https://doi.org/10.47470/0869-7922-2023-31-6-413-417
Оригинальная статья

Гусева Е.А.

Проблема отбора релевантных дескрипторов при прогнозировании токсичности химических веществ

Федеральное государственное автономное образовательное учреждение высшего образования Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет), 199911, Москва, Российская Федерация

РЕЗЮМЕ

Введение. Математические модели широко применимы при проведении токсикологических исследований и могут использоваться для заполнения пробелов, возникающих при оценке химической безопасности. Большая часть внимания уделяется вопросам изучения алгоритмов построения моделей, а не подходам к выбору наиболее информативных признаков. Поэтому, *цель* настоящей работы – осветить аспекты проблемы выбора полезных переменных при проведении математического моделирования.

Материал и методы. В интерактивной среде Google Colaboratory на основании программного кода при помощи обеспечения RDKit, Mordred были сгенерированы SMILES и молекулярные дескрипторы для фосфорорганических инсектицидов. С помощью инструментов библиотеки scikit-learn Ver. 1.2.2 происходил отбор признаков методом фильтрации и методом рекурсивного исключения признаков. Из официальных информационных источников о химических веществах были взяты значения параметров острой пероральной токсичности. Полученные модели прошли процедуру внутренней валидации, проведена сравнительная оценка производительности моделей.

Результаты. Необходимо отметить, что модели, где использовалось рекурсивное исключение признаков, обладают лучшими характеристиками, чем модели на основе дескрипторов, отобранных методом фильтрации. В частности, модель прогнозирования острой токсичности для органотиофосфатов на основе метода дерева принятия решения с рекурсивным исключением признаков обладает высоким коэффициентом детерминации (R^2=0,91713), сравнительно небольшой среднеквадратичной ошибкой (RMSE= 0,35099), а также высоким значением коэффициента детерминации кросс-валидации (Q^2LOO= 0,79756).

Ограничения исследования. Полученные результаты могут быть использованы только при прогнозировании токсичности указанной группы химических веществ со сходным механизмом действия.

Заключение. Использование математического моделирования – перспективный инструмент оценки токсичности химических веществ, имеющий ряд особенностей: с одной стороны, это быстрый и удобный ресурс для проведения скрининга токсичности веществ, с другой – модель необходимо обучить на основе не только надежных данных исследований, но и провести процедуру качественного отбора признаков, вносящих значительный вклад в функционирование прогностической модели.

Ключевые слова: *токсичность; прогнозирование; дескрипторы*

Соблюдение этических стандартов. Исследование не требует представления заключения комитета по биомедицинской этике или иных документов.

Для цитирования: Гусева Е.А. Проблема отбора релевантных дескрипторов при прогнозировании токсичности химических веществ. *Токсикологический вестник.* 2023; 31(6): 413–417. https://doi.org/10.47470/0869-7922-2023-31-6-413-417

Для корреспонденции: *Гусева Екатерина Андреевна*, ассистент кафедры экологии человека и гигиены окружающей среды Института общественного здоровья им. Ф.Ф. Эрисмана ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет), 199911, Москва, Россия. E-mail: guseva_e_a@staff.sechenov.ru

Конфликт интересов. Авторы заявляют об отсутствии конфликтов интересов.
Финансирование. Исследование не имело спонсорской поддержки.

Дата поступления: 21 сентября 2023 / Дата принятия к печати: 03 декабря 2023 / Дата публикации: 29 декабря 2023

https://doi.org/10.47470/0869-7922-2023-31-6-413-417
Original article

Guseva E.A.

The problem of selecting relevant descriptors in predicting the toxicity of chemicals

Federal State Autonomous Educational Institution of Higher Education I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, 199911, Russian Federation

ABSTRACT

Introduction. Mathematical models are widely applicable in conducting toxicological studies and can be used to fill gaps that arise in the assessment of chemical safety. Most of the attention is paid to the study of algorithms for constructing models, rather than approaches to choosing the most informative features.

The purpose of this study is to highlight aspects of the problem of choosing useful variables during mathematical modeling.

Material and methods. SMILES and molecular descriptors for organothiophosphates were generated in the interactive Google Colaboratory environment based on the program code using the RDKit, Mordred software. Using the tools of the scikit-learn Ver. 1.2.2 library, features were selected by filtering and by recursive feature exclusion. The values of acute oral toxicity parameters were taken from official information sources about chemicals. The obtained models are subjected to an internal validation procedure to evaluate the performance of the models.

Results. It should be noted that models where recursive exclusion of features was used have better characteristics than models based on descriptors selected by the filtering method. In particular, the acute toxicity prediction model for organothiophosphates based on the decision tree method with recursive exclusion of features has a high coefficient of determination (R^2=0,91713), a relatively small root-mean-square error (RMSE= 0,35099), as well as high values of the cross-validation coefficient of determination (Q^2LOO= 0,79756).

Limitations. The results obtained can be used only in predicting the toxicity of the specified group of chemicals with a similar mechanism of action.

Conclusion. The use of mathematical modeling is a promising tool for assessing the toxicity of chemicals, which has a number of features: on the one hand, it is a quick and convenient resource for screening the toxicity of substances, on the other hand, the model needs to be trained based not only on reliable research data, but also to carry out a qualitative selection procedure for signs that make a significant contribution to the functioning of the prognostic model.

Keywords: *toxicity; prediction; descriptors*

Compliance with ethical standards. The study does not require the submission of the conclusion of the Biomedical ethics committee or other documents.

For citation: Guseva E.A. The problem of selecting relevant descriptors in predicting the toxicity of chemicals. *Toksikologicheskiy vestnik / Toxicological Review.* 2023; 31(6): 413–417. https://doi.org/10.47470/0869-7922-2023-31-6-413-417 (In Russian)

For correspondence: *Ekaterina A. Guseva*, Assistant of the Department of Human Ecology and Environmental Hygiene of the Institute of Public Health named after F.F. Erisman, Sechenov First Moscow State Medical University of the Ministry of Health of Russia (Sechenov University), Moscow, 199911, Russian Federation. E-mail: guseva_e_a@staff.sechenov.ru

Conflict of interest. Author declare no conflict of interest.
Funding. The study had no sponsorship.

Date of receipt: September 21, 2023 / Date of acceptance for printing: December 3, 2023 / Date of publication: December 29, 2023

https://doi.org/10.47470/0869-7922-2023-31-6-413-417
Оригинальная статья

Введение

Математические модели широко применимы при проведении токсикологических исследований: они могут использоваться в качестве инструмента прогнозирования токсичности химических веществ для заполнения пробелов, возникающих при оценке химической безопасности в условиях недостатка данных. Существуют исследования, подтверждающие то, что значения показателей острой токсичности, полученные на основании компьютерного моделирования и в эксперименте на животных, являются достаточно близкими при соответствующем пути введения [1].

Одним из доминирующих, перспективных и наиболее разработанных методов прогнозирования токсичности является метод количественного соотношения структура-активность, КССА (англ. *QSAR* – quantitative structure-activity relationship) [2–5]. Разработку модели для прогнозирования некоторого заданного свойства химического вещества на основе уже имеющейся информации можно описать как математическую задачу аппроксимации функции отображения *(f)* от входных переменных *(X)* к выходным переменным *(y)*.

Этапы разработки модели прогнозирования токсичности включают:

- (1) сбор значений и обработка информации о «конечных точках» токсичности;
- (2) генерация молекулярных дескрипторов;
- (3) разработка моделей прогнозирования «конечных точек» токсичности при помощи методов машинного обучения;
- (4) оценка и валидация моделей [6, 7].

В настоящее время большая часть внимания уделяется вопросам изучения алгоритмов построения моделей, а не подходам к селекции наиболее информативных дескрипторов для целей моделирования. Поэтому *цель* настоящей работы – осветить аспекты проблемы выбора «полезных» переменных при проведении математического моделирования.

Материал и методы

Для исследования была взята информация о группе веществ, относящихся к классу органотиофосфатов – группе органических соединений пятивалентного фосфора, применяемых в качестве инсектицидов. В интерактивной среде Google Colaboratory на основании программного кода при помощи программного обеспечения RDKit, Mordred были сгенерированы SMILES и молекулярные дескрипторы для веществ [8]. При помощи инструментов библиотеки scikit-learn Ver. 1.2.2 происходил отбор признаков методом фильтрации и методом рекурсивного исключения признаков. Из официальных информационных источников о химических веществах были взяты значения параметров острой пероральной токсичности (DL_{50}) и выражены в виде $lg(1/DL_{50})$ (моль/кг) в соответствии со стандартной процедурой для построения прогностических моделей [9]. Полученные модели подвергнуты процедуре внутренней валидации для оценки производительности [10–12].

Результаты

Было вычислено 1826 независимых переменных для каждого соединения из набора. Однако такое количество дескрипторов совершенно излишне при моделировании токсичности веществ небольшой выборки и может привести к переобучению модели. Поэтому нами были выбраны 2 наиболее распространённых метода выбора признаков: метод фильтрации и метод рекурсивного исключения признаков. Метод фильтрации заключался в удалении дескрипторов с нечисловыми значениями; с константными или полуконстантными значениями; дескрипторов, имеющих хотя бы одно пропущенное значение; дескрипторов автокорреляции. После этого из набора признаков исключались дескрипторы со стандартным отклонением менее 0,01, а также дескрипторы, имеющие высокую взаимную корреляцию.

Метод рекурсивного исключения включал в себя часть «фильтрационного» этапа в области удаления дескрипторов с нечисловыми и/или постоянными, и/или пропущенными значениями. Размерность полученных признаков уменьшалась рекурсивным исключением признаков с использованием ансамблевого метода Gradient Boosting Regressor. В таблице представлены основные метрики внутренней валидации математических моделей при разных методах отбора значимых признаков (в круглых скобках указаны целевые значения показателей).

Как видно из таблицы, в случае применения метода фильтрации были отобраны такие дескрипторы, как ABC (индекс связности между атомами, англ. «atom-bond connectivity index»), nS (число атомов S, англ. «number of S atoms»), BCUTZ-1h (первое по величине собственное значение матрицы Бёрдена, взвешенное по атомному номеру, англ. «first heighest eigenvalue of Burden matrix weighted by atomic

https://doi.org/10.47470/0869-7922-2023-31-6-413-417
Original article

Процедура внутренней валидации моделей, полученных на основе разных подходов к отбору «полезных» дескрипторов

The procedure of internal validation of models obtained on the basis of different approaches to the selection of "useful" descriptors

Метод отбора	Отобранные дескрипторы	Модель	Коэффициент детерминации, R^2 (≥0,6)	Среднеквадратичная ошибка, RMSE, (RMSE →0)	Коэффициент детерминации кроссвалидации, Q^2LOO (≥0,5)
Фильтрация	ABC nS BCUTZ-1h BCUTi-1h	Линейная регрессия	0,0795	1,16986	2,32047
		Метод к-ближайших соседей	–0,211	1,34217	1,85127
		Метод дерева принятия решений	0,0462	1,19083	1,75576
Рекурсивное исключение признаков	BCUTc-1h MINsCH3 AETA_beta_s	Линейная регрессия	0,65050	0,80683	0,65566
		Метод к-ближайших соседей	0,56214	0,70136	1,83621
		Метод дерева принятия решений	0,91713	0,35099	0,79756

number»), BCUTi-1h (первое по величине собственное значение матрицы Бёрдена, взвешенное по потенциалу ионизации, *англ*. «first heighest eigenvalue of Burden matrix weighted by ionization potential»); при применении метода рекурсивного исключения признаков – BCUTc-1h (первое по величине собственное значение матрицы нагрузки, взвешенное по заряду Гастайгера, англ. «first heighest eigenvalue of Burden matrix weighted by Gasteiger charge»), MINsCH3 (минимальное значение sCH3, *англ*. «min of sCH3»), AETA_beta_s (усредненный сигма-вклад в количество подвижных валентных электронов, англ. «averaged sigma contribution to valence electron mobile count»).

Модель прогнозирования острой токсичности для органотиофосфатов на основе метода дерева принятия решения с рекурсивным исключением признаков обладает высоким коэффициентом детерминации (R^2 = 0,91713), сравнительно небольшой среднеквадратичной ошибкой (RMSE = 0,35099), а также высоким значением коэффициента детерминации кросс-валидации (Q^2LOO = 0,79756). Модель на основе дерева принятия решений объясняет 91,7% изменения дисперсии десятичного логарифма $1/DL_{50}$.

Обсуждение

При сравнении полученных значений параметров внутренней валидации можно отметить, что модели, в которых релевантные дескрипторы были отобраны на основе метода рекурсивного исключения признаков, обладают лучшими характеристиками производительности на тренировочном наборе данных, чем модели на основе дескрипторов, отобранных методом фильтрации.

Ограничения исследования. Полученные результаты могут быть использованы только при прогнозировании токсичности указанной группы химических веществ со сходным механизмом действия.

Заключение

Использование математического моделирования – перспективный инструмент оценки токсичности химических веществ, имеющий ряд особенностей: с одной стороны, это быстрый и удобный ресурс для проведения скрининга токсичности веществ, с другой – модель необходимо обучить на основе не только надежных данных исследований, но и провести процедуру качественного отбора признаков, вносящих значительный вклад в функционирование прогностической модели.

ЛИТЕРАТУРА

(пп. 2–5, 7–12 см. в References)

1. Сухачев В.С., Иванов С.М., Филимонов Д.А., Поройков В.В. Альтернативные методы исследования. Компьютерная оценка острой токсичности для грызунов. *Лабораторные животные для научных исследований*. 2019; 4. https://doi.org/10.29296/2618723X-2019-04-04
6. Жолдакова З.И., Харчевникова Н.В. Система ускоренной оценки токсичности и опасности химических веществ в воде. *Здоровье населения и среда обитания*. 2014; 8(257): 21–3.

REFERENCES

1. Suhachev V.S., Ivanov S.M., Filimonov D.A., Porojkov V.V. Al'ternativnye metody issledovaniya. Komp'yuternaya ocenka ostroj toksichnosti dlya gryzunov. Laboratornye zhivotnye dlya nauchnyh issledovanij. 2019; 4. https://doi.org/10.29296/2618723X-2019-04-04 (in Russian)
2. Gramatica P., Papa E., Sangion A. QSAR modeling of cumulative environmental endpoints for the prioritization of hazardous chemicals. *Environmental Science: Processes & Impacts*. 2018; 20(1): 38–47. https://doi.org/10.1039/C7EM00519A
3. Carrio P., Sanz F., Pastor M. Towards a unifying strategy for predicting toxicological endpoints based on structure. *Archive of Toxicology*. 2016; 90: 2445–460. https://doi.org/10.1007/s00204-015-1618-2
4. Reyes A.B., Bayich V.B. In silico toxicology: computational methods for predicting chemical toxicity. *Interdisciplinary reviews of Wiley: Computational Molecular Science*. 2016; 6(2): 147–72. https://doi.org/10.1002/wcms.1240

https://doi.org/10.47470/0869-7922-2023-31-6-413-417
Оригинальная статья

5. Villaverde J.J., Sevilla-Moran B., Lopez-Goti S., Alonso-Prados J.L., Sandin-España P. QSAR/QSPR models based on quantum chemistry for assessing the risk of pesticides in accordance with current European legislation. *SAR and QSAR in environmental studies*. 2020; 31(1): 49–72. https://doi.org/10.1080/1062936X.2019.1692368
6. Zholdakova Z.I., Harchevnikova N.V. Sistema uskorennoj ocenki toksichnosti i opasnosti himicheskih veshchestv v vode.Zdorov'e naseleniya i sreda obitaniya. 2014, 8(257): 21–3. (in Russian)
7. Idakwo G., Luttrell J., Chen M., et al. A review on machine learning methods for in silico toxicity prediction. *Journal of environmental science and health. Part C. Environmental carcinogenesis & ecotoxicology reviews*. 2018; 36(4): 169–91. https://doi.org/10.1080/10590501.2018.153711
8. Moriwaki H., Tian Y.S., Kawashita N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* 2018; 10, 4. https://doi.org/10.1186/s13321-018-0258-y
9. Gallagher M.E. Toxicity testing requirements, methods and proposed alternatives. *Environmental Law and Policy Journal.* 2003; 26(2): 257–73.
10. Dearden J.C., Cronin M. T.D., Kaiser K.L.E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research.* 2009; 20(3–4): 241–66. https://doi.org/10.1080/10629360902949567
11. Golbraikh A., Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design*. 2002; 16(5/6): 357–69. https://doi.org/10.1023/A:1020869118689
12. Frimayanti N., Yam M.L., Lee H.B., Othman R., Zain S.M., Rahman N.A. Validation of Quantitative Structure-Activity Relationship (QSAR) Model for Photosensitizer Activity Prediction. *International Journal of Molecular Sciences*. 2011; 12(12): 8626–44. https://doi.org/10.3390/ijms12128626

ОБ АВТОРАХ:

Гусева Екатерина Андреевна *– ассистент кафедры экологии человека и гигиены окружающей среды Института общественного здоровья им. Ф.Ф. Эрисмана, ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет), 199911, Москва, Россия. E-mail: guseva_e_a@staff.sechenov.ru https://orcid.org/0000-0001-8389-7981*

INFORMATION ABOUT THE AUTHORS:

Ekaterina A. Guseva *– Assistant of the Department of Human Ecology and Environmental Hygiene of the Institute of Public Health named after F.F. Erisman, Sechenov First Moscow State Medical University of the Ministry of Health of Russia (Sechenov University), Moscow, 199911, Russian Federation. E-mail: guseva_e_a@staff.sechenov.ru https://orcid.org/0000-0001-8389-7981*