Automatic Recognition System for Pejibaye palm DNA using SVM

CARLOS M. TRAVIESO[1], JUAN C. BRICENO[2], JOSE L. VASQUEZ[3], JAVIER VASQUEZ[2], ELENA CASTILLO[4]
[1] Department of Signals and Communications, University of Las Palmas de Gran Canaria, Campus de Tafira, E-35017, (Pabellón B – Office 111) Las Palmas de G.C., SPAIN
[2] Department of Computer Science, Ciudad Universitaria Rodrigo Facio,2500, University of Costa Rica, COSTA RICA
[3] Sede del Atlántico, University of Costa Rica, COSTA RICA
[4] Research Center on Molecular Biology, Ciudad Universitaria Rodrigo Facio, 2500, University of Costa Rica, COSTA RICA

Abstract: -This paper account for a correct DNA simplified sound, and inexpensive parameterization of pejibaye palm races, and a SVM automatic correct classifier obtaining 100% classes identification. The biochemical proposed parameterization, based on 89 RAPD primer markers applied on haplotypes of pejibaye races, has been proved correct by the classifier rate success. Previous result leads to a substantial description reduction over the DNA total chain description. Finally the interesting combination of this techniques (biochemical and computational), gives rise to a formulation of an inexpensive and handy method of origin denomination plant certification.

Key-Words: - SVM, DNA, RAPD, Pejibaye, plant, certification.

1 Introduction

The pejibaye palm belongs to the monocotyledons, family of the Arecaceae, tribe of the cocoids, sub tribe Bactridinae and Bactris genus [9]. This palm is the only domesticated one of the neotropic and produce: fruit, wood, and the most common and know heart-of-palm "palmito" present on international markets. This palm presents a large variety of morphology genus and large distribution over Central and South America. Since last century, due to the crop origin controversy [2][12] till now unsolved, mayor concern has been to identify biologically, domestic races and the research has been aimed to obtain genetic improvement and preservation instead of varieties identification. Till now, there is not known literature on an automatic pejibaye identification system. Economically, because different "landraces" (varieties), promote more or less one or other product and, in order to obtain origin denominations, there is an evident interest to correctly certify each one of different seed varieties.

For this study we considered six landraces pejibaye palms: Utilitis (Costa Rica), Tuira (Panamá), Putumayo (Colombia), Yurimagua (Perú), Tembé (Bolivia) and Pará (Brasil). Selected criterion considered races proponed by Clement and Mora-Urpi [2][10][11][12]. Such races have enough general representation on the germ plasma banc and were previously characterized by morphological markers [13]. Original population considered 191 palms with 18 to 10 individuals per race mean, evaluated with the RAPD technique.

On this study we have obtained three important results. First a validation of RAPDS (Random Amplified polymorphic DNA) traces analysis technique, obtaining an inexpensive straight forward method to correct pejibaye palm parameterization of DNA chains, and obtaining similar grouping on selected landraces than morphological methods. Second a substantial reduction of parameters to account for, and therefore concluding in a real time system response. And finally a 100% correct identification of each palm variety.

2 Pejibaye Palm Database

The germ plasma banc of the University of Costa Rica has been stabilized about 30 years ago and account for more than 1200 different introductions of pejibaye palms from Central and South America, becoming one of the most World wide completed.

In this present work, we have used a database win 6 classes of pejibaye (Utilitis - Costa Rica, Tuira - Panama, Putumayo - Colombia, Yurimagua - Peru, Tembé - Bolivia and Pará - Brazil), and each one has 13 samples with 89 RAPS primer marks per sample.

3 DNA Parameterization

Deoxyribonucleic acid (DNA is a long polymer of nucleotides, with a backbone made of sugars and phosphate groups joined by ester bonds. Attached to each sugar is one of four types of bases molecules and, it is the sequence of these four bases along the backbone that encodes information. This code is read by copying stretches of DNA into the related nucleic acid RNA.

Raw DNA analysis is a very expensive and time consuming technique but, the interest of such analysis is based on the fact that it is used on decision making, handled and preservation of genetic resources, taxonomy and systematic molecular studies.

Several techniques have been developed in order to diminish this description extension. RAPDS trace analysis (Random Amplified polymorphic DNA) is one of those finger printing technique based on PCR (Polymerase Chain Reaction) [5][6][13] [14] [15][16]. See figure 1.

This study was realized over each individual's genetic material, with 89 OPC primers (from the Operon Company) obtaining information variables with clear and well defined fragments, after multiples reactions amplifications. That is, for each individual, a 89 long parameter binary description vector. That is to say, markers and individuals produced a binary matrix, indicating enough presence of a particular RAPDS primer, from the six different pejibaye races considered.

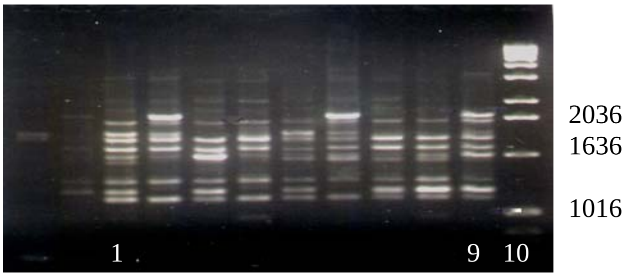

Figure 1. Some examples of Utilitis-Tucurrique pejibaye amplified DNA description, through the application of the PCR OPC-20 primer, with the RAPDS

On Figure 1, we see some examples of Utilitis-Tucurrique pejibayes amplified DNA description, through the application of the PCR OPC-20 primer, with the RAPD technique. From left to right Columns 1 to 9: samples of the amplified DNA. Column 10 shows the molecular weight marker (1Kb).

4 Classification Systems

On this research we have used two different types of classifiers in order to evaluate and analyze performance and behavior face to DNA parameterization of pejibaye palms. Used classifiers were Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

4.1 Support Vector Machines (SVM)

For the classification system based on the SVM [7][8], in order to establishing efficiency, we have calculated error, success and rejected rates on recognition.

Particularly, we have used an implementation of Vapnik`s Support Vector Machine known as SVM light[7][8] which is a fast optimization algorithm for pattern recognition, regression problem, and learning retrieval functions from unobtrusive feedback to propose a ranking function. The optimization algorithms used in SVM light are described in [7][8]. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently.

In the next figure, we can see the detection of support vectors and the creation of a boundary, one per each class, because it is a bi-class classifier. In our implementation, we have built a multi-classes classification module, from this SVM light.

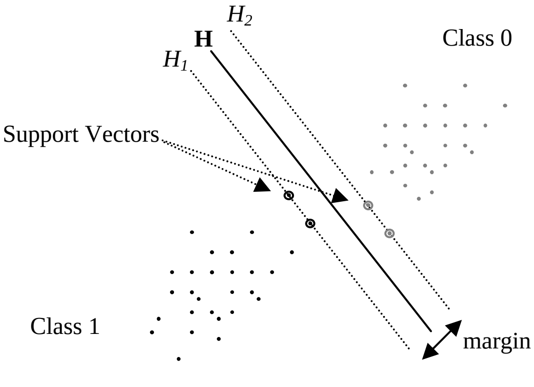

Figure 2. Separate lineal Hyperplane in SVM

4.2. Artificial Neural Networks

In recent years several classification systems have been implemented using classifying techniques, such as Neural Networks. The widely used Neural Networks techniques are much known on applications of pattern recognition.

The perceptron of a simple layer establishes its correspondence with a rule of discrimination between classes, based on the lineal discriminant. However, it is possible to define discriminations for not lineally separable classes using multilayer perceptrons that are networks without refreshing (feed-forward) with one or more layers of nodes between the input layer and exit layer. These additional layers contain hidden neurons or nodes, are directly connected to the input and output layer [1] [3] [3].

A neural network multilayer perceptron (NN-MLP) of three layers is shown on figure 3, with two layers of hidden neurons. Each neuron is associated with weights and biases. These weights and biases are set to each connections of the network and, are obtained from training in order to make their values suitable for the classification task between the different classes.

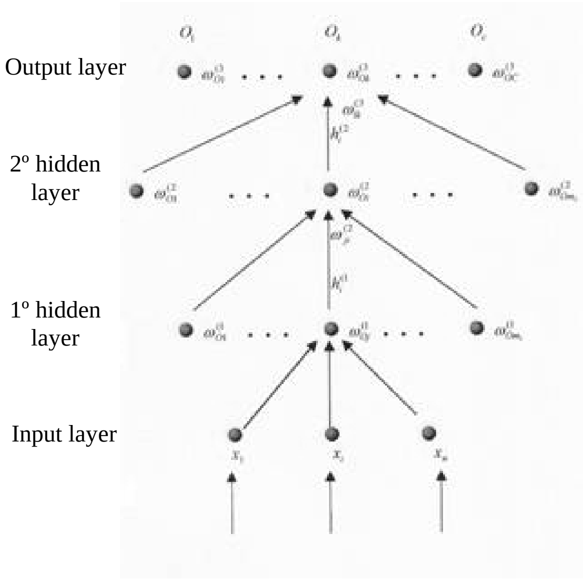

Figure 3. Multilayer Perceptron

5 Experiments and results

It has been developed a supervised recognition system, using the crossing validation method. We have used between 8% and 23% (from 1 to 3 samples/class) from our database in order to do training process, and the rest of them (from 92% - 12 samples/class to 77% - 10 samples/class) for test processing. Experiments have been done 10 times, and therefore, the success rate is shown with mean and variance.

Two different classifiers have been used for these experiments. We have searched the best configuration for both classifiers and on tables 1 and 2 best results are shown. In table 1, it can be observed the success rates of SVM classifier. SVM produces a binary classification; therefore, we have built a multi-class classifier for each one of the 6 different classes. The technique implemented is one-versus-all Multiclass SVM. Two different kernels have been used: Radial Basis Function RBF and Lineal kernel, and too, it is shown the maximum separation margin between the two classes known as gamma (γ), γ shows the exponential variable of RBF kernel.

This is an important parameter for a good configuration of SVM classifier.

In table 2, it is shown the success rates for ANN classifier. For our experiments, we have used a Multilayer Perceptron (MLP) Feed-Forward with Back-Propagation training algorithm, and with only one hidden layer.

The number of input neurons fits in with the number of DNA elements, and the number of output neurons with the Pejibaye palms races

Experimentation has been conducted with different number of neurons for the hidden layer, from 1 to 60 neurons. The best results are shown in table 2.

Table 1. Results with the SVM classifier

Training Percentage	Success Rates	Type of kernel	γ
8%	98.89% ± 0.34	Lineal	---
15%	99.85% ± 0.23		
23%	100% ± 0		
8%	99.02% ± 0.88	RBF	5×10^{-2}
15%	100% ± 0		8×10^{-2}
23%	100% ± 0		8×10^{-2}

Table 2. Results with the ANN classifier

Training Percentage	Success Rates	Number of neurons (hidden layer)
8%	61,43% ± 15.57	8
15%	72.14% ± 10.69	6
23%	66.74% ± 7.24	9

Table 3. Computational cost of proposed system.

Kind of classifier	Computational Load (milliseconds)	
	Training Mode	Test Mode
SVM	7012 ms	>0.1 ms per sample
ANN	6810 ms	>0.1 ms per sample

Results observation on those tables shows SVM as a better fit technique than the NN, showing also best fit results for binary coding.

Also, noisiest elements are barely influent for the SVM classifier and heavily for the NN. Despite the fact of using non lineal classification in both cases, the SVM adapt better for this type of parameter using the one against all structure.

It is interesting to observe the low system computational cost in both training and testing modes.

Table 3 shows, in milliseconds, reached values. These times indicate only training and testing performances and not parameterization time costs are considered. Shown training time, includes all training samples and, for testing mode, only consider one sample time processing. The system was implemented in Mat Lab and, time results are mean time for must samples training (3 individuals). Considering time consuming in test mode, this application may be considered real time performing.

6 Conclusions

In this paper we present a robust well performing system and innovative parameterization for automatic identification of haplotypes RAPD of pejibaye races, using a SVM classifier.

We have verified that the use of that classifier offers better guaranties than the NN when establishing the system.

On the other hand, the proposed approach on primers and RAPDs markers selection of the pejibaye DNA, among thousands and thousands elements of the DNA chain description joint with the SVM classifier, constitute a useful system for the scientific molecular biological community for automatic identification and origin denomination certification. The produced model reduces the need to evaluate only 89 primers without quality lost.

Considering economical savings, as well as system real time performance, the present work has a particular biological interest.

7 Acknowledgment

This work was supported by the Spanish Government with program "*Programa José Castillejo*" in 2007/08 and the University of Costa Rica.

References:

[1] B.H. Juang & L.R. Rabiner, "Spectral representations for speech recognition by neural networks-a tutorial", *Proceedings of the Workshop Neural Networks for Signal Processing*, 1992, pp. 214 – 222.

[2] Clement, C.R., J. Aguiar, D.B. Arkcoll, J. Firmino and R. Leandro. "Pupunha brava (Bactris dahlgreniana Glassman): progenitora da pupunha (Bactris gasipaes H.B.K.)" *Boletim do Museu Paraense Emilio Goeldi, Botánica* 1989 5(1) pp 39-55.

[3] Christopher M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[4] D.R. Hush & B.G. Horne, "Progress in supervised neural networks", *IEEE Signal Processing Magazine*, 10(1), 1993, pp. 8-39.

[5] Dellaporta, S. L., J. Wood y J. B. Hick**.** Plant DNA minipreparation. Version II*: Plant Mol. Biol.* Rep. 1, 1983. pp 19-21.

[6] Ferrer, M., Eguiarte, L.E. & C. Montana. "Genetic structure and outcrossing rates in Flourensia cernua (Asteraceae) growing at different densities in the South-western Chihuahuan Desert". *Annals of Botany* 94, 2004. pp 419–426,

[7] G. Cui, G. Feng & S. Shan, "Face Recognition Based on Support Vector Method", *Proceedings of 5th Asian Conference on Computer Vision*, 2002, pp 23-28.

[8] G. Guo, S. Z. Li, , and C. Kapluk. "Face recognition by support vector machines". *Image and Vision Computing*, 19(9- 10), 2001, pp. 631–638.

[9] Henderson, A. Bactris (Palmae). *Flora Neotropica Monograph* 79, 2000. pp: 1-181.

[10] Mora-Urpí, J. 1993. "Diversidad Genética en Pejibaye: II. Origen y Evolución". IV *Congreso Internacional sobre Biología, Agronomía e Industrialización del Pijuayo. Universidad de Costa Rica*. 1993. pp. 21-29.

[11] Mora-Urpí, J. y C. Arroyo. "Sobre origen y diversidad en pejibaye. Serie Técnica Pejibaye (Guilielma)". *Boletín Informativo. Editorial de la Universidad de Costa Rica*. 1996 5(1): 18-25.

[12] Mora-Urpí, J., C. Clement y V. Patiño. "Diversidad Genética en Pejibaye: I. Razas e Híbridos". IV *Congreso Internacional sobre Biología, Agronomía e Industrialización del Pijuayo. Universidad de Costa Rica*. 1993. p. 11-20.

[13] Mattos, L. “Diferenciación Taxonómica de Diez Razas de Pejibaye Cultivado (Bactris(Guilielma) Gasipaes Kunth) y su relación con otras Especies de Bactris”. *Magister Scientiae, Universidad de Costa Rica*. 1992. pp 197.

[14] Porebski, S., L. Grant y B. Baun. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* 15. 1997. pp: 8-15.

[15] Ravishankar, K. V., Anand L. and Dinesh M. R**.** “Assessment of genetic relatedness among mango cultivars of India using RAPD markers”. *Journal of Horticultural Sci. & Biotechnology*, 75, 2000. pp 198-201.

[16] Williams, J. G. K. “DNA polymorphisms amplified by arbitrary oligonucleotide primers are useful as genetics markers” *Nucleic Acids Research* 18, 1990, pp 6531-6535.