Advances in Applied Mathematics 应用数学进展, 2021, 10(12), 4454-4468
Published Online December 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1012474

Hans汉斯

抗乳腺癌候选药物的生物活性预测模型的构建

徐　浦，张振东，李　晨

上海理工大学机械工程学院，上海

收稿日期：2021年11月27日；录用日期：2021年12月17日；发布日期：2021年12月29日

摘　要

乳腺癌是目前发病率较高的疾病之一，选择能够拮抗ERα活性的化合物对治疗乳腺癌具有重要的意义。本文综合考虑ERα拮抗剂的生物活性即pIC_{50}值(与生物活性具有正相关性)，筛选治疗乳腺癌的候选药物。为了构建能够使化合物对抑制ERα具有更好的生物活性的预测模型，本文通过随机森林算法和距离相关系数算法，寻求主要变量进行降维处理；通过建立基于BP神经网络的pIC_{50}预测模型并进行训练与验证，为寻找处理后变量的全局最优解，采用粒子群优化算法，以pIC_{50}的最大值作为目标函数，设定参数运行求得优化结果。研究结果表明，pIC_{50}的最大值与其对应的分子描述符都在合理的区间范围内，说明此次建立的模型具有一定的稳定性与合理性。

关键词

抗乳腺癌药物，随机森林，距离相关系数，BP神经网络，粒子群算法

Construction of Bioactivity Prediction Models for Breast Cancer Candidate Drugs

Pu Xu, Zhendong Zhang, Chen Li

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 27th, 2021; accepted: Dec. 17th, 2021; published: Dec. 29th, 2021

Abstract

Breast cancer is one of the diseases with high incidence at present. It is important to select compounds that can antagonize ERα activity in the treatment of breast cancer. In this study, the bioactivity of ERα antagonists (pIC_{50} value), which was positively correlated with biological activity, was considered to screen candidate drugs for breast cancer. In order to construct a compound that can inhibit ERα with better biological activity, this paper seeks the main variables for dimensionality reduction through random forest algorithm and distance correlation coefficient algo-

文章引用：徐浦，张振东，李晨．文章标题抗乳腺癌候选药物的生物活性预测模型的构建[J]．应用数学进展，2021, 10(12): 4454-4468. DOI: 10.12677/aam.2021.1012474

rithm; Through the establishment, training and verification of the pIC_{50} prediction model based on BP neural network, in order to find the global optimal solution of the processed variables, the particle swarm optimization algorithm is adopted, the maximum value of pIC_{50} is taken as the objective function, and the parameters are set to run to obtain the optimization results. The results show that the maximum value of pIC_{50} and its corresponding molecular descriptor are within a reasonable range, indicating that the model established this time has certain stability and rationality.

Keywords

Anti-Breast Cancer Drugs, Random Forest, Distance Correlation Coefficient, BP Neural Network, Particle Swarm Optimization

1. 引言

乳腺癌作为常见的癌症之一，拥有较高的致死率。而乳腺癌作为经典的雌激素依赖性肿瘤[1]，其细胞的增殖、迁移等生物活性与雌激素受体(Estrogen receptors，以下简称 ER)密切相关，最新研究发现，雌激素受体 α 亚型(ERα)有 50%~80%在肿瘤细胞中表达，而在正常乳腺细胞中表达的不足 10%，通过对 ERα 基因缺失小白鼠的实验得到 ERα 对乳腺发育有重要作用[2]。目前 ERα 常用作治疗乳腺癌的标靶，通过调节 ER 活性来遏止癌细胞的生物活性，所以拥有拮抗 ERα 活性的化合物可能会是治疗乳腺癌的候选药物，比如他莫昔芬、芳香化酵抑制剂等。

目前，新药物的研发多为建立化合物活性预测模型来筛选潜在的活性化合物。以本题为例子，针对与癌症相关的几个重要靶标，整合出作用于靶标的化合物及其生物活性数据，然后将每一个分子结构描述符作为自变量，化合物的生物活性值作为因变量，推定化合物的定量结构－活性关系(Quantitative Structure-Activity Relationship, QSAR)模型，然后使用建立的模型对新的化合物分子进行预测，或者对已有化合物进行结构优化。

2. 寻找建模主要变量

2.1. 数据来源及变量降维思路

本文用 2021 数学建模竞赛 B 题所提供的数据集作为原始数据集。在数据集中，给出了 1974 个化合物的 729 个分子描述符信息及对 ERα 的生物活性(pIC_{50})的数据。

由于所给的数据集变量过多，因此需要对变量进行降维处理。在 1974 个化合物的 729 个分子描述符中，找到具有代表性和独立性的 20 个主要变量，使得选择的主要变量可以对生物活性有显著影响。

本节将解决选取主要变量和确保变量独立性两个问题。首先使用随机森林算法对 1974 个化合物的 729 个分子的变量信息进行操作，获得各变量的贡献度，贡献度即变量对生物活性影响的重要程度，根据贡献度从高到低进行排序，然后对贡献度高的变量进行距离相关性计算，剔除相关性高的耦合变量，最后成功提取到 20 个显著影响生物活性的变量。最后，从变量降维过程中采用的算法及处理流程以及变量降维的最终结果两方面对所选择变量的合理性进行评价。变量降维的思路流程图如图 1 所示：

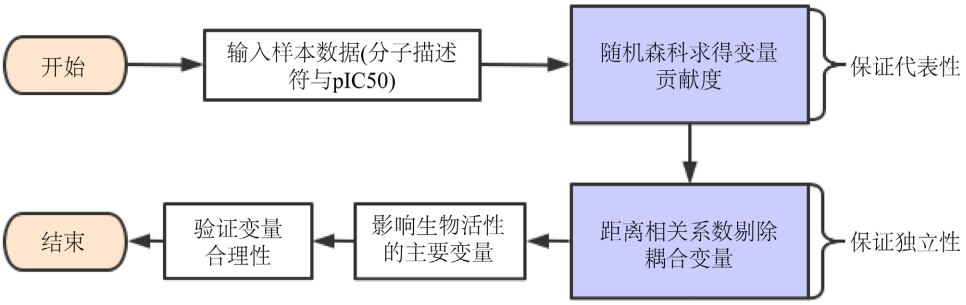

Figure 1. Flow chart of variable dimensionality reduction
图 1. 变量降维思路流程图

2.2. 随机森林算法评价变量的重要程度

随机森林[3]是由 Leo Breiman 和 Adele Cutler 发展出推论出的算法，在机器学习中，随机森林常常作为一个含有多个决策树的分类器，其输出的类别是由个别树输出的类别的众数确定的。随机森林是基于 bagging 思想延伸而来。

Bagging [4]思想：bagging 即 boostrap aggregating，其中的 boostrap 代指一种有放回的抽样方法，抽样的策略是简单随机抽样。其原理是把多个基础模型放到一起，最后求其平均值，这里可以把决策树当作基础模型，其实基本上所有集成策略都是以树模型为基础的，公式如下：

$$f(x)=\frac{1}{M}\sum_{m=1}^{M} f_m(x) \tag{1}$$

首先对数据集进行随机采样，分别训练多个树模型，最终将其结果整合在一起即可。

训练随机森林的过程就是训练各个决策树的过程，由于各个决策树的训练是相互独立的，因此随机森林的训练可以通过并行处理来实现，这将大大提高生成模型的效率。随机森林的整体训练过程以及任意一个训练样本 m 的训练过程如图 2 所示。

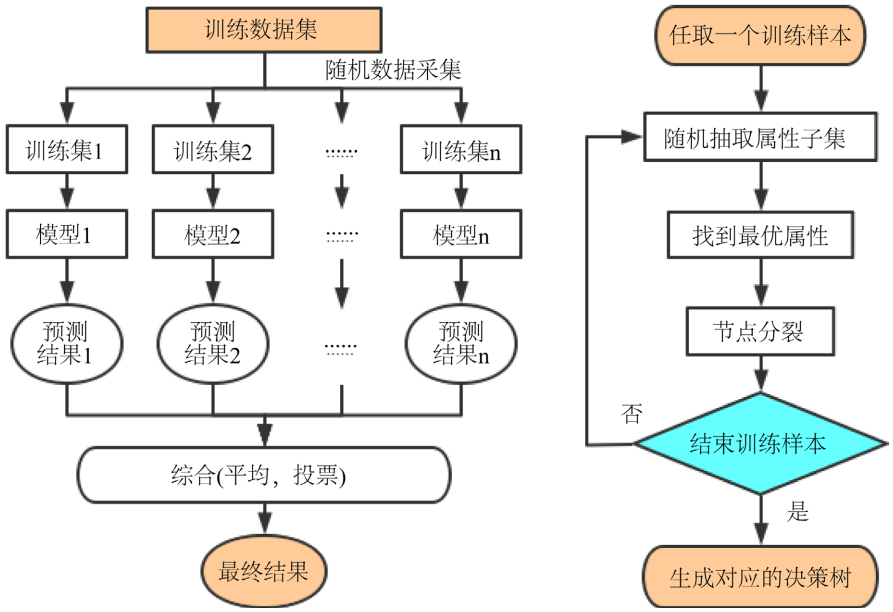

Figure 2. Training process of random forest overall training tree and single decision tree
图 2. 随机森林整体训练树以及单个决策树的训练过程

通过以上图示我们可以得到随机森林有以下特点：

1) 数据集采用随机采样的方式，其中包括样本随机采样与特征随机采样，满足了多样性的要求，使得每一个树模型都有个性。

2) 随机森林代表了一种并联思想，它同时创造多个相互独立的树模型，并且使用了一样的参数，只有输入上的不同。

3) 将所有的树模型集合在一起，求得众数即为最后的分类结果。

集成算法中综合考虑了所有树模型，这就能带来一个很实用的参数——特征重要性。特征重要性指代了数据中每一个特征的重要程度，特征重要性越大，表明该特征对预测结果的影响越大，这是因为树模型会优先考虑最优价值的特征。

使用 Python [5]运行随机森林算法，运行程序得到 729 个变量的贡献度排名，将贡献度由大到小进行排序得到图 3。考虑到下一步的高相关性滤波操作会对进一步变量进行降维，选择贡献度从大到小并贡献度总和达 90%的 70 个分子描述符作为主变量，如表 1 所示。

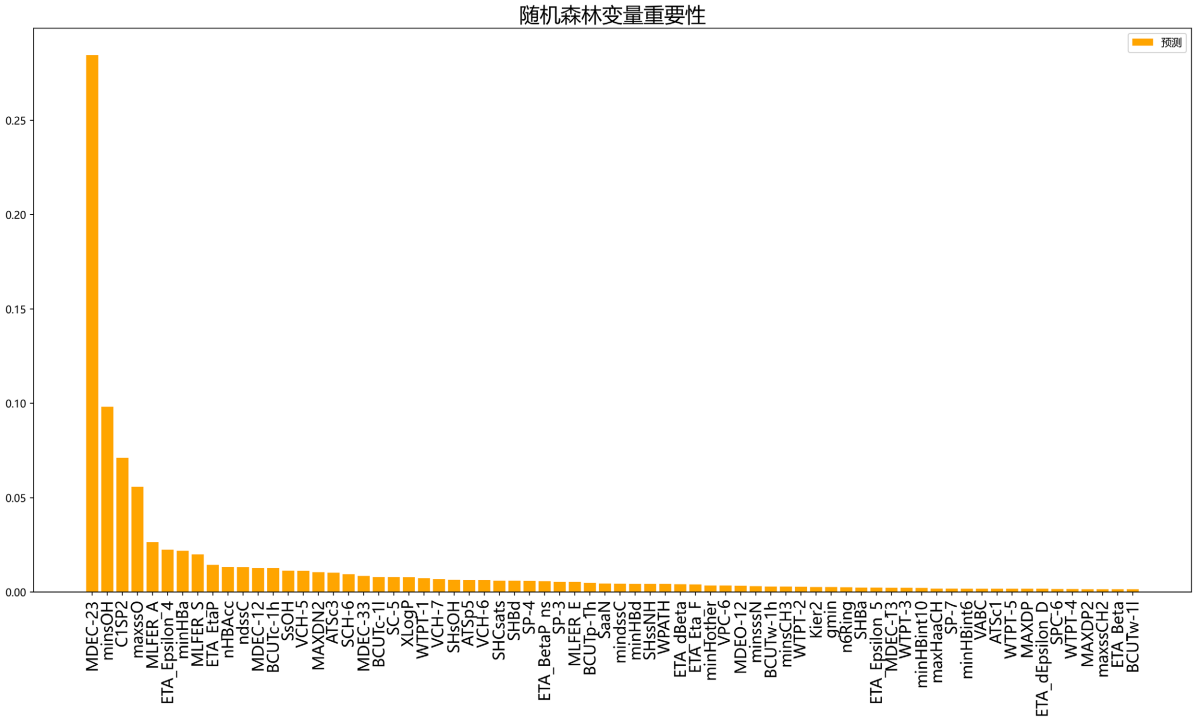

Figure 3. The contribution ranking of independent variables is obtained from random forest
图 3. 随机森林得到自变量贡献度排名

Table 1. The main variables with a total contribution of 90% and the size of the contribution
表 1. 贡献度总和达 90%的主要变量及贡献度大小

名次	变量	贡献度大小
1	MDEC-23	27.09%
2	minHsOH	8.6%
3	C1SP2	8.09%
...	...	...
70	apol	0.17%
变量个数	70	贡献度总和	90%

2.3. 相关性距离算法——耦合变量的筛除

考虑到变量分子描述符之间具有高耦合与非线性的特点，故需要进行高相关性滤波，在高度相关的变量中选出并保留贡献值最大的变量，默认其能代表所耦合变量的全部信息。

本题变量之间的关系多为非线性，传统的皮尔逊[6] (Person)相关系数仅适用于两变量呈线性关系的情况。所以我们选择距离相关系数[7] (Distance Correlation)作为衡量变量间相关性的指标。

利用距离相关系数研究独立变量下 x, y 的独立性，记录二者间的距离相关系数为 $\mathrm{dcorr}(x,y)$，当 $\mathrm{dcorr}(u,v)=0$ 时，说明变量 x 与变量 y 相互独立，设 $\{(x_i,y_i), i=1,2,\cdots,n\}$ 为总体 (x,y) 的随机样本，Szekely [8]等定义随机变量 x 与 y 的样本距离相关系数的估计值为：

$$\hat{\mathrm{d}}\mathrm{corr}(x,y)=\frac{\hat{\mathrm{d}}\mathrm{cov}(x,y)}{\sqrt{\hat{\mathrm{d}}\mathrm{cov}(x,x)\hat{\mathrm{d}}\mathrm{cov}(y,y)}} \tag{2}$$

其中 $\hat{\mathrm{d}}\mathrm{cov}^2(x,y)=\hat{S}_1+\hat{S}_2-2\hat{S}_3$，$\hat{S}_1$，$\hat{S}_2$ 和 $\hat{S}_3$ 分别为：

$$\hat{S}_1=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|x_i-x_j\right\|_{dx}\left\|y_i-y_j\right\|_{dy} \tag{3}$$

$$\hat{S}_2=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|x_i-x_j\right\|_{dx}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|y_i-y_j\right\|_{dy} \tag{4}$$

$$\hat{S}_3=\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\left\|x_i-x_l\right\|_{dx}\left\|y_j-y_l\right\|_{dy} \tag{5}$$

同理可求得 $\hat{\mathrm{d}}\mathrm{cov}(x,x)$ 和 $\hat{\mathrm{d}}\mathrm{cov}(y,y)$。

任取两个变量，若它们的距离相关系数为 0~0.2，则互相独立，即通过两个随机变量之间的距离相系数就能判断出二者相关性的强弱，统计学中相关性系数与相关性强弱关系见表 2：

Table 2. Correlation measurement table
表 2. 相关性度量表

相关性	相关系数
极强相关	0.8~1.0
强相关	0.6~0.8
中相关	0.4~0.6
弱相关	0.2~0.4
极弱或无关	0~0.2

题目要求筛选出 20 个对生物活性最具有显著影响的变量，显然筛选后的变量应为中相关及以下相关程度，所以将本题距离相关系数的阈值初始设置成 0.48，对表 1 中的变量进行相关性检验。

使用 python 程序执行此项操作，首先计算任意两变量间的距离相关系数大小，之后与设定的阈值进行比对。大于则说明两者具有高相关性，剔除两者之间贡献度排名低的，反之小于则将两者保留。程序重复上面流程，直至运算完所有数据。

2.4. 模型结果及验证

在 python 中对表 3 中的 70 个变量进行耦合变量的剔除，保留贡献度较高的变量，剔除相关联的贡献度低的变量，得到表 3 所示的 20 个主要变量，将其作为降维后的对生物活性最具有显著影响的变量供

下文使用。

Table 3. The remaining main variables after eliminating the coupling variables
表 3. 剔除耦合变量后剩下的主要变量

名次	变量名称	名次	变量名称
1	MDEC-23	11	VC-4
2	minHsOH	12	MLFER_S
3	C1SP2	13	SHBint10
4	maxssO	14	WTPT-5
5	minHssNH	15	CrippenLogP
6	ETA_Shape_Y	16	BCUTc-1h
7	VP-5	17	nHBAcc
8	maxHBd	18	ATSc2
9	VC-5	19	maxHaaCH
10	VPC-5	20	VABC

变量的降维采用了随机森林算法，得到变量贡献度的高低排名，贡献度高的保留下来，低的剔除，这样既保证了自变量的重要性又保证了其与因变量的相关性；通过距离相关系数进行判断保证了降维后变量的独立性，将变量计算后的距离相关系数进行数据可视化，得到热力图如图 4 所示。显而易见，颜色越深，变量之间相关性越小，说明保留的 20 个变量具有较好的代表性与独立性。

3. BP 神经网络的生物活性定量预测模型

通过找寻主要变量我们得知变量之间存在非线性问题，而 BP 神经网络作为人工智能的经典算法之一，强劲的非线性映射能力能够有效解决此类难题，因此变量经过降维处理后再通过神经网络进行运算更加有效的保证了模型的准确性。

3.1. BP 神经网络模型

BP (Back Propagation)神经网络[9]是一种基于误差逆向传播算法训练的多层前馈神经网络，是目前使用最为广泛的神经网络。

1) 单个神经元模型

在神经元模型中，每一个神经元接受来自其他神经元的输入信号，输入信号要经过一个带权重的连接传递，神经元接受信号进行求和运算得到总输入值，之后再将总输入值与其阈值进行比较，最后通过一个“激活函数”处理进行最终的输出，而这个输出又会作为下一个神经元的输入，就这样一层层的传递下去。单个神经元模型如图 5 所示。

2) 神经网络模型

BP 神经网络能够进行多层训练得到输入输出的映射关系。在正向传播中，输入的数据经过输入层、隐含层和输出层。其结构示意图如图 6 所示。

3) BP 学习算法

要想在模型中引入非线性需要设置激活函数，这样可以使得神经网络能够逼近任何非线性函数。若缺少激活函数，则无论多少隐含层都只有单一线性映射，即无法解决线性不可分的问题。

常见的激活函数有 Sigmoid、RelU、Thah 三种，如 Sigmoid 函数为：

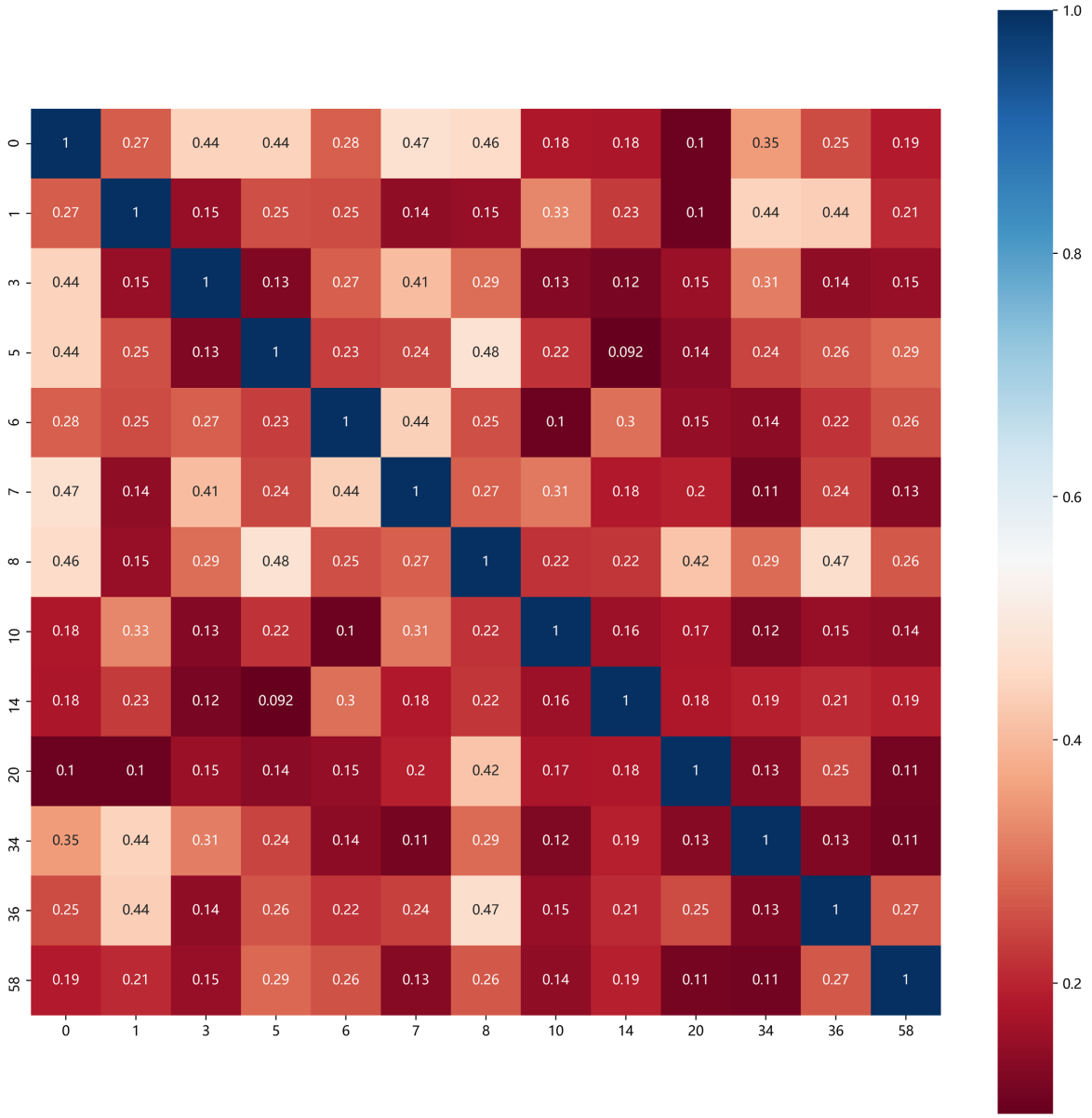

Figure 4. Thermodynamic diagram of distance correlation coefficient between variables
图 4. 变量间距离相关系数热力图

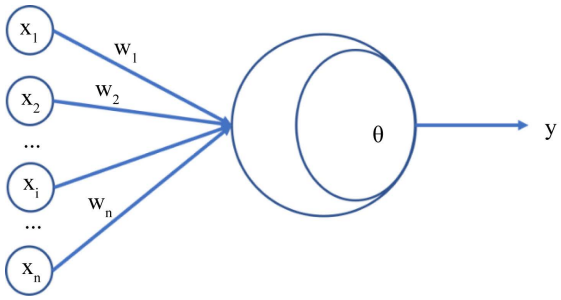

Figure 5. Single neuron model
图 5. 单个神经元模型

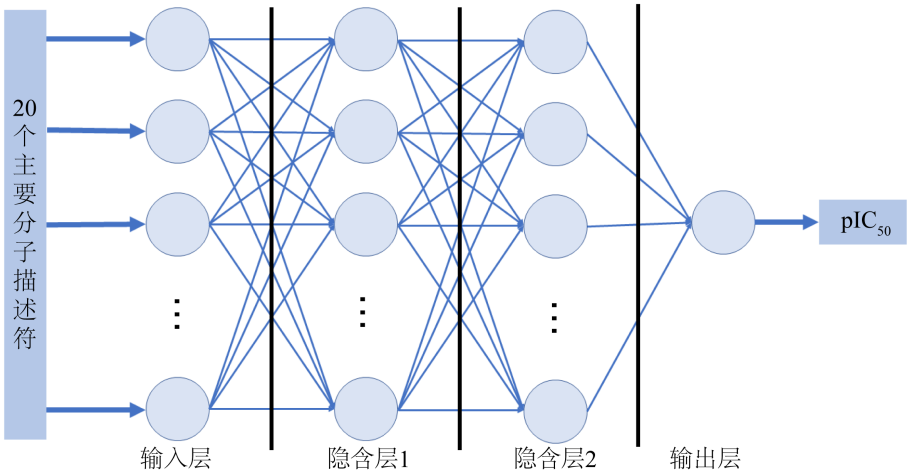

Figure 6. BP neural network model
图 6. BP 神经网络模型

$$f(x)=\frac{1}{1+e^{-Qx}} \tag{6}$$

其中，Q 为增益值，用以表征神经元非线性的参数，Q 值越大，S 曲线越陡峭；反之，Q 值越小，S 曲线越平坦。

隐含层第 i 个节点的输出 r_i，其中，节点 i 与节点 j 之间的权重为 w_{ij}：

$$r_i = f\left(\sum_{j=1}^{n} w_{ij} p_j + \theta_i\right) \tag{7}$$

输出层第 k 个神经元的输出为 a_k，b_k 为输出层节点 k 的阈值：

$$\alpha_k = f\left(\sum_{j=1}^{\delta i} w_{kj} r_i + b_k\right) \tag{8}$$

正常情况下，实际输出值与期望值会存在差异，可以通过误差函数实行校正，误差函数：

$$E(w,b)=\frac{1}{2}\sum_{k=1}^{s_2}(t_k-b_k)^2 \tag{9}$$

输出层中从第 i 个输入值第 k 个权值为：

$$\Delta\omega_{ki}=-\eta\frac{\partial E}{\partial\omega_{ki}} \tag{10}$$

隐含层中第 j 个输入值第 i 个权值为：

$$\Delta\omega_{ji}=-\eta\frac{\partial E}{\partial\omega_{ji}} \tag{11}$$

神经网络的程序框图如图 7 所示。

3.2. 生物活性 pIC_{50} 预测模型的建立

通过上述式子我们建立了含有两个隐含层的四层多输入单输出 BP 神经网络模型，并将这个模型用

于 pIC_{50} 的预测。

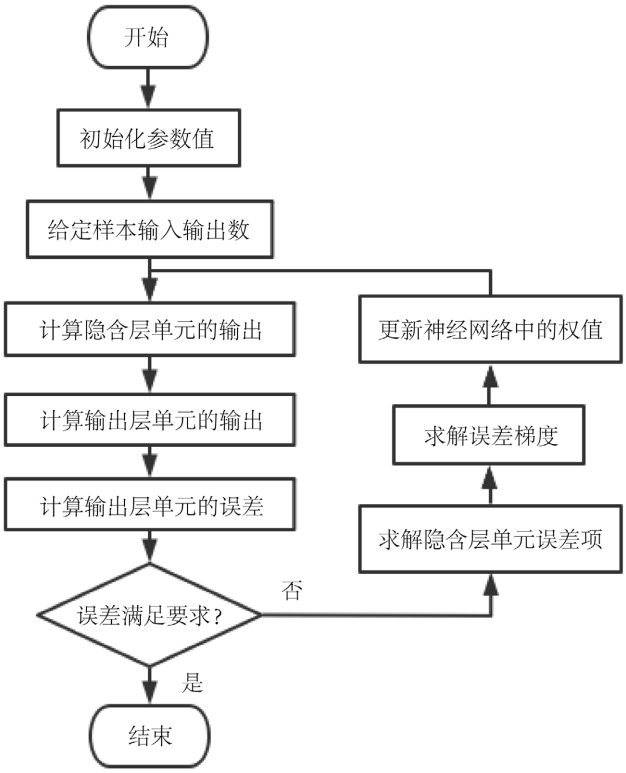

Figure 7. BP neural network program block diagram
图 7. BP 神经网络程序框图

1) 数据的准备

首先，通过上文求得的 20 个分子描述符作为本题的自变量，选择 pIC_{50} 为因变量，pIC_{50} 是 IC_{50} 值的负对数，能够清楚的表明其与生物活性的正相关性，即 pIC_{50} 数值越大，生物活性越高。之后对文件“ERα_activity.xlsx”的 train 表中 1974 个 pIC_{50} 值数据进行整理，随机选取 70%作为训练集数据，剩余 30%作为测试集数据。

2) 搭建网络结构

输入输出层：将上文得到的 20 个分子描述符作为输入，生物活性作为输出。所以有输入层神经元个数 $n = 20$，输出层神经元个数 $m = 1$。

隐含层：选择合适的隐含层神经元个数，避免计算冗杂、过度拟合等情况。本文综合考虑，在保证精度的前提下，反复调试，最终确定隐层神经元个数为 30 个。

3) 确定激活函数

本文选择 ReLU 函数作为此次神经网络的激活函数，其优点是收敛速度快，且该函数还能够稀疏化网络效果。其表达形式如下：

$$f(x)=\begin{cases}0, x \le 0\\ x, x > 0\end{cases} \tag{12}$$

4) 模型实现

网络结构为四层结构的 BP 神经网络，输入层神经元个数是 20；输出层神经元个数是 1；隐含层神

经元个数设置成 30；激活函数选择 ReLU 函数；最大迭代次数设为 1000，期望误差为 0.00000001，学习速率设为 0.001.

5) 构建数学关系式

通过上面的预测模型我们求得 pIC_{50} 数据，又从原文得知 IC_{50} 与 pIC_{50} 存在直接数学换算关系式：

$$pIC_{50} = -\lg\left(IC_{50} \cdot 10^{-9}\right) \tag{13}$$

3.3. 验证预测模型

1) 均方误差

MSE (Mean Square Error)是参数估计值与真值之差平方的期望值，常用来检测模型预测值与真实值之间的偏差，MSE 越大，误差越大，预测效果越差。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{14}$$

2) R^2

R^2 衡量的是回归方程整体的拟合度，是表达因变量与所有自变量之间的总体关系。R^2 等于回归平方和在总平方和中所占的比率，即回归方程所能解释的因变量变异性的百分比。R^2 最大值为 1。R^2 的值越接近 1，说明回归直线对观测值的拟合程度越好；反之，R^2 的值越小，说明回归直线对观测值的拟合程度越差。

$$R^2 = \frac{\left(y-\bar{y}\right)^2 - \left(y-\hat{y}\right)^2}{\left(y-\bar{y}\right)^2} \tag{15}$$

将生物化学预测模型的测试集与预测集进行 MSE 与 R^2 求解，用均方误差来描述模型预测的准确性，用 R^2 来描述模型的拟合程度，如图 8 得到此预测模型 MSE 为 0.76，图 9 得到预测模型的 R^2 为 0.8，从两图可以清楚看出真实值与预测值的准确性和拟合度都很高。

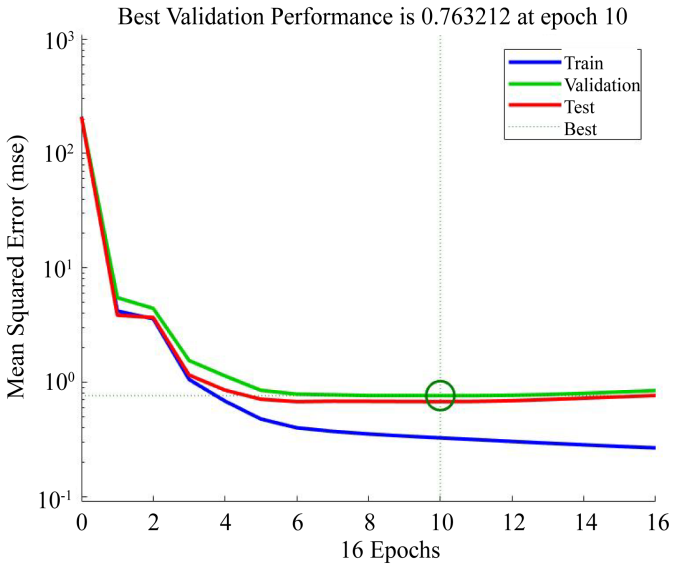

Figure 8. Mean square error of prediction model
图 8. 预测模型的均方误差

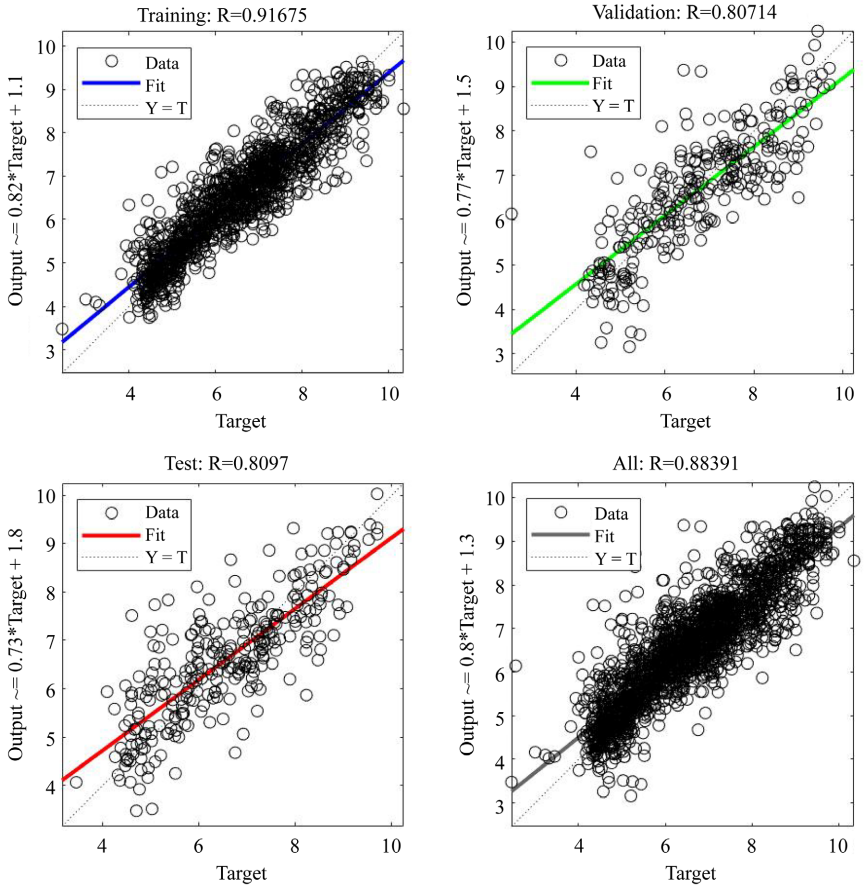

Figure 9. R-square of prediction model
图 9. 预测模型的 R 方

3) 误差比对

将附件中 pIC_{50} 的真实值与构建预测模型求得的预测值进行比较，绘制折线对比图与预测误差图，见图 10 所示。

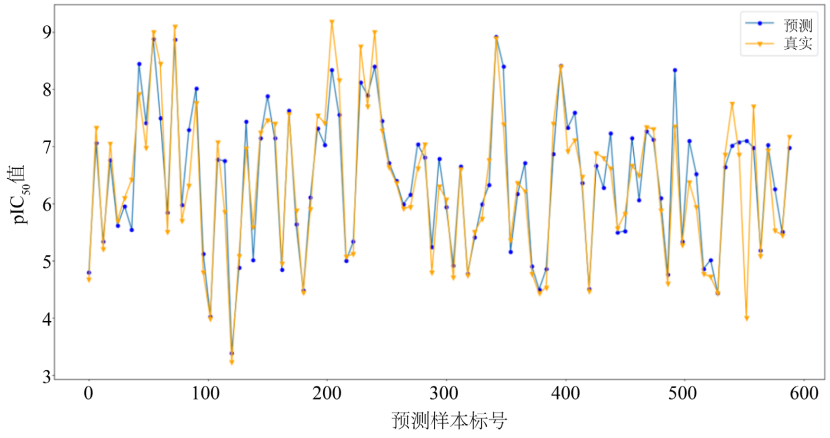

Figure 10. Comparison between the real value of pIC_{50} and the predicted value of the prediction model
图 10. pIC_{50} 真实值与预测模型预测值对比

测试集 pIC_{50} 真实值与所建模型的预测值之间存在误差，平均误差为 5.01%，如图 11 所示。

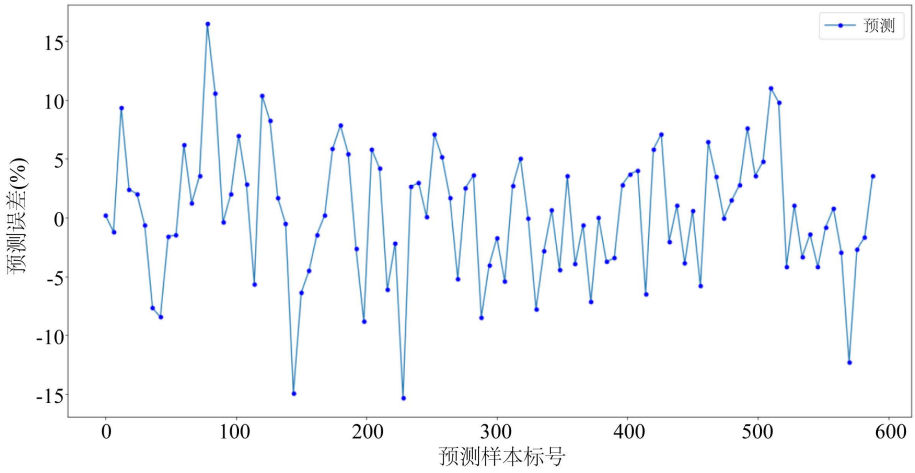

Figure 11. Error distribution of prediction model (%)
图 11. 预测模型误差分布(%)

4. 基于粒子群算法的方案优化

建立好预测模型后，为寻找处理后变量的全局最优解，采用粒子优化算法，按照贡献度选择粒子群优化算法的决策变量，以 pIC_{50} 的最大值作为目标函数，设定参数运行求得优化结果。

4.1. 基于粒子群对分子描述符的优化

粒子群优化算法(Particle swarm optimization，简称 PSO)源自对鸟群捕食的生物行为研究[10]。其基本思想是在群体中通过个体间的协作和共享信息来快速搜寻最优解，粒子能够不断更新自己位置，其位置更新方式如图 12：

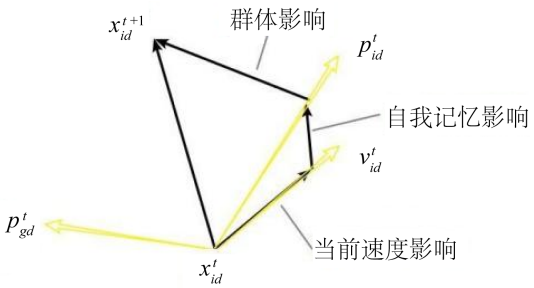

Figure 12. Particle motion diagram
图 12. 粒子运动图

假设某个 D 维空间中有 M 个粒子组成一个群落，则第 i 个粒子的运动轨迹图如上图，x 表示粒子某时刻 t 的初始位置，v 为粒子在此刻运动的速度，p 为粒子搜寻的最优解位置。

粒子 i 搜寻到的个体最优解记为 p_{best}，公式：

$$p_{best} = \left(p_{i1}, p_{i2}, p_{i3}, \cdots, p_{iD}\right), i = 1, 2, \cdots, N \tag{16}$$

整个搜寻到的全局最优解记为 g_{best} 公式：

$$g_{best} = \left(p_{g1}, p_{g2}, p_{g3}, \cdots, p_{gD} \right), i = 1, 2, \cdots, N \tag{17}$$

当粒子搜寻到这两个极值时，会根据下面式子更新自己位置与速度：

$$v_{id}^{t+1} = \omega v_{id}^{t} + c_1 r_1 \left(p_{id} - x_{id} \right) + c_2 r_2 \left(p_{gd} - x_{gd} \right) \tag{18}$$

其中 c_1 和 c_2 为学习因子，r_1 和 r_2 为范围[0, 1]内的随机数。

粒子群算法的算法流程：初始化粒子群，群体规模、随机位置等；计算每个粒子的适应度值；将每个粒子的适应度值与其个体极值对比，根据要求决定是否替换个体适应值；将每个粒子的适应度值与全局极值对比，根据要求决定是否替换个体适应值；更新粒子的速度和位置；判断是否满足终止条件，不满足返回第 2 步(算法终止通常要到达最大迭代次数或者最佳适应度值的增量小于某个给定阈值)，如图 13 所示。

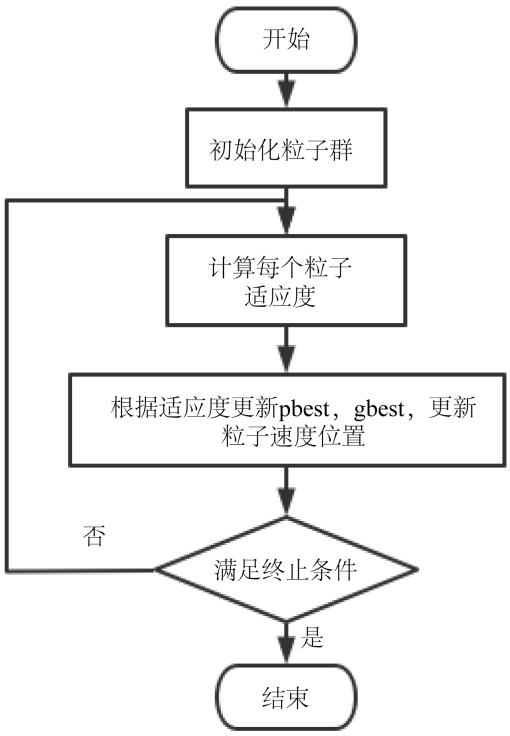

Figure 13. Flow chart of particle swarm optimization algorithm
图 13. 粒子群算法流程图

4.2. 粒子群算法的设定及求解

1) 决策变量：

通过前面数据处理得到 20 种分子描述符，将这 20 个分子描述符记为主要变量(题目未强调分字符中有不可变量，默认均为可变量)，记为：

$$X = \left\{ x_1, x_2, x_3, \cdots, x_{20} \right\} \tag{19}$$

2) 目标函数：

以 BP 神经网络的输出最大值(即 pIC_{50} 的最大值)作为粒子群优化算法的目标函数。

3) 约束条件：

针对此优问题，数据降维后 20 个变量存在最大值 N_{max} 与最小值 N_{min}，所以有约束：

$$N_{min} \le x_i \le N_{max}, i = 1,2,\cdots,50 \tag{20}$$

4) 确定粒子群参数：

采用局部粒子群算法求解本文问题，综合考虑后算法的参数值设置如表 4 所示：

Table 4. Parameter setting of particle swarm optimization algorithm
表 4. 粒子群算法的参数设置

参数名称	符号	数值
群体大小	m	300
维数	D	50
权重因子	ω	0.1~1.1 自适应
学习因子	c_1c_2	1.49 (默认)
粒子最大运动速度	v_{max}	依照附件二
粒子最大迭代步数	t	1000
各粒子起始位置	x_0	随机数
各粒子起始速度	v_0	随机数

针对上述建立的复杂多约束优化模型，本文使用 MATLAB 语言编写粒子群算法的程序对模型进行求解。对目标函数 pIC_{50} 最大值进行迭代可视化处理，由于粒子群优化算法求得的结果是最小值，通过取反处理可得到 pIC_{50} 最大值，即 13.1345，迭代图见图 14。

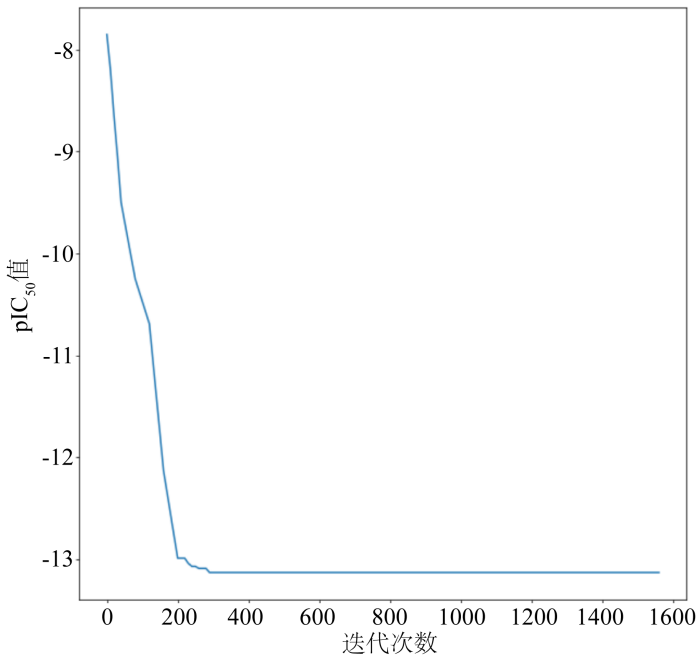

Figure 14. Iterative graph of pIC_{50} maximum
图 14. pIC_{50} 最大值的迭代图

选取多次不同的迭代次数运算后，pIC_{50} 的最大值始终稳定在 13.1345，且对应的分子描述符都在合理的区间范围内，说明此次建立的模型具有一定的稳定性与合理性。

5. 结束语

本文提出一整套为优化 ERα 拮抗剂的生物活性提供预测服务的数学方法。从选取数据的主要变量，建立到生物活性 pIC_{50} 的预测，最后确定分子描述符及其取值范围。每一步对选取 ERα 拮抗剂的化合物都有重要意义。

本文模型多处满足药物研发过程中对活性化合物的实际需求，如筛选主要的分子描述符时保证代表性和独立性的要求；通过已有化合物的生物活性对其他化合物进行定量预测；以及对分子描述符量值的确定。综上可知，在工程应用方面，本文对药物研发有较高的应用价值；在学术研究方面，本文对数据降维和模型预测提出创新思想，对其他相关研究具有一定的参考意义。

参考文献

[1] 后梦玥, 黄照权. 雌激素受体-α36 与乳腺癌的关系研究进展[J]. 中国医药导报, 2021, 18(14): 36-38+50.

[2] 王勇, 王晓东, 陈文捷, 于兆进, 吴慧哲, 赵琳, 魏敏杰. 散发性乳腺癌中 DNMT3a、DNMT3b 表达与 ERα 基因启动子甲基化状态及蛋白表达的关系[J]. 天津医药, 2015, 43(5): 500-504.

[3] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[4] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[5] 王娟, 华东, 罗建平. Python 编程基础与数据分析[M]. 南京: 南京大学出版社, 2019.

[6] Li, J., Wang, B. and Li, H. (2021) Research on Computer Forecast Model Using BP Neural Network and Pearson Correlation Coefficient. .*Journal of Physics: Conference Series*, **2033**, Article ID: 012091. https://doi.org/10.1088/1742-6596/2033/1/012091

[7] Li, W., Meng, X. and Huang, Y. (2021) Fitness Distance Correlation and Mixed Search Strategy for Differential Evolution. *Neurocomputing*, **458**, https://doi.org/10.1016/j.neucom.2019.12.141

[8] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007) Measuring and Testing Dependence by Correlation of Distances. *Annals of Statistics*, **35**, 2769-2794. https://doi.org/10.1214/009053607000000505

[9] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Internal Representation by Back-Propagation Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[10] 张刘, 叶楠, 马灵玲, 汪琪, 吕雪莹, 章家保. 改进粒子群优化算法的高光谱波段选择[J]. 光谱学与光谱分析, 2021, 41(10): 3194-3199.