第33卷，第6期
2013年6月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 33, No. 6, pp1537-1541
June, 2013

近红外光谱结合主成分分析和BP神经网络的转基因大豆无损鉴别研究

吴 江[1]，黄富荣[1*]，黄才欢[2]，张 军[1]，陈星旦[1,3]

1. 暨南大学光电工程系，广东 广州 510632
2. 暨南大学食品科学与工程系，广东 广州 510632
3. 中国科学院长春光学精密机械与物理研究所，吉林 长春 130033

摘 要 为探究无损鉴别转基因大豆的可行性，利用近红外光谱分析仪对大豆扫描得到反射光谱，应用主成分分析结合BP神经网络方法进行分析鉴别。首先应用主成分分析法，得到包含大豆99.03%的光谱信息的6个主成分，再将其作为BP神经网络的输入，对应的大豆种类作为输出，建立一个三层BP神经网络模型。该模型对于转基因大豆的正确识别率为100%，说明近红外光谱结合主成分分析和BP神经网络的方法能无损快速准确地鉴别转基因大豆。

关键词 近红外光谱；转基因大豆；主成分分析；BP神经网络

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2013)06-1537-05

引 言

转基因大豆作为主要的转基因作物，占据全球转基因作物种植面积的47%[1]。海关总署发布的数据显示，2011年全年中国累计进口大豆5 264万t，其中大部分为转基因大豆。目前我国进口转基因大豆主要用作加工原料，例如生产豆油、豆腐、豆奶等制品。作为重要的转基因食品，转基因大豆一直受到很大的争议，其对人类健康的危害主要包括是否引起部分人过敏、转基因食品的毒性、抗生素抗性风险问题以及营养品质改变等问题[2]。

转基因产品的传统检测方法主要有两大类：DNA检测和蛋白质检测[3]，但存在较多不足：进行DNA检测前，必须要了解被检测对象的转入基因的序列信息，否则无法进行对比检测；而蛋白质检测只适合未经加工的产品，原因是由于蛋白质对温度较为敏感，在加热过程中极容易变性，导致检测失败。并且，DNA与蛋白质检测方法在提取DNA和蛋白质时，转基因成分会受到不同程度的损失和破坏，因此所获得的信息不能代表转基因产品的成份所有信息。对于转基因大豆的鉴别，需寻求一种无损、准确、绿色、快捷的鉴别方法。

现代近红外光谱分析技术是近年来将光谱测量技术、计算机技术、化学计量学技术与基础测试技术的有机结合的高新分析技术，具有高效、快速、无损、可多组分同时测量等特点。现已成功地应用于食品[4]、医药[5]、农业[6]等各科学领域。随着转基因作物的迅速发展，陆续出现了近红外光谱分析技术用于转基因作物鉴别的报道，但在转基因大豆鉴别上的应用尚未见报道。针对我国目前大量进口转基因大豆的现状，开展转基因大豆无损鉴别具有重要的意义。

本工作采用近红外光谱结合主成分分析与BP神经网络方法探讨无损鉴别转基因大豆的可行性。

1 实验部分

1.1 仪器设备

光谱采用丹麦FOSS公司的XDS光栅型近红外光谱分析仪的反射附件采集，探测器为Si(400～1 100 nm)和PbS(1 100～2 500 nm)。光谱扫描范围为400～2 500 nm，光谱带宽8 nm。光谱仪开机半小时后进行仪器性能测试，性能测试通过后，开始对样品采集光谱。

1.2 样品

黑龙江非转基因大豆1号(以下简称H01)来自网购，产地黑龙江七台河；黑龙江非转基因大豆2号(以下简称H02)来自广州天河天娱广场百佳超市，产地黑龙江；北京非转基

收稿日期：2012-07-05，修订日期：2012-10-07

基金项目：高等学校博士学科点专项科研基金项目(20124401120005)，广东省自然科学基金项目(S2011040001850)，广东高校优秀青年创新人才培养计划项目(LYM11026)和中央高校基本科研业务费专项资金项目(21612436，21612353)资助

作者简介：吴 江，1990年生，暨南大学光电工程系本科生

*通讯联系人 e-mail：furong_huang@163.com

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

因大豆 1 号(以下简称 B01)来自广州天河岗顶好又多超市，产地北京；转基因大豆 1 号(以下简称 Z01)由暨南大学理工学院食品系提供。以上 4 种大豆经过蛋白质检测法检验，证实前 3 种为非转基因大豆，第 4 种为转基因大豆。

1.3 光谱采集

4 种大豆分别取样 500 g，经 40 ℃烘干 48 h。光谱采集在装有空调的恒温(23±2)℃实验室进行，对样品的近红外光谱在全光谱范围内(400～2 500 nm)进行采集。样品统一盛放在统一尺寸的圆形样品池中，每次扫描 3 次计算平均值，记为 1 个原始光谱。H01，H02 和 B01 分别采集 20 个光谱，Z01 采集 30 个光谱。选取 400～2 500 nm 光谱区作为鉴别转基因大豆的波长范围，每隔 2 nm 记录一个点，4 种大豆分别得到 1 050 个吸光度值。

1.4 主成分分析

主成分分析方法[7]是一种重要的多元统计分析方法，通过研究指标体系的内在结构关系，把多指标转化为少数几个互相独立并且包含原来指标大部分信息的综合指标。利用方差最大原则，对原始光谱数据所包含的多个自变量进行线性拟合，以新的低维变量代替原始高维变量，进而达到数据降维的目的[8]。在实际应用中一般取前几个主成分，基本包含了全部测量指标所具有信息的百分率，可使高维空间的数据降到低维，易于数据的观察，一般推荐 T 值应高于 80%(本文中取高于 99%的 T 值)。一方面，用少于原有指标个数的不相关主成分因子来代替原有指标，可以减少指标提供信息的交叉和冗余；另一方面，根据各项指标自身数据相关性与变异来确定权重，基于数据分析得到指标之间的内在结果关系，不需要人为来确定，使检测结果更加合理化。

1.5 BP 神经网络

人工神经网络(artificial neural network，ANN)是通过人工建立的以有向图组成拓扑结构的动态系统[9]。通过输入连续或离散的初始信息，进行计算后按误差逆传播(back propagation，BP)算法[10]的前馈神经网络，最终实现输出和输入之间的高度的非线性映射。

人工神经网络由大量处理单元互联组成，每一个神经元既是信息存储单元，又是处理单元，信息处理的能力分布在各处理单元上[11]。每个神经元的结构和功能比较简单，但数量大的神经元组合产生的系统行为却非常复杂，使其具有自适应、自组织、自学习的能力。

图 1 是 BP 神经网络的结构示意图。BP 算法是由信号的正向传播与误差的反向传播两个过程组成[12]：第一阶段(正向传播过程)输入样本从输入层传入，经各隐层逐层处理后，计算出各单元的实际输出值，传向输出层；第二阶段(反向过程)若输出层的实际输出与期望的输出不符，则转入误差的反向传播阶段，逐层计算实际输出与要求输出之差值(即误差)，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据，在不断学习和修正过程中，使网络的学习误差达到最小。

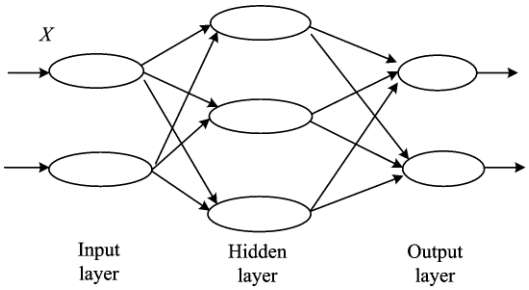

Fig. 1 The architecture of three layer back propagation artificial neural network

2 结果与讨论

2.1 近红外光谱图

图 2 为 90 个大豆的近红外光谱图，由图可以看出，大豆的近红外光谱形状基本相似，无法从原始光谱直接区分出转基因和非转基因大豆。

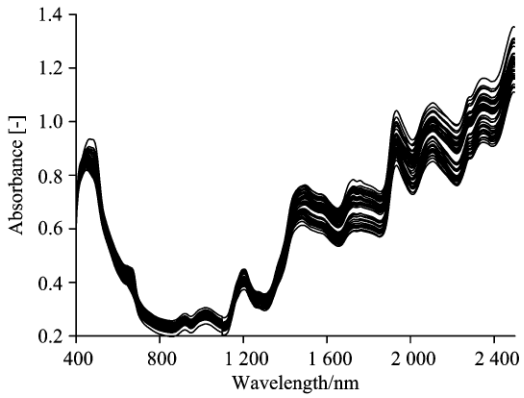

Fig. 2 Near-infrared spectra of soybean samples

2.2 数据预处理

采集的光谱受到高频随机噪声、基线漂移、光散射、样品不均匀等影响[13]，仪器扫描所得到的原始近红外光谱无法直接用于计算。因此，采用矢量归一化法对原始光谱进行预处理，以消除噪声。

具体计算步骤为：(1)对一条原始光谱，计算其平均吸光度值；(2)用原始光谱值减去平均吸光度值，得到处理后的光谱值；(3)计算处理后的光谱值的平方和，再开平方根，得到的值设为 m；(4)将处理后的光谱值除以 m，完成一条光谱的矢量归一化。

2.3 主成分分析

经过数据预处理后，大豆光谱在全谱区内仍然有 1 050 个数据点，数据量还是很大，并且冗余信息也较多。当采用全谱区计算时，不仅计算量大，而且在某些光谱区域样品的光谱信息很弱，与样品组成和性质缺乏相关关系，引入这样的波长会造成模型精度降低甚至错误。所以经过预处理后，再对 90 个样品的光谱数据进行主成分分析。

图 3 是四种大豆的第 1 和第 2 主成分的得分图。图中横坐标表示每个样本的第 1 主成分的得分值，纵坐标表示每个样本第 2 主成分的得分值。

从图 3 看出，Z01(转基因大豆)聚类效果较好，主要分布在横坐标的负半轴区域，但存在 1 个误判。而 H01 主要分

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

布在横坐标的正半轴，H02 和 B01 的分布比较靠近，较难区别，但都远离 Z01。

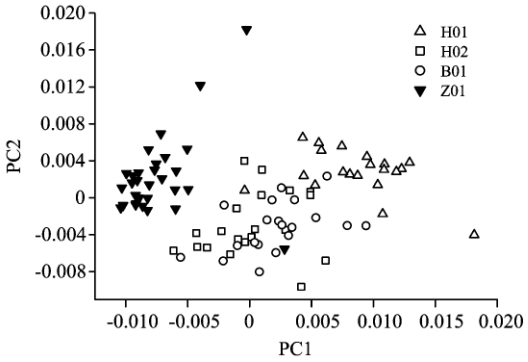

Fig. 3　Two-dimension map of PC1 *vs*. PC2 of all samples

由此可见，主成分分析方法对转基因大豆和非转基因大豆的区分是比较明显的，但存在误判，并且对于不同地区的非转基因大豆较难准确地区分，还需要建立大豆区分的定量分析模型。

2.4　BP 神经网络

表 1 是经过计算后包含绝大部分大豆光谱信息的前 6 个主成分的累积贡献率。表中数据显示，前 6 个主成分的累计贡献率已超过 99%，所以选取前 6 个主成分就可以表示原始光谱的主要信息，这样数据阵就从 90×1 050 减少到 90×6。从而大大降低运算量。

Table 1　PCs and reliabilities

主成分	PC1	PC2	PC3	PC4	PC5	PC6
累计贡献率/%	68.68	86.89	93.60	96.42	97.91	99.03

把样本分为训练集和测试集，按照 3∶1 的比例从每种非转基因大豆样中随机选取 15 个样本用来建立模型，5 个样本用来预测结果，转基因大豆样中随机选取 21 个样本用来建立模型，7 个样本用来预测结果。

在主成分分析基础上，把前 6 个主成分作为 BP 的输入节点，输出节点为 2 即大豆的品种值(将 H01，H02，B01 和 Z01 种类值分别设为 00，01，10 和 11)。通过多次建模实验，发现隐含层节点取 8 时，建立模型输出检测结果效果达到最好。故建立一个 6(输入节点)－8(隐含层节点)－2(输出节点)的三层 ANN-BP 模型，训练 1 000 次，误差目标 1e-5。以预测结果偏差±0.2 设定为大豆区分的界限。

检测结果显示，66 个大豆建模样本的学习结果输出鉴别正确率达到 100%，具体结果见表 2。

模型对 22 个预测样本品种鉴别率为 100%，如表 3 所示。结果表明，本方法所建立的 BP 神经网络模型能够准确地鉴别转基因大豆。

Table 2　The analytical results of soybean calibration samples

序号	真实值		预测值		结果	序号	真实值		预测值		结果
1	0	0	0.002 5	0.000 7	正确	34	1	0	1.000 6	−0.000 5	正确
2	0	0	−0.004 4	0.001 0	正确	35	1	0	1.002 5	−0.002 4	正确
3	0	0	−0.000 2	0.000 4	正确	36	1	0	0.996 3	−0.000 3	正确
4	0	0	0.000 1	0.000 4	正确	37	1	0	1.000 9	0.005 5	正确
5	0	0	0.000 3	−0.002 7	正确	38	1	0	1.000 7	−0.000 5	正确
6	0	0	−0.000 2	−0.000 3	正确	39	1	0	1.000 5	−0.0004	正确
7	0	0	0.001 2	−0.000 5	正确	40	1	0	1.002 0	0.002 8	正确
8	0	0	0.000 1	0.002 7	正确	41	1	0	1.000 9	0.000 1	正确
9	0	0	0.000 4	0.000 5	正确	42	1	0	0.999 7	0.001 1	正确
10	0	0	0.000 2	−0.001 2	正确	43	1	0	0.996 8	−0.005 4	正确
11	0	0	0.000 5	0.000 8	正确	44	1	0	1.000 0	0.000 5	正确
12	0	0	0.000 0	−0.000 4	正确	45	1	0	0.998 7	−0.000 1	正确
13	0	0	0.000 3	−0.000 5	正确	46	1	1	0.999 6	0.996 5	正确
14	0	0	−0.000 1	0.000 4	正确	47	1	1	0.999 4	1.001 1	正确
15	0	0	0.000 6	−0.002 1	正确	48	1	1	0.999 8	1.000 0	正确
16	0	1	0.002 1	0.998 9	正确	49	1	1	0.995 8	1.000 1	正确
17	0	1	−0.001 3	1.000 4	正确	50	1	1	1.000 0	0.999 7	正确
18	0	1	−0.000 2	1.002 1	正确	51	1	1	0.998 5	0.999 8	正确
19	0	1	−0.000 3	0.999 7	正确	52	1	1	1.000 4	1.000 2	正确
20	0	1	0.000 1	0.998 9	正确	53	1	1	1.000 7	1.000 3	正确
21	0	1	0.000 5	0.999 3	正确	54	1	1	1.001 2	1.000 2	正确
22	0	1	0.001 8	1.001 3	正确	55	1	1	1.000 6	1.000 3	正确
23	0	1	0.001 5	0.999 9	正确	56	1	1	1.002 0	1.000 4	正确
24	0	1	0.000 1	1.000 0	正确	57	1	1	1.000 6	1.000 2	正确
25	0	1	0.001 7	0.998 0	正确	58	1	1	1.000 4	1.000 1	正确
26	0	1	−0.002 8	1.001 9	正确	59	1	1	1.000 5	0.999 9	正确
27	0	1	0.000 3	0.999 0	正确	60	1	1	0.999 1	1.000 3	正确

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

续表 2

28	0	1	0.000 0	0.998 9	正确	61	1	1	0.999 9	0.999 6	正确
29	0	1	−0.002 2	1.000 8	正确	62	1	1	1.000 5	1.000 1	正确
30	0	1	−0.000 6	0.999 0	正确	63	1	1	0.999 7	1.005 1	正确
31	1	0	0.998 6	0.000 9	正确	64	1	1	1.000 7	0.998 1	正确
32	1	0	1.000 8	−0.000 3	正确	65	1	1	0.999 9	0.997 1	正确
33	1	0	1.000 8	−0.000 7	正确	66	1	1	1.000 0	1.000 6	正确

Table 3 The analytical results of soybean prediction samples

序号	真实值		预测值		结果	序号	真实值		预测值		结果
1	0	0	0.001 2	−0.000 1	正确	12	1	0	0.999 5	0.000 1	正确
2	0	0	0.000 0	0.012 4	正确	13	1	0	0.999 6	0.000 1	正确
3	0	0	0.000 1	−0.000 1	正确	14	1	0	1.030 6	−0.030 6	正确
4	0	0	−0.000 1	0.000 8	正确	15	1	0	0.999 5	0.000 2	正确
5	0	0	0.060 6	0.005 6	正确	16	1	1	0.999 6	0.999 9	正确
6	0	1	0.056 6	1.042 6	正确	17	1	1	1.000 5	1.000 1	正确
7	0	1	−0.000 6	1.000 1	正确	18	1	1	1.000 1	0.999 9	正确
8	0	1	0.019 9	1.000 1	正确	19	1	1	0.930 5	1.069 0	正确
9	0	1	−0.001 7	1.002 5	正确	20	1	1	1.000 0	1.000 0	正确
10	0	1	−0.000 4	1.000 1	正确	21	1	1	1.000 2	1.000 0	正确
11	1	0	0.999 7	0.000 1	正确	22	1	1	1.000 7	1.008 6	正确

3 结 论

利用近红外光谱仪对大豆进行扫描，得到反射光谱数据，结合主成分分析和 BP 神经网络建立了大豆鉴别模型，对转基因大豆进行了快速鉴别。

首先大豆通过直接的无损近红外光谱检测，得到反射光谱数据，其近红外光谱信息为高维空间的复杂数据，单一的模式识别模型难以进行大豆品种的鉴别。将光谱数据预处理后，再利用主成分分析法对数据进行降维，这样就可以降低各类噪声的干扰，得到低维空间的大豆光谱信息。通过 BP 神经网络进一步分析，把前 6 个主成分作为模型的输入，对 22 个未知样本的识别率达到 100%。

综上所述，近红外光谱结合主成分分析和 BP 神经网络的方法具有很好的分类和鉴别作用，为转基因大豆的鉴别提供了一种新的途径，有良好的应用前景。

References

[1] Clive James. Chinese Journal of Biomedical Engineering(中国生物工程杂志), 2012, 32(1): 3.

[2] YE Ke-ping, ZHOU Guang-hong, XU Xing-lian, et al(叶可萍, 周光宏, 徐幸莲, 等). Food Science(食品科学), 2010, 31(13): 312.

[3] LUO A-dong, JIAO Yan-chao, CAO Yun-heng, et al(罗阿东, 焦彦朝, 曹云恒, 等). Journal of Southern Agriculture(南方农业学报), 2012, 43(3): 290.

[4] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄, 赵龙莲, 韩东海, 等). Foundation and Application of Near-infrared Spectroscopy Analysis(近红外光谱分析基础与应用). Beijing: Light Industry Press of China(北京: 中国轻工业出版社), 2005. 101.

[5] ZHONG Jian-li, RAO Wei-wen, ZHANG Zhi-jun(钟建理, 饶伟文, 张治军). China Pharmaceuticals(中国药业), 2009, 18(4): 32.

[6] HUANG Fu-rong, PAN Tao, ZHANG Gan-lin, et al(黄富荣, 潘 涛, 张甘霖, 等). Optics and Precision Engineering(光学精密工程), 2010, 18(3): 586.

[7] LI Liang, WANG Lei-ming, DING Wu(李 亮, 王雷鸣, 丁 武). Foodstuff Industry(食品工业), 2009, 6: 67.

[8] LIU Wei, CHANG Qing-rui, GUO Man, et al(刘 炜, 常庆瑞, 郭 曼, 等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2011, 30(1): 48.

[9] XIONG Xing-chuang, FANG Xiang, OUYANG Zheng, et al(熊行创, 方 向, 欧阳证, 等). Chinese Journal of Analytical Chemistry(分析化学), 2012, 1(4): 43.

[10] FAN Yuan-yuan, SANG Ying-jun, SHEN Xiang-heng(范媛媛, 桑英军, 沈湘衡). Journal of Applied Optics(应用光学), 2011, 6: 1150.

[11] MATLAB Chinese Forum(MATLAB 中文论坛). MATLAB Neural Network of 30 Cases Analysis(MATLAB 神经网络 30 个案例分析). Beijing: Beijing University of Aeronautics and Astronautics Press(北京: 北京航空航天大学出版社), 2010. 1.

[12] JIAO Shu-fei, XIANG Yu-hong, HUANG An-min, et al(焦淑菲, 相玉红, 黄安民, 等). Journal of Capital Normal University · Natural Sciences Edition(首都师范大学学报 · 自然科学版), 2010, 31(1): 30.

[13] HUANG Fu-rong, LUO Yun-han, ZHENG Shi-fu, et al(黄富荣, 罗云瀚, 郑仕富, 等). Acta Optica Sinica(光学学报), 2011, 31(10): 3001.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Study on Near Infrared Spectroscopy of Transgenic Soybean Identification Based on Principal Component Analysis and Neural Network

WU Jiang[1], HUANG Fu-rong[1*], HUANG Cai-huan[2], ZHANG Jun[1], CHEN Xing-dan[1, 3]

1. Opto-Electronic Department of Jinan University, Guangzhou 510632, China
2. Department of Food Science and Engineering of Jinan University, Guangzhou 510632, China
3. Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

Abstract In order to explore a rapid identification method for transgenic soybeans, non-transgenic and transgenic soybeans were tested as the experimental samples via near infrared spectroscopy (NIR) and principal component analysis (PCA) combined with back propagation artificial neural network (BP-ANN) model. The spectrum data was collected after NIRS scanning the samples, and then analyzed by PCA plus BP-ANN model. The accumulative reliabilities of the six components were 99.03% through the PCA. Then BP-ANN model was used to further test these six components and a three-layer BP-ANN model was developed. The final result achieved a 100% recognition rate of all 22 test samples respectively. In conclusion, the measure of NIRS and PCA combined with BP-ANN model has proved to be a rapid and accurate method to detect transgenic soybean nondestructively.

Keywords NIR; Transgenic soybean; PCA; ANN-BP

(Received Jul. 5, 2012; accepted Oct. 7, 2012)

* Corresponding author

关于《光谱学与光谱分析》收取审稿费的通知

尊敬的《光谱学与光谱分析》广大作者、读者同志们，本刊自 2006 年底采用由“北京玛格泰克科技发展有限公司”开发的投稿系统实现网络采编以来，进一步扩展了审稿专家队伍。本刊参考同类期刊的现行做法，决定自 2010 年 12 月 1 日以后登记的稿件向投稿作者收取审稿费 100 元/篇，在您投稿之前，为免受经济损失，请您必须考虑：

1. 没有创新的一般性稿件，请您不要投稿。
2. 没有国家级基金资助的稿件，请您不要投稿。
3. 不是光谱专业的稿件，请您不要投稿。
4. 与其他文章重合率超过 10%的稿件，请您不要投稿。

作者在投稿后，将会收到缴纳审稿费的通知。请作者及时从我刊网站(http://www.gpxygpfx.com)查询稿件是否处于交审稿费状态，在收到通知后，请及时缴纳审稿费；如在 10 天之内没有收到您的审稿费，被视为自动放弃，本刊不再受理。汇款时，请写明详细通信地址、邮政编码、收件人姓名等信息，以便准确寄回发票。

汇款方式(在附言里写明审稿费)：

邮局汇款：北京市海淀区学院南路 76 号，《光谱学与光谱分析》期刊社(收)

邮政编码：100081 联系电话：010-62181070，62182998

电子邮箱：chngpxygpfx@vip.sina.com

感谢您多年来对《光谱学与光谱分析》的支持和厚爱！

《光谱学与光谱分析》期刊社

2010 年 12 月 1 日

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net