Automatic Extraction of DNA Profiles in Polyacrilamide Gel Electrophoresis Images

Francisco Silva-Mata[1], Isneri Talavera-Bustamante[1], Ricardo González-Gazapo[1], Noslén Hernández-González[1], Juan R. Palau-Infante[1], and Marta Santiesteban-Vidal[2]

[1] Advanced Technologies Applications Center, MINBAS, Cuba
{fjsilva, italavera, rgazapo, nhernandez, jpalau}@cenatav.co.cu
http://www.cenatav.co.cu/
[2] Central Criminologist Laboratory, Cuba
gordillo@mn.mn.co.cu

Abstract. In this paper is presented a method for the automatic DNA spots classification and extraction of profiles associated in DNA polyacrilamide gel electrophoresis based on image processing. A software which implements this method was developed, composed by four modules: Digital image acquisition, image preprocessing, feature extraction and classification, and DNA profile extraction. The use of different types of algorithms as: C4.5 Decision Trees, Support Vector Machines and Leader Algorithm are needed 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.

1 Introduction

DNA profiling has attracted a good deal of public attention in the last years. The practical application of DNA technology to the identification of biological material has had a significant impact on forensic biology, because it enables much stronger conclusions of identity or non-identity to be made [1].

For human identity, scientists use Short Tandem Repeat ("STR") loci [2]. Each STR locus exhibits variation in DNA molecule length. One person will inherit two specific lengths from their parents, which is likely to be different from the pair of lengths of another person. STR locus of an individual has two "alleles," each corresponding to a true DNA. To form a DNA profile, scientists generate and analyze STR data. Such data is derived from a blood (or other) sample taken from a person or obtained from the crime scene. It is common to build a DNA profile using 10 STR loci (20 alleles). Therefore, when (for example) ten loci are used, it is extremely improbable 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 [3].

During laboratory data generation, the forensic scientist conducts experiments to transform these unknown DNA lengths into observable data [4]. This process has 3 main steps: 1) Perform polymerase chain reaction ("PCR") amplification on the DNA sample to transform the STR lengths into PCR products.2) Size separates the

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 242 – 251, 2005.

amplified PCR products on a DNA sequencer to form electrophoretic bands (two bands per loci one band for each allele). The locations of these bands are related to their size.3) Detect the bands to acquire data. Each band in loci has a number, related to its side; therefore we obtain a pair of numbers per loci, to build at last de DNA profiles per samples.

There are two chemical techniques in order to take to end the two last steps [5], one using the Capillary Electrophoresis Analysis, and the other applying Electrophoresis on Polyacrilamide Gels with tintion reagents. The first is a very expensive technique, and no many laboratories have the possibilities to apply it. An automatic module for the data processing based on signal processing accompanies the system. The second is a more common analysis, as an output, is obtained DNA sequencers in the form of electrophoretic bands on a Polyacrilamide Gel plate, the bands are visualized with a tintion reagent, one of them is the silver tintion reagent, and in this case we detect the DNA bands as black spots.

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 tedious, inefficient and inhuman form to do the task if we have under consideration that only one plate can contain more than 32 samples, plus 12 loci, plus 2 alleles per loci, 768 measurements are necessary to obtain the correspondent profiles.

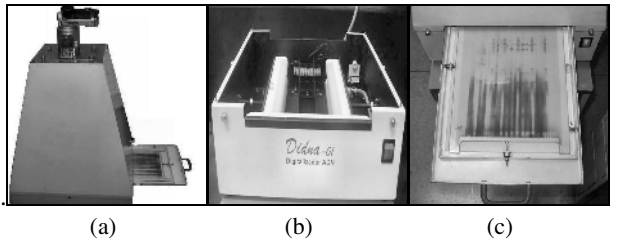
In this article an automatic solution is presented for DNA profile extraction in Polyacrilamide Gel Electrophoresis Images, integrating image processing, pattern recognition techniques and the associated image acquisition module.

2 Image Acquisition Module

To acquire the images a digital camera Sony DSC-F717 was place on a controllable illumination system. The Polyacrilamide gel plate is placed in a mobile gate between a diffuser plate and the digital camera and the light sources are in the bottom, below the diffuser plate. Figure 1, shows a view of different parts of the module.

The light sources are conformed by Leuci Lamps 8 watt cool-white 4500 ^{0}K.To obtains a uniform illumination in the acquire images, the fulfillment of the equation (1), is necessary [6]:

$$E=(I_i \cos\lambda_i)*r_i^{-1} \quad (1)$$



(a) (b) (c)

Fig. 1. Acquisition module: (a) General view, (b) Light Sources, (c) Mobile gate

Where E is the light emission of set light sources, I_i the Light intensity, λ_i the angle between the direction of the luminous flux and the normal to the surface and r_i the distance to the surface.

This condition guarantees that the dimension of the luminous bundle emitted for the source of illumination is little in relation to the distance that separates the diffusion plate from the lamps, and as a result a uniform illumination is obtained.

3 Image Preprocessing

Data artefact can be introduced at every step of the data generation process. There are dozens potential artefacts, some include: Low-level intensity spots, contaminating DNA material, bands reflexion, shifts in the baseline, colour background of the gels and other size distortions. Some of them can be corrected at a preprocessing step in order to enhance the image quality [7].

First, the source image obtained from the acquisition system is RGB, but colour, does not give us any useful information; therefore a conversion to halftones is convenient.

One of the main tasks of preprocessing is the removal or reduction of noise. In order to find the best suited one for this kind of images some linear and non-linear filtering methods, and also filtering methods in the wavelet domain [8] were tested. The best results were obtained using a Homomorphic Filtering [9, 10]. In our case, this filter acts to reduce the low frequency multiplicative noise that it is produced as a result of a non homogeneity illumination or a non homogeneity developed chemical process.

The application of the Fourier transform to the logarithm of the image, gives:

$$F\{ln\, I(x, y)\} = F\{ln\, L(x, y)\} + F\{lnR(x, y)\}. \tag{2}$$

Where L is the luminance and R the reflectance. This can be written as the sum of two functions in the frequency domain as:

$$Z(u, v) = F_L(u, v) + F_R(u, v). \tag{3}$$

F_R is composed of mostly high frequency components and F_L of mostly low frequency components. Z can be convolved with a filter of transfer function $H(u, v)$ that reduces the low frequencies and amplifies high frequencies, thus improving contrast and compressing dynamic range,

$$H(u,v).Z(u,v) = H(u,v).F_L(u,v) + H(u,v).F_R(u,v). \tag{4}$$

The processed image can be found by inverse Fourier transforming the previous equation and taking the exponential,

$$I'(x, y) = e^{F^{-1}}\{[H(u, v).Z(u, v)]\}. \tag{5}$$

Next step contemplates the process of spots segmentation. In order to carry out this task, we apply a Sobel Edge Detector; it uses a special mask [11] to approximate digitally the first derivatives Gx and Gy. In other words, the gradient at the center point in a neighbourhood is computed as follows:

$$g = [Gx^2 + Gy^2]\ ½ \tag{6}$$

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

Where z_1.....z_9 conform the image neighbourhood.. Then we say that a pixel at location(x, y) is an edge pixel if $g \geq T$ at that location, where T is a specified threshold.

The segmentation process finished applying automatically a Global Threshold following the iterative procedure proposed by González and Woods [11].

4 Feature Selection

Once finished the spot's segmentation, the next step is to represent and describe them in a form suitable for further computer processing. A representation using 14 boundary and region descriptors was chosen: Area, Complementary area, Perimeter, Rectified perimeter, Compacness, Maximum width, Maximum Height, 2-D moment invariants (φ_1, φ_2, φ_3, φ_4, φ_5), θ (angle of the principal axis), and Height-Width ratio. An automatic tool was developed to assign the descriptor at each spot in the image gel after segmentation.

To know which of these descriptors or features are the most significant to describe DNA spots, a data set formed of 4 images gels with more than 1890 spots were marked by an expert selecting only DNA spots according his experience, at last 965 DNA from the total of spots was marked . A C4.5 Decision tree [12] was used to do the task. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies the test of some feature of the instance, and each branch descending from that node corresponds to one of the possible values for this feature. An instance is classified by starting at the root node of the tree, testing the feature species by this node, then moving down the tree branch corresponding to the value of the feature in the given example. This process is then repeated for the sub tree rooted at the new node. The features that are situated in the roots nodes of the tree will be the most significant.

After the training, a decision tree with an effectiveness of 94.7% in the classification among DNA spots and No-DNA spots are obtained. The most significant features probe to be: φ_3, φ_1, φ_4, and area.

5 Classification

In our method all spots present on the polyacrilamide gel images, are described automatically, using the most significant features obtained in section 4. For the profile extraction only DNA spots are useful, therefore a two–class classification problem among DNA spots and No-DNA spots is necessary to solve. In order to realize the classification process with a high velocity, effectiveness, and robustness, a Support Vector Machine$_s$ classifier was selected.

Supports Vector Machines (SVM$_s$) are kernel based learning algorithm introduced by Vapnik [13, 14]. SVM$_s$ classifiers are introduced to solve two-class pattern recognition problems using the Structural Risk Minimization principle. Burges; Cristianini

& Shawe-Taylor [15, 16] worked given a training set in a vector space, SVM_s find the best decision hyperplane that separates two classes. The quality of a decision hyperplane is determined by the distance (i.e. hard or soft margin) between two hyperplanes defined by the support vectors. The best decision hyperplane is the one that maximizes this margin. The mapping to higher dimensional spaces is done using appropriate kernels such as Gaussian kernel and polynomial kernel [17]. SVM_s lend themselves well to accurate non-linear modelling and are very powerful and rapid learners. Good results in the application of SVM_s for different classification task of DNA were reported by Xu and Buckles [18].

In our case a non-linear SVM_s using a radial kernel offered the best results.

6 DNA's Profile Extraction

After the Classification process, an image with only DNA spots is obtained. For the profile extraction, first it is necessary to determine the regions in the image that contains the STR loci patterns, remember that usually we need twelve STR loci patterns in order to obtain the profiles. These patterns contain all the posibles alleles presents in a population and are possible to visualize them in the image as a sequence of DNA black spots for each STR loci. It is used to put the set of these 12 STR loci patterns more than one time in the plate intercalating the set each four samples investigated.

For the determination of these regions the first step is the detection of the candidate's regions according to the intensities histogram along the x axis. The second step is the determination inside of these regions of the periodic sequence of the image according the characteristics of the patterns. The third step is the validation of the results in correspondence with the data position given by the specialist and we finish assigning the coordinates at start and ending of the regions founded and it is marked in the image.

Using as reference the coordinates contributed by the patterns, the next step is the division of the image in lanes, each lane contains one sample or the set of STR loci patterns according to the distribution above mentioned. Normally we have more or less 32 lanes per image.

Inside the pattern's lanes we have different STR loci each of them have a specific sequence of spots always with the same quantity of spots, each of them have assigned a number, it is necessary the determination of these sub regions in the image each of them contains one STR loci with their spots. To solve this task a Sequential Leader algorithm was used [19]. It performs in two basic steps:

1. Chose a cluster threshold value.
2. For every new sample vector (DNA spot centroid that appears in patterns lanes):
   - -Compute the distance between the new vector and every cluster's codebook vector.
   - -If the distance between the closest codebook vector and the new vector is smaller than the chosen threshold, then recomputed the closest codebook vector with the new vector.
   - -Otherwise, make a new cluster with the new vector as its codebook vector.

Sometimes as a consequence of a malfunction of the classification algorithm, or by difficulties in the electrophoresis chemical process, one or more spots inside a sequence of a STR loci pattern were missing and in other cases two of them join up. A

restoration of the sequence of spots in the pattern is essential in order to obtain, in next steps, the DNA profiles of samples .To restore the missing spots a new algorithm was developed, which can be described as follows:

```
1) Comment: In the initial conditions the clustered spots in the STR LOCI PATTERN
           sequence are DNA spots and all spots for this analysis are in the same
           lane, therefore for all clustered spots the value of Cluster.Spot.DNA is
           true
2) Comment: Sort the clustered DNA spots in ascendant order by 'y' coordinate value
           of its centroide.
3) Cluster. Sort ();
4) Comment: We denote the clustered spots as spotc, and the others as spot
5) for (all clustered spots) do
6)   Comment: There is a hole (DNA spot not present)
7)   if (distance (spotc[i].centroid. y), spotc[i+1].centroid. y) ≥ threshold) do begin
8)     Comment (Case 1): At this point when a spot is not clustered it means that is
                         not DNA, therefore we want to know which spots not
                         clustered are between spotc[i] and spotc[i+1]
9)     for (all pair (spot[j], spot[k]), not clustered)
10)      Comment: Case 1: There is a spot divided in two neighbouring spots
                  (spot[j] & spot[k]) situated between (spotc[i] & spotc[i+1])
11)      if (((spot[j].centroid. y > spotc[i].centroid. y) &
            (spot[k].centroid. y > spotc[i ].centroid. y) &
           ((spot[j].centroid. y < spotc[i+1].centroid. y) &
            (spot[k].centroid. y < spotc[i+1].centroid. y)))
12)        if ((Abs (spot[j].centroid. y- spot[k].centroid. y) <=2) & ((spot[j].area +
             spot[k].area)>=area threshold)) do begin
13)          Comment: join the spot[j] & spot[k] into one and eliminate them
14)          Cluster. Add (newElement (spot[j], spot[k]));
15)          Cluster. Sort ();
16)          Delete (spot[j], spot[k]);
17)          end;
18)    for (all not clustered spot)
19)      Comment (Case 2): There are some spots (spot[j]) between spotc[i] &
                           spotc[ i+1] that are not divided (NOT Case1), but they
                           are near enough of spotc[i], in this case the solution is
                           adding to the cluster the most similar of all of them
20)      if ((spot[j].centroid. y > spotc[i].centroid. y) &
            (spot[j].centroid. y < spotc[i+1].centroid. y) &
            (distance (spot[j].centroid. y), spotc[i].centroid. y) < threshold)) do
            begin
21)           Distance[j] =EuclideanDistance between spot[j].featurevector and
                           spotc[i].featurevector]
22)           if ( Distance[j] < distance threshold) do begin
23)             K=index of the minimal Distance[j]
24)             Cluster. add (newElement(spot[k]))
25)             Cluster. Sort ();
26)             end;
```

```
27)                end;
28)     end;
29) Comment: if the restoration of STR Loci Pattern was not completely possible
30)  if (Cluster. count < total )
      ErrorMessage ("STR LOCI NOT COMPLETED")
```

The joined spots are separated by means of the detection of Freeman's chain typical segments of the contour [20], for example: 033332…03332…2110…21110…, the calculation of the horizontal dividing halfback line among them, permits an effective separation.

The final step is to assign the corresponded number to the spots that represents the two alleles per STR loci to conform the DNA profile of each sample (24 pair of numbers are obtained per 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 number, that is specific for each STR loci pattern, 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 distance from the centroid of each spots, (alleles), to the lines of the patterns spots nearest to them. The number assigned to the alleles are the same assigned to the lines of the patterns whose distances are the minors to them.

7 Data Set

A set of 20 DNA polyacrilamide gel electrophoresis plates, containing 200 real samples investigated by the National Forensic Laboratory of Cuba were used for the experimentation. The Plates have been directly recorded with the acquisition module, and the images obtained were automatically store in the computer for the process.

8 System Implementation

For the preprocessing step, we used software in C# based on the algorithms and procedures proposed by Rafael Gonzalez [11].. The feature selection using the Decision Tree C4.5 was implemented by the pack of classes that offers Software WEKA [12] 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 our method. Classifications with SVMs were done using SVM. NET Version 0.8b[21].

9 Results and Discussion

The SVM_s were training to classify the spots obtained after the electrophoresis process on the gel in DNA and No-DNA spots. For training the same data set used for the feature selection was employed, for testing we used the data set explained in point 7.

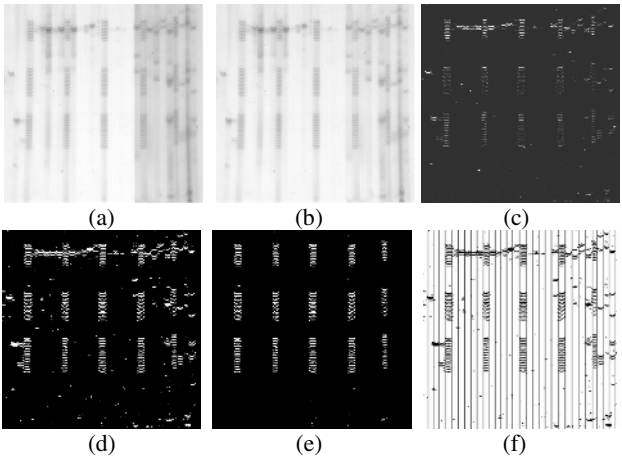
The classification accuracy was calculated by taken the number of correctly classified spots by the SVMs, and divided by the total number of samples into the test data set. Table 1 shows the results obtained in the classification.

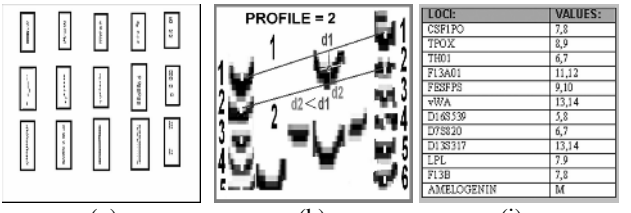
Table 1. DNA spot classification

Type of spots	#of spots	Confusion matrix		Classification
		ADN	NoADN	
ADN	2019	**1997**	22	
NoADN	4201	101	**4100**	
Total	**6220**			**98.02%**

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, only 30 sec. fundamentally because their structure is automatically determined on the basis of the training data and relatively few parameters are needed; in the other hand training involves optimisation of a function that relates to a quadratic convex programming problem, hence generating a completely reproducible solution (a major drawback of Neural Networks); overfitting can be avoided without using a validation set.

The set of the original plates, were processed by the expert using the standardized manual procedure and the results of the profile extraction were compare with the results obtained applying the automatic method taking into account the success rate and the time of response. Table 2 shows the results obtained in this comparison.





(g) (h) (i)

Fig. 2. (a) Original image, (b) Homomorphic filtering, (c) Sobel edge detector, (d) Global threshold, (e) STR loci patterns regions, (f) Division in lanes, (g) Determination sub regions, (h) Assigning number to alleles, (i) DNA profile

Table 2. 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

Added to the previous tables only 5 profiles was not possible to extract, 4 caused by the presence of mix samples (DNA of two persons are present in the same sample) with 4 different alleles present in each STR Loci causing that it is not possible to determine, which of the 6 pairs of alleles is the correct by the automatic method. The other one was caused by misclassification errors in the lanes corresponded to the samples, given that the misclassification errors in the lane of the patterns are restored by the algorithm developed for this purpose.

Another significant result is the decrease in the time's response of the task that influences not only in the increase of the available time of the expert but also in the decrease of the cost of the analysis. Fig.2 (a-i) shows a set of images representatives of all the process.

10 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, 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. Gill, P. Urquhart, A., Millican E., Oldroyd, N., Watson, S. Sparkers.: Criminal intelligence Databases and interpretation of STR_s Advances in Forensic Haemogenetics, 1996; 6:235-42.
2. Lander, E.S.: DNA fingerprinting: The NRC report, Science, vol.260, pp 1221. (1993).
3. Lewontin, R.C., Hartl, D.L.: Population genetics in forensic DNA typing, Science, vol. 254, pp. 1745-1750. (1991).
4. Weber, J., May, P.: Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am.. J. Hum. Genet. 1989:44; 388-96.
5. Estrada, C.: Techniques for DNA analysis in forensic genetics. http//: www.ugr.es/~ eianez biotecnología/forensetec.htm#1.(2001).
6. Shortley, G.. Dudly, W. Elements of Physics. B.E.E, Third Ed. (1966), Chap.24 Illumination and Photometry, pp 506.
7. Kacmazmarek, B.Walczak, B., Jong, S. Vandeginste, B.G.M.: Preprocessing of 2-D gel electrophoresis images, Analytical Chemistry, 75 (2003) 3631-3636.
8. Kacmazmarek, B.Walczak B., Jong, S. Vandeginste, B.G.M: Enhancement of images from 2-D gel electrophoresis. Proceedings 9th International Conference, CAC 2004.pp.171.
9. Stockham, T.G.: Image processing in the context of a Visual Model. Proc, IEEE, vol.60, No. 7, pp 828-842, (1972).
10. Short, J., Kittler, J., Messer, K. A comparison of photometric normalization algorithms for face verification. Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition.(FGR'04) 2004.
11. Gonzalez, R., Woods, R. "Digital Image Processing using MATLAB" Prentice Hall, Second Ed. (2004), pp 385-387.
12. Quinlan, R. J. C4.5: Programs for Machine Learnig (Morgan Kaufmann Series in Machine Learning). Paperback- January 15, (1993).
13. Vapnik, V., Chervonenkis, A.: Theory of Pattern Recognition. Nauka, Moscow, (1974).
14. Vapnik, V.: The nature of Statistical Learning Theory.. New York: Springer Verlag (1995).
15. Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2): 121-167 (1998).
16. Cristianini., Shawe-Taylor, J.: An introduction to Support Vector Machine. Cambridge University Press. (2000).
17. Scholkopf, C., Burges, J., Smola, A.: Advances in Kernel methods: Support Vector Learning. MIT. Press. (1999).
18. Xu, Z., Buckles, B.: DNA Sequence Classification by using Support Vector Machine. EECS, Tulane University..
19. Hartigan J.: "Clustering Algorithm". John Wiley and Sons. New York, (1975)
20. Alvarez, A. Ruiz J., Sanchiz, M.: Typical Segment Descriptors: A new method for shape description and classification. LNCS 2905, pp. 512–520, (2003).
21. Ching-Huei, Tsou: A.NET Implementation of Support Vector Machine.IESL MIT Version 0.8b October 25, (2004).