Journal of Pharmaceutical Analysis 2012;2(2):117–122

Contents lists available at ScienceDirect

Journal of Pharmaceutical Analysis

www.elsevier.com/locate/jpa
www.sciencedirect.com

ORIGINAL ARTICLE

The expert system of genotype discrimination for D5S818 locus based on near-infrared spectroscopy–principal discriminant variate

Zai-Zhen Wu[a], Jian-Hua Tang[a], Bin Zhang[a], Li-Ping Guo[b], Hong-Ping Xie[a,*], Bing-Ren Gu[c,*]

[a]*College of Pharmaceutical Sciences, Soochow University, Suzhou 215123, China*
[b]*College of Chemistry and Chemical Engineering, Chongqing University of Science and Technology, Chongqing 401331, China*
[c]*Suzhou Institute for Drug Control, Suzhou 215002, China*

Received 16 August 2011; accepted 24 October 2011
Available online 10 November 2011

KEYWORDS

Short tandem repeat;
Near-infrared spectroscopy;
Principal discriminant variate;
Genotyping-tree structure;
Expert system

Abstract This paper studied the expert system of genotype discrimination for the STR locus D5S818 based on near-infrared spectroscopy–principal discriminant variate (PDV). Six genotypes, i.e. genotypes 10–10, 10–11, 11–11, 11–12, 11–13 and 13–13, were selected as research subjects. Based on the optimum polymerase chain reaction (PCR) conditions, about 54 measuring samples for each genotype were obtained; these samples were tested by near-infrared spectroscopy directly. With differences between homozygote genotypes and heterozygote ones, and differences of the total number of core repeat units between the six genotypes, two types of genotyping-tree structure were constructed and their respective PDV models were studied using the near-infrared spectra of the samples as recognition variables. Finally, based on the classification ability of these two genotyping-tree structures, an optimum expert system of genotype discrimination was built using the PDV models. The result demonstrated that the built expert system had good discriminability and robustness; without any preprocessing for PCR products, the six genotypes studied could be discriminated rapidly and correctly. It provided a methodological support for establishing an expert system of genotype discrimination for all genotypes of locus D5S818 and other STR loci.

*Corresponding authors.
E-mail addresses: hpxie@suda.edu.cn (H.-P. Xie), gubingren@163.com (B.-R. Gu)

Peer review under responsibility of Xi'an Jiaotong University.
doi:10.1016/j.jpha.2011.10.007

Production and hosting by Elsevier

1. Introduction

Short tandem repeats (STRs), also known as micro-satellites or simple sequence repeats, are polymorphic sites containing core repeat units of between two and seven nucleotides in length that are tandemly repeated from approximately a half dozen to several dozen times [1]. Because of high polymorphism and genetic stability, they have been widely used in the construction of genomic maps [2], paternity tests in forensic medicine [3,4] and gene diagnosis of diseases [5,6]. At present, many methods have

been reported for detection of STRs, including capillary electrophoresis [7], polyacrylamide gel electrophoresis [8], microdevice electrophoresis [9] and mass spectrometry [10]. For high throughput analysis, the capillary array electrophoresis chip has also been developed [11]. For most of the methods, polymerase chain reaction (PCR), which is used to amplify information from small amounts of available biological materials, is needed. For analysis of PCR products, uorescent dye markers are needed in the electrophoresis-based methods, and purification of PCR products is necessary in mass spectrometry analysis. Our laboratory has investigated the feasibility of near-infrared spectroscopy–principal discriminant variate (PDV) method in the STR genotyping from the methodology [12]. The results showed that this method has advantages of good fitting, stability and strong prediction. Compared with the afore-mentioned methods, not only uorescent dye markers but also purification of PCR products is not necessary. At the same time, it is simple, rapid and low-cost. In reference [12], only three genotypes of D16S539 locus with middle degree of difference were selected for methodology research, but in real usage each STR locus has more than three genotypes, which could not be discriminated by only one PDV model. Therefore, this paper would study the expert system of genotype discrimination for many genotypes of one STR locus in methodology based on the method in reference [12].

The national database in US has recommended the 13 STR loci for paternity test, including D3S1358, THO1, D21S11, D18S51, VWA, D8S1179, TPOX, FGA, D5S818, D13S317, D7S820, D16S539 and CSF1PO. Since alleles of D5S818 locus have the smallest difference of only one core repeat unit (i.e. the alleles 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16), the requirement for the classification ability of PDV models would be higher than others. Therefore, the genotypes 10–10, 10–11, 11–11, 11–12, 11–13 and 13–13 of the D5S818 locus with middle number of core repeat units and high frequency (i.e. the alleles 10 for 0.195, 11 for 0.33, 12 for 0.23 and 13 for 0.18) were selected as the research subjects.

2. Theory

PDV method is often used to establish the spectra-based chemical pattern recognition model for multivariate classification because it can effectively handle the collinear problem, which often appears in multi-variable spectra.

Supposing there are N objects $\boldsymbol{x_n}$ ($n=1, 2, \ldots, N$) from K classes C_k, ($k=1, 2, \ldots, K$) with the k-th class containing N_k objects. The object of the PDV method is to find a direction, called the principal discriminant variate $\boldsymbol{a}$, which maximizes the following principal discriminant criterion:

$$\mathbf{J} = \frac{\boldsymbol{a}^T[\lambda\mathbf{B} + (1-\lambda)\mathbf{T}]\boldsymbol{a}}{\boldsymbol{a}^T[\lambda\mathbf{T} + (1-\lambda)\mathbf{I}]\boldsymbol{a}} \tag{1}$$

where $\mathbf{I}$ is the identity (unitary) matrix of the same size as the covariance matrix $\mathbf{T}$. λ, with a value varied between 0 and 1, is a weight controlling the balance between Fisher linear discriminant analysis (FLDA) and principal component analysis (PCA). That is, it can find a balance between the separability and the stability. By setting $\lambda=1$, the principal discrimination criterion (1) becomes

$$\mathbf{J} = \frac{\boldsymbol{a}^T\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}^T\mathbf{T}\boldsymbol{a}} \tag{2}$$

and Eq. (2) is a criterion of FLDA. If $\lambda=0$, the principal discrimination criterion (1) becomes

$$\mathbf{J} = \frac{\boldsymbol{a}^T\mathbf{T}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} \tag{3}$$

which turns out to be the variance criterion maximized in PCA.

In Eq. (1), $\mathbf{B}$ is the between-class covariance matrix, which is defined as

$$\mathbf{B} = \frac{1}{N}\sum_{k=1}^{K} N_k(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^T \tag{4}$$

$\mathbf{T}$ is the total covariance matrix given by

$$\mathbf{T} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^T \tag{5}$$

In Eqs. (4) and (5), $\boldsymbol{m}$ and $\boldsymbol{m}_k$ are the mean vectors of all objects and those from C_k, respectively, as given by

$$\boldsymbol{m} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \tag{6}$$

$$\boldsymbol{m}_k = \frac{1}{N_k}\sum_{x_i \in C_k}\boldsymbol{x}_i \tag{7}$$

After determining the first PDV $\boldsymbol{a}_1$, one can find successively other directions by Eq. (1) under the orthogonality constraint that $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_j=0$ ($i>j$). Therefore, the PDV method can find a sequence of discriminant variates $\boldsymbol{a}$ until additional discriminant variates do not provide discriminatory information.

3. Experimental

3.1. Genomic DNA extraction

EDTA blood samples available for research purposes at the Department of Forensic Medicine of Soochow University were selected for this research. Genomic DNA samples were extracted according to the Chelex-100 method [12].

3.2. PCR ampli cation

PCR was performed in a final volume of 25 μL reaction mixture containing 1 μL of genomic DNA template, 0.25 mM of each primer (GenScript Inc., Nanjing in China. Primer A: 5′–GAA TGA TTT TCC TCT TTG GT-3′ and primer B: 5′–TGA TTC CAA TCA TAG CCA CA-3′ for D5S818), 0.625U of Taq DNA polymerase (Fermentas Inc., USA), 1 × Taq buffer with 1.5 mM of $MgCl_2$, 0.2 mM of each dNTP (GenScript Inc., Nanjing in China) and 16.625 mL of redistilled water.

PCR was carried out using the PTC-200 Thermal cycler (BIO-RAD, USA) under the following conditions: held for 3 min at 95 °C, followed by 30 cycles of 30 s at 95 °C, 30 s at 58 °C and 30 s at 72 °C, and then held for 7 min at 72 °C.

3.3. Electrophoresis

Agarose gel electrophoresis was performed using the POWER BC6003En electrophoresis apparatus (Shanghai Shenergy

Biocolor Bioscience and Technology Company, China). 3 μL of each PCR product was electrophoresed on the 2% agarose gel containing 0.5 μg/mL EtBr for 7 min at 200 V in 1 × TAE buffer. Image of the gel was taken using GeneGenius BioImaging Systems (SynGene, England).

3.4. *Near-infrared spectral analysis*

20 μL of each PCR product was diluted with water to 450 μL as measuring sample, and then placed in the quartz cell (path length 10 mm, inside width 2 mm). With air as the background, near-infrared spectrum (NIRS) was measured using the Fourier transformation NEXUS infrared spectrometer (Thermo Electron Company, USA). Spectrometer parameters were as follows: the wave number range of 4000–9400 cm^{-1}, the number of scans that are averaged 32 and the resolution of 8 cm^{-1}.

4. Results and discussion

4.1. *Optimization of PCR conditions*

Major factors in uence on the specificity and efficiency of PCR amplification containing concentrations of dNTP, Mg^{2+}, primers and Taq DNA polymerase as well as annealing temperature. In this study, agarose gel electrophoresis was used to detect specificity and efficiency. We deemed that the PCR was non-efficient when there was no specific band for an amplified product, and then the product should be removed. Certainly, for an efficient amplification, the more was the brightness ratio of specific band to respective nonspecific one, the higher was the amplification specificity. Based on the specificity and efficiency, PCR conditions were optimized. Because of very low genomic DNA quantities available for the six studied samples, a whole genome amplification procedure was performed primarily to obtain enough DNA samples. These DNA samples were diluted to 1000 times with water, which then was used as the DNA templates of the second PCR amplification. The gel electrophoresis image of the second PCR products is shown in Fig. 1.

4.2. *The preprocessing of spectral data*

Using the method in Section 3.4, the NIRS-s of all the measuring samples were obtained. The spectra of all the six genotypes were very similar in shape (as shown in Fig. 2 for example the measuring samples of the genotype 10–10).

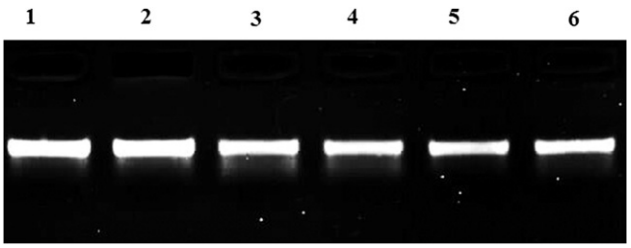

Figure 1 Agarose gel electrophoretograms of the samples of the six different genotypes of the locus D5S818 (Lanes 1, 2, 3, 4, 5 and 6: Genotypes 10–10, 10–11, 11–11, 11–12, 11–13 and 13–13, respectively).

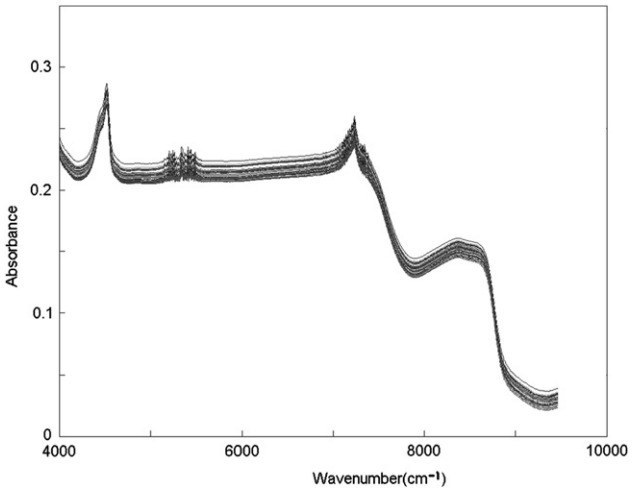

Figure 2 Measured spectra for the 57 samples of the genotype 10–10.

Table 1 Data sets of six different genotypes of STR samples.

Genotype	Calibration set	Prediction set	Prediction accuracy (%)
10–10	38	19	100
10–11	36	18	100
11–11	38	19	100
11–12	37	19	100
11–13	36	19	100
13–13	37	19	100

Obviously, in the no signal range or poor one of the spectra in Fig. 2, it can be found that the NIRS-s of the measuring samples had the shift. In order to ensure that the spectral variation should be only related to genotype differences, the spectral drift needed to be eliminated with the help of the base-corrected method according to the information of no signal range. To eliminate the in uence of concentration on the spectra, this paper used a normalization method for the base-corrected spectra.

4.3. *Establishing PDV discriminant model*

For genotypes 10–10, 10–11, 11–11, 11–12, 11–13 and 13–13, the spectral differences resulted from the differences of their genotypes were very small. So the parallel property of PCR amplifications should be very important. Three times of parallel PCR amplification were carried out for each genotype. The small differences of parallel PCR amplifications would be contained in all the measuring samples, which would ensure the robustness of discriminant models. For all the measuring samples of each genotype, two-thirds of sample were selected randomly as calibration sample set and the rest as prediction one (shown in Table 1).

In Ref. [12], both PDV and support vector machine (SVM) methods were used to establish discriminant models. SVM can better classify a small number of calibration samples, and PDV can effectively handle the collinear problems that often

appear in multi-variable spectra and spectra-characterized samples with only small difference in composition. In this research, the spectral differences between the six genotypes were very small, which could result in high degree of the collinearity. At the same time, considering the number of the measuring samples were enough large, we selected PDV method to establish discriminant models. For example, the two genotypes 10–10 and 10–11 with minimal difference were discriminated successfully using the optimum PDV model with the weigh $\lambda = 10^{-6}$ (Fig. 3).

4.4. *The expert system of genotype discrimination*

For genotyping of the six genotypes studied, two kinds of genotyping-tree structure were studied. One was constructed according to the difference between homozygote genotypes and heterozygote ones, which is shown in Fig. 4. The first layer of this genotyping-tree structure contained one class of homozygous genotypes 10–10, 11–11 and 13–13, and the other

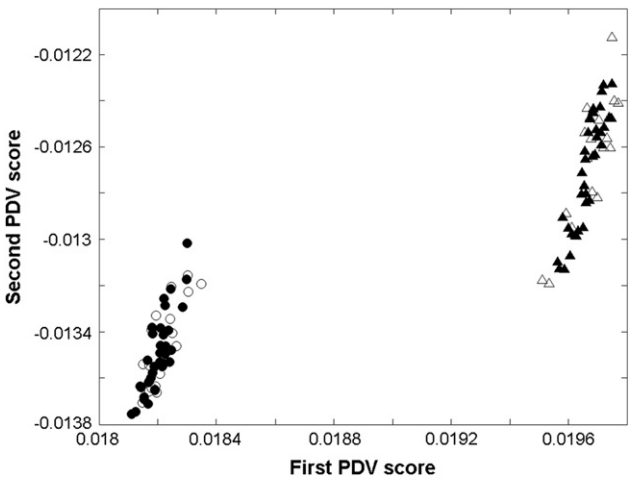

Figure 3 Optimum PDV discriminant model between the genotypes 10–10 and 10–11. (▲" the calibration sample and △" the prediction sample for the 10–10 genotype; ●" the calibration sample and ○" the prediction sample for the 10–11 genotype).

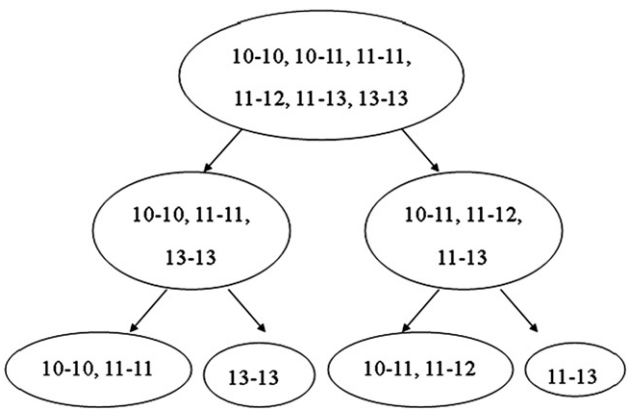

Figure 4 Built genotyping-tree structure of the six studied genotypes based on the difference of homozygote and heterozygote.

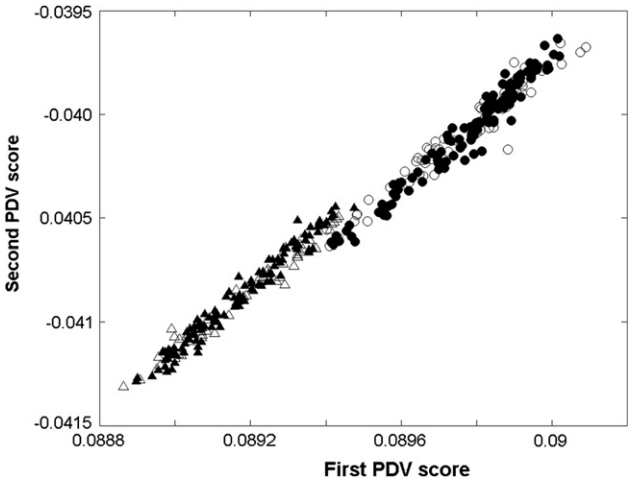

Figure 5 Optimum PDV discriminant model between the class of 10–10, 11–11 and 13–13 and the one of 10–11, 11–12 and 11–13. (▲" the calibration sample and △" the prediction sample for the genotypes 10–10; 11–11 and 13–13; ●" the calibration sample and ○" the prediction sample for the ones 10–11, 11–12 and 11–13).

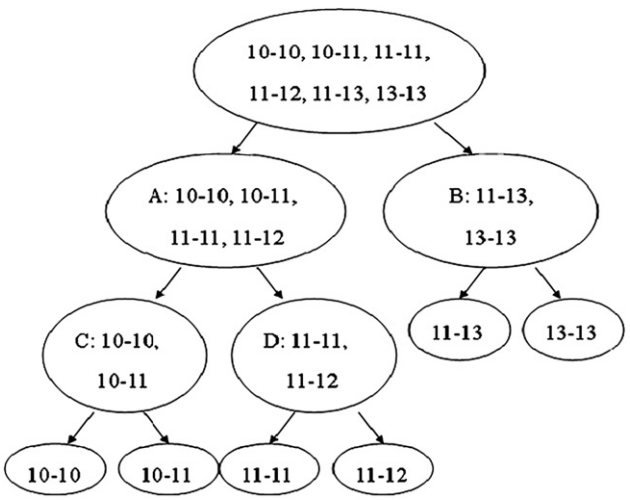

Figure 6 Built genotyping-tree structure of the six studied genotypes based on the different total numbers of the core repeat units of the genotypes.

class of heterozygous ones 10–11, 11–12 and 11–13. The PDV discriminant model of these two classes was optimized by adjusting the weigh parameter λ. The optimum one is shown in Fig. 5, with the weigh $\lambda = 10^{-6}$. It could be found that the model did not have good discriminability, which indicated that genotypes could not be classified based on the difference between homozygote genotypes and heterozygote ones.

The other genotyping-tree structure is shown in Fig. 6, which was established with the different total number of core repeat units of the six genotypes. In the first layer of this genotyping-tree structure, one class had a range 20–23 of total number of the core repeat units and the other was 24–26. The PDV discriminant models for these two classes were established with the decrease of λ (not shown). When the weight

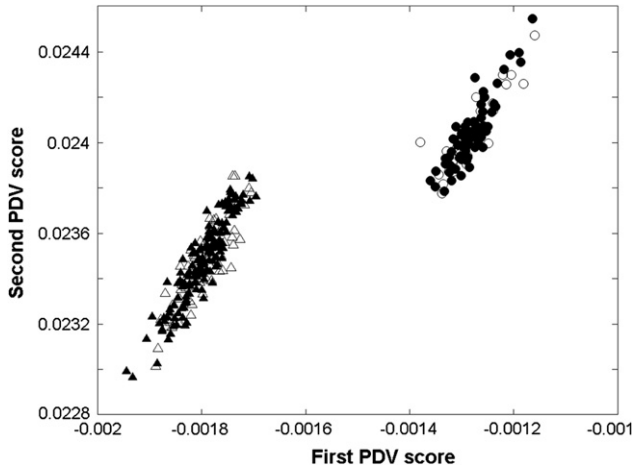

Figure 7 Optimum PDV discriminant model between the class of 10–10, 10–11, 11–11, 11–12 and the one of 11–13, 13–13. (▲" the calibration sample and △" the prediction sample for the genotypes 10–10, 10–11, 11–11 and 11–12; ●" the calibration sample and ○" the prediction sample for the ones 11–13 and 13–13).

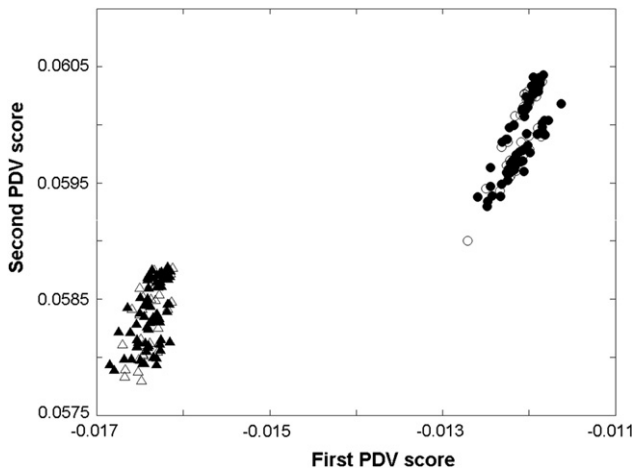

Figure 8 Optimum PDV discriminant model between the class of 10–10, 10–11 and the one of 11–11, 11–12. (▲" the calibration sample and △" the prediction sample for the genotypes 10–10 and 10–11; ●" the calibration sample and ○" the prediction sample for the ones 11–11 and 11–12).

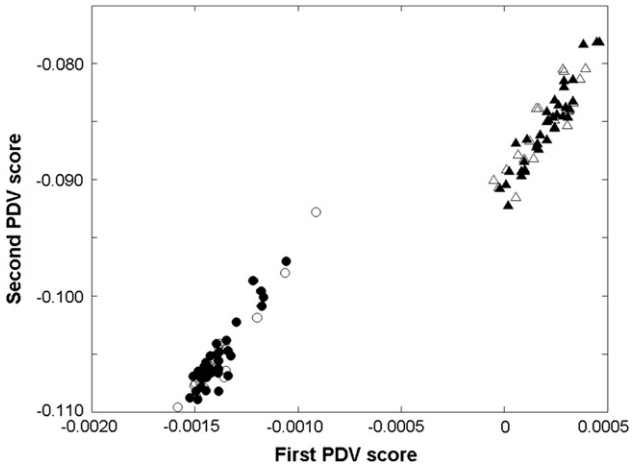

Figure 9 Optimum PDV discriminant model between the genotypes 11–13 and 13–13. (▲" the calibration sample and △" the prediction sample for the 11–13 genotype; ●" the calibration sample and ○" the prediction sample for the 13–13 genotype).

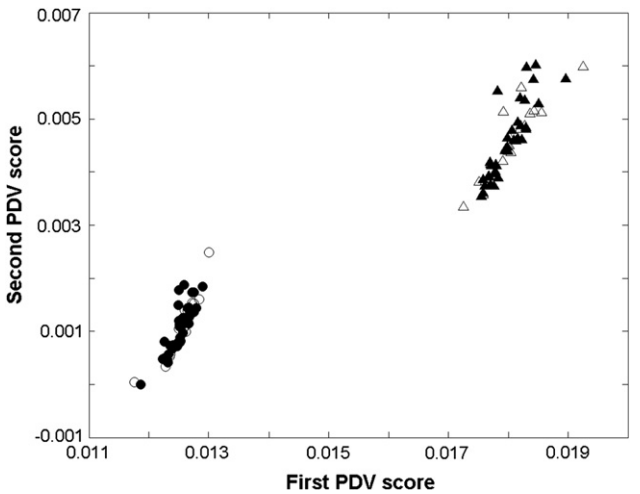

Figure 10 Optimum PDV discriminant model between the genotypes 11–11 and 11–12. (▲" the calibration sample and △" the prediction sample for the 11–11 genotype; ●" the calibration sample and ○" the prediction sample for the 11–12 genotype).

$\lambda = 10^{-6}$ (Fig. 7), there were a large between-class distance and small within-class distances, and the predictive samples were in the range of the calibration ones. So this model was the best one. All other optimum PDV models (Fig. 3, Figs. 8–10) in this tree-shape structure had the similar properties with Fig. 7, and the weight parameters λ of Figs. 8, 9 and 10 were 10^{-5}, 10^{-6} and 10^{-5}, respectively. Based on the Figs. 3, 9 and 10, the discriminant accuracies of the six genotypes were calculated. In the three models, although some predictive samples were out of the range of the calibration ones, it would not contribute to the prediction accuracy because of the good discriminant ability of the PDV1/PDV2. The prediction accuracies are shown in the last column of Table 1. In this genotyping-tree structure, six genotypes were discriminated successfully, which indicated that the established expert system of genotype discrimination based on the PDV models in this genotyping-tree could be used to genotyping of STR.

5. Conclusions

This paper studied the expert system of genotype discrimination for the six genotypes of STR locus D5S818 in methodology based on near-infrared spectroscopy–PDV method. The expert system of genotype discrimination in the genotyping-tree structure established with differences of total number of core repeat units had a good discriminability and robustness. Without any preprocessing for PCR products, the six genotypes of D5S818 locus could be classified rapidly and successfully, which provided a methodological support for

establishing the expert system of genotype discrimination for other STR loci.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (Grant no. 81001686).

References

[1] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, second ed., Elsevier, New York, 2005.
[2] M. Morgante, A. Rafalski, P. Biddle, et al., Genetic mapping and variability of seven soybean simple sequence repeat loci, Genome 37 (1994) 763–769.
[3] C.L. Holt, C. Stauffer, J.M. Wallin, et al., Practical applications of genotypic surveys for forensic STR testing, Forensic Sci. Int. 112 (2000) 91–109.
[4] L. Roewer, Y chromosome STR typing in crime casework, Forensic Sci. Med. Pathol. 5 (2009) 77–84.
[5] R.A. Lea, A. Dohy, K. Jordan, et al., Evidence for allelic association of the dopamine b-hydroxylase gene (DBH) with susceptibility to typical migraine, Neurogenetics 3 (2000) 35–40.
[6] M.A. Trnitat, B.L. Rosa, F. Esther, et al., Preimplantation genetic diagnosis of P450 oxidoreductase deficiency and Huntington disease using three different molecular approaches simultaneously, J. Assist. Reprod. Genet. 26 (2009) 263–271.
[7] J.M. Butler, R. Schoske, P.M. Vallone, et al., A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers, Forensic Sci. Int. 129 (2002) 10–24.
[8] U. Ricci, M. Klintschar, F. Neuhuber, et al., Study on the STR TPOX in an Italian and an Austrian population using two different primer pairs and three different electrophoretic methods, Int. J. Legal. Med. 111 (1998) 212–214.
[9] N. Goedecke, B. McKenna, S. El-Difrawy, et al., Microdevice DNA Forensics by the Simple Tandem Repeat Method, J. Chromatogr. A 1111 (2006) 206–213.
[10] Z.J. Meng, A. Tracey, S. Willis, et al., The use of mass spectrometry in genomics, Biomol. Eng. 21 (2004) 1–13.
[11] A. Mast, M. de Arruda, Invader assay for single-nucleotide polymorphism genotyping and gene copy number evaluation, Methods Mol. Biol. 335 (2006) 173–186.
[12] L. Ren, Y.Z. Gao, J.H. Yin, et al., The determination for the three genotypes of D16S539 locus based on near-infrared spectroscopy and chemical pattern recognition, Anal. Chim. Acta 638 (2009) 202–208.