玉米品种近红外光谱的DFT特征分析方法

李阳鹏，李卫军*，来疆亮

中国科学院半导体研究所，北京 100083

摘 要 提出了一种基于离散傅里叶变换(discrete Fourier transform，DFT)的玉米品种特征分析新方法。实验数据为37个玉米品种种子的近红外漫反射光谱数据，波段范围为4 000～12 000 cm^{-1}。文中通过对原始数据进行分析，发现扫描频率较高的部分噪声也比较大。文中首先定义了一种类间、类内差异度Q_m的计算方法，以度量特征选择的有效性；然后利用Q_m对原始数据和DFT变换后的数据进行分段分析。实验结果表明，4 000～7 085 cm^{-1}波段的DFT数据相对于全波段原始数据，Q_m曲线均值、峰值明显提高。均值从原始的4.804 9提高到8.513 8，峰值最大值从原始的35.924 0提高到60.821 6，峰值最小值从原始的2.891 8提高到3.741 5。且变换后数据特征点(即Q_m值大的点)较原始数据集中，最有利于提取玉米品种特征。

关键词 近红外光谱；玉米种子；离散傅里叶变换；差异度；特征分析

中图分类号：S132 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)01-0119-04

引 言

玉米是当前农业生产中最重要的农作物之一，目前农业生产中玉米种子品种鉴别是一个迫切需要解决的重要问题。近红外光是介于可见光区和中红外光区间的电磁波，其谱区波长范围为780～2 500 nm[1]。通过近红外光谱检验，可以提取出种子样品中所有有机分子氢基团的特征信息[2-4]。当今应用近红外光谱检测农产品的技术已经相当成熟[5-7]，应用在玉米种子检测领域也有很多相关研究。目前常用的玉米种子光谱数据预处理方法有一阶导数、矢量归一化、窗口平滑等处理方法[8-13]。

本文提出了一种基于离散傅里叶变换(discrete Fourier transform，DFT)的玉米品种近红外光谱特征分析的新方法。针对37个玉米品种的籽粒漫反射光谱数据，首先分段对原始数据进行DFT预处理，然后分别对预处理数据进行特征分析，并与原始数据对比，发现选取前800点数据做变换后，其特征点特征值相对原始数据有显著提高，且特征点位置较原始数据集中。

1 玉米种子近红外光谱数据分析

本文所用玉米品种近红外光谱数据由国家科技支撑计划高产优质玉米项目组提供，包含37个玉米品种，每个品种包含25组近红外光谱测量数据，每组数据包含2 075个给定波数下的吸光度数值。

图1是37种种子种内25组数据求均值后的曲线组(共37组)。

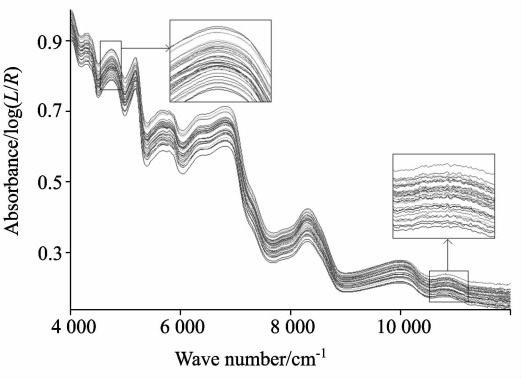

Fig. 1 Mean curve for 37 sets of data

(1) 从总体来看，37组种子数据曲线起伏趋势相近，不同品种之间只存在细微的差别，故直接提取特征点有很大的困难；

(2) 从细节来看，当扫描频率较低时，曲线数据的相对比较光滑；而当扫描频率较高时，曲线数据相对比较粗糙。

收稿日期：2010-02-22，**修订日期**：2010-07-08

基金项目：国家自然科学基金项目(90920013，60753001)资助

作者简介：李阳鹏，1987年生，中国科学院半导体研究所硕士研究生 e-mail：fog.2000@yahoo.com.cn

*通讯联系人 e-mail：wjli@semi.ac.cn

故判断随着扫描频率的升高，数据噪声也随之升高。后文将分段对玉米光谱数据进行分析。

以上分析表明，仪器分析得到的原始近红外光谱数据无法直接用于样品的定性分析。因此，有必要对原始光谱数据进行有效的预处理。

本文选用基于离散傅里叶变换(DFT)的预处理方法，它具有以下特点：(1)离散傅里叶变换是原始数据的频谱在[0，2π]上的 N 点等间隔采样，也就是对序列频谱的离散化，通过 DFT 变换可以很好的分离原始数据各频谱信息；(2)极大地保留原始数据信息，可以通过傅里叶反变换得到原始数据；(3)DFT 变化后的幅值为偶函数，相角为奇函数。目前通常认为数据信息主要集中在幅值部分，相角部分信息研究尚未有突破性进展，故通常只选取幅值部分信息近似代表原始数据。如果只取 DFT 变化的幅值部分，可节省近一半的数据量。

2 差异度计算

定义特征点为种类差异度小而种间差异度大的点，并定义特征点值由 Qm 表示。定义计算公式如下

$$Qm_i^k = \frac{\sum_{j=1}^{37}(ave_i^k - ave_j^k)^2}{\sum_{l=1}^{25}(data_{il}^k - ave_i^k)^2} \times \frac{25}{36} \tag{1}$$

$$ave_i^k = \frac{\sum_{j=1}^{25} data_{ij}^k}{25} \tag{2}$$

其中，Qm_i^k 为 37 种玉米品种中的第 i 种中的第 k 个点的 Qm 值；ave_i^k 为第 i 种玉米第 k 个点 25 个样本的均值；$data_{ij}^k$ 为第 i 种玉米第 j 次采样样本第 k 点的数值。

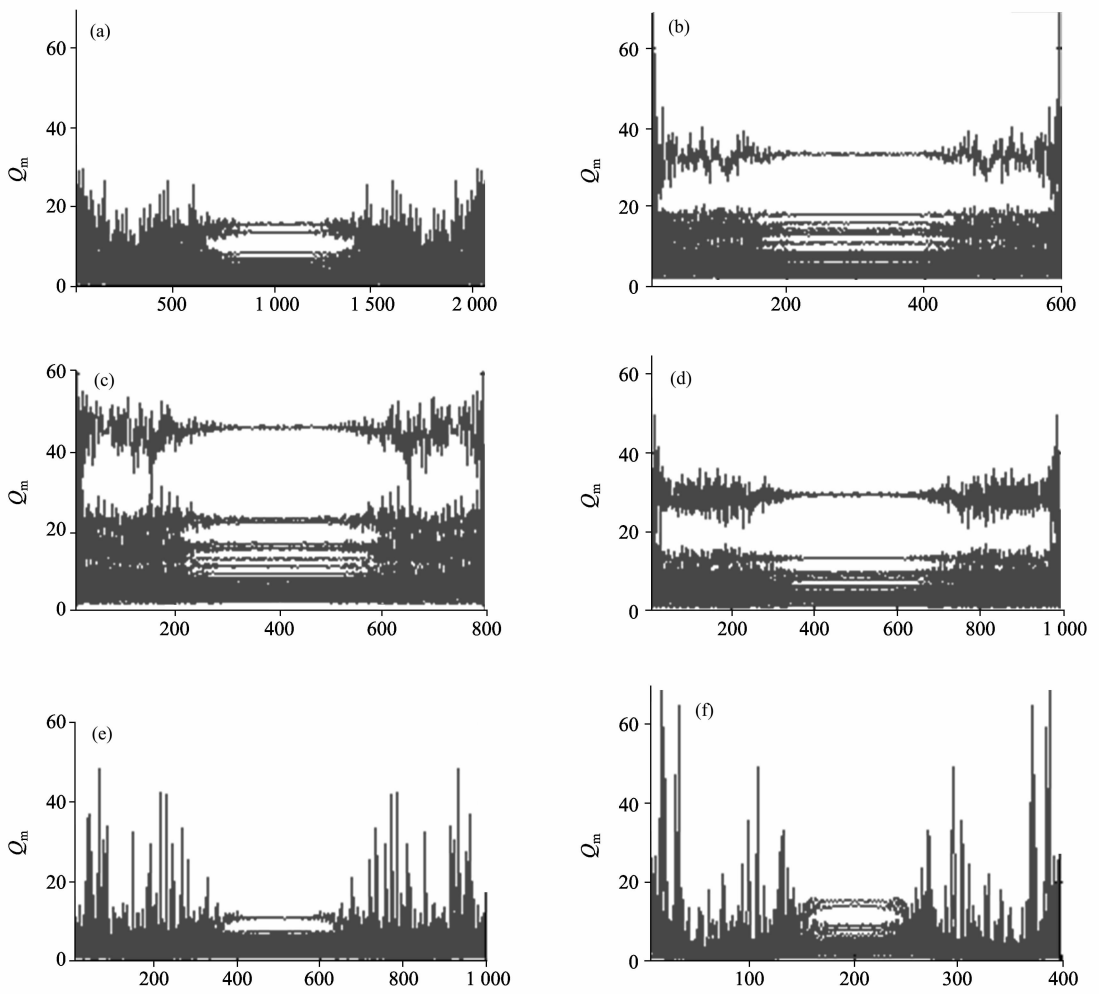

Fig. 2 Origin(curves set of (a) segment); Previous 600(curves set of (b) segment); Previous 800(curves set of (c) segment); Previous 1000(curves set of (d) segment); Last 1000(curves set of (e) segment); Last 400(curves set of (f) segment); *Q*m curves after DFT

从计算公式中可以看出，当 Qm 越大时种间差异度越大，种内差异度越小；当 Qm 越小时种间差异度越小，种内差异度越大。

3　分段处理原始数据

前文已经提到，随着扫描频率的升高，数据噪声也随之升高，故选取前面部分数据可能更有利于提取不同品种的特征。据此推测，实验中对原始数据进行分段 DFT，并利用上文定义的差异度 *Q*m 对其进行比较分析。

分别提取原始数据的前 600 点(波段范围 4 000～6 310 cm^{-1})、前 800 点(波段范围 4 000～7 085 cm^{-1})、前 1 000 点(波段范围 4 000～7 853 cm^{-1})、后 1 000 点(波段范围 8 146～12 000 cm^{-1})、后 400 点(波段范围 10 460 到 12 000 cm^{-1})数据做 DFT 变换，并求出相应的 *Q*m 曲线，如图 2 所示。

从图中可以看出，提取前 800 点数据作为特征数据，相对其他部分数据具有以下优点：

(1) 前 800 点数据选取的数据范围是扫描频率较低的部分，即从原始的 4 000～12 000 cm^{-1}波段数据中选取 4 000～7 085 cm^{-1}波段数据作为特征数据。由于分析得出扫描频率越高时噪声越大，故尽量选取前面部分的数据可以将噪声的影响降低；

(2) 数据量适中，在体现原始数据特征的前提下，尽可能多的减轻了数据量过大所引发的数据处理负担。由于原始数据一共有 37 种种子，而每一种种子又有 25 组数据，每一组数据为 2075 个 double 型数据，选取前 800 点数据可节省近 2/3 的运算量；

(3) 前 800 点数据的 *Q*m 曲线数值从统计意义上来说最好，从表 1 可以看出，前 800 数据 *Q*m 曲线整体均值最大，峰值最大值、最小值较大，有利于数据特征点的提取。

Table 1　*Q*m values comparison of data in different spectral region

	37 组曲线整体均值	37 组曲线峰值中的最大值	37 组曲线峰值中的最小值
DFT 前 800	8.513 8	60.821 6	3.741 5
DFT 原始	3.039 9	29.334 4	2.663 1
DFT 前 600	7.484 0	69.609 8	2.0245
DFT 前 1000	5.653 5	49.752 4	2.126 4
DFT 后 1000	3.043 6	48.468 8	3.155 1
DFT 后 400	2.915 0	69.060 8	3.765 6
原始数据	4.804 9	35.924 0	2.891 8

4　前 800 点 DFT 数据与原始数据 *Q*m 曲线对比

从图 3 和图 4 可以看出，选取前 800 点数据做 DFT 处理后，计算得出的 *Q*m 曲线明显优于原始数据 *Q*m 曲线，具体体现在如下几个方面：

(1) 前 800 点数据 DFT 处理后的 *Q*m 曲线均值、峰值都明显大于原始数据，见表 1；

(2) DFT 处理后差异度大的点较原始数据集中。从图 3 可以看出，图中曲线两侧的点数值比较大，而曲线中间部分的点相对数值较小(个别曲线中间和两边数值都大)。可以直接提取前部分某些点作为特征点，而忽略中间差异度不大的点，有利于后续特征点提取。而图 4 曲线差异度大的点分布几乎毫无规律，不利于特征点的提取；

(3) 由于 DFT 变换后数据幅值关于数据中轴对称，相应的 *Q*m 的曲线也关于数据中轴对称，故处理起来只需要考虑前一半数据即可，减少了近一半的数据运算量和存储量。

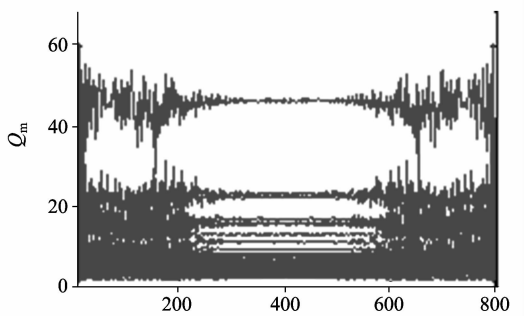

Fig. 3　*Q*m curves of previous 800 after DFT

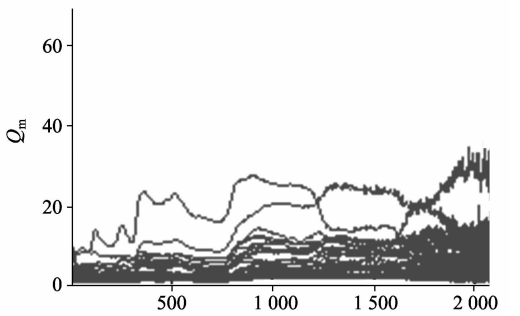

Fig. 4　Qm curves of the origin

5　总　结

本文提出了一种基于离散傅里叶变换(DFT)的玉米品种特征分析新方法。实验数据为 37 个玉米品种种子的近红外漫反射光谱数据，波段范围为 4 000～12 000 cm^{-1}。本文通过对原始数据进行分析，发现扫描频率较高的部分噪声也比较大，随着扫描频率的升高，数据噪声也随之增大。

文中首先定义了一种类间、类内差异度 *Q*m 的计算方法，以度量特征选择的有效性，当 *Q*m 值越大时种间差异度越大，种内差异度越小；当 *Q*m 值越小时种间差异度越小，种内差异度越大。然后，选取六个不同波段范围的数据，分别为 4 000～12 000，4 000～6 310，4 000～7 085，4 000～7 853，8 146～12 000 和 10 460～12 000 cm^{-1}，分别利用 *Q*m 对原始数据和各段数据 DFT 变换后的数据进行分析。

实验结果表明，4 000～7 085 cm^{-1}波段的 DFT 数据相对于原始数据，*Q*m 曲线均值、峰值明显提高，最有利于提取玉米品种特征。均值从原始的 4.804 9 提高到 8.513 8，峰值最大值从原始的 35.924 0 提高到 60.821 6，峰值最小值从原始的 2.891 8 提高到 3.741 5。选取 4 000～6 310 cm^{-1}(前

600 点数据)，4 000～7 853 cm^{-1}(前 1 000 点数据)波段的 DFT 数据 *Q*m 曲线均值分别为 7.484 0 和 5.653 5，相对于原始数据的 4.804 9 都有所提高。选取4 000～12 000 cm^{-1}(原始全波段)，8 146～12 000 cm^{-1}(后 1 000 点数据)，10 460～12 000 cm^{-1}(后 400 点数据)波段的 DFT 数据 *Q*m 曲线均值分别为 3.039 9，3.043 6 和 2.915 0，相对于原始数据的 4.804 9 都有所下降。进一步证实了扫描频率较高的部分噪声也比较大，数据噪声不利于数据特征的提取。

4 000～7 085 cm^{-1}波段的 DFT 数据相对于原始数据数据特征点(即 *Q*m 值大的点)更为集中，易于提取特征数据点。此外，由于 DFT 变换数据关于数据中轴对称，相应计算得出的 *Q*m 数据也关于数据中轴对称，故处理时提取前半部分数据即可。这样不但能够提高处理速度，同时还可以节省数据存储空间。

致谢：感谢中国农业大学信息与电气工程学院安冬副教授、郭婷婷博士的指导。

References

[1] LI Xiao-li, TANG Yue-ming, HE Yong, YING Xia-fang(李晓丽，唐月明，何 勇，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2008, 28(3): 578.
[2] 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: China Light Industry Press(北京：中国轻工业出版社), 2005.
[3] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍，袁洪福，徐广通，等). Modern Near Infrared Spectroscopy Analytical Technology(Second Edition)(现代近红外光谱分析技术，第 2 版). Beijing: China Petrochemical Press(北京：中国石化出版社), 2007.
[4] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立，袁洪福，陆婉珍). Progress in Chemistry(化学进展), 2004, 16(4): 528.
[5] WANG Tie-gu, LIU Xin-xiang, KU Li-xia, et al(王铁固，刘新香，库丽霞，等). Journal of Maize Sciences(玉米科学), 2008, 16(3): 57.
[6] FANG Li-min, LIN Min(方利民，林 敏). Chinese Journal of Analytical Chemistry(分析化学), 2008, 36(6): 815.
[7] Baye Tesfaye M, Pearson Tom C, Settles A Mark. Journal of Cereal Science, 2006, 43: 236.
[8] LI Shang-yu, CHEN Yang, WANG Chun-yan, et al(李尚禹，陈 阳，王春艳，等). Journal of Molecular Science(分子科学学报), 2007, 23(3): 220.
[9] DING Nian-ya, LI Wei, FENG Xin-wei, et al(丁念亚，黎 薇，冯昕韡，等). Computers and Applied Chemistry(计算机与应用化学), 2008, 25(4): 499.
[10] ZHAO Jie-wen, HU Huai-ping, ZOU Xiao-bo(赵杰文，呼怀平，邹小波). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2007, 23(4): 149.
[11] HAO Yong, CAI Wen-sheng, SHAO Xue-guang(郝 勇，蔡文生，邵学广). Chemical Journal of Chinese Universities(高等学校化学学报), 2009, 30: 28.
[12] ZHANG Hui, eu al(张 卉，等). Chinese Journal of Spectroscopy Laboratory(光谱实验室), 2007, 24(3): 380.
[13] HAN Liang-liang, MAO Pei-sheng, WANG Xin-guo, et al(韩亮亮，毛培胜，王新国，等). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2008, 27(2): 86.

DFT Feature Analysis of Corn Varieties Based on Near Infrared Spectra

LI Yang-peng, LI Wei-jun*, LAI Jiang-liang
Institute of Semiconductor, Chinese Academy of Sciences, Beijing 100083, China

Abstract The present paper develops a new approach to the analyse of corn based on discrete Fourier transform (DFT). The experiment data is of 37 varieties of corn seed with the Fourier transform near infrared spectrometer in the wave number range from 4 000 to 12 000 cm^{-1}. Analyse of the origin data found that as the wave number increases, the data noise also increases. Firstly, the paper defines a calculation method of interspecific and intraspecific differences *Q*m to measure the effectiveness of feature selection. Secondly, *Q*m was used to analyse the original data and DFT-section data. Experimental results show that by choosing data of DFT with wave number range from 4 000 to 7 085 cm^{-1}, the mean value and the peak value of the the *Q*m curve markedly improved relative to the full band original data. The mean value was enhanced from the original 4.804 9 to 8.513 8, and the max of the peak value was enhanced from the original 35.924 0 to 60.821 6, while the min of the peak value was enhanced from the original 2.891 8 to 3.741 5. Data feature points (*Q*m value of large point) are more concentrated than the original data after DFT. Such a result is most conducive to extracting the characteristics of corn seed.

Keywords Near infrared spectra (NIRS); Corn seed; Discrete fourier Transform (DFT); Difference degree; Character analyse

* Corresponding author

(Received Feb. 22, 2010; accepted Jul. 8, 2010)