Биомедицинская химия, 2016 том 62, вып. 2, с. 173-179.

БИОИНФОРМАТИКА

УДК 541.69+519.25+518.5

КЛАССИФИКАЦИОННЫЕ МОДЕЛИ ВЗАИМОСВЯЗИ СТРУКТУРЫ ЛЕКАРСТВЕННЫХ СОЕДИНЕНИЙ И ИХ Р-ГЛИКОПРОТЕИНОВОЙ АКТИВНОСТИ

В.Ю. Григорьев*, С.Л. Солодова, Д.Е. Полианчик, О.А. Раевский

Институт физиологически активных веществ Российской академии наук,
142432, Московская область, Черноголовка, Северный пр. 1; эл. почта: beng@ipac.ac.ru

С использованием методов линейного дискриминантного анализа, случайного леса и опорных векторов созданы 33 классификационные модели субстратной специфичности 177 лекарственных соединений по отношению к Р-гликопротеину. QSAR моделирование проведено с использованием двух стратегий. Первая стратегия заключалась в переборе всех возможных комбинаций из 1÷5 дескрипторов на основе 7 наиболее значимых молекулярных дескрипторов с ясной физико-химической интерпретацией. Во втором случае было использовано последовательное включение в модель до 5 дескрипторов, начиная с лучшего одиночного дескриптора. Эту стратегию применяли для набора из 387 DRAGON дескрипторов. В результате было установлено, что только одна из 33-х моделей обладает необходимыми статистическими параметрами. Данная модель была сконструирована с помощью метода линейного дискриминантного анализа на основе одиночного дескриптора Н-связи (ΣC_{ad}). Она имеет хорошие статистические характеристики, как внутренней кросс-валидацией, так и внешней валидацией с применением 44 новых соединений. Это подтверждает важную роль водородной связи в процессах, связанных с проникновением химических соединений через гематоэнцефалический барьер.

Ключевые слова: QSAR, водородная связь, ГЭБ, Р-гликопротеин, HYBOT

DOI: 10.18097/PBMC20166202173

ВВЕДЕНИЕ

Р-гликопротеин (Pgp) представляет собой крупный трансмембранный белок с молекулярной массой 170 кДа, состоящий из 1280 аминокислотных остатков [1]. Он является распространённым транспортером, входящим в состав люминальных мембран эндотелиоцитов, который осуществляет транспорт большого числа структурно разнородных цитотоксических лекарств и других липофильных соединений через мембраны [2], препятствуя при этом достижению терапевтических концентраций ряда физиологически активных веществ в головном мозге [3]. Недавно установлено, что Pgp экспрессируется не только на люминальной мембране, но и на внутриклеточных мембранах [4].

Взаимодействие низкомолекулярных соединений с Pgp происходит по весьма сложному механизму, который обусловлен тем, что этот белок содержит несколько сайтов связывания, а также тем, что в результате связывания Pgp с АТРазой происходит изменение его конформации, что также влияет на связывание белка с субстратом [5].

Пространственная структура этого белка в настоящее время мало изучена. Так, в работе [6] была представлена трёхмерная структура Pgp с разрешением ~ 8 Е в комплексе с AMP-PNP, а в публикации [7] структура Pgp мыши – с разрешением около 4 Е. Хотя эти результаты и дают представление об общей трёхмерной структуре и функционировании Pgp, достигнутая точность ещё далека от необходимой для молекулярного моделирования взаимодействия Pgp с субстратами и ингибиторами. Ввиду этого понятно наличие в литературе многих публикаций (см. например, обзор [8]), связанных с моделированием *in silico* взаимосвязи между структурой соединений и их транспортом в центральную нервную систему (ЦНС).

Анализ этих публикаций показывает, что к исследованиям в этой области привлечены практически все современные QSAR технологии (методы, дескрипторы, модели). Тем не менее, здесь можно отметить два важных обстоятельства: (1) использование при построении моделей достаточно большого числа дескрипторов, что затрудняет интерпретацию полученных результатов; (2) недостаточно глубокое описание водородной связи. Первое обстоятельство препятствует созданию устойчивых предсказательных механистических моделей структура-активность для модификации структуры химических соединений с целью получения веществ с заданными свойствами. Что же касается водородной связи, в настоящее время это взаимодействие считается одним из важных (если не самым важным) в межмолекулярных взаимодействиях вообще, и в субстрат-рецепторных комплексах в частности [8]. Однако количественное описание этого эффекта в рамках QSAR или молекулярного моделирования подчас проводится недостаточно корректно.

В настоящем исследовании поставлена задача создания устойчивых предсказательных классификационных QSAR моделей взаимодействия химических соединений и лекарств с Pgp и их транспорта через гематоэнцефалический барьер (ГЭБ) на основе небольшого числа

* - адресат для переписки

хорошо интерпретируемых физико-химических дескрипторов с включением в эти модели дескрипторов водородной связи.

МЕТОДИКА

Данные

В настоящей работе были использованы тестовые литературные данные по определению субстратной специфичности соединений к Pgp, основанные на изучении клеточной проницаемости. При этом в качестве основной характеристики были использованы величины ER (Efflux Ratio), которые представляют собой отношение коэффициентов проницаемости соединений в двух противоположных направлениях. Соединения, имеющие величины ER>2, рассматривались как субстраты, остальные – как несубстраты. В обучающую выборку были включены данные по субстратной специфичности соединений, полученные на основе линии клеток MDCK-MDR1 [9], а в тестовую выборку вошли данные из работы [10], основанные на использовании клеток Caco-2, исключая общие соединения из этих двух публикаций. Таким образом, данные по классификации субстрат/несубстрат включали 221 соединение. 177 соединений обучающей выборки содержали 71 субстрат и 106 несубстратов, а 44 соединения тестового ряда содержали 20 субстратов и 24 несубстрата. На рисунке представлено распределение соединений по наличию в них 17 наиболее распространенных в органической химии функциональных групп. Из рисунка очевидно пропорциональное присутствие одних и тех же функциональных групп в обучающей и тестовой выборках.

Дескрипторы

С использованием компьютерных программ HYBOT [11], MOLTRA [12], DRAGON [13] для описания молекулярной структуры соединений было рассчитано два ряда дескрипторов. Первый ряд состоял из сознательно выбранных 15 физико-химических дескрипторов, связанных в основном с описанием межмолекулярного взаимодействия, включая характеристики Н-связи и липофильность молекул. Выбор этих дескрипторов обусловлен решающей ролью липофильности [14] и особенно способности соединений к образованию водородной связи [14, 15] при создании QSAR моделей взаимодействия соединений с Pgp. Второй ряд представлял из себя набор из 1064 доступных 2D дескрипторов, рассчитанных на основе программы DRAGON [13], список которых приведён в Приложении.

Дальнейший отбор дескрипторов был связан с необходимостью обеспечения линейной независимости переменных. Для этого был проведен анализ корреляционной матрицы дескрипторов с использованием итерационной процедуры, которая состояла из ряда шагов: 1) выбор наиболее

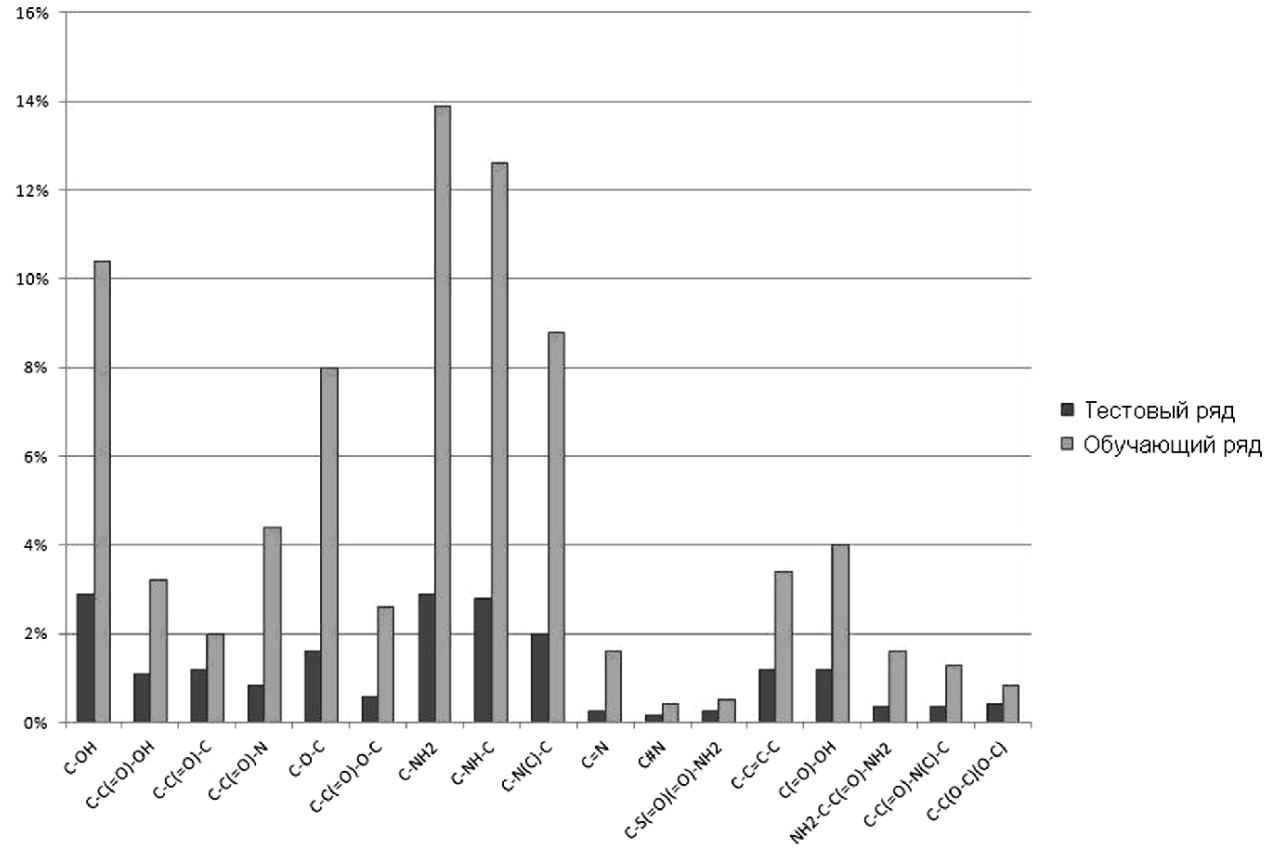

Рисунок. Частота встречаемости (%) 17 наиболее распространённых функциональных групп в соединениях обучающего и тестового ряда.

информативного дескриптора с максимальной величиной коэффициента вариации; 2) формирование вокруг этого выбранного дескриптора кластера родственных дескрипторов на основе граничной величины коэффициента корреляции (0,90-0,95); 3) удаление из матрицы родственных дескрипторов; 4) повторение шагов 1-3 до остановки процедуры. В результате из первого ряда было отобрано 7 дескрипторов (табл. 1), а из второго ряда – 387 дескрипторов.

Методы классификации

Бинарную классификацию соединений проводили с использованием 3-х методов: линейного дискриминантного анализа (ЛДА), случайного леса (СЛ) и опорных векторов (ОВ). Для расчёта классификационных моделей на основе первого ряда из 7 дескрипторов применяли метод полного перебора всех возможных комбинаций дескрипторов. Расчёты на основе второго ряда из 387 дескрипторов выполняли с помощью метода последовательного включения от одного до пяти дескрипторов в модель, выбирая лучшие из них. В качестве критерия для отбора наиболее значимых моделей использовали коэффициент корреляции Мэтьюза (MCC).

Линейный дискриминантный анализ проводили с использованием компьютерной программы LDA [16]. В качестве метрики в исследуемом пространстве дескрипторов использовали расстояние Махаланобиса.

Для классификации на основе метода случайного леса использовали программу rf5new [17]. При этом параметры метода имели следующие значения: число деревьев (jbt=500), число случайно выбираемых переменных (mtry0=$(M+0,5)^{0,5}$), где M – общее число дескрипторов в модели.

Для создания классификационных моделей на основе опорных векторов применяли компьютерную программу flssvm [18]. Для решения задачи использовали радиальную базисную функцию с заданными параметрами (без их оптимизации). Дескрипторы автошкалировали на основе формулы: $B_i=(A_i-Г)/S$, где A_i и B_i – исходные и нормалированные величины дескриптора для i-ой молекулы, B – средняя величина и S – стандартное отклонение дескриптора.

Бинарные классификационные модели оценивали на основе TP (число правильно распознанных соединений, принадлежащих к первому классу (субстраты)), FN (число ошибочно распознанных соединений среди первого класса), TN (число правильно распознанных соединений, принадлежащих ко второму классу (несубстраты)) и FP (число ошибочно распознанных соединений среди второго класса). С использованием величин TP, FN, TN и FP рассчитывали общую точность ACC=(TP+TN)/(TP+FN+TN+FP) и MCC=$(TP\cdot TN–FP\cdot FN)/((TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN))^{0,5}$.

Валидацию полученных классификационных моделей проводили на основе двух подходов. В первом случае использовали внутреннюю кросс-валидацию (CV) обучающей выборки с выбором по одному (leave-one-out). (В методе СЛ аналогичную роль выполняла процедура OOB (out-of-bag)). Во втором случае применяли внешнюю валидацию на базе тестовой выборки.

Оценку структурного сходства молекул осуществляли с помощью коэффициентов Танимото (T_c) [19] с использованием компьютерной программы MOLDIVS [20].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Классификационные модели, полученные с использованием 7 физико-химических дескрипторов

Среди специалистов по QSAR моделированию уже достаточно длительное время ведется дискуссия о предпочтительности внутренней или внешней валидации для оценки предсказательной способности моделей [21]. В связи с этим при анализе качества классификационных моделей мы использовали оба этих подхода. Результаты работы на основе первого ряда из семи дескрипторов представлены в таблице 2.

Как видно из представленных данных, при использовании ЛДА среди одиночных дескрипторов наибольшим дискриминирующим эффектом обладает ΣC_{ad}, то есть дескриптор, который характеризует суммарную H-акцепторную и H-донорную способность молекул. Использование одного этого дескриптора позволило построить

Таблица 1. Физико-химические молекулярные дескрипторы

Дескриптор	Интерпретация	Интервал	Программа
ΣQ^-	Сумма отрицательных атомных зарядов	–7,78÷ –0,66	HYBOT
ΣC_{ad}	Сумма свободноэнергетических H-акцепторных и H-донорных дескрипторов	2,70÷38,50	HYBOT
MW	Молекулярная масса, а.е.м.	141,17÷1202,63	HYBOT
PSA_{ea}	Парциальный поверхностный дескриптор суммарной энтальпийной H-акцепторной способности, E^2	1,80÷45,96	MOLTRA
PSA_{ca}	Парциальный поверхностный дескриптор суммарной свободноэнергетической H-акцепторной способности, E^2	1,64÷54,53	MOLTRA
PSA_{ed}	Парциальный поверхностный дескриптор суммарной энтальпийной H-донорной способности, E^2	0,00÷15,95	MOLTRA
$TPSA_{tot}$	Общая поверхность полярных взаимодействий, E^2	3,24÷278,80	DRAGON

Таблица 2. Статистические характеристики классификационных моделей, рассчитанных на основе 7 дескрипторов

№	Метод	Дескрипторы	Обучающий ряд				Тестовый ряд	
			ACC	MCC	ACC_{cv}	MCC_{cv}	ACC	MCC
1	ЛДА	ΣC_{ad}	0,791	0,558	0,785	0,546	0,818	0,639
2	ЛДА	ΣC_{ad}, MW	0,808	0,597	0,802	0,584	0,727	0,456
3	ЛДА	ΣQ^-, ΣC_{ad}, MW	0,819	0,621	0,814	0,609	0,636	0,262
4	ЛДА	ΣQ^-, ΣC_{ad}, MW, PSA_{ea}	0,825	0,634	0,808	0,597	0,659	0,31
5	ЛДА	ΣQ^-, ΣC_{ad}, MW, PSA_{ca}, $TPSA_{tot}$	0,814	0,611	0,768	0,509	0,591	0,232
6	ЛДА	ΣQ^-, ΣC_{ad}, MW, PSA_{ea}, PSA_{ca}, PSA_{ed}, $TPSA_{tot}$	0,768	0,51	0,734	0,434	0,591	0,232
7	СЛ	PSA_{ca}	0,989	0,976	0,644	0,264	0,75	0,494
8	СЛ	MW, $TPSA_{tot}$	0,994	0,988	0,791	0,562	0,591	0,203
9	СЛ	ΣC_{ad}, MW, $TPSA_{tot}$	0,994	0,988	0,774	0,524	0,659	0,324
10	СЛ	MW, PSA_{ea}, PSA_{ca}, PSA_{ed}	0,994	0,988	0,78	0,536	0,636	0,283
11	СЛ	ΣC_{ad}, MW, PSA_{ea}, PSA_{ca}, $TPSA_{tot}$	0,994	0,988	0,78	0,54	0,682	0,365
12	СЛ	ΣQ^-, ΣC_{ad}, MW, PSA_{ea}, PSA_{ca}, PSA_{ed}, $TPSA_{tot}$	0,994	0,988	0,763	0,5	0,659	0,324
13	ОВ	MW	0,729	0,437	0,729	0,437	0,659	0,309
14	ОВ	MW, $TPSA_{tot}$	0,746	0,474	0,734	0,442	0,682	0,356
15	ОВ	MW, PSA_{ca}, PSA_{ed}	0,757	0,508	0,706	0,381	0,659	0,309
16	ОВ	ΣC_{ad}, MW, PSA_{ed}, $TPSA_{tot}$	0,785	0,571	0,74	0,458	0,636	0,258
17	ОВ	ΣC_{ad}, MW, PSA_{ca}, PSA_{ed}, $TPSA_{tot}$	0,785	0,578	0,712	0,391	0,659	0,306
18	ОВ	ΣQ^-, ΣC_{ad}, MW, PSA_{ea}, PSA_{ca}, PSA_{ed}, $TPSA_{tot}$	0,819	0,638	0,684	0,316	0,705	0,402

модель (1) с достаточно высокими статистическими характеристиками как описательной, так и предсказательной способности. В первом случае ACC=0,791, MCC=0,558. Во втором случае при использовании внутренней валидации ACC_{cv}=0,785, MCC_{cv}=0,546, а внешней – ACC=0,818, MCC=0,639. Дальнейшее увеличение размера дескрипторного пространства немного улучшало кросс-валидированную статистику. Максимальные величины были получены для модели (3) из 3-х дескрипторов (ACC_{cv}=0,814, MCC_{cv}=0,609). Однако статистические характеристики, связанные с внешним тестированием, при этом оставались на неудовлетворительном низком уровне (ACC<0,75, MCC<0,50).

Похожая ситуация была и при последовательном конструировании моделей на основе СЛ. Однако в случае одиночного дескриптора (PSA_{ca}, модель 7) статистические параметры кросс-валидации оказались неудовлетворительными (ACC_{cv}=0,644, MCC_{cv}=0,264). Дальнейшее увеличение числа дескрипторов (модели 8÷12) приводило к получению максимальных величин описательной статистики (ACC=0,994, MCC=0,988), удовлетворительных величин предсказательной статистики в случае внутреннего тестирования и неудовлетворительных показателей при внешнем тестировании.

При создании классификационных ОВ моделей наиболее “весомым” среди одиночных дескрипторов оказалась молекулярная масса MW (модель 13). Ни одна из разработанных моделей (13÷18) не может быть признана удовлетворительной. Только в случае моделей (15÷18) наблюдаются приемлемые величины описательной статистики: ACC>0,75 и MCC>0,50. При этом предсказательная способность моделей не достигала минимально необходимых показателей.

Интересно отметить разное поведение статистических показателей внешней тестовой выборки в зависимости от используемого метода и размерности пространства. Так, в случае ЛДА при переходе от одномерной к оптимальной (с учётом кросс-валидации) многомерной классификации величины ACC и MCC резко уменьшаются: ACC=0,818, MCC=0,639 (модель 1) и ACC=0,636, MCC=0,262 (модель 3). Аналогичное поведение демонстрируют СЛ модели: ACC=0,750, MCC=0,494 (модель 7) и ACC=0,591, MCC=0,203 (модель 8). Напротив, при использовании ОВ классификации переход от одномерного (модель 13) к семимерному (модель 18) дескрипторному пространству сопровождался увеличением степени общего правильного распознавания (ACC) с 65,9% до 70,5%.

Классификационные модели, полученные с использованием 387 DRAGON дескрипторов

В таблице 3 представлены результаты конструирования классификационных моделей с использованием метода последовательного включения дескрипторов на основе максимальных величин описательной статистики MCC.

Из анализа этих данных следует, что выбранный метод классификации оказывает большое влияние на состав дескрипторов. При использовании трёх методов: ЛДА, СЛ и ОВ общее число дескрипторов равно 15. Большинство из них относится к топологическим индексам и атом-центрированным фрагментам. При этом каждому методу присущ свой уникальный набор дескрипторов, пересечений не наблюдается.

Таблица 3. Статистические характеристики классификационных моделей, рассчитанных на основе 387 дескрипторов

№	Метод	Дескрипторы	Обучающий ряд				Тестовый ряд	
			ACC	MCC	ACC_{cv}	MCC_{cv}	ACC	MCC
19	ЛДА	QW	0,763	0,504	0,763	0,504	0,682	0,356
20	ЛДА	QW, EEig02r	0,791	0,56	0,791	0,56	0,614	0,218
21	ЛДА	QW, EEig02r, H-049	0,808	0,599	0,802	0,586	0,591	0,169
22	ЛДА	QW, EEig02r, H-049, GATS8m	0,814	0,611	0,802	0,586	0,568	0,126
23	ЛДА	QW, EEig02r, H-049, GATS8m, F06[O-O]	0,825	0,636	0,808	0,597	0,568	0,119
24	СЛ	MAXDP	0,994	0,988	0,599	0,148	0,455	-0,056
25	СЛ	MAXDP, CIC1	0,994	0,988	0,74	0,451	0,568	0,126
26	СЛ	MAXDP, CIC1, O-060	0,994	0,988	0,791	0,558	0,591	0,169
27	СЛ	MAXDP, CIC1, O-060, X5v	0,994	0,988	0,802	0,585	0,591	0,203
28	СЛ	MAXDP, CIC1, O-060, X5v, JGI7	0,994	0,988	0,831	0,643	0,591	0,203
29	ОВ	DECC	0,751	0,501	0,74	0,478	0,659	0,391
30	ОВ	DECC, C-016	0,797	0,587	0,763	0,521	0,682	0,388
31	ОВ	DECC, C-016, F02[N-N]	0,831	0,649	0,774	0,532	0,659	0,324
32	ОВ	DECC, C-016, F02[N-N], MLOGP	0,853	0,694	0,774	0,526	0,682	0,388
33	ОВ	DECC, C-016, F02[N-N], MLOGP, C-026	0,893	0,776	0,718	0,405	0,568	0,175

Примечание. QW - квази-Винеровский индекс; EEig02r - характеристическое число, полученное на основе матрицы связности; H-049 - атом-центрированный фрагмент H/C; GATS8m - автокорреляция Geary, взвешенная на основе атомных масс; F06[O-O] - частота O-O при топологической дистанции 6; MAXDP - максимальный электротопологический индекс с положительной вариацией; CIC1 - комплементарное информационное содержание 1-го порядка; O-060 - атом-центрированный фрагмент Al-O-Ar/Ar-O-Ar/R..O..R/R-O-C=X; X5v - индекс валентной связности chi-5; JGI7 - средний топологический зарядный индекс 7-го порядка; DECC - эксцентричность; C-016 - атом-центрированный фрагмент =CHR; F02[N-N] - частота N-N при топологической дистанции 2; MLOGP - коэффициент распределения в системе октанол-вода Моригучи; C-026 - атом-центрированный фрагмент R--CX--R.

Все представленные модели (19÷33) характеризуются хорошей описательной статистикой: величины АСС меняются от 0,751 до 0,994, а величины МСС от 0,501 до 0,988. Параметры предсказательной кросс-валидированной статистики большинства анализируемых моделей также выглядят вполне удовлетворительно. Только в 4-х случаях из 15 (модели 24, 25, 29, 33) величины АСС и МСС были ниже приемлемых значений. Однако при использовании внешнего тестового ряда приходится констатировать, что ни одна из разработанных классификационных моделей не удовлетворяет минимальным требованиям в отношении величин АСС и МСС.

Необходимо отметить разный характер общей корреляции величин МСС(АСС) тестовой выборки и соответствующих величин MCC_{cv}, (ACC_{cv}) обучающей выборки для сконструированных моделей. Так, в отличие от моделей (1)÷(18), где рост MCC_{cv}, сопровождался падением величин МСС тестовой выборки с коэффициентом корреляции Пирсона r=-0,2, в ряду моделей (19)÷(33) наблюдалась положительная корреляция с величиной r=0,4.

Ранее в литературе уже были опубликованы результаты бинарной классификации 187 соединений (177 из которых использованы в настоящей работе в качестве обучающего ряда) [9]. При этом с применением методов СЛ, ОВ и к-ближайшего соседа было создано 23 классификационные модели на основе 30 различных дескрипторов. Лучшие результаты были получены с использованием 3D молекулярных дескрипторов VS+. При этом минимальное число дескрипторов было равно 4, предсказательная способность, оцененная на основе кросс-валидации и внешнего тестирования, имела близкие значения и составляла для АСС и МСС ~0,80 и ~0,60 соответственно. Разработанная в настоящей работе классификационная ЛДА модель (1) при сопоставимых статистических характеристиках выгодно отличается тем, что она имеет один дескриптор (ΣC_{ad}) с ясной физико-химической интерпретацией. Кроме того, использованный дескриптор является двумерным, что снимает ряд вопросов, связанных с моделированием структуры молекул.

Необходимой характеристикой конструируемых QSAR моделей является определение их области применимости (ОП). Что касается модели (1), то в качестве дескрипторной ОП может быть использован соответствующий интервал для ΣC_{ad} (табл. 1): 2,70÷38,50. Для оценки структурной ОП может быть применён способ [22], основанный на расчете и сопоставлении максимальных величин структурного сходства (например, коэффициентов Танимото (T_c)) тестируемых молекул и молекул обучающей выборки. Так, в частности, в использованном в данной работе тестовом ряду (44 соединения) доля молекул, имеющих величины T_c>0,5, составила более 77%.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

С использованием методов линейного дискриминантного анализа, случайного леса и опорных векторов на основе двух рядов из 7 и 387 дескрипторов сконструировано 33 классификационные модели. Для их создания была использована обучающая выборка из 177 соединений. Поиск лучших моделей осуществлялся с применением двух стратегий: путём перебора всех возможных комбинаций из 1÷5 дескрипторов или путём их последовательного включения в модель. При этом в качестве критерия для отбора лучших QSAR моделей использовался корреляционный коэффициент Мэтьюза обучающей выборки. Для валидации моделей использовали внутреннюю кросс-валидацию и внешний тестовый ряд из 44-х молекул. В результате было установлено, что только одна модель, полученная с использованием метода ЛДА на основе одного дескриптора ΣC_{ad} (табл. 2, модель 1), имеет удовлетворительные характеристики как описательной, так и предсказательной способности и может быть рекомендована к практическому применению с учётом ее области применимости. Полученный результат подтверждает важную роль водородной связи в процессах Pgp транспорта лекарственных соединений через гематоэнцефалический барьер.

Приложение доступно в электронной версии статьи на сайте журнала (pbmc.ibmc.msk.ru).

ЛИТЕРАТУРА

1. *Li Y., Yuan H., Yang K., Xu W., Tang W., Li X.* (2010) Curr. Med. Chem., **17**, 786-800.
2. *Ramachandra M., Ambudkar S.V., Chen D., Hrycyna C.A., Dey S., Gottesman M.M., Pastan I.* (1998) Biochemistry, **37**, 5010-5019.
3. *Miller D.S., Bauer B., Hartz A.M.S.* (2008) Pharmacol. Rev., **60**, 196-209.
4. *Bendayan R., Ronaldson P.T., Gingras D., Bendayan M.* (2006) J. Histochem. Cytochem., **54**, 1159-1167.
5. *Higgins C.F., Linton K.J.* (2004) Nat. Struct. Mol. Biol., **11**(10), 918-926.
6. *Rosenberg M.F., Callaghan R., Modok S., Higgins C.F., Ford R.C.* (2005) J. Biol. Chem., **280**, 2857-2862.
7. *Aller S.G., Yu J., Ward A., Weng Y., Chittaboina S., Zhuo R., Harrell P.M., Trinh Y.T., Zhang Q., Urbatsch I.L., Chang G.* (2009) Science, **323**, 1718-1722.
8. *Раевский О.А., Солодова С.Л., Лагунин А.А., Поройков В.В.* (2014) Биомед. химия, **60**, 161-181.
9. *Broccatelli F.* (2012) J. Chem. Inf. Model., **52**, 2462-2470.
10. *Crivori P., Reinach B., Pezzetta D., Poggesi I.* (2006) Molecular Pharmaceutics, **3**, 33-44.
11. *Раевский О.А., Григорьев В.Ю., Трепалин С.В.* Свидетельство об официальной регистрации программы для ЭВМ HYBOT (Hydrogen Bond Thermodynamics) № 990090 от 26 февраля 1999 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.
12. *Раевский О.А., Трепалин С.В. Раздольский А.Н.* Свидетельство об официальной регистрации программы для ЭВМ MOLTRA (Molecular Transform Analysis) № 990092 от 26 февраля 1999 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.

13. URL: http://www.talete.mi.it
14. *Seelig A., Landwojtowicz E.* (2000) Eur. J. Pharm. Sci., **12**, 31-40.
15. *Ecker G., Huber M., Schmid D., Chiba P.* (1999) Mol. Pharmacol., **56**, 791-796.
16. *Murtagh F., Heck A.* (1987) Multivariate Data Analysis, D. Reidel Publ. Co., Dordrecht, pp. 128-133.
17. URL: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_examples/prog.f
18. URL: https://github.com/jbcolme/fortran-ls-svm
19. *Rogers D.J., Tanimoto T.T.* (1960) Science, **132**, 1115-1118.
20. *Раевский О.А., Герасименко В.А., Трепалин С.В.* Свидетельство об официальной регистрации программы для ЭВМ MOLDIVS (Molecular Diversity & Similarity) № 990093 от 26 февраля 1999 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.
21. *Veerasamy R., Rajak H., Jain A., Sivadasan S., Varghese C.P., Agrawal R.K.* (2011) Int. J. Drug Des. Discov., **2**(3), 511-519.
22. *Sahigara F., Mansouri K., Ballabio D., Mauri A., Consonni V., Todeschini R.* (2012) Molecules, **17**, 4791-4810.

Поступила: 25. 06. 2015.
Принята к печати: 12. 10. 2015.

CLASSIFICATION MODELS OF STRUCTURE - P-GLYCOPROTEIN ACTIVITY OF DRUGS

V.Yu. Grigorev, S.L. Solodova, D.E. Polianczyk, O.A. Raevsky

Institute of Physiologically Active Compounds, Russian Academy of Science,
Moscow region, Chernogolovka, 1 Severniy proezd, 142432, Russia; e-mail: beng@ipac.ac.ru

Thirty three classification models of substrate specificity of 177 drugs to P-glycoprotein have been created using of the linear discriminant analysis, random forest and support vector machine methods. QSAR modeling was carried out using 2 strategies. The first strategy consisted in search of all possible combinations from 1÷5 descriptors on the basis of 7 most significant molecular descriptors with clear physico-chemical interpretation. In the second case forward selection procedure up to 5 descriptors, starting from the best single descriptor was used. This strategy was applied to a set of 387 DRAGON descriptors. It was found that only one of 33 models has necessary statistical parameters. This model was designed by means of the linear discriminant analysis on the basis of a single descriptor of H-bond (ΣC_{ad}). The model has good statistical characteristics as evidenced by results to both internal cross-validation, and external validation with application of 44 new chemicals. This confirms an important role of hydrogen bond in the processes connected with penetration of chemical compounds through a blood-brain barrier.

Key words: QSAR, Hydrogen bond, BBB, P-glycoprotein, HYBOT