Available online at www.sciencedirect.com
ScienceDirect
Analytica Chimica Acta 595 (2007) 43–50
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Application of chemometric tools for automatic classification and profile extraction of DNA samples in forensic tasks

Isneri Talavera Bustamante [a,*], Francisco Silva Mata [a], Noslen Hernández González [a], Ricardo González Gazapo [a], Juan Palau [a], Marcia M. Castro Ferreira [b]

[a] *Advanced Technology Application Center, 7ma, No. 21812, Siboney, Playa 12200, Cuba*
[b] *Instituto de Química, Universidad Estadual de Campinas (UNICAMP), 13083-970 Campinas, SP, Brazil*

Received 12 October 2006; received in revised form 23 December 2006; accepted 8 January 2007
Available online 13 January 2007

Abstract

In this paper a method for the automatic DNA spots classification and extraction of profiles associated in DNA polyacrilamide gel electrophoresis is presented and it integrates the use of image processing techniques and chemometrics tools. A software which implements this method was developed; for feature extraction a combination of a PCA analysis and a C4.5 decision tree were used. To obtain good results in the profile extraction only DNA spots are useful; therefore, it was necessary to solve a two-class classification problem among DNA spots and no-DNA spots. In order to perform the classification process with high velocity, effectiveness and robustness, comparative classification studies among support vector machine (SVM), K-NN and PLS-DA classifiers were made. The best results obtained with the SVM classifier demonstrated the advantages attributed to it in the literature as a two-class classifier. A Sequential Cluster Leader Algorithm and another one developed for the restoration of pattern missing spots were needed to conclude the profiles extraction step. The experimental results show that this method has a very effective computational behavior and effectiveness, and provide a very useful tool to decrease the time and increase the quality of the specialist responses.

Keywords: Classification; Support vector machine; Image analysis

1. Introduction

For human identity, scientists use the DNA profile. To obtain a profile it is necessary to generate and analyze a Short Tandem Repeat ("STR") data. Such data is derived from a blood (or other) sample taken from a person or obtained from the crime scene [1]. Each data contains several STR loci. A STR loci of an individual has two "alleles," each corresponding to a true DNA. It is common to build a DNA profile using 10 STR loci (20 alleles), and it is extremely unlikely that the 20 numbers (i.e., 10 length pairs or alleles) from one individual will identically match the 20 numbers of an unrelated individual. This uniqueness serves as a "fingerprint" of genetic identity [2,3]. During laboratory data generation, the forensic scientist conducts experiments to transform these unknown DNA samples into observable data [4].

The polyacrilamide gel electrophoresis is a common analysis to develop this transformation, in which DNA sequences are obtained in the form of electrophoretic bands on a polyacrilamide gel plate, the bands are visualized with a silver tintion reagent, and are detected as black spots [5].

There is a standardized method to manually detect the spots of DNA and make the numbers designations of the pair alleles per loci, but it is a very inefficient form to perform the task.

Results obtained by Walczak and Kaczmarek in the preprocessing and matching of 2D gel electrophoresis images for protein studies offered us important information to undertake the work with DNA samples [6,7].

In this paper a method for the automatic DNA spots classification and extraction of profiles associates in DNA polyacrilamide gels is presented. It integrates the use of image processing techniques and chemometric tools. A software which implements this method was developed.

2. Image acquisition and preprocessing

To acquire the images a digital camera Sony DSC F717 was placed on a controllable illumination system. The

* Corresponding author. Tel.: +537 272 1628; fax: +537 273 0045.
E-mail address: italavera@cenatav.co.cu (I.T. Bustamante).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.01.010

Fig. 1. Acquisition module: (a) general view, (b) light sources, (c) mobile gate.

-1	-2	-1
0	0	0
1	2	1

$G_x = (z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3)$

-1	0	1
-2	0	2
-1	0	1

$G_y = (z_3 + 2z_6 + z_9) - (z_1 + 2z_4 + z_7)$

z_1	z_2	z_3
z_4	z_5	z_6
z_7	z_8	z_9

Image neighborhood

Fig. 2. Sobel edge detector mask and the first order derivatives that it implements.

polyacrilamide gel plate is placed on a mobile gate between a diffuser plate and the digital camera and the light sources are in the bottom, below the diffuser plate as Fig. 1 shows.

One of the main tasks of pre-processing is the removal or reduction of noise. In order to find the most suitable one for this kind of images some linear and non-linear filtering methods, were compared. Best results were obtained using a homomorphic filtering [8]. In this case, the filter acts to reduce the low frequency multiplicative noise that is produced as a result of a non-homogeneity illumination or a non-homogeneous chemical process.

The next step involves the spots segmentation. In order to carry out this task, a Sobel Edge Detector was applied. It uses the special mask in Fig. 2 to approximate digitally the first derivatives *Gx* and *Gy* of the image [9].

In other words, the gradient at the center point in a neighbourhood is computed as follows:

$$g = [\mathrm{Gx}^2 + \mathrm{Gy}^2]^{1/2}$$
$$g = \{[(z_7 + 2z_8 + z_9) - (z_1 + 2z_2 + z_3)]^2 + [(z_3 + 2z_6 + z_9) - (z_l + 2z_4 + z_7))]^2\}^{1/2} \quad (1)$$

where $z_1, \ldots, z_9$ conform the image neighbourhood. A pixel at location (x, y) is an edge pixel if $g \geq T$ at that location, where T is a specific threshold. The segmentation process finishes by applying an automatically global threshold following the iterative procedure proposed by González and Woods [9].

3. Feature selection

Once the spot's segmentation is finished, the next step is to represent and describe it in a suitable form for further computer processing. A representation using 14 boundary and region descriptors was chosen as can be seen in Table 1.

Table 1
Boundary and region descriptors and their symbols used to represent the spots

Descriptors	Symbol
Area	A
Complementary area	Ac
Complementary area–area ratio	Rac
Perimeter	Pe
Rectified perimeter	Per
Compactness	Cp
Maximum width	Anmax
Maximum height	Almax
2D moment invariants	P1
	P2
	P3
	P4
Height–width ratio	Raa
Area bounding box	Abb

Feature selection is an important problem in machine learning. In high dimensional data it is critical to weed out noisy features, which have no discriminatory power, before applying standard learning algorithms.

PCA focuses on the search for features, which explain most of the variance in the data [10]; this is clearly not enough for feature selection because there is no clear correlation between variance and discriminatory power. To take into account all the factors for a good feature selection in this case a combination of PCA and a C4.5 decision tree were used. Both techniques were applied separately on a set of samples ("*S*") described by original features, composed by spots from different images in order to guarantee a high representativeness. It contains 638 spots, 305 in DNA class and 325 in no-DNA class.

3.1. PCA

On way to look for the relevant features using PCA is to obtain their modeling and discriminatory power.

The modeling power varies with the number k of principal components selected but is variable-oriented. Typically, it is not helpful in determining the optimal number of factors to retain, but it does point out important variables.

The modeling power of variable "*j*" is defined as:

$$MP_j = 1 - \frac{\hat{S}_j}{S_{0j}} \quad (2)$$

where $\hat{S}_j$ is the square root of variable residual variance $\hat{S}_j^2$, which is calculated using the *j*th column of the residual matrix E_k defined above:

$$\hat{S}_j^2 = \frac{\hat{e}_j^T \hat{e}_j}{n-k-1} \quad (3)$$

and S_{0j} is the square root of the total variance of that variable calculated as:

$$\hat{S}_{0j}^2 = \frac{1}{n-1}\sum_i^n (x_{ij} - \bar{x}_j)^2 \quad (4)$$

and "*n*" is the number of samples.

When the power of a variable to model information in the data increases, MP approaches 1; however when it decreases, MP approaches 0. Even with random data, some features will exhibit high modeling power, so *an absolute threshold cannot be specified*. Instead, the different variables should be compared based on their relative modeling power.

It may also be instructive to know which variables are best at discriminating between training set classes. For each variable, comparing the average residual variance of each class fit to all other classes and the residual variance of all classes fit to themselves provides an indication of how much a variable discriminates between "correct" and "incorrect" classification. The discriminatory power of variable "*j*" between the two-classes *q* and *r* is thus defined as:

$$DP_j^{r,q} = \left[\frac{(S_{j,r}^q)^2 + (S_{j,q}^r)^2}{(S_{j,r}^r)^2 + (S_{j,q}^q)^2}\right]^{1/2} - 1 \quad (5)$$

where $S_{j,r}^q$ is the corresponding standard deviation of variable "*j*" and it is defined as follows:

$$S_{j,r}^q = \left[\frac{J}{n_r(J-k_q)}\sum_{p=1}^{n_r}(e_{jp}^q)^2\right]^{1/2} \quad (6)$$

It is important to note that the sum is over n_r (the number of samples in class *r*) and e_{jp}^q is the residual standard deviation of the *p*th object in class *r* when fitted to the class model *q*. *J* is the number of variables and k_q the number of principal components selected in class model *q*.

A value of $DP_j^{r,q}$ close to zero indicates a low discriminatory power and much above one a good power [11].

PCA has been applied to the set of samples "*S*". With the aim of giving the same importance to all features the data was scaled before PCA to zero mean and unit variance. Four PC_s explaining 96.05% of total variance, with a RMSECV value of 1.238 was selected to describe the data, as can be seen in Tables 2 and 3. An exploratory view of the samples/scores plot on PC_1 versus PC_2 in Fig. 3, shows a real possibility of separation among DNA and no-DNA spots, likely to improve with a further feature selection.

Table 2
Principal components selected and % of variance explained

PC_s selected	Var. *E* (%)
PC_1	67.70
PC_2	11.40
PC_3	10.34
PC_4	4.52
Total	96.05

Table 3
Factor selection, principal components and RMSECV values

Principal Components	RMSECV
PC_1	2.18112
PC_2	1.90516
PC_3	1.54119
PC_4	**1.23802**
PC_5	1.28845
PC_6	1.09052
PC_7	1.45241
PC_8	2.24155
PC_9	2.36414
PC_{10}	2.99329

Fig. 4 shows a comparison of all original features by their modeling and discriminatory power values. To select the best features according to PCA, the criteria to use only the features with high discriminatory power values to enhance the classification process is not a good choice, because it is possible to exaggerate the difference among the classes. A better approach is to remove the ones with low modeling and discriminatory power values.

According to this criterion the features selected as relevant using PCA were: A, Pe, Per, Anmax, P_1, P_3, Raa, and Abb.

Fig. 5 shows the PCA plot obtained using only selected variables, the results prove the improvement in the separation among DNA and no-DNA spots.

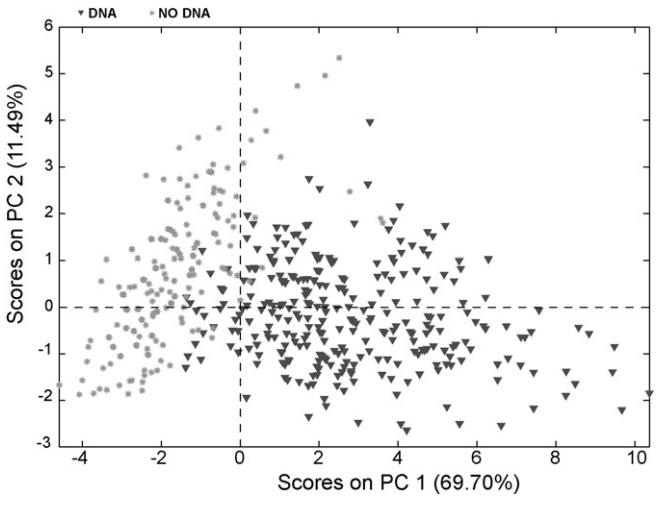


Fig. 3. Sample/scores plot on PC_1 vs. PC_2 of training set "*S*".

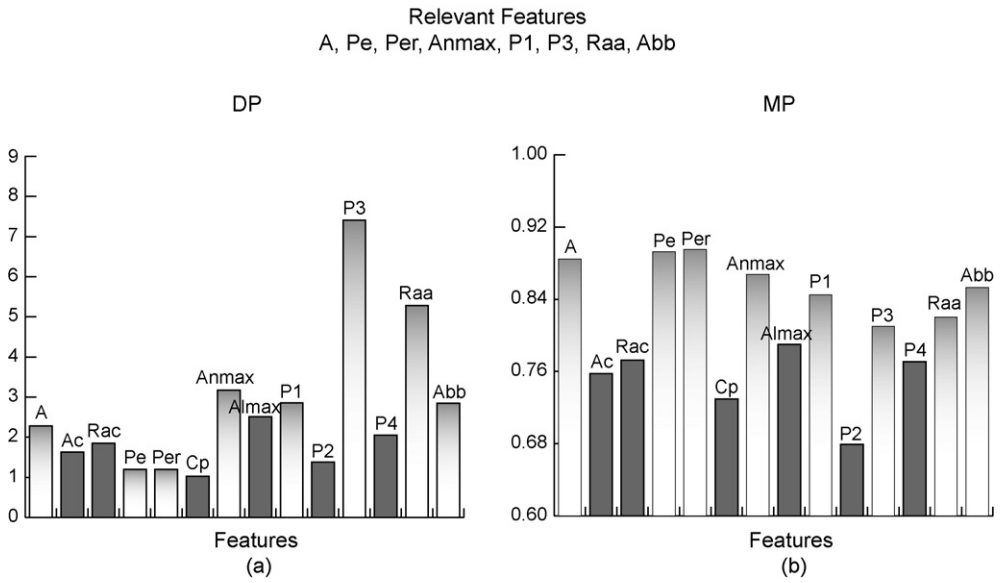


Fig. 4. Comparison of all original features by their modeling and discriminatory power values. (a) Modeling power vs. features. (b) Discriminatory power vs. features.

3.2. C4.5 decision tree

In order to complement the results obtained with PCA with more information about the effectiveness of the discriminatory capabilities of the original features aiming at obtaining a good separation among the two-classes, a C4.5 decision tree was applied.

In general, decision trees represent a disjunction of conjunctions of constraints on the attribute-values of examples. The selection is based on a statistical property called *information gain* that measures how well a given attribute separates the training examples according to their classification. In this sense the tree nodes will contain the most relevant features, being the most important feature located at the high levels of the tree [12].

An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node and moving down by the tree branch corresponding to the value of the attribute. This process is then repeated at the node on this branch and so on, until a leaf node is reached which provides the classification of the instance.

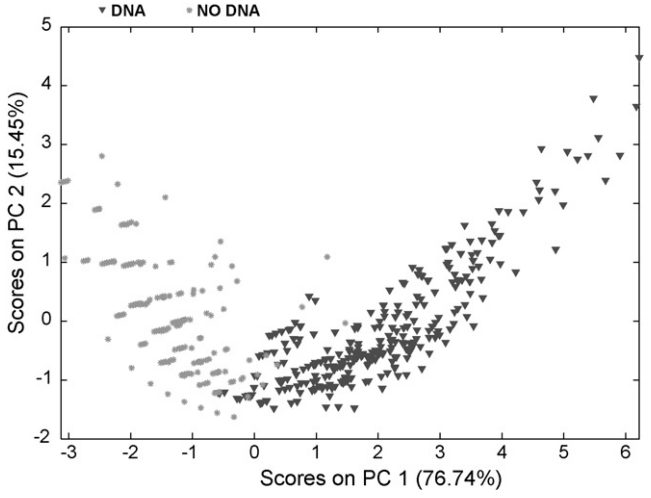


Fig. 5. Sample/scores plot on PC_1 vs. PC_2 of training set "*S*" with selected variables.

Three different experiments were designed and applied to the set "*S*". In the Experiment 1, "*S*" was randomly divided in four disjoint training sets which have the same number of samples and also a good balance among DNA and no-DNA spots. Four decision trees were built; each of them validated using 10 folds cross validation. In Experiment 2, four different sets were prepared using resampling on "*S*"; each of them was divided in 75% of the samples for training and the other 25% for testing. This means that each classifier was built with training sets that could contain common samples. In the Experiment 3, only one decision tree was built using the set "*S*" as training set and it was validated with 10 folds cross validation.

The results in Fig. 6 show that some of the initial features do not appear in any of the trees. This means that they are not relevant for DNA characterization and recognition. Finally, Anmax, Pe, Pl, Raa, P3, A were selected as a result of the union of the partial results of each decision tree.

For final feature selection, a combined interpretation of the results obtained with PCA and decision tree was made, features selected as relevant must have high modeling power values with discriminatory capabilities. Taking into account the high correlation that exists between pe and per, one of them was rejected. The selection of pe instead of per was based on the fact that pe appears more times in the experiments carried out with decision trees. The features selected Anmax, Pe, P_1, Raa, P_3, and A were used to build an automatic classifier.

4. Classification algorithms

All spots present in the training set "*S*", are described automatically, using the most significant features obtained in Section 3. For the profile extraction only DNA spots are use-

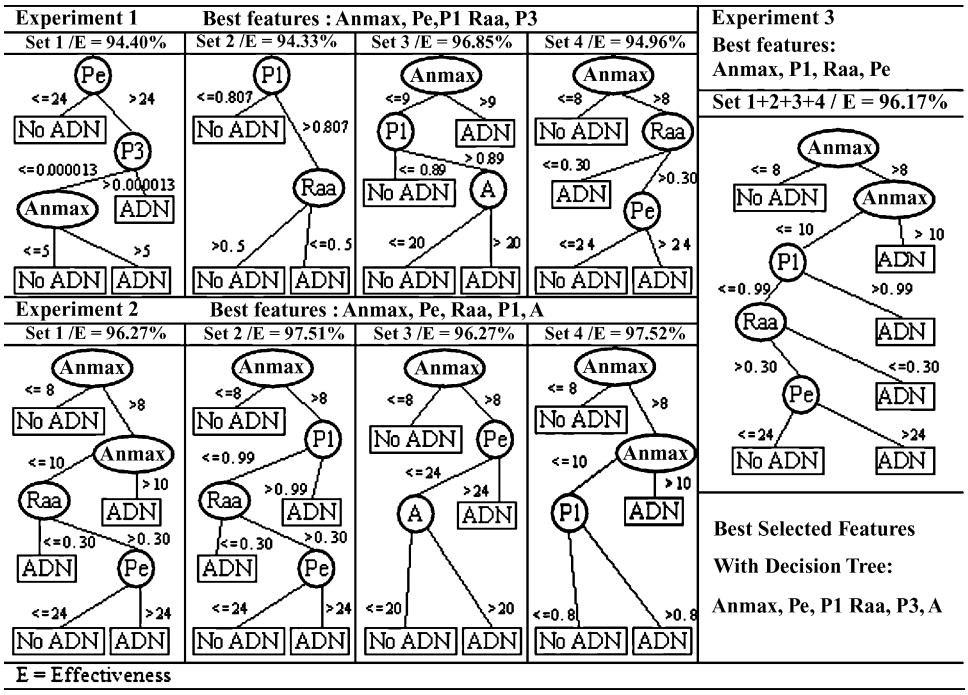


Fig. 6. Features selection with C4.5 decision tree.

ful, therefore, it is necessary to solve a two-class classification problem among DNA spots and no-DNA spots. In order to do the classification process with a high velocity, effectiveness, and robustness, a study of the behavior of three classifiers was made. A SVM classifier was selected and the results were compared with two other classical classifiers as KNN, and PLS-DA.

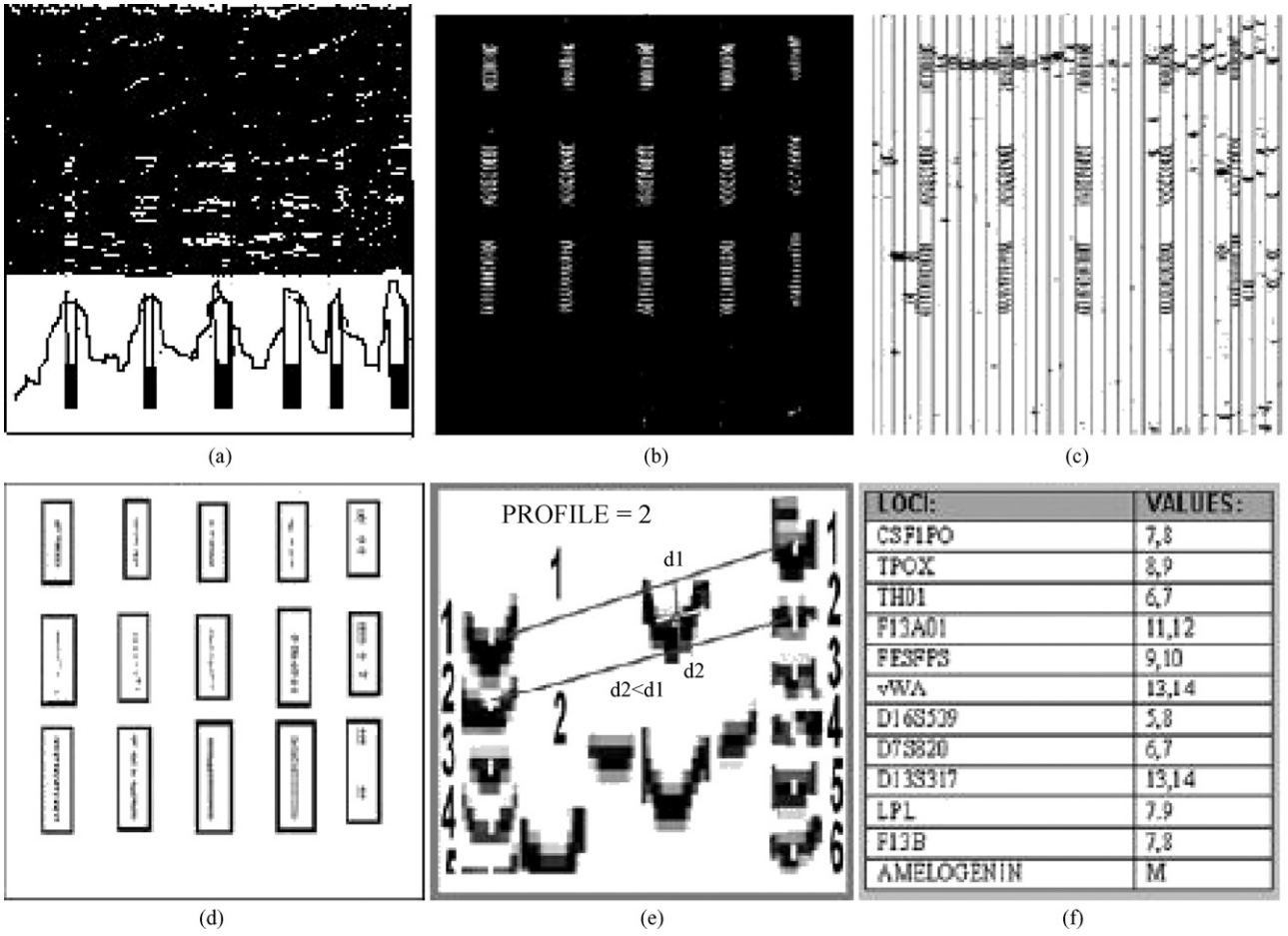
SVMs are kernel based learning algorithms introduced by Vapnik [13,14]. In a two linearly separable class problem the principal aim of the SVMs classifiers, is to find a separating "maximal margin" hyperplane which gives the smallest generalization error among the infinite number of possible hyperplanes. The data on margin and/or the closest ones are called support vectors. They are found by solving a quadratic programming (QP) problem.

Sometimes the separation function between the classes is nonlinear. In this case, the data will be mapped from an input space into a high dimensional feature space by a nonlinear transformation $\phi(x)$. The QP problem in a feature space depends only on a dot product $\phi(x_i)^T\phi(x_j)$ for that reason learning can be performed by using Mercer theorem [15] for positive definite functions, which allows replacing the product $\phi(x_i)^T\phi(x_j)$ by a positive definite symmetric kernel function $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$. In our case a non-linear SVM_s with a Gaussian radial basis function (RBF) kernel was used [15]. PLS-DA and K-NN models were built using the PLS Tool Box Matlab V.3.5.

5. DNA's profile extraction

After the Classification process, an image with only DNA spots is obtained. The profile extraction process was developed following five steps:

1. Detection of the candidate's regions in the image that contains the STR loci patterns, according to the intensities histogram along the x axis.
2. Division of the image in lanes, using as reference the coordinates contributed by the patterns. Each lane contains one sample or the set of STR loci patterns according to the distribution of samples and patterns applied by the specialist in the polyacrilamide gel.
3. Determination inside the pattern's lanes the different sub regions that contain the STR loci's, each of them has a specific sequence of spots with their correspondent assigned numbers. To solve this task a Sequential Leader Cluster algorithm was used [16].
4. Restoration of the sequence of spots inside the STR loci patterns, in case of counting with missing spots or with joined ones, as a consequence of a malfunction of the classification algorithm, or by difficulties in the electrophoresis chemical process. To restore the missing spots a new algorithm was developed [17]. The joined spots are separated by means of the detection of Freeman's chain typical segments of the contour [18]. The calculation of the horizontal dividing halfback line among them permits an effectiveness separation.
5. Assignation of the corresponded number to the spots that represents the two alleles per STR loci in order to obtain the DNA profile of each sample. To solve this task, it is necessary first the layout of the horizontal lines that join the centroide of each spot in the sequence of the STR loci patterns with their matches distributed in the plate (remember that each of these spots in a sequence of a STR loci has a unique and specific number), therefore all the spots in the same line have the same number assigned. Applying the formula of distance of one point to a straight line, it is possible to evaluate the



LOCI:	VALUES:
CSF1PO	7,8
TPOX	8,9
TH01	6,7
F13A01	11,12
FES/FPS	9,10
vWA	13,14
D16S539	5,8
D7S820	6,7
D13S317	13,14
LPL	7,9
F13B	7,8
AMELOGENIN	M

Fig. 7. (a) Candidate's regions in the image according to the intensities histogram along the *x* axis, (b) STR loci patterns regions, (c) division in lanes, (d) determination sub-regions, (e) assigning number to alleles, (f) DNA profile.

distance from the centroid of each spots (alleles), to the lines of the patterns spots nearest to them. The numbers assigned to the alleles are the same assigned to the lines of the patterns whose distances are the shortest to them. A summary of all this process is displayed in Fig. 7.

6. System implementation

For the preprocessing step, we used software in C# based on the algorithms and procedures proposed by Gonzalez and Woods [9]. The feature selection using the decision tree C4.5 was implemented by the pack of classes that offers Software WEKA [19] specifically Weka classifier tree J48. As this software is programmed in Java # a DLL that permits the conversion to Visual Studio C# was developed in order to guarantee the compatibility with this method.

Classifications with SVMs were done using LIBSVM [20].

7. Results and discussion

For training the SVMs and to build the PLS-DA and KNN models, the same data set used for the feature selection was employed, for testing a set of 20 DNA polyacrilamide gel electrophoresis plates, containing 200 real samples were used. The plates have been directly recorded with the acquisition module, and the images obtained were automatically stored in the computer for the process. All models were built using original (selected) variables. The best PLS-DA model was obtained with the following specifications: preprocessing: autoscale, validation: cross (5), optimal factors: 5; for the KNN model the best specification were: preprocessing: autoscale, validation: none and five optimal neighbors. In order to obtain the best parameters value to train the SVM a grid search using 10 folds cross validation was done. The ranges assigned to each hyperparameter were $\log_2 C = (-5, -4, \ldots, 15, 16)$ and $\log_2 \gamma = (-4, -3, \ldots, 4)$. Then, with the values obtained by grid search, the SVM classifier was trained again and the model was validated using 10 folds cross validation, giving us an accuracy of 98.26%.

The accuracy of SVM trained using different training parameters combination and the characteristic of the obtained model are shown in Fig. 8.

The classification accuracy was calculated by taking the number of correctly classified spots for each classifier, and divided by the total number of samples into the test data set. Table 4 shows the results obtained in the classification.

The good results obtained in the classification task demonstrated the advantages attributed in the literature to the SVM_s as a two-class classifier. The training process was very fast,

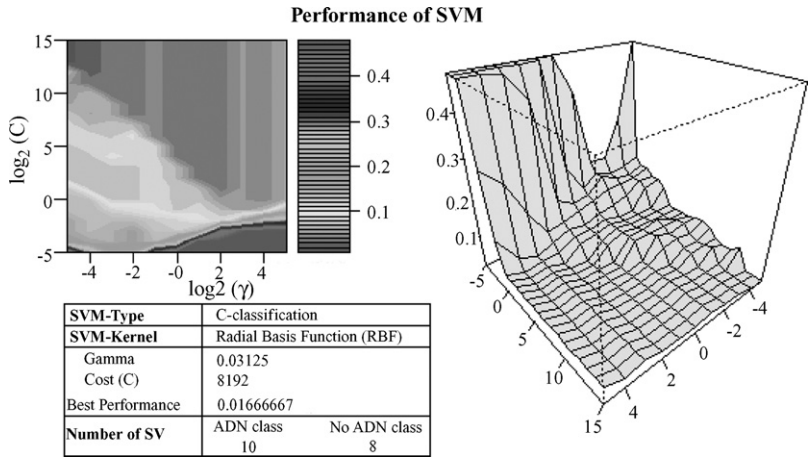


Fig. 8. Performance of SVM, the best pair of parameter (C, γ) values were extracted from the area with more accuracy. The general characteristic of the obtained model are presented.

Table 4
DNA spot classification results with three classifiers

	# of spots	Confusion matrix SVMs		Confusion matrix PLS-DA		Confusion matrix K-NN	
		ADN	NoADN	ADN	NoADN	ADN	NoADN
ADN	433	409	24	407	26	391	42
NoADN	4938	6	4932	84	4854	36	4902
Total	5371	CA—99.44%		CA—97.95%		CA—98.54%	

only 30 s, fundamentally because their structure is automatically determined on the basis of the training data and relatively few parameters are needed; overfitting can be avoided without using a validation set.

The set of the original plates, was processed by the expert using the standardized manual procedure and the results of the profile extraction were compared with the results obtained applying the automatic method taking into account the success rate and the time of response. Table 5 shows the results obtained in this comparison.

Added to the table only five profiles was not possible to extract, caused by the presence of mix samples (DNA of two persons are present in the same sample) with four different alleles present in each STR Loci thus, not being able to determine which of the six pairs of alleles is the correct by the automatic method.

Another significant result is the decrease in the time's response of the task that influences not only the increase of the available time of the expert but also the decrease of the cost of the analysis.

Table 5
Automatic profile extraction results vs. manual method results

# of samples	Profiles detected by expert	System success	Success rate	Time of response	
				Expert	Automat.
200	204	199	97.54%	20 days	15 min

8. Conclusions

The development and implementation of an effective method for the automatic DNA spots classification and extraction of profiles associated in DNA polyacrilamide gel electrophoresis, combining image process and pattern recognition techniques are obtained.

Different types of algorithms as: C4.5 decision trees, PCA, support vector machines, Leader Algorithm and the contribution with a new one for restoration purposes are used to resolve all the tasks.

The experimental results show that this method has a very nice computational behavior and effectiveness, and provide a very useful tool to decrease the time and increase the quality of the specialist responses.

References

[1] P. Gill, A. Urquhart, E. Millican, E. Oldroyd, N. Watson, S. Sparkers, Adv. For. Haemogenet. (1996) 235–242.
[2] E.S. Lander, Science 260 (1993) 1221.
[3] C. Lewontin, D. Hartl, Science 254 (1991) 1745–1750.
[4] J. Weber, P. May, Am. J. Hum. Genet. 44 (1989) 388–396.
[5] C. Estrada, http://www.ugr.es/~eianez biotecnologia/forensetec htm 1, 2001.
[6] B. Kacmazmarek, B. Walczak, S. Jong, Anal. Chem. 75 (2003) 3631–3636.
[7] B. Kacmazmarek, B. Walczak, S. Jong, B. Vandeginste, Proceeding of CAC-2004, p. 171.
[8] E. Garea, J. Silva, I. Talavera, N. Hernández, R. González, Proceeding of CAC-2006.

[9] R. Gonzalez, R. Woods, Digital Image Processing using MATLAB, second ed., Prentice Hall, 2004, pp. 385–387.
[10] J.E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, New York, 1991.
[11] S. Wold, M. Sjostrom, Chemometrics: Theory and Application, in: B.R. Kowalski (Ed.), ACS Symposium Series, vol. 52, 1977, pp. 243–282, Chapter 12.
[12] R.J. Quinlan, C4.5 Programs for Machine Learnig, Morgan Kaufmann Series in Machine Learning, January 15, 1993.
[13] V. Vapnik, A. Chervonenkis, Theory of Pattern Recogn, Nauka, Moscow, 1974.
[14] V. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, 1995.
[15] C. Scholkopf, J. Burges, A. Smola, Advances in Kernel methods, MIT Press, 1999.
[16] J. Hartigan, Clustering Algorithm, John Wiley and Sons, New York, 1975.
[17] F. Silva, I. Talavera, I.R. González, N. Hernández, J. Palau, M. Santiesteban, LNCS 3773 (2005) 242–251.
[18] A. Álvarez, J. Ruiz, M. Sanchiz, LNCS 2905 (2003) 512–520.
[19] I.H. Witten, E. Frank, Data Mining Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann Publisher, 2005, Part II The WEKA Machine Learning workbench.
[20] C. Chang, C. Lin, Library for support vector machines, Eng. National Taiwan University, 2003. Available in Internet http://www.csie.ntu.edu.tw/~cjlin/libsvm/. May 14, 2006.