表皮生长因子受体抑制剂的理论预测研究

李秉轲[a,b],付明杰[a],邹钰嵘[a],康晓康[a],王 林[a],马艺飞[a]

(成都师范学院 a.化学与生命科学学院; b.功能分子研究所,成都 611130) *

摘 要:表皮生长因子受体(Epidermal Growth Factor Receptor, EGFR)的扩增或突变与人类实体肿瘤密切相关。三种机器学习方法:随机森林(random forest, RF),k 最近邻和 C4.5 决策树,被用于建立理论模型,来预测 EGFR 的抑制剂。结果表明,这些模型的预测结果均取得了令人满意的精度。又通过对比分析,发现 RF 模型具有较好的预测性能,并进一步修改和优化了 RF 模型的参数。此外,还采用特征选择程序筛选出了 25 个和 EGFR 的抑制最相关的分子描述符。最后对得到的最优 RF 模型进行了评价。

关键词:EGFR 抑制剂;机器学习方法;随机森林;分子描述符

doi:10.3969/**j.issn.**2095-5642.2018.09.104

中图分类号:R979.1 **文献标志码**:A **文章编号**:2095-5642(2018)09-0104-08

蛋白激酶,特别是受体酪氨酸激酶,在信号转导通路中发挥着重要作用,调控着细胞的多种功能,包括增殖,分化,迁移和血管生成。表皮生长因子受体(Epidermal Growth Factor Receptor, EGFR)是一种跨膜糖蛋白。EGFR 由其同源配体通过与 EGFR 家族的其他成员形成同二聚体或异二聚体而被激活[1]。当 EGFR 被激活时,几个信号转导级联被启动,然后导致 DNA 合成和细胞增殖。特别是 EGFR 扩增或突变时,DNA 合成和细胞增殖将发生异常并导致癌症。目前,EGFR 的扩增或突变已在人类实体肿瘤如神经胶质瘤,肺癌,卵巢癌和乳腺癌中发现。因此,EGFR 也被认为是这些疾病的潜在治疗靶点[2]。

许多 EGFR 抑制剂已被开发,并被 FDA 批准上市[3]。如拉帕替尼,被用于治疗乳腺癌。而替莫唑胺,洛莫司汀,厄洛替尼和吉非替尼,则被用于治疗神经胶质瘤。然而,由于选择性、毒性和副作用,现有的 EGFR 抑制剂超出了人们的预期。因此,有必要设计和合成新的高效 EGFR 抑制剂。

而综合现今的文献报导[4-7],也有许多关于 EGFR 抑制剂虚拟计算方面的研究。Zhao 等[4]介绍了应用二维和三维定量构效关系(quantitative structure activity relationship, QSAR)方法区分 EGFR 抑制剂的方法。Shinde 等[7]经比较分子力场分析法,对酞菁类抗癌药物恶嗪衍生物进行 3D-QSAR 研究,确定了该类 EGFR 抑制剂的药效团结构,并与分子相似性指数方法进行比较,显示了较好的预测效果。

1 计算方法

1.1 机器学习方法

本文采用的机器学习方法有 k 最近邻(k-nearest neighbor, k-NN)[8],C4.5 决策树(decision tree, DT)[9]和随机森林(random forest, RF)[10]。

* **收稿日期**:2018-05-26

作者简介:李秉轲(1985—),男,河南南阳人,讲师,博士,研究方向:计算机辅助药物分子设计;
付明杰(1997—),男,四川新都人,研究方向:计算机辅助药物分子设计;
邹钰嵘(1998—),男,江西青云谱人,研究方向:计算机辅助药物分子设计;
康晓康(1997—),男,四川南溪人,研究方向:计算机辅助药物分子设计;
王 林(1994—),男,四川射洪人,研究方向:计算机辅助药物分子设计;
马艺飞(1994—),男,四川达州人,研究方向:计算机辅助药物分子设计。

1.1.1 k－NN 方法

k－NN 将训练的样本作为平面坐标上的点进行处理,所有的样本对应于不同维度空间的点,再将训练集的样本进行分类处理,然后对一个给定的待分类的预测集,通过计算每个元素与划分等级之间的距离,来判断每个元素所处的位置,进而对样本进行分类。k 值选择、距离度量和分类决策规则是 k－NN 算法的三个基本要素:(1)k 值的选择对算法的结果有很大的影响。小 k 值意味着只有训练实例接近输入实例是有效的预测,而且非常容易发生过拟合。如果 k 值较大,该方法的优点是可以减少学习估计误差,但缺点是学习误差随近似误差的增大而增大,且训练样本远离输入实例也会影响预测,导致预测误差。(2)算法的分类规则主要是多数选票,即输入实例的 k 最近决定输入实例的类别。(3)距离测量一般使用欧几里德距离。在测量之前,应该对每个属性的值进行标准化。

1.1.2 C4.5 DT 方法

C4.5DT 方法是应用概率分析的直观图形方法。在已知概率和决策分析方法的基础上,构造 DT 来评价项目风险,确定其可行性。这是一个有监督的学习,所谓的监督学习有一堆样本,每个样本有一组属性和一个类别,这些类别是预先确定的,然后由分类器学习,给出正确分类的对象分类器。以二分任务为例,DT 算法希望从给定的训练数据集中学习一个模型,并对新实例进行分类,可以认为是“当前样本属于正样本”的“决策”过程。DT 是一个树结构,其中每个内部节点代表一个属性上的一个测试,每个分支代表一个测试输出,每个叶节点代表一个类别。

1.1.3 RF 方法

RF 方法由 Breiman[10] 提出,是一个由 DT 分类器集合 $\{h(x,\theta_k),k=1,2,\Lambda\}$ 构成的组合分类器模型,其中参数集 $\{\theta_k\}$ 是独立同分布的随机向量,x 是输入向量。当给定输入向量时每个 DT 有一票投票权来选择最优分类结果。每一个 DT 是由分类回归树算法构建的未剪枝的 DT。每棵树都依赖于一个提取的样本,森林中的每棵树都有相似的分布。分类误差则取决于树的分类能力及其相关性。特征选择使用一个随机的方法来分割每个节点,然后比较在不同情况下所犯的错误。可以检测到的内在估计误差、分类的能力和相关性决定了所选特征的数量。单个树的分类能力可能相对较小,但在随机生成大量 DT 后,根据每棵树的分类结果,可以统计出一个测试样本来选择最可能的分类。

1.2 分子描述符

分子描述符可以反映化合物的物理和化学性质,即分子在某一方面性质的度量[11]。本文采用实验室自己发展的一套分子描述符计算程序基于数据集中每个化合物的 3D 结构来实现对每个分子的描述符的计算。这些描述符一共有 189 个,大致可以分为五大类:18 个简单分子性质描述符、27 个分子连接性和形状描述符、97 个电拓扑态描述符、22 个量子化学性质描述符和 25 个几何性质描述符。

2 结果与讨论

2.1 几种机器学习方法预测精度的比较

TP(True positive)表示预测正确的阳性样本的数量,TN(True negative)表示预测正确的阴性样本的数量,FP(False positive)表示错误预测为阳性的阴性样本的数量,FN(False negatives)表示错误预测为阴性的阳性样本的数量。对于本文而言,TP 是预测正确的 EGFR 抑制剂的数量,FN 则是预测错误的 EGFR 抑制剂的数量,TN 是预测正确的 EGFR 非抑制剂的数量,FP 则是预测错误的 EGFR 非抑制剂的数量。同时,在衡量预测性能方面还有以下几个精度函数,包括灵敏度(Sensitivity, SE),特异性(Specificity, SP),总预测精度(Q)和马氏相关系数(Markov correlation coefficient, MCC)。这些函数与前面的变量之间存在有以下关系:

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

$$SP=\frac{TN}{TN+FP} \tag{2}$$

$$Q=\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

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

本文用 RF、k－NN 和 C4.5 DT 三种方法对同一训练集建立了理论预测模型，并用同一测试集对模型的性能进行评估，把数据总结在了表格中，如表 1 所示。

表 1　RF，k－NN 和 C4.5 的 EGFR 抑制剂和非抑制剂预测准确率的比较

		测试集							
		EGFR 抑制剂			EGFR 非抑制剂				
方法	参数	TP	FN	SE(%)	TN	FP	SP(%)	Q(%)	MCC
RF	$M_{try}=24$	293	11	96.38	285	3	98.95	97.64	95.31%
k－NN	k=30	294	10	96.71	266	22	92.36	94.59	89.24%
C4.5 DT	/	292	12	94.19	270	18	95.74	94.93	89.87%

从表 1 中可以看出，RF 模型的总预测精度最高，马氏相关系数值最大，预测正确率最高，与 k－NN 和 C4.5 DT 两种方法建立的模型相比，具有较大的优势。其中，M_{try}是 RF 方法的参数，这里取的是优化之后的数值 24。k 是 k－NN 方法的参数，是通过内在的参数选择程序优选出来的。

2.2 描述符的筛选

表 2　RF 最优模型中特征选择得到的 25 个对 EGFR 抑制剂的预测最相关的描述符

描述符序号	描述符名称	描述符代表的意义
84	S(35)	Atom－type Estate sum for :N:
51	S(2)	Atom－type H Estate sum for =NH
83	S(34)	Atom－type Estate sum for =N－
76	S(27)	Atom－type Estate sum for :C::
28	$^{6}\chi_{CH}$	Simple molecular connectivity Chi indices for cycles of 6 atom
157	$Q_{N,\ SS}$	Sum of squares of charges on N atoms
149	$Q_{H,\ Min}$	Most negative charge on H atoms
145	$Q_{H,\ Max}$	Most positive charge on H atoms
89	S(40)	Atom－type Estate sum for =O
146	$Q_{C,\ Max}$	Most positive charge on C atoms
38	$^{6}\chi^{v}_{CH}$	valence molecular connectivity Chi indices for cycles of 6 atoms
75	S(26)	Atom－type Estate sum for :C:－
69	S(20)	Atom－type Estate sum for =CH－
61	S(12)	Atom－type H Estate sum for CH_n(Saturated)
153	$A_{Q,\ max}$	Most positive charge in a molecule
59	S(10)	Atom－type H Estate sum for :CH:(sp^2, aromatic)
70	S(21)	Atom－type Estate sum for :CH:(aromatic)
14	nnitro	Count of N atoms
58	S(9)	Atom－type H Estate sum for =CH－(sp^2)
71	S(22)	Atom－type Estate sum for >CH－
162	Mac	Mean absolute charge
147	$Q_{N,\ Max}$	Most positive charge on N atoms
74	S(25)	Atom－type Estate sum for =C<
148	$Q_{O,\ Max}$	Most positive charge on O atoms
163	Rpc	Relative positive charge

按照 RF 方法特有的程序,本文还对最优参数对应的模型进行了描述符的特征选择处理,即从 189 个描述符中,筛选得到了与 EGFR 抑制剂分子的性质最相关的 25 个描述符,全部列在了表 2 中。这些描述符每

个都有相应的贡献率,其相对重要性排名可见图 1。

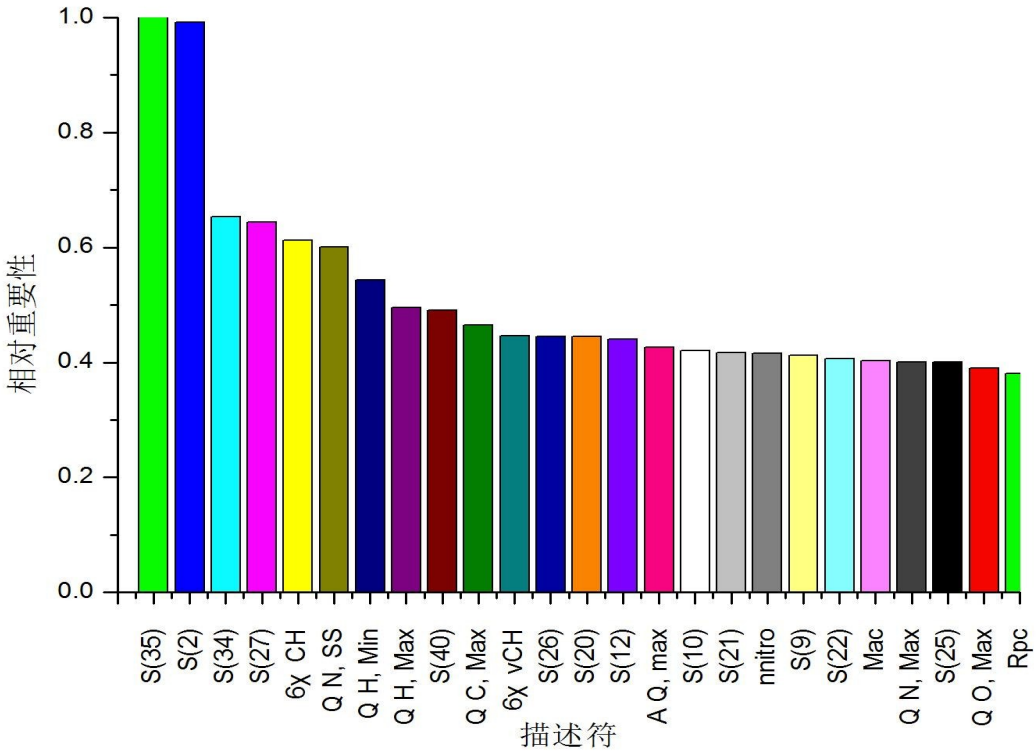

图 1 RF 最优模型中特征选择得到的 25 个对 EGFR 抑制剂的预测最相关的描述符的相对重要性排名

如图 1 所示,25 个选定的描述符的重要性依次降低,排在前三位的分别是 S(35)、S(2)和 S(34),分别代表:N:原子类型拓扑状态之和,=NH 原子类型拓扑状态之和,以及=N—原子类型拓扑状态之和。可以看出,这三个特征对于预测 EGFR 的潜在抑制剂具有非常重要的参考价值。

2.3 RF 最优模型的评价

图 2 显示了在建立的 RF 最优化模型中,测试集中 592 个分子的分布情况。从图中可以看出,已建立模型的分类边界可以很好地将 EGFR 抑制剂与非抑制剂分离。仔细观察图表中的散点分布,我们发现有 11 个抑制剂的点得分在 0.5 以下,有 3 个非抑制剂的点得分在 0.5 以上,直观地表现出这些点被模型错误的预测。而图中显示的结果也与表 1 中 RF 方法预测错误的分子个数相一致,说明模型的打分函数具有相当的可靠性。

通过绘制受试者工作特征(receiver operating characteristic, ROC)曲线,本文进一步分析和评价了该二元分类模型的判别效果。ROC 曲线将 SE 与 SP 结合,随着预测概率阈值的变化,将产生许多对 SE 和"1—SP"的组合。如果我们把 SE 作为纵坐标,"1—SP"作为横坐标,并将每个点连接起来,我们就可以画出 ROC 曲线。曲线上的点表示当预测的概率阈值发生变化时 SE 和 SP 之间的差值。

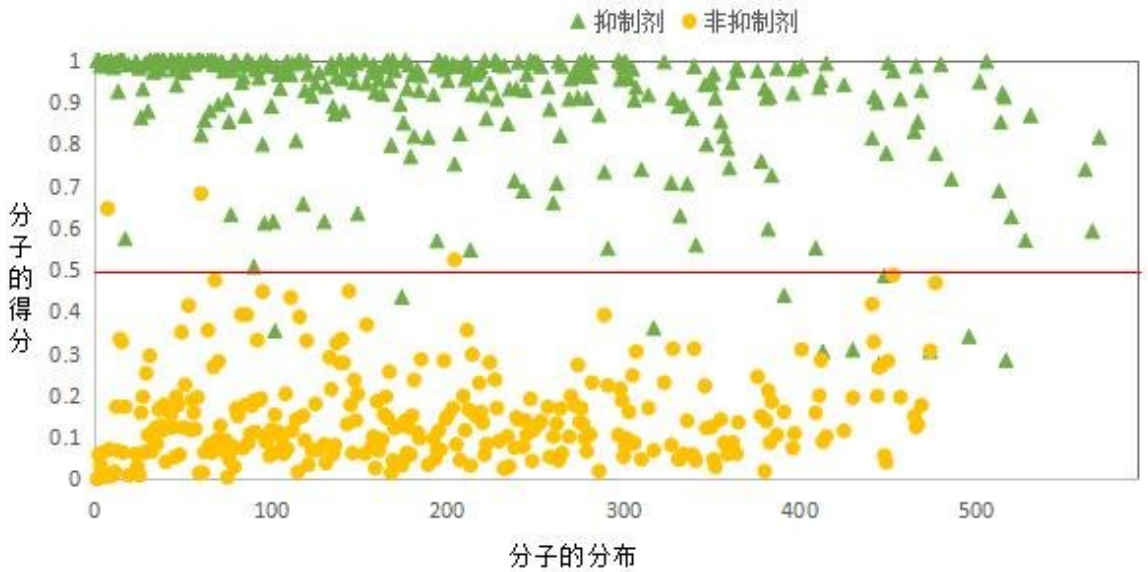

图 2　测试集的总共 592 个分子在所建立的 RF 最优模型中的可视化分布

评价分类模型的预测能力还有一个非常重要的指标:ROC 曲线下的面积(area under roc curve,AUC),其值介于 0.5 到 1 之间,越大表示模型的分类性能越好。

本文中的最优 RF 模型对训练集和测试集的 ROC 曲线分别如图 3 和图 4 所示。通过曲线拟合可知,训练集 ROC 曲线的 AUC 为 0.978,测试集 ROC 曲线的 AUC 为 0.997,均反映了该 RF 最优模型非常优越的预测性能。

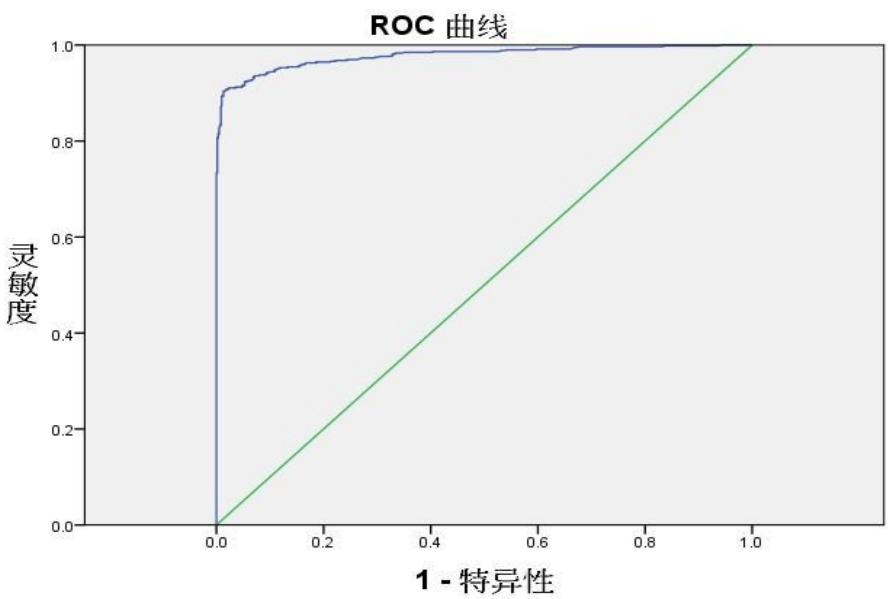

图 3　最优 RF 模型对训练集的 ROC 曲线

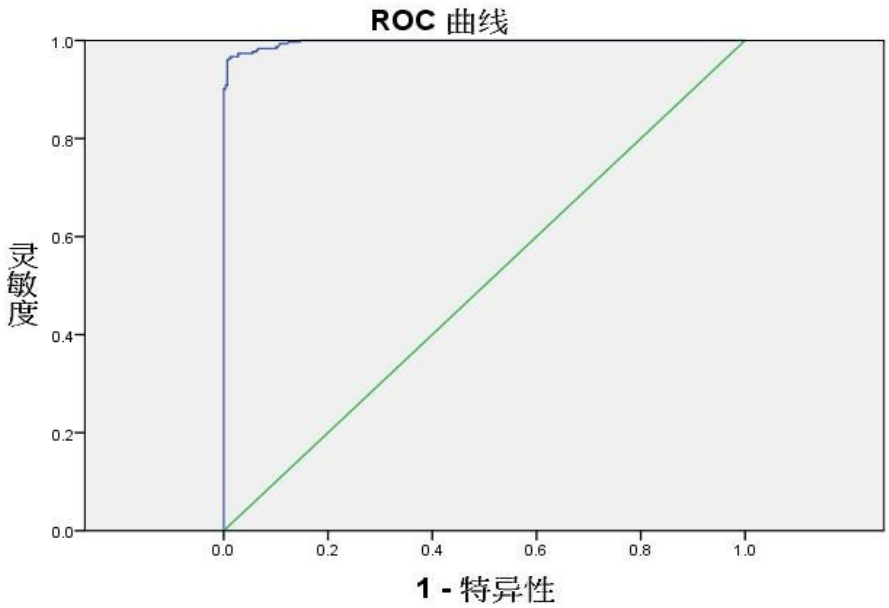

图 4 最优 RF 模型对测试集的 ROC 曲线

3 结论

本文采用三种机器学习方法(RF、k－NN 和 C4.5 DT)对 EGFR 的抑制剂与非抑制剂进行了理论预测模型的建立,取得了比较令人满意的结果。通过几种模型的对比分析,本文发现 RF 模型具有更好的预测性能,并进一步对模型参数进行调试,得到了最优的 RF 模型和相对应的具有重要作用的分子描述符。本文所建立的模型和筛选出来的分子描述符,可以为后续的分子设计和相关合成工作提供依据,并最终推动 EGFR 先导化合物的发现。

参考文献:

[1] ZHU W, CHEN H, WANG Y, et al. Design, synthesis, and pharmacological evaluation of novel multisubstituted pyridin－3－amine derivatives as multitargeted protein kinase Inhibitors for the treatment of non－small cell lung cancer[J]. Journal of Medicinal Chemistry, 2017, 60(14): 6018－6035.

[2] SONG Z, HUANG S, YU H, et al. Synthesis and biological evaluation of morpholine－substituted diphenylpyrimidine derivatives (Mor－DPPYs) as potent EGFR T790M inhibitors with improved activity toward the gefitinib－resistant non－small cell lung cancers (NSCLC)[J]. European Journal of Medicinal Chemistry, 2017, 133: 329－339.

[3] BUTTERWORTH S, CROSS D A E, FINLAY M R V, et al. The structure－guided discovery of osimertinib: the first U.S. FDA approved mutant selective Inhibitor of EGFR T790M[J]. MedChemComm, 2017, 8(5): 820－822.

[4] ZHAO M, WANG L, ZHENG L, et al. 2D－QSAR and 3D－QSAR analyses for EGFR inhibitors[J]. BioMed Research International, 2017, 2017: 1－11.

[5] AKHTAR M J, SIDDIQUI A A, KHAN A A, et al. Design, synthesis, docking and QSAR study of substituted benzimidazole linked oxadiazole as cytotoxic agents, EGFR and erbB2 receptor inhibitors[J]. European Journal of Medicinal Chemistry, 2017, 126: 853－869.

[6] EL－SAYED M A A, EL－HUSSEINY W M, ABDEL－AZIZ N I, et al. Synthesis and biological evaluation of 2－styrylquinolines as antitumour agents and EGFR kinase inhibitors: molecular docking study[J]. Journal of enzyme inhibition and medicinal chemistry, 2018, 33(1): 199－209.

[7] SHINDE M G, MODI S J, KULKARNI V M. QSAR and molecular docking of phthalazine derivatives as epidermal growth factor receptor (EGFR) inhibitors[J]. Journal of Applied Pharmaceutical Science, 2017, 7(4): 181－191.

[8] DHANABAL S, CHANDRAMATHI S. A review of various k－nearest neighbor query processing techniques[J]. International Journal of Computer Applications, 2011, 31(7): 14－22.

[9] ROKACH L, MAIMON O. Top－down induction of decision trees classifiers－a survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2005, 35(4):476－487.

[10] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5－32.

[11] LI B K, HE B, TIAN Z Y, et al. Modeling, predicting and virtual screening of selective inhibitors of MMP－3 and MMP－9 over MMP－1 using random forest classification[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 147: 30－40.

Theoretical Prediction on Inhibitors of the Epidermal Growth Factor Receptor

LI Bingke, *FU Mingjie*, *ZOU Yurong*, *KANG Xiaokang*, *WANG Lin*, *MA Yifei*

(Institute of Functional Molecules, School of Chemistry and Life Science,
Chengdu Normal University, Chengdu 611130, China)

Abstract: Amplification or mutation of Epidermal Growth Factor Receptor (EGFR) is closely related to human solid tumors. In this paper, three machine learning methods were used: random forest (RF), k－nearest neighbor and C4.5 decision tree. Some theoretical models were established to predict the inhibitors of EGFR. The results manifested that the predicting outcomes of these models have achieved satisfying accuracy. Through contrastive analysis, we found that the RF model had better predictive performance. After further modifying and optimizing the parameters of RF model, we evaluated the optimal model. Besides, we also screened 25 molecular descriptors most relevant to the inhibitors of EGFR by feature selection process, which can provide evidence for the discovery and synthesis of EGFR's lead compounds.

Key words: EGFR inhibitors; machine learning method; random forest; molecular descriptors

(实习编辑:杨晓玲　责任校对:曲　比)