DOI: 10. 3724/ SP. J. 1096. 2011. 01701

人工神经网络-紫外光谱法短串联重复序列基因分型检测

汪雪娇 牟红元 鲁 辉 豆兴茹 邱 婷 谢洪平*
(苏州大学医学部药学院, 苏州 215123)

摘 要 以 STR 基因座 D16S539 中的总核心重复串数相差较小的 10-11, 10-12, 11-11 和 10-13 基因型为研究对象, 以紫外光谱为判别变量, 建立了以人工神经网络(ANN)提取富信息变量为基础的 ANN 基因分型方法。在优化条件下, 对 4 个基因型样本进行了聚合酶链式反应扩增, 以扩增样本在 200~ 310 nm 范围内的检测光谱进行预处理和偶合的 ANN-ANN 网络优化。结果表明, 提取富信息变量和基因分型的 ANN 的最优网络结构分别为 391-50-391 和 50-6-4, 该结构下的判别模型的校正相对均方根误差(RMS)和预测 RMS 分别是 0. 0279 和 0. 0418, 模型表现出了良好的稳健性和 100% 的基因型正确预测率。成功实现了基于紫外光谱对 STR 基因型的快速、简单和低成本检测。

关键词 短串联重复序列; 紫外光谱; 人工神经网络; 基因分型

1 引 言

短串联重复序列(STR)是一类广泛分布在人和动物基因组中的 DNA 重复序列片段, STR 的核心序列为 2~ 6 bp, 呈串联重复排列, 重复 15~ 30 次左右, 其总长度一般位于 100~ 400 bp 之间, 多位于基因编码区附近、基因内含子和非翻译区。按孟德尔规律呈显性遗传[1]。常见的 STR 形式有二核苷酸重复$(CA)_n$, $(GA)_n$, $(AA)_n$, $(GG)_n$, 三核苷酸重复$(CAG)_n$, $(CGG)_n$ 和四核苷酸重复$(GATA)_n$, $(AGAT)_n$ 等。据估计, 人类基因组约有 50 万个 STR 位点。STR 具有种类多、分布广、多态信息量大、杂合度高、片段较短、易于聚合酶链式反应(Polymerase chain reaction, PCR)扩增等优点[2], 因而在人类学、生命科学和医学等领域得到广泛应用[3~5]。目前, STR 基因型的检测常采用凝胶电泳法[6,7]、毛细管电泳法[8~10]、微芯片电泳法[11,12]、和基因测序[13~15]等方法, 这些方法可以较准确地对 STR 基因型进行检测, 但通常需要较为复杂的样品前处理(如分离、纯化)或 PCR 引物标记(如荧光标记、放射线标记)。对于 STR 基因型质谱检测法[16,17], 虽然不需要标记, 但仪器较为昂贵。结合化学模式识别的近红外光谱法实现了一次 PCR 扩增、无需分离纯化、无需荧光标记、低成本的基因型快速检测[18]。但是, 在光谱分析方法中, 近红外光谱法的应用广谱性相对较低。本研究以 STR 的紫外光谱(UVS)为识别变量, 采用人工神经网络(ANN)化学模式识别方法, 建立了不同 STR 基因型的判别模型。

2 实验部分

2.1 仪器与试剂

PTC-225 型 PCR 扩增仪(美国 Bio-Rad 公司); POWER BC6003En 型电泳仪(上海申能博彩生物科技有限公司); GeneGenius 凝胶成像系统(英国 Syngene 公司); 离心机(日本 Tomy 公司); 移液器(德国 Eppendorf 公司); 微型混合器(上海沪西仪器厂); UV2401 型紫外分光光度计(日本 Shimazdu 公司)。

Taq DNA 聚合酶(美国 Fermentas 公司); PCR 引物(PAGE 级, 金思特科技有限公司); 琼脂糖(LP0028A, 英国 Oxoid 公司); 溴化乙锭(Sangon 生物公司); 人血液样本(苏州大学司法鉴定所); 其它试剂均为分析纯, 所有溶液由灭菌去离子水配制。

2.2 基因组 DNA 的提取

按 Chelex 100 法[19]提取基因组 DNA。取 6 名以 EDTA 抗凝的 D16S539 基因座已知基因型的受试

2011-01-15 收稿; 2011-04-19 接受
* E-mail: hpxie@suda. edu. cn

者的全血 0.2~1.5 mL 于离心管中, 滴加红细胞裂解液 1 mL, 以 10000 r/min 离心 5 min; 弃去上清液, 加入磷酸盐缓冲液 1 mL, 离心; 弃去上清液, 重复上述过程 1 次。加入 5% Chelex100 悬浮液 0.2 mL, 悬浮细胞核, 在 56 ℃保温 30 min; 在沸水浴中保温 8 min, 立即置于- 20 ℃冰箱保存备用。

2.3 PCR 扩增

PCR 扩增条件为: 94 ℃预变性 3 min, 94 ℃变性 20 s, 58 ℃退火 30 s, 72 ℃延伸 30 s, 扩增 30 个循环; 在 72 ℃下延伸 7 min, 使扩增完全。每个 25 mL 的 PCR 反应体系中含有 0.25 mmol/L D16S539 引物两条(引物 A: 5′-GAT CCC AAG CTC TTC CTC TT-3′; 引物 B: 5′-ACG TTT GTG TGT GCA TCT GT-3′), 0.2 mmol/L dNTP, 1.5 mmol/L $MgCl_2$, 0.625 U Taq DNA 聚合酶, 2.5 mL 配套缓冲液和 2 mL DNA 模板。经过上述处理即得 PCR 扩增产物。

2.4 紫外光谱检测

取 20 mL PCR 扩增产物于 1.5 mL 离心管中, 加 500 mL 水稀释, 混匀后, 置于石英比色皿, 检测其光谱。测定参数为: 波长 200~310 nm, 分辨率 0.1 nm, 光程 10 mm, 光斑宽 2 mm, 以水为参比, 测定 PCR 产物的紫外光谱。

3 人工神经网络模型的建立

人工神经网络是一种模拟人脑神经网络行为特征, 以数学网络拓扑结构为理论基础, 进行分布式并行信息处理的算法数学模型[20]。其中, 误差反向传播算法(BP)可分为两个阶段: 正向传播和反向传播。正向为构造非线性的数学模型过程, 而反向是进行网络中权重的修正。在不断的学习和修正过程中, 使网络的学习误差达到最小。本研究采用误差反向传播算法的 ANN(BPANN)方法, 以紫外光谱为识别变量, 建立 D16S539 基因座的 10-11, 10-12, 11-11 和 11-13 基因型的分类判别模型, 实现简单、快速、低成本地分型检测 STR 的 4 种基因型。

以紫外光谱为 BPANN 的输入变量和目标变量, 利用隐含层神经元提取富信息, 并以该信息为另一个 BPANN 的输入变量, 建立基因型的分类模型, 即前一个 BPANN 为富信息提取的 ANN(RIE-ANN), 后一个则为判别模型建立的 ANN(DMB-ANN)。在 DMB-ANN 中, 对于 10-12, 11-11, 10-11 和 11-13 基因型的目标变量分别为(1,0,0,0), (0,1,0,0), (0,0,1,0)和(0,0,0,1)。对于 10-11 基因型的 33 个样本, 随机选择 24 个为校正样本, 其余 9 个为预测样本, 而其它 3 个基因型均为 34 个样本。每个基因型随机选择 10 个样本构成预测集, 其余样本构成校正集。在 210~280 nm 波长范围内, 每个样本的量测光谱共有 781 个数据点, 以"每隔一点取一点"的方法, 得 391 个数据点, 代表样本的量测光谱, 作为 RIE-ANN 输入层神经元。优化的最佳网络结构: RIE-ANN 为 391-50-391, DMB-ANN 为 50-6-4。对于建立的分类模型, 相对于基因型的目标变量, 校正集的相对均方根误差(RMS)为 0.0275, 预测集 RMS 为 0.0418, 预测准确率达到 100%。ANN 算法来自 Matlab(V6.0)工具箱, 其它计算程序在 Matlab(V6.0)环境下自行编写。

4 结果与讨论

4.1 光谱预处理

光谱的变化由基因型差异、被测样本组成和浓度产生。为了使光谱的变化主要由基因型的差异所决定, 需要消除同一基因型不同样本间的浓度差异。以 10-11 基因型的样本的紫外光谱为例(见图 1), 相同基因型样本的光谱存在明显差异, 表明样本间有浓度差异, 为了消除这种差异对以基因型差异为主的判别模型的影响, 本研究对每个样本光谱均进行了浓度归一化处理, 使得同一基因型样本的光谱差异性显著减小。同时, 在图 1 的 300 nm 以上的无信息区域, 其光谱存在明显的基线漂移(图 1 内插图), 对于同一基因型的样本光谱的这种差异也会严重影响模式识别模型对基因型差异的判别, 对其进行了基线漂移校正, 使所有样本在无信息区域的光谱均值均等于零, 有效消除了此漂移。

但是, 尚存在极小的光谱差异, 可能是因为被测样本纯度为达到 100%。虽然这种差异不可能消除, 但通过 PCR 扩增条件的优化而保证了该差异的最小化。同时, 为了使各次扩增的操作差异也包含在模型中, 对每个基因型样本平行扩增, 每次获得 18 个样本, 共获得测试样本 36 个。

4.2　BP-ANN 提取富信息

紫外光谱数据通常由极多的数据点构成, 如果将这些数据点都作为模式判别的 ANN 网络输入变量, 则神经网络的结构庞大, 太多的输入神经元易于造成过拟合, 并使迭代时间过长。因此有必要对紫外光谱数据进行降维处理, 提取富信息, 再以富信息变量建立 ANN 判别模型。本研究以 RIE-ANN 提取量测光谱的富信息, 作为 DMB-ANN 的输入变量, 建立了判别模型, 两个网络隐含层的神经元数通过试差法优化。

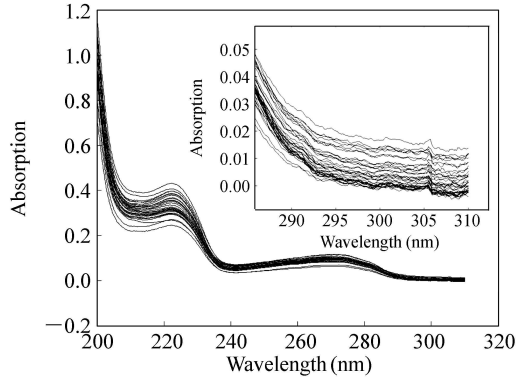

图 1　10-11 基因型 33 个样本的量测紫外光谱

Fig. 1　Measured ultraviolet spectra (UVS-s) of 33 samples for 10-11 genotype

对于每个样本的量测光谱, 由于在 210~ 280 nm 范波长围内有 781 个波长点, 数据量较大, 为了不丢失信息, 采用了"每隔一点取一点"的方法, 得到了 391 个量测数据点。若将以 391 个量测光谱数据直接用于建立基因型的判别模型, 相对于本研究使用的 24 个校正样本, 这个变量数远远大于样本数, 极易于出现过拟合。为了避免这种现象, 本研究利用 RIE-ANN, 将 391 个量测光谱数据作为输入层变量, 同时将这些变量作为目标输出变量, 即网络结构为 391-m-391, 数据量远远小于 391 的 m 个隐含层变量即为提取的富信息变量。以此富信息变量作为基因型的判别变量建立判别模型, 即在建立判别模型的 DMB-ANN 中, 采用网络结构 m-n-4, 其中隐含层神经元为 n 个。

设定 RIE-ANN 的目标误差为均方误差(MSE) 0.0005(应小于 DMB-ANN 的目标误差), 最大循环 500 次, 最小梯度 10^{-6}, 随机初始化权重。隐含层的激发函数选用 S 型函数, 输出层的激发函数为线性函数。实验表明, 隐含层神经元数目在 30~ 100 较为适宜, 分别对隐含层的神经元数为 30, 40, 50, 60, 70, 80, 90 和 100 进行了优化。由于本研究建立的基因型判别模型实际上是 RIE-ANN 与 DMB-ANN 两个 ANN 的偶合, 网络结构参数 m 和 n 即为模型参数, 可采用"预测准确率、较小的类内距和较大的类间距"为标准, 优化获取该参数。

4.3　ANN-ANN 对 STR 的 4 种基因型的判别能力

以 RIE-ANN 提取的 m 个富信息变量为建立判别模型的 DMB-ANN 的输入变量, 网络结构为 m-n-4, 其中 4 个目标输出变量为基因型的表达变量, 即 D16S539 基因座的 10-12, 11-11, 10-11 和 11-13 基因型的模式表达变量分别为(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0) 和(0, 0, 0, 1)。以分类预测率和模型稳健性为基础, 优化 DMB-ANN 网络, 并选择 RIE-ANN 的最佳隐含层神经元数。本研究选用 Levenberg-Marquardt 算法, 隐含层的激发函数选 S 型函数, 输出层的激发函数为线性函数, 建立判别模型, 其目标优化参数为: 目标误差 MSE 为 0.001、最大循环次数 500 和最小梯度 10^{-6}。对于初步试探的模型参数 m[30, 100] 和 n[2, 10], 以预测集的相对均方根误差 RMS 最小为标准, 分别以步长 Δm= 10 和 Δm= 1 进行全局优化。预测集 RMS 0.0418 为上述范围内的最小值, 此时校正集的 RMS 为 0.0275, 模型参数 m= 50, n= 6, 该模型对基因型的预测准确率达到 100%(表 1)。从表 1 可见, 判别模型对基因型的预测准确率较高, 基因型表达变量的预测范围[0.8177, 1.0000] 和[0.0000, 0.2318] 分别与目标表达变量 1 和 0 非常接近, 两者之间有明显差异, 不存在类间重叠, 表明模型能够准确判别基因型。由于预测集与校正集的 RMS 相差较小, 后者略小于前者, 预示着校正模型不存在过训练现象, 从三维表达图也证明了此现象。

5　结　论

对于 STR 不同基因型的紫外光谱的强相似性, 本研究以偶合 ANN-ANN 的网络为基础, 提取光谱富信息和建立基因分型的判别模型。以模型稳健性为前提、最小预测 RMS 对应的 ANN-ANN 即为最优网络。研究表明, 用于压缩变量的 ANN 的最佳网络结构是 391-50-391, 用于分类的 ANN 的最佳网络结构是 50-6-4, 由此建立的模型表现出了良好的稳健性和 100% 的正确判别率。实现了对 PCR 扩增

样本在不经分离提纯等预处理条件下的快速、简单、低成本的紫外光谱基因型判别。

表1 对于预测集样本 ANN 判别模型的预测结果

Table 1 Predictions of ANN discriminant model for prediction samples sets

样本 Sample	基因型 Genotype	目标输出 Target output	实际输出 Actual output				预测基因 Genotype prediction
1	10-12	1 0 0 0	0.9980	0.0017	0.0057	0.0008	10-12
2	10-12	1 0 0 0	0.9948	0.0005	0.0019	0.0052	10-12
3	10-12	1 0 0 0	0.9291	0.0370	0.0042	0.0006	10-12
4	10-12	1 0 0 0	0.9982	0.0002	0.0043	0.0044	10-12
5	10-12	1 0 0 0	0.9969	0.0003	0.0023	0.0043	10-12
6	10-12	1 0 0 0	0.9998	0.0017	0.0054	0.0003	10-12
7	10-12	1 0 0 0	0.9745	0.0004	0.0027	0.0157	10-12
8	10-12	1 0 0 0	0.9749	0.0000	0.0044	0.0124	10-12
9	10-12	1 0 0 0	0.9987	0.0069	0.0050	0.0003	10-12
10	10-12	1 0 0 0	0.9985	0.0055	0.0216	0.0001	10-12
11	11-11	0 1 0 0	0.0006	0.8177	0.2318	0.0001	11-11
12	11-11	0 1 0 0	0.0001	0.9409	0.0448	0.0004	11-11
13	11-11	0 1 0 0	0.0003	0.9970	0.0007	0.0009	11-11
14	11-11	0 1 0 0	0.0000	0.9922	0.0014	0.0017	11-11
15	11-11	0 1 0 0	0.0005	0.9926	0.0001	0.0076	11-11
16	11-11	0 1 0 0	0.0024	0.9807	0.0003	0.0026	11-11
17	11-11	0 1 0 0	0.0011	0.9911	0.0174	0.0001	11-11
19	11-11	0 1 0 0	0.0036	0.8653	0.0593	0.0001	11-11
20	11-11	0 1 0 0	0.0000	0.9754	0.0032	0.0011	11-11
21	10-11	0 0 1 0	0.0000	0.0500	0.8623	0.0007	10-11
22	10-11	0 0 1 0	0.0000	0.0865	0.9943	0.0014	10-11
23	10-11	0 0 1 0	0.0001	0.0188	0.9629	0.0001	10-11
24	10-11	0 0 1 0	0.0001	0.1735	0.9422	0.0000	10-11
25	10-11	0 0 1 0	0.0000	0.1903	0.9452	0.0021	10-11
26	10-11	0 0 1 0	0.0000	0.0377	0.9919	0.0001	10-11
27	10-11	0 0 1 0	0.0000	0.0851	0.9440	0.0114	10-11
28	10-11	0 0 1 0	0.0000	0.0275	0.9930	0.0001	10-11
29	10-11	0 0 1 0	0.0000	0.0227	0.9735	0.0017	10-11
30	10-13	0 0 0 1	0.0002	0.0000	0.0050	0.9604	10-13
31	10-13	0 0 0 1	0.0000	0.0307	0.0015	0.8696	10-13
32	10-13	0 0 0 1	0.0070	0.0000	0.0061	0.9876	10-13
33	10-13	0 0 0 1	0.0659	0.0000	0.0091	0.9909	10-13
34	10-13	0 0 0 1	0.0133	0.0000	0.0045	0.9307	10-13
35	10-13	0 0 0 1	0.0000	0.0000	0.0110	1.0000	10-13
36	10-13	0 0 0 1	0.0000	0.0000	0.0045	0.9985	10-13
37	10-13	0 0 0 1	0.0000	0.0000	0.0050	0.9893	10-13
38	10-13	0 0 0 1	0.0003	0.0000	0.0060	0.9766	10-13
39	10-13	0 0 0 1	0.0006	0.0000	0.0064	0.9878	10-13

References

1 Alford R L, Hammond H A, Coto I, Caskey C T. *Am. J. Hum. Genet.*, **1994**, 55(1): 190~195

2 Hao F, Jia Y C. *Geno. Prot. Bioinfo.*, **2007**, 5(1): 7~14

3 Lev A Z, Peter A U, Cengiz C, Manfred K, Bharti M, Toomas K, Rosaria S, Fulvio C, Giovanni D, Gabriella S, Geoffrey K C, Rene J H, Kiau K Y, David G, Ivailo T, Marcus W F, Luba K. *Am. J. Hum. Genet.*, **2004**, 74(1): 50~61

4 Asamura H, Tsukada K, Ota M, Sakai H, Takayanagi K, Kobayashi K, Saito S, Fukushima H. *Int. Congr. Ser.*, **2006**, 1288: 610~612

5 Matthias G, Tanja S Y. *Forensic. Sci. Int.*, **2000**, 113: 43~46

6 Riad A B, Lihadh I A, Uzma J, Mohamed S A, Nurekamal, A D, Abdulbari B, Valsamma E, Bruce B. *Electrophoresis*, **1997**, 18(9): 1637~1640

7 Victoria L, Carmela P, Christopher P, Francisco B, Angel C, Denise S C, Patrick L. *Electrophoresis*, **1998**, 19(10): 1566~1572

8 Stephanie H I Y, Peng L, Nadia D B, Susan A G, Richard A M. *Anal. Chem.*, **2009**, 81(1): 210~217
9 LIU Yong, WANG Rong, GAO Lan, JIA Zheng-Ping, XIN Xiao-Ting, XIE Hua, MA Jun(刘勇, 王荣, 高岚, 贾正平, 辛晓婷, 谢华, 马骏). *Chinese J. Anal. Chem.*(分析化学), **2009**, 37(10): 1494~1498
10 Stephanie H I Y, Tae S S, Cecelia A C, Susan A G, Thomas N C, Jeff D B, Richard A M. *Electrophoresis*, **2008**, 29(11): 2251~2259
11 Stephanie H I Y, Peng L, Nadia D B, Susan A G, Richard A M. *Anal. Chem.*, **2009**, 81(1): 210~217
12 Yi N S. *Electrophoresis*, **2006**, 27(19): 3703~3711
13 Wei M H, Kwang J H, Cheng Y C. *Taiwanese. J. Obstet. Gynecol.*, **2005**, 44(1): 52~56
14 Antonio B, Maria J A, Manuela A N, Zarapuz L, Blanco L, Garc a-Sl nchez F, Vicario J L. *Hum Immunol*, **2005**, 66(8): 903~911
15 ZHANG Xiao-Dan, WU Hai-Ping, CHEN Zhi-Yao, ZHOU Guo-Hua(张晓丹, 武海萍, 陈之遥, 周国华). *Chinese J. Anal. Chem.*(分析化学), **2009**, 37(8): 1107~1112
16 Herbert O, Walther P, Roswitha M, Christian G H. *Anal. Chem.*, **2001**, 73(21): 5109~5115
17 Allison P N, James C H, David C M. *Anal. Chem.*, **2001**, 73(18): 4514~4521
18 Yuan G L, Ren L, Gao Y Z, Wang W P, E X, Xie H P. *Anal. Chim. Acta.*, **2009**, 28(11): 1245~1249
19 Sweet D, Lorente M, Valenzuela A, Lorente J A, Alvarez J C. *Forensic Sci. Int.*, **1996**, 83(3): 167~177
20 PANG Tao-Tao, YAO Jian-Bin, DU Li-Ming(庞涛涛, 姚建斌, 杜黎明). *Sepctroscopy and Spectral Analysis*(光谱学与光谱分析), **2007**, 27(7): 1336~1339

Genotyping of Short Tandem Repeat Based on Ultraviolet Spectroscopy Combined with Artificial Neural Network

WANG Xue-Jiao, MOU Hong-Yuan, LU Hui, DOU Xing-Ru, QIU Ting, XIE Hong-Ping*
(College of Pharmaceutical Sciences, Soochow University, Suzhou 215123)

Abstract Taking genotypes 10-11, 11-11, 10-12 and 10-13 of short tandem repeat (STR) locus D16S539 commonly used in forensic medicine as study objects, an ANN genotyping method was developed based on ultraviolet spectra of the measured samples and another ANN was used to extract the variables of rich information. Under the optimal conditions, each of the genotypes was amplified. The ultraviolet spectra of the samples that were produced by polymerase chain reaction, which was measured at length range of 200~310 nm, were pretreated and optimized by coupled ANN-ANN. The results showed that the best network structures of the rich information extraction ANN and the discriminant model built ANN were 391-50-391 and 50-6-4, respectively. The root mean square error for the training and the prediction samples sets was obtained to be 0.0279 and 0.0418. It was indicated that the models had a good ability of the robustness and big discriminating power for the prediction samples (the accuracy was 100%). The detection of STR genotypes by UVS was rapid, simple and low-cost.

Keywords Short tandem repeat; Ultraviolet spectroscopy; Artificial neural network; Genotyping
(Received 15 January 2011; accepted 19 April 2011)