中国科学: 化学 2011 年 第 41 卷 第 5 期: 893 ~ 899
SCIENTIA SINICA Chimica www.scichina.com chem.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

基于启发式变量筛选方法研究与雌激素β受体结合的化合物结构与活性之间的关系

张一鸣[①], 杨旭曙[②*], 孙成[③], 王连生[③]

① 南京医科大学基础医学院, 南京 210029
② 南京医科大学药学院, 南京 210029
③ 污染控制与资源化研究国家重点实验室; 南京大学环境科学与工程学院, 南京 210093
*通讯作者, E-mail: yangxushu@sina.com

收稿日期: 2010-05-05; 接受日期: 2010-05-27; 网络版发布日期: 2010-09-14
doi: 10.1007/s11426-010-4077-x

摘要 雌激素类化合物由于其对人和野生动物健康的负面影响而受到广泛关注. 雌激素受体存在两种亚型(ERα 和 ERβ), 化合物与两种受体亚型在结合活性和化合物结构特征方面存在差异. 以31种与雌激素β受体亚型(ERβ)结合的化合物为研究对象, 采用启发式变量筛选方法, 从 1524 个变量中筛选出 5 个与化合物活性(lgRBA)最相关的变量, 然后采用多元线性回归(MLR)建立最佳预测模型. 模型相关性显著, 而且具有良好的稳健性和预测能力(r^2 = 0.829, q^2_{LOO} = 0.742, r^2_{pred} = 0.772, q^2_{ext} = 0.724, RMSEE = 0.395). 同时揭示了影响化合物与 ERβ 受体结合的配体化合物分子的结构特征, 并对模型的应用域进行了研究.

关键词
雌激素 β 受体(ERβ)
定量结构与活性相关(QSAR)
启发式变量筛选
模型应用域

1 引言

研究表明, 许多雌激素类化合物能模拟内源性雌激素化合物的功能, 与雌激素受体(ER)结合, 使雌激素受体(ER)的活性构象发生改变, 激活或抑制有关细胞生长和发育的靶细胞的转录, 产生各种雌激素效应, 从而对人和野生动物的健康产生严重的负面影响[1]. 雌激素受体有两种亚型—— 雌激素α受体(ERα)和雌激素 β 受体(ERβ). 这两种受体在组织细胞中的分布和调节基因转录方面存在明显差异[2]. ERα 主要是作为基因转录的激动剂, 而 ERβ 主要是作为基因转录的抑制剂[3]. 探讨雌激素类化合物与两种雌激素受体亚型的不同结合活性和结构特征是环境激素领域研究的难点. 目前对雌激素类化合物的结构和活性方面的研究主要集中在与 ERα 结合的化合物上[4~9], 而对与 ERβ 结合的化合物的研究相对较少.

定量结构与活性相关技术(QSAR)可以直接基于化合物分子的结构信息, 实现对化合物活性的准确预测, 并能揭示影响化合物活性的分子结构特征, 因此在环境毒理学领域有着广泛的应用[4~9]. QSAR 方法大致可分成两类—— 传统的 QSAR(C-QSAR)方法和三维 QSAR(3D-QSAR)方法. 目前, C-QSAR 方法和 3D-QSAR 方法已广泛应用于研究雌激素类化合物结构与活性之间的定量关系[4~13]. 相对于 3D-QSAR 方法, C-QSAR 方法无需进行分子叠合, 且具有计算速度快等优势[5, 8, 13]. 目前表征化合物结构的描述符有几千种, 如何选择最佳描述符构建稳健性高、预测能力强的预测模型是 C-QSAR 领域研究的难点. 目前的变量选择方法有偏最小二乘法(PLS)[14]、k 个最

近邻域法(kNN)[9]、人工神经网络方法(ANN)[6, 8]、遗传算法(GA)[5]和启发式变量筛选(HM)[13]等方法. 采用 PLS 和 kNN 方法筛选出的描述符较多, 难以对模型进行有效的解释, ANN 方法存在黑箱.

目前 QSAR 研究对构建的模型有以下几个要求[5, 13]: (1)确定的终点(化合物的活性); (2)明确的算法; (3)确定的应用域; (4)显著的相关性、良好的稳健性和预测能力; (5)模型易于解释. Tong 等[10]和 Xing 等[11]分别应用比较分子力场分析方法(CoMFA)研究了与 ERβ 结合的化合物结构与活性之间的定量关系, 均建立了具有一定预测能力的 QSAR 模型, 但 CoMFA 模型存在分子叠合方面的困难, 且模型均未采用独立的外部样本集进行外部验证, 并缺少模型应用域方面的研究.

本文应用启发式变量筛选方法对表征 31 种与 ERβ 受体结合化合物的 1524 种描述符进行变量选择, 筛选出 5 种与化合物活性最相关的描述符, 建立了相关性显著、稳健性和预测能力强的 QSAR 模型. 同时揭示了影响化合物与 ERβ 受体结合活性的分子结构特征, 并对模型进行了外部验证和应用域方面的研究.

2 方法与步骤

2.1 化合物与活性数据

本文用于构建模型的训练集化合物及其活性数据取自文献[15], 用于模型验证的测试集化合物活性数据取自文献[16]. 化合物活性指标为其与 ERβ 受体相对结合力的对数(lgRBA). 训练集包括 31 个化合物, 其中含 15 个类固醇、3 个 1,2-二苯乙烯雌激素、4 个三苯乙烯雌激素、4 个雄激素、3 个植物雌激素和 2 个环境雌激素. 训练集典型化合物的结构与活性见表 1. 测试集包括 15 个化合物为(见表 2), 其中含 4 个类固醇、4 个羟基多氯联苯、4 个植物雌激素、1 个 DDT 类、1 个取代苯酚和 1 个合成雌激素. 训练集和测试集两组活性数据虽来自同一实验室, 但其活性数据存在差异, 所以在对模型进行验证之前, 要对测试集数据进行标准化处理[4, 5]. 具体过程是: 选取两文献中相同化合物的两批活性数据, 建立相关方程, 然后以训练集活性数据为基准, 通过建立的相关方程对测试集活性数据进行标准化处理, 最后以标准化后的活性数据对所建模型质量进行评价[4, 5, 17]. 本文选取文献[15]和[16]中 9 个相同且具有活性的化合物, 根据其活性数据建立相关方程, 相关关系显著(见图 1).

2.2 分子模拟和描述符计算

采用 Chemoffice8.0 软件构建化合物结构, 采用 MOPAC 中的 AM1 方法对所有化合物结构进行能量最优化处理. 再用 Dragon2.1[18]和 Chemoffice8.0 软件计算化合物分子描述符, 共算得全面表征化合物结构信息的共 1524 个分子描述符. 由 Dragon2.1 软件算得的描述符包括化合物组成描述符、一维官能团描述符、一维原子中心碎片描述符、二维拓扑描述符、二

表 1 训练集化合物分类、数量和典型化合物结构名称

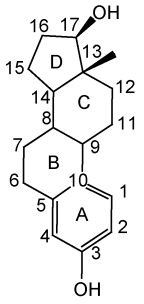
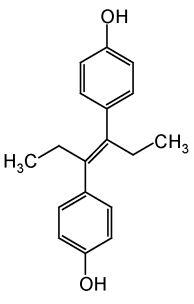
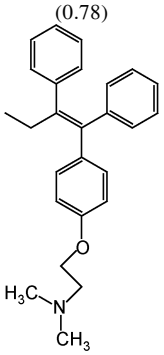
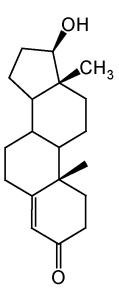
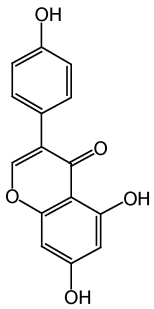
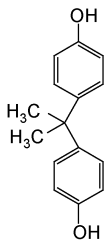
化合物类别	类固醇	二苯乙烯雌激素	三苯乙烯雌激素	雄激素	植物雌激素	环境雌激素
数目	15	3	4	4	3	2
典型化合物 (LgRBA)	17β-雌二醇 (2.00)	DES (2.47)	它莫西芬 (0.78)	睾酮 (−2.10)	5,4′-三羟(基)异黄酮 (1.56)	双酚 A (−0.48)
分子结构						

表 2　训练集和测试集化合物名称、活性(LgRBA)的实验值与预测值及残差

名称	实验值	预测值	残差	名称	实验值	预测值	残差
E2	2.00	1.32	0.68	estriol	1.32	1.42	−0.10
testosterone	−2.10	−1.53	−0.57	norethynodrel	−0.66	−0.71	0.05
estrone	1.57	−0.26	1.83	coumestrol	2.27	2.34	−0.07
2-hydroxy-estradiol	1.04	1.27	−0.23	genistein	1.56	1.52	0.04
4-hydroxy-estradiol	0.85	1.11	−0.26	β-zearalanol	1.15	1.21	−0.06
moxestrol	0.70	0.32	0.38	nafoxidine	1.20	1.31	−0.11
ICI 164384	2.22	1.78	0.44	bisphenol A	−0.48	0.16	−0.64
17α estradiol	1.04	1.21	−0.17	p,p′-methoxychlor	−0.89	−0.85	−0.04
3α-androstanediol	−0.52	0.00	−0.52	2-OH-Estrone [a)]	0.26	0.40	−0.14
3β-androstanediol	0.85	0.06	0.79	17-epiestriol [a)]	1.98	1.46	0.52
4-androstenediol	−0.22	0.77	−0.99	5-androstenediol	1.23	0.34	0.89
16-keto-17β-estradiol [a)]	0.69	0.43	0.26	16α-bromo-17β-estradiol [a)]	1.38	0.76	0.62
dehydroepiandrosteron	−1.15	−1.04	−0.11	2′,3,3′,5′, 6′-pentac hloro-4-biphenylol [a)]	−0.20	−0.81	−0.61
5α-dihydrotestoster	−0.77	−1.55	−0.78	o,p′-DDT [a)]	−0.40	−0.66	−0.26
nandrolone	−0.64	−1.36	−0.72	raloxifene [a)]	1.52	1.43	0.09
norethindrone	−2.00	−1.17	−0.83	clomiphene	1.08	1.20	−0.12
4-hydroxy-tamoxifen	2.53	1.99	0.54	2′,4′, 6′-trichloro -4-biphenylol [a)]	1.17	1.28	−0.11
2′,3,3′,4′, 5′-pentachloro-4-biphenylol [a)]	0.09	−0.33	0.42	2,3,3′,4′, 5-pentachloro-4-biphenylol [a)]	−0.40	−0.38	0.02
progesterone	−3.50	−2.47	−1.03	4-tert-octylphenol [a)]	−0.28	−0.28	0.00
DES	2.47	2.84	−0.37	zearalenone [a)]	1.19	1.54	−0.35
hexestrol	2.37	2.80	−0.43	daidzein [a)]	0.53	0.46	0.26
dienestrol	2.61	2.28	0.33	apigenin [a)]	1.24	1.41	−0.17
tamoxifen	0.78	1.59	−0.81	naringenin [a)]	0.09	0.99	−0.90

a)为测试集, 其活性为经标准化处理后的 LogRBA, 其他化合物为训练集

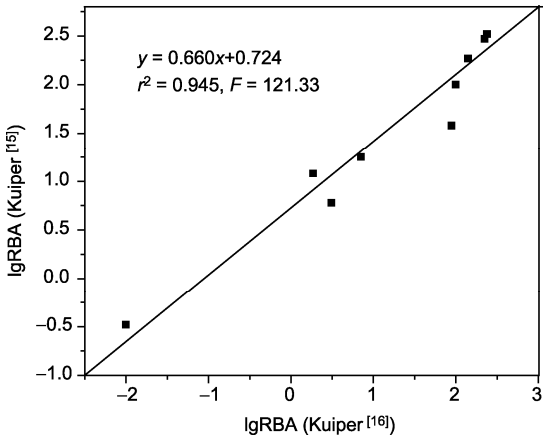


图 1　Kuiper[15, 16]化合物活性数据相关图

维自相关描述符、二维连接性指数描述符、三维几何描述符、三维 WHIM 描述符和 GATAWAY 描述符等共 1481 个, 由 Chemoffice8.0 软件算得的描述符包括物理化学描述符、电子特征描述符、立体特征描述符和量子化学描述符等共 43 个.

2.3　启发式变量筛选和模型构建

客观变量筛选: (1)剔除对某些化合物不能计算的描述符; (2)剔除对所有化合物其值都相同或几乎相同的描述符; (3)剔除相关系数高的描述符($r > 0.9$).

主观变量筛选: 先剔除在单变量关系式中不具有统计学意义的变量($p > 0.05$), 然后采用逐步变量筛选方法, 根据使化合物活性与变量的相关系数(r^2)和 F-检验值(F-test)最大的原则, 逐步挑选 1 个、2 个直至 n 个描述符变量. 为防止模型过度拟合, 一旦 r^2 的增加值不超过 0.02(即 $r^2_n - r^2_{n-1} \leqslant 0.02$), 则停止变量筛选[19].

模型构建: 由挑选出的最佳变量, 应用最佳子集回归[20]构建最佳 QSAR 模型. 用非交叉验证相关系数(r^2)、F-检验值(F-test)、统计学显著性(p)表征模型的拟合程度. 用交叉验证相关系数(q^2_{LOO})、预测相关系数(r^2_{pred})、外部验证系数(q^2_{ext})、交叉验证均方根误差(RMSEP)和测试集均方根误差(RMSEE)表征模型的稳健性和预测能力[5, 7, 12, 21, 22].

2.4 模型的应用域

本文采用臂比方法, 即以化合物的臂比值(h_i)为横坐标、以化合物活性的交叉验证标准残差为纵坐标绘制 Williams 图研究模型的应用域. 化合物的臂比值(h_i)定义为

$$h_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i \quad (i = 1, 2, \ldots, n) \qquad (1)$$

式(1)中, x_i 为求解的化合物描述符向量; X 为由训练集描述符组成的矩阵. 预警臂比 h^*定义为:

$$h^* = 3\times(p'+1)/n \quad (i = 1, 2, \ldots, n) \qquad (2)$$

式(2)中, p'为模型参数的数量; n 为训练集化合物的数量. 如果某一化合物的臂比值 h_i 大于预警臂比值 h^*, 则由模型所预测的化合物响应可能是不可靠的.

3 结果与讨论

3.1 模型的构建与验证

经过客观变量筛选后, 共保留了 94 个描述符. 然后采用主观变量筛选, 筛选出与化合物活性最相关的 GATS1e、MATS3m、Mor16u、Du、nDB 和 MATS8v 共 6 个描述符. 由于在变量筛选过程中, 描述符 GATS8v 和 Mor23v 与筛选出的其他描述符具有相近的 r^2 和 F-test 值, 所以再加上这两个描述符, 共 8 个描述符用于最佳子集回归[20]. 允许的变量自相关系数设定为 $r_{\mathrm{int}} = 0.70$, 在不同变量数目下运行最佳子集回归程序, 所得到的变量优化结果及相应统计量见表 3. 由于 $r^2_6 - r^2_5 < 0.02$, 故选择最佳变量数 $m = 5$, 最佳子集为 GATS1e、MATS3m、Mor16u、Du 和 nDB. 用最佳子集建立最佳多元线性回归模型(MLR):

LgRBA = −(42.14±13.64)+(4.41±0.68)GATS1e+(41.61±13.01)MATS3m+(1.32±0.41)Mor16u−(13.13±5.13)Du+(0.57±0.28)nDB (3)

p: 0.000001, 0.004, 0.003, 0.02, 0.05

VIF(方差膨胀因子): 1.8, 1.3, 1.2, 1.5, 2.2

$N = 31$, $m = 5$, $r^2 = 0.829$, RMSEE = 0.62, $q^2_{\mathrm{LOO}} = 0.742$, RMSEP = 0.77, F-test = 24.20, $p < 0.0001$.

数据经过标准化后, 得(4)式:

lgRBA = 0.73GATS1e + 0.30MATS3m + 0.29Mor16u− 0.26Du + 0.25nDB (4)

对式(4)中 5 个描述符再进行相关分析, 其相关矩阵见表 4. 由表 4 可见, 式(4)中任意两变量线性不相关(相关系数 $r < 0.8$).

为更好地说明模型的预测能力, 需要对模型进行外部验证[5, 21, 22]. 本文以 15 个化合物组成的独立样本数据集为测试集, 对所建模型进行验证. 外部验证的统计参数如下: 预测相关系数 $r^2_{\mathrm{pred}} = 0.772$, 外部验证系数 $q^2_{\mathrm{ext}} = 0.724$, 测试集均方根误差 RMSEE = 0.395. 这说明模型具有稳健的预测能力. 训练集中 31 个化合物和测试集中 15 个化合物活性的实验值与预测值结果见表 2, 实验值和预测值的相关图见图 2.

3.2 模型的应用域

由 Williams 图(图 3), 训练集中所有化合物的臂

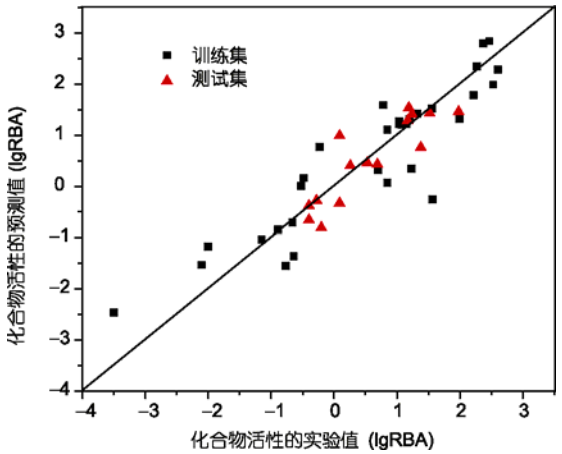


图 2 训练集和外部测试集实验值与预测值的相关图

表 3 与 ERβ 结合化合物的 LogRBA 最佳变量子集及相关统计参数

m	r^2	RMSEE	q^2_{LOO}	RMSEP	变量名称
1	0.520	1.04	0.448	1.12	GATS1e
2	0.700	0.82	0.636	0.91	GATS1e, MATS3m
3	0.773	0.72	0.692	0.84	GATS1e, MATS3m, Mor16u
4	0.799	0.67	0.713	0.81	GATS1e, Mor16u,Du, GATS8v
5	0.829	0.62	0.742	0.77	GATS1e, MATS3m, Mor16u, Du, nDB
6	0.840	0.60	0.747	0.76	GATS1e, MATS3m, Mor16u, Du, nDB, GATS8v
7	0.846	0.59	0.724	0.80	GATS1e, MATS3m, Mor16u, Du, nDB, GATS8v, Mor23v

表 4　最佳子集自相关系数

	GATS1e	MATS3m	Mor16u	Du	nDB
GATS1e	1.000				
MATS3m	0.143	1.000			
Mor16u	0.217	0.168	1.000		
Du	−0.188	-0.479	−0.253	1.000	
nDB	−0.700	-0.234	−0.356	0.409	1.000

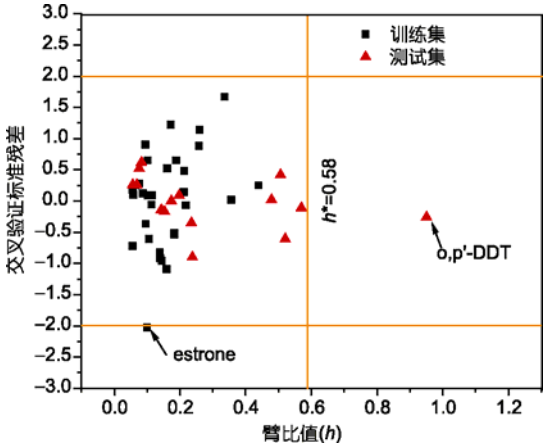


图 3　训练集和测试集的 Williams 图

比值均小于临界值(h^*), 化合物的交叉验证标准残差也均介于±2.5σ(标准差的 2.5 倍), 如果将化合物的交叉验证标准残差设为 2.0σ, 则化合物雌酮(estrone)的值为离群值, 删除化合物雌酮后, 模型的相关性和稳健性以及内部预测能力都有显著的提高, r^2 = 0.880, RMSEE = 0.53, q^2_{LOO} = 0.812, RMSEP = 0.66. 测试集中所有化合物的交叉验证标准残差都介于±2.0σ, 但化合物 o,p′-DDT 的臂比值超过临界值(h^*). 这说明模型(4)对此化合物活性的预测结果可能是不可靠的, 但预测的结果显示模型还是对化合物 o,p′-DDT 的活性作了较为准确的预测.

3.3　模型的解释

模型(3)和(4)中, GATS1e、MATS3m 均是二维自相关描述符[23]. 反映沿分子拓扑结构的原子特征信息. GATS1e 表示拓扑步长为 1、经原子电负性权重后的 Geary 自相关指数. MATS3m 表示拓扑步长为 3、经原子质量权重后 Moran 自动相关指数. Mor16u 为基于电子衍射的分子结构描述符, 反映分子的骨架和取代基的结构[24]. Du 为 WHIM 描述符[23], 表示未权重的总体可及性指数. nDB 为组成描述符, 表示分子中双键的个数[23]. 由模型(4)可知, 影响化合物与 ERβ 结合活性最重要的描述符为 GATS1e, 化合物与 ERβ 的结合活性随 GATS1e 值的增加而增大. GATS1e 反映配体化合物原子电负性方面的结构信息, 这可能是 GATS1e 所表征的分子结构信息影响配体化合物与 ERβ 受体结合腔中氨基酸残基之间的分子间作用力. 分子对接研究表明在雌激素 β 受体的配体结合腔中, 在雌二醇 A 环 C2 和 C3 以及 B 环 C6 附近分别出现残基 Leu339、Glu305 和 Leu298 的负电性强的羰基氧原子, 而在 A 环 C3 和 D 环 C-17β 位置附近分别出现正电性强的残基 Arg346 氨基氢原子和残基 His475 咪唑环亚氨基氢原子, 因此这些位置原子或基团的电负性可能会对化合物与 ERβ 受体的结合活性产生重要的影响[12]. Tong[10]等和 Xing[11]等的 3D-QSAR 研究也均表明配体化合物周围的静电场对配体与 ERβ 受体之间结合力的贡献最大. 描述符 MATS3m 反映分子大小的结构信息. 配体化合物取代基的大小和位置可能会影响化合物与 ERβ 受体之间的结合活性. 研究表明在雌激素 β 受体的配体结合腔中, 在雌二醇 C-7α 位置由残基 Asp303、Leu306、Trp335、Leu476 和 Val485 形成一个缺口, 此处可容纳大的取代基; B 环 C-6β 位置靠近残基 Leu298, 故此处只能容纳小的取代基[12]. 由模型(4), 双键的个数(nDB)也会影响化合物与 ERβ 受体的结合活性. 这可能是在 ERβ 受体的配体结合腔中含有较多的疏水性残基, 较多的双键可能会增加化合物的疏水性, 从而有利于提高化合物与 ERβ 受体的结合活性. 综上所述, 配体化合物中原子或取代基的电负性、分子大小、疏水性等决定了其与 ERβ 受体的结合活性.

3.4　与其他文献的比较

Tong 等[10]和 Xing 等[11]也分别用比较分子力场分析方法(CoMFA)对 Kuiper 数据集[15]做过研究. 由于 CoMFA 方法要求对训练集中所有化合物根据其立体场和静电场进行最大叠合, 而与 ERβ 受体结合的化合物来源广泛, 结构复杂, 对化合物分子进行最佳叠合存在困难, 需要研究者具备大量的化学和生物学知识, 同时要消耗较多的时间, 而且模型的稳健性和预测能力往往并不很高. Tong 等[10] CoMFA 模型的交叉验证系数 q^2_{LOO} 只有 0.60, Xing 等[11]虽对化合物的分子叠合方法进行改进, q^2_{LOO} 值有所提高, 但也只达到 0.646. 而本文采用启发式变量筛选方法筛选出 5 个最佳变量所构建的模型其 q^2_{LOO} 值达到了 0.742.

且 Tong 等[10]和 Xing 等[11]模型缺少外部验证, 同时也没有模型应用域方面的研究. 本文采用独立的外部样本数据集对模型进行验证, 结果模型具有良好的预测能力(r^2_{pred} = 0.772, 外部验证系数 q^2_{ext} = 0.724, 测试集均方根误差 RMSEE = 0.395). 这说明本模型的预测能力和稳健性要明显优于 Tong 等[10]和 Xing 等[11]采用比较分子力场分析方法(CoMFA)所构建的模型.

4 结论

本文以 31 种与 ERβ 受体结合的化合物为研究对象, 采用启发式变量筛选方法筛选出 5 种与化合物活性最相关的描述符表征化合物的分子结构, 然后采用多元线性回归(MLR)建立配体分子结构与其活性之间的定量模型. 通过交叉验证和外部测试集验证, 表明此模型具有良好的预测能力和稳健性. 通过文献比较, 此模型的预测能力和稳健性明显优于 3D-QSAR 模型. 模型还表明化合物与 ERβ 受体的结合活性与化合物中原子或取代基的电负性、分子大小和疏水性等因素有关. 同时对模型的应用域进行了相应的探讨.

致谢 本工作得到南京医科大学科技发展基金重点项目(09NJMUZ16)资助, 特此致谢.

参考文献

1 Kavlock RJ, Daston GP, Derosa C, Fenner-Crisp P, Gray LE, Kaattari S, Lucier G, Luster M, Mac MJ, Maczka C, Miller R, Moore J, Rolland R, Scott G, Sheehan DM, Sinks T, Tilson HA. Research needs for the risk assessment of health and environmental effects of endocrine disruptors: A report of the U.S. EPA-sponsored workshop. *Environ Health Perspect*, 1996, 104 (suppl 4): 715–740

2 Diel P. Tissue-specific estrogenic response and molecular mechanisms. *Toxicol Lett*, 2002, 127: 217–224

3 Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, Greene GL. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 1998, 95: 927–937

4 Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair R, Branham W, Sheehan D. QSAR models using a large diverse set of estrogens. *J Chem Inf Comput Sci*, 2001, 41: 186–195

5 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

6 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

7 杨旭曙, 王晓栋, 季力, 李荣, 孙成, 王连生. 分子对接结合比较分子相似性指数分析用于雌激素类化合物活性预测和分子机理研究. 科学通报, 2008, 11: 2735–2741

8 季力, 王晓栋, 杨旭曙, 刘树深, 王连生. 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. 科学通报, 2007, 52: 2116–2121

9 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: 6724–6729

10 Tong W, Perkins P. QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology*, 1997, 138: 4022–4025

11 Xing L, Welsh WJ, Tong W, Perkins R, Sheehan DM. Comparison of estrogen receptor α and β subtypes based on comparative molecular field anylysis (CoMFA). *SAR QSAR Environ Res*, 1999, 10: 215–237

12 杨旭曙, 王晓栋, 罗斯, 季力, 秦良, 李荣, 孙成, 王连生. 基于雌激素 β 受体结构雌激素化合物的三维定量结构-活性关系与分子对接研究. 中国科学 B 辑: 化学, 2009, 39: 459–468

13 Li F, Chen J, Wang Z , Li J, Qiao X. Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR. *Chemosphere*, 2009, 74: 1152–1157

14 Kurunczi L, Seclaman E, Oprea TI, Crisan L, Simon Z. MTD-PLS: A PLS variant of the minimal topologic difference method. III. Mapping interactions between estradiol derivatives and the alpha estrogenic receptor. *J Chem Inf Model*, 2005,45: 1275–1281

15 Kuiper GGJM, Carlsson B**,** Grandien K, Enmark E, Häggblad J, Nilsson S, Gustafsson JÅ. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β. *Endocrinology*, 1997, 138: 863–870

16 Kuiper GGJM, Lemmen JG, Carlsson B, Corton JC, Safe SH, van der Saag PT, van der Burg B, Gustafsson JÅ. Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor β. *Endocrinology*, 1998, 139: 4252–4263
17 Fang H, Tong W, Perkins R, Soto AM, Prechtl NV, Sheehan DM. Quantitative comparison of in vitro assays for estrogenic activity. *Environ Health Perspect*, 2000, 108: 723–729
18 Todeschini R, Consonni V, Mauri A, Milano PM. DRAGON, Version 2.1 for Windows, Software for the calculation of molecular descriptors. 2002, Talete srl, Milan, Italy
19 Katritzky AR, Fara DC, Yang HF, Karelson M, Suzuki T, Solovév VP. Quantitative structure-property relationship modeling of β-cyclodextrin complexation free energies. *J Chem Inf Comput Sci*, 2004, 44: 529–541
20 Liu SS, Liu HL, Yin CS, Wang LS. VSMP: A novel variable selection and modeling method based on the prediction. *J Chem Inf Comput Sci*, 2003, 43: 964–969
21 Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*, 2003, 22: 69–76
22 Golbraikh A, Tropsha A. Beware of *q*2! *J Mol Graph Model*, 2002, 20: 269–276
23 Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Weinheim: Wiley-VCH, 2000
24 Gasteiger J, Sadowski J, Schuur J. Chemical Information in 3D Space. *J Chem Inf Comput Sci*, 1996, 36: 1030–1037

Quantitative structure-activity relationship of compounds binding to estrogen receptor β based on heuristic method

ZHANG YiMing[1], YANG XuShu[2], SUN Cheng[3] & WANG LianSheng[3]

1 School of Basic Medical Sciences, Nanjing Medical University, Nanjing 210029, China
2 School of Pharmacy, Nanjing Medical University, Nanjing 210029, China
3 State Key Laboratory of Pollution Control and Resources Reuse, School of Environment, Nanjing University, Nanjing 210093, China

Abstract: Estrogen compounds may pose a serious threat to the health of humans and wildlife. The estrogen receptor (ER) exits as two subtypes, ERα and ERβ. Compounds might have different relative affinities and binding modes for ERα and ERβ. In this study, heuristic method was performed on 31 compounds binding to ERβ to select 5 variances most related to the activity (LogRBA) from 1524 variances, which were then employed to develop the best model with the significant correlation and the best predictive power (r^2 = 0.829, q^2_{LOO} = 0.742, r^2_{pred} = 0.772, q^2_{ext} = 0.724, RMSEE = 0.395) using multiple linear regression (MLR). The model derived identified critical structural features related to the activity of binding to ERβ. The applicability domain (AD) of the model was assessed by Williams plot.

Keywords: estrogen receptor β (ERβ), quantitative structure-activity relationship (QSAR), heuristic method, applicability domain