BIOSYSTEMS ENGINEERING 172 (2018) 144–153

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Non-destructive identification of maize haploid seeds using nonlinear analysis method based on their near-infrared spectra

Lina Yu [a,c], *Wenjie Liu* [a,c], *Weijun Li* [a,c,*], *Hong Qin* [a,c], *Jian Xu* [a,c], *Min Zuo* [b,**]

[a] Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
[b] Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology & Business University, Beijing 100048, China
[c] School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 28 November 2017
Received in revised form 8 May 2018
Accepted 28 May 2018

Keywords:
Maize haploid seeds
Identification
Nonlinear analysis method
Near-infrared spectra
Non-destructive

The identification of maize haploid seeds is a significant process in genetic research and modern maize breeding. Adopting near-infrared spectroscopy technology to distinguish haploid seeds from hybrid seeds has the advantages of being non-destructive, rapid and low cost. However, due to the influence of light, temperature, humidity, near-infrared intensity, instrument and dynamic change of seed activity, the near-infrared spectra of maize seeds showed high dimensional nonlinear characteristics. In this study, to make full use of the class label information, a nonlinear feature analytical method for haploid maize seeds identification based on Supervised Virtual Sample Kernel Locality Preserving Projection (SVSKLPP) has been proposed. The experimental results showed that the nonlinear identification model SVSKLPP achieved strong classification performance to identify the maize haploid seeds. Moreover, compared with the linear feature extraction methods Principal Component Analysis, Orthogonal Linear Discriminant Analysis, Locality Preserving Projection, Supervised Virtual Sample Locality Preserving Projection and nonlinear feature extraction methods Kernel Locality Preserving Projection, Isomap, Locally Linear Embedding, Laplacian Eigenmaps and Local Tangent Space Alignment, the SVSKLPP model achieved a better performance. The average accuracy, sensitivity and specificity using method SVSKLPP were 97.1%, 98.8% and 95.4% respectively, and it also had high robustness. The overall results show that the SVSKLPP-NIR methodology was efficient in accurately identifying haploid maize seeds, thus demonstrating its capabilities for application in haploid breeding for crop variety improvement.

* *Corresponding author*. Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China. Fax: +86 010 82304573.
** *Corresponding author*. Fax: +86 010 82304573.
E-mail addresses: wjli@semi.ac.cn (W. Li), 13311392158@163.com (M. Zuo).
https://doi.org/10.1016/j.biosystemseng.2018.05.011
1537-5110/

Nomenclature

KLPP	Kernel locality preserving projection
LE	Laplacian eigenmaps
LLE	Locally linear embedding
LPP	Locality preserving projection
LTSA	Local tangent space alignment
NIR	Near-infrared
OLDA	Orthogonal linear discriminant analysis
PCA	Principal component analysis
SVSLPP	Supervised virtual sample locality preserving projection
SVSKLPP	Supervised virtual sample kernel locality preserving projection

1. Introduction

Haploid breeding is combines biotechnology with conventional breeding methods. Genetic improvement using haploid breeding techniques has not only contributed to yield increases but also to the tolerance to diseases, pests and the environment (Germanà, 2011). Maize is the most important grain and feed crop, and its haploid breeding is highly significant to meet the future grain supply needs (Liu et al., 2015). For the acquisition of homozygote, in contrast to 5–7 generations of self-cross mating by conventional methods, the use of haploid can shorten this breeding cycle significantly in 1 year (2–3 generations) (Prasanna, Chaikam, & Mahuku, 2012). For crop improvement, haploid enables the achievement of rapid homozygosity and enhances selection efficiency for recessive genes. However, the average probability of the occurrence of maize haploid seeds is only 0.1% (Geiger, Gordillo, & Koch, 2013). Even with artificial induction, the highest rate is only about 10%. Therefore, it is very important to identify haploid seeds rapidly and precisely from a large number of maize kernels. The methods of maize haploid identification mainly include morphology, cytogenetics, anatomy, molecular biology, radiation, genetic markers, etc (Belhumeur, Hespanha, & Kriegman, 2002; Comon, 1994; Liu & Wechsler, 2000; Snape & Zafeiriou, 2014; Swets & Weng, 1996). But, due to the complex process, long time span, high cost and the need for professional input for these methods, the application of haploid breeding technology has been limited.

Near-infrared (NIR) is a region of the electromagnetic spectrum between visible light (Vis) and mid infrared (MIR). The American Society for Testing and Materials (ASTM) defined the NIR spectrum as 780–2526 nm. The absorption region of the NIR spectrum is consistent with that of the hydrogen groups (OH, NH, and CH) vibrations in organic molecules (Ghosh et al., 2016; Pan, Liu, Chen, Zhang, & Zhao, 2013). By scanning the NIR spectrum of the samples, the characteristic information of the hydrogen groups in the organic molecule can be obtained. Moreover, using NIR spectroscopy (NIRS) to analyse samples is convenient, fast, efficient, non-destructive, chemical reagent free and environmental friendly (Cheng et al., 2013; Geleta et al., 2014; Song et al., 2015). These characteristics provide a new approach to the identification of maize haploid seeds.

The NIR spectra of maize haploid seeds are affected by many factors, such as light, temperature, humidity, NIR intensity, collecting instrument, and the seed itself is also changing (Zhou, Fu, & Ying, 2007). These factors are coupled together rather than simply linear superposition. Therefore, the NIR spectra of maize seeds exhibit high dimensional nonlinear characteristics. Manifold learning is a very suitable method for solving nonlinear problems (Raducanu & Dornaika, 2012; Yin & Huang, 2010). There are many algorithms about manifold learning, but the most representative ones are Isomap (Balasubramanian & Schwartz, 2002; Seung & Lee, 2001), Locally Linear Embedding (LLE) (Kokiopoulou & Saad, 2005; Roweis & Saul, 2000), Laplacian Eigenmaps (LE) (Belkin & Niyogi, 2014), Local Tangent Space Alignment (LTSA) (Zhang & Zha, 2004) and Kernel Locality Preserving Projection (KLPP) (Jiang, Fu, Wen, Hao, & Hong, 2016; Luo, Bao, Mao, & Tang, 2016; Su, Chiu, Kuo, Yeh, & Hsu, 2014; Wong & Zhao, 2012). Isomap is a global dimensionality reduction algorithm, which can keep the high dimensional geometry structure of the original sample data very well in low dimensional space. However, the method requires that the subset of manifolds corresponding to the lower dimensional space is a convex set, otherwise, the geodesic distance using the shortest path approximation cannot characterise the inner manifold structure. Moreover, Isomap has poor noise immunity (Tenenbaum, De, & Langford, 2000). LLE can learn the local manifold structure of the sample set efficiently, and the sample set does not need to be a convex set. It has low computational complexity and strong robustness (Roweis & Saul, 2000). Rotation, translation and expansion have little influence on it. However, LLE needs dense sampling and is sensitive to noise points. The manifold structure of the data set cannot be closed and needs to be locally linear, and the choice of parameters is needed. LE is a local nonlinear dimensionality reduction algorithm based on spectral graph theory (Belkin & Niyogi, 2014). It can keep the local neighbourhood relationship between sample points very well, and the computational complexity is low. But, LE is still sensitive to noise points and needs to select parameters. The LTSA can effectively reconstruct a subset of manifolds of low dimensional space based on global and local features. It does not require subsets of low dimensional spaces to be convex sets. It is difficult to handle large data sets and newly added sample data with LTSA, which is also sensitive to noise points. KLPP introduces the idea of kernel functions to preserve the local structural properties of data. However, it does not take full advantage of sample class information.

The objective of this paper is to propose a novel method for identifying maize haploid seeds, which is Supervised Virtual Sample Kernel Locality Preserving Projection (SVSKLPP) and to demonstrate its advantages.

2. Materials and methods

2.1. *Samples and spectral acquisition*

The seeds of cultivar “Zheng Dan 958” (which has Navajo genetic marker) were used as experimental samples as shown in