Биомедицинская химия, 2017 том 63, вып. 3, с. 278-283.

КОРРЕКЦИЯ ЭЛЕКТРОФОРЕТИЧЕСКОГО СДВИГА В ВИРТУАЛЬНОМ 2D SDS-PAGE ЭЛЕКТРОФОРЕЗЕ

В.С. Скворцов*, Н.Н. Алексейчук, А.В. Рыбина

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

Виртуальный электрофорез в протеомике может быть использован для поиска локализации конкретных белков и протеоформ, особенно если они представлены в низкой концентрации, для формирования гипотез, какие протеоформы обнаруживаются в эксперименте и других задач. В то время как задача предсказания изоэлектрической точки хорошо исследована, вопрос коррекции электрофоретического сдвига относительно молекулярного веса, рассчитанного по аминокислотной последовательности белка, как правило, игнорируется. Для формирования уравнения, корректирующего данную величину, использовались 4 выборки данных, взятых из литературных источников и базы данных SWISS-2DPAGE (123, 72, 118 и 470 наблюдения соответственно). Было построено 2 группы моделей. Первая базировалась на аминокислотном спектре белков, вторая – на анализе параметров, рассчитанных по аминокислотной последовательности (теоретический молекулярный вес, гидрофобность, распределение зарядов, способность образовывать спиральные структуры). В пределах отдельных выборок коэффициент детерминации колебался от 0,35 до 0,75, причём кросс-предсказание между выборками не давало хорошего результата. Однако направление, в какую сторону следует провести коррекцию, в 74% случаев предсказывалось правильно. При объединении выборок и делении коэффициент детерминации при настройке колебался от 0,44 до 0,51, в то же время, R^2 предсказания был не хуже 0,39, а направление коррекции предсказывалось правильно в 80% случаев. Созданные модели предсказания интегрированы в программу pIPredict v.2, доступную по адресу http://www.ibmc.msk.ru/LPCIT/pIPredict.

Ключевые слова: электрофоретический сдвиг, виртуальный электрофорез, статистический анализ

DOI 10.18097/PBMC20176303278

ВВЕДЕНИЕ

2D SDS-PAGE электрофорез – метод, широко применяемый в протеомике. Одним из направлений в вычислительной биологии является так называемый “виртуальный электрофорез”, данные которого используются для сравнительного анализа экспериментальных данных [1], предсказания положения целевых белков на геле и поиска локализации белков, представленных в низких концентрациях. Виртуальный электрофорез базируется на широком спектре методов предсказания величины pI, причём, средняя точность предсказания для заданного белка обычно не хуже ±0,15 pH единиц. Наиболее широко используют методы на базе уравнения Хендерсона-Хассельбаха. Самый популярный – калькулятор на Bioinformatics Resource Portal (http://www.expasy.org), созданный на основе шкалы, разработанной Bjellkvist и соавт. [2] для области pH от 4 до 7. Нами также была ранее разработана собственная шкала, реализованная в программе pIPredict [3], для диапазона pH от 3 до 10. В то же время по поводу предсказания молекулярного веса традиционный подход – это использование вычислений по формуле. Известно, что кажущийся молекулярный вес (M_r), рассчитанный из наблюдаемого электрофоретического сдвига, только коррелирует с рассчитанным по брутто-формуле (M_f) и зависит от большого числа факторов (часть из которых прямо влияет на количество связанного SDS): молекулярного веса (точнее даже длины последовательности), наличия элементов вторичной структуры (известно, например, что часть спиралевидных структур при денатурации в SDS сохраняется), гидрофобности последовательности, общего заряда последовательности и условий проведения электрофореза (pH, концентрации ПААГ [4]) и других факторов [5]. Проблема достаточно известная, однако, в области предсказаний электрофоретического сдвига при 2D SDS-PAGE электрофорезе сделано мало, и, если проанализировать литературу, в основном одной группой авторов [5, 6]. В то же время, предсказание электрофоретического сдвига для других видов электрофореза (и особенно для пептидов) даёт очень хорошие результаты. Например, в одной из последних работ Krokhin и соавт. [7] для CZE электрофореза пептидов создали модель с высокой предсказательной силой, в которой учитывались молекулярный вес, число заряженных групп пептида при заданном pH, гидрофобность и склонность к образованию спиральных структур. Из общих соображений несложно понять, что большинство исследователей в области экспериментального электрофореза не сильно интересует отличие электрофоретического сдвига от расчётного, поскольку, как правило, исследуется сравнение двух или более электрофоретических карт, полученных при различных исходных условиях. Однако, в протеомных исследованиях, связанных с использованием виртуальных электрофоретических карт, ошибка при определении виртуального электрофоретического сдвига может быть критической. В данной работе мы предприняли попытку разработать

* - адресат для переписки

соответствующие инструменты предсказания электрофоретического сдвига, позволяющие минимизировать возможную ошибку.

МЕТОДИКА

В работе были использованы 2 набора данных. Первый, выборка, взятая из статьи [6] (обозначим её S0). Данная выборка представляет собой коллекцию в основном трансмембранных белков, подобранных авторами из литературных источников, на основании которой авторы рассчитывали коррекционные факторы для электрофоретического сдвига (ЭС). Так как целью работы [6] было обоснование введения отдельного стандарта для гидрофобных белков, то собственно предсказаний величины ЭС авторы не делали, но показали связь этой величины с наличием предсказанных фрагментов последовательности – "трансмембранных миметиков". Расчёт вероятности появления и количества таких "трансмембранных миметиков" по алгоритму программы TM Finder [8] (использованному в работе [6]) мы применили при создании моделей предсказания ЭС. Использовалась собственная программная реализация данного алгоритма. При анализе данной выборки возникли разночтения между вычисленными нами значениями M_f и представленными в статье. Вероятно, это связано с тем, что авторы в ряде случаев анализировали кристаллографические структуры, содержащие урезанные последовательности. Оставляя за границами данной статьи дискуссию, что более адекватно, мы решили не использовать данные, отличающиеся от расчётных по последовательности взятой из UNIPROT (http://www.uniprot.org) более чем на 2 кДа. Так как предполагаемая область применения данных моделей – анализ белков, для которых в общем случае ничего неизвестно, кроме последовательности из базы данных, то именно величину M_f, рассчитанную на основании полной последовательности, мы в дальнейшем и будем использовать. Другим ограничением стал факт, что во втором наборе данных практически отсутствовали данные, где расхождения между M_r и M_f были больше чем 50%. Так что во всех случаях данные, где это различие было больше 50% при определении обучающей выборки, эти наблюдения были отброшены. Таким образом, выборка S0 сокращена со 174 до 123 наблюдений.

Второй набор данных был получен из серии 2D электрофоретических карт, депонированных в базе данных (БД) SWISS-2DPAGE (http://world-2dpage.expasy.org/swiss-2dpage). В качестве условий отбора использовались количество идентифицированных белков (чем больше – тем лучше), а также качество аннотации (рис. 1). Нередки случаи, когда при аннотации гелей авторы используют не величины, вычисленные из электрофоретической карты, а расчётные данные, кроме того в данных нередко встречаются выбросы, которые скорее всего представляют собой опечатки или ошибки ввода. Теоретически даже при неправильной аннотации из координат положения "точек" на фотографии геля можно рассчитать M_r, но в данном случае в БД достаточно хорошо аннотированных карт, чтобы не вносить собственные ошибки. Всего было использовано 3 карты (номер карты соответствует номеру в БД, число наблюдений дано после фильтрации по критерию "не больше 50%-го отклонения" и отбрасывания "точек", для которых ID белка и M_r совпадали): S176 [9] – 72 наблюдения (75 "точек", 65 белков); S182 [10] – 118 наблюдений (120 "точек", 105 белков); S188 [11] – 470 наблюдений (538 "точек", 295 белков). Несложно заметить, что некоторые белки представлены несколькими протеоформами, но из-за отсутствия каких-либо достоверных указаний, что это за протеоформы, отфильтровать их достоверно нельзя. Это вносит существенный шум. Впрочем, именно для анализа этих отклонений и нужны предсказания M_r.

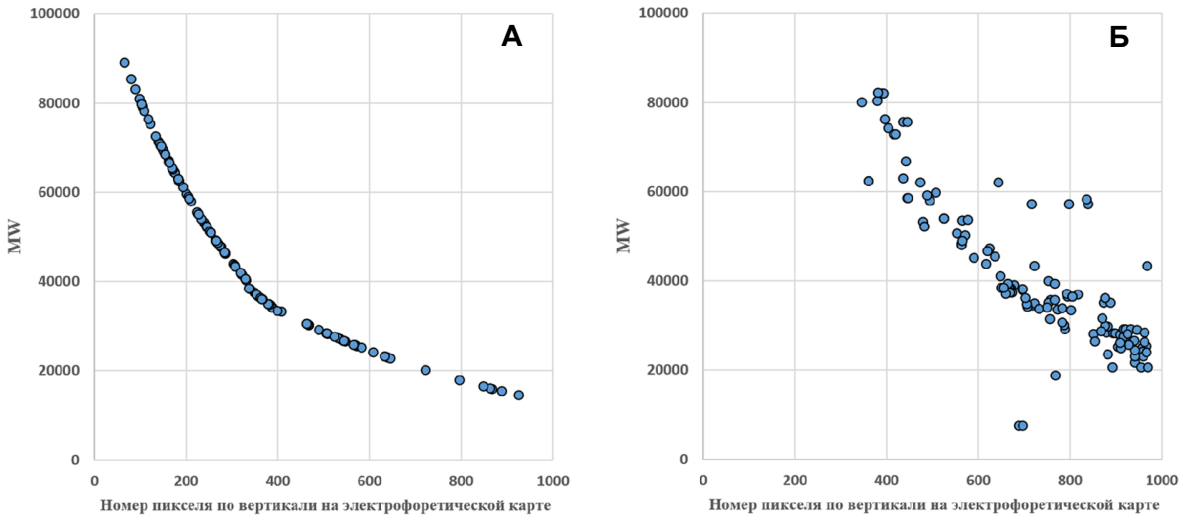

Рисунок 1. Пример разного подхода к аннотации геля авторами в БД SWISS-2DPAGE. Зависимость молекулярного веса от координаты по вертикали на карте геля. А. MW в аннотации соответствует вычисленному из электрофоретической карты. Б. Используется величина MW, вычисленная для аминокислотной последовательности. В данном случае калибровочная функция неизвестна. Её можно рассчитать, но это возможный источник ошибок.

Для предсказания M_r строились 2 группы моделей. Первая предсказывала ΔM, равную M_f - M_r, с использованием простой линейной регрессии, где в качестве независимых переменных выступал аминокислотный спектр белка (20 значений). Во второй – использовалась величина "доля отклонения" (D_r), равная (M_f - M_r)/M_f, она же использовалась для фильтрации данных (50% отклонение это $|D_r|>0,5$. В качестве независимых переменных использовались 9 параметров, рассчитываемых по последовательности, взятой из UNIPROTE:

1. Величина M_f. Тот факт, что, скорее всего, все цистеины были модифицированы, не учитывался.

2. Средний заряд белка при pH 8,3, рассчитанный согласно уравнению Хендерсона-Хассельбаха по шкале Bjellkvist, выполненный программой pIPredict.

3. Условная "гидрофобность", рассчитанная по алгоритму метода предсказания времени удержания пептида (SSRCalc. [12]), выполненная программой ProteoCat [13].

4. Количество положительно заряженных аминокислотных остатков.

5. Количество отрицательно заряженных аминокислотных остатков.

6. Общая гидрофобность последовательности, рассчитанная аддитивным методом по шкале из работы [8].

7. Общая величина "спиральности" белка, рассчитанная аддитивным методом по соответствующей шкале из работы [8].

8. Число миметиков "трансмембранных фрагментов", рассчитанная также по аналогии с работой [8].

9. Локальная величина "спиральности", рассчитанная только для аминокислотных остатков, входящих в участки последовательности, выявленные как миметики "трансмембранных фрагментов".

В данном случае, каждая величина использовалась в регрессионном уравнении не напрямую, а модифицированная по сигмовидной зависимости, параметры которой также подбиралась в процессе обучения (рис. 2). Вычисления проводились программой NNC [14]. Несмотря на то, что внешне схема на рисунке 2 напоминает нейронную сеть, по сути она таковой не является, так как узлы друг на друга не влияют. Сравнение полученных результатов проводилось как между предсказываемыми величинами, так и после пересчёта в предсказанные величины ΔM и M_r.

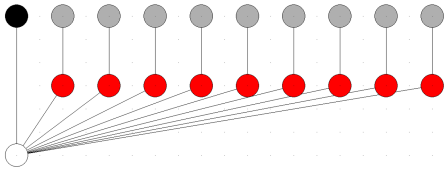

Рисунок 2. Схематическое изображение предсказательной модели группы 2. Первый ряд кроме чёрного узла (константа) - входные узлы. Средний ряд узлы, преобразующие сигнал по сигмовидной функции независимо друг от друга. Белый узел - выходной.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Для начала рассмотрим результаты анализа, полученные для выборки S0, представленные на рисунке 3 и в таблице (первая строка). Видно, что в обеих моделях можно достичь достаточно хороших статистических показателей, в том числе

Таблица. Параметры моделей, предсказывающих величину электрофоретического сдвига

№	Выборка	Тип модели	Размер выборки	Диапазон значений MW (эксп.)	Медиана MW	Целевой параметр	Диапазон значений целевого параметра	Медиана значений целевого параметра	R^2	SE	R^2 (пересчёт в ΔM)	Q^2
1	S0	1	123	3,6 - 120,0	24,0	ΔM	-23,8 - 14,6	-1,57	0,75	2,57		0,66
2	S0	2	123	3,6 - 120,0	24,0	D_r	-0,41 - 0,48	-0,08	0,67	0,05	0,95	0,54
3	S176	1	72	10,34 - 89,73	45,42	ΔM	-41,38 - 19,17	2,96	0,63	4,36		0,40
4	S176	2	72	10,34 - 89,73	45,42	D_r	-0,37 - 0,43	0,06	0,74	0,06	0,79	0,52
5	S182	1	118	14,42 - 88,92	43,48	ΔM	-17,85 - 19,07	-1,23	0,42	2,28		0,30
6	S182	2	118	14,42 - 88,92	43,48	D_r	-0,25 - 0,4	-0,04	0,49	0,05	0,62	0,38
7	S188	1	470	14,91 - 109,93	57,77	ΔM	-39,38 - 26,53	6,9	0,36	5,6		0,28
8	S188	2	470	14,91 - 109,93	57,77	D_r	-0,44 - 0,49	0,16	0,5	0,08	0,53	0,48
9	O1	1	392	4,0 - 109,93	49,08	ΔM	-41,38 - 26,53	3,06	0,45	4,93		0,39
10	O1	2	392	4,0 - 109,93	49,08	D_r	-0,44 - 0,49	0,08	0,47	0,1	0,48	0,43
11	O2	1	391	3,6 - 120,0	49,09	ΔM	-41,38 - 26,52	3,06	0,49	4,88		0,43
12	O2	2	391	3,6 - 120,0	49,09	D_r	-0,43 - 0,49	0,08	0,51	0,1		0,45
13	O	1	783	3,6 - 120,0	49,09	ΔM	-41,38 - 26,53	3,06	0,44	4,9	0,51	0,42
14	O	2	783	3,6 - 120,0	49,09	D_r	-0,44 - 0,49	0,08	0,5	0,8		0,47

Примечание. Все величины MW и ΔM приведены в кДа. R^2 - коэффициент детерминации при обучении. SE - средняя ошибка при обучении. Для величины D_r проводился пересчёт в ΔM. Q^2 - коэффициент детерминации в процедуре скользящего контроля методом "выкидывания по одному".

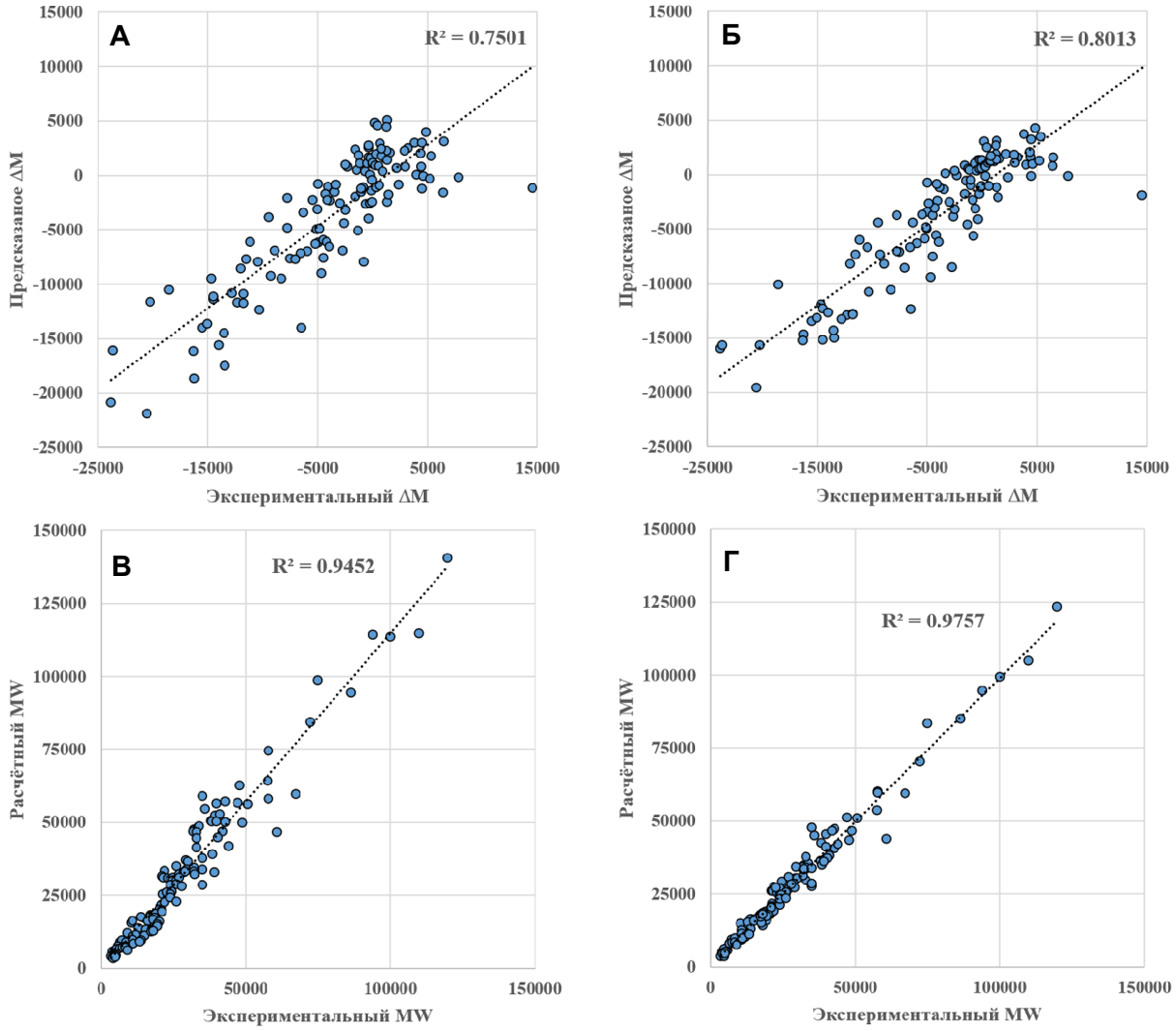

Рисунок 3. Предсказание электрофоретического сдвига для выборки S0. А. Сравнение экспериментально определённого ΔM и предсказанного по модели 1. Б. Сравнение экспериментально определённого ΔM и предсказанного (пересчитан из D_r, предсказанного по модели 2). В. Сравнение экспериментально определённого (M_r) и рассчитанного из брутто формулы (M_f) молекулярного веса. Г. Сравнение экспериментально определённого (M_r) и расчётного молекулярного веса (с учётом предсказанной величины D_r).

и при проверке методом выбрасывания по одному. Предсказание величины D_r внешне имеет худшие параметры, однако из-за относительности этой величины при пересчёте в M_r она демонстрирует лучший вариант предсказания. Важно отметить, что направление изменений (то есть совпадение знака предсказанных величин и расчётных) составляет 80% для модели M, до 91% для D_r, причём около 60% отличий приходится на диапазон изменений |M|<1 кДа, что вполне укладывается в среднюю точность оценки M_r по электрофоретической карте.

К сожалению, при анализе данных, полученных из SWISS-2DPAGE, результаты (таблица) не столь оптимистичные. Из трёх выборок только в одном случае (S176) уравнения имеют хорошие статистические характеристики. Тем не менее, используя данные уравнения, можно уверенно предсказать направление изменений M_r относительно M_f со средней точностью в 74%, причём 45% “ошибок” также попадают в диапазон |M|<1 кДа. Кросс-предсказания также не дают удовлетворительного результата: R^2 предсказания колеблется от 0,12, до 0,37 в самом лучшем случае (предсказание выборки S0 по модели для S176). Последнее связано, скорее всего, с двумя обстоятельствами: различия в методиках эксперимента, которые не учитываются при создании регрессионных уравнений, и в свойствах самих выборок, например, выборка S0 и выборки из SWISS-2DPAGE достоверно различаются по размеру белков, в S0 они в основном небольшие.

К сожалению, ввести в уравнение дополнительные параметры, связанные с особенностями проведения эксперимента не так просто (это потребует набора данных по нескольким десяткам электрофоретических карт). Что, разумеется, мы планируем сделать в будущем. А вот самый очевидный вариант – создание смешанной выборки (O), был проанализирован. В общей сумме 4 выборки в сумме дают 783 наблюдения, что, даже при условии деления

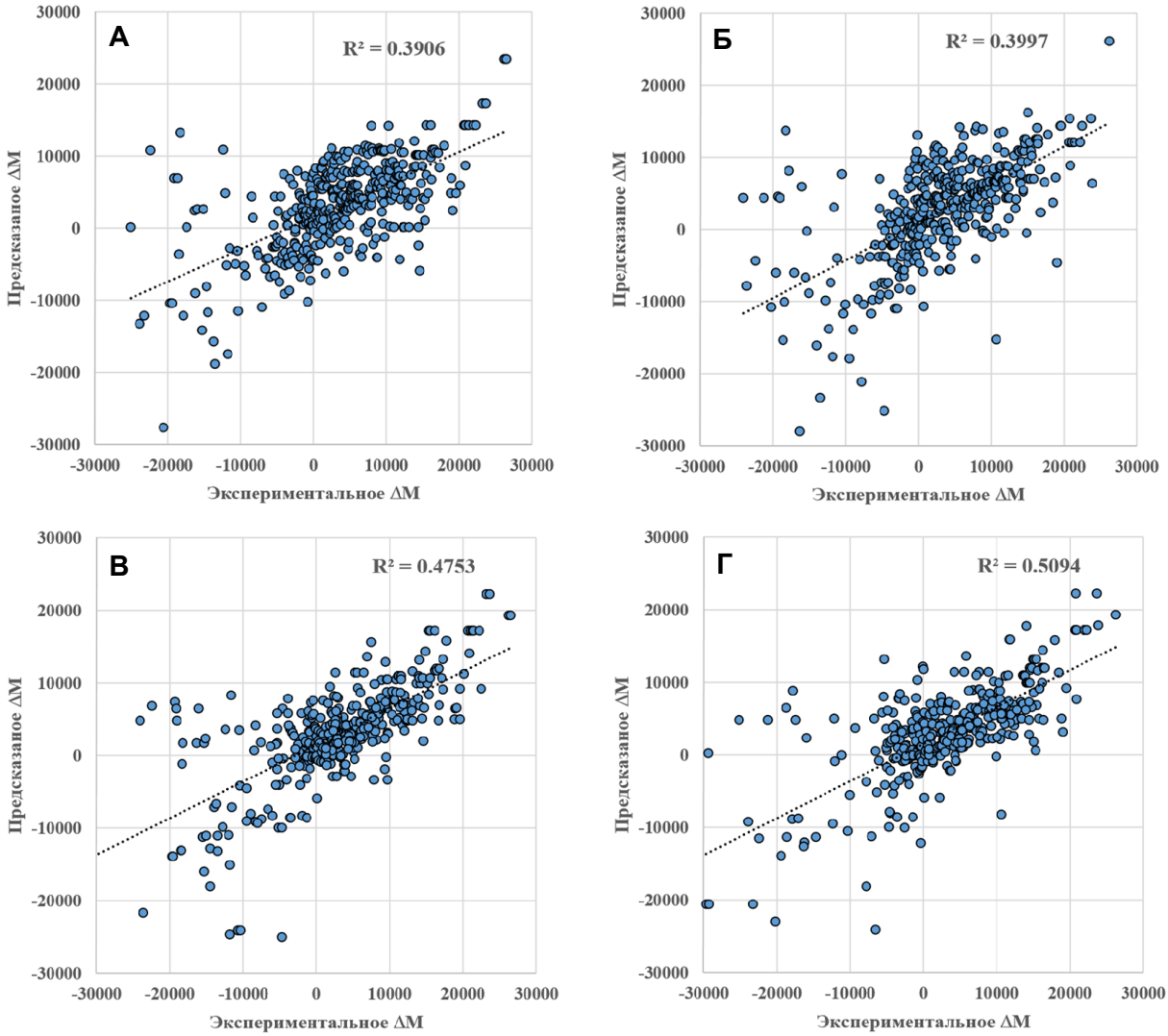

Рисунок 4. Перекрёстное предсказание по моделям, построенным для объединённой выборки. А. Сравнение экспериментально определённого ΔM и предсказанного по модели 1. Б. Сравнение экспериментально определённого ΔM и предсказанного (пересчитан из D_r, предсказанного по модели 2). В. Сравнение экспериментально определённого (M_r) и рассчитанного из брутто формулы (M_f) молекулярного веса. В. Сравнение экспериментально определённого (M_r) и расчётного молекулярного веса (с учётом предсказанной величины D_r).

пополам, должно давать репрезентативные модели. Выборка была поделена пополам, данные были отсортированы по значению ΔM. Соответственно, первую выборку образовали нечётные номера (O1), а вторую – чётные (O2). Результаты представлены в таблице и на рисунке 4. Если сравнивать с одиночными моделями для S0 и S176, то результаты не производят впечатления. Однако, попытка объединить только две “хорошие” выборки S0 и S176 даёт близкий по статистическим параметрам результат. Основное улучшение наблюдается в предсказательной силе моделей. Если предсказывать “чётную” выборку по “нечётной” и наоборот, то R^2 предсказания колеблется от 0,39 до 0,51. Направление изменений, даже не отбрасывая изменения меньше 1 кДа, предсказываются с точностью от 74% до 81%. А при предсказании целевой величины M_r R^2 предсказания 0,91 и 0,92 (при сравнении для M_f 0,86 и 0,85 соответственно).

Таким образом, можно построить уравнение предсказания кажущегося молекулярного веса и сделать, таким образом, картину виртуальной электрофоретической карты более реалистичной. Представленные модели интегрированы в программу pIPredict v.2, доступную по адресу http://www.ibmc.msk.ru/LPCIT/pIPredict, которая также демонстрирует различия в картине виртуальных электрофоретических карт при использовании различных подходов к вычислениям величин pI и M_r.

БЛАГОДАРНОСТИ

Работа выполнена в рамках Государственного задания ИБМХ Программы фундаментальных научных исследований государственных академий наук на 2013-2020 годы.

ЛИТЕРАТУРА

1. *Naryzhny S.N., Zgoda V.G., Maynskova M.A., Novikova S.E., Ronzhina N.L., Vakhrushev I.V., Archakov A.I.* (2016) Electrophoresis, **37**(2), 302-309.
2. *Bjellqvist B., Hughes GJ., Pasquali Ch., Paquet N., Ravier F., Sanchez J.-Ch., Frutiger S., Hochstrasser D.F.* (1993) Electrophoresis, **14**, 1023-1031.
3. *Скворцов В.С., Алексейчук Н.Н., Худяков Д.В., Ромеро Рейес И.В.* (2015) Биомед. химия, **61**(1), 83-91. DOI: 10.18097/PBMC20156101083
4. *Rath A., Cunningham F., Deber C.M.* (2013) Proc. Nat. Acad. Sci., **110**(39), 15668-15673.
5. *Rath A., Glibowicka M., Nadeau V.G., Chen G., Deber C.M.* (2009) Proc. Nat. Acad. Sci., **106**(6), 1760-1765.
6. *Rath A., Debe C.M.* (2013) Anal. Biochem., **434**(1), 67-72.
7. *Krokhin O.V., Anderson G., Spicer V., Sun L., Dovichi N.J.* (2017) Anal. Chem., **89**(3), 2000-2008.
8. *Deber C.M., Wang C., Liu L.P., Prior A.S., Agrawal S., Muskat B.L., Cuticchia A.J.* (2001) Protein Sci., **10**(1), 212-219.
9. *Bogaerts A., Temmerman G., Boerjan B., Husson S.J., Schoofs L., Verleyen P.* (2010) Develop. Compar. Immunol., **34**(6), 690-698.
10. *Bogaerts A., Beets I., Temmerman L., Schoofs L., Verleyen P.* (2010) Biology Direct, **5**, 11.
11. *D'Hertog W., Maris M., Thorrez L., Waelkens E., Overbergh L., Mathieu C.* (2011) Proteomics, **11**(7), 1365-1369.
12. *Krokhin O.V.* (2006) Anal. Chem., **78**(22), 7785-7795.
13. *Скворцов В.С., Алексейчук Н.Н., Худяков Д.В., Микурова А.В., Рыбина А.В., Новикова С.Е., Тихонова О.В.* (2015) Биомед. химия, **61**(6), 770-776. DOI: 10.18097/PBMC20156106770
14. *Belkina N.V., Krepets V.V., Shakin V.V.* (2002) Autom. Remote Control, **63**, 66-75.

Поступила: 30. 05. 2017.
Принята к печати: 09. 06. 2017.

CORRECTION OF THE ELECTROPHORETIC SHIFT IN VIRTUAL 2D SDS-PAGE ELECTROPHORESIS

V.S. Skvortsov, N.N. Alekseychuk, A.V. Rybina

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

Virtual electrophoresis in proteomics can be used to search localization of proteins and their proteoforms (especially those existing in low concentrations), to identify proteoforms found in experiments etc. Although the problem of predicting the isoelectric point is well studied, the need of electrophoretic shift correction is usually ignored. Researchers simply use the brutto molecular weight of the protein. In this study four data sets taken from the literature sources and the SWISS-2DPAGE database have been used to build correction equations for prediction of the electrophoretic shift (123, 72, 118 and 470 points, respectively). Two groups of models were built. The first model was based on the amino acid composition of proteins, the second one, on analysis of parameters calculated by amino acid sequences (theoretical molecular weight, hydrophobicity, charge distribution, ability to form helix structures). The coefficient of determination ranged from 0.35 to 0.75 in each single set, but cross-prediction between samples did not gave satisfactory results. At the same time, the direction of correction was predicted correctly in 74% of cases. After combining of the samples and dividing pooled data into 2 representative sets, the coefficient of determination during in the process of learning ranged from 0.44 to 0.51, and R^2 of predictions were not less than 0.39. The direction of correction was predicted correctly in 80% of cases. This prediction models have been integrated into the program pIPredict v.2, freely available at http://www.ibmc.msk.ru/LPCIT/pIPredict.

Key words: electrophoretic shift, virtual electrophoresis, statistical analysis