Vol. 30 高等学校化学学报 No. 10
2009 年 10 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 1960 ~ 1964

纤维堆囊菌发酵液中埃博霉素含量的 HPLC 法分析

孟凡欣, 郭伟良, 逯家辉, 杜林娜, 李又欣, 滕利荣
(吉林大学生命科学学院, 长春 130012)

摘要 采用反馈神经网络结合遗传算法(BPANN-GA)对高效液相色谱(HPLC)法同时测定纤维堆囊菌(*Sorangium cellulosum*)代谢物中埃博霉素 A(Epo A)和埃博霉素 B(Epo B)含量的条件进行优化, 采用均匀设计(U_{12}^{3})方案对流动相中乙腈的体积分数、色谱柱温度和流动相的 pH 等 3 个因素进行实验设计; 以色谱函数(COF)值为优化指标, 运用双层反馈神经网络建立色谱优化函数(COF)值, 考察因素间的预测模型, 采用 Levenberg-Marquardt backpropagation 算法对所建立的神经网络预测模型进行训练, 以逼近度(D_a)为优化参数, 选择预测模型的最适隐含层节点数. 最优预测模型预测的 COF 值与实验值之间的相关系数(R)达到 0.98165, 采用遗传算法在实验考察范围内进行全局寻优, 得到最优化的 HPLC 分析条件: 流动相中乙腈体积分数为 29.2%, 色谱柱温度为 34 ℃, 流动相 pH 为 4.23. 在此最优条件下对纤维堆囊菌代谢产物进行 HPLC 分析, 结果表明, 该方法对两种埃博霉素色谱峰均具有较好的分离度.

关键词 反馈神经网络(BPANN); 遗传算法(GA); 埃博霉素; 高效液相色谱(HPLC)
中图分类号 O657.7 **文献标识码** A **文章编号** 0251-0790(2009)10-1960-05

埃博霉素(Epothilone, Epo)是从纤维堆囊菌(*Sorangium cellulosum*)发酵液中分离出来的细胞毒化合物, 具有与紫杉醇相似的生物活性. 埃博霉素是一类 16 碳的大环内酯化合物, 已报道有多种生物活性的类似物, 如埃博霉素 A(Epo A)、埃博霉素 B(Epo B)和埃博霉素 D(Epo D). Epo 具有水溶性好、结构简单、有良好的化学修饰潜力及对紫杉醇耐药的肿瘤细胞具有高活性等优点, 是极具有市场潜力的新型抗癌新药[1~4]. 纤维堆囊菌是一种黏细菌, 是最高等的革蓝氏阴性单细胞细原核生物类菌, 具有多细胞行为. 黏细菌可产生丰富的次级代谢产物, 目前已从中发现大约 360 多种生物活性物质, 约占微生物来源总数的 3.5%, 而溶纤维素群的黏细菌如纤维堆囊菌 95% 的菌株可产生生物活性物质. 由于纤维堆囊菌发酵液中代谢产物种类繁多, 且多数产量很低, 在采用高效液相色谱(HPLC)分析时目标产物峰与杂质峰难以分开, 因此给纤维堆囊菌生物活性物质高产菌株的筛选、改良和培养条件优化等工作造成很大困难[5~8]. 人工神经网络(ANN)是一种模拟人脑功能并由大量神经元相互作用形成网络的高度非线性动力学系统, 常用于考察因素与响应值之间关系高度非线性或不明确的数学建模, 具有高度的鲁棒性和容错能力, 能充分逼近复杂的非线性关系. 最常见的人工神经网络为反馈神经网络(BPANN)[9~13]. 遗传算法(GA)是根据生物进化的模型提出的一种优化算法, 它是基于自然选择和基因遗传学原理的高度并行、随机和自适应搜索算法, 具有无需知道目标函数的具体形式且得到的优化解为全局最优等特点[14,15]. 色谱优化函数(COF)则综合考虑所考察目标产物组分的色谱峰与相邻杂质峰的分离程度, COF 值越高, 目标产物组分的色谱峰分离度越好. 本文采用均匀设计方法进行实验设计, 再采用双层的反馈神经网络建立了流动相中乙腈体积分数、色谱柱温度和流动相 pH 与两种 Epo 色谱优化函数值间的相关模型, 结合遗传算法对所建立模型在实验考察范围内进行全局寻优, 得到利用 HPLC 法同时测定纤维堆囊菌代谢产物中 Epo A 和 Epo B 含量的最优条件.

收稿日期: 2009-01-24.
基金项目: 中国医学基金会新药发展基金(批准号: 20061108)资助.
联系人简介: 滕利荣, 男, 教授, 博士生导师, 主要从事微生物与生化药学研究. E-mail: tenglr543@gmail.com

1　实验部分

1.1　试剂与仪器

乙腈和甲醇(Fisher CHemAlertGuide 公司，色谱纯)，乙酸乙酯(国产分析纯)，超纯水，Epo A 和 Epo B 标准品(美国 Sigma 公司)，HLD-16 中性大孔树脂(上海华羚树脂有限公司)，其它试剂为分析纯. HZQ-F160 全温振荡培养箱(哈尔滨东联电子技术开发有限公司)；Shimadzu SCL-10A vp 高效液相色谱系统(日本岛津公司)，包括 LC-6AD vp 泵、SPD-A vp 紫外-可见检测器、AT-330 柱温箱(天津奥特塞恩斯仪器有限公司)和 Nova-pak C_{18}色谱柱(3.9 mm×150 mm).

1.2　实验过程

样品液的制备：采用亚硝酸-紫外复合诱变纤维堆囊菌 ATCC15384，得到 Epo 高产菌株纤维堆囊菌 ATCC15384-UN16H127. 在 500 mL 的摇瓶中装入 200 mL ATCC 液体培养基，按 10% 的接种量接入培养 16 h 的纤维堆囊菌 ATCC15384-UN16H127 种子液中，加入 3% 的 HLD-16 中性大孔树脂，30 ℃摇床，以转速 150 r/min 培养 5 d，取出，静置至 HLD-16 中性大孔树脂完全沉至瓶底，弃去上层发酵液，用去离子水将 HLD-16 中性大孔树脂洗涤至水无色，放入 60 ℃烘箱中烘干，加入 10 倍体积的乙酸乙酯进行解吸过夜，解吸 2 次，合并解吸液；用 45 ℃旋转蒸发仪将乙酸乙酯全部蒸出，用 2 mL 甲醇复溶，过 0.22 μm 的微孔滤膜，作为待测样品.

标准溶液的配制：用甲醇(色谱纯)配制 4 μg/mL 的 Epo A 和 Epo B 标准溶液.

高效液相色谱条件：流速 1 mL/min，检测波长 250 nm，进样量 10 μL，流动相采用超纯水和乙腈按一定比例配制，采用磷酸调节流动相的 pH，流动相中乙腈的体积分数、色谱柱温度和流动相 pH 按照 3 因素 12 水平的均匀设计(U_{12}^3)表进行设计，设计方案如表 1 所示.

色谱优化函数(COF)值按下式计算：

$$\mathrm{COF}=\sum_{i=1}^{n}A_i\ln(R_i/R_{id}) \tag{1}$$

式中，n 为分析组分数目，A_i为第 i 组分的加权因子，R_i和 R_{id}分别为第 i 组分的分离度和理想分离度，本文中 R_{id}取 1.5，Epo A 和 Epo B 两个组分的 A_i值均设为 0.5.

Table 1　U_{12}^3 uniform design and experiment results

Run	Volume fraction of cetonitrile(%)	Temperature/℃	pH	COF value	Run	Volume fraction of cetonitrile(%)	Temperature/℃	pH	COF value
1	26.0	34	3.4	−0.051	7	29.0	31	4.0	0.025
2	26.5	40	3.7	−1.023	8	29.5	37	4.3	−0.158
3	27.0	33	4.0	−4.940	9	30.0	30	3.4	−4.523
4	27.5	39	4.3	−0.561	10[a]	30.5	36	3.7	−4.205
5	28.0	32	3.4	−0.328	11[b]	31.0	29	4.0	−0.395
6	28.5	38	3.7	−1.383	12	31.5	35	4.3	−1.949

a. Validation set; *b*. prediction set.

2　结果与讨论

2.1　BPANN 模型的建立

以乙腈体积分数、色谱柱温度和流动相 pH 为输入变量，COF 值作为输出变量(如表 1 所示)，对输入变量进行归一化处理；采用 Matlab R2008a 的 Neural Fitting 工具箱构建双层反馈神经网络(BPANN)，其拓扑结构如图 1 所示. 隐含层传递函数为 Sigmoid 传递函数，输出层传递函数为线性传递函数. 随机将表 1 所示的 12 组实验分成训练集样本、预测集样本和验证集样本，将训练集样本(10 组实验)用于模型的训练，验证集样本(1 组试验)用于验证模型的泛化能力. 在模型训练过程中，如果模型的泛化能力不再提升，模型将停止训练. 将预测集样本(1 组实验)用于检验模型的预测能力，其本身不参与模型的训练，采用 Levenberg-Marquardt Backpropagation 算法对模型进行训练，均方差(MSE)作为模型的性能参数，其计算公式如下：

$$\mathrm{MSE}=\left[\sum_{i=1}^{n}(\mathrm{COF}_{\mathrm{BPANN}i}-\mathrm{COF}_{\mathrm{ACTU}i})^2\right]/n,\ i=1,2,3\cdots n \tag{2}$$

式中，$\mathrm{COF}_{\mathrm{BPANN}i}$和 $\mathrm{COF}_{\mathrm{ACTU}i}$分别为神经网络预测的 COF 值和实验得到的 COF 值；n 为样本数.

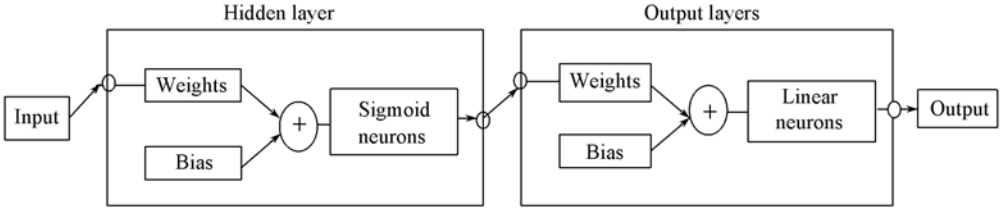

Fig. 1 Topology structure of BPANN

2.2 BPANN 模型的优化

在构建 BPANN 模型时，拓扑结构参数对模型的拟合性能和预测性能及泛化能力均有很大的影响. 输入层以流动相中的乙腈体积分数、色谱柱温度和流动相 pH 为输入节点，COF 值作为输出节点. 不同隐含层节点数对网络的性能参数有很大的影响，隐含层节点数越多，模型的拟合就越充分，但隐含节点数过多则会造成过拟合现象，即模型可以无限逼近目标值，但模型的预测能力和泛化能力会大大下降. 为此，本文引进逼近度(D_a)作为评价参量，并综合考虑模型的拟合度、模型的预测性能和泛化能力，可有效地避免模型出现过拟合现象，逼近度(D_a)计算公式如下：

$$D_a=\frac{c}{\frac{n_c}{n}\times\mathrm{MSE}_c+\frac{n_t}{n}\times\mathrm{MSE}_t+|\mathrm{MSE}_c-\mathrm{MSE}_t|} \tag{3}$$

式中，n，n_c，n_t分别为所有样品数、校正集样本数和验证集样本数，MSE_c和 MSE_t分别为校正集和验证集的均方差，c 是根据作图需要而设定的常量，本文取 $c=1$. 不同的隐含节点数对 BPANN 模型的 D_a影响如图 2 所示. 由图 2 可以看出，隐含节点数与 D_a并不是简单的线性相关，这是由于 BPANN 模型的性能也受其随机赋予的初始权值的影响，在隐含节点数为 9 时，D_a最大值为 3.74362，BPANN 的最适隐含节点数为 9.

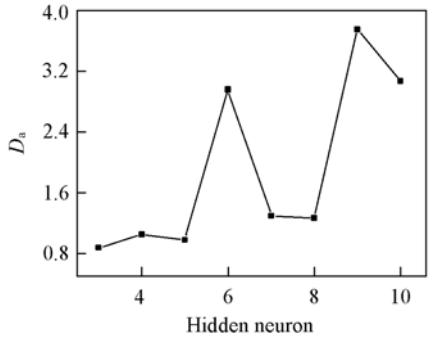

Fig. 2 Effect of number of hidden neurons on D_a

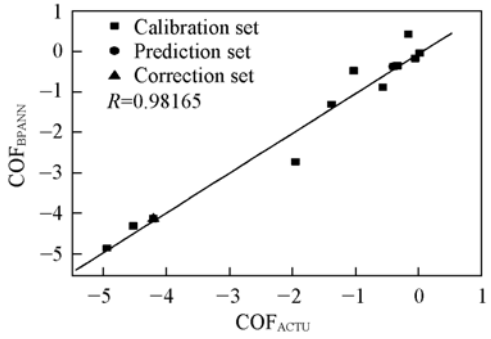

Fig. 3 Correlation between $\mathrm{COF_{ACTU}}$ and $\mathrm{COF_{BPANN}}$

2.3 BPANN 最优模型的建立

经过上述优化，BPANN 模型最优拓扑结构(输入节点-隐含层节点-输出节点)为 3-9-1. 隐含层采用 Sigmoid 传递函数，输出层采用线性传递函数，采用 Levenberg-Marquardt Backpropagation 算法对模型进行训练，将最大训练次数设为 1000. 所建立的 BPANN 模型对样本的 COF 值进行预测，预测值与实验值间的相关性如图 3 所示. 结果表明，预测值与实验值间的相关系数(R)达到 0.98165；校正集均方差($\mathrm{MSE_c}$)为 0.14258，表明模型预测值与实验值吻合度好；预测集均方差($\mathrm{MSE_v}$)为 0.00073，表明预测能力均很好；验证集均方差($\mathrm{MSE_t}$)为 0.00559，表明模型的泛化能力很好，可以进行下一步优化.

2.4 遗传算法全局寻优

采用遗传算法全局寻优的步骤如图 4 所示.

（1）将各个变量约束在考察的范围内；（2）选择编码策略，把参数集合转换成位串结构空间，确定遗传策略，包括选择群体大小（n）、确定选择、交叉、变异方法以及确定交叉概率和变异概率等参数；（3）随机生成初始化群体；（4）将随机生成的初始化群体解码得到输入参数，用 BPANN 模型计算，得到目标函数值（Fitness）；（5）按照遗传策略，运用选择、交叉、变异算法作用于群体，形成下一代群体；（6）判断群体性能是否满足某一指标，或者已经完成预定叠代次数，若不满足则返回步骤（4），将形成的子代群体代替随机生成的初始化群体；（7）把遗传算法运行结束后的最佳个体解码所得参数作为模型的最优解，对应的网络输出值即为最优化值；采用 Marlab R2008a 的 Optimization tool 工具箱中的 ga-Genetic Algorithm 部分对所建立的 BPANN 模型在实验考察范围内进行最小值全局寻优，群体类型选择双精度向量类型，群体的每代染色体种群数为 20，将适应度（Fitness）函数定义为

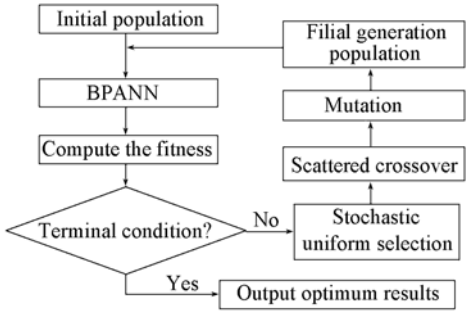

Fig. 4　Process of genectic algorithm

$$\begin{cases}\text{Fitness}=15, & \text{COF}\leqslant 0\\ \text{Fitness}=1/\text{COF}, & \text{COF}>0\end{cases} \tag{4}$$

在求出所有 20 个染色体的适应度之后，采用 Stochastic Uniform 选择法来进行种群选择，每个个体之间通过散点式交叉（Scattered crossover）和变异，可产生新的子代种群，设定交叉概率为 0.8，最大遗传代数为 100，应用遗传算法在表 1 的实验考察范围内进行全局最低的适应度（Minimum fitness）搜索，经过 51 次遗传操作后，由于平均 Fitness 变化率过小而终止遗传操作. 图 5 为每次遗传操作过程中 Minimum fitness 与遗传代数的相关图. 由图 5 可以看出，当遗传代数达到 20 代以上时，Minimum fitness 不再降低，进而搜索到全局最优条件值. 经过遗传算法搜索得到最优的 HPLC 条件为：乙腈体积分数 29.2%，柱温箱温度 34 ℃，流动相 pH 4.23，BPANN 模型预测的最优 COF 值 1.83.

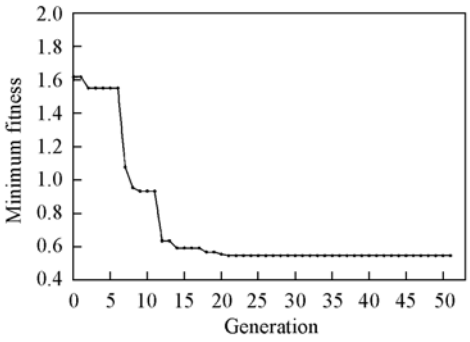

Fig. 5　Correlation between minimum fitness and generation

2.5　最优 HPLC 分析条件的检验

采用 BPANN 模型结合遗传算法优化，得到的 HPLC 法测定纤维堆囊菌 ATCC15384-UN16H127 代谢物中 Epo A 和 Epo B 的最优分析条件为：乙腈体积分数 29.2%，柱温箱温度 34 ℃，流动相 pH 4.23，在此最优条件下对样品进行分析，得到最优的 COF 值 0.6171 比优化之前最优的 COF 值 0.025 提升了 24 倍多，大大改善了两组分的分离效果. 虽然实验 COF 值与模型的预测值相差较大，但如果每个考察组分的色谱峰均达到理想分离度后，COF 值的大小意义将不是很大. 在最优 HPLC 条件下得到 Epo A 和 Epo B 的分离度 $R_{\text{Epo A}}$ 和 $R_{\text{Epo B}}$ 分别为 2.148 和 3.197，两个组分的分离度均超出理想分离度 R_{id} 值，得到的 HPLC 色谱图如图 6 所示. 由图 6 可以看出，Epo A 和 Epo B 两个组分的峰能较好的与相邻的峰分开，能满足 HPLC 定量分析的要求，表明采用 BPANN 结合遗传算法优化 HPLC 测定条件是可行的.

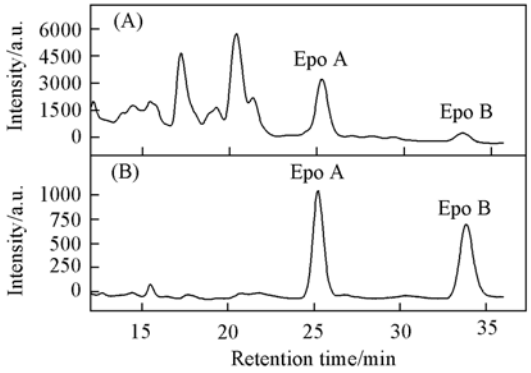

Fig. 6　Chromatograms of the sample（A）and the standards（B）in the optimum HPLC conditions

参 考 文 献

[1] Fumoleau P., Coudert B., Isambert N., *et al.*. Ann. Oncol. [J], 2007, **18**(Supplement 5): 9—15

[2] Christopher T., Walsh S. E., O'Connor Tanya L. S.. J. Ind. Microbiol. Biotechnol. [J], 2003, **30**: 448—455

[3] Altmann K., Gertsch J.. Nat. Prod. Rep. [J]. 2007, **24**: 327—357

[4] Bode H. B., Muller R.. J. Ind. Microbiol. Biotechnol. [J], 2006, **33**: 577—588

[5] Christopher T. W., Sarah E. O., Tanya L. S.. J. Ind. Microbiol. Biotechnol. [J], 2003, **30**: 448—455

[6] Gerth K., Pradella S., Perlova O., *et al.*. J. Biotechnol. [J], 2003, **106**: 233—253

[7] Beyer S., Kunze B., Silakowski B., *et al.*. Biochimical et Biophysica Acta[J], 1999, **1445**: 185—195

[8] Klaus G., Heinrich S., Gerhard H., *et al.*. J. Antibiot. [J], 2000, **53**(12): 1373—1377

[9] CHEN Xiao-Mei(陈晓梅), RAO Han-Bing(饶含兵), HUANG Wen-Li(黄文丽), *et al.*. Chem. J. Chinese Universities(高等学校化学学报)[J], 2007, **28**(11): 2171—2178

[10] YIN Chun-Sheng(印春生), SHEN Yang(沈阳), LIU Shu-Shen(刘树深), *et al.*. Chem. J. Chinese Universities(高等学校化学学报)[J], 2000, **21**(1): 49—52

[11] Bassheer L. A., Hajmeer M.. J. Microbiol. Meth. [J], 2003, **43**: 3—31

[12] Pons M. N., Bonte A. L., Potier O.. J. Biotechnol. [J], 2004, **113**: 211—230

[13] Mukta P., Usha A. K.. Expert. Syst. Appl. [J], 2009, **36**: 2—17

[14] Franco-Lara E., Link H., Weuster-Botz D.. Process Biochem. [J], 2006, **41**: 2200—2206

[15] Potocnik P., Grabec I.. Math. Comput. Simulat. [J], 1999, **49**: 363—379

HPLC Optimization for Analysis of Epothilones in *Polyangium Cellulosum* Fermentation Metabolites

MENG Fan-Xin, GUO Wei-Liang, LU Jia-Hui, DU Lin-Na, LI You-Xin, TENG Li-Rong*

(*College of Life Science*, *Jilin University*, *Changchun* 130012, *China*)

Abstract Back-propagation artificial neural network combined with genetic algorithm (BPANN-GA) was applied to optimize the high performance liquid chromatography (HPLC) conditions for the determination of epothilone A (Epo A) and epotholine B (Epo B) simultaneously in *Polyangium cellulosun* metabolites. The concentration of acetonitrile in mobile phase, column temperature and the pH of mobile phase were selected as casual factors and a three-factor-twelve-level uniform design (U_{12}^{3}) was used for experiment design. A two-layer back-propagation artificial neutral network (BPANN) was applied to model for the correlation between the casual factors and chromatography optimization function (COF) values, which was chosen as the criterion. Levenberg-Marquardt algorithm was used for training the BPANN. The BPANN model was optimized by selecting the most suitable numbers of hidden neurons depending on the degree of approximation (D_a). The correlation coefficient (R) between the COF values obtained by BPANN model and the experiment values was 0.98165. While the optimum BPANN model was developed, genetic algorithm (GA) was applied to find out global dissolution in modeling range. The optimum HPLC conditions obtained by BPANN-GA were as follows: the concentration of acetonitrile in mobile phase was 29.2% (volume fraction); the column temperature was 34 ℃ and pH of mobile phase was 4.23. The validation experiment at the optimum conditions was performed, and the satisfied chromatogram was obtained.

Keywords Back-propagation artificial neural network (BPANN); Genetic algorithm (GA); Epothilone; High performance liquid chromatography (HPLC)

(Ed.: H, J, Z)