На правах рукописи

Захаров Алексей Владимирович

ПРОГНОЗ КОЛИЧЕСТВЕННЫХ СВОЙСТВ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ НА ОСНОВЕ ДЕСКРИПТОРОВ АТОМНЫХ ОКРЕСТНОСТЕЙ

03.00.28 – биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2008 г.

Работа выполнена в Государственном учреждении Научно-исследовательском институте биомедицинской химии им. В.Н. Ореховича Российской академии медицинских наук

Научные руководители: кандидат биологических наук
Лагунин Алексей Александрович
кандидат физико-математических наук
Филимонов Дмитрий Алексеевич

Официальные оппоненты: доктор химических наук, профессор
Раевский Олег Алексеевич
доктор биологических наук
Веселовский Александр Владимирович

Ведущая организация: Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН

Защита состоится 11 декабря 2008 года в 11:00 часов на заседании Диссертационного совета Д 001.010.01 при Государственном учреждении Научно-исследовательском институте биомедицинской химии имени В.Н. Ореховича Российской академии медицинских наук по адресу: 119121, г. Москва, Погодинская ул., д.10

С диссертацией можно ознакомиться в библиотеке Государственного учреждения Научно-исследовательского института биомедицинской химии имени В.Н. Ореховича Российской академии медицинских наук.

Автореферат разослан «___» ноября 2008 г.

Ученый секретарь Диссертационного совета,
кандидат химических наук Е.А. Карпова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Методы количественного анализа взаимосвязей структура-активность (QSAR) широко применяются для поиска и конструирования лекарств, а также для оценки безопасности химических веществ. В основе QSAR моделирования лежит предположение, что свойство химического соединения определяется его структурой. Для описания структуры химического соединения используют так называемые дескрипторы – разнообразные характеристики молекул вещества. Известно более 3000 дескрипторов, которые применяются для построения QSAR моделей (Todeschini R. et al., 2000; Olah M. et al., 2004). Одной из проблем, активно исследуемых в данной области, является поиск оптимального набора дескрипторов, которые смогли бы описать взаимосвязь структура-активность для разнообразных видов биологической активности органических соединений (Nicholls A. et al., 2004; Todeschini R. et al., 2000). Использование разных дескрипторов, приводящих к различным моделям для одних и тех же веществ, порождает проблему выбора лучшей модели, что часто приводит к ее переобучению (Tetko I. et al., 1995; Johnson S., 2008).

Наряду с методами QSAR, основанными на структурных формулах химических соединений (2D QSAR), успешно применяются для создания новых биологически активных молекул методы, использующие пространственное описание химических структур (3D QSAR) (Li A. et al., 1999; Korhonen L. et al., 2005; Korhonen L. et al., 2007). Для этих методов необходимы данные о пространственной структуре белков и/или лигандов, и их отличительной особенностью является то, что они учитывают стереоспецифичность лиганд-белкового взаимодействия. В то же время, для построения предсказательных моделей необходимо пространственное выравнивание молекул, которое часто бывает неоднозначным, а если сами лиганды гибкие, то необходимо проводить поиск наилучших конформаций молекул, что приводит к увеличению сложности вычислений, и результат также может быть неоднозначным (Kubinyi H., 1993; Raha K. et al., 2007).

Одним из важнейших аспектов QSAR является алгоритм построения модели. Изначально в QSAR доминирующим методом была обычная множественная линейная регрессия. С ростом количества дескрипторов и с появлением проблемы их выбора стали использовать другие методы: проекция на скрытые переменные, искусственные нейронные сети, метод опорных векторов и пр. (Livingstone D. et al., 1995). Эти методы имеют как преимущества, так и недостатки, например, для искусственных нейронных сетей хорошо известна проблема оптимальной остановки обучения, для метода опорных векторов необходим поиск оптимальных параметров (Hughes L. et al., 2008), для других методов существует проблема выбора наилучшей модели (Golbraikh A. et al., 2002).

Наряду с QSAR-моделированием отдельных свойств химических соединений, все актуальней становится проблема одновременной оценки множества разнообразных свойств для больших массивов химических структур: проблема оценки ADME/T (абсорбция, распределение, метаболизм, выведение и токсичность) (Beresford A. et al., 2002; Tingjun H. et al., 2007), компьютерное предсказание действия лигандов на множество мишеней одновременно (Rizzo S. et al., 2008), предсказание побочных эффектов лекарств (Kamenska V. et al., 2006; Shamovsky I. et al., 2008). В то время, как качественное предсказание спектра биологической активности уже известно и широко применяется (Poroikov V. et al., 2000), множественное количественное предсказание биологических активностей, фармакологических свойств и токсичности вывело бы на новый уровень количественный анализ взаимосвязей структура-активность органических соединений. Эффективная реализация множественного прогноза требует использования универсальных дескрипторов единого типа, примером которых могут служить дескрипторы многоуровневых атомных окрестностей (Filimonov D. et al., 1999; Филимонов и др., 2006).

Цель работы: разработка и валидация метода количественного прогноза биологической активности органических соединений на основе дескрипторов атомных окрестностей.

Задачи исследования

1. Исследовать применимость дескрипторов атомных окрестностей для описания органических соединений в задачах количественного анализа взаимосвязей структура-активность.
2. Разработать эффективный алгоритм количественного прогноза биологической активности органических соединений на основе дескрипторов атомных окрестностей и реализовать его в виде компьютерной программы.
3. Сравнить точность и прогностическую способность предложенного алгоритма с другими методами количественного анализа взаимосвязей структура-активность для разных видов биологической активности.

Положения, выносимые на защиту

Количественный анализ взаимосвязей структура-активность на основе оценки среднего вклада атомов молекулы в ее активность, как функция количественных дескрипторов атомных окрестностей, и самосогласованной регрессии, обладает более высокой прогностической способностью по сравнению с другими методами QSAR.

Научная новизна. Впервые исследована возможность применения дескрипторов атомных окрестностей для количественного анализа взаимосвязей структура-активность. Разработано три новых метода построения QSAR моделей. Показана эффективность применения количественных дескрипторов атомных окрестностей для QSAR анализа лиганд-ферментного, лиганд-рецепторного взаимодействий и прогноза острой токсичности. Также показана эффективность применения многоуровневых дескрипторов атомных окрестностей для QSAR анализа лиганд-рецепторных взаимодействий.

Продемонстрирована возможность использования разработанного нами оригинального метода для одновременного прогноза количественных значений разных типов биологической активности для одной молекулы.

Предложен новый метод оценки применимости модели, имеющий существенное преимущество по сравнению с наиболее распространенными методами.

Научно-практическая значимость. Исследование возможности применения дескрипторов атомных окрестностей для количественного анализа взаимосвязей структура-активность позволило выявить области применимости каждого типа дескрипторов.

Разработанные методы позволяют более точно предсказывать количественные величины биологической активности, что дает возможность существенно снизить временные и финансовые затраты при поиске базовых структур новых лекарств.

В разработанной нами программе количественного прогноза биологической активности реализована возможность предсказания количественных значений разных видов биологической активности одновременно по многим моделям, что позволяет выявлять соединения, обладающие множественными механизмами действия, оценивать ADME свойства химических соединений, а также побочные и токсические эффекты.

Новый метод оценки области применимости модели повышает эффективность отбора химических соединений при виртуальном скрининге.

Работа выполнена при поддержке государственного контракта Роснауки № 02.434.11.1014, гранта РФФИ №06-03-39015, гранта INTAS № 03-51-5218, гранта шестой Европейской рамочной программы № LSHB-CN-2007-037590, гранта МНТЦ № 3777.

Апробация работы. Основные положения диссертации были доложены на следующих симпозиумах и конференциях: “The 15th European Symposium on Quantitative Structure-Activity Relationships & Molecular Modelling, Turkey, 2004”, “XII Российский национальный конгресс "Человек и лекарство", Москва, 2005”, “The 12th International Workshop on Quantitative Structure – Activity Relationships in Environmental Toxicology, France, 2006”, “3 Международная конференция «Геномика, Протеомика, Биоинформатика и Нанотехнологии для Медицины», Новосибирск, 2006”, “4th Eurasian Meeting on Heterocyclic Chemistry, Greece,

2006", "Московская Международная конференции «Биотехнология и Медицина», Москва, 2006", "XIII Российский национальный конгресс «Человек и Лекарство», Москва, 2006", "XIV Российский национальный конгресс «Человек и лекарство», Москва, 2007", "4th International Symposium «Computational Methods in Toxicology and Pharmacology Integrating Internet Resources», Moscow, 2007", "3rd German Conference on Chemoinformatics, Goslar, Germany, 2007", "Helmholtz-Russian-German Workshop on Systems Biology, Moscow, 2008", "XV Российский национальный конгресс «Человек и лекарство», Москва, 2008", "8th International Conference on Chemical Structures, Noordwijkerhout, the Netherlandes, 2008", "The 17th European Symposium on Quantitative Structure-Activity Relationships & Omics Technologies and Systems Biology, Uppsala, Sweden, 2008".

Публикации. По материалам диссертации опубликовано 18 научных работ в отечественных и зарубежных научных изданиях, в том числе 4 статьи в рецензируемых научных журналах, 2 статьи в трудах конференций и 12 публикаций в сборниках научных конференций. Получено свидетельство Роспатента о регистрации компьютерной программы.

Объем и структура диссертации. Диссертация состоит из введения, 4 глав: литературный обзор, объекты и методы исследования, результаты и обсуждение, заключение, – выводов, и списка литературы, включающего 152 публикации. Диссертационная работа изложена на 120 страницах машинописного текста и содержит 22 рисунка и 23 таблицы.

ОБЪЕКТЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Для оценки и сравнения методов QSAR-моделирования, согласно критериям:

1) разнообразие типов биологической активности;

2) широкий диапазон значений измеренной биологической активности;

3) разнообразие химических структур;

4) наличие в литературе QSAR моделей для данных выборок, построенных с помощью различных методов;

в данной диссертационной работе было отобрано десять выборок низкомолекулярных органических соединений, для которых уже были построены модели с использованием различных методов QSAR, таких, как CoMFA, CoMSIA, HQSAR, EVA, GRID/GOLPE и другие. Эти выборки представляют следующие типы биологической активности:

1) лиганд-ферментные взаимодействия:

- Ингибиторы циклин-зависимой киназы 2 (обучающая выборка $CDK2_{обуч}$ – 29 структур и тестовая $CDK2_{тест}$ – 7 структур). Эти соединения являются производными бисарилмалеимида. Экспериментальные данные, представленные значениями IC_{50}

(50% ингибирующая концентрация, М), были переведены в логарифмы $pIC_{50} = - \log_{10}IC_{50}$, диапазон pIC_{50} от 5,057 до 8,194.

- Ингибиторы дигидрофолат редуктазы (выборки $DHFR_{обуч}$ – 237 структур и $DHFR_{тест}$ – 124 структуры), были представлены 11 химическими классами (пиримидины, хиназолины и др.). Экспериментальные данные, представленные значениями IC_{50} (М), были переведены в логарифмы, pIC_{50} в диапазоне от 3,3 до 9,8.
- Ингибиторы ангиотензин-превращающего фермента (выборки $ACE_{обуч}$ – 76 структур и $ACE_{тест}$ – 38 структур). Эти соединения являются производными карбоновых кислот, содержащих в своей структуре различные гетероциклы (пирролидин, индол, имидазол и др.). Экспериментальные данные, представленные значениями IC_{50} (М), были переведены в логарифмы, pIC_{50} в диапазоне от 2,1 до 9,9.
- Ингибиторы изоформы 2A5 цитохрома P450 (выборки $CYP2A5_{обуч}$ – 23 структуры и $CYP2A5_{тест}$ – 5 структур). Экспериментальные данные, представленные значениями IC_{50} (М), были переведены в логарифмы, pIC_{50} в диапазоне от 1,73 до 5,68.
- Ингибиторы изоформы 2A6 цитохрома P450 (выборки $CYP2A6_{обуч}$ – 23 структуры и $CYP2A6_{тест}$ – 5 структур). Экспериментальные данные, представленные значениями IC_{50} (М), были переведены в логарифмы, pIC_{50} в диапазоне от 0,46 до 4,52.

2) лиганд-рецепторные взаимодействия:

- Соединения, действующие на альфа-2 адренорецепторы (выборка $ADREN_{обуч}$ – 30 структур). Экспериментальные данные, представленные значениями K_i (мкМ), были переведены в логарифмы – $\log_{10}(K_i)$, их диапазон от 3,33 до 6,66.
- Соединения, действующие на эстрогеновые рецепторы (выборка $ESTR_{обуч}$ – 21 структура). Эти соединения являются производными тетрагидроизохинолина. Экспериментальные данные, представленные значениями IC_{50} (мкМ), были переведены в логарифмы, pIC_{50} в диапазоне от –0,567 до 0,983.

3) острая токсичность:

- Соединения, проявляющие острую токсичность для *Chlorella vulgaris* (выборка $ALGAE_{обуч}$ – 65 структур). В выборке представлены производные фенолов, анилинов, бензальдегидов, нитробензолов и др. Экспериментальные данные, представленные значениями EC_{50} (мМ), были переведены в логарифмы $\log_{10}(1/EC_{50})$, их диапазон от -1,46 до 3,10.
- Соединения, проявляющие острую токсичность для *Vibrio fischeri* (выборка $VIBRIO_{обуч}$ – 56 структур). Эти соединения являются фенилсульфонилкарбоксилатами. Экспериментальные данные, представленные

значениями EC_{50} (мкМ), были переведены в логарифмы – $logEC_{50}$, их диапазон от -0,44 до 2,28.

- Соединения, проявляющие острую токсичность для *Tetrahymena pyriformis* (выборка $TETRA_{обуч}$ – 200 структур). Эти соединения являются алкильными производными фенола, галоген-замещенными фенолами и др. Экспериментальные данные, представленные значениями IGC_{50} (мМ), были переведены в логарифмы – $\log_{10}(1/IGC_{50})$, их диапазон от -1,50 до 2,71.

Необходимо отметить, что 3 из 10 выборок являются гетерогенными по химической структуре (ингибиторы дигидрофолат редуктазы, ингибиторы ангиотензин-превращающего фермента, соединения, проявляющие острую токсичность к *Tetrahymena pyriformis*), что важно для проверки прогностической способности модели. Количество структур в обучающих выборках варьирует от 23 до 237 соединений, а в тестовых выборках от 5 до 124 соединений.

Для описания структур химических соединений в данной работе использовались дескрипторы многоуровневых атомных окрестностей (MNA - Multilevel Neighborhoods of Atoms) (Filimonov D. et al., 1999) и количественных атомных окрестностей (QNA - Quantitative Neighborhoods of Atoms) (Filimonov D. et al., 2004).

Дескрипторы атомных окрестностей основаны на представлении структурной формулы молекул вещества, в которой, согласно валентностям и зарядам атомов, явно указаны все водороды, и не учитываются типы связей. Для расчета QNA дескрипторов используются значения потенциала ионизации (*IP* - ionization potential) и сродства к электрону (*EA* - electron affinity) каждого атома молекулы (значения приведены в диссертации, раздел 2.2.1.2.). QNA вычисляются по следующим формулам:

$$P_i = B_i \sum_k \left(\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)\right)_{ik} B_k\,, \quad Q_i = B_i \sum_k \left(\mathrm{Exp}\left(-\tfrac{1}{2}\mathbf{C}\right)\right)_{ik} B_k A_k$$

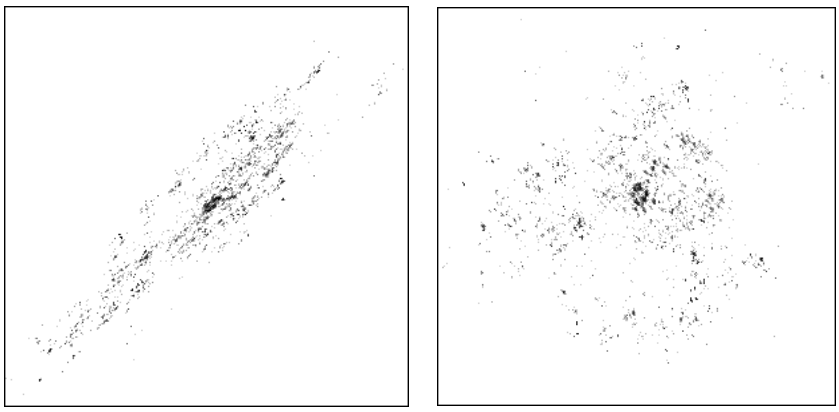
где i, k – номера атомов в молекуле, $A_k = \frac{1}{2}(IP_k + EA_k)$, $B_k = (IP_k - EA_k)^{-\frac{1}{2}}$, $\mathbf{C}$ - матрица смежности молекулы. QNA дескрипторы описывают каждый атом молекулы, и при этом каждое значение *P* или *Q* зависит от строения молекулы в целом.

Для химических структур, взятых из десяти выборок, используемых в данной диссертационной работе, было рассчитано 16617 QNA дескрипторов для всех атомов, имеющих два или более непосредственных соседей. На рисунке 1а они отображены как точки в QNA пространстве. Рисунок 1а показывает, что значения *P* и *Q* сильно коррелируют, поэтому они были дополнительно нормализованы.

Нормализация QNA дескрипторов, после расчета среднего (E_P и E_Q), стандартного отклонения (D_P и D_Q) и корреляции (R_{PQ}) значений P и Q, осуществлялась по следующим формулам:

$$P' = \frac{P - E_P}{D_P}, \qquad Q' = \frac{Q - E_Q}{D_Q}, \qquad U = \frac{P' + Q'}{\sqrt{2(1 + R_{PQ})}}, \qquad V = \frac{P' - Q'}{\sqrt{2(1 - R_{PQ})}}$$

Ортонормальные U и V имеют нулевое среднее, единичную вариацию и они не коррелированны между собой (Рис. 1б).



а) б)

Рис. 1. Распределение 16617 значений P и Q на матрице 300x300 пикселов: а) начальные значения P (ось абсцисс) и Q (ось ординат) в границах (-0,0579, 0,0784) для P и (-0,581, 0,666) для Q; б) нормализованные значения QNA в границах (-3, 3) для U и V.

Дескрипторы атомных окрестностей, как MNA, так и QNA, не отражают информацию об объеме и форме молекулы, хотя она может быть важна для анализа взаимосвязей структура-активность. Поэтому в данной работе дополнительно использовались два дескриптора: топологические длина и объем молекулы. Топологическая длина рассчитывалась как максимальное расстояние между двумя атомами молекулы. Объем молекулы рассчитывался как сумма объемов атомов молекулы, которые вычислялись как $\frac{4}{3}\pi R^3$, где R – это радиус атома. Значения атомных радиусов приведены в диссертации (раздел 2.2.2.).

Количество дескрипторов атомных окрестностей (MNA и QNA), рассчитанных для определенной молекулы, зависит от числа атомов в ней. Поскольку при регрессионном анализе необходимо одинаковое количество переменных для всех анализируемых структур, нами были разработаны специальные алгоритмы преобразования дескрипторов атомных окрестностей для построения QSAR моделей. Для MNA дескрипторов был использован

метод нечетких градаций (детальное описание представлено в диссертации – раздел 2.2.3.1), а для QNA – преобразования с помощью квантилей (детальное описание представлено в диссертации – раздел 2.2.3.2) и полиномов Чебышева.

Метод с полиномами Чебышева – конкретный частный случай преобразования, при котором для каждого атома молекулы вычисляются значения $g_i(P,Q)$ функций величин P и Q для этого атома, и структура молекулы описывается дескрипторами, значения которых равны среднему по атомам молекулы значению $g_i(P,Q)$:

$$f_i = \frac{1}{m}\sum_k g_i\left(P_k, Q_k\right).$$

где m – количество атомов, по которым выполняется усреднение, k – номер атома молекулы, P_k и Q_k – QNA дескрипторы для атома k.

В диссертационной работе в качестве семейства функций $g_i(P,Q)$ использованы полиномы Чебышева, средние значения которых вычислялись по атомам молекулы, имеющих два или более соседей. Ортонормальные значения U и V были дополнительно преобразованы с помощью гиперболического тангенса, так что нормированные QNA варьируют от -1 до 1, и значения полиномов Чебышева рассчитываются по формуле:

$$g_i\left(P,Q\right) = T_{uv}\left(P,Q\right) = \mathrm{Cos}\left(u * \mathrm{ArcCos}\left(\mathrm{TanH}\left(U\right)\right)\right) * \mathrm{Cos}\left(v * \mathrm{ArcCos}\left(\mathrm{TanH}\left(V\right)\right)\right),$$

где целые числа u, v = 0, 1, 2, ... определяют степень двумерного полинома Чебышева. Для большей эффективности и отсутствия эффекта переобучения число полиномов Чебышева на каждую выборку бралось меньше, чем количество соединений в обучающей выборке.

Для QSAR анализа, помимо дескрипторов, необходим математический метод построения зависимостей структура-активность. В диссертационной работе мы использовали метод самосогласованной регрессии, так как он имеет ряд преимуществ по сравнению с другими подходами (Филимонов Д. и др., 2004).

Нами были разработаны следующие новые методы QSAR анализа на основе дескрипторов атомных окрестностей и самосогласованной регрессии:

1) Метод 1 (MNA_SCR) использует MNA дескрипторы и преобразование с помощью нечетких градаций.

2) Метод 2 (qQNA_SCR) использует QNA дескрипторы и преобразование с помощью квантилей.

3) Метод 3 (QNA_Cheb_SCR) использует QNA дескрипторы и преобразование с помощью полиномов Чебышева.

Эти методы были использованы при построении QSAR моделей для десяти отобранных выборок.

Для оценки области применимости модели был разработан и использован метод, который основан на оценке расстояния от прогнозируемого соединения до его ближайшего соседа в обучающей выборке. Эта оценка рассчитывается как:

$$h_i = (x - x_i)^T (X^T X)^{-1} (x - x_i), \; (i = 1,2 \ldots n)$$

где x – вектор дескрипторов предсказываемого соединения, x_i – вектор дескрипторов i-го соединения из обучающей выборки, X является центрированной матрицей $n*k$, k – количество дескрипторов, используемых в модели, n – количество соединений в обучающей выборке. При этом для прогнозируемого соединения отбирается минимальное значение оценки по всем соединениям обучающей выборки.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Оценку точности и прогностической способности предложенных в диссертационной работе методов и их сравнение с другими методами QSAR выполняли на основе следующих критериев:

R^2 – квадрат коэффициента корреляции ($R^2 = 1 - \sum(y_{набл} - y_{расч})^2 / \sum(y_{набл} - y_{сред})^2$).

Q^2 – квадрат коэффициента корреляции, рассчитанный при процедуре скользящего контроля с исключением по одному.

$R^2_{тест}$ – квадрат коэффициента корреляции для тестовой выборки.

Ингибиторы циклин-зависимой киназы 2

Модели были построены для обучающей выборки $CDK2_{обуч}$ и проведена валидация полученных моделей на тестовой выборке $CDK2_{тест}$. Для данной выборки Дессалев и соавторы построили QSAR модель с использованием метода CoMFA (Dessalew N. et al., 2007). В таблице 1 представлены характеристики моделей, полученных методом CoMFA и при помощи методов, предложенных в диссертационной работе.

Таблица 1. Оценка прогностической способности QSAR моделей для ингибиторов CDK2.

Метод	R^2	Q^2	$R^2_{тест}$
MNA_SCR	0,85	0,78	0,72
qQNA_SCR	0,88	0,83	0,49
QNA_Cheb_SCR	0,84	0,77	0,89
CoMFA	0,94	0,56	0,86

Лучшая CoMFA модель была получена при использовании стандартных значений стерических и электростатических полей. Ей соответствует высокое значение R^2 (0,94), которое превосходит значения, полученные другими методами. Однако значения Q^2 (0,56)

для CoMFA модели существенно ниже, чем для остальных моделей. При этом предсказательная способность у моделей, полученных методами MNA_SCR и qQNA_SCR (0,72 и 0,49, соответственно), оказалась гораздо ниже, чем у CoMFA модели (0,86). Наилучший результат по предсказанию для тестовой выборки показан для метода QNA_Cheb_SCR (0,89). Таким образом, метод QNA_Cheb_SCR обладает лучшей точностью предсказания по сравнению с методами CoMFA, MNA_SCR и qQNA_SCR для выборки ингибиторов CDK2.

Ингибиторы дигидрофолат редуктазы

QSAR-моделирование выполнено для обучающей выборки $DHFR_{обуч}$, а валидация полученных моделей была проведена на тестовой выборке $DHFR_{тест}$. Эти выборки были использованы Джефреем и соавторами (Jeffrey J. et al., 2004) для QSAR моделирования на основе PLS с использованием различных дескрипторов, реализованных в методах CoMFA, CoMSIAbasic, CoMSIAextra HQSAR, EVA, и в программе Cerius2. В таблице 2 представлены характеристики моделей, полученных Джефреем и в диссертационной работе.

Таблица 2. Оценка прогностической способности QSAR моделей для ингибиторов дигидрофолат редуктазы.

Метод	R^2	Q^2	$R^2_{тест}$
MNA_SCR	0,86	0,63	0,60
qQNA_SCR	0,68	0,63	0,03
QNA_Cheb_SCR	0,78	0,72	0,61
CoMFA	0,79	0,65	0,59
CoMSIAbasic	0,76	0,63	0,52
CoMSIAextra	0,75	0,65	0,53
HQSAR	0,81	0,69	0,63
EVA	0,81	0,64	0,57
2D Cerius2	0,61	0,51	0,47
3D Cerius2	0,65	0,53	0,49

Модели, полученные с помощью CoMFA, EVA и HQSAR, имеют значение R^2 немного выше (0,79, 0,81 и 0,81, соответственно), чем у QNA_Cheb_SCR модели (0,78). При этом MNA_SCR модели соответствует самый высокий показатель R^2 (0,86). Наилучший результат по процедуре скользящего контроля с исключением по одному, по сравнению со всеми методами, был получен для QNA_Cheb_SCR модели (0,72). Модели, полученные на основе CoMSIAbasic и CoMSIAextra, сопоставимы с моделями MNA_SCR и qQNA_SCR (значения Q^2 0,63, 0,65, 0,51 и 0,53, соответственно). Самые худшие результаты по Q^2 были получены

для 2D и 3D дескрипторов программы Cerius2 (0,51 и 0,53, соответственно). Из трех моделей, основанных на дескрипторах атомных окрестностей, лучшая прогностическую точность получена для модели QNA_Cheb_SCR – $R^2_{тест}$ = 0,61. Такая точность, полученная на гетерогенной выборке ингибиторов дигидрофолат редуктазы, сопоставима с результатами моделей CoMSIAbasic, CoMSIAextra, EVA, CoMFA и MNA_SCR, незначительно меньше, чем результат, полученный HQSAR, и существенно лучше, чем значения, полученные на 2D/3D дескрипторах Cerius2 и модели qQNA_SCR.

Ингибиторы ангиотензин-превращающего фермента

QSAR-моделирование выполнено для обучающей выборки $ACE_{обуч}$, а валидация полученных моделей проведена на тестовой выборке $ACE_{тест}$. Эти выборки были использованы Джефреем и соавторами (Jeffrey J. et al., 2004) для QSAR моделирования на основе PLS и различных дескрипторов, реализованных в методах CoMFA, CoMSIAbasic, CoMSIAextra HQSAR, EVA, и в программе Cerius2. В таблице 3 представлены характеристики моделей, полученных Джефреем и в диссертационной работе.

Таблица 3. Оценка прогностической способности QSAR моделей для ингибиторов ангиотензин-превращающего фермента.

Метод	R^2	Q^2	$R^2_{тест}$
MNA_SCR	0,81	0,52	0,46
qQNA_SCR	0,79	0,74	0,16
QNA_Cheb_SCR	0,85	0,80	0,54
CoMFA	0,80	0,68	0,49
CoMSIAbasic	0,76	0,65	0,52
CoMSIAextra	0,73	0,66	0,49
HQSAR	0,84	0,72	0,30
EVA	0,84	0,70	0,36
2D Cerius2	0,76	0,68	0,47
3D Cerius2	0,82	0,72	0,51

Из таблицы 3 видно, что лучшие характеристики модели были получены с помощью метода QNA_Cheb_SCR. Разница между значениями R^2 QNA_Cheb_SCR модели и другими моделями существенна только в одном случае: CoMSIAextra. Разница между значениями Q^2 QNA_Cheb_SCR модели и другими моделями существенна в пяти случаях: CoMFA, CoMSIAbasic, CoMSIAextra, 2D Cerius2 и MNA_SCR, так как эта разница больше 0,1. Помимо этого, метод QNA_Cheb_SCR показал хорошую прогностическую способность на

выборке $ACE_{тест}$ – $R^2_{тест}$ = 0.54. Таким образом, для данных выборок точность метода QNA_Cheb_SCR по R^2, Q^2, $R^2_{тест}$ выше, чем у других методов.

Ингибиторы цитохрома P450 2A5

QSAR-моделирование проведено для выборок $CYP2A5_{обуч}$ и $CYP2A5_{тест}$. Данные выборки были также исследованы в работе Явонена и соавторов (Juvonen R. et al., 2001). QSAR моделирование было выполнено ими при помощи 3D QSAR методов: CoMFA и GRID/GOLPE. Результаты сравнения характеристик моделей, полученных Явоненом и в диссертационной работе, представлены в таблице 4.

Таблица 4. Оценка прогностической способности QSAR моделей для ингибиторов CYP2A5.

Метод	R^2	Q^2	$R^2_{тест}$
MNA_SCR	0,78	0,29	0,69
qQNA_SCR	0,87	0,80	0,89
QNA_Cheb_SCR	0,91	0,88	0,93
CoMFA	0,94	0,79	0,83
GRID/GOLPE	0,94	0,86	0,90

Из таблицы 4 видно, что разница между величинами R^2, Q^2 и $R^2_{тест}$ моделей qQNA_SCR, QNA_Cheb_SCR, CoMFA и GRID/GOLPE незначительна. Самые низкие характеристики у модели MNA_SCR. Наилучшие результаты по значениям Q^2 и $R^2_{тест}$ были получены с помощью QNA_Cheb_SCR модели. Наивысшие значения R^2 были получены для методов CoMFA и GOLPE. Таким образом, можно сделать вывод, что прогностическая способность модели QNA_Cheb_SCR лучше, чем у моделей CoMFA и GRID/GOLPE.

Ингибиторы цитохрома P450 2A6

QSAR-моделирование проведено для выборок $CYP2A6_{обуч}$ и $CYP2A6_{тест}$. Эти выборки были использованы Явоненом и соавторами (Juvonen R. et al., 2001). QSAR моделирование было выполнено при помощи CoMFA и GRID/GOLPE. Результаты сравнения характеристик моделей, полученных Явоненом и в диссертационной работе, представлены в таблице 5.

Таблица 5. Оценка прогностической способности QSAR моделей для ингибиторов CYP2A6.

Метод	R^2	Q^2	$R^2_{тест}$
MNA_SCR	0,85	0,60	0,76
qQNA_SCR	0,91	0,79	0,57
QNA_Cheb_SCR	0,80	0,72	0,79
CoMFA	0,97	0,81	0,77
GRID/GOLPE	0,93	0,78	0,76

Из таблицы 5 видно, что разница между величинами Q^2 и $R^2_{тест}$ моделей QNA_Cheb_SCR, CoMFA и GRID/GOLPE незначительна. Самые низкие характеристики Q^2 у модели MNA_SCR, $R^2_{тест}$ у модели qQNA_SCR. Наилучшие результаты, по значениям $R^2_{тест}$, были получены с помощью QNA_Cheb_SCR модели. Наивысшие значения R^2 были получены для методов CoMFA и GRID/GOLPE. Исходя из важности характеристик моделей – Q^2 и $R^2_{тест}$, – можно сделать вывод, что метод QNA_Cheb_SCR сопоставим с методами CoMFA и GRID/GOLPE.

Соединения, действующие на альфа-2 адренорецепторы

QSAR-моделирование проведено для обучающей выборки $ADREN_{обуч}$. Данная выборка была исследована Грюневолдом и соавторами (Grunewald G. et al., 1999) методом CoMFA. В качестве пробного атома при CoMFA анализе был взят C_{sp3}. В таблице 6 представлены характеристики моделей, полученных Грюневолдом и в диссертационной работе.

Таблица 6. Оценка прогностической способности QSAR моделей для выборки $ADREN_{обуч}$.

Метод	R^2	Q^2
MNA_SCR	0,85	0,07
qQNA_SCR	0,52	-0,07
QNA_Cheb_SCR	0,86	0,78
CoMFA	0,92	0,69

Результаты, полученные с помощью метода QNA_Cheb_SCR, сопоставимы с результатами CoMFA модели. R^2 для QNA_Cheb_SCR модели меньше R^2 для CoMFA модели, но Q^2, полученный на QNA_Cheb_SCR модели, значительно выше, чем Q^2 CoMFA модели. Таким образом, модель, построенная методом QNA_Cheb_SCR, лучше, чем CoMFA модель.

Соединения, действующие на эстрогеновые рецепторы

QSAR-моделирование проведено для обучающей выборки $ESTR_{обуч}$. Данная выборка была исследована Роем и соавторами (Roy K. et al., 2005), которые QSAR-моделирование выполнили на сочетании «E-states» дескрипторов и множественной линейной регрессии (MLR). Характеристики моделей, полученных Роем и в диссертационной работе, представлены в таблице 7.

Таблица 7. Оценка прогностической способности QSAR моделей для выборки $ESTR_{обуч}$.

Метод	R^2	Q^2
MNA_SCR	0,79	0,74
qQNA_SCR	0,87	0,78
QNA_Cheb_SCR	0,88	0,81
MLR и E-states дескрипторы	0,82	0,77

Таблица 7 показывает, что лучшие характеристики модели получены с помощью метода QNA_Cheb_SCR. При этом разница между значениями R^2 и Q^2 QNA_Cheb_SCR модели и остальных моделей незначительна. Важно отметить, что модель на основе qQNA_SCR также превосходит MLR модель. Таким образом, результаты, полученные с помощью метода QNA_Cheb_SCR сопоставимы с результатами, полученными методом, основанным на E-states дескрипторах и MLR.

Соединения, проявляющие острую токсичность для *Chlorella vulgaris*

QSAR-моделирование проведено для обучающей выборки $ALGAE_{обуч}$. Данная выборка была ранее исследована Кронином и соавторами (Cronin M. et al., 2004), которые для QSAR моделирования использовали MLR и PLS, а для описания структур 102 дескриптора, рассчитанных с помощью программ ClogP, MOPAC93, TSAR 3.3 и QSARis ver. 1.1. В таблице 8 представлены характеристики моделей, полученных Кроним и в диссертационной работе.

Таблица 8. Характеристики QSAR моделей острой токсичности для *Chlorella vulgaris*.

Метод	R^2	Q^2
MNA_SCR	0,74	0,30
qQNA_SCR	0,84	0,80
QNA_Cheb_SCR	0.93	0,88
MLR	0,84	0,82
PLS	0,86	0,84

Из таблицы 8 видно, что характеристики QNA_Cheb_SCR модели лучше, чем значения, полученные с помощью MLR и PLS. При этом разница в значениях R^2 и Q^2 QNA_Cheb_SCR модели и остальных моделей, за исключением MNA_SCR модели, незначительна. Таким образом, точность, полученная для нашего метода QNA_Cheb_SCR, выше точности, полученной другими авторами.

Соединения, проявляющие острую токсичность для *Vibrio fischeri*

QSAR-моделирование выполнено для обучающей выборки $VIBRIO_{обуч}$. Данная выборка была исследована ранее другими авторами, использовавшими CoMFA, множественный регрессионный анализ (MLR), регрессию на главные компоненты (PCR) и генетический алгоритм (GFA) в сочетании с ETA дескрипторами и другими 2D дескрипторами (Roy K. et al., 2004). К 2D дескрипторам относились: топологические индексы Винера, Хосойа, молекулярная связность, индекс Балабана, E-State дескрипторы, а также физико-химические дескрипторы, такие, как AlogP98, MolRef и акцептор водородных связей. Комбинации ETA и других 2D дескрипторов также были использованы для построения модели с помощью обратной пошаговой регрессии (Factor score, PCR, MLR). В таблице 9 представлены характеристики моделей, полученные разными методами.

Таблица 9. Характеристики QSAR моделей острой токсичности для *Vibrio fischeri*.

Метод	R^2	Q^2
MNA_SCR	0,66	0,46
qQNA_SCR	0,89	0,83
QNA_Cheb_SCR	0,88	0,84
CoMFA	0.92	0.79
PCR, MLR, ETA дескрипторы	0.84	0.73
PCR, MLR, 2D дескрипторы	0.80	0.76
PCR, MLR, ETA и 2D дескрипторы	0.80	0.76
Factor score, PCR, MLR, ETA дескрипторы	0.89	0.82
Factor score, PCR, MLR, 2D дескрипторы	0.87	0.83
Factor score, PCR, MLR, ETA и 2D дескрипторы	0.91	0.85
GFA, ETA дескрипторы	0.86	0.77
GFA, 2D дескрипторы	0.82	0.81
GFA, ETA и 2D дескрипторы	0.87	0.78

Таблица 9 показывает, что значения R^2 и Q^2, полученные для метода QNA_Cheb_SCR, близки к таковым, полученным при помощи обратной пошаговой регрессии (Factor score, PCR, MLR), использующей комбинацию 2D дескрипторов с ETA дескрипторами. При этом разница между значениями R^2 и Q^2 этих методов незначительна. Метод qQNA_SCR также показал высокие значения R^2 и Q^2 на уровне лучших значений других моделей. Таким образом, методы QNA_Cheb_SCR и qQNA_SCR пригодны для QSAR моделирования значений острой токсичности для *Vibrio fischeri*.

Соединения, проявляющие острую токсичность для *Tetrahymena pyriformis*

QSAR-моделирование проведено для обучающей выборки $TETRA_{обуч}$. Данная выборка была исследована Кроним и соавторами (Cronin M. et al., 2002), которыми для QSAR-моделирования были использованы MLR, PLS и пошаговая регрессия (SWR), а для описания структур – 108 дескрипторов, рассчитанных с помощью программ ACD/Labs, Chem-X version 2000.1, MOPAC ver. 6.49, TSAR 3.3 и QSARis ver. 1.1. В таблице 10 представлены характеристики моделей, полученных Кроним и в диссертационной работе.

Таблица 10. Оценка прогностической способности QSAR моделей для выборки $TETRA_{обуч}$.

Метод	R^2	Q^2
MNA_SCR	0,78	0,47
qQNA_SCR	0,69	0,65
QNA_Cheb_SCR	0,80	0,75
MLR	0,69	0,67
SWR	0,65	0,63
PLS	0,77	не представлено

Таблица 10 показывает, что результаты, полученные при помощи методов QNA_Cheb_SCR и qQNA_SCR, существенно лучше результатов, полученных с помощью MLR, SWR, PLS и MNA_SCR моделей на гетерогенной выборке. При этом точность метода QNA_Cheb_SCR также превосходит точность qQNA_SCR.

Статистическое сравнение QSAR методов

Метод MNA_SCR показал хорошие результаты только на пяти из десяти выборок (CDK2, ACE, DHFR, ESTR и CYP2A6). Таким образом, можно констатировать, что MNA дескрипторы, в основном, пригодны для QSAR-моделирования лиганд-ферментных взаимодействий, и частично пригодны для моделирования лиганд-рецепторных взаимодействий.

Анализ результатов, полученных методом qQNA_SCR, показал, что он дает хорошую точность для шести выборок из десяти (VIBRIO, ALGAE, TETRA, ESTR, CYP2A5 и CYP2A6). Исходя из видов биологической активности в этих шести выборках, можно сделать вывод, что преобразование QNA дескрипторов в виде квантилей пригодно для QSAR-моделирования в случаях острой токсичности и взаимодействия ингибиторов с ферментами, метаболизирующими ксенобиотики, и частично пригодно для случая лиганд-рецепторных взаимодействий.

Характеристики выше упомянутых моделей показали, что метод QNA_Cheb_SCR имеет сопоставимую или лучшую точность предсказания для всех десяти исследованных выборок по различным видам биологической активности. Для метода QNA_Cheb_SCR значения $R^2_{тест}$ на всех 10 выборках выше значений $R^2_{тест}$, полученных методами MNA_SCR и qQNA_SCR, что свидетельствует о лучшей прогностической способности данного метода. Величина Q^2 у метода QNA_Cheb_SCR была ниже, чем у методов MNA_SCR и qQNA_SCR, только в трех случаях (один и два раза соответственно). При этом во всех трех случаях разница в значениях Q^2 была незначительна.

Для десяти исследованных выборок мы сравнили значения Q^2 и $R^2_{тест}$ QNA_Cheb_SCR моделей с известными из литературы параметрами QSAR моделей, полученными различными методами. При этом было всего 35 значений Q^2 и 19 значений $R^2_{тест}$. В случае, если метод QNA_Cheb_SCR ничем не отличается от других методов, можно ожидать, что метод QNA_Cheb_SCR в половине случаев будет иметь лучшие значения Q^2 ($R^2_{тест}$), а в другой половине случаев – худшие. Примем данное утверждение за нулевую гипотезу. Вероятность наблюдения *m* или более успехов в *n* сравнениях, при условии нулевой гипотезы, выражается следующим образом:

$$\Pr(m \mid n) = 2^{-n} \sum_{k=m}^{n} \frac{n!}{k!(n-k)!}$$

В 32 из 35 случаев значения Q^2, полученные методом QNA_Cheb_SCR, выше, то есть вероятность нулевой гипотезы $\Pr(32 \mid 35) = 2{,}09 * 10^{-7}$. Значения $R^2_{тест}$, полученные методом QNA_Cheb_SCR, в 18 случаях из 19 выше, и, вероятность нулевой гипотезы $\Pr(18 \mid 19) = 3{,}81 * 10^{-5}$. Данная оценка показывает, что точность метода QNA_Cheb_SCR достоверно превосходит точность других 2D и 3D QSAR методов, использованных для сравнения в данной работе.

Также необходимо отметить, что значения R^2, полученные методом QNA_Cheb_SCR, в 36 сравнениях оказались лучшими только 22 раза. При этом вероятность нулевой гипотезы – $\Pr(22 \mid 36) = 0{,}12$. Такой результат можно объяснить гораздо меньшим переобучением метода QNA_Cheb_SCR по сравнению с другими методами.

Реализация метода – программа GUSAR

На основе всесторонней валидации разработанных и представленных в диссертационной работе методов QSAR, основанных на дескрипторах атомных окрестностей, был отобран лучший метод – QNA_Cheb_SCR. Он был реализован в компьютерной программе GUSAR (General Unrestricted Structure-Activity Relationships),

которая была создана с использованием среды программирования Delphi 5.0 Professional и работает под операционной системой Windows.

Программа позволяет создавать QSAR модели на основе структурных формул соединений обучающей выборки, без необходимости привлечения данных о трехмерной структуре белка или информации о трехмерных структурах молекул лигандов.

Благодаря разработанному уникальному алгоритму и единому типу дескрипторов с помощью программы GUSAR можно создавать отдельную QSAR модель для каждого вида биологической активности, и затем сохранять все полученные модели в одном файле. Сохраненные модели можно использовать для одновременного прогноза нескольких видов биологической активности для новых химических соединений. На рисунке 2 приведен пример интерфейса программы GUSAR, с загруженными десятью моделями, созданными по десяти вышеописанным выборкам.

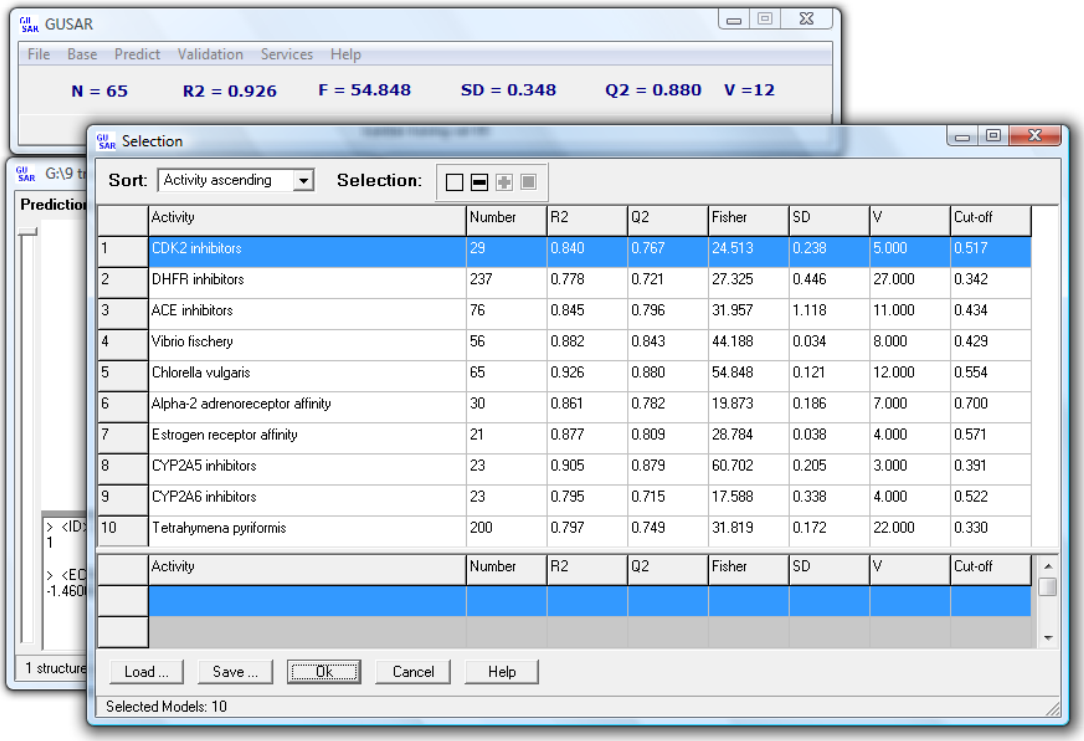


Рис. 2. Интерфейс программы GUSAR с сохраненными моделями.

Как видно из рисунка 2, программа GUSAR позволяет не только прогнозировать количественные значения по всем моделям, но и выбирать нужные модели для прогноза, а также сортировать данные модели по их параметрам.

Проверка устойчивости прогноза количественных свойств органических соединений

Для проверки устойчивости метода QNA_Cheb_SCR, реализованного в программе GUSAR, была выполнена процедура случайного двадцатикратного разбиения каждой

исследуемой выборки на обучающие и тестовые. При этом 20% соединений из начальной выборки относились к тестовой, а 80% соединений – к обучающей выборке. Затем по полученной обучающей выборке строилась модель, по которой предсказывалась тестовая выборка. Процедуры обучения и предсказания выполнялись для каждого разбиения.

Перед началом проверки устойчивости метода пять исследуемых обучающих выборок ($CDK2_{обуч}$, $DHFR_{обуч}$, $ACE_{обуч}$, $CYP2A5_{обуч}$ и $CYP2A6_{обуч}$) были объединены с тестовыми выборками ($CDK2_{тест}$, $DHFR_{тест}$, $ACE_{тест}$, $CYP2A5_{тест}$ и $CYP2A6_{тест}$). Результаты исследования устойчивости метода представлены в таблице 11.

Таблица 11. Оценка устойчивости метода QNA_Cheb_SCR.

Выборки	**Мин. $R^2_{тест}$**	**Макс. $R^2_{тест}$**	**Сред. $R^2_{тест}$**	**$Q^2_{нач.выб.}$**
Ингибиторы циклин-зависимой киназы 2	0,54	0,97	0,81	0,86
Ингибиторы дегидрофолат редуктазы	0,52	0,75	0,62	0,72
Ингибиторы ангиотензин-преврашающего фермента	0,53	0,82	0,67	0,74
Токсичность на *Vibrio fischeri*	0,60	0,95	0,84	0,83
Токсичность на *Chlorella vulgaris*	0,57	0,92	0,78	0,85
Токсичность на *Tetrahymena pyriformis*	0,41	0,73	0,60	0,74
Сродство к альфа 2 адренорецептору	0,50	0,99	0,83	0,78
Сродство к эстрогеновому рецептору	0,56	0,99	0,86	0,84
Ингибиторы цитохрома P450 изоформы 2A5	0,65	0,99	0,93	0,89
Ингибиторы цитохрома P450 изоформы 2A6	0,54	0,97	0,78	0,75
Среднее значение	**0,54**	**0,91**	**0,77**	**0,80**

Таблица 11 содержит: минимальные, максимальные и средние значения $R^2_{тест}$, полученные при двадцатикратном разбиении, и значения Q^2, полученные на всей выборке до разбиения. Результаты, представленные в таблице 11, свидетельствует о высокой устойчивости и предсказательной способности метода QNA_Cheb_SCR, так как разница между средними значениями $R^2_{тест}$ (0,77) и Q^2 (0,80) по всем десяти выборкам всего 0,03.

Сравнение методов оценки области применимости модели

Нами было проведено сравнение разработанного метода оценки области применимости модели (модифицированная оценка) с наиболее популярными методами оценки области применимости модели – классическая оценка и методом ближайшего соседа.

Для сравнения была использована выборка ингибиторов дигидрофолат редуктазы ($DHFR_{обуч}$ и $DHFR_{тест}$), так как она содержит наибольшее количество соединений (237 в обучающей и 124 в тестовой выборках). Для сопоставления методов мы построили график

зависимости ошибки предсказания активности соединений тестовой выборки от величины оценки области применимости модели для данных соединений (рис. 3).

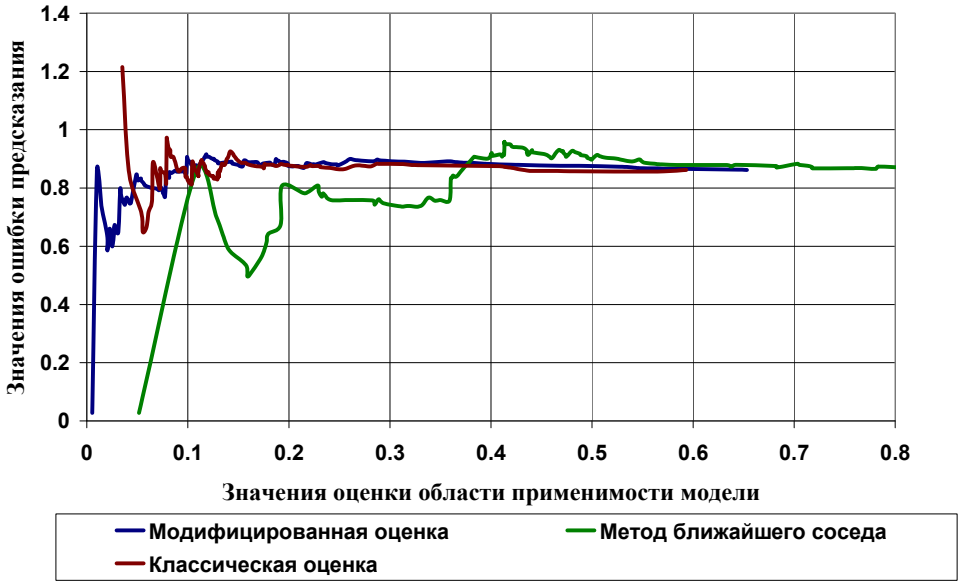


Рис. 3. График зависимости ошибок предсказания от оценки области применимости модели.

Как видно из рисунка 3, наиболее гладкий график зависимости значений ошибок предсказания от оценок области применимости модели получился при использовании метода модифицированной оценки (синий цвет). На двух графиках (синий и зеленый цвет), полученных по результатам модифицированной оценки и метода ближайшего соседа, видна тенденция, что чем меньше оценка области применимости модели, тем выше точность предсказания. При использовании классической оценки такая тенденция отсутствует (график красного цвета).

Помимо представленных графиков было подсчитано, сколько соединений попадает в область применимости модели при разных порогах. Результаты представлены в таблице 12.

Таблица 12. Число соединений тестовой выборки в зависимости от порога отсечения по оценке применимости модели.

	Модифицированная оценка		**Классическая оценка**		**Метод ближайшего соседа**	
Порог отсечения	**Количество соединений**	**Ошибка предсказания (RMSE)**	**Количество соединений**	**Ошибка предсказания (RMSE)**	**Количество соединений**	**Ошибка предсказания (RMSE)**
0,1	44	0,897	29	0,827	1	0,027
0,2	85	0,875	98	0,879	13	0,807
0,3	105	0,894	112	0,882	27	0,747
0,4	115	0,882	114	0,878	42	0,900
0,5	118	0,876	122	0,857	79	0,898

Как видно из таблицы 12, ошибка предсказания для соединений тестовой выборки в зависимости от порога отсечения для всех трех методов примерно одинаковая. Исключение составляет только порог отсечения, равный 0,1: у метода ближайшего соседа наилучший показатель ошибки, но при этом в область применимости попало только одно соединение. Анализируя таблицу 12, можно сделать вывод, что при использовании методов модифицированной и классической оценки в область применимости попадает гораздо большее количество соединений из тестовой выборки, чем при методе ближайшего соседа, при этом точность методов примерно одинаковая.

Подводя итог полученным результатам, можно сделать вывод, что наилучшим из трех методов оценки области применимости модели является предложенный нами метод модифицированной оценки, так как он объединяет в себе лучшие качества двух других методов (классической оценки и ближайшего соседа).

ЗАКЛЮЧЕНИЕ

В диссертационной работе исследована возможность количественного предсказания биологической активности органических соединений на основе дескрипторов атомных окрестностей. Разработано три метода построения QSAR моделей, сравнительный анализ которых помог выявить универсальный метод QSAR моделирования (QNA_Cheb_SCR), пригодный для количественного анализа взаимосвязей структура-активность различных типов биологической активности. Показано, что метод MNA_SCR пригоден для QSAR моделирования лиганд-белковых взаимодействий, а метод qQNA_SCR пригоден для QSAR моделирования острой токсичности и взаимодействий с ферментами, метаболизирующими ксенобиотики. Разработанные методы могут предсказывать количественные значения биологической активности химических соединений по их структурной формуле, и при этом не требуют использования информации о трехмерной структуре химического соединения и/или белка-мишени. Они основаны на одном типе дескрипторов и едином алгоритме, в отличие от классических методов QSAR. Точность разработанных методов была сопоставлена с широко применяемыми 3D и 2D QSAR методами на десяти разных выборках. Результаты исследования показали, что прогностическая способность метода QNA_Cheb_SCR в большинстве исследованных случаев лучше других QSAR методов, как на разнородных выборках (ингибиторы дигидрофолат редуктазы; ингибиторы ангиотензин-превращающего фермента; соединения, обладающие острой токсичностью для *Tetrahymena pyriformis*), так и на однородных (ингибиторы циклин-зависимой киназы 2; соединения, обладающие острой токсичностью для *Vibrio fischeri*; соединения, обладающие острой токсичностью для *Chlorella vulgaris*; соединения, действующие на альфа-2 адренорецептор;

соединения, действующие на эстрогеновый рецептор; ингибиторы цитохрома P450 2A5; ингибиторы цитохрома P450 2A6). Предложенный метод реализован в компьютерной программе GUSAR, показана его устойчивость и высокая предсказательная способность. В программе GUSAR реализована возможность предсказания количественных значений биологической активности одновременно по многим моделям, что является необходимым при поиске соединений, обладающих множественным механизмом действия, при оценке их ADMET свойств и возможного побочного действия. Предложен и реализован в программе GUSAR новый метод оценки области применимости модели, существенное преимущество, которого было продемонстрировано путем сопоставления с другими методами.

ВЫВОДЫ

1. Дескрипторы атомных окрестностей применимы для описания органических соединений в задачах количественного анализа взаимосвязей структура-активность.
2. Разработаны и программно реализованы три алгоритма прогноза количественных свойств органических соединений: MNA_SCR использует MNA дескрипторы и преобразование с помощью нечетких градаций; qQNA_SCR использует QNA дескрипторы и преобразование с помощью квантилей и QNA_Cheb_SCR использует QNA дескрипторы и преобразование с помощью полиномов Чебышева.
3. Точность алгоритмов MNA_SCR и qQNA_SCR сопоставима с другими QSAR методами. Алгоритм на основе QNA дескрипторов, полиномов Чебышева и самосогласованной регрессии – наилучший из исследованных, его точность достоверно выше использованных для сравнения других известных QSAR методов. Он реализован в виде компьютерной программы GUSAR.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Поройков В.В., Филимонов Д.А., Лагунин А.А., Глориозова Т.А., Рудик А.В., Степанчикова А.В., Акимов Д.В., Захаров А.В., Дмитриев А.В. Компьютерная оценка спектра биологической активности химических соединений с целью минимизации рисков их применения в медицине. – Москва, 2004. – С. 167-169.
2. Захаров А.В., Лагунин А.А., Поройков В.В., Филимонов Д.А. Компьютерный поиск соединений, ингибирующих рост и развитие опухолевых клеток. Материалы XII Российского национального конгресса «Человек и лекарство». – Москва (Россия). – 2005. – С. 665.
3. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. Computer prediction of human carcinogenicity for chemical compounds according to the IARC classification. // – Ankara: CADD & D Society, 2005. – P. 106.
4. Захаров А.В., Лагунин А.А., Филимонов Д.А., Поройков В.В. Новый метод количественного анализа взаимосвязей «структура – активность»: апробация на примере ингибиторов циклин-зависимой киназы. // Материалы Московской международной конференции «Биотехнология и медицина». – Москва (Россия). – 2006. – С. 51.
5. Захаров А.В., Лагунин А.А., Филимонов Д.А, Поройков В.В. Прогноз количественных свойств химических соединений на основе дескрипторов атомных окрестностей. // Материалы XIII Российского национального конгресса «Человек и лекарство». – Москва (Россия). – 2006. – С. 16.
6. Захаров А.В., Лагунин А.А., Филимонов Д.А., Поройков В.В. Количественный анализ взаимосвязи «структура-активность» ингибиторов циклин-зависимой киназы 1. // Биомедицинская химия. – 2006. – Т. 52. № 1. – С. 3-18.
7. Lagunin A.A., Zakharov A.V., Filimonov D.A., Poroikov V.V. New approach for QSAR modeling of the acute toxicity. // The 12th International Workshop on Quantitative Structure – Activity Relationships in Environmental Toxicology (Euro QSAR 2006). – Paris (France). – 2006. – P. 37.
8. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. Validation of support compounds set based approach for QSAR modelling. // The 3 rd International Conference «Genomics, Proteomics, Bioinformatics and Nanotechnologies for Medicine». – Novosibirsk (Russia). – 2006. – P. 127.
9. Захаров А.В., Филимонов Д.А., Лагунин А.А., Поройков В.В. Свидетельство об официальной регистрации программы для ЭВМ GUSAR (General Unrestricted Structure

Activity Relationships) № 2006613591 от 16 октября 2006 г., Москва, Федеральная служба по интеллектуальной собственности, патентам и товарным знакам.

10. Захаров А.В., Лагунин А.А., Филимонов Д.А., Поройков В.В. Прогноз количественных свойств химических соединений на основе QNA дескрипторов. // Материалы XIV Российского национального конгресса «Человек и лекарство». – Москва (Россия). – 2007. – С. 285.
11. Lagunin A.A., Zakharov A.V., Filimonov D.A., Poroikov V.V. A new approach to QSAR modelling of acute toxicity. // SAR and QSAR in Environmental Research. – 2007. – V. 18. № 3-4. – P. 285-298.
12. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. QSAR modelling of rat's carcinogenic toxicity. // Abstr. 4rd Internat. Symp. «Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2007)». – Moscow (Russia). – 2007. – P. 170.
13. Poroikov V.V., Filimonov D.A., Lagunin A.A., Gloriozova T.A., Zakharov A.V. PASS: identification of probable targets and mechanisms of toxicity. // SAR and QSAR in Environmental Research. – 2007. – V. 18. № 1-2. – P. 101-110.
14. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. QSAR Modelling of acute toxicity in the Fathead Minnow. // Abstr. 3rd German conference on chemoinformatics. – Goslar (Germany). – 2007. – P. 71.
15. Захаров А.В., Лагунин А.А., Филимонов Д.А., Поройков В.В. Прогноз количественных свойств органических соединений на основе QNA дескрипторов. // Материалы XV Российского национального конгресса «Человек и лекарство». – Москва (Россия). – 2008. – С. 521.
16. Zakharov A.V. QSAR modeling of antineoplastic activities using NIH Roadmap Data. // Abstr. Helmholtz-Russian-German Workshop on Systems Biology. – Moscow (Russia). – 2008. – P. 76.
17. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. QSAR modeling of antineoplastig activities using NIH roadmap data. // Abstr. 8th International Conference on Chemical Structures. – Noordwijkerhout (the Netherlandes). – 2008. – P. 100.
18. Zakharov A.V., Lagunin A.A., Filimonov D.A., Poroikov V.V. GUSAR: new approach for multiple QSAR. // Abstr. 8th International Conference on Chemical Structures. – Noordwijkerhout (the Netherlandes). – 2008. – P. 101.
19. Geronikaki A., Druzhilovsky D.S., Zakharov A.V., Poroikov V.V. Computer-aided prediction for medicinal chemistry via the Internet. // SAR and QSAR in Environmental Research. – 2008. – V. 19. № 1-2. – P. 27-38.