Active learning efficiently converges on rational limits of toxicity prediction and identifies patterns for molecule design(Abstract_要旨)

AUTHOR(S):

Ahsan, Habib Polash

CITATION:

Ahsan, Habib Polash. Active learning efficiently converges on rational limits of toxicity prediction and identifies patterns for molecule design. 京都大学, 2021, 博士(医学)

ISSUE DATE:

2021-03-23

URL:

https://doi.org/10.14989/doctor.k23092

RIGHT:

https://doi.org/10.1016/j.comtox.2020.100129

京都大学	博士（医学）	氏名	Ahsan Habib Polash
論文題目	Active learning efficiently converges on rational limits of toxicity prediction and identifies patterns for molecule design （能動的機械学習による、化学構造から毒性を予測する手法の開発、および、予測能力の限界を合理的に説明する研究）		

（論文内容の要旨）

Government organizations utilizes different assays to assess the safety of chemicals. Most of the established assays have potential drawbacks which includes but are not limited to: lack of cost effectiveness, long evaluation times, false negative results, so forth. Moreover, animal-based assays are increasingly becoming discouraged by different animal welfare organizations. As a consequence, toxicologists are encouraging the development of new types of toxicity detection assays which would overcome the drawbacks of the current toxicity assays.

A number of recent scientific studies have employed machine learning (ML) to predict binding of compounds to proteins. Such studies are accelerating computational chemistry, methods in toxicology, and usage of Artificial Intelligence in drug discovery. In 2017, Reker *et al* employed a ML method called Active Learning (AL) to model several kinases and G-protein coupled receptor proteins with high performances. One advantage of AL is that, instead of learning all of the available bioactivity data, it iteratively selects a subset of the data and builds a reduced-size model that is as good as a model constructed from the whole dataset. By doing this AL can reduce computational resources substantially. In 2019, Polash and co-workers employed AL in predicting highly selective inhibitors for matrix metalloproteinases, a protein family known to play various roles in cancer cell proliferation. They also deconvoluted ML model's decision-making process.

Following the elucidation of a ML model architecture, in this study AL was applied to a dataset of approximately 9000 compounds which were tested for acute oral toxicity in rats. The data have been curated by the US government and made publicly available for the evaluation and development of predictive methodologies. Particularly notable is the fact that compared to biochemical assay data, sources of in-vivo toxicity are diverse, and thus the predictive challenge is amplified compared to biochemical assay data. Unlike many previous studies with mathematically complex ML algorithms and full activity dataset, this study showed that only a strategically subset of data was sufficient to build a model that could predict toxic compounds with high performance.

Instead of developing a `black box` model which lacks insight into model building steps, the authors tried to deconvolute the decision-making steps. In depth analyses showed that some of the compounds were predictable from the early stages of model building, whereas some compounds became predictable gradually. However, some of the compounds never could be correctly predicted; subsequent analysis revealed that these compounds frequently formed a "toxicity cliff". A toxicity cliff can be described as a minor change in structure leading to a large change in toxicity (activity). Apparently, it was found that some of the toxic compounds in the validation data have nearest neighbors in training data that are not toxic and as a result the model failed to classify them accurately; this suggests the rational limits of toxicity prediction. Furthermore, it has been shown that the removal of the compounds from the data that had toxicity values near the borderline separating toxic and non-toxic classification yielded even higher performance. Finally, compound structure analysis revealed that some compound substructures are differentially present or absent across toxic and nontoxic compounds.

In summary, the study demonstrated efficient selection of compounds toward generating computational models for toxicity prediction in rats and provided insights about the chemical substructures and patterns that are crucial for classifying toxic compounds. Computational toxicity prediction is still a fairly nascent discipline with a variety of challenges. However, insights from this research will contribute further to develop better toxicity prediction models and the understanding of chemical fragment-toxicity association will contribute to the flagging of risk-associated structures for drug discovery to avoid potential toxicity.

（論文審査の結果の要旨）

毒性判定試験を目的とした実験動物には、倫理的な懸念がある。そのため実験動物をしない毒性判定試験手法の開発が望まれている。化学物質の毒性を予測する計算モデル（シミュレーション）は、実験動物を減らす（代替する）ことが可能な手法として注目を集めている。

本研究では、アクティブラーニング（AL）法を採用し、米国国立環境保健科学研究所（NIEHS）が提供された in vivo ラット経口急性毒性の化学物質データセット(約 9,000 種類)をコンピューターに機械学習させて毒性予測モデルを構築した。また、Matthews correlation coefficient 値を用いて、構築したモデルの予測の正確性を先行研究の機械学習（ML）法を採用したモデルと比較評価した。その結果、AL 法を用いて構築した毒性予測モデルは、データセットの 20−40%だけを用いて、すなわち、より少ない計算量で、既存の ML 法を用いて構築した毒性予測モデルと同等の正確性を示した。加えて、本研究では、従来の ML 法によるモデルでは不可能である、AL 法の決定過程の背後にあるロジックを解析することによって、毒性を示す化学物質の持つ共通の特徴量(common bits)を見つけることができた。

以上のように、本研究で提案した AL 法を用いた毒性予測モデル構築アプローチは、化学物質の学習における優先順位付けを可能とし、より良い毒性予測モデル構築に貢献する。

したがって、本論文は博士（　医学　）の学位論文として価値あるものと認める。

なお、本学位授与申請者は、令和２年１１月６日実施の論文内容とそれに関連した試問を受け、合格と認められたものである。

要旨公開可能日：　　年　　月　　日　以降