结合机器学习与代谢组学的方法再分析农药暴露对环境归趋、生态毒性和人类健康的影响

Further analysis of the effects of pesticide exposure on environmental fate, ecotoxicity, and human health through a combination of machine learning and metabolomics methods

摘要

农药暴露对环境和人类的影响在近年来受到广泛关注。农药的持久性、生态毒性、对人类健康的影响不可忽视。近年来研究人员使用各种方法研究农药暴露对诸如鱼、小鼠、大鼠等模式生物的影响。然而，这些研究都是针对单个或者几个农药。本文的第一部分基于PPDB数据库建立了QSAR模型，使用不同的分子描述符建立不同的机器学习分类模型（随机森林，支持向量机，逻辑回归，朴素贝叶斯，K近邻，梯度提升决策树），通过953个农药的分子结构预测农药暴露对环境归趋、生态毒性和人类健康的影响；同时建立了使用分子描述符建立了基于7种机器学习方法（支持向量回归、多元线性回归、随机森林、偏最小二乘回归、LASSO回归、岭回归、梯度提升回归）的QSRR模型，预测了367种代谢物在HILIC色谱柱上的保留时间，得到了准确的结果，辅助鉴定代谢物，使得农药暴露对模式生物代谢通路的研究更快速准确。在第二部分的研究中，创作了一个R包meTool以处理农药代谢组学的数据，另外有一个Python软件neuturalloss寻找质谱数据的中性丢失。同时以已经发表的文章“妊娠期烯啶虫胺暴露诱导后代雌性小鼠肠道菌群和粪便代谢物失衡”为例，详细解释了meTool包的使用方法。主要研究结果如下：

1. 有了全新的发现：农药对环境归趋和人类健康的影响很大程度取决于该农药中叠氮基团的数量；而叠氮化合物是有毒物质，暴露于叠氮化物化合物会导致线粒体中细胞色素 c 氧化酶的抑制，导致急性致命中毒；并且直至现在叠氮化合物都没有解毒剂。
2. 农药的生态毒性主要取决于该农药的亲水性/亲脂性，这可能是因为PPDB数据库中的农药生态毒性由对鱼类或是藻类等水生生物的影响决定。
3. 代谢物在HILIC色谱柱上的保留时间和代谢物的分子量、分子中氮氧原子数量、分子表面上的电子状态等因素密切相关。
4. 随机森林建立的QSAR模型预测环境归趋和生态毒性的效果最好；由互信息的特征选择方式和随机森林的机器学习方法建立的QSRR模型能最好得预测367个代谢物在HILIC色谱柱上的保留时间，预测全部数据的平均绝对误差仅仅为0.616，并且仅仅用了63个分子描述符，便于解释。
5. R包meTool的优势在于：它清晰得显示了实验组和对照组数据PCA和PLS-DA等模型在主成分1和主成分2上的分布；可以和肠道菌群联合分析，这是其他代谢组学工具里没有的；比起SIMCA-P，它是免费的，且功能更多；比起R包Ropls，它作图更美观且自由度更高；比起metaboanalyst 5.0，它克服了网络卡顿的烦恼，而且同样的，自由度更高。当然，meTool也有局限性，它也有其他代谢组学工具没有的功能。
6. neuturalloss作为一款寻找质谱中性丢失的软件，它很好解决了原先寻找中丢失依赖全部人力的问题。

In recent years, the impact of pesticide exposure on the environment and human health has been widely recognized. The persistence of pesticides, ecotoxicity, and their effects on human health cannot be ignored. In recent years, researchers have used various methods to study the effects of pesticide exposure on model organisms such as fish, mice, and rats. However, these studies are mostly focused on individual or a few pesticides.

In the first part of this article, QSAR models were established based on the PPDB database. Different machine learning classification models (random forest, support vector machine, logistic regression, naive Bayes, K-nearest neighbors, gradient boosting decision tree) were built using different molecular descriptors to predict the effects of pesticide exposure on environmental fate, ecological toxicity, and human health using the molecular structures of 953 pesticides. Additionally, QSRR models were established using molecular descriptors and seven machine learning regression methods (support vector regression, multiple linear regression, random forest, partial least squares regression, LASSO regression, ridge regression, gradient boosting regression) to predict the retention times of 367 metabolites on HILIC columns. Accurate results were obtained, aiding in the identification of metabolites and facilitating faster and more accurate studies of pesticide exposure on metabolic pathways in model organisms.

In the second part of the study, an R package “meTool” was developed to process pesticide metabolomics data, along with a Python software “neuturalloss” for finding neutral losses in mass spectrometry data. Using the published article "Pregnancy pyrethroid exposure induces dysbiosis in offspring female mice gut microbiota and fecal metabolites" as an example, the usage of the meTool package was explained in detail. The main research findings are as follows:

1. A novel discovery was made: The impact of pesticides on environmental fate and human health largely depends on the number of azido groups in the pesticide. Azido compounds are toxic substances, and exposure to azides can lead to inhibition of cytochrome c oxidase in mitochondria, causing acute lethal poisoning. Furthermore, there are currently no antidotes for azido compounds.
2. The ecotoxicity of pesticides mainly depends on the hydrophilicity/lipophilicity of the pesticide, possibly because the ecological toxicity of pesticides in the PPDB database is determined by their effects on aquatic organisms such as fish or algae.
3. The retention time of metabolites on HILIC chromatographic columns is closely related to factors such as the molecular weight of the metabolite, the number of nitrogen and oxygen atoms in the molecule, and the electronic state on the molecular surface.
4. Random forest-based QSAR models performed the best in predicting environmental fate and ecotoxicity. QSRR models built using mutual information feature selection and random forest machine learning methods performed best in predicting the retention times of 367 metabolites on HILIC chromatographic columns, with an average absolute error of only 0.616, using only 63 molecular descriptors for easy interpretation.
5. The advantages of the meTool package include: It clearly displays the distribution of experimental and control group data on PCA and PLS-DA models in principal component 1 and principal component 2; it can be jointly analyzed with gut microbiota, which is not available in other metabolomics tools; compared to SIMCA-P, it is free and has more functions; compared to the R package Ropls, its plotting is more aesthetic and has higher freedom; compared to metaboanalyst 5.0, it overcomes the frustration of network lagging, and similarly, it has higher freedom. Of course, meTool also has limitations and features that other metabolomics tools do not have.
6. As a software for finding neutral losses in mass spectrometry data, neuturalloss effectively solves the problem of dependence on manual labor for finding neutral losses in the past.

关键词：农药，生态毒性，叠氮基团，HILIC色谱柱，数据处理

Keywords：Pesticides, Ecotoxicity, Azide, HILIC Column, Data Processing