УДК 541.6

Одноклассовый подход: модели для виртуального скрининга ненуклеозидных ингибиторов обратной транскриптазы ВИЧ-1 на основе концепции непрерывных молекулярных полей*

П. В. Карпов,[а] И. И. Баскин,[а]★ Н. И. Жохова,[а] М. Б. Навроцкий,[б] А. Н. Зефиров,[а] А. С. Яблоков,[б] И. А. Новаков,[б] Н. С. Зефиров[а]

[а]Московский государственный университет имени М. В. Ломоносова, Химический факультет, Российская Федерация, 119991 Москва, Ленинские горы, 1, стр. 3. Факс: (495) 939 0290. E-mail: igbaskin@gmail.com; zhokhovann@gmail.com
[б]Волгоградский государственный технический университет, Российская Федерация, 400131 Волгоград, просп. Ленина, 28. Факс: (8442) 23 8125. E-mail: kholstaedt@yandex.ru

Впервые в рамках одноклассового подхода с использованием метода опорных векторов построены одноклассовые модели для виртуального скрининга потенциальных ненуклеозидных ингибиторов обратной транскриптазы ВИЧ. Обучающая выборка включала данные по 786 структурам производных 2-замещенных пиримидинонов и их ингибирующей активности в отношении фермента для дикого и мутантных (K103, IRLL98, Y188L) штаммов ВИЧ-1. Представление молекулярной структуры органических лигандов на основе непрерывных молекулярных полей позволяет получать классификационные модели более высокого качества по сравнению с традиционными подходами, базирующимися на использовании фрагментных дескрипторов Кархарта, «молекулярных отпечатков» и спектрофоров.

Ключевые слова: органические соединения, ингибиторы обратной транскриптазы, ВИЧ, HIVRT, одноклассовая классификация, виртуальный скрининг, непрерывные молекулярные поля, моделирование биологической активности.

Поиск новых высокоэффективных ингибиторов обратной транскриптазы вируса иммунодефицита человека (HIVRT)** представляет одну из важных задач современной медицинской химии. Обратная транскриптаза ВИЧ на основе РНК вириона синтезирует комплементарную ей одноцепочную ДНК, которая после достраивания второй цепи, встраивается в молекулу ДНК хозяина, поражая его здоровые клетки. Ингибирование процесса обратной транскрипции приводит к подавлению распространения вируса, и воздействие на него является важным этапом современной противовирусной терапии. При создании новых соединений-лидеров анти-ВИЧ-препаратов особое внимание исследователей привлекает широкая группа органических соединений, которая относится к типу ненуклеозидных ингибиторов HIVRT и включает разные структурные классы (НЕРТ***, DABO**** и др.)[1,2]. Эти соединения обладают низкой токсичностью, высокой активностью и селективностью в ингибировании HIVRT, а их использование в комплексной высокоактивной антивирусной терапии позволяет значительно уменьшить риск смертности от СПИДа. Однако несмотря на то что применение известных соединений этой группы приводит к значительному снижению скорости репликации ВИЧ, полностью подавить ее не удается, а развитие резистентности вируса ставит задачу разработки новых более эффективных ненуклеозидных ингибиторов HIVRT с различными профилями резистентности[3].

Виртуальный скрининг библиотек органических лигандов является эффективным инструментом при поиске и конструировании новых соединений-лидеров лекарств. Одним из этапов при создании системы виртуального скрининга является построение моделей в рамках методов 2D-3D QSAR*. В литературе описано большое число регрессионных QSAR-моделей, полученных для оценки ингибирующей активности узких серий органических соединений по отношению к HIVRT, которые имеют весьма высокие статистические характеристики прогноза IC_{50} и EC_{50}

*Посвящается академику Российской академии наук О. М. Нефедову в связи с его 80-летием.

**HIVRT — Human immunodeficiency virus reverse transcriptase.

***НЕРТ — 1-[(2-гидроксиэтокси)метил]-6-(тиофенил)тимин (1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine).

****DABO — дигидроксиалкоксибензилоксопиримидины (dihydroalkoxybenzyloxopyrimidines).

*QSAR — количественное соотношение структура—активность (quantitative structure activity relationship).

($q^2 \approx 0.8$)[4]. Тем не менее такие модели не позволяют проводить виртуальный скрининг новых потенциальных ингибиторов внутри широких подклассов лигандов, а использование специальных методов по объединению узкоспециализированных моделей[5] сопряжено с излишней сложностью и снижением точности прогноза за счет добавления ошибки классификации. QSAR-модели, построенные на широких выборках, включающих органические структуры разных классов, как правило, имеют значительно более низкие статистические характеристики, и поэтому их применение приводит к большим ошибкам прогноза и к «засоренности» результатов виртуального скрининга ложно активными структурами.

Отдельную проблему количественных QSAR-моделей представляет корректная оценка их порога активности, варьирование которого влияет на набор соединений, получаемых в результате виртуального скрининга, что может как результат ошибки прогноза привести к потере истинно активных перспективных соединений.

Альтернативным подходом, который отчасти позволяет избежать проблем регрессионных QSAR-моделей, является использование для виртуального скрининга классификационных моделей, которые позволяют на качественном уровне прогнозировать, будут ли тестируемые соединения обладать активностью или нет. В данном случае вывод об отнесении соединения к классу активных или неактивных структур осуществляется на основе специальных процедур оценки принадлежности тестируемой структуры к моделируемым классам. Однако для корректного построения двухклассовой модели необходимо иметь в наличии две представительные выборки для активных и неактивных соединений. К сожалению, при моделировании многих видов биологической активности представительная выборка неактивных соединений не всегда доступна[6–8].

В работе[9] нами был апробирован новый подход к построению моделей для виртуального скрининга, основанный на одноклассовой классификации, при котором для построения моделей необходима только выборка примеров активных соединений. В хемоинформатике принцип одноклассовой классификации применяется для определения областей применимости QSAR/QSPR*-моделей[10,11].

В настоящей работе мы применили одноклассовый подход при построении моделей для виртуального скрининга потенциальных ненуклеозидных ингибиторов HIVRT. В рамках этой задачи с целью выбора оптимального решения при построении моделей представляло интерес сопоставление различных способов представления молекулярной структуры органических лигандов (как на основе традиционных дескрипторов: «молекулярных отпечатков», спектрофоров[12,13], фрагментных дескрипторов Кархарта[14,15], так и непрерывных молекулярных полей[16]).

*QSPR — количественные соотношения структура—свойство (Quantitave structure activity relationship).

Методика расчетов

Обучающая выборка органических лигандов содержала данные по ингибирующей активности (EC_{50}) HIVRT дикого и мутантных (K103, IRLL98, Y188L) штаммов ВИЧ-1 для 786 соединений, включающих производные 2-замещенных (6-арилметил)пиримидин-4(3*H*)-онов (DABO). Молекулярные массы соединений варьируются в диапазоне от 202 до 550 дальтонов, липофильность — от 3.4 до 8.0. В выборке представлены как оригинальные данные (предоставлены группой М. Б. Навроцкого), так и взятые из литературы.

При построении моделей для представления молекулярной структуры соединений использовали следующие дескрипторы: «молекулярные отпечатки», спектрофоры и дескрипторы Кархарта. Первые два типа рассчитывали на основе стандартных функций свободно распространяемой программы OpenBabel[17]. Модифицированные дескрипторы Кархарта вычисляли с использованием разработанного нами дескрипторного блока CARHART[9]. В этом блоке принцип генерации цепочечных фрагментов по Кархарту мы объединили со схемой кодирования блока FRAGMENT[18], которая учитывает в структуре соединения гибридизацию, формальный заряд, связевое окружение, а также число соседних атомов водорода. Такая комбинация позволяет улучшить качество описания молекулярной структуры соединений.

Одноклассовые модели, как на основе традиционных дескрипторов, так и с помощью непрерывных молекулярных полей, строили с использованием разрабатываемого нами[16] программного комплекса MCMF*. В качестве ядерного метода машинного обучения использовали метод опорных векторов (MOB, 1-SVM, Support Vector Machine), реализованный в библиотеке LIBSVM**. Оценку прогнозирующей способности моделей осуществляли с помощью процедуры скользящего контроля с исключением по одному. Качество одноклассовых моделей оценивали по площади под ROC-кривой*** (см. лит.[19]). Для оптимизации параметров метода опорных векторов, статистических ядер и параметров непрерывных молекулярных полей мы использовали реализованные в библиотеке NlOpt**** функции эмпирического поиска экстремумов функции в определенной области изменения параметров.

Выравнивание базы органических структур проводили при помощи разработанной нами программы на основе алгоритма SEAL[20].

Обсуждение полученных результатов

Особенность построения моделей с использованием одноклассовой классификации заключается в том, что в распоряжении имеются лишь примеры активности, поэтому обученный классификатор может рассчитывать значения для истинно положитель-

*MCMF — метод непрерывных молекулярных полей (method of continuous molecular fields).

**LIBSVM — Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

***ROC — операционная характеристика приемника (receiver operating characteristic).

****http://ab-initio.mit.edu/nlopt.

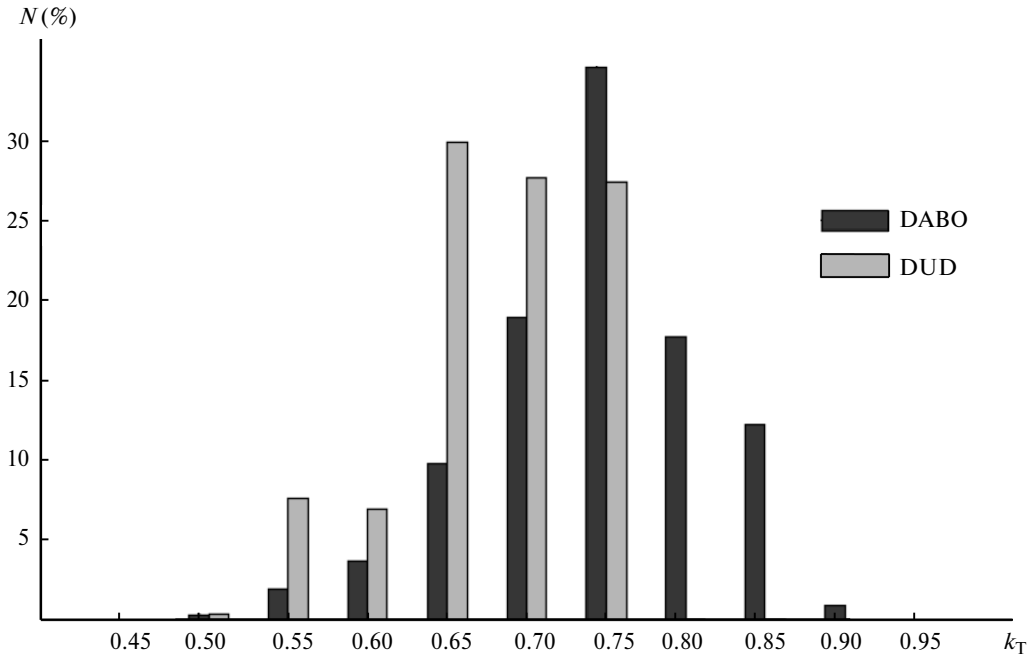

Рис. 1. Гистограмма молекулярного сходства активных лигандов исследуемой базы, а также лигандов тестового набора DUD для HIVRT с соединениями-приманками базы DUD (k_T — коэффициент Танимото, N(%) — количество неактивных молекул).

ных примеров (предсказанных положительно (TP)* или отрицательно (FN))**. Однако для оценки его эффективности, вычисляемой по площади под ROC-кривой, необходимо также рассчитать значения отрицательных примеров (предсказанных положительно (FP)*** или отрицательно (TN))*****), т.е. соединений заведомо не проявляющих активности. При этом качество модели характеризуют следующие параметры: чувствительность (TP/(TP + FN)) и специфичность (TN/(FP + TN)). В настоящей работе в качестве отрицательных примеров ингибиторной активности в отношении HIVRT использовали библиотеку соединений, так называемых «приманок» (decoys), из базы данных DUD***** (см. лит.[21]). Использованный набор «приманок» состоял из 1519 соединений, которые по физико-химическим свойствам похожи на известные лиганды HIVRT, но значительно отличаются от них по химической структуре. Следует отметить, что структуры «приманок» используются только для расчета эффективности получаемых одноклассовых моделей, в отличие от двухклассовой классификации, где они непосредственно участвуют в построении модели.

На рисунке 1 представлена гистограмма, характеризующая молекулярное сходство активных лигандов исследуемой базы, а также лигандов тестового набора базы DUD для HIVRT с используемыми соединениями-приманками базы DUD. По оси ординат указан процент неактивных лигандов, для которых минимальные значения молекулярного подобия с активными лигандами попадают в интервал изменения коэффициента Танимото (k_T)[22], отложенного по оси абсцисс. Из рисунка видно, что соединения класса DABO имеют большее молекулярное сходство с соединениями-приманками базы DUD, чем лиганды тестового набора DUD для HIVRT. Наибольшее количество соединений исследуемой базы имеет молекулярное сходство с приманками порядка ≤0.75, что характеризует ее как разнородную.

При построении моделей методом опорных векторов помимо выбора оптимальных параметров статистических ядер требуется также оптимизация параметра ν, ограничивающего число опорных векторов и участвующего в построении модели. А в случае представления молекулярной структуры на основе комбинированного молекулярного поля как линейной комбинации электростатического, стерического и гидрофобного молекулярных полей при построении модели необходимо оптимизировать значения семи параметров (коэффициенты смешения для трех типов полей, три коэффициента аттенуации и параметр ν). Эта оптимизация проводилась путем максимизации площади под ROC-кривой (AUC*). На рисунке 2 приведена схема построения одноклассовых моделей с использованием «приманок». Каждое активное соединение последовательно исключали из обучающей выборки и его активность предсказывали на основе модели, построенной на оставшихся структурах. Затем строили модель с использованием всех примеров активности и прогнозировали структуры «приманок». Результаты прогноза объединяли и рассчитывали пло-

*TP — количество активных соединений, спрогнозированных как активные (true positive).

**FN — количество активных соединений, спрогнозированных как неактивные (false negative).

***FP — количество неактивных соединений спрогнозированных как активные (false positive).

****TN — количество неактивных соединений, спрогнозированных как неактивные (true negative).

*****DUD — каталог «приманок» (directory of useful decoys).

*AUC — площадь под кривой (area under curve).

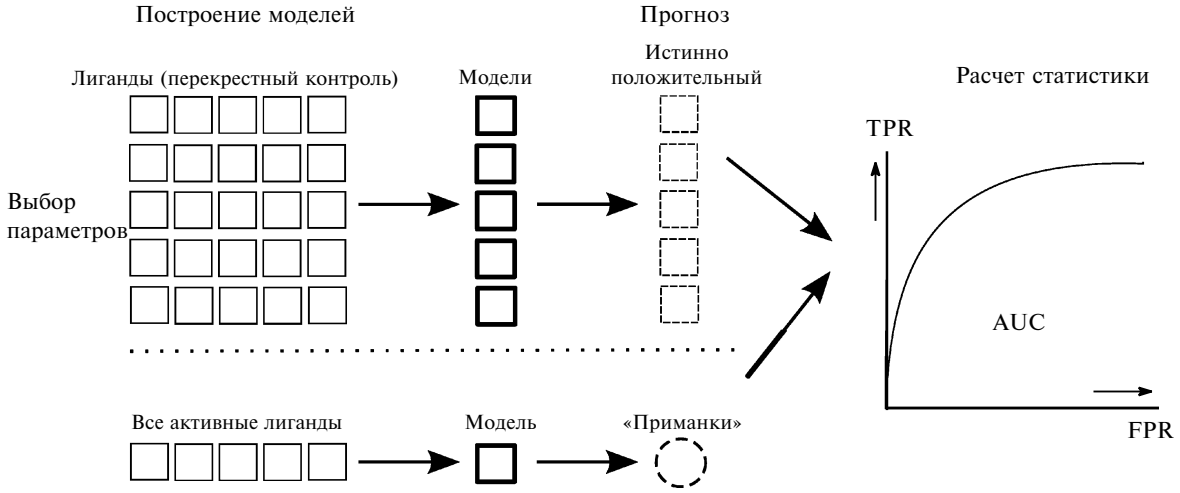

Рис. 2. Схема оптимизации параметров одноклассового классификатора с использованием пятикратного скользящего контроля и «приманок». TPR — доля положительных примеров, спрогнозированных как положительные (True positive rate), FPR — доля отрицательных примеров, спрогнозированных как положительные (False postivie rate).

щадь под ROC-кривой (AUC). Чем она выше, тем лучше классификатор справляется с поставленной задачей. Идеальный классификатор имеет AUC = 1.0.

С помощью приведенной схемы мы получили набор одноклассовых моделей, построеннных на основе различных подходов к представлению молекулярной структуры, как традиционных методов — модифицированный вариант фрагментных дескрипторов Кархарта, «молекулярные отпечатки», спектрофоры, так и в рамках предложенной нами ранее методологии непрерывных молекулярных полей[16]. Традиционное описание структуры в рамках общепринятых методов SAR/QSAR* подразумевает представление ее в качестве набора дескрипторов — чисел, которые описывают какие-либо характеристики рассматриваемого соединения[23]. В настоящее время для виртуального скрининга широко используются фрагментные (подструктурные) дескрипторы[24]. Однако такое описание имеет ряд существенных недостатков, среди которых главным является невозможность предложить активные соединения из других структурных типов, отличных от тех, которые были использованы для построения модели. В связи с этим активно развиваются методы «бездескрипторного» описания структур, т.е. расчет моделируемых свойств непосредственно из математического описания структуры органического соединения. Например, структура представляется в виде молекулярного графа, и на ее основе строятся различные статистические молекулярные ядра[25]. Нами был предложен альтернативный способ построения статистических ядер на основе использования непрерывных молекулярных полей[16]. Его успешное применение продемонстрировано при построении регрессионных QSAR-моделей для прогнозирования биологической активности, а также одноклассовых моделей для виртуального скрининга. Так, использование этого метода при построении моделей для стандартного набора лигандов базы DUD для фермента HIVRT позволяет получить модели более высокого качества, чем простой поиск по подобию[26].

Предлагаемый нами подход одноклассовой классификации, в котором используется при построении прогностической модели только образцы активных лигандов, нельзя сравнивать с двухклассовыми моделями, применяемыми в основном для целей виртуального скрининга. В литературе представлена одна работа[27], в которой исследовали базу DUD с помощью стандартных процедур поиска по подобию, используя различные оценочные функции для учета 2D- и 3D-информации. Описанный поиск по подобию учитывает только один активный лиганд, в данном случае кристаллографический, и ранжирует относительно него все структуры из тестируемой базы. По своей сути эту процедуру можно рассматривать как упрощенный вариант одноклассовой классификации.

Статистические характеристики полученных одноклассовых моделей лигандов класса DABO приведены в таблице 1, а соответствующие ROC-кривые — на рисунке 3. Как видно из таблицы, модели, построенные с использованием фрагментных дескрипторов (модифицированных дескрипторов Кархарта и на основе «молекулярных отпечатков»), характеризуются высокими значениями AUC. Так, для модели на основе «молекулярных отпечатков» эта величина составляет 0.97 (статистическое ядро — функция Танимото), а на основе дескрипторов Кархарта — 0.99 (ядро Гаусса), что практически соответствует идеальному классификатору. В качестве иллюстрации в таблице 1 приведены параметры одноклассовых моделей, при построении которых мы использовали стандартный набор лигандов HIVRT и соединения-приманки из базы DUD. Большие различия в количестве соединений двух выборок (число лигандов стандартного набора DUD составляет 40 соединений, а исследуемой базы — 786),

*SAR/QSAR — structure activity relationships/quantitative structure activity relationships.

а также в их разнообразии не позволяют корректно сравнивать результаты, полученные с использованием этих баз данных. Тем не менее сравнение ROC-кривых (см. рис. 3), построенных по результатам прогноза с использованием стандартного набора лигандов DUD для HIVRT на основе модели 1-SVM на базе дескрипторов Кархарта, а также модели 1-SVM в сочетании со стерическим молекулярным полем в качестве описания молекулярных структур, с данными работы[27] показывает, что предлагаемый подход одноклассовой классификации эффективнее простого поиска по подобию на основе индекса Танимото, а в случае стерического поля наблюдаются примерно равные AUC классификаторов. Так, модель на основе спектрофоров (кривая *1*) характеризуется наибольшей площадью под ROC-кривой. Хотя использование стерического молекулярного поля для описания структур лигандов (кривая *2*) сравнимо по эффективности со стандартной процедурой поиска по подобию (кривая *3*), однако на начальной стадии, которая представляет интерес при практическом проведении виртуального скрининга, первая из этих кривых находится существенно выше. Это свидетельствует о большей эффективности применения модели 1-SVM в сочетании со стерическим молекулярном полем по сравнению со стандартной процедурой поиска по подобию на наиболее важных начальных стадиях виртуального скрининга. При этом предлагаемый метод одноклассовой классификации в обоих случаях характеризуется большей площадью под ROC-кривой.

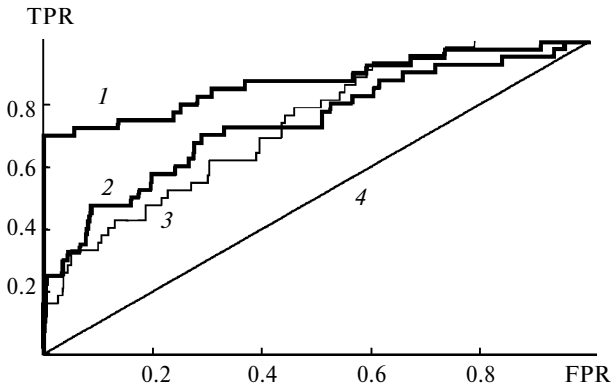

Рис. 3. ROC-кривые одноклассовых моделей для стандартного набора лигандов и приманок базы DUD для HIVRT: *1* — модель 1-SVM, ядро Гаусса, дескрипторы — спектрофоры; *2* — модель 1-SVM, стерическое молекулярное поле в качестве описания структур; *3* — стандартный поиск по подобию на основе молекулярных отпечатков[27], *4* — случайный классификатор.

Как отмечалось выше, применение подструктурного описания ограничивает набор структур, получаемых в результате виртуального скрининга, классами соединений, использованных для построения модели. Теоретически этого недостатка можно избежать, используя такие дескрипторы, как спектрофоры, которые вычисляют значения поверхностных полей в точках искусственно сконструированной вокруг лиганда ячейки. Однако в данном исследовании модель на основе спектрофоров характеризуется довольно низким значением AUC = 0.76 (ядро Гаусса).

Модели, полученные на основе непрерывных молекулярных полей, имеют более высокие статистические характеристики по сравнению с моделями, построенными с помощью модифицированных дескрипторов Кархарта (AUC = 0.99), «молекулярных отпечатков» (AUC = 0.97) и спектрофоров (AUC = 0.76). Модели, построенные с использованием гидрофобного (AUC = 1.00) и стерического полей (AUC = 1.00), обладают максимальными значениями площадей под ROC-кривыми. Высокие статистические показатели этих моделей позволяют предложить их в качестве инструмента для виртуального скрининга потенциальных ингибиторов HIVRT.

Таблица 1. Статистические характеристики одноклассовых моделей, полученных с использованием различных способов представления молекулярной структуры соединений для лигандов класса DABO и стандартного набора базы DUD для HIVRT

Способ	Лиганды DABO							Тестовый набор лигандов базы DUD для HIVRT						
	AUC	TN	TP	FN	FP	Чувствительность	Специфичность	AUC	TN	TP	FN	FP	Чувствительность	Специфичность
Спектрофоры	0.71	993	514	272	526	65.4	65.4	0.87	1100	31	9	345	76.1	77.5
«Молекулярные отпечатки»	0.98	1477	765	21	42	97.2	97.3	0.83	1157	32	8	288	80.1	80.8
Дескрипторы Кархарта	0.99	1489	770	16	30	98.0	98.0	0.80	1511	24	16	8	99.5	62.8
Электростатическое поле	0.99	1505	779	7	14	99.1	99.1	0.60	831	23	17	614	57.5	57.5
Гидрофобное поле	1.00	1519	786	0	0	100.0	100.0	0.75	1012	28	12	433	70.0	70.0
Стерическое поле	1.00	1519	786	0	0	100.0	100.0	0.65	851	24	16	594	58.9	60.0

На рисунке 4 изображены поля коэффициентов (направляющих косинусов перпендикуляра к разделяющей гиперплоскости в характеристическом пространстве) моделей 1-SVM*, построенных для ингибиторов HIVRT с использованием электростатического и стерического полей. Найденные поля коэффициентов показывают конфигурацию молекулярных полей, которыми должна обладать молекула органического соединения для того, чтобы быть ингибитором HIVRT.

Стандартная процедура проведения виртуального скрининга на основе поиска по молекулярному подобию заключается в выборе некоторого активного соединения в качестве реперной структуры и ранжирования тестируемой базы структур относительно него на основе индекса Танимото. Формально такой метод можно рассматривать как вариант одноклассового классификатора. Поэтому мы сравнили эффективность стандартного поиска по молекулярному подобию с рассматриваемым в работе одноклассовым подходом. Для этого последовательно выбирали каждую структуру из множества активных лигандов исследуемой базы и на ее основе рассчитывали величину молекулярного подобия для всех остальных лигандов и «приманок». В результате был получен набор ROC-кривых, площади под которыми оценивают эффективность проведения виртуального скрининга при помощи поиска по подобию с соответствующей реперной структурой. Примеры подобных ROC-кривых представлены на рисунке 5 (кривые *3*, *6* и *7*). Наименьшей площадью под ROC-кривой характеризуется кривая *7* (AUC = 0.31). Средняя площадь под всеми построенными таким образом ROC-кривыми, тем не менее, составляет 0.97. Таким образом, предлагаемый нами подход приводит к несколько более высоким статистическим показателям классификатора (0.99 и 1.00, в зависимости от типа молекулярного поля). Кроме того, в отличие от метода поиска по подобию, предлагаемый нами подход позволяет: во-первых, учитывать информацию о множестве активных структур, во-вторых, настраивать меру подобия для получения максимальной эффективности процедуры виртуального скрининга.

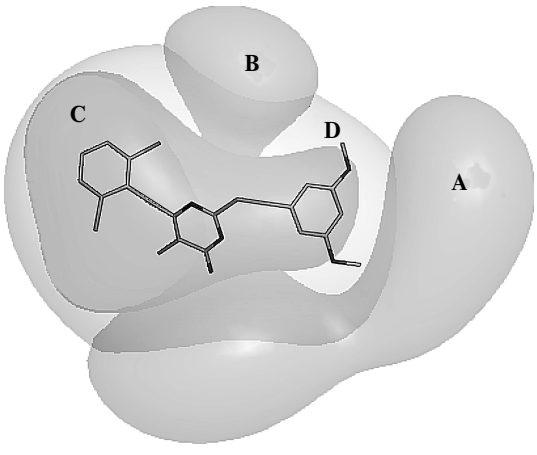

Рис. 4. Поверхности уровней полей коэффициентов 1-SVM моделей, построенных на основе электростатического и стерического полей для ингибиторов HIVRT: A и B — области отрицательного электростатического потенциала, C — область положительного электростатического потенциала, D — область стерического потенциала, определяющая форму молекулы.

*SVM — метод опорных векторов (Support vector machine).

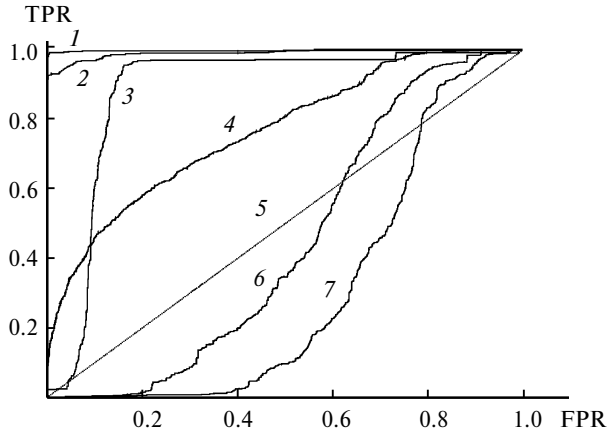

Рис. 5. ROC-кривые одноклассовых моделей, построенных с использованием 1-SVM и электростатического поля, а также «молекулярных отпечатков» и дескрипторов Кархарта (*1*); 1-SVM и стерического/гидрофобного полей (*2*); традиционного поиска по подобию на основе индекса Танимото с использованием разных реперных структур (*3*, *6* и *7*); 1-SVM и спектрофоров (*4*); случайного классификатора (*5*). FPR (False Positive Rate) = 1 — [TN(FP + TN)], TPR (True Positive Rate) = TP/(TP + FN), где TN(FP + TN) — специфичность, TP/(TP + FN) — чувствительность.

Таким образом, в настоящей работе впервые в рамках одноклассового подхода с использованием метода опорных векторов построены модели для виртуального скрининга потенциальных ингбиторов обратной транскриптазы ВИЧ-1. Показано, что представление молекулярной структуры органических лигандов на основе непрерывных молекулярных полей позволяет получать классификационные модели более высокого качества по сравнению с подходами, базирующимися на использовании «молекулярных отпечатков», спектрофоров и фрагментных дескрипторов Кархарта. Наилучшие модели, построенные на основе непрерывных молекулярных полей, имеют параметры, близкие к идеальному классификатору, эти модели можно рекомендовать для проведения широкомасштабного виртуального скрининга.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 10-07-00201), а также Совета по грантам при Президенте Российской Федерации (программа государственной поддержки молодых российских ученых—кандидатов наук, грант МК-1351.2011.3) и Министерства образования и науки РФ (Федеральная целевая программа «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России

на 2007—2012 годы» 2011-1.2-512-055, заявка «2011-1.2-512-055-011»).

Список литературы

1. G. Barbaro, A. Scozzafava, A. Mastrolorenzo, C. T. Supuran, *Curr. Pharm. Design*, 2005, **11**, 1805.
2. E. De Clercq, *J. Med. Chem.*, 2005, **48**, 1.
3. M. B. Nawrozkij, D. Rotili, D. Tarantino, G. Botta, A. S. Eremiychuk, I. Musmuca, R. Ragno, A. Samuele, S. Zanoli, M. Armand-Ugón, I. Clotet-Codina, I. A. Novakov, B. S. Orlinson, G. Maga, J. A. Esté, M. Artico, A. Mai, *J. Med. Chem.*, 2008, **51**, 4641.
4. R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K. Debnath, C. Hansch, *Chem. Rev.*, 1999, **99**, 3525.
5. G. Gini, M. V. Craciun, C. König, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1897.
6. L. Bruno-Blanch, J. Galvez, R. García-Domenech, *Bioorg. & Med. Chem. Lett.*, 2003, **13**, 2749.
7. S. Rodgers, *J. Chem. Inf. Model.*, 2006, **46**, 569.
8. B. Su, M. Shen, E. X. Esposito, A. J. Hopfinger, Y. J. Tseng, *J. Chem. Inf. Model.*, 2010, **50**, 1304.
9. П. В. Карпов, И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Докл. АН*, 2011, **437**, 642 [*Dokl. Chem.* (*Engl. Transl.*), 2011, **437**, 642].
10. I. I. Baskin, N. Kireeva, A. Varnek, *Mol. Inf.*, 2010, **29**, 581.
11. N. Fechner, A. Jahn, G. Hinselmann, A. Zell, *J. Cheminform.*, 2010, **2**, 2.
12. P. Bultinck, W. Langenaeker, P. Lahorte, F. De Proft, P. Geerlings, C. Van Alsenoy, J. P. Tollenaere, *J. Phys. Chem. A*, 2002, **106**, 7895.
13. P. Bultinck, W. Langenaeker, R. Carbó-Dorca, J. P. Tollenaere, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 422.
14. R. E. Carhart, D. H. Smith, R. Ventkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64.
15. П. М. Васильев, А. А. Спасов, *Рос. хим. журн.*, 2006, № 2, 108 [*Mendeleev Chem. J.* (*Engl. Transl.*), 2006, No. 2 (in Russian)].
16. Н. И. Жохова, И. И. Баскин, Д. К. Бахронов, В. А. Палюлин, Н. С. Зефиров, *Докл. АН*, 2009, **429**, 201 [*Dokl. Chem.* (*Engl. Transl.*), 2009, **429**, 273].
17. R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, E. L. Willighagen, *J. Chem. Inf. Model.*, 2006, **46**, 991.
18. Н. В. Артеменко, И. И. Баскин, В. А. Палюлин, Н. С. Зефиров, *Докл. АН*, 2001, **381**, 203 [*Dokl. Chem.* (*Engl. Transl.*), 2001, **381**, 317].
19. T. Fawcett, *Pattern Recognition Lett.*, 2006, **27**, 861.
20. S. K. Kearsley, G. M. Smith, *Tetrahedron Computer Methodology*, 1990, **3**, 615.
21. N. Huang, K. Shoichet, J. J. Irwing, *J. Med. Chem.*, 2006, **49**, 6789.
22. Pang-Ning Tan, M. Steinback, V. Kumar, *Introduction to Data Mining*, Publisher: Addision-Wesley, 2006, 769 pp.
23. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000, 667.
24. I. Baskin, A. Varnek, in *Chemoinformatics Approaches to Virtual Screening*, RCS Publishing, 2008, **338**, 1.
25. P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert, *J. Chem. Inf. Model.*, 2005, **45**, 939.
26. П. В. Карпов, И. И. Баскин, Н. И. Жохова, Н. С. Зефиров, *Докл. АН*, 2011, **440**, 480—483 [*Dokl. Chem.* (*Engl. Transl.*), 2011].
27. V. Venkatraman, V. I. Perez-Nueno, L. Mavridis, D. W. Ritchie, *J. Chem. Inf. Model.*, 2010, **50**, 2079.

Поступила в редакцию 10 июня 2011;
после доработки — 1 сентября 2011