Application of Chemometrics Tools for Automatic Classification and Profile Extraction of DNA Samples in Forensic Tasks

Isneri Talavera[1]*, Francisco Silva[1], Noslén Hernández[1], Ricardo González[1], Juan Palau[1], Marcia M.Castro Ferreira[2] italavera@cenatav.co.cu

[1]Advanced Technologies Applications Center. MINBAS. Havana, Cuba. [2]Instituto de Química, Universidade Estadual de Campinas (UNICAMP),13083-970, Campinas, SP, Brazil.

Keywords: Classification, Support Vector Machine, Image Analysis

DNA profiling has attracted a good deal of public attention in recent years. The practical application of DNA technology to the identification of biological material has a significant impact on forensic biology, because it enables much stronger conclusions in genetic identity tasks[1]. During laboratory data generation, the forensic scientist conducts experiments to transform the biological samples into observable DNA data[2]. To carry out this transformation our forensic especialists apply Electrophoresis Analysis on Polyacrilamide Gels with silver tintion reagent, and the DNA sequences are visualized as black spots on the gels. There is a standardized method to manually detect spots of DNA and make the number designations, but it is a very tedious, and inefficient way to do the task. In this paper an automatic solution is presented which integrates image processing, and pattern recognition techniques.

After digital image acquisition and image preprocessing[3] the next step is the description of the spots. A representation using 14 boundary and regions descriptors was chosen. To find out which of them are the most significant to characterize DNA spots, a combination of a PCA analysis and a C4.5 Decision Tree were used. All the spots present on the polyacrilamide gel images were described automatically, using the most significant features obtained. For 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, 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[4]

After the Classification process, an image whith only DNA spots was obtained. For the profile extraction, first it was necessary to determine the regions and sub regions in the image that contain the DNA sequences patterns (STR loci). These patterns contain all the posibles alleles present in a population with especific numbers and it is possible to visualize them in the image as sequence of DNA black spots for each STR Loci. It is necessary to use a set of 12 STR Loci in order to be able to conform the profiles.To solve the task, the first step was the detection of the candidate's regions according to the intensities histogram along the x axis; second the determination inside of these regions of the periodic sequence of the image according the characteristics of the patterns and then we finished with the determination of the sub regions aplying a Sequential Cluster Leader Algorithm[5]. Sometimes, as a consequence of a malfuntion of the classifiction algorithm, or by dificulties in the electrophoresis chemical process, one or more spots inside a sequence of SRT Loci pattern were missing; to restore them a new algorithm was developed. The process finalizes with the profile extraction of the experimental DNA samples. To solve this task, the formula of distance of one point to a straight line was applied. The numbers assigned to the experimental spots were the same assigned to the lines of the patterns whose distance are the minors to them.

A set of original plates were 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. A success rate of 97% and a significant decrease in the time's response, indicated that this method has a very computational behavior, effectiveness, and provides a very useful tool to reduce time an increase the quality of the forensic specialist responses. A software which implements all of the method was developed.

[1] Gill U.; Millican E.; Oldroyd N.; Watson S.; Sparkers. Advances in Forensic Haemogenetics, 1996; 6:235
[2] Weber J.; May P. Am. J. Hum. Genet. 1989:44; 388-96.
[3] Silva F.;Talavera I.; González R.; Hernández N.; Palau J.; Santiesteban M. LNCS 3773, 2005; 242-251.
[4] Nianyi.; Wencong L.; Jie Y.; Guozheng L. Support Vector Machine in Chemistry. World Scientific Publishing Co. 2004
[5] Hartigan J. Clustering Algorithm. John Wliley and Sons. New York 1975.