УДК: 61:615.1(043.2)
ББК: 5:52.82
А 43
ISBN: 978-985-21-0765-5

Евстафьева В. А.

ПОСТРОЕНИЕ QSAR МОДЕЛИ ДЛЯ ПРЕДСКАЗАНИЯ АКТИВНОСТИ ИНГИБИТОРОВ РЕНИНА

Научный руководитель: ассист. Кашкур Ю. В.
Кафедра фармакологии
Белорусский государственный медицинский университет, г. Минск

Актуальность. Ренин – это протеолитический фермент, осуществляющий гидролиз ангиотензиногена до ангиотензина I. Он является первым звеном ренин-ангиотензин-альдостероновой системы, участвуя в регуляции артериального давления. Ингибиторы ренина - одна из перспективных групп лекарственных средств для лечения артериальной гипертензии. На данный момент единственным препаратом на рынке из этой группы является алискирен. Для предсказания активности ингибиторов ренина можно использовать методы машинного обучения, что значительно облегчает поиск новых потенциальных лекарственных соединений.

Цель: построить модель машинного обучения на основе алгоритма “случайных лесов” (random forest), которая позволит предсказывать активность ингибиторов ренина, основываясь на структуре молекул.

Материалы и методы. Для отбора ингибиторов ренина использовалась база данных ChEMBL. Всего было найдено 5154 лиганда, из которых после обработки данных осталось 2190 соединений, на основе которых и была построена модель машинного обучения. Скрипты для обработки данных и построения модели были написаны на языке программирования Python. Для 1D представления структуры молекул использовался генератор фингерпринтов (FingerprintGenerator) из библиотеки RDkit. Построение модели машинного обучения осуществлялось с помощью алгоритма “случайных лесов” (random forest) из программной библиотеки scikit-learn.

Результаты и их обсуждение. В качестве меры активности соединений в данной работе использовалась IC50. После обработки данных, полученных из базы данных ChEMBL, было отобрано 2190 уникальных соединений, для которых в качестве меры активности в базе данных была указана IC50. Для упрощения анализа данной меры активности был получен ее логарифмический показатель - pIC50. Активными считались соединения, у которых $pIC50 \geq 6$, а неактивными - соединения, у которых $pIC50 < 6$.

Для одномерного представления молекул использовались фингерпринты (RDkit fingerprints), полученные с помощью генератора фингерпринтов. Набор данных был разделен случайным образом на тестовую (30 % данных) и тренировочную (70 %) части (train_test_split, библиотека scikit-learn). Далее была построена QSAR модель “случайных лесов”. Обучение модели проводилось на тренировочных данных. В качестве независимых переменных передавались значения фингерпринтов, в качестве зависимой – класс соединения: активное или неактивное.

Для поиска оптимальных значений гиперпараметров модели использовался метод RandomizedSearchCV. Лучшей оказалась модель с числом “деревьев” (n_estimators), равным 14, глубиной “деревьев” (max_depth), равной 19, и минимальным количеством соединений для разделения узла “дерева” (min_samples_split), равным 17. При этом средняя точность предсказаний на тренировочных данных составила 94,72%, а на тестовых – 94,82%. Метрики качества модели составили: precision - 0,949; recall - 0,997; f1-score - 0,972.

Выводы. С помощью методов машинного обучения была построена QSAR модель “случайных лесов”, предсказывающая активность ингибиторов ренина на основе особенностей их структуры. Точность модели составляет около 94,8 %. Данную модель можно использовать для in-silico поиска новых потенциальных ингибиторов ренина и предсказания их активности. Модель доступна на GitHub.