

DOI:10.7524/j.issn.0254-6108.2020092304

王宇飞, 曹慧明, 梁勇. 基于膜雌激素受体 (GPER) 结合化合物能力的分类预测模型[J]. 环境化学, 2022, 41(2): 417-428.

WANG Yufei, CAO Huiming, LIANG Yong. Classification prediction model based on GPER binding ability of membrane estrogen receptor[J]. Environmental Chemistry, 2022, 41 (2): 417-428.

基于膜雌激素受体 (GPER) 结合化合物能力的 分类预测模型*

王宇飞 曹慧明** 梁 勇

(持久性有毒污染物环境与健康危害湖北省重点实验室, 环境与健康研究院, 江汉大学, 武汉, 430056)

摘 要 近年来, 计算毒理学的方法被广泛应用于潜在的环境内分泌干扰物 (EDCs) 的筛选. 膜雌激素受体 (GPER), 作为一种可以快速响应内源性配体雌激素的关键靶蛋白, 调控其介导的多项生理功能. 但是针对 GPER 的化合物毒性预测模型仍未见报道. 因此, 本研究收集了 130 个化合物对 GPER 的结合活性数据, 主要包括双酚类、多溴联苯类以及农药杀虫剂类环境污染物. 利用随机森林 (RF)、支持向量机 (SVM)、人工神经网络 (ANN)、K 最近邻 (KNN)、朴素贝叶斯 (NB) 以及逻辑回归 (LG) 等 6 种机器学习算法构建二分类模型. 结果显示, 所有被测试算法的测试集准确率均达到 85% 以上, 其中 SVM、RF、ANN、KNN 等 4 种算法的训练集准确率高于 93%, 10 折交叉验证准确率高于 80%, 说明得到的模型具有优秀的分类预测性能. 因此, 本研究基于机器学习算法构建的分类模型, 可以用来快速、准确地预测环境污染物是否通过结合 GPER 产生内分泌干扰效应. 为评估环境污染物的潜在健康风险提供了理论依据.

关键词 内分泌干扰物, 膜雌激素受体, 机器学习算法, 分类预测模型.

Classification prediction model based on GPER binding ability of membrane estrogen receptor

WANG Yufei CAO Huiming** LIANG Yong

(Hubei Key Laboratory of Environmental and Health Effects of Persistent Toxic Substances, Institute of Environment and Health, Jiangnan University, Wuhan, 430056, China)

Abstract In recent years, computational toxicology has been widely applied to the screening of potential environmental endocrine disruptors (EDCs). Membrane estrogen receptor (GPER), as a key target protein that can rapidly respond to endogenous ligand estrogen, regulates its mediated multiple physiological functions. However, the prediction model of compound toxicity for GPER has not been reported. Therefore, the binding activity data of 130 compounds against GPER were collected in this study, mainly including bisphenols, polybrominated biphenyls and environmental pollutants like pesticides and insecticides. Six machine learning algorithms, including random forest (RF), support vector machine (SVM), artificial neural network (ANN), K-nearest neighbor (KNN), Naive Bayes (NB) and logistic regression (LG), were used to construct the dichotomous model. The results showed that the test set accuracy of all the algorithms tested reached more than 85%, among which

2020 年 9 月 23 日收稿 (Received: September 23, 2020).

* 国家自然科学基金 (21806058) 资助.

Supported by the National Natural Science Foundation of China (21806058).

** 通信联系人 Corresponding author, Tel: 17707237430, Email: hmcao1986@126.com

the training set accuracy of SVM、RF、ANN and KNN algorithm was 93% higher than that of the four algorithms, and the 10-fold cross validation accuracy was 80% higher than that of the four algorithms, indicating that the model obtained had excellent classification prediction performance. Therefore, the classification model built based on machine learning algorithm in this study can be used to quickly and accurately predict whether environmental pollutants will produce endocrine disrupting effects by combining with GPER. It provides a theoretical basis for evaluating the potential health risks of environmental pollutants.

Keywords endocrine disruptors, membrane estrogen receptor, machine learning algorithm, classification prediction model.

环境内分泌干扰物(EDCs)^[1]是20世纪90年代以来引起广泛关注的一类新型化学物质。分为天然和人工合成两类,除铅、汞、砷等金属类外,其余均为有机化学物。按其用途可分为工业原材料、洗涤用品、化妆品、农药、塑料包装和儿童玩具等,因工业生产和使用被不断释放到空气、水体、土壤等环境介质中,可经呼吸空气、饮水、食物摄入及皮肤接触等途径进入人体,影响生物体内激素的合成、释放和代谢,甚至可与激素竞争结合相应受体,或者影响激素受体上下游的调控基因,干扰激素正常的调节功能,造成内分泌系统的紊乱,从而引发各种疾病,对人类健康产生潜在的危害和不利影响^[2]。

已有大量研究报道 EDCs 可产生类雌激素效应^[3],如邻苯二甲酸脂类、多氯联苯类、双酚类等化合物。这些 EDCs 在环境浓度下发挥着与天然雌激素类似的生物学活性,从而影响女性的生殖健康^[4]。已知的不利健康危害包括:月经周期变化、子宫内膜异位症、子宫肌瘤、多囊卵巢综合征、不孕不育、乳腺癌、子宫内膜癌及卵巢癌。此外,近年来的流行病学调查报告进一步证实女性雌激素敏感癌症发病率与环境中 EDCs 的浓度存在密切相关性。

EDCs 诱导产生的类雌激素效应主要通过经典的核雌激素受体 ER α 与 ER β 介导^[5]。随着结构生物学的深入,有报道证实存在膜雌激素受体(GPER),属于7-跨膜G蛋白偶联受体(GPCR)家族^[6]的一员,作为雌激素信号的靶蛋白,参与了雌激素在生殖、神经、内分泌、免疫和心血管系统中的介导作用,对于包括癌症在内的一系列疾病,GPER正成为一种新的治疗靶点和预后指标^[7-9]。G蛋白偶联雌激素受体(GPER)最初被称为GPR30^[10],其结构未被结晶实验确认,激动和拮抗的分子机制及其结构特征的研究仍处在初步阶段。不同于经典的核雌激素受体亚型,GPER被认为是介导快速细胞信号的媒介^[11],参与介导了雌激素快速非基因组效应^[12]。内源性化合物,如雌激素可与细胞膜上的GPER结合,激活快速细胞效应反应^[13],包括环磷酸腺苷的产生、细胞内钙离子的动员、多种激酶的激活,如细胞外信号调节激酶、肌醇磷脂3激酶、离子通道以及内皮型一氧化氮合酶等途径。影响下游效应分子在相应的靶组织中发挥其生物学效应^[14],从而导致细胞的增值与分化。类似地,环境污染物也可模拟内源性物质结合GPER^[15]。如BPA可在低浓度下激活GPER,产生较强的雌激素效应^[16]。

因此,构建快速筛查环境污染物是否结合GPER的分类预测模型,可为理论评估化合物健康风险与毒性效用提供重要的依据。虽然针对环境污染物的毒性预测,已有报道显示机器学习算法可表现出良好的分类预测性能^[17-20],但是结合GPER的小分子却未有可用的分子数据库和已知的预测模型,这限制了构建针对GPER分类预测模型的发展。为了解决这一问题,本研究系统地总结了已报道的有机小分子结合GPER的数据,并进一步评测了随机森林(random forest, RF)、支持向量机(support vector machine, SVM)、神经网络(neural network, ANN)、K-最近邻(K-nearest neighbour, KNN)、朴素贝叶斯(naive bayes, NB)、逻辑回归(logistic, LG)等6种典型机器学习算法的分类预测性能,其中基于RF算法构建的分类模型展示出了优秀的分类预测表现。

1 材料与方法(Materials and methods)

1.1 数据集构建

从文献中查询了130种对GPER检测结合能力的化合物^[6, 11, 16, 21-38],分别包括7种双酚类化合物、

30 种多溴联苯类化合物、18 种类固醇类化合物、15 种植物激素、30 种吡啶及其衍生物、10 种农药杀虫剂以及 20 种 GPER 选择性配体化合物。其中具有结合能力的化合物有 91 种; 不具有结合能力的化合物有 39 种。将所有化合物以 4 : 1 的比例随机分成训练集和测试集, 其中训练集具有结合能力的有 73 种, 不具有结合能力的有 31 种; 测试集具有结合能力的有 18 种, 不具有结合能力的有 8 种。利用训练集构建分类预测模型, 通过测试集定性评估模型的预测能力。

1.2 分子描述符筛选

利用软件 ChemDraw 程序包绘制测试化合物相应的 2D 分子结构, 随后通过 Chem 3D 程序包转化为相应的 3D 分子结构, 并对其进行能量最小化优化处理。基于优化后的 SDF 格式的分子结构文件, 通过 PaDEL 软件计算得到 1538 个 1D 和 2D 分子描述符。根据以下两个原则进行初步的描述符筛选: (1) 剔除常数和至少有一个缺失值的描述符; (2) 剔除描述符之间相关系数大于 0.9 的描述符, 最终得到 369 个分子描述符特征。随后利用 RF 算法和递归特征删减法的描述符重要性评价, 将重要性前五位的描述符保留, 作为构建机器学习分类预测模型的特征值。

1.3 分类预测模型的构建与表征

为了系统地比较不同机器学习算法的表现, 筛选最优算法构建 GPER 配体分类预测模型, 我们选择了 6 种机器学习算法进行模型的构建^[17], 分别为: (1) RF 算法, 是一种集成分类器, 它将大量的决策树拟合到一个数据集中, 然后将所有树的预测组合起来; (2) SVM 算法, 通过核函数将输入向量映射到高维特征空间, 优化特征空间内的线性划分, 构造出具有最大边距的超平面来分离不同类别的数据; (3) ANN 算法, 采用最常见的前馈反向传播算法, 其主要思想是: 输入学习样本, 使用反向传播算法对网络的权值和偏差进行反复的调整训练, 使输出的向量与期望向量尽可能地接近; (4) KNN 算法, 考虑到数据中的所有情况, 并根据局部邻域来获得分类预测。局部领域的大小由参数 k 决定, 即 k 个最近邻数据点, 而类别则由 K -最近邻邻居的多数票决定; (5) NB 算法, 是基于现有的特征并假设特征独立, 判断分类类别的概率, 进行分类; (6) LG 算法, 主要研究二分类的响应变量与相应特征值之间的相互关系, 并建立相应的预测模型。

所有的数据处理和模型建立都是基于统计编程环境 R (版本 4.0.2) 实现。分别使用 randomForest 和 caret 程序包执行特征选择。使用 randomForest、kernlab、nnet、kknn、e1071 程序包执行 RF、SVM、ANN、KNN、NB 和 LG 算法的分类模型构建。

对于分类模型的评估^[17], 将模型输出的预测分类与实际结果相比, 利用混淆矩阵产生真阳性 (true positive, TP)、真阴性 (true negative, TN)、假阳性 (false positive, FP)、假阴性 (false negative, FN), 并计算相应指标敏感性 (Sensitivity, S_n)、特异性 (Specificity, S_p)、精确度 (Accuracy, Q) 以及马修斯相关系数 (matthews correlation coefficient, MCC), 相应计算公式如下:

$$S_n = \frac{TP}{TP + FN} \quad (1)$$

$$S_p = \frac{TN}{TN + FP} \quad (2)$$

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN) \times (TN + FP) \times (TP + FN) \times (TP + FP)}} \quad (4)$$

敏感性 S 、特异性 S_p 、精确度 Q 以及马修斯相关系数 MCC 的值越高模型评价越好, 精确度在 0.9 以上为预测结果较好。

而受试者工作特征曲线 (receiver operator characteristic curve, ROC) 则是另外一个评价二分类模型的重要指标。使用 pROC 程序包生成 ROC 曲线并计算每个分类模型的训练集和预测集的 AUC 值。此外, 我们也采用 10 折交叉验证, 来评估模型的精度。最终采用 AMBIT Discover 软件计算训练集和测试集中化合物的欧几里德距离, 评价基于分子描述符的预测模型的应用域。

2 结果与讨论(Results and discussion)

采用 10 折交叉验证来评估模型的可靠性和预测性能,图 1 显示了整个分类模型构建的流程.

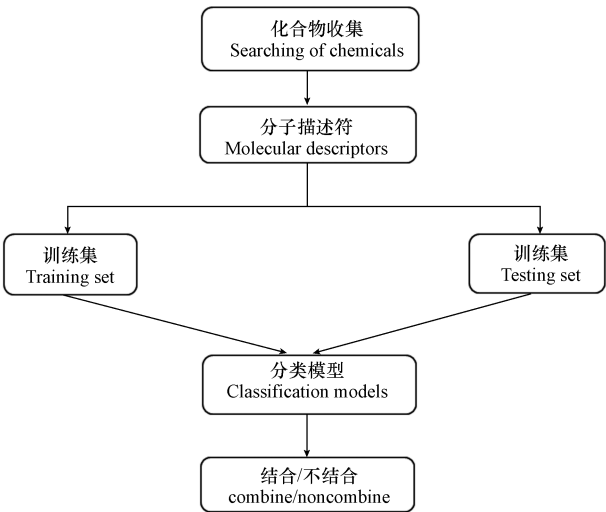


图 1 模型构建流程

Fig.1 Model building process

2.1 分类模型使用的描述符

联合 RF 算法的重要性评价和 caret 程序包的递归特征删减法进行描述符选取,得到如表 1 所示的 5 个最重要的描述符作为模型特征值.

表 1 模型选定的描述符

Table 1 Selected descriptors for models

变量 Variables	描述 Description
GATS4c	Geary autocorrelation - lag 4 / weighted by charges
GATS4s	Geary autocorrelation - lag 4 / weighted by I-state
ETA_Eta	Composite index Eta
AATSC6i	Average centered Broto-Moreau autocorrelation - lag 6 / weighted by first ionization potential
MATS2i	Moran autocorrelation - lag 2 / weighted by first ionization potential

其中,自变量 GATS4c、GATS4s、AATSC6i 和 MATS2i 为基于 Geary 自相关指数、Broto-Moreau 自相关指数以及 Moran 自相关指数的描述符. 自相关指数反应原子性质沿拓扑结构的分布,其中 GATS4c 和 GATS4s 描述符反映了化合物的静电性质对于结合的影响. AATSC6i 和 MATS2i 描述符代表了化合物第一电离能的影响,ETA_Eta 表示扩展拓扑化学原子指数对于分子结合能力的影响.

2.2 机器学习算法参数的优化

对于不同机器学习算法,采用 caret 程序包的格点搜索法进行超参数的优化,提高分类模型的统计学评价标准.

对于 RF 模型,通过调整参数 trees 和 mtry 的值,改变森林“树”的数量,从而达到对模型的优化效果. trees 表示森林“树”的数量,mtry 表示每次迭代变量抽样的数量. 最终选定 RF 模型的参数为 ntree=500、mtry=5,此时模型预测精度最优.

对于 SVM 模型,采用高斯 RBF 核函数进行训练. 参数 gamma 决定了数据映射到新的特征空间后的分布;而参数 cost 表示为惩罚因子,即对误差的容忍度. 最终选定 SVM 模型的参数为 gamma=0.35、cost=10.

针对人工神经网络, 分别对 size、decay 和 maxit 等 3 个参数进行优化, 其中 size 代表隐藏层神经元数, decay 代表输入权重的修正参数、maxit 代表最大迭代次数. 最终选择 ANN 模型的参数为 size=17、decay=0.01、maxit=1000.

对于 K-最近邻, 通过对邻居数量 k 的优化, 从而增加模型的精度. 最终选择 5 个邻居数的 KNN 算法建立模型, 精确度较高.

对于 NB 模型, 为避免由于样本数据较少而导致整个数据零概率问题的出现, 使用拉普拉斯技术做平滑处理, 避免零概率问题.

对于 LG 模型, 我们使用逐步回归的方法进行变量重要性的筛选, 得到最优模型.

2.3 模型预测及其表征

通过对模型参数的优化, 最终得到最优模型预测结果, 如表 2 所示.

表 2 模型预测结果
Table 2 Model prediction results

数据集 Dataset	化合物数 Chemical number (n)	真阳性 True positive (TP)	真阴性 True negative (TN)	假阴性 False negative (FN)	假阳性 False positive (FP)
模型 RF					
训练集	104	73	30	0	1
测试集	26	17	6	1	2
模型 SVM					
训练集	104	72	29	1	2
测试集	26	16	7	2	1
模型 ANN					
训练集	104	72	27	1	4
测试集	26	15	7	3	1
模型 KNN					
训练集	104	71	28	2	3
测试集	26	15	7	3	1
模型 NB					
训练集	104	69	16	4	15
测试集	26	16	6	2	2
模型 LG					
训练集	104	70	13	3	18
测试集	26	16	7	2	1

根据结果, 表 3 列举了模型 RF、SVM、ANN 和 KNN 的 4 组预测结果较好的数据, 其中 Y 表示分子结合 GPER、N 表示不结合. 可以看到, 测试集的 α -E2、训练集的 2'-OH-BDE-003 和 BPF 被四个模型全部预测错误. 推测原因为 α -E2 和 BPF 在结构上与具有结合效应的分子 E2 和 BPA 有相似之处, 2'-OH-BDE-003 在结构上与不具有结合效应的分子 BDE-003 有相似之处, 由此造成了本研究中的所有分类模型都不能对其进行正确的分类.

依据表 4, 计算模型评价指标 S_n 、 S_p 、 Q 和 MCC 值, 如下所示. 可以看到 RF 的精确度最高, 训练集达到 99%, 测试集达到 88.5%, 敏感性和特异性也较高, 说明 RF 的预测效果优秀. SVM、ANN、KNN 的精度次之, 训练集精确度分别达到 97.1%、95.2%、95.2%; 测试集精确度分别达到 88.5%、84.6%、84.6%, 敏感性和特异性也较高. 而 NB 和 LG 的预测结果较差, 训练集精度分别为 81.7% 和 79.8%; 测试集精度分别为 84.6% 和 88.5%, 由此得到最优模型为 RF. 同时其它模型的统计学评价指标(S_n 、 S_p 、 Q 和 MCC)也显示为 RF 为最优的分类预测模型.

表 3 模型化合物预测情况
Table 3 Prediction of model compounds

序号 No.	化合物 Compound	参考文献 Reference	GPER配体分子(是/否) GPER ligand molecule(Y/N)				
			观测值Observed	SVM	RF	ANN	KNN
1	SK0*	[25]	Y	Y	Y	Y	Y
2	SK0P	[25]	Y	Y	Y	Y	Y
3	G-1	[6, 21]	Y	Y	Y	Y	Y
4	G-15	[6,21]	Y	Y	Y	Y	Y
5	G-36	[6,21]	Y	Y	Y	Y	Y
6	Oleuropein	[24]	Y	Y	Y	Y	Y
7	Hydroxytyrosol	[24]	Y	Y	Y	Y	Y
8	MIBE	[24]	Y	Y	Y	Y	Y
9	4-hydroxytamoxifen	[24]	Y	Y	Y	Y	Y
10	GPER-L1	[24]	Y	Y	Y	Y	Y
11	GPER-L2	[24]	Y	Y	Y	Y	Y
12	17β-estradiol	[22]	Y	Y	Y	Y	Y
13	E3	[24]	N	N	N	Y	N
14	Tamoxifen	[22]	Y	Y	Y	Y	Y
15	Fulvestrant	[22]	Y	Y	Y	Y	Y
16	Epi*	[27]	Y	Y	Y	Y	Y
17	Epi-prop	[27]	Y	Y	Y	Y	Y
18	Epi-4-prop	[27]	Y	Y	Y	Y	Y
19	Epi-5-prop	[27]	Y	Y	Y	Y	Y
20	Epi-Ms	[27]	N	N	N	Y	N
21	C4PY	[26]	Y	Y	Y	Y	Y
22	7β-OH-EpiA*	[24]	Y	Y	Y	Y	Y
23	G-DOTA	[37]	Y	Y	Y	Y	Y
24	G-Bz-DOTA	[37]	N	N	N	N	N
25	G-Bz-DTPA	[37]	N	N	N	N	N
26	Atrazine	[24]	Y	Y	Y	Y	Y
27	PBX1	[34]	Y	Y	Y	Y	Y
28	PBX2	[34]	Y	Y	Y	Y	Y
29	ZINC65156419(1)	[29]	Y	Y	Y	Y	N
30	ZINC65156419(2)	[29]	N	N	N	N	N
31	ZINC65156419(3)	[29]	N	N	N	N	N
32	ZINC65156419(4)	[29]	N	N	N	N	N
33	ZINC65156419(5)	[29]	Y	Y	Y	Y	Y
34	ZINC65156419(6)	[29]	N	N	N	N	N
35	ZINC65156419(7)	[29]	N	N	N	N	N
36	ZINC65156419(8)	[29]	N	N	N	N	N
37	ZINC65156419(9)*	[29]	Y	Y	Y	N	Y

续表 3

序号 No.	化合物 Compound	参考文献 Reference	GPER配体分子(是/否) GPER ligand molecule(Y/N)				
			观测值Observed	SVM	RF	ANN	KNN
38	E2-NH3 ⁺	[13]	Y	Y	Y	Y	Y
39	E2-COO ⁻	[13]	Y	Y	Y	Y	Y
40	E2-NMe3 ⁺	[13]	Y	Y	Y	Y	Y
41	E2-NB	[13]	Y	Y	Y	Y	Y
42	<i>o,p'</i> -DDE	[6]	Y	Y	Y	Y	Y
43	E1*	[24]	N	N	Y	N	N
44	α-E2	[24]	N	Y	Y	Y	Y
45	Genistein	[6]	Y	Y	Y	Y	Y
46	<i>p,p'</i> -DDT	[6]	Y	Y	Y	Y	Y
47	BPA*	[6]	Y	Y	Y	Y	Y
48	quercetin	[24]	Y	Y	Y	Y	Y
49	Resveratrol*	[24]	Y	Y	Y	Y	Y
50	Raloxifene	[24]	Y	Y	Y	Y	Y
51	zearalnone	[6]	Y	Y	Y	Y	Y
52	Nonylphenol	[6]	Y	Y	Y	Y	Y
53	kepone	[6]	Y	Y	Y	Y	Y
54	STX	[24]	Y	Y	Y	Y	Y
55	PPT*	[24]	Y	Y	Y	Y	Y
56	2,2',5'-PCB-4-OH	[6]	Y	Y	Y	Y	Y
57	equol	[24]	Y	Y	Y	Y	Y
58	2-methoxye stradiol	[24]	Y	Y	Y	Y	Y
59	niacin	[24]	Y	Y	Y	Y	Y
60	daidzein	[24]	Y	Y	Y	Y	Y
61	BDE-003	[28]	N	N	N	N	N
62	BDE-007*	[28]	N	N	N	N	N
63	BDE-028	[28]	N	N	N	N	N
64	BDE-047	[28]	N	N	N	N	N
65	BDE-049*	[28]	N	N	N	N	N
66	BDE-085*	[28]	N	N	N	N	N
67	BDE-099*	[28]	N	N	N	N	N
68	BDE-100	[28]	N	N	N	N	N
69	BDE-154	[28]	N	N	N	N	N
70	BDE-180	[28]	N	N	N	N	N
71	BDE-187	[28]	N	N	N	N	N
72	BDE-201	[28]	N	N	N	N	N
73	2'-OH-BDE-003*	[28]	Y	N	N	N	N
74	3'-OH-BDE-007	[28]	Y	Y	Y	Y	Y
75	3'-OH-BDE-028	[28]	Y	Y	Y	Y	Y

续表 3

序号 No.	化合物 Compound	参考文献 Reference	GPER配体分子(是/否) GPER ligand molecule(Y/N)				
			观测值Observed	SVM	RF	ANN	KNN
76	3'-OH-BDE-047*	[28]	Y	Y	Y	Y	Y
77	3'-OH-BDE-154	[28]	Y	Y	Y	Y	Y
78	4'-OH-BDE-049	[28]	Y	Y	Y	Y	Y
79	5'-OH-BDE-099*	[28]	Y	Y	Y	Y	Y
80	2'-OH-BDE-007	[28]	N	N	N	N	N
81	2'-OH-BDE-028*	[28]	N	N	N	N	N
82	3-OH-BDE-100	[28]	Y	N	Y	N	Y
83	4-OH-BDE-187	[28]	Y	Y	Y	Y	Y
84	4'-OH-BDE-201	[28]	Y	Y	Y	Y	Y
85	5-OH-BDE-047	[28]	N	N	N	N	N
86	5'-OH-BDE-100	[28]	N	N	N	N	N
87	6-OH-BDE-047	[28]	N	N	N	N	N
88	6-OH-BDE-085*	[28]	N	N	N	N	N
89	6'-OH-BDE-099	[28]	N	N	N	N	N
90	6-OH-BDE-180	[28]	Y	Y	Y	Y	N
91	BPAF	[16]	Y	Y	Y	Y	Y
92	BPB*	[16]	Y	Y	Y	Y	Y
93	BPF*	[16]	N	Y	Y	Y	Y
94	BPS*	[16]	Y	Y	Y	N	N
95	TBBPA	[16]	N	N	N	N	Y
96	TCBPA	[16]	Y	Y	Y	Y	Y
97	Diethylstilbestro	[24]	N	N	N	N	N
98	2-Hydroxy stradiol*	[24]	Y	N	Y	Y	N
99	Aldosterone	[24]	Y	Y	Y	Y	Y
100	Tectoridin	[24]	Y	Y	Y	Y	Y
101	Apigenin*	[24]	Y	Y	Y	Y	Y
102	Methoxychlor	[24]	Y	Y	Y	Y	Y
103	<i>p,p'</i> -DDE*	[24]	Y	Y	Y	Y	Y
104	<i>o,p'</i> -DDT*	[24]	Y	Y	Y	Y	Y
105	DPN	[24]	Y	Y	Y	Y	Y
106	Ethinylestradiol	[24]	Y	Y	Y	Y	Y
107	3MC	[32]	Y	Y	Y	Y	Y
108	AB-1	[31]	N	N	N	N	N
109	CIMBA-5*	[36]	Y	Y	Y	Y	Y
110	CIMBA-6	[36]	Y	Y	Y	Y	Y
111	CIMBA-7	[36]	Y	Y	Y	Y	Y
112	CIMBA-8	[36]	Y	Y	Y	Y	Y
113	CIMBA-9	[36]	N	Y	N	Y	N

续表 3

序号 No.	化合物 Compound	参考文献 Reference	GPER配体分子(是/否) GPER ligand molecule(Y/N)				
			观测值Observed	SVM	RF	ANN	KNN
114	CIMBA-10	[36]	Y	Y	Y	Y	Y
115	CIMBA-11	[36]	Y	Y	Y	Y	Y
116	CIMBA-12	[36]	Y	Y	Y	Y	Y
117	CIMBA-13	[36]	Y	Y	Y	Y	Y
118	CIMBA-14	[36]	N	N	N	N	N
119	CIMBA-15	[36]	Y	Y	Y	Y	Y
120	CIMBA-16	[36]	N	N	N	N	Y
121	CIMBA-17	[36]	Y	Y	Y	Y	Y
122	CIMBA-18	[36]	Y	Y	Y	Y	Y
123	CIMBA-19	[36]	N	N	N	N	N
124	CIMBA-20	[36]	Y	Y	Y	Y	Y
125	CIMBA-21	[36]	Y	Y	Y	Y	Y
126	CIMBA-22	[36]	Y	Y	Y	Y	Y
127	CIMBA-23	[36]	Y	Y	Y	Y	Y
128	CIMBA-24	[36]	Y	Y	Y	Y	Y
129	CIMBA-25	[36]	Y	Y	Y	Y	Y
130	Carbhydraz	[35]	Y	Y	Y	Y	Y

注: *测试集化合物.
Note: *Testing set compounds.

表 4 模型评价

Table 4 Model performance

数据集 Dataset	化合物数 Chemical number(<i>n</i>)	敏感性 Sensitivity(<i>S_n</i>)	特异性 Specificity(<i>S_p</i>)	精确度 Accuracy(<i>Q</i>)	马修斯相关系数 Matthews correlation coefficient(MCC)
模型 RF					
训练集	104	1	0.968	0.99	0.977
测试集	26	0.944	0.75	0.885	0.723
模型 SVM					
训练集	104	0.986	0.935	0.971	0.931
测试集	26	0.889	0.875	0.885	0.741
模型ANN					
训练集	104	0.986	0.871	0.952	0.884
测试集	26	0.833	0.875	0.846	0.672
模型 KNN					
训练集	104	0.972	0.903	0.952	0.884
测试集	26	0.833	0.875	0.846	0.672
模型 NB					
训练集	104	0.945	0.516	0.817	0.535
测试集	26	0.889	0.75	0.846	0.639
模型 LG					
训练集	104	0.959	0.419	0.798	0.480
测试集	26	0.889	0.875	0.885	0.741

此外通过受试者工作特征(ROC)曲线以及 AUC 曲下面积对模型评价进行进一步说明. AUC 曲下面积越接近 1 模型性能越高. 由图 2 可知, RF 训练集和测试集的 AUC 值均在 0.9 以上, 说明模型性能最优; SVM、ANN、KNN 次之, 训练集 AUC 值分别达到 0.961、0.987、0.938, 测试集 AUC 分别达到 0.882、0.889、0.854, 模型性能较好; NB 和 LG 的训练集 AUC 值分别为 0.795 和 0.765, 测试集 AUC 均为 0.799, 说明模型预测性能较差.

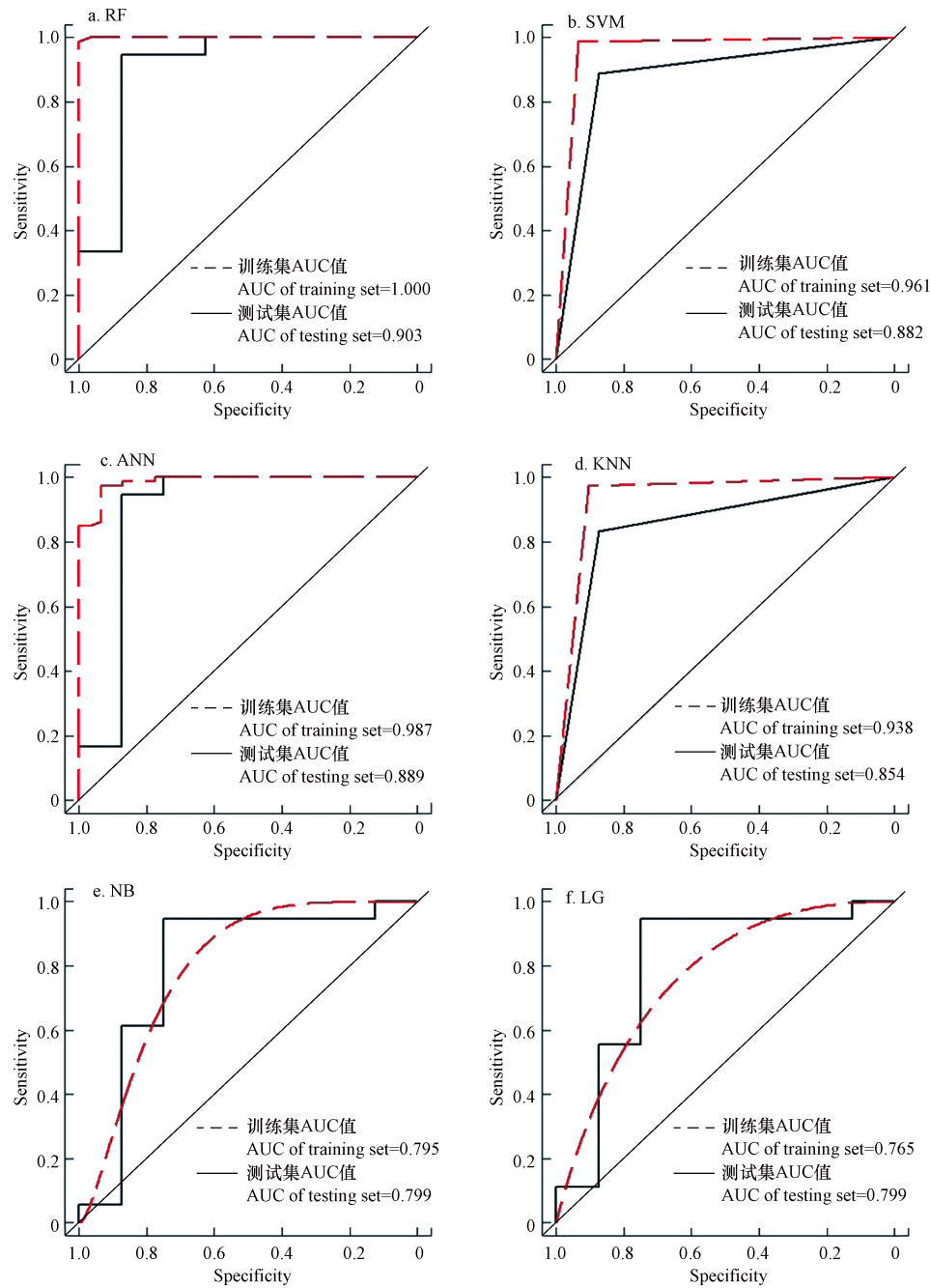


图 2 分类模型的受试者工作特征(ROC)曲线(虚线为训练集、实线为测试集)
(A)RF 模型; (B)SVM 模型; (C)ANN 模型; (D)KNN 模型; (E)NB 模型; (F)LG 模型

Fig.2 Receiver operating characteristics(ROC) graphs of the classification model

最后, 通过 10 折交叉验证来计算模型的精度, 其中 RF、SVM、ANN、KNN 精度分别为 81%、81%、81%、80%, 而 NB 和 LG 的精度较低, 分别为 75% 和 72%, 说明 RF 和 SVM 模型的精度较高, 这和前面模型性能评估得出的结论是一致的.

2.4 应用域表征

本文使用基于欧几里得距离的标准来探索化合物是否适用于所建模型, 利用训练集计算可知化合

物中最大的欧几里得距离为 0.88(截断值), 而计算得到的测试集中所有化合物的欧几里得距离均小于 0.88, 表明测试集化合物均适用于预测模型. 这也为测试活性未知的化合物是否适用于所建模型提供了评价标准.

3 结论(Conclusion)

目前对于外源性化合物对膜雌激素受体 GPER 的干扰效应研究方兴未艾, 仍有大量潜在的化合物有待验证. 因此, 在我们的研究中, 搜集了 130 种已知的化合物结合 GPER 的情况, 评价了 RF、SVM、ANN、KNN、NB、LG 的 6 种机器学习算法构建的分类预测模型的表现. 通过特征选择、模型优化以及模型评价, 结果显示 RF 为最优模型, 为评价环境污染物是否通过结合 GPER 而产生内分泌干扰效应提供了快速筛查的预测模型与理论支撑.

参考文献 (References)

- [1] FENICHEL P, CHEVALIER N. Is testicular germ cell cancer estrogen dependent? The role of endocrine disrupting chemicals [J]. *Endocrinology*, 2019, 160(12): 2981-2989.
- [2] BALABANIC D, RUPNIK M S, KLEMENCIC A K. Negative impact of endocrine-disrupting compounds on human reproductive health [J]. *Reproduction Fertility and Development*, 2011, 23(3): 403-416.
- [3] BRAUN J M. Early-life exposure to EDCs: Role in childhood obesity and neurodevelopment [J]. *Nature Reviews Endocrinology*, 2017, 13(3): 161-173.
- [4] FOWLER P A, BELLINGHAM M, SINCLAIR K D, et al. Impact of endocrine-disrupting compounds (EDCs) on female reproductive health [J]. *Molecular and Cellular Endocrinology*, 2012, 355(2): 231-239.
- [5] HARRIS H A, KATZENELLENBOGEN J A, KATZENELLENBOGEN B S J E. Characterization of the biological roles of the estrogen receptors, ER α and ER β , in estrogen target tissues in vivo through the use of an ER α -selective ligand [J]. *Endocrinology*, 2002(11): 4172-4177.
- [6] THOMAS P, DONG J. Binding and activation of the seven-transmembrane estrogen receptor GPR30 by environmental estrogens: A potential novel mechanism of endocrine disruption [J]. *The Journal of Steroid Biochemistry and Molecular Biology*, 2006, 102(1): 175-179.
- [7] PROSSNITZ E R, BARTON M. The G-protein-coupled estrogen receptor GPER in health and disease [J]. *Nature Reviews Endocrinology*, 2011, 7(12): 715-726.
- [8] PROSSNITZ E R, BARTON M. Signaling, physiological functions and clinical relevance of the G protein-coupled estrogen receptor GPER [J]. *Prostaglandins Other Lipid Mediat*, 2009, 89(3-4): 89-97.
- [9] WANG C, LI J, YE S, et al. Oestrogen inhibits VEGF expression and angiogenesis in triple-negative breast cancer by activating GPER-1 [J]. *Journal of Cancer*, 2018, 9(20): 3802-3811.
- [10] NIRO S, PEREIRA E, PELISSIER M A, et al. The DHEA metabolite 7 β -hydroxy-epiandrosterone exerts anti-estrogenic effects on breast cancer cell lines [J]. *Steroids*, 2012, 77(5): 542-551.
- [11] PROSSNITZ E, BARTON M. Estrogen biology: New insights into GPER function and clinical opportunities [J]. *Molecular and Cellular Endocrinology*, 2014, 389(1): 71-83.
- [12] ALBANITO L, MADEO A, LAPPANO R, et al. G protein-coupled receptor 30 (GPR30) mediates gene expression changes and growth response to 17 β -estradiol and selective GPR30 ligand G-1 in ovarian cancer cells [J]. *Cancer Research*, 2007, 67(4): 1859-1866.
- [13] REVANKAR C M, MITCHELL H D, FIELD A S, et al. Synthetic estrogen derivatives demonstrate the functionality of intracellular GPR30 [J]. *Acs Chemical Biology*, 2007, 2(8): 536-544.
- [14] TSUGAWA Y, HIRAMOTO M, IMAI T. Estrogen induces estrogen receptor α expression and hepatocyte proliferation in late pregnancy [J]. *Biochemical & Biophysical Research Communications*, 2019, 511(3): 592-596.
- [15] PROSSNITZ E R. GPER modulators: Opportunity nox on the heels of a class akt [J]. *The Journal of Steroid Biochemistry and Molecular Biology*, 2018, 176: 73-81.
- [16] CAO L Y, REN X M, LI C H, et al. Bisphenol AF and bisphenol B exert higher estrogenic effects than bisphenol A via G protein-coupled estrogen receptor pathway [J]. *Environmental Science & Technology*, 2017, 51(19): 11423-11430.
- [17] RUSSO D P, ZORN K M, CLARK A M, et al. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction [J]. *Molecular Pharmaceutics*, 2018, 15(10): 4361-4370.
- [18] GRISONI F, CONSONNI V, BALLABIO D. Machine learning consensus to predict the binding to the androgen receptor within the compara project [J]. *Journal of Chemical Information and Modeling*, 2019, 59(5): 1839-1848.

- [19] SHEFFIELD T, JUDSON R. Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure [J]. *Environmental Science & Technology*, 2019, 53(21): 12793-12802.
- [20] WANG Z, CHEN J, HONG H. Applicability domains enhance application of PPARG agonist classifiers trained by drug-like compounds to environmental chemicals [J]. *Chemical Research in Toxicology*, 2020, 33(6): 1382-1388.
- [21] ARNATT C K, ZHANG Y. G protein-coupled estrogen receptor (GPER) agonist dual binding mode analyses toward understanding of its activation mechanism: a comparative homology modeling approach [J]. *Molecular Informatics*, 2013, 32(7): 647-658.
- [22] WANG D, HU L, ZHANG G, et al. G protein-coupled receptor 30 in tumor development [J]. *Endocrine*, 2010, 38(1): 29-37.
- [23] BARTON M, PROSSNITZ E R. Emerging roles of GPER in diabetes and atherosclerosis [J]. *Trends in Endocrinology & Metabolism*, 2015, 26(4): 185-192.
- [24] PROSSNITZ E, ARTERBURN J. International union of basic and clinical pharmacology. XC VII. G protein-coupled estrogen receptor and its pharmacologic modulators [J]. *Pharmacological Reviews*, 2015, 67(3): 505-540.
- [25] KHAN S U, AHMED N, CHUAH L H, et al. Sequential ligand and structure-based virtual screening approach for the identification of potential G protein-coupled estrogen receptor-1 (GPER-1) modulators [J]. *Rsc Advances*, 2019, 9(5): 2525-2538.
- [26] LAPPANO R, ROSANO C, PISANO A, et al. A calixpyrrole derivative acts as an antagonist to GPER, a G-protein coupled receptor: mechanisms and models [J]. *Disease models & mechanisms*, 2015, 8(10): 1237-1246.
- [27] SARMIENTO V, SANCHEZ T R, ULLOA A M, et al. Synthesis of novel (-)-epicatechin derivatives as potential endothelial GPER agonists: Evaluation of biological effects [J]. *Bioorganic & Medicinal Chemistry Letters*, 2018, 28(4): 658-663.
- [28] CAO L Y, REN X M, YANG Y, et al. Hydroxylated polybrominated biphenyl ethers exert estrogenic effects via non-genomic G protein-coupled estrogen receptor mediated pathways [J]. *Environmental Health Perspectives*, 2018, 126(5): 057005.
- [29] O'DEA A, SONDERGARD C, SWEENEY P, et al. A series of indole-thiazole derivatives act as GPER agonists and inhibit breast cancer cell growth [J]. *Acs Medicinal Chemistry Letters*, 2018, 9(9): 901-906.
- [30] ALBANITO L, LAPPANO R, MADEO A, et al. Effects of atrazine on estrogen receptor α and G protein-coupled receptor 30-mediated signaling and proliferation in cancer cells and cancer-associated fibroblasts [J]. *Environmental Health Perspectives*, 2015, 123(5): 493-499.
- [31] REVANKAR C M, BOLOGA C G, PEPERMANS R A, et al. A selective ligand for estrogen receptor proteins discriminates rapid and genomic signaling [J]. *Cell Chemical Biology*, 2019, 26(12): 1692-1702.
- [32] CIRILLO F, LAPPANO R, BRUNO L, et al. AHR and GPER mediate the stimulatory effects induced by 3-methylcholanthrene in breast cancer cells and cancer-associated fibroblasts (CAFs) [J]. *Journal of Experimental & Clinical Cancer Research*, 2019, 38(1): 335-353.
- [33] ALDO M U, DAVID M L, ERNESTO B R, et al. The effects of (-)-epicatechin on endothelial cells involve the G protein-coupled estrogen receptor (GPER) [J]. *Pharmacological Research*, 2015, 100: 309-320.
- [34] MAGGIOLINI M, SANTOLLA M, AVINO S, et al. Identification of two benzopyrroloxazines acting as selective GPER antagonists in breast cancer cells and cancer-associated fibroblasts [J]. *Future medicinal chemistry*, 2015, 7(4): 437-448.
- [35] ROSANO C, PONASSI M, SANTOLLA M F, et al. Macromolecular modelling and docking simulations for the discovery of selective GPER ligands [J]. *The AAPS Journal*, 2016, 18(1): 41-46.
- [36] DELEON C, WANG H, GUNN J, et al. A novel GPER antagonist protects against the formation of estrogen-induced cholesterol gallstones in female mice [J]. *Journal of Lipid Research*, 2020, 61(5): 767-777.
- [37] NAYAK T K, DENNIS M K, RAMESH C, et al. Influence of charge on cell permeability and tumor imaging of GPR30-targeted ¹¹¹In-labeled nonsteroidal imaging agents [J]. *ACS Chemical Biology*, 2010, 5(7): 681-690.
- [38] LUO L J, LIU F, LIN Z K, et al. Genistein regulates the IL-1 beta induced activation of MAPKs in human periodontal ligament cells through G protein-coupled receptor 30 [J]. *Archives of Biochemistry and Biophysics*, 2012, 522(1): 9-16.