# СРАВНЕНИЕ ЭФФЕКТИВНОСТИ ПРЕДИКТОРОВ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ В МОДЕЛИРОВАНИИ ОТНОШЕНИЯ «СТРУКТУРА – АКТИВНОСТЬ»

**Б. Н. Кузиев**

*Джизакский политехнический институт, Республика Узбекистан*

**Р. Р. Давронов, Ф. Т. Адылова**

*Институт математики АН РУЗ, г. Ташкент, Республика Узбекистан*

**Б. А. Абдурахмонов**

*Ташкентский фармацевтический институт, Республика Узбекистан*

Виртуальный скрининг (VS) является распространенным и эффективным подходом к открытию новых соединений. VS-методы классифицируются как методы на основе лиганда (LBVs) и на основе структуры (SBVS) в зависимости от наличия кристаллических структур для интересующей цели. Любой инструмент LBVs основан на принципе сходства, т. е. соединения со сходными химическими структурами, которые, как ожидается, имеют сходные биологические свойства. Тогда можно прогнозировать специфическую биологическую активность молекулы химически подобных соединений, для которых уже известны активности [1]. Два основных подхода LBVs включают поиск соединений на основе химического сходства [2] и предсказания на основе QSAR [3], такие как метод и программное обеспечение PASS.

Целью работы является апробация применения различных наборов дескрипторов в режиме kNN-QSAR, в том числе дескрипторов SiRMS, позволяющих интерпретировать построенную модель; показан пример интерпретации на исследуемом наборе соединений.

**Материал и методы.** В качестве данных для исследования были взяты 90 нитросоединений и значения их токсичности. Для вычислительных экспериментов данные были представлены в формате .sdf и стандартизованы программой Chemaxon.

**Результаты и обсуждение.** Все вычислительные эксперименты (ВЭ) проводились в рамках метода kNN-QSAR. Для этого 90 соединений разделяли на обучающую выборку (48 соединений), тестовую (22 соединения) и внешнюю (20 соединений). В качестве статистических критериев достоверности моделей были использованы стандартные статистические критерии. *Первый вычислительный эксперимент* (*ВЭ1*)

имел целью исследовать различные наборы дескрипторов, генерируемые программой Rcdk, из которых процедурой (Simulated Annealing) отбираются разные по числу наборы дескрипторов, на которых строятся регрессионные модели. В зависимости от разных наборов дескрипторов были получены около 10 моделей, из которых 2 согласно критериям kNN-QSAR можно считать приемлемыми. Из табл. 1 видно, что модель № 2 можно считать наилучшей.

*Таблица 1*

**Модели kNN-QSAR на дескрипторах Rcdk**

| Модель | $q^2$ | $R^2$ | RMSE | *F* | *p-value* | MAE | Число дескрипторов |
|---|---|---|---|---|---|---|---|
| 1 | 0,5170509 | 0,631 | 0,0004 | 24307,270 | $1{,}019853e^{-05}$ | 0,4557880 | 8 |
| **2** | **0,5333849** | **0,782** | **0,0105** | **1280,645** | $\mathbf{4{,}723404e^{-08}}$ | **0,4750792** | **16** |

Во *втором вычислительном эксперименте* (*ВЭ2*) использовали другие системы генерации дескрипторов – Dragon, Sirms и их комбинации с системой генерации Rcdk. В табл. 2 дана одна модель из многих, удовлетворяющая критериям приемлемости kNN-QSAR (дескрипторы Dragon).

*Таблица 2*

**Модель kNN-QSAR на дескрипторах Dragon**

| $q^2$ | $R^2$ | RMSE | *F* | *p-value* | MAE | Число дескрипторов |
|---|---|---|---|---|---|---|
| 0,6791584 | 0,649 | 0,0109 | 822,346 | $6{,}02503e^{-06}$ | 0,4136384 | 16 |

В табл. 3 представлена одна модель, построенная на дескрипторах Sirms, удовлетворяющая критериям kNN-QSAR.

*Таблица 3*

**Модели kNN-QSAR на дескрипторах Sirms**

| Модель | $q^2$ | $R^2$ | RMSE | *F* | *p-value* | MAE | Число дескрипторов |
|---|---|---|---|---|---|---|---|
| 1 | 0,7730982 | 0,655 | 0,0447 | 176,336 | $5{,}084503e^{-06}$ | 0,4163070 | 16 |

Из табл. 3 видно, что наилучшей является модель 2, построенная на 18 дескрипторах.

*Третий вычислительный эксперимент* был проведен на дескрипторах Sirms.

Исходные 90 соединений в формате sdf и их активности в форме log(1/C) были загружены в программу SPCI с целью получить структурную интерпретацию. Были построены четыре модели регрессии с использованием методов Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting Regression (GBR), Partial least Squares (PLS). В табл. 4 приведены их статистические характеристики. Используя эти модели, были найдены вычисленные значения активностей каждого фрагмента. Пусть минимальное количество фрагментов равно *N*, минимальное количество молекул, содержащих один и тот же фрагмент, равно *M*. Здесь мы положим $M = N = 10$.

*Таблица 4*

**Модели регрессии на дескрипторах Sirms**

| Модель | $R^2$ | RMSE | MAE |
|---|---|---|---|
| GBM | 0,28 | 0,74 | 0,55 |
| **RF** | **0,44** | **0,66** | **0,43** |
| SVM | 0,27 | 0,75 | 0,56 |
| PLS | 0,17 | 0,80 | 0,64 |

Из табл. 4 видно, что лучшей по определенности вкладов фрагментов является модель RF.

Таким образом, данное исследование еще раз подтвердило необходимость выбора подходящей системы дескрипторов в каждом конкретном случае, что неоднократно подчеркивалось и другими авторами. Кроме этого в работе показан пример интерпретации построенных моделей.

Литература

1. Адылова, Ф. Т. Сравнение компьютерных предикторов биологической активности органических соединений (аналитический обзор) / Ф. Т. Адылова // Проблемы вычислит. и приклад. математики. – 2017. – № 2. – С. 76–81.
2. Tropsha, A. Golbraikh / A. Tropsha // Curr. Pharm. Des. – 2007. – № 13. – P. 3494–3504.
3. Structural and Physico-Chemical nterpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis / P. Polishchuk [et al.] // J. Chem. Inf. Model. – 2016. – № 56. – P. 1455−1469.