DOI:10.7524/j.issn.0254-6108.2013.07.012

QSAR 模型内部和外部验证方法综述*

覃礼堂[1,2] 刘树深[1,3]** 肖乾芬[1,3] 吴庆生[1,2]

(1. 同济大学长江水环境教育部重点实验室, 上海, 200092;2. 同济大学化学系, 上海, 200092;
3. 同济大学环境科学与工程学院, 上海, 200092)

摘 要 验证定量-结构活性相关(QSAR)模型,是保证模型对未知样本的生物活性具有可靠预测能力的重要前提. 然而,目前部分 QSAR 论文没有对模型进行有效验证. 因此,本文详细综述 QSAR 模型的内部验证方法和外部验证方法. 内部验证方法包括留一法(leave-one-out,LOO)交叉验证,留多法(leave-many-out,LMO)或留 N 法(leave-N-out,LNO)交叉验证,y 随机化验证和自举法. 评价模型外部预测能力的统计量包括 Q^2_{F1} 、Q^2_{F2} 、Q^2_{F3} 、一致性相关系数(concordance correlation coefficient,CCC)、r^2_m 和 Golbraikh-Tropsha 方法. 此外,从文献中总结出可接受 QSAR 模型对应的统计量参考数值,从而为 QSAR 建模者提供指导与帮助.

关键词 QSAR, 内部验证, 外部验证.

最近几十年,国内外大量文献报道定量结构-活性/属性相关(QSAR/QSPR)模型. 王连生教授作为我国有机污染物定量构效关系研究领域的开创者,为我国的 QSAR 研究做出了突出的贡献. 为了检索国内已报道的 QSAR 相关论文,以主题为“QSAR”或“QSPR”或“定量构效”或“定量结构-活性相关”,在中国知识基础设施工程(CNKI)上检索期刊论文,截止 2013 年 1 月 21 日,共 2584 篇文献.

经济合作与发展组织(OECD)提出 QSAR 模型需遵循 5 个法则[1-3]:(1)确定的终点;(2)明确的运算方法;(3)定义应用范围;(4)适当验证模型拟合优度、稳健性和预测能力;(5)如果可能,进行机理解释. 建立 QSAR 模型的目的通常是为了:(1)预测未测定或新化合物的生物活性;(2)确定哪些分子结构属性决定化合物的生物活性[4]. 例如,在药物学研究中,通过 QSAR 研究可以修改药物分子结构进而提高药效或更进一步理解生物学机理[5]. 在 OECD 的 5 个法则中,法则(4)是 QSAR 模型最重要的法则之一. 然而,已报道的部分 QSAR 模型并没有进行充分地验证. 在 CNKI 的 2584 篇 QSAR 期刊论文中,为了检索已验证的论文,以主题为“QSAR”或“QSPR”或“定量构效”或“定量结构-活性相关”,并且主题为“验证”或“检验”,在 CNKI 上检索期刊论文,截止 2013 年 1 月 21 日,仅 584 篇文献. 从 1994 年至 2012 年各年度的已验证和未验证的 QSAR 期刊论文发表情况列于图 1. 由此可知,国内大部分 QSAR 论文没有严格地进行验证或检验.

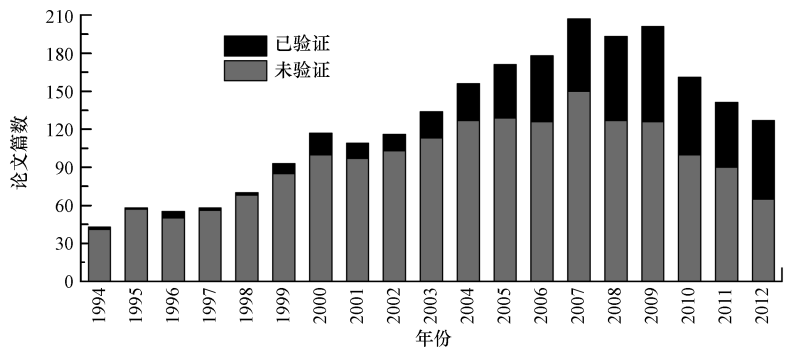

图 1 1994—2012 年度 CNKI 的 QSAR 期刊论文发表情况
Fig. 1 Journal papers of QSAR published in CNKI between 1994 and 2012

2013 年 1 月 25 日收稿.

* 国家自然科学基金(21177097)资助;中国博士后科学基金(2012M520932)资助.

** 通讯联系人, Tel:021-65982767;E-mail:ssliuhl@263.net

为此,本文详细综述 QSAR 模型的内部验证和外部验证方法,为 QSAR 建模者提供指导与帮助. 这些内部验证和外部验证可作为 QSAR 的验证方法,以保证回归模型的可靠性和有效性. 此外,从文献中总结可接受的 QSAR 模型对应的统计量参考值.

1 QSAR 模型内部验证方法

严格的 QSAR 模型验证程序应包括内部验证和外部验证. 内部验证方法包括留一法(LOO)交叉验证、留多法(leave-many-out,LMO)或留 N 法(leave-N-out,LNO)交叉验证、y 随机化验证和自举法等.

1.1 LOO 交叉验证

LOO 交叉验证是模型内部验证最简单的方法之一[6]. 假设对于含 n 个样本的数据集,LOO 交叉验证步骤如下:

(1) 抽出第 1 个样本作为外部检验样本,余下的 $n-1$ 个样本作为训练集建立回归模型,并用这个模型去预测抽出的作为外部检验样本的因变量值.

(2) 将第 1 个样本放回原样本数据集,依次抽出第 2 个样本作为外部检验样本,同样以余下的 $n-1$ 个样本作为训练集建立回归模型,并预测第 2 个样本的因变量值.

(3) 将第 2 个样本放回原样本数据集,依次按照"抽出 1 个样本→余下样本建模→预测抽出样本→放回抽出样本"的顺序对原样本集进行操作,直到所有样本均被抽出一次并进行预测为止.

完成 LOO 交叉验证后,计算 n 次抽出样本的因变量 LOO 预测值($\hat{y}_i$)与原抽出样本的因变量实验值(y_i)之间的相关系数(Q^2_{LOO})及 LOO 交叉验证均方根误差(RMSECV),以评价模型内部预测能力.

1.2 LMO/LNO 交叉验证

LMO 或 LNO 交叉验证[7-10]也是检验模型稳健性的另一种方法. LMO 与 LOO 的区别是,LMO 的计算过程每次从数据集中抽出多个样本,用剩余的样本建模并预测被抽出的多个样本,该过程重复多次. 假设有 n 个样本的数据集,将数据集分成相等大小的 m 组($m = n/G$),其中 G 常常选择 2—10. 然后以 $n-m$个训练集样本建立多个 QSAR 模型,并预测被抽出的 m 个样本,计算 LMO 交叉验证相关系数(Q^2_{LMO}). 如果一个模型在 LMO 验证中得到一个高的 Q^2_{LMO},那么模型是稳健的.

在 LOO 交叉验证中,对于样本数为 n 的训练集,需要 n 次交叉验证. 在 LMO 交叉中,训练集中 n 个样本的顺序对 LMO 的结果将产生一定的影响. 假设取 $M=2$,即 L2O 交叉验证,对于给定顺序的 n 个样本训练集,需要进行 $n/2$ 次交叉验证并获得 $n/2$ 个模型. 然而,该验证仅是所有可能 2 个样本组合中($n!/(n-2)!$)的一种组合. 因此,Kiralj 和 Ferreira 建议将数据集中样本随机排序后再进行 LMO 交叉验证[6]. 在一些 LMO 交叉验证中,数据集进行多次随机化(如 10 次),取多个 Q^2_{LMO}值的平均值和标准偏差作为评价模型的稳健性[11]. 在 LMO 交叉验证中,M 的取值目前仍然没有固定的说法. 对于大数据集,M 可以取较大的数值,只要剩余的样本数足够用于建立一个有意义的模型. 对于中度或较小的数据集($n < 50$),M 的取值不应过大,最好的 LMO 交叉验证是 LMO 30% ($M = n \times 30\%$,n 为数据集样本数)[3].

1.3 y 随机化验证

y 随机化验证是确保模型稳健性常用的方法[12-13],其目的是检验因变量和自变量之间的偶然相关. 在该验证中,因变量 Y 被随机排序并使用原始自变量矩阵 X 建立新的模型,该过程重复多次,例如随机化 10—25 次[6]. 可以期望,产生的 QSAR 模型通常应具有低的 R^2_{yrand}(y 随机化相关系数)和低的 LOO 交叉验证 Q^2_{yrand}值(y 随机化 Q^2). 如果 y 随机化得到的所有模型都具有高的 R^2_{yrand}和 Q^2_{yrand}值,那么意味着对于给定的数据集,用当前的建模方法不可能得到一个可接受的 QSAR 模型.

1.4 自举法

自举法(Bootstrapping)[14]的基本假设是:抽出总体样本的代表性数据集. 在一个典型的自举法验证中,从原始数据集中随机选择 K 组,且每组的样本数都为 m. 某些样本可能被多次选取,而其它的一些样本不会被选择. 对于 m 个随机选择样本建立的模型用来预测那些被排除在外样本的活性. 在一个典型的模型验证中,重复抽取 10—25 次已足够[6]. 自举法验证中获得高的平均相关系数(R^2_{bstr}和 Q^2_{bstr}),则表明模型具有高的稳健性.

2　QSAR 模型外部验证方法

模型外部验证的最好办法是利用具体代表性和足够大的检验集(也称为预测集)来验证,并且该检验集的预测值可以与观测值(实验值)相比较. 外部验证通常把整体数据集拆分为训练集(training set)和检验集(test set),用检验集验证训练集模型[15]. Tropsha[16]将整体数据集拆分为训练集、检验集和外部验证集(external validation sets),进而验证模型的预测能力. 模型外部预测能力通过不同统计量或方法进行评价,这些统计量包括 Q^2_{F1}[13](或 R^2_{pred}[17-18])、Q^2_{F2}[19]、Q^2_{F3}[20-22]、CCC[23-24]、r^2_m[18, 25-26]和 Golbraikh 和 Tropsha 方法[27]等. 不同统计量的数学表达式详细列于表 1.

表 1　QSAR 模型统计量

Table 1　Statistical parameters of QSAR model

统计量名称	数学等式	统计量名称	数学等式
估计相关系数(训练集)	$R^2 = 1 - \sum_{i=1}^{n_{TR}} (\hat{y}_i - y_i)^2 / \sum_{i=1}^{n_{TR}} (y_i - \bar{y})^2$		$R = \dfrac{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2 \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2}}$
标准偏差(训练集)	$RSD = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n_{TR}} (y_i - \bar{y})^2}$		$k = \sum_{i=1}^{n_{EXT}} y_i \hat{y}_i / \sum_{i=1}^{n_{EXT}} \hat{y}_i^2$, $y_i^{r0} = k\hat{y}_i$
Fisher 统计量(训练集)	$F = \dfrac{\frac{1}{m} \sum_{i=1}^{n_{TR}} (\hat{y}_i - \bar{y})^2}{\frac{1}{n-m-1} \sum_{i=1}^{n_{TR}} (y_i - \hat{y}_i)^2}$	Golbraikh 和 Tropsha 方法[27](检验集)	$k' = \sum_{i=1}^{n_{EXT}} y_i \hat{y}_i / \sum_{i=1}^{n_{EXT}} y_i^2$, $\hat{y}_i^{r0} = k' y_i$
LOO 交叉验证相关系数(训练集)	$Q^2_{LOO} = 1 - \dfrac{\sum_{i=1}^{n_{TR}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y})^2}$		$R_0^2 = 1 - \sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i^{r_0})^2 / \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2$
			$R'^2_0 = 1 - \sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i^{r_0})^2 / \sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2$
均方根误差[28](训练集和检验集)	$RMSE = \sqrt{\dfrac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$	Q^2_{F1}[13](检验集)	$Q^2_{F1} = 1 - \dfrac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2}$
Pearson 相关系数[6](训练集)	$r = \dfrac{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2} \sqrt{\sum_{i=1}^{n_{TR}} (\hat{y}_i - \bar{\hat{y}})^2}}$	Q^2_{F2}[19](检验集)	$Q^2_{F2} = 1 - \dfrac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2}$
Pearson 预测相关系数[6](检验集)	$r_{ext} = \dfrac{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} \sqrt{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2}}$	Q^2_{F3}[20-21](检验集)	$Q^2_{F3} = 1 - \dfrac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2 / n_{EXT}}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2 / n_{TR}}$
一致性相关系数 CCC[23-24](检验集)	$CCC = 2 \sum_{i=1}^{n_{EXT}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) / YYY$	r^2_m[18,25](检验集)	$r^2_m = r^2(1 - \sqrt{r^2 - r_0^2})$

注:n 是样本数;m 是模型参数个数;$\hat{y}_i$ 是计算值;y_i 是观测值;$\bar{y}_{TR}$ 是训练集观测值平均值;$\bar{y}_{EXT}$ 是检验集观测值平均值;$\bar{\hat{y}}$ 是计算值的平均值;k 和 k' 为斜率;n_{EXT} 和 n_{TR} 分别是检验集和训练集的样本数;统计量 CCC(concordance correlation coefficient)中,$YYY = \sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT} (\bar{y} - \bar{\hat{y}})^2$;如果 RMSE 用于训练集,则表示为校正均方根误差(Root Mean Square Error of Calibration, RMSEC),如果用于训练集的交叉验证(如 LOO 或 LMO 交叉验证),则表示为交叉验证均方根误差(Root Mean Square Error of Cross-validation, RMSECV),如果用于检验集(预测集),则表示为预测均方根误差(Root Mean Square Error of Prediction, RMSEP)[6].

此外,Golbraikh 和 Tropsha[27]提出 4 个条件(简称 Golbraikh 和 Tropsha 方法)评价检验集预测值与观测值之差. 对于检验集,他们推荐使用下列统计特征:预测与观测活性之间的相关系数 R 应接近于 1;相关系数 R_0^2 和 R'^2_0(预测对观测活性的 R_0^2 和观测对预测的 R'^2_0)至少一个(最好两个)接近于 R^2;通过

原点的回归线斜率 k 和 k'应该接近于 1. 该方法的相关统计量表达式详见表 1.

统计量 Q^2_{F1}的数学等式类似于 Q^2_{LOO},Q^2_{F1}等式中的分母是检验集观测值(y_i)与训练集观测值的平均值($\bar{y}_{TR}$)之差的平方和. Q^2_{F1}可以直接评价模型预测值与实验值是否一致. Schuurmann 等[19]以检验集的平均值($\bar{y}_{EXT}$)替代 Q^2_{F1}分母的 $\bar{y}_{TR}$,提出统计量 Q^2_{F2}. 在 Golbraikh 和 Tropsha 方法[27]的基础上,Roy 等[18, 25, 29-30]提出统计量 r^2_m,该方法利用 Golbraikh 和 Tropsha 方法中的统计量 R 和 R^2_0(或 R'^2_0)进行计算,获得检验集的两个统计量 r^2_m 和 r'^2_m 以及两者的差值(Δr^2_m)和平均值 $\overline{r^2_m}$. Consonni 等[20-21]提出另一个统计量 Q^2_{F3}. Q^2_{F3}的分母是检验集的观测值和平均值之差的平方和,并且分子和分母分别除以训练集样本数(n_{EXT})和检验集样本数(n_{TR}).

3 统计量参考数值

利用表 1 中的统计量评价 QSAR 模型的内部预测能力和外部预测能力,当统计量的数值满足一定条件时,则认为模型可接受. 根据文献中的经验值,统计量的参考数值列于如下:

(1) 模型样本数和变量数的比值建议大于等于 5∶1[31].

(2) $R^2 > 0.6$[13, 27];Q^2大于 0.5 认为模型好,大于 0.9 则模型优秀[4]. Tropsha 等[16]建议 R^2和 Q^2均大于 0.6.

(3) $R^2 > Q^2$;校正均方根误差(RMSEC) < 交叉验证均方根误差(RMSECV);

$R^2 - Q^2 < 0.3$,如果差值大于 0.3[6],则模型过拟合和有不相关的自变量或数据有离群值[4].

(4) 在 y 随机化中,$R^2_{yrand} > Q^2_{yrand}$[4];原始 Y 与随机化后 Y 的 Pearson 相关系数的绝对值$|r|$与 R^2_{yrand}的回归线的截距(a_R)小于 0.3,$|r|$与 Q^2_{yrand} 的回归线的截距(a_Q)小于 0.05[4].

(5) Golbraikh 和 Tropsha 方法[13, 27]:$Q^2 > 0.5$; $R^2 > 0.6$; $(R^2 - R^2_0)/R^2 < 0.1$ 或 $(R^2 - R'^2_0)/R^2 < 0.1$;$0.85 \leqslant k \leqslant 1.15$ 或 $0.85 \leqslant k' \leqslant 1.15$.

(6) Roy 的 r^2_m 统计参数:$\Delta r^2_m < 0.2$ 和 $\overline{r^2_m} > 0.5$[29].

(7) Chirico 和 Gramatica[34]建议外部统计量的阈值为:Q^2_{F1}、Q^2_{F2}和 Q^2_{F3}大于 0.6,CCC > 0.85.

需要指出的是,以上的参考数值来自不同的文献,对于相同的统计量,不同作者可能使用不同的数值判定模型是否可接受.

4 评价 QSAR 模型验证方法

一个可接受的 QSAR/QSPR 模型,其必备条件之一是具有高的估计相关系数(R^2)和低的标准偏差. 然而,高的 R^2和低的标准偏差对模型的验证是不够的,因为回归模型可能包含很多参数[5]. 相关系数可能并不能反映变量间的真实关系. 相关系数与样本数和自变量数有关. 大量样本,其相关系数较小,但可能很显著. 小量样本(例如小于 10),其相关系数较高,但可能不显著. 相同的样本数,自变量数增加,模型 R^2值增加(最大等于 1). 因此,必须验证 QSAR 模型的稳定性和预测能力.

对于一个 QSAR 模型,数据集(包括样本数、自变量和因变量等)应该满足一定条件,才能保证模型具有显著的统计意义和可预测能力. 首先,所有化合物的活性值(因变量)分布不能集中一点或两点,活性值应该均匀分布且具有变化较大的特点. 其次,应该避免使用少量样本建模,少量样本不能满足数据变化较大的特征,可能导致模型存在偶然相关和较低数值的统计量. 再次,线性回归模型不应包含太多的描述符(自变量),从而使得模型解释更加复杂. 对于多元线性回归模型,一般认为样本数和描述符数的比值至少大于 5 倍(Topliss 比例)[31]. 最后,对于线性回归模型,描述符之间应没有明显的相关性.

LOO 交叉验证是模型内部验证最常用的方法,LMO 和自举法技术也被用于 QSAR 模型内部验证. 为了验证模型的稳定性,除了 LOO 或 LMO(LNO)交叉验证与自举法验证,建议使用 y 随机化方法检验模型稳定性,通过统计量是否满足参考数值($|r|$与 R^2_{yrand}的回归线的截距小于 0.3,$|r|$与 Q^2_{yrand}的回归线的截距小于 0.05)判定模型是否存在偶然相关.

研究表明相关系数 R^2与留一法(LOO)交叉验证相关系数(Q^2_{LOO})并没有相关性[6, 27]. 同样地,内部预测能力(Q^2_{LOO})和外部预测能力(R^2_{pred})之间也没有相关性. Q^2_{LOO}不能用于评价模型的外部预测能力[28]. QSAR 模型具有高的内部预测能力,但外部预测能力可能很低,反之亦然[7]. 因此,QSAR 模型必

须通过有效的外部验证,才能保证模型对外部样本的预测能力. 部分国际期刊,如 SAR QSAR Environ Res、QSAR Comb Sci、J Med Chem 和 J Chem Inf Model,明确表示每一篇 QSAR/QSPR 论文必须经过外部验证.

利用训练集以外的数据进行外部验证被认为是唯一的方法保证 QSAR/QSPR 模型的预测能力[3, 32-33]. 在模型外部验证方法中,比较简单的方法是以一个统计量评价模型的预测能力,如使用 Q^2_{F1}、Q^2_{F2}、Q^2_{F3}或 CCC. Roy 等提出统计量 r^2_m评价模型的外部预测能力,该参数需要计算 Golbraikh 和 Tropsha 方法[5]中的统计量 R^2、R^2_0和 R'^2_0. 研究表明 r^2_m能够更加严格地验证模型的外部预测能力[30]. 对于不同的数据集,Q^2_{F1}、Q^2_{F2}、Q^2_{F3}和 r^2_m的数值有差异. 当训练集的平均值与检验集的平均值有明显差别时,Q^2_{F1}有可能大于训练集的相关系数 R^2,从而导致过高估计模型的外部预测能力. 从统计量 Q^2_{F1}的数学表达式可知,该参数需要训练集观测值的平均值($\bar{y}_{TR}$). 因此,当训练集的平均值未知时(例如,一些软件中的隐含式模型),不建议使用该统计量. Golbraikh 和 Tropsha 方法需要计算预测对观测活性的 R^2_0和观测对预测的 R'^2_0,以及对于的回归斜率 k 和 k',与单一的统计量 Q^2_{F1}、Q^2_{F2}、Q^2_{F3}和 r^2_m相比,该方法计算过程较为复杂.

Q^2_{F2}统计量的缺点是没有考虑检验集与训练集之间的距离,其优点是当训练集平均值未知时(例如通过商业软件直接预测外部数据集),可通过 Q^2_{F2}评估模型的外部预测能力. Roy 等[18]指出 r^2_m能够更加严格地评价模型外部预测能力. 然而,该方法没有考虑检验集预测值和观测值回归线的斜率,可能导致过高地估计模型的外部预测能力[34]. 虽然统计量 Q^2_{F2}总是小于等于 Q^2_{F1}[19],并不能说明 Q^2_{F2}是评价模型外部预测能力的准确方法. Consonni 等[20]证明了 Q^2_{F2}依赖于数据分布,当检验集的活性值相互接近时,Q^2_{F2}不能真实反映模型的预测能力. 而 Q^2_{F3}不依赖于检验集的数据分布和样本数[21]. Q^2_{F2}统计量随着检验集样本数的增加而增加,当检验集样本数增加至一定程度时,Q^2_{F1}、Q^2_{F2}和 Q^2_{F3}的数值接近相同. 然而,与 Q^2_{F1}、Q^2_{F2}和 r^2_m类似,Q^2_{F3}仍然没有考虑检验集预测值和观测值回归线的斜率. Gramatica 等[34]认为统计量 CCC 比 Q^2_{F1}、Q^2_{F2}、Q^2_{F3}、r^2_m和 Golbraikh 和 Tropsha 方法更加严格,因此建议使用 CCC(> 0.85)作为外部验证参数评价 QSAR 模型. 然而,作者提高 CCC 阈值至 0.85 作为模型可接受的标准,而其它统计量的阈值设为 0.6. 因此,统计参数 CCC 过于乐观地估计了模型的外部预测能力.

除了统计量 Q^2_{F1}、Q^2_{F2}、Q^2_{F3}和 r^2_m,RMSEP[28]作为一个辅助验证标准应用于简单地检验模型外部预测能力. 但是 RMSEP 不能有效评估和比较不同模型,由于它取决于活性值(因变量)范围[34]. 评估模型的预测能力,一般也需要计算 RMSE(RMSEC、RMSECV 和 RMSEP)数值. 这些统计量仅仅用于判定模型的内部或外部预测能力,该统计量具有高的数值,并不代表已建立的模型是可靠的或可接受的. 一个可靠的 QSAR/QSPR 模型与多种因素具有紧密的联系,例如模型使用的描述符[35-37]、使用不恰当的终点数据、数据过拟合、数据集中有重复化合物[32]、训练集和检验集选择方法[38-41]、变量选择方法[42-45]、分子结构是否正确[46-47]、模型包含离群值[32, 48]等因素.

因此,对于模型的内部验证,我们建议至少使用 LOO 交叉验证($Q^2_{LOO} > 0.5$)和 y 随机化验证(截距 $a_R < 0.3$ 和 $a_Q < 0.05$)评价模型的内部预测能力和稳健性. 此外,QSAR 模型应该给出统计量 RMSEC 和 RMSECV. 对于模型的外部检验,我们建议使用较为严格的统计参数 Δr^2_m (< 0.2)和 $\overline{r^2_m}$ (> 0.5)[29]评价模型的外部预测能力. 为了获得完整和可靠地 QSAR/QAPR 模型,所有模型应该符合 OECD 建议的 5 个法则[1].

总之,一个没有通过严格内部验证和外部验证的模型,或仅通过 LOO 或 LMO 进行内部验证,不能保证模型对外部样本的真正预测能力. 因此,在 QSAR 模型的实际应用或解释之前,建议严格地进行内部和外部验证.

参 考 文 献

[1] Organisation for Economic Co-operation and Development (OECD), Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models[EB/OL]. [2013-03-20]. http://search. oecd. org/officialdocuments/displaydocumentpdf/? cote = env/jm/mono(2007)2&doclanguage = en

[2] Rucki M, Tichy M. Validation of QSAR models for legislative purposes [J]. Interdiscip Toxicol, 2009, 2(3): 184-186

[3] Gramatica P. Principles of QSAR models validation: internal and external [J]. QSAR Comb Sci, 2007, 26(5): 694-701

[4] Eriksson L, Jaworska J, Worth A P, et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs [J]. Environ Health Perspect, 2003, 111(10): 1361-1375
[5] Wold S. Validation of QSAR's [J]. Quant Struct-Act Rel, 1991, 10(3): 191-193
[6] Kiralj R, Ferreira M M C. Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application [J]. J Braz Chem Soc, 2009, 20(4): 770-787
[7] Geisser S. The predictive sample reuse method with applications [J]. J Am Stat Assoc, 1975, 70: 320-328
[8] Konovalov D A, Llewellyn L E, Heyden Y V, et al. Robust cross-validation of linear regression QSAR models [J]. J Chem Inf Model, 2008, 48(10): 2081-2094
[9] Clark R D. Boosted leave-many-out cross-validation: The effect of training and test set diversity on PLS statistics [J]. J Comput Aid Mol Des, 2003, 17(2): 265-275
[10] Besalu E. Fast computation of cross-validated properties in full linear leave-many-out procedures [J]. J Math Chem, 2001, 29(3): 191-204
[11] Qin L T, Liu S S, Chen F, et al. Chemometric model for predicting retention indices of constituents of essential oils [J]. Chemosphere, 2013, 90(2): 300-305
[12] Rucker C, Rucker G, Meringer M. *y*-Randomization and its variants in QSPR/QSAR [J]. J Chem Inf Model, 2007, 47(6): 2345-57
[13] Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models [J]. QSAR Comb Sci, 2003, 22(1): 69-77
[14] Wehrens R, Putter H, Buydens L M C. The bootstrap: A tutorial [J]. Chemometr Intell Lab Syst, 2000, 54(1): 35-52
[15] Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection [J]. J Comput Aid Mol Des, 2002, 16(5/6): 357-369
[16] Tropsha A. Best practices for QSAR model development, validation, and exploitation [J]. Mol Inf, 2010, 29(6/7): 476-488
[17] Roy K. On some aspects of validation of predictive quantitative structure-activity relationship models [J]. Expert Opin Drug Discovery, 2007, 2(12): 1567-1577
[18] Roy P P, Paul S, Mitra I, et al. On two novel parameters for validation of predictive QSAR models [J]. Molecules, 2009, 14(5): 1660-1701
[19] Schuurmann G, Ebert R U, Chen J W, et al. External validation and prediction employing the predictive squared correlation coefficient-test set activity mean *vs* training set activity mean [J]. J Chem Inf Model, 2008, 48(11): 2140-2145
[20] Consonni V, Ballabio D, Todeschini R. Comments on the definition of the Q^2 parameter for QSAR validation [J]. J Chem Inf Model, 2009, 49(7): 1669-1678
[21] Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques [J]. J Chemometr, 2010, 24(3/4): 194-201
[22] Chirico N and Gramatica P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection [J]. J Chem Inf Model, 2012, 52(8): 2044-2058
[23] Lin L I. Assay validation using the concordance correlation coefficient [J]. Biometrics, 1992: 599-604
[24] Lin L I. A concordance correlation coefficient to evaluate reproducibility [J]. Biometrics, 1989, 45(1): 255-268
[25] Mitra I, Roy P P, Kar S, et al. On further application of r^2_m as a metric for validation of QSAR models [J]. J Chemometr, 2010, 24(1): 22-33
[26] Roy P P, Roy K. On some aspects of variable selection for partial least squares regression models [J]. QSAR Comb Sci, 2008, 27(3): 302-313
[27] Golbraikh A, Tropsha A. Beware of q^2! [J]. J Mol Graph Model, 2002, 20(4): 269-276
[28] Aptula A O, Jeliazkova N G, Schultz T W, et al. The better predictive model: High q^2 for the training set or low root mean square error of prediction for the test set? [J]. QSAR Comb Sci, 2005, 24(3): 385-396
[29] Ojha P K, Mitra I, Das R N, et al. Further exploring r^2_m metrics for validation of QSPR models [J]. Chemometr Intell Lab Syst, 2011, 107(1): 194-205
[30] Roy K, Mitra I, Kar S, et al. Comparative studies on some metrics for external validation of QSPR models [J]. J Chem Inf Model, 2012, 52(2): 396-408
[31] Topliss J G, Edwards R P. Chance factors in studies of quantitative structure-activity relationships [J]. J Med Chem, 1979, 22(10): 1238-1244
[32] Dearden J C, Cronin M T, Kaiser K L. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR) [J]. SAR QSAR Environ Res, 2009, 20(3/4): 241-266
[33] Benigni R, Bossa C. Predictivity of QSAR [J]. J Chem Inf Model, 2008, 48(5): 971-980
[34] Chirico N, Gramatica P. Real external predictivity of QSAR models: How to evaluate it? comparison of different validation criteria and proposal of using the concordance correlation coefficient [J]. J Chem Inf Model, 2011, 51(9): 2320-2335
[35] Kiralj R, Ferreira M M C. Is your QSAR/QSPR descriptor real or trash? [J]. J Chemometr, 2010, 24(11/12): 681-693
[36] Hechinger M, Leonhard K, Marquardt W. What is wrong with quantitative structure-property relations models based on three-dimensional descriptors? [J]. J Chem Inf Model, 2012, 52(8): 1984-1993
[37] Paster I, Shacham M, Brauner N. Investigation of the relationships between molecular structure, molecular descriptors, and physical properties

[J]. Ind Eng Chem Res,2009,48(21):9723-9734

[38] Puzyn T, Mostrag-Szlichtyng A, Gajewicz A, et al. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models [J]. Struct Chem,2011,22(4):795-804

[39] Roy P P, Leonard J T, Roy K. Exploring the impact of size of training sets for the development of predictive QSAR models [J]. Chemometr Intell Lab Syst,2008,90(1): 31-42

[40] Rajer-Kanduc K, Zupan J, Majcen N. Separation of data on the training and test set for modelling: A case study for modelling of five colour properties of a white pigment [J]. Chemometr Intell Lab Syst,2003,65(2): 221-229

[41] Orfi L, Szantai-Kis C, Kovesdi I, et al. Validation subset selections for extrapolation oriented QSPAR models [J]. Mol Divers,2003,7(1): 37-43

[42] Goodarzi M, Heyden Y V, Funar-Timofei S. Towards better understanding of feature-selection or reduction techniques for quantitative structure-activity relationship models [J]. TrAC, Trends Anal Chem,2013,42:49-63

[43] Goodarzi M, Dejaegher B, Vander Heyden Y. Feature selection methods in QSAR studies [J]. J Aoac Int,2012,95(3):636-651

[44] Eklund M, Norinder U, Boyer S, et al. Benchmarking variable selection in QSAR [J]. Mol Inf,2012,31(2):173-179

[45] Andersen C M, Bro R. Variable selection in regression-a tutorial [J]. J Chemometr,2010,24(11/12):728-737

[46] Young D, Martin T, Venkatapathy R, et al. Are the chemical structures in your QSAR correct? [J]. QSAR Comb Sci,2008,27(11/12): 1337-1345

[47] Li J Z, Gramatica P. The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders [J]. Mol Divers,2010,14(4):687-696

[48] Furusjo E, Svenson A, Rahmberg M, et al. The importance of outlier detection and training set selection for reliable environmental QSAR predictions [J]. Chemosphere,2006,63(1):99-108

Internal and external validtions of QSAR model: Review

QIN Litang[1,2]　　*LIU Shushen*[1,3*]　　*XIAO Qianfen*[1,3]　　*WU Qingsheng*[1,2]

(1. Key Laboratory of Yangtze River Water Environment, Ministry of Education, Tongji University, Shanghai, 200092, China;
2. Department of Chemistry, Tongji University, Shanghai, 200092, China;
3. College of Environmental Science and Engineering, Tongji University, Shanghai, 200092, China)

ABSTRACT

Validation of a quantitative structure-activity relationship (QSAR) model plays an important role for ensuring the reliable predictive ability of activity of untested chemicals. Currently, a number of QSAR models, however, lack reliable validation. The present study reviews the internal validation and external validation methods that exist for QSAR model. The internal validations include leave-one-out (LOO) cross-validation, leave-many-out (LMO) or leave-*N*-out (LNO) cross-validation, *y*-randomization test and bootstrapping, while the external validations include the statistical parameters Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, concordance correlation coefficient (CCC), $\bar{r}^2_m$, and Golbraikh-Tropsha method. Furthermore, the cutoff values of the different statistical parameters for an acceptable model were recommended according to the references. The internal and external validations addressed in this study together with the recommended cutoff values of statistical parameters may help researchers to develop QSAR models.

Keywords: QSAR, internal validation, external validation.