ISBN 978-985-21-0916-1
УДК 61:615.1(06)
ББК 5:52.8
И 66

СРАВНЕНИЕ QSAR МОДЕЛЕЙ ДЛЯ ПРЕДСКАЗАНИЯ АКТИВНОСТИ ИНГИБИТОРОВ РЕНИНА

Евстафьева В.А., Кашкур Ю.В.

Белорусский государственный медицинский университет, кафедра фармакологии, г. Минск

Ключевые слова: *ингибиторы ренина, QSAR, случайные леса, дескрипторы.*

Резюме: *в данной работе были построены QSAR модели на основе различных групп дескрипторов с помощью алгоритма случайных лесов. Полученные модели были сравнены по своему качеству. Наибольшую предсказательную способность показала комбинация дескрипторов из различных групп. Модели, построенные на дескрипторах из одних и тех же групп показали меньшую предсказательную способность.*

Resume: *in this research, QSAR models were constructed based on the random forest algorithm, which predicts the activity of renin inhibitors, depending on different groups of descriptors. The constructed models were compared in terms of their metrics of quality. The best model was based on combination of descriptors from different groups. Models based on descriptors from the same groups showed less predictive power.*

Актуальность. Сердечно-сосудистые заболевания являются основной причиной смерти и инвалидизации во всем мире. Поиск новых способов дизайна лекарственных веществ для лечения данной группы заболеваний является важной проблемой медицины. Ингибиторы ренина являются одной из перспективных групп лекарственных средств для лечения такой сердечно-сосудистой патологии как артериальная гипертензия. Ренин – это протеолитический фермент, осуществляющий гидролиз ангиотензиногена до ангиотензина I. Он является первым звеном ренин-ангиотензин-альдостероновой системы, которая участвует в регуляции артериального давления. На данный момент единственным препаратом на рынке из этой группы является алискирен. Для предсказания активности ингибиторов ренина можно использовать методы машинного обучения, что значительно облегчит поиск новых потенциальных лекарственных соединений. Большой проблемой является выбор алгоритмов и дескрипторов для обучения моделей. В данной работе рассматривается вопрос выбора дескрипторов для обучения модели.

Цель: построить несколько моделей машинного обучения на основе алгоритма случайных лесов, которые позволят предсказывать активность ингибиторов ренина, основываясь на различных дескрипторах, и сравнить полученные модели между собой.

Задачи: 1. Собрать данные об уже изученных лигандах ренина; 2. Провести обработку данных, отобрав только подходящие для построения модели лиганды; 3.Построить таблицу, включающую дескрипторы различных групп для каждого лиганда; 4. Построить модели машинного обучения на основе алгоритма случайных лесов, предсказывающие активность ингибиторов ренина на основе различных комбинаций дескрипторов; 5.Сравнить метрики качества моделей; 6. Использовать

ISBN 978-985-21-0916-1
УДК 61:615.1(06)
ББК 5:52.8
И 66

лучшую модель для предсказания активности найденных in-silico потенциальных ингибиторов ренина.

Материал и методы. Машинное обучение (machine learning, ML) – совокупность методов искусственного интеллекта, позволяющих строить алгоритмы (модели), которые способны обучаться на каких-либо данных. QSAR (Quantitative Structure–Activity Relationship) – частный случай применения машинного обучения для построения моделей, способных по химическому строению молекул предсказывать их различные свойства. QSAR для предсказания активности соединений можно использовать, как для задачи классификации (то есть отнесения молекулы к классу активных, либо неактивных соединений), так и для задачи регрессии (прогнозирования числовых показателей активности соединения). В данной работе методы машинного обучения использовались для классификации молекул на активные и неактивные. Скрипты для обработки данных и построения модели были написаны на языке программирования Python. Построение модели машинного обучения осуществлялось с помощью алгоритма случайных лесов (random forest) из программной библиотеки scikit-learn. Для отбора ингибиторов ренина использовалась база данных ChEMBL. Для предсказания активности (IC50) было отобрано 2190 соединений.

Для каждого соединения был получен набор дескрипторов. Использовались следующие группы дескрипторов:

- физико-химические (молекулярная рефракция(MR), логарифм коэффициента разделения октанол/вода (logP), относительная молекулярная масса (MW));
- топологические (индекс Балабана (BalabanJ), индекс Бертца (BertzCT));
- поверхностные дескрипторы (топологическая площадь полярной поверхности (TPSA), доступная для растворителя площадь поверхности (LabuteASA));
- структурные дескрипторы (число акцепторов водородных связей (numHBA), число доноров водородных связей (numHBD)).

Для обучения модели был выбран алгоритм случайных лесов. Случайный лес (random forest) – алгоритм машинного обучения, основанный на использовании множества решающих деревьев. Решающее дерево (дерево принятия решений, decision tree) – алгоритм машинного обучения, структура которого представляет собой “узлы” и “листья”, где каждый узел – это какое-либо условие, а “лист” – это результат, получаемый при соблюдении либо несоблюдении данного условия. Метод случайных лесов позволяет получить более точные результаты, чем использование одного решающего дерева, так как невысокое качество прогноза каждого отдельного дерева корректируется предсказаниями других деревьев.

В качестве меры активности молекул использовалась IC50. Для удобства анализа данной меры активности был получен ее логарифмический показатель - pIC50. Активными считались соединения, у которых $pIC50 \geq 6$, а неактивными - соединения, у которых $pIC50 < 6$.

Результаты и их обсуждение. Перед отбором дескрипторов для моделей была построена таблица коэффициентов корреляции Пирсона дескрипторов между собой

ISBN 978-985-21-0916-1
УДК 61:615.1(06)
ББК 5:52.8
И 66

(табл. 1), для того чтобы в последующем не комбинировать дескрипторы, которые коррелируют между собой, во избежание мультиколлинеарности.

Табл. 1. Коэффициенты корреляции Пирсона.

	MR	logP	MW	BalabanJ	BertzCT	TPSA	LabuteASA	numHBA	numHBD
MR	**1.000**	0.256	**0.979**	-0.276	0.711	0.672	**0.993**	0.610	0.571
logP	0.256	**1.000**	0.238	-0.058	0.358	-0.423	0.235	-0.411	-0.373
MW	**0.979**	0.238	**1.000**	-0.289	0.728	0.686	**0.993**	0.609	0.583
BalabanJ	-0.276	-0.058	-0.289	**1.000**	-0.012	-0.210	-0.278	-0.241	-0.160
BertzCT	0.711	0.358	0.728	-0.012	**1.000**	0.335	0.729	0.314	0.212
TPSA	0.672	-0.423	0.686	-0.210	0.335	**1.000**	0.684	0.810	**0.906**
LabuteASA	**0.993**	0.235	**0.993**	-0.278	0.729	0.684	**1.000**	0.618	0.577
numHBA	0.610	-0.411	0.609	-0.241	0.314	0.810	0.618	**1.000**	0.564
numHBD	0.571	-0.373	0.583	-0.160	0.212	**0.906**	0.577	0.564	**1.000**

Согласно полученным результатам не рекомендуется комбинировать между собой MR и MW, MR и LabuteASA, MW и LabuteASA, TPSA и numHBD.

Для проверки качества модели перед ее построением набор данных был разделен случайным образом на тестовую (30 % данных) и тренировочную (70 %) части (train_test_split, библиотека scikit-learn). Тренировочные и тестовые данные были сравнены с помощью t-теста, статистически значимых различий между выборками не обнаружено.

С помощью метода RandomForestClassifier были построены QSAR модели случайных лесов. Обучение моделей проводилось на тренировочных данных. В качестве независимых переменных передавались значения дескрипторов, в качестве зависимой – класс соединения: активное или неактивное. Для поиска оптимальных значений гиперпараметров моделей использовался метод RandomizedSearchCV, позволяющий задать определенный диапазон параметров, в рамках которых будет идти поиск лучшей модели.

Для сравнения эффективности различных моделей использовались следующие метрики: точность предсказания на тренировочных данных, точность предсказания на тестовых данных, Precision (отношение true positives к сумме true positives и false positives), Recall (отношение true positives к сумме true positives и false negatives), F1-score (двойное произведение precision и recall, деленное на их сумму). Метрики полученных моделей представлены в таблице 2.

Табл. 2. Метрики качества построенных моделей.

Дескрипторы	Точность на тренировочных данных	Точность на тестовых данных	Precision	Recall	F1-score
MR+logP+MW	0,936	0,919	0,924	0,993	0,957
BalabanJ + BertzCT	0,953	0,913	0,922	0,988	0,954
TPSA + LabuteASA	0,94	0,921	0,925	0,993	0,958
numHBD + numHBA	0,913	0,907	0,911	0,995	0,951
logP+MW+BalabanJ+BertzCT+	0,978	0,924	0,939	0,98	0,96

ISBN 978-985-21-0916-1
УДК 61:615.1(06)
ББК 5:52.8
И 66

TPSA+numHBA					
logP+MW+BalabanJ+BertzCT+ TPSA	0,954	0,932	0,934	0,995	0,964
logP+MW+BalabanJ+TPSA	0,953	0,924	0,933	0,987	0,959
logP+MW+TPSA	0,932	0,918	0,924	0,992	0,956
logP+MW+BertzCT+TPSA	0,953	0,925	0,932	0,99	0,96

Лучшей моделью оказалась модель, построенная на комбинации физико-химических (logP и MW), топологических (BalabanJ и BertzCT) и поверхностных (TPSA) дескрипторов. На следующем этапе работы полученная модель была использована для проверки активности соединений, найденных с помощью построения фармакофора.

Для построения фармакофора использовались только активные молекулы (pIC50 ≥ 6). Была проведена кластеризация активных лигандов с помощью Butina Clustering. Всего было получено 229 кластеров. Для построения фармакофора использовались 5 молекул (рис.1) из первого кластера, содержащего 194 лиганда.

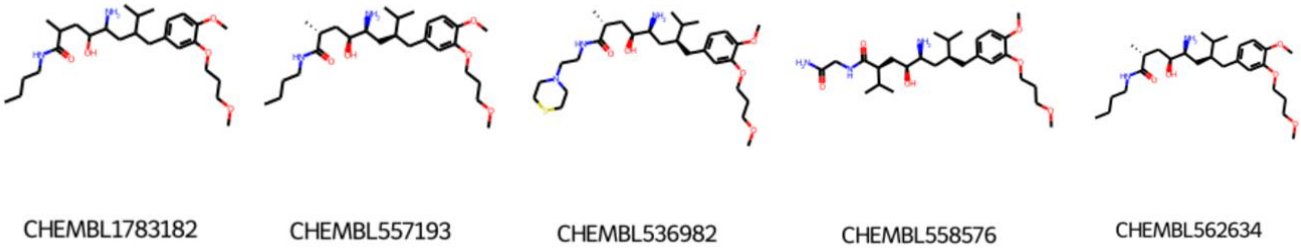

Рис. 1 – Лиганды, использованные для построения фармакофора.

Построение фармакофора проводилось с помощью алгоритма MAPex. Полученный фармакофор (рис. 2) имеет 6 фармакофорных центров: 2 акцептора водорода (показаны красным цветом) и 4 гидрофобных центра (желтый цвет).

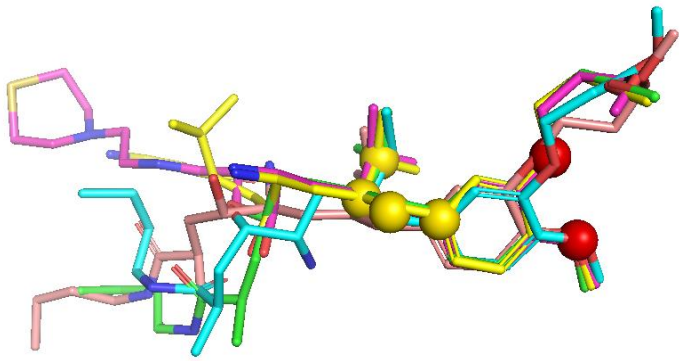

Рис. 2 – Визуальное представление выровненных молекул и фармакофорных центров.

ISBN 978-985-21-0916-1
УДК 61:615.1(06)
ББК 5:52.8
И 66

По данному фармакофору был проведен поиск в базе данных ZINCPharmer, было найдено 119 молекул. Далее был проведен анализ данных соединений на предмет совпадения с лигандами, которые уже анализировались. Оставшиеся 87 соединений были проверены на соответствие правилам Липински. 60 соединений, которые удовлетворяли критериям Липински были проверены по фильтру PAINS (Pan-assay interference compounds). Через фильтр прошло 58 соединений. Далее они были проверены на предмет наличия нежелательных структур (длинных алифатических цепей, нитрогрупп и т.п.). “Чистых” структур осталось 34.

Для оставшихся 34 соединений были получены дескрипторы. С использованием QSAR модели, которая была построена ранее, были предсказаны классы найденных молекул. Среди данных молекул 25 соединений оказались активными.

Выводы: 1. QSAR моделирование – это современный способ анализа свойств потенциальных лекарственных веществ, основанный на методах машинного обучения; 2. В ходе данной работы были построены QSAR модели на основе различных дескрипторов, которые позволяют достаточно точно предсказывать активность ингибиторов ренина; 3. Код для построения моделей доступен на GitHub: https://github.com/walking-chaos/QSAR_random_forest.git; 4. Данная модель была применена для предсказания активности найденных с помощью фармакофора молекул. Среди этих молекул потенциально активными (согласно построенным моделям) и удовлетворяющими критериям лекарственных веществ являются 25 соединений.

Литература

1. Программирование на Python [Электронный ресурс] / Stepik. – Режим доступа: https://stepik.org/67 (дата обращения: 05.02.21).

2. Введение в Data Science и машинное обучение [Электронный ресурс] / Stepik. – Режим доступа: https://stepik.org/4852 (дата обращения: 05.02.21).

3. Баскин, И. И. Введение в хемоинформатику : учеб. пособие. Ч. 3. Моделирование «структура-свойство» / И. В. Баскин, Т. И. Маджидов, А. А. Варнек. – Казань: Изд-во Казан. ун-та, 2015. – 304с.