isoelectric point; 2D electrophoresis; data collection

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: 24.02.2022, revised: 11.03.2022, accepted: 14.03.2022