遗传算法结合共轭梯度法改进 BP 算法人工神经网络用于环境雌激素的 QSAR 研究

季力 王晓栋* 杨旭曙 刘树深 王连生

(污染控制与资源化研究国家重点实验室, 南京大学环境学院, 南京 210093; 长江水环境教育部重点实验室, 同济大学环境科学与工程学院, 上海 200092. *联系人, E-mail: wangxd@nju.edu.cn)

摘要 目前应用最广泛的BP人工神经网络存在收敛速度低、有局部震荡的危险。采用共轭梯度法(CG)改进BP算法, 并且结合遗传算法筛选分子描述符, 用于环境内分泌干扰物的定量结构-活性相关(QSAR)研究, 有效地克服了以上缺点, 得到了稳健、准确的预测模型, 模型的 $R^2 = 0.845$, 预测集的 $q^2_{\text{pred}} = 0.81$, 均方根误差(RMSE) = 0.688。所得结果说明这种方法能够为筛选有机物的雌激素活性提供一种迅速、可行的工具。

关键词 BP 共轭梯度法 遗传算法 环境内分泌干扰物 定量结构-活性相关

实验室和流行病学研究表明, 一些人造的或天然产生的化学物质, 即“内分泌干扰物”(EDCs)^[1], 由于能够干扰人和野生动物内分泌系统的正常功能, 威胁到人类和野生生物的生存和繁衍, 因而引起全世界的广泛关注。美国环境保护署为此成立了内分泌干扰物审查和试验咨询委员会(简称EDSTAC), 该委员会提出了“双层方案”对大量的化学物质来进行筛选和测试它们的内分泌干扰活性。在第一层中, 将会采用多终点的方法对化学物质进行筛选, EDSTAC推荐了20多种离体和活体测试方法^[2]。鉴于目前大量的有毒化学品需要进行筛选, 采用完全基于试验的筛选体系是不现实的。

研究表明, 定量结构-活性相关(QSAR)方法是预测有机物内分泌活性的有效工具, 自20世纪90年代以来, 国外学者对此做了大量研究工作。由于内分泌干扰物的结构多样性以及与受体作用机理的复杂性, 更多的研究集中于采用三维或者准三维的QSAR模型, 例如比较分子场分析法(CoMFA)^[3-6]、比较分子相似性指数分析法(COMSIA)^[7]、自组织分子场分析法(SOMFA)^[8]、光谱QSAR法(Spectroscopic QSAR)^[8]、分子全息QSAR法(HQSAR)^[5,7]、迅速随机排除描述子/亚结构键法(FRED/SKEYS)^[9]等, 这些模型从不同角度揭示了环境内分泌干扰物的主要作用受体——雌激素受体和配体间的相互作用, 但是对于结构复杂

的大量数据集, 多维方法显示出操作复杂、规则难以确定等弱点, 难以适应高通量筛选的要求。

人工神经网络是模拟人类生理神经机制的计算模型, 能够识别高度非线性的关系, 为复杂的、高度非线性问题的模式识别提供了可选手段^[10]。目前最常用的是误差反传前馈神经网络(BP神经网络), 初始权值一般随机赋值, 训练循环进行, 通过不断调整各连接权之间的权值达到最终构建网络的目的^[11]。Marini等人^[12]曾经用BP网络对内分泌干扰物进行研究, 取得了比较好的结果, 但是通常认为BP网络类似于黑箱, 网络的一个或多个权值对每个输出都有影响, 每次随机赋值都会得到不同的预测结果, 致使网络不够稳健。而循环训练的结果导致了学习速率的降低、学习时间的冗长^[13-16]。

本文中, 首先使用客观变量选择和主观变量选择——遗传算法, 对125种雌激素的分子描述符进行筛选, 随后在训练以上雌激素活性值与分子结构描述符之间关系的人工神经网络模型中, 当BP算法收敛速度达到缓慢阶段时, 使用一种重要的无约束最优化方法——共轭梯度法(CG)^[17]进行改进, 这种两阶段训练法, 可以充分利用两种算法的优点, 有效地克服了单纯使用BP网络所存在的缺点, 建立了具有较高预测性能的QSAR模型, 可以对环境雌激素类物质进行快速、比较准确的预测。

2007-02-02 收稿, 2007-06-26 接受

国家自然科学基金(批准号: 20507008, 20477018)、江苏省自然科学基金创新人才项目(批准号: BK200418)和国家重点基础研究发展计划(批准号: 2003CB415002)资助

1 材料和方法

1.1 研究对象

本文所研究的125种雌激素的活性数据全部取自文献[18], 雌激素活性指标为雌激素受体相对亲和力(取对数, 为logRBA). 此数据集包括类固醇雌激素类、双酚A类衍生物、乙烯雌酚类衍生物、烷基酚类、植物雌激素类、苯甲酸酯类、多氯联苯类、滴滴涕类等, 其代表性结构见表1. 将这些物质随机分为训练集、监测集和预测集. 其中随机选取的20种物质作为监测集监测过拟合现象的发生, 20种物质作为预测集以验证模型对外部样本的预测能力, 在选取监测集和预测集时, 保证了结构的代表性和多样性.

1.2 分子描述符的计算

采用了MOPAC-AM1方法对所有分子进行了结构最优化, 用DRAGON软件(<http://www.taletel.mi.it/>)计算每种物质的1481种分子描述符, 包括拓扑描述符、分子连接性指数、几何描述符等18类[19].

1.3 变量选择(feature selection)[20,21]

在大量的分子描述符中, 很多描述符提供了重

复的结构信息, 因此变量选择是必要的. 采用客观变量选择和主观变量选择两个流程进行变量选择. 客观变量选择方法是用DRAGON软件本身筛选为常数或者近似常数的结构描述符以及相关系数达到0.95以上的两个描述符中的一个; 主观变量选择采用遗传算法, 它是一种根据生物学中所谓自然选择和遗传机理设计的随机优化算法, Rogers和Hopfinger[22]首先将该方法用于QSAR研究中. 在本文中联合使用遗传算法和人工神经网络, 以遗传算法作为优化工具, 以神经网络方法所建立模型的均方根误差作为遗传算法的适应值, 以筛选出与分子活性最相关的描述符[23].

1.4 共轭梯度法改进BP算法人工神经网络

BP算法[11]: 它的训练过程是将输入信息沿网络正向传播, 误差信号反向传播来修改网络连接权的组成. 具体地讲, 它是利用均方误差和梯度下降法来实现网络连接权的修正, 对网络连接权修正的目标是网络误差平方和最小. 从输入层输入信号经隐层处理, 并传向输出层, 每一层神经元状态只是影响下一层神经元状态, 如果在输出层得不到期望输出, 则

表1 研究对象的结构类别与代表化合物

类别	代表物质	类别	代表物质
类固醇类		双酚 A 类衍生物
乙烯雌酚类衍生物		烷基酚类
植物雌激素类		多氯联苯类
滴滴涕类		苯甲酸酯类
其他

将误差信号沿原来连接通路反向传播,进而修正网络原来的连接权,使得网络误差信号最小.权值调整公式为

$$\Delta W(k+1) = \eta \delta_j p_j + \mu \Delta W(k),$$

式中, k 为训练次数, η 为学习率, p_j 为第 j 个输入矢量, δ_j 为反向传播误差导数, μ 为动量因子.

共轭梯度算法^[17]: 共轭梯度法是重要的无约束最优化方法,其基本思想是利用已知点处的梯度构造一组共轭方向,并据此搜索目标函数极值.在BP算法中,相邻迭代的搜索方向正交,当接近极值时会发生“锯齿”形震荡,利用这种正交性又可以构成称为共轭方向的搜索方向,大大改善收敛性能与收敛速度,化不利为有利.

将两者结合,把训练过程分成两个阶段,在开始阶段采用BP算法,到收敛速度缓慢阶段时(即接近最优点时),改用共轭梯度法,便可以充分发挥二者的优点,彼此取长补短.这就是共轭梯度法改进BP算法的基本原理^[13,14,24].在实际应用中,BP网络存在着网络训练易于过度、收敛速度慢、预测结果不稳定、易陷入局部极小的问题,使用共轭梯度法改进BP算法,可以大大加快收敛速度,有效地避免局部极小值,所建立的模型也更加稳健.共轭梯度法改进BP算法人工神经网络的构建和训练在STATISTICA Neural Networks 7.0(SNN)神经网络工具软件中进行^[23].

在人工神经网络训练中,为了避免训练过程中过拟合现象的发生,从训练集中分离出监测集来决定训练的终点.在训练过程中,始终周期性地计算监测集的均方根误差(RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_{\text{预测}} - y_{\text{实验}})^2}{n}},$$

当均方根误差开始上升的时候则说明出现过拟合现象.此时停止训练所得到的神经网络用于未来的预测具有合理的可信度^[20,21].

1.5 模型质量的验证^[20,21]

采用预测集的均方根误差来表征模型的泛化能力,采用预测集的 q_{pred}^2 来评估QSAR模型的预测能力: $q_{\text{pred}}^2 = 1 - \text{PRESS}/\text{SD}$, PRESS是预测集的实验值与预测值差值的平方的加和,SD是预测集实验值与训练集平均值差值的平方的加和^[25].

模型验证的第二步是使用蒙特卡罗实验方法,^[doi/10.1360/1000-0607/2007-09-002118]所确定隐含层个数为1,隐含层神经元

其基本原理就是因变量被完全打乱,并且被重新分配给每个化合物.随后使用相同的分子描述符,相同的神经网络结构,相同的训练集和监测集,并且使用相同的预测集用于模型的验证.通过打乱因变量,真正的构效关系被破坏了,如果原模型不是偶然相关,那么打乱后所建模型的质量会大大恶化.

2 结果与讨论

2.1 遗传算法结合共轭梯度法(CG)改进BP算法神经网络的建立

通过客观变量选择和主观变量选择——遗传算法,筛选出45个描述符作为人工神经网络的输入值.这45个描述符包括Aromaticity描述符3个,表征分子的芳香性;3D-MoRSE描述符1个,是基于电子衍射的三维分子结构表征;GETAWAY描述符31个,是几何、拓扑和原子分子量集合描述符;Functional描述符4个,是功能性基团描述符;Atom-centered-fragment描述符6个,是原子中心片段描述符.详见表2.

表2 遗传算法对分子描述符的筛选结果

描述符	筛选结果
Aromaticity 描述符	HOMA, RCI, AROM
3D-MoRSE 描述符	Mor06v
GETAWAY 描述符	H5u, H6u, H7u, H6v, H5e, H6e, H7e, H5p, H6p, RCON, R5u, R6u, R7u, R8u, Rtu, R3v, R4v, R5v, R6v, R7v, RTv, R5e, R6e, R7e, R8e, Rte, R4p, R5p, R6p, R7p, RTP
Functional 描述符	nCt, nCrH2, nCrR2, nCOOR
Atom-centred-fragment 描述符	C-003, C-004, C-013, C-029, H-052, O-058

学习率和动量值: 在BP人工神经网络的训练中,学习率的选取对训练的效果有着重要影响.学习率过大将使误差函数发生振荡,学习率过小,收敛过慢,经过反复实验,本文中该值取0.06;附加动量项有助于使网络从误差曲面的局部极小值中跳出,该值取0.69.

隐含层(中间层)层数和神经元数的确定: 在BP算法中,人工神经网络拓扑结构中的输入节点与输出节点数是由问题的本身决定的,关键在于隐含层的层数与隐含层神经元的数目的确定.相对于隐含层的层数,神经元数目的选取更困难,神经元数目少了,学习过程不可能收敛;神经元数目多了,网络性能下降,节点冗余.本文在网络的学习过程中,经过反复训练,调整结构,最后得到一个大小合适的神经网络模型,所确定隐含层个数为1,隐含层神经元

目为 23.

综上, 本文设计 3 层神经网络模型, 分别为输入层、隐含层和输出层, 其神经元数分别为 45, 23 和 1, 用 BP 算法进行 150 次学习, 用 CG 算法进行 45 次学习后, 监测集的误差曲线明显出现上升的趋势, 立刻停止训练.

此时, 所得到的模型的训练集均方根误差为 0.716 ($R^2 = 0.845$), 监测集的均方根误差为 0.763 ($R^2 = 0.895$), 预测集的均方根误差为 0.688 ($R^2 = 0.810$). 预测集具有较小的误差值, 说明所建模型具有良好的泛化能力. 预测值相对实验值的散点图见图 1. 人工神经网络预测值和实验值的比较, 详见表 3. q_{pred}^2 被用来比较不同 QSAR 模型的预测能力, $q_{\text{pred}}^2 = 1 - \text{PRESS}/\text{SD}$, 经计算, 本文的 q_{pred}^2 为 0.81.

图 1 用共轭梯度法改进 BP 算法神经网络模型计算所得的预测值相对实验值的散点图

而单纯采用 BP 网络, 用相同的参数值训练 1000 次后, 所得到的结果如下: 训练集均方根误差为 1.188 ($R^2 = 0.818$), 监测集均方根误差为 1.368 ($R^2 = 0.710$), 预测集均方根误差为 1.199 ($R^2 = 0.604$), 训练集和预测集的均方根误差分别提高了 56% 和 74%. 可以看出 BP 网络相对于本文提出的方法, 训练周期长, 而且效果相对较差.

2.2 验证模型的蒙特卡罗实验

对模型进行蒙特卡罗实验, logRBA 值被杂混并被重新分配给化合物后, 与以上神经网络具有相同描述符和相同网络拓扑结构的模型被搭建. 此时, 训练集均方根误差为 1.317 ($R^2 = 0.757$), 监测集(交叉验证集)的均方根误差为 3.088 ($R^2 = 0.303$), 预测集的均方根误差为 2.677 ($R^2 = 0.344$). 可以看出, 蒙特卡罗实验所得模型的质量相对于真实的模型, 是大大恶化了, 说明本文所建立的模型并非偶然相关.

2.3 与文献数据的比较

Shi 等人 [5] 曾经用 CoMFA 和 HQSAR 两种方法, Asikainen 等人 [26] 用 KNN QSAR 分别对与本文相同的雌激素数据集建立了 QSAR 模型, 与本文所建模型质量的比较见表 4.

从表 4 中的比较可以看出, 我们所建的共轭梯度法改进 BP 算法人工神经网络模型的 R^2 明显优于 HQSAR 方法所得的结果, 稍逊于 CoMFA 和 KNN, 但是在预测能力上, 本文中的 q_{pred}^2 达到 0.81, 高于 CoMFA 和 KNN 模型. 值得强调的是在 Shi 等人 [5] 的研究中所采用的两个测试集, 均不是来自于和训练集同一实验室的数据, 而是取自其他实验室, 经过标准化以后的测试集, 因此, 如果和本文一样采用取自相同实验室的数据集作为测试集, 那么 CoMFA 方法所得的 q_{pred}^2 还能在现有基础上得到一定程度的提高. 尽管如此, 本文所取得的 q_{pred}^2 值仍然是相当优异的. 应该指出的是, CoMFA 等方法在操作时, 需要确定分子的活性构象和叠加规则, 尤其是在其分子叠加过程中, 需要有丰富的化学和生物学知识, 在处理大量结构多样性的物质时, 需要消耗大量的时间, 而本文所提出的共轭梯度法改进 BP 算法人工神经网络, 尽管无法详细地从微观角度揭示污染物的毒性作用机理, 但是操作简单、迅速、预测性能高, 相比较具有很大的优越性.

3 结论

在本文中, 建立了用来预测分子内分泌活性并且基于结构多样性的雌激素 QSAR 模型, 这些雌激素的结构特征完全由 DRAGON 软件计算所得的分子描述符来表征. 通过客观变量选择, 以及主观变量选择——遗传算法筛选出与分子活性最相关的描述符. 使用共轭梯度法改进 BP 算法人工神经网络建立了分子描述符和雌激素受体相对亲和力的定量结构-活性关

表3 125种物质 logRBA 的实验值和共轭梯度法改进 BP 算法神经网络的预测值^{a)}

名称	实验值	预测值	残差	名称	实验值	预测值	残差
17 α -estradiol	0.49	0.76	0.27	4,2,4-trihydroxychalcone	-1.26	-1.83	-0.57
17 β -estradiol	2	1.09	-0.91	<i>phlorein</i>	-1.16	-1.80	-0.64
4-OH-estradiol	1.82	1.12	-0.70	α -zearalanol	1.48	0.63	-0.85
2-OH-estradiol	1.47	1.11	-0.36	β -zearalanol	-0.19	0.32	0.51
17-deoxyestradiol	1.14	2.02	0.88	zearalanone	0.32	-0.71	-1.03
3-deoxyestradiol	-0.3	0.73	1.03	α -zearalenol	1.63	0.85	-0.78
<i>estriol</i>	0.99	0.54	-0.45	β -zearalenol	-0.69	0.01	0.70
ethynylestradiol	2.28	1.09	-1.19	<i>bisphenol B</i>	-1.07	-1.09	-0.02
mestranol	0.35	0.70	0.35	<i>bisphenol A</i>	-2.11	-2.13	-0.02
3-methyl-estriol	-1.65	-0.28	1.38	cumyl phenol	-2.3	-2.15	0.15
estrone	0.86	-0.31	-1.17	diphenolic acid	-3.13	-2.16	0.97
3-deoxy-estrone	-2.2	-0.75	1.45	2,2'-methylenebis(4-chlorophenol)	-2.45	-1.70	0.75
<i>3-OH-estra-1,3,5(10)-trien-16-one</i>	-0.29	-0.89	-0.60	phenolphthalein	-3.67	-3.47	0.20
6 α -OH-estradiol	-0.15	0.20	0.35	phenolphthalin	-1.87	-2.18	-0.31
16 β -OH-16-methyl-3-methyl-estradiol estradiol	-1.48	-1.07	0.41	aurin	-1.5	-1.39	0.11
moxestrol	1.14	0.70	-0.44	<i>phenol red</i>	-3.25	-5.56	-2.31
ICI 182 780	1.57	1.34	-0.23	2,4-dihydroxybenzophenone	-2.61	-1.86	0.75
ICI 164 384	1.16	0.77	-0.39	4,4-dihydroxybenzophenone	-2.46	-2.05	0.41
<i>diethylstilbestrol (DES)</i>	2.6	2.48	-0.12	4,4-sulfonyldiphenol	-3.07	-3.23	-0.16
diethylstilbestrol monomethyl ether	1.31	0.68	-0.63	<i>monohydroxy methoxychlor</i>	-0.89	-1.73	-0.84
diethylstilbestrol dimethyl ether	-1.25	-0.29	0.96	HPTE	-0.6	-1.61	-1.01
dimethylstilbestrol	1.16	0.44	-0.72	monohydroxymethoxychlor olefin	-0.63	-0.47	0.16
4,4'-dihydroxystilbene	-0.55	-1.83	-1.28	dihydroxymethoxychlor olefin	0.42	0.68	0.26
meso-hexestrol	2.48	2.57	0.09	o,p'-DDT	-2.85	-2.43	0.42
<i>monomethyl ether hexestrol</i>	0.97	0.34	-0.63	2,3,4,5-tetrachloro-4'-biphenylol	-0.64	-1.57	-0.93
p-meso-phenol	0.6	0.34	-0.26	<i>2,5-dichloro-4-biphenylol</i>	-1.44	-1.32	0.12
2,6-dimethylhexestrol	1.11	1.82	0.71	4-chloro-4'-biphenylol	-2.18	-2.47	-0.29
4,4'-(1,2-ethanediy)bisphenol	-1.44	-2.22	-0.78	2-chloro-4-biphenylol	-2.77	-2.08	0.69
4-phenethylphenol	-2.69	-2.60	0.09	3,3,5,5-tetrachloro-4,4-biphenyldiol	-3.25	-2.82	0.43
tamoxifen	0.21	1.15	0.94	2,4-dichlorobipheny	-3.61	-2.84	0.77
4-OH-tamoxifen	2.24	1.38	-0.86	4-phenylphenol	-3	-3.27	-0.27
droloxifene citrate	1.18	1.66	0.48	3-phenylphenol	-3.44	-2.18	1.26
toremifene citrate	0.14	0.97	0.83	octylphenol	-2.31	-2.12	0.19
clomiphene	-0.14	-0.28	-0.14	nonylphenol	-1.53	-1.82	-0.29
<i>nafoxidine</i>	-0.14	0.48	0.62	<i>4-dodecylphenol</i>	-1.73	-1.49	0.24
triphenylethylene	-2.78	-2.56	0.22	4-tert-butylphenol	-3.61	-3.39	0.22
3,6,4'-trihydroxyflavone	-0.35	-2.59	-2.24	4-tert-amylphenol	-3.26	-3.39	-0.13
6,4-dihydroxyflavone	-0.82	-2.68	-1.86	4-tert-octylphenol	-1.82	-2.42	-0.60
apigenin	-1.55	-2.55	-1.00	4-sec-butylphenol	-3.37	-3.41	-0.04
kaempferol	-1.61	-2.36	-0.75	2-sec-butylphenol	-3.54	-3.47	0.07
<i>fisetin</i>	-2.35	-2.63	-0.28	4-ethylphenol	-4.17	-4.11	0.06
myricetin	-2.75	-2.28	0.47	3-ethylphenol	-3.87	-4.43	-0.56
baicalein	-3.05	-2.61	0.44	<i>4-chloro-3-methylphenol</i>	-3.38	-3.72	-0.34
morin	-3.35	-2.30	1.05	4-chloro-2-methylpheno			
rutin	-4.09	-4.11	-0.02	4-chloro-2-methylphenol	-3.67	-3.64	0.03
6-hydroxyflavone	-3.41	-2.94	0.47	2-chloro-4-methylphenol	-3.66	-2.92	0.74
7-hydroxyflavanone	-3.73	-2.79	0.94	Methyl paraben	-3.44	-2.92	0.52
6-hydroxyflavanone	-3.05	-2.75	0.30	ethylparaben	-3.22	-3.09	0.13
<i>3-hydroxyflavanone</i>	-2.78	-2.94	-0.16	<i>propylparaben</i>	-3.22	-3.62	-0.40
4-hydroxyflavanone	-2.65	-2.95	-0.30	butylparaben	-3.07	-3.12	-0.05
naringenin	-2.13	-2.47	-0.34	heptylparaben	-2.09	-2.34	-0.25
genistein	-0.36	-1.87	-1.51	benzylparaben	-2.54	-2.96	-0.42
<i>daidzein</i>	-1.65	-2.00	-0.35	4-heptyloxyphenol	-2.88	-2.69	0.19
3',4',7-trihydroxy isoflavone	-2.35	-1.88	0.47	4-benzyloxphenol	-3.44	-3.40	0.04
biochanin A	-2.37	-2.19	0.18	α,α -dimethyl- β -ethylallenolic acid	-0.02	-1.39	-1.37
equol	-0.82	-1.29	-0.47	doisynestrol	-2.74	-2.19	0.55
prunetin	-2.74	-2.06	0.68	<i>2,2,4,4-tetrahydroxybenzil</i>	-0.68	-1.70	-1.02
formononetin	-2.98	-2.19	0.79	nordihydroguaiaretic acid	-1.51	-0.59	0.92
coumestrol	-0.05	0.27	0.32	1,3-diphenyltetramethyldisiloxane	-3.16	-2.07	1.09
4-ethyl-7-OH-3-coumarin-2-one	-0.05	-0.52	-0.47	3α-androstanediol	-2.67	-2.23	0.44
chalcone	-2.82	-2.34	0.48	<i>3β-androstanediol</i>	-0.92	-2.05	-1.13
4'-hydroxychalcone	-2.43	-2.22	0.21	norethynodrel	-0.67	-0.73	-0.06
4-hydroxychalcone	-2.55	-1.64	0.91	dihydrotestosterone	-2.89	-2.96	-0.07

a) 表中斜体为监测集, 黑体为测试集/engine.scichina.com/doi/10.1360/csb2007-52-18-2116

表 4 本文结果与文献数据的比较

	CoMFA ^{a)}	CoMFA	HQSAR	KNN	ANN
R^2	0.90	0.91	0.76	0.88	0.845
q^2_{pred}	0.71 ^{b)} , 0.62 ^{c)}	0.68 ^{b)} , 0.63 ^{c)}	0.15, 0.22	0.73	0.81

a) 在传统的立体场和静电场中加入酚环指示变量; b) 测试集为 Waller 数据集; c) 测试集为 Kuiper 数据集

系, 而此 QSAR 模型与单纯采用 BP 网络所建立的模型比较, 无论在稳定性、预测能力和训练周期上, 均有显著改善, 克服了传统的 BP 人工神经网络模型的缺点. 通过与文献中用其他方法对相同雌激素数据集所建的 QSAR 模型的比较, 发现本研所得到的结果, 尤其在对外部样本的预测能力上是非常优异的. 除了令人满意的统计学结果外, 我们所使用的方法简单、迅速, 从活性预测与大样本筛选的角度来看, 可以应用于内分泌干扰物的大规模初筛.

参 考 文 献

- Cooper R L, Kavlock R J. Endocrine disruptors and reproductive development: A weight-of-evidence overview. *J Endocrinol*, 1997, 152(2): 159—166[DOI]
- Gray L E. Tiered screening and testing strategy for xenoestrogens and antiandrogens. *Toxicol Lett*, 1998, 102-103: 677—680[DOI]
- Waller C L, Oprea T I, Chae K, et al. Ligand-based identification of environmental estrogens. *Chem Res Toxicol*, 1996, 9(8): 1240—1248[DOI]
- Tong W, Perkins R, Xing L, et al. QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. *Endocrinology*, 1997, 138(9): 4022—4025[DOI]
- Shi L M, Fang H, Tong W, et al. QSAR models using a large diverse set of estrogens. *J Chem Inf Comput Sci*, 2001, 41(1): 186—195[DOI]
- Yu S J, Keenan S M, Tong W, et al. Influence of the structural diversity of data sets on the statistical quality of 3D-QSAR Models: Predicting the estrogenic activity of xenoestrogens. *Chem Res Toxicol*, 2002, 15(10): 1229—1234[DOI]
- Colemana K P, Toscano W A, Wiese T E. QSAR models of the in vitro estrogen activity of bisphenol A analogs. *QSAR Comb Sci*, 2003, 22: 78—88[DOI]
- Asikainen A, Ruuskanen J, Tuppurainen K. Spectroscopic QSAR methods and self-organizing molecular field analysis for relating molecular structure and estrogenic activity. *J Chem Inf Comput Sci*, 2003, 43(6): 1974—1981[DOI]
- Waller C L. A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J Chem Inf Comput Sci*, 2004, 44: 758—765[DOI]
- Haykin S. 神经网络原理. 叶世伟, 史忠植, 译. 北京: 机械工业出版社, 2004
- Salomon R, Hemmen J. Accelerating back propagation through dynamic self-adaptation. *Neural Netw*, 1996, 9(4): 589—601[DOI]
- Marini F, Roncaglioni A, Novic M. Variable selection and interpretation in structure-affinity correlation modeling of estrogen receptor binders. *J Chem Inf Model*, 2005, 45: 1507—1519[DOI]
- Bolanca T, Stefanovic S C, Regelja M, et al. Development of an inorganic cations retention model in ion chromatography by means of artificial neural networks with different two-phase training algorithms. *J Chromatogr A*, 2005, 1085: 74—85[DOI]
- Engin M, Demirag S, Engin E Z, et al. The classification of human tremor signals using artificial neural network. *Expert Syst Appl*, 2007, 33: 754—761[DOI]
- Yao X J, Panaye A, Doucet J P. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. *J Chem Inf Model*, 2004, 44: 1257—1266
- Liu H, Papa E, Gramatica P. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. *Chem Res Toxicol*, 2006, 19: 1540—1548[DOI]
- Shepherd A J. Second-Order Methods for Neural Networks. New York: Springer, 1997
- Fang H, Tong W, Shi L M, et al. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem Res Toxicol*, 2001, 14: 280—294[DOI]
- Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH, 2000
- Johhnson S R. Prediction of physicochemical properties and biological activities from molecular structure and the use of computational neural networks for the analysis of sensor apply data. Doctor Dissertation. Pennsylvania: The Pennsylvania State University, 1999
- Mcelroy N R. The prediction of physical properties and biological activities of organic compounds from their molecular structures. Doctor Dissertation. Pennsylvania: The Pennsylvania State University, 2003
- Rogers D, Hopfinger A J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J Chem Inf Comput Sci*, 1994, 34: 854—866[DOI]
- Statsoft. Statistica Neural Networks. Tulsa: Statsoft Inc, 1999
- 闻新, 周露, 李翔, 等. MATLAB 神经网络仿真与应用. 北京: 科学出版社, 2003. 272—274
- Cramer R D, Patterson D E, Bunce J D. Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc*, 1988, 110: 5959—5967[DOI]
- Asikainen A, Ruuskanen J, Tuppurainen K. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. *Environ Sci Technol*, 2004, 38(24): 6724—6729[DOI]