ISSN 1561-8323 (Print)
ISSN 2524-2431 (Online)

УДК 575.112:004.4+615.011
https://doi.org/10.29235/1561-8323-2023-67-5-388-398

Поступило в редакцию 29.05.2023
Received 29.05.2023

М. И. Шаладонова[1], Я. В. Диченко[2], член-корреспондент С. А. Усанов[2]

[1]Университет Национальной академии наук Беларуси, Минск, Республика Беларусь
[2]Институт биоорганической химии Национальной академии наук Беларуси, Минск, Республика Беларусь

ПРОГНОСТИЧЕСКАЯ МОДЕЛЬ ИДЕНТИФИКАЦИИ НОВЫХ ЛИГАНДОВ CYP19A1 НА АНАЛИТИЧЕСКОЙ ПЛАТФОРМЕ KNIME

Аннотация. Сформирована база данных химических соединений – низкомолекулярных лигандов CYP19A1 (ароматазы) человека на основании проанализированных данных, полученных *in vitro*. С использованием полученной базы данных при помощи метода машинного обучения «случайный лес деревьев принятия решений» на аналитической платформе KNIME построены две прогностические модели для идентификации активности лигандов стероидной (I типа) и нестероидной (II типа) структуры. В качестве обучающих данных при построении модели применялись топологические дескрипторы химической структуры, учитывающие корреляцию между структурой молекулы и биологическим эффектом. Для каждой модели был осуществлен отбор наиболее значимых признаков (дескрипторов), произведено вычисление оптимальных параметров и найдена область применимости моделей. На основании результатов показателей качества AUC проведена оценка способности моделей предсказывать результаты тестовой выборки. Полученные показатели качества свидетельствуют о достаточно высокой прогностической способности моделей и перспективности их использования для идентификации новых лигандов CYP19A1 человека. Найденные таким способом соединения могут рассматриваться как потенциальные к созданию лекарственные препараты для лечения гормон-зависимых опухолей.

Ключевые слова: CYP19A1 человека, ингибиторы ароматазы, лиганд, топологические дескрипторы, машинное обучение, прогностическая модель, область применимости, идентификация препаратов

Для цитирования. Шаладонова, М. И. Прогностическая модель идентификации новых лигандов CYP19A1 на аналитической платформе KNIME / М. И. Шаладонова, Я. В. Диченко, С. А. Усанов // Докл. Нац. акад. наук Беларуси. – 2023. – Т. 67, № 5. – С. 388–398. https://doi.org/10.29235/1561-8323-2023-67-5-388-398

Marina I. Shaladonova[1], Yaraslau V. Dzichenka[2], Corresponding Member Sergei A. Usanov[2]

[1]University of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus
[2]Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus

PREDICTIVE MODEL FOR IDENTIFYING NEW CYP19A1 LIGANDS ON THE KNIME ANALYTICAL PLATFORM

Abstract. The purpose of this study was to create a database of the chemical compounds – ligands of human steroid-hydroxylating cytochrome CYP19A1 (aromatase) in order to build a predictive model. The idea was to create a model on the basis of the machinery learning method such as random forest for two types of ligands – with steroidal (I type) and non-steroidal structure (II type). Two predictive models were built with the help of the KNIME analytical platform. Topological descriptors of the chemical structure were used as training data when building a model that takes into account their correlation between the structure of the molecule and the biological effect. The selection of the feature importance of the descriptors, optimal parameters of random forest and the definition of applicability domain of the models were carried out. The assessment of the ability to predict the results of a test sample was performed for each model. The quality marks of the obtained models indicated a rather high predictive ability of the models and the prospects of their use for identification of new human CYP19A1 ligands as potential drugs for treatment of hormone-dependent tumors.

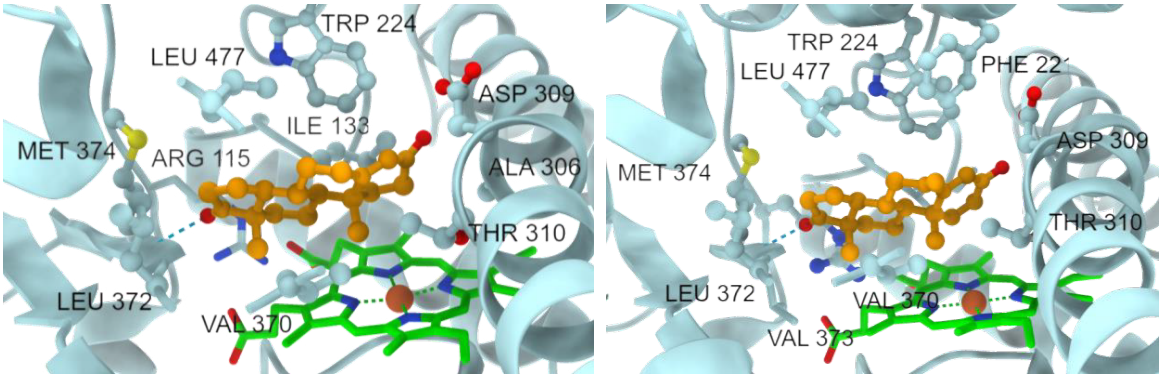
Keywords: human CYP19A1, aromatase inhibitors, ligand, topological descriptors, machinery learning, predictive model, applicability domain, drug identification

For citation. Shaladonova M. I., Dzichenka Ya. V., Usanov S. A. Predictive model for identifying new CYP19A1 ligands on the KNIME analytical platform. *Doklady Natsional'noi akademii nauk Belarusi = Doklady of the National Academy of Sciences of Belarus*, 2023, vol. 67, no. 5, pp. 388–398 (in Russian). https://doi.org/10.29235/1561-8323-2023-67-5-388-398

Введение. Создание структур новых лекарственных препаратов в современной практике зачастую осуществляется с использованием различных методов компьютерного моделирования, позволяющих значительно уменьшить количество образцов, которые необходимо протестировать

in vitro. Но вместе с тем современные методы компьютерного моделирования часто не позволяют с достаточной точностью описать процесс взаимодействия лиганда с мишенью и оценить его эффективность, что, в свою очередь, приводит к ошибочному прогнозу при скрининге низкомолекулярных биорегуляторов. Использование методов машинного обучения является одним из возможных способов решения этой проблемы. Значительное улучшение предсказательной способности компьютерной модели в данном случае происходит за счет того, что при ее обучении неявно учитываются различные типы межмолекулярных взаимодействий, корректно смоделировать которые с использованием «классических» методов молекулярной механики или молекулярного докинга не представляется возможным [1].

Ароматаза (CYP19A1 человека) принадлежит к семейству стероид-гидроксилирующих цитохромов Р450 и является скорость-лимитирующим ферментом биосинтеза эстрогенов из андрогенов, катализируя реакцию ароматизации цикла А и образование эстрона и эстрадиола из андростендиона и тестостерона соответственно. CYP19A1 человека является основной мишенью при терапии некоторых видов опухолей: высокий уровень эстрогена в организме приводит к росту и пролиферации раковых клеток в молочной железе и эндометрии, а также к рецидивам и метастазированию. Снижение уровня эстрогенов путем ингибирования процесса их биосинтеза считается одной из эффективных стратегий в лечении гормон-зависимых злокачественных опухолей [2]. На рис. 1 отображено взаимодействие CYP19A1 с субстратом (рис. 1, *a*) и лигандом (рис. 1, *b*).



a *b*

Рис 1. Фрагменты полости активного центра пространственных структур CYP19 человека в комплексе с андростендионом (*a*, PDB ID: 3EQM) и экземестаном (*b*, PDB ID: 3S7S). Отмечены аминокислотные остатки, с которыми образуются контакты при связывании лиганда; синим пунктиром обозначена водородная связь с Met374

Fig. 1. Fragments of the cavity of the active site of human CYP19 spatial structures in complex with androstenedione (*a*, PDB ID: 3EQM) and exemestane (*b*, PDB ID: 3S7S). The amino acid residues with which contacts are formed upon binding of the ligand are marked; the blue dotted line indicates the hydrogen bond with Met374

Ингибиторы ароматазы признаны одними из широко используемых лекарственных препаратов для лечения эстроген-зависимого рака [2–8]. По химической структуре и механизму действия различают два основных типа ингибиторов ароматазы: стероидные (I тип) и нестероидные (II тип). Среди наиболее широко известных и используемых в терапии стероидных ингибиторов ароматазы можно привести примеры таких соединений, как форместан, экземестан [4], в качестве примеров нестероидных ингибиторов – производные триазолов, как летрозол, анастрозол, ворозол, а также производные флавоноидов: флавон, флавонон, апигенин, кверцетин [3].

Особенностью всех известных ингибиторов ароматазы является развитие резистентности при длительном применении, а также возникновение ряда побочных эффектов: резорбция костной ткани (остеопороз), артралгии, желудочно-кишечные расстройства, гипертония, отеки, гиперхолестеринемия [6]. Таким образом, остается актуальной потребность в поиске и открытии новых эффективных, но менее токсичных молекул ингибиторов CYP19A1 человека.

В рамках данной работы построены две прогностические модели, основанные на использовании методов машинного обучения, для скрининга и идентификации лигандов I и II типа CYP19A1 человека в обширных базах данных химических соединений.

Работа включала в себя следующие этапы:

создание электронной базы данных лигандов I и II типа с известной активностью по отношению к CYP19A1;

вычисление ряда топологических дескрипторов для каждого лиганда и внесение их значений в базу данных;

кросс-валидация рассчитанных значений топологических дескрипторов, разделение выборки на обучающую и тестовую;

построение прогностических моделей для лигандов I и II типа с использованием метода машинного обучения «случайный лес деревьев принятия решений»;

оценка качества полученных моделей и определение области применимости моделей.

Материалы и методы исследования. *Формирование базы данных химических соединений.* Базу данных лигандов CYP19A1 человека формировали на основании информации, представленной в открытом доступе в научных статьях. В результате поиска был создан файл в формате Excel, содержащий данные о 300 соединениях стероидной и нестероидной структуры, которые изучались на предмет связывания с ферментом CYP19A1 человека и способности ингибировать его функции: название соединения согласно номенклатуре IUPAC; строка SMILES, соответствующая молекуле (получена с использованием инструмента PubChem Sketcher V2.4 (https://pubchem.ncbi.nlm.nih.gov)); значение IC_{50} (концентрация полумаксимального ингибирования – показатель эффективности лиганда при ингибирующем биохимическом или биологическом взаимодействии); величина K_d (константа диссоциации фермент-субстратного комплекса); субстрат, который использовался при определении параметра IC_{50}; тип связывания (I – для стероидных ингибиторов, II – для нестероидных ингибиторов); способ получения фермента CYP19A1; ссылка на источник информации.

Вычисление топологических дескрипторов химической структуры для соединений из базы данных. Для вычисления дескрипторов использовали сервис ChemoPy (http://www.scbdd.com/chemopy_desc/index), представленный на платформе ChemDes (http://www.scbdd.com/chemdes). Данный инструмент позволяет рассчитать 35 различных топологических дескрипторов на основании строки SMILES, кодирующей соединение.

Выбор топологических дескрипторов в качестве параметров, характеризующих структуру молекулы, обусловлен, прежде всего, оптимальным соотношением между легкостью (по сравнению, например, с квантово-химическими) их вычисления и ценностью при построении прогностической модели. Для расчета топологических дескрипторов не нужна информация о биоактивной конформации молекулы: они характеризуют структуру молекулы с точки зрения связности ее атомов, степени разветвленности, наличия гетероатомов и химических связей различного типа. Несмотря на их кажущуюся простоту, неоднократно было показано, что существует достаточно сильная корреляция величин топологических дескрипторов со свойствами молекулы: физико-химическими, токсикологическими, фармакологическими, биологическими.

Топологические дескрипторы, которые были использованы для обучения предсказательных моделей, представлены в табл. 1. После расчета каждого из них соответствующая информация была добавлена в базу данных.

Построение прогностической модели. В рамках данной работы строили прогностическую модель для классификации: соединения, для которых величина IC_{50} была меньше 1 мкМ, относили к «активным» (1) и наоборот, если величина была больше 1 мкМ – к «неактивным» (0).

В работе использовали алгоритм машинного обучения по методу «случайного леса». Данный метод является одним из наиболее часто используемых для решения задач классификации и регрессии в хемоинформатике [9; 10]. Основными его преимуществами являются высокая прогнозирующая способность моделей, простота их построения, отсутствие большого количества свободных параметров, значение которых необходимо оптимизировать, надежность и высокая вычислительная эффективность [1; 9; 10].

Т а б л и ц а 1. **Топологические дескрипторы платформы ChemoPy, используемые для обучения модели**
T a b l e 1. **ChemoPy platform topological descriptors used for model training**

Сокращенное название дескриптора (используемое в базе данных и на платформе ChemoPy) Abbreviated descriptor name (used in the database and ChemoPy platform)	Расшифровка названия дескриптора Decoding the descriptor name
W	Индекс Винера
AW	Средний индекс Винера
J	Индекс Балабана
Thara	Граф Харари
Tsch	Индекс Шульца
Tigdi	Индекс расстояния графа
Platt	Индекс Платта
Xu	Индекс Xu
Pol	Индекс полярности
Dz	Индекс Поглиани
Ipc	Индекс информационного наполнения
BertzCT	Разновидность индекса сложности молекулы
GMTI	Молекулярно-топологический индекс Гутмана
ZM1	Первый индекс Загреба
ZM2	Второй индекс Загреба
MZM1	Модифицированный первый индекс Загреба
MZM2	Модифицированный второй индекс Загреба
Qindex	Квадратичный индекс
Diametert	Топологический диаметр
Radiust	Топологический радиус
Petitjeant	Индекс Петиджан на основе топологии
Sito	Логарифм простого топологического индекса
Hato	Гармонизированный топологический индекс Наруми
Geto	Геометрический топологический индекс Наруми
Arto	Арифметический топологический индекс Наруми
ISIZ	Общий индекс информации о размере молекулы
TIAC	Показатель атомного состава
DET	Индекс равенства расстояний
IDE	Средний индекс равенства расстояний
IVDE	Индекс равенства вершин
Sitov	Логарифм простого топологического индекса Наруми валентных степеней вершин
Hatov	Гармонизированный топологический индекс Наруми валентных степеней вершин
Getov	Геометрический топологический индекс Наруми валентных степеней вершин
Gravto	Гравитационный топологический индекс, основанный на топологической дистанции
GMTIV	Молекулярно-топологический индекс Гутмана валентных степеней вершин

Построение модели осуществляли с использованием локальной версии аналитической платформы KNIME (https://www.knime.com/knime-analytics-platform). Рабочее пространство платформы KNIME (Workspace) представляет собой систему упорядоченных узлов (Nodes), предназначенных для решения отдельных подзадач. Логика обработки данных закладывается через создание потока данных (Workflow): узлов, связанных друг с другом стрелками (Connections), показывающими направление движения данных. После создания Workflow запускается на исполнение, и каждый из узлов выполняет свои заданные функции. Настройка параметров отдельных узлов осуществляется вручную.

Результаты и их обсуждение. *Отбор наиболее значимых признаков для моделей лигандов I и II типа.* Для улучшения прогностической способности моделей на первом этапе их построения

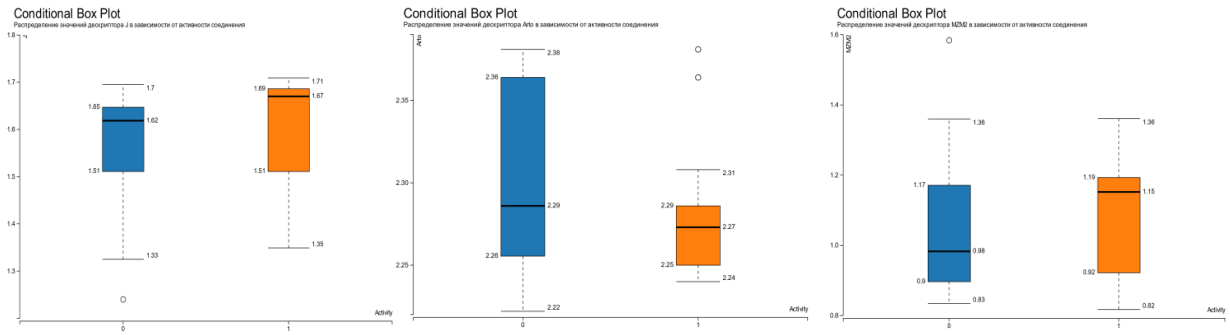
проведен отбор наиболее значимых признаков (дескрипторов), которые в большей степени коррелируют со значением активности исследуемой молекулы. Отбор осуществляли с использованием узла Global Feature Importance, сущность которого заключается в вычислении коэффициента значимости признака-дескриптора с использованием модели «случайного леса». Более высокое значение коэффициента значимости указывает на то, что данный признак является важным с точки зрения построения классификационной модели с бинарным разделением.

При вычислении наиболее значимых признаков для модели лигандов I типа были найдены и отобраны 15 дескрипторов с наилучшими показателями коэффициента значимости.

Соответствующие величины для каждого из 15 дескрипторов составили: J – 0,781; Arto – 0,736; MZM2 – 0,71; Ipc – 0,617; Geto – 0,536; Hato – 0,483; Getov – 0,442; Xu – 0,433; Hatov – 0,43; ZM2 – 0,407; MZM1 – 0,389; BertzCT – 0,369; AW – 0,368; GMTIV – 0,333; Sito – 0,308.

Согласно полученным данным наиболее значимое влияние на активность лигандов I типа оказывают дескрипторы J и Arto. Первый из них имеет отрицательную корреляцию с величиной липофильности молекулы [11], а второй свидетельствует о наличии в структуре молекулы стероидного фрагмента, так как его значение коррелирует со степенями вершин молекулярного графа. Важную роль при бинарном разделении играет и дескриптор MZM2, который характеризует степень разветвленности молекулы.

Статистический анализ диапазона разброса численных значений трех наиболее значимых дескрипторов (рис. 2) показал, что медиана значений индекса Балабана для активных соединений составляет 1,67, а для неактивных – 1,62, т. е. чем выше значение индекса Балабана у соединения, тем, соответственно, выше вероятность того, что оно является активным по отношению к ароматазе человека. Медиана значений дескриптора арифметического топологического индекса Наруми составила 2,27 для активных соединений и 2,29 для неактивных соединений, причем из графика видно, что чем меньше значение дескриптора, тем более вероятна активность соединения. Для значений дескриптора MZM2 медиана активных соединений составила 1,15, а для неактивных соединений – 0,98, что говорит о прямой корреляции значений величины MZM2 по отношению к активности.



a *b* *c*

Рис. 2. Область распределения значений дескрипторов для активных и неактивных соединений (модель для лигандов I типа): *a* – индекс Балабана; *b* – арифметический топологический индекс Наруми; *c* – модифицированный второй индекс Загреба

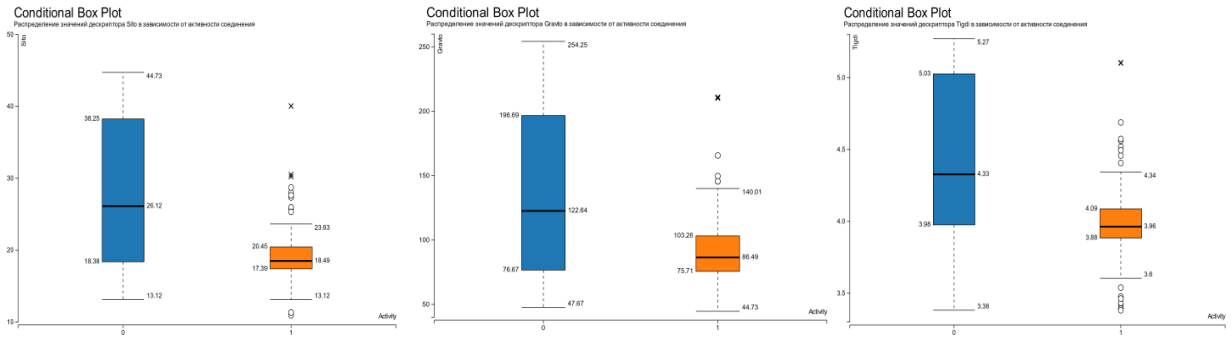
Fig. 2. Distribution area of descriptor values for active and inactive compounds (model for the I type ligands): *a* – Balaban index; *b* – arithmetic topological index by Narumi; *c* – modified Zagreb index with order 2

При вычислении наиболее значимых признаков для модели лигандов типа II отобраны следующие дескрипторы: Sito – 0,734; Gravto – 0,579; Tigdi – 0,577; Pol – 0,566; Thara – 0,539; Platt – 0,532; Quindex – 0,523; Diametert – 0,466; Ipc – 0,462; ZM2 – 0,421; Sitov – 0,42; IDE – 0,402; ZM1 – 0,38; DZ – 0,352; AW – 0,352.

В данном случае, наиболее значимыми являются дескрипторы Sito, который отображает степень разветвленности химической структуры и информацию о заместителях у триазольной, тетразольной или флавоновой структуры лиганда; Gravto, учитывающий взаимное притяжение

и взаимодействие атомов в молекуле; Tigdi, значение которого отражает расстояние между атомами, расположение кратных связей в молекуле и наличие гетероатомов в структуре соединения.

Статистический анализ числовых значений для каждого из данных дескрипторов (рис. 3) показал, что медиана значений логарифма простого топологического индекса (Sito) для активных соединений составила 18,49, в то время как для неактивных соединений значение равно 26,12 (границы верхнего и нижнего квартилей для активных соединений равны 20,45 и 17,39, для неактивных соединений – 38,25 и 18,38 соответственно). Медиана распределения значений дескриптора гравитационного топологического индекса (Gravto) для активных соединений равна 86,49, для неактивных – 122,64 (границы верхнего и нижнего квартилей для активных соединений равны 103,26 и 75,17, для неактивных соединений значения составили 196,69 и 76,67 соответственно). Распределение значений индекса расстояния графа (Tigdi) описано следующими показателями: медиана для активных соединений – 3,96, для неактивных – 4,33; границы верхнего и нижнего квартилей для активных соединений равны 4,09 и 3,88, для неактивных соединений значения составили 5,03 и 3,98 соответственно. Таким образом, можно сделать вывод, что увеличение данных индексов для молекулы приводит к снижению ее активности по отношению к ароматазе человека.



a *b* *c*

Рис. 3. Область распределения значений дескрипторов для активных и неактивных соединений (модель для лигандов II типа): *a* – логарифм простого топологического индекса; *b* – гравитационный топологический индекс; *c* – индекс расстояния графа

Fig. 3. Distribution area of descriptor values for active and inactive compounds (model for the II type ligands): *a* – logarithm of the simple topological index; *b* – gravitational topological index based on topological distance; *c* – graph distance index

Если сравнивать значимость дескрипторов для двух моделей, то можно отметить, что как для лигандов I, так и II типа значимыми являются дескрипторы, характеризующие степень разветвленности молекулы и структуру имеющихся заместителей (для лигандов I типа через дескриптор MZM2, а для лигандов II типа через дескриптор Sito). Для лигандов II типа степень разветвленности молекул, характеризуемая величиной дескриптора Sito (рис. 3, *a*) имеет обратную корреляцию с величиной активности: чем меньше величина дескриптора, тем менее выражен стерический эффект и более выражена активность соединений. Исходя из данных, представленных на рис. 2, *c*, можно сделать вывод, что для лигандов I типа разветвленность молекулы будет, наоборот, способствовать тому, что молекула будет отнесена к классу «активных». Для модели лигандов I типа к наиболее значимому из всех дескрипторов относится индекс Балабана (рис. 2, *a*), характеризующий липофильность стероидной молекулы: чем больше значение индекса Балабана и, соответственно, меньше липофильность молекулы, тем больше вероятность того, что лиганд I типа будет активным.

Построение прогностических моделей. При построении моделей для лигандов I и II типа использовали следующие узлы (рис. 4).

Excel Reader – считывание информации о химических соединениях из созданной базы данных.

Column Filter, Row Filter – фильтрация данных согласно заданным критериям (тип лиганда, набор признаков и т. д.).

Color Manager – цветовая маркировка активных и неактивных соединений по отношению к CYP19A1: активные соединения в таблице интерфейса KNIME отмечены зеленым цветом, неактивные – красным.

X-Partitioner и X-Aggregator – разделение выборки на обучающую и тестовую с использованием *k*-блочной кросс-валидации со значением $k = 10$, в параметрах установлены настройки для стратифицированной выборки по показателю «Активность».

Random Forest Learner – построение и обучение модели с использованием метода «случайного леса». В качестве критерия разделения использовали индекс Джини (Gini Index), который, согласно научным публикациям, является наилучшим показателем разделения для «случайного леса деревьев принятия решений» при построении моделей классификации и регрессии в сравнении с методом Informational Gain Ratio [12].

Parameter Optimization Loop Start, Parameter Optimization Loop End – узлы для вычисления оптимальных параметров модели согласно величине используемой скоринговой функции (в данном случае – значение AUC). В настройках узла выбрана стратегия поиска Brute Force, при которой проверяются все возможные комбинации параметров с учетом заданных интервалов и шага, а затем выбирается наилучшая комбинация. Выбранный диапазон для параметра «глубина дерева принятия решений» для обеих прогностических моделей составил от 5 до 100, а для параметра «число деревьев» – от 10 до 300, размер шага для каждого из параметров – 1.

Random Forest Predictor – прогнозирование активности для тестовой выборки на основании данных, полученных с использованием построенных моделей при выполнении узла Random Forest Learner.

ROC Curve – оценка качества построенной модели с использованием рабочей характеристики приемника (ROC curve). Количественная характеристика, использовавшаяся для оценки качества – площадь под кривой (AUC): чем ближе показатель AUC к 1, тем более качественная модель.

Table Row to Variable – передача значений показателя AUC в узел Parameter Optimization Loop End.

В результате построения Workflow на аналитической платформе KNIME было получено 2 модели (для лигандов I и II типа) с оптимальными параметрами глубины «деревьев принятия решений» и количества построенных моделей «деревьев случайного леса» (рис. 4).

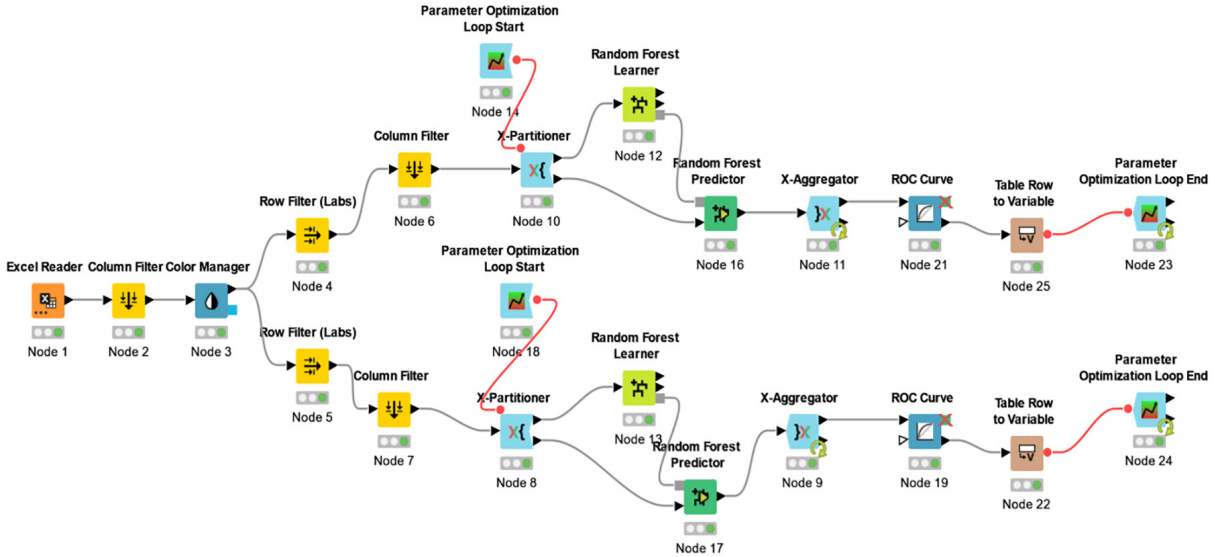


Рис. 4. Поток данных, отображающий последовательность разработки прогностической модели и ее дальнейшую оценку

Fig. 4. Workflow showing the sequence of predictive model development and its further evaluation

Оптимальные параметры, позволяющие получить максимально высокое значение показателя качества AUC, найдены с применением узлов Parameter Optimization Loop Start и Parameter Optimization Loop End и представлены в табл. 2.

Т а б л и ц а 2. **Оптимальные параметры для двух прогностических моделей**

T a b l e 2. **Optimal parameters for two predictive models**

Настраиваемый параметр Configurable parameter	Наилучший параметр Best parameter	
	Лиганды I типа Type I ligands	Лиганды II типа Type II ligands
Критерий разделения	Gini index	Gini index
Глубина дерева решений	46	5
Количество моделей деревьев	17	272
Значение показателя AUC	0,817	0,905

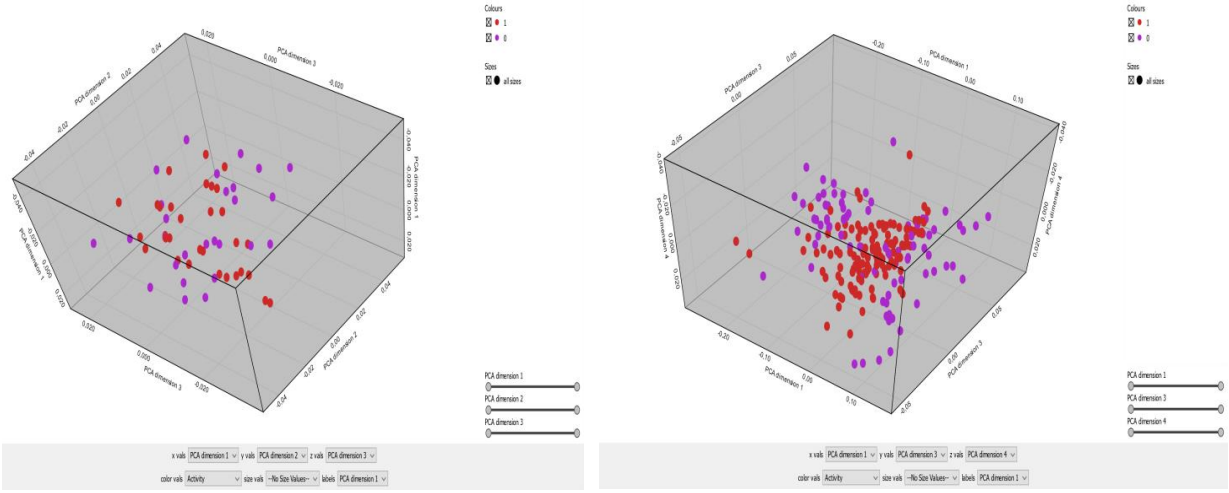
Так как при построении моделей решетчатый поиск параметров вместе с кросс-валидацией проведен только один раз (в процессе поиска параметров не задавали явно значение параметра Random Seed), полученные значения показателя AUC в обоих случаях, строго говоря, не являются корректными с точки зрения статистики и, скорее всего, немного завышены. Для получения наиболее приближенных к реальности значений оценки качества необходимо рассчитать среднее значение показателя AUC для нескольких результатов предсказания одной и той же модели. Для нахождения среднего значения показателя качества AUC были получены ROC-кривые и рассчитаны показатели AUC для 10 случайных выборок при использовании оптимальных параметров для лигандов I и II типа. Среднее значение показателя AUC для модели лигандов I типа составило 0,752, а для модели лигандов II типа – 0,893. Согласно литературным данным, полученные значения свидетельствуют о достаточно высоком качестве построенных моделей [13].

Нахождение области применимости модели. Как правило, полученные модели могут быть использованы для предсказания свойств только лишь молекул, сходных с теми, которые были использованы при построении моделей, т. е. возникает задача нахождения так называемой области применимости модели – области в пространстве признаков молекул из обучающей выборки, для которой прогноз будет статистически значим.

В рамках данной работы для нахождения области применимости моделей применяли метод главных компонент (PCA). Для этого был добавлен соответствующий узел (PCA) к Workflow на платформе KNIME. С целью приведения значений дескрипторов в один масштаб применена нормализация данных при помощи узла Normalize, где использовался метод Decimal Scaling, позволяющий провести десятичное масштабирование путем перемещения десятичной точки на число разрядов, соответствующее порядку числа-значения дескриптора. Область применимости находилась при использовании трех главных компонент с последующим построением 3D-графиков.

На рис. 5, *a* представлен 3D-график области применимости для модели лигандов I типа. Красным цветом на рисунке обозначены активные соединения, фиолетовым – неактивные. Исходя из графика определяется, что основная часть химического пространства расположена в рамках значений первой главной компоненты в пределах диапазона от –0,031 до 0,026, в пределах диапазона второй главной компоненты – от –0,034 до 0,038 и третьей главной компоненты – от –0,02 до 0,024. При рассмотрении химического пространства распределения структур соединений можно сделать вывод о том, что активные соединения располагаются в непосредственной близости друг от друга, формируя небольшие группы, в то время как неактивные расположены хаотично по всей области применимости. Данная особенность говорит о структурном сходстве активных соединений со схожей химической структурой, которая описана в выбранных наиболее значимых топологических дескрипторах для модели лигандов I типа.

На рис. 5, *b* представлен 3D-график области применимости для модели лигандов II типа. Красным цветом на рисунке обозначены активные соединения, фиолетовым – неактивные. Результаты, представленные на графике, показывают, что основная часть химического пространства находится в рамках значений первой главной компоненты в пределах диапазона от –0,232 до 0,128, в пределах диапазона второй главной компоненты – от –0,055 до 0,058 и третьей главной компоненты – от –0,038 до 0,033. При анализе химического пространства распределения



a *b*

Рис. 5. Области применимости для моделей лигандов I (*a*) и II (*b*) типа

Fig. 5. Applicability domain of the models for 1st (a) and 2nd (*b*) type ligands

активных и неактивных соединений необходимо выделить тот факт, что активные соединения в основном сосредоточены посреди области применимости модели, а неактивные соединения замыкают ее. Такие результаты объясняют структурные различия, описанные в выбранных дескрипторах модели для лигандов II типа, позволяющие классифицировать молекулы как активные или неактивные с большей достоверностью, по сравнению с лигандами типа I.

Для определения достоверности результатов PCA модели применяли узел Domain-Leverage, определяющий положение объекта при заданном числе главных компонент PCA модели. Эта величина равна квадрату расстояния Махаланобиса от центра модели до определенного химического объекта тестовой выборки в пространстве и характеризует то, как далеко находится каждый объект в гиперплоскости главных компонент. Считается, что объект тестовой выборки принадлежит области применимости модели и результаты прогноза для него будут надежными, если значение размаха не превышает пороговое. Для модели лигандов I типа пороговое значение показателя размаха составляет 0,138 (для 87,5 % соединений тестовой выборки результаты прогноза будут считаться надежными в найденной области применимости). Для модели лигандов II типа пороговое значение показателя размаха составило 0,044 (для 95,7 % соединений тестовой выборки результаты прогноза будут считаться надежными в найденной области применимости).

Заключение. В проведенной работе с использованием машинного обучения (алгоритм «случайного леса деревьев принятия решений») на аналитической платформе KNIME построены две предсказательные модели для лигандов ароматазы человека первого и второго типа. В качестве обучающих данных для построения модели использовали топологические дескрипторы, характеризующие структуру молекулярного графа. С применением встроенных функций платформы KNIME отобраны наиболее значимые дескрипторы, обладающие наилучшей дискриминирующей способностью при построении каждой из моделей. Методом 10-блочной проверки с решетчатым поиском подобраны оптимальные значения параметров моделей. Значение AUC полученных моделей для лигандов первого типа составило 0,752, а для лигандов второго типа – 0,893, что согласно литературным данным является достаточно высокими показателями и свидетельствует о перспективности использования данных моделей для нахождения новых потенциальных ингибиторов ароматазы. При помощи метода главных компонент оценена область применимости моделей. Полученные модели будут использованы далее для скрининга обширных библиотек химических соединений и идентификации новых лигандов CYP19A1 человека – перспективных молекул для разработки лекарственных препаратов против гормон-зависимых опухолей.

Список использованных источников

1. Guha, R. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors / R. Guha, P. C. Jurs // J. Chem. Inf. Comput. Sci. – 2004. – Vol. 44, N 6. – P. 2179–2189. https://doi.org/10.1021/ci049849f

2. Novel triazole-tetrahydroisoquinoline hybrids as human aromatase inhibitors / C. Chamduang [et al.] // Bioorg. Chem. – 2019. – Vol. 93. – Art. 103327. https://doi.org/10.1016/j.bioorg.2019.103327

3. Brueggemeier, R. W. Aromatase Inhibitors in the Treatment of Breast Cancer / R. W. Brueggemeier, J. C. Hackett, E. S. Diaz-Cruz // Endocrine Rev. – 2005. – Vol. 26, N 3. – P. 331–345. https://doi.org/10.1210/er.2004-0015

4. Bertelli, G. Sequencing of aromatase inhibitors / G. Bertelli // Br. J. Cancer. – 2005. – Vol. 93, N S1. – P. 6–9. https://doi.org/10.1038/sj.bjc.6602689

5. Studies on non-steroidal inhibitors of aromatase enzyme; 4-(aryl/heteroaryl)-2-(pyrimidin-2-yl) thiazole derivatives / Z. Sahin [et al.] // Bioorg. Med. Chem. – 2018. – Vol. 26, N 8. – P. 1986–1995. https://doi.org/10.1016/j.bmc.2018.02.048

6. Aromatase Inhibitors Evolution as Potential Class of Drugs in the Treatment of Postmenopausal Brest Cancer Women / S. Avvaru [et al.] // Mini-Rev. Med. Chem. – 2018. – Vol. 18, N 7. – P. 609–621. https://doi.org/10.2174/1389557517666171101100902

7. Determining the IC_{50} Values for Vorozole and Letrozole, on a Series of Human Liver Cytochrome P450s, to Help Determine the Binding Site of Vorozole in the Liver / L. Raymond [et al.] // Enzyme Research. – 2015. – Vol. 2015. – P. 1–4. https://doi.org/10.1155/2015/321820

8. Synthesis of Aromatase Inhibitors and Dual Aromatase Steroid Sulfatase Inhibitors by Linking an Arylsulfamate Motif to 4-(4H-1,2,4-triazol-4-ylamino)benzonitrile: SAR, Crystal Structures, *in vitro* and *in vivo* Activities / C. Bubert [et al.] // ChemMedChem. – 2008. – Vol. 3, N 11. – P. 1708–1730. https://doi.org/10.1002/cmdc.200800164

9. Баскин, И. И. Введение в хемоинформатику / И. И. Баскин, Т. И. Маджидов, А. А. Варнек. – М., Казань, Страсбург, 2020. – Ч. 4: Методы машинного обучения. – 321 с.

10. Application of the Random Forest Method in Studies of Local Lymph Node Assay Based Skin Sensitization Data / S. Li [et al.] // J. Chem. Inf. Model. – 2005. – Vol. 45, N 4. – P. 952–964. https://doi.org/10.1021/ci050049u

11. Применение метода количественных корреляций структура–свойство (ККСС) с использованием топологического индекса Балабана на примере группы сульфаниламидов / А. В. Сыроешкин [и др.] // Вестн. Рос. ун-та дружбы народов. Сер. Медицина. – 2000. – № 2. – С. 80–83.

12. Optimisation and evaluation of the random forest model in the efficacy prediction of chemoradiotherapy for advanced cervical cancer based on radiomics signature from high-resolution T2 weighted images / D. Liu [et al.] // Arch. Gynecol. Obstet. – 2021. – Vol. 303, N 3. – P. 811–820. https://doi.org/10.1007/s00404-020-05908-5

13. Janitza, S. An AUC-based permutation variable importance measure for random forests / S. Janitza, C. Strobl, A.-L. Boulesteix // BMC Bioinformatics. – 2013. – Vol. 14, N 1. – P. 1–11. https://doi.org/10.1186/1471-2105-14-119

References

1. Guha R., Jurs P. C. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *Journal of Chemical Information and Computer Sciences*, 2004, vol. 44, no. 6, pp. 2179–2189. https://doi.org/10.1021/ci049849f

2. Chamduang C., Pingaew R., Prachayasittikul V., Prachayasittikul S., Ruchirawat S., Prachayasittikul V. Novel triazole-tetrahydroisoquinoline hybrids as human aromatase inhibitors. *Bioorganic Chemistry*, 2019, vol. 93, art. 103327. https://doi.org/10.1016/j.bioorg.2019.103327

3. Brueggemeier R. W., Hackett J. C., Diaz-Cruz E. S. Aromatase Inhibitors in the Treatment of Breast Cancer. *Endocrine Reviews*, 2005, vol. 26, no. 3, pp. 331–345. https://doi.org/10.1210/er.2004-0015

4. Bertelli G. Sequencing of aromatase inhibitors. *British Journal of Cancer*, 2005, vol. 93, no. S1, pp. 6–9. https://doi.org/10.1038/sj.bjc.6602689

5. Sahin Z., Ertas M., Berk B., Biltekin S. N., Yurttas L., Demirayak S. Studies on non-steroidal inhibitors of aromatase enzyme; 4-(aryl/heteroaryl)-2-(pyrimidin-2-yl)thiazole derivatives. *Bioorganic & Medicinal Chemistry*, 2018, vol. 26, no. 8, pp. 1986–1995. https://doi.org/10.1016/j.bmc.2018.02.048

6. Avvaru S. P., Noolvi M. N., Aminbhavi T. M., Chkraborty S., Dash A., Shukla S. S. Aromatase Inhibitors Evolution as Potential Class of Drugs in the Treatment of Postmenopausal Brest Cancer Women. *Mini-Reviews in Medicinal Chemistry*, 2018, vol. 18, no. 7, pp. 609–621. https://doi.org/10.2174/1389557517666171101100902

7. Raymond L., Rayani N., Polson G., Sikorski K., Lian A., VanAlstine-Parris M. A. Determining the IC_{50} Values for Vorozole and Letrozole, on a Series of Human Liver Cytochrome P450s, to Help Determine the Binding Site of Vorozole in the Liver. *Enzyme Research*, 2015, vol. 2015, pp. 1–4. https://doi.org/10.1155/2015/321820

8. Bubert C., Woo L. W. L., Sutcliffe O. B., Mahon M. F., Chander S. K., Purohit A., Reed M. J., Potter B. V. L. Synthesis of Aromatase Inhibitors and Dual Aromatase Steroid Sulfatase Inhibitors by Linking an Arylsulfamate Motif to 4-(4H-1,2,4-triazol-4-ylamino)benzonitrile: SAR, Crystal Structures, *in vitro* and *in vivo* Activities. *ChemMedChem*, 2008, vol. 3, no. 11, pp. 1708–1730. https://doi.org/10.1002/cmdc.200800164

9. Baskin I. I., Madzhilov T. I., Varnek A. A. *Introduction to Chemoinformatics. Vol. 4: Machine learning methods*. Moscow, Kazan, Strasburg, 2020. 321 p. (in Russian).

10. Li S., Fedorowicz A., Singh H., Soderholm S. C. Application of the Random Forest Method in Studies of Local Lymph Node Assay Based Skin Sensitization Data. *Journal of Chemical Information and Modeling*, 2005, vol. 45, no. 4, pp. 952–964. https://doi.org/10.1021/ci050049u

11. Syroeshkin A. V., Kovaleva A. N., Kandalaft E., Pleteneva T. V. Application of a method of quantitative correlations frame – property with usage of a topological coefficient on an example of group of sulfanilamidums. *Vestnik Rossiiskogo universiteta druzhby narodov. Seriya: Meditsina = RUDN Journal of Medicine*, 2000, no. 2, pp. 80–83 (in Russian).

12. Liu D., Zhang X., Zheng T., Shi Q., Cui Y., Wang Y., Liu L. Optimisation and evaluation of the random forest model in the efficacy prediction of chemoradiotherapy for advanced cervical cancer based on radiomics signature from high-resolution T2 weighted images. *Archives of Gynecology and Obstetrics*, 2021, vol. 303, no. 3, pp. 811–820. https://doi.org/10.1007/s00404-020-05908-5

13. Janitza S., Strobl C., Boulesteix A.-L. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 2013, vol. 14, no. 1, pp. 1–11. https://doi.org/10.1186/1471-2105-14-119

Информация об авторах

Шаладонова Марина Игоревна – магистрант. Университет НАН Беларуси (ул. Радиальная, 38Б, 220070, Минск, Республика Беларусь). E-mail: shalmari@tut.by.

Диченко Ярослав Владимирович – канд. хим. наук, доцент, вед. науч. сотрудник. Институт биоорганической химии НАН Беларуси (ул. Купревича, 5/2, 220084, Минск, Республика Беларусь). E-mail: dichenko@iboch.by.

Усанов Сергей Александрович – член-корреспондент, д-р хим. наук, профессор. Институт биоорганической химии НАН Беларуси (ул. Купревича, 5/2, 220084, Минск, Республика Беларусь). E-mail: usanov@iboch.by.

Information about the authors

Shaladonova Marina I. – Master's Student. University of the National Academy of Sciences of Belarus (38B, Radialnaya Str., 220070, Minsk, Republic of Belarus). E-mail: shalmari@tut.by.

Dzichenka Yaraslau V. – Ph. D. (Chemistry), Associate Professor, Leading Researcher. Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus (5/2, Kuprevich Str., 220084, Minsk, Republic of Belarus). E-mail: dichenko@iboch.by.

Usanov Sergei A. – Corresponding Member, D. Sc. (Chemistry), Professor. Institute of Bioorganic Chemistry of the National Academy of Sciences of Belarus (5/2, Kuprevich Str., 220084, Minsk, Republic of Belarus). E-mail: usanov@iboch.by.