Chromosome Classification Using Continuous Hidden Markov Models*

César Martínez[1], Héctor García[2], Alfons Juan[3], and Francisco Casacuberta[3]

[1] Facultad de Ingeniería, Universidad Nacional de Entre Ríos
Ruta 11, Km. 10. CP 3100 Paraná, Entre Ríos (Argentina)
cmartinez@fi.uner.edu.ar

[2] Instituto Valenciano de Investigaciones Económicas (IVIE)
46020 Valencia (Spain)
hector.garcia@ivie.es

[3] DSIC/ITI, Universitat Politècnica de València
Camí de Vera s/n, E-46071 València (Spain)
{ajuan,fcn}@iti.upv.es

Abstract. Up-to-date results on the application of Markov models to chromosome analysis are presented. On the one hand, this means using *continuous Hidden Markov Models (HMMs)* instead of discrete models. On the other hand, this also means to conduct empirical tests on the same large chromosome datasets that are currently used to evaluate state-of-the-art classifiers. It is shown that the use of *continuous HMMs* allows to obtain error rates that are very close to those provided by the most accurate classifiers.

1 Introduction

A common task in cytogenetics is the *karyotye analysis of a cell*. It consists of labelling each chromosome of the cell with its class label, in order to have the genetic constitution of individuals. This analysis provides important information about number and shape of the chromosomes, which serves as a basis to study the possible abnormalities the individual could have.

In a normal, nucleated human cell there are 46 chromosomes. The *karyogram* is a standard format which shows the complete set organized into 22 classes (each one consisting of a matching pair of two *homologous* chromosomes), ordered by decreasing length, and two sex chromosomes, XX in females or XY in males.

The first attempts to automate this task were made in the early 1960s, motivated by the fact that manual analysis is very tedious and labour-intensive. Since then, many classification techniques have been tried, including both statistical and structural approaches. Most of them, however, are conventional, statistical

* Work supported by the Valencian "Oficina de Ciència i Tecnologia" under grant CTIDIA/2002/80, the Spanish "Ministerio de Ciencia y Tecnología" under grant TIC2000-1703-CO3-01 and the Argentinian "Ministerio de Educación de la Nación" under fellowship FOMEC 1108.

F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 494–501, 2003.

classification techniques in the sense that they reduce each chromosome image to a point in some multi-dimensional feature space [6]. Instead, the use of structurally richer approaches has been rare and focused mainly on *discrete Markov models* [3, 7].

The aim of this paper is to provide up-to-date results on the application of Markov models to chromosome analysis. On the one hand, this means using *continuous Hidden Markov Models (HMMs)* instead of discrete models. On the other hand, this also means to conduct empirical tests on the same large chromosome datasets that are currently used to evaluate state-of-the-art classifiers [6]. It is worth noting that this is a new application of standard Speech Recognition technology and, in particular, a new application of the well-known and widely available standard *HMM Tool Kit (HTK)* [8]. However, this is not a straightforward application of HTK since we also have to take care of preprocessing, feature extraction and HMM topology design. These aspects are covered in the next section. Empirical results and the main conclusions drawn are given in sections 3 and 4, respectively.

2 The Approach

The basic steps of our HTK-based approach are illustrated in Figure 1. They are described in what follows.

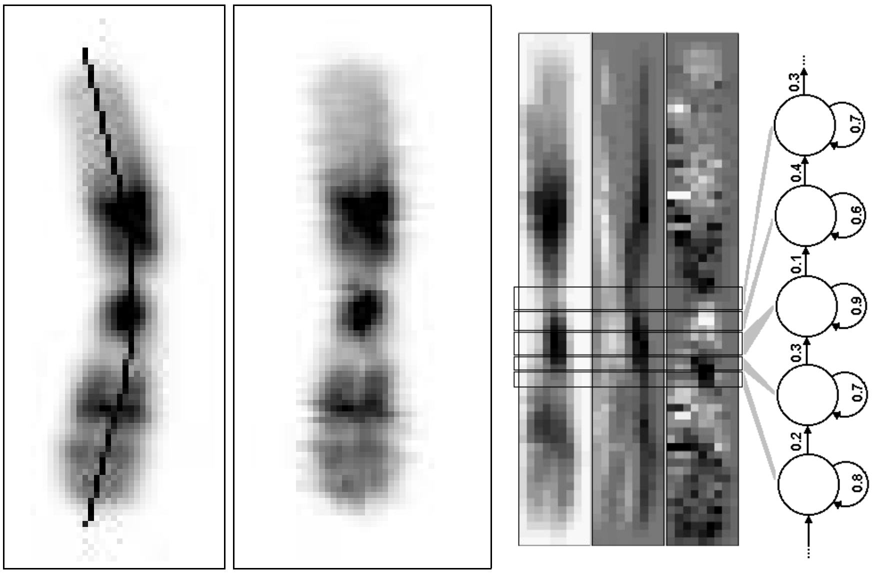

Fig. 1. Basic steps of our HTK-based approach. From left to right: computation of the longitudinal axis, unfolding, feature extraction and HMM modelling

Computation of the Longitudinal Axis and Unfolding

The computation of the longitudinal axis of a chromosome is a standard preprocessing step in chromosome analysis. However, no precise definition has been widely accepted and, in fact, it is a matter of current research [6]. In our case, we have used a rather standard procedure that includes the classical Hilditch's thinning algorithm for *medial axis* computation and some refinements [2].

Once the longitudinal axis has been computed, it is traversed at unit speed and a perpendicular slice is cut at each sampled axis point to obtain an unfolded, straight version of the chromosome. After this *chromosome unfolding*, feature extraction reduces to compute an appropriate set of features from each image row.

Feature Extraction

Feature extraction for *local* characterization of chromosomes is an interesting, open problem. Based on our previous experience [5] and some informal tests, we have considered the following four types of features:

- **9p:** grey densities
- **D:** horizontal derivative
- **A:** horizontal acceleration
- **V:** vertical derivative

The set of features referred to as 9p corresponds to 9 equidistant Gaussian-filtered points. Concretely, each point was filtered by convolution with a 5×5 filter mask with weights: 16 in the center, 2 at a 1-pixel distance and 1 at a 2-pixel distance. Derivatives and acceleration were computed from successive 9p vectors. As an example, the sequence of feature vectors shown Fig. 1 comprises grey densities plus horizontal and vertical derivatives (9p+D+V).

It must be noted that these types of features come from previous work on HTK-based handwriting recognition. Please see [1, 4] for more details on the computation of these types of features.

HMM Chromosome Modeling

As illustrated in Fig. 1, chromosomes are modelled by *continuous left-to-right HMMs*. Basically, an HMM for a chromosome class is a stochastic finite-state device aimed at modelling the succession, along the longitudinal axis, of feature vectors extracted from instances of the chromosome class. Each HMM state generates feature vectors following an adequate parametric probabilistic law; typically a *mixture of Gaussian densities*. The number of states and number of densities per state that are appropriate to model each chromosome class depend on the class variability and the amount of training data available. So, these numbers need some empirical tuning. The training process is carried out with the HTK toolkit, using conventional re-estimation formulas [8].

3 Experiments

The data set used in the experiments was the Cpa, a corrected version of the Cpr corpus (the complete *Copenhagen* data set) [6]. The corpus contains the segmented chromosome images of 2804 human cells, 1344 of which are female and 1460 are male.

3.1 Context-Free Classification

The usual method for classifying a test chromosome is to find the HMM with the highest probability in the Viterbi decoding. This is the so-called *context-free* classification, because each chromosome is classified with independence of each other. The experiments reported in this subsection were done under this context-free framework.

As discussed before, one of the two basic parameters characterising continuous left-to-right HMMs is the number of states chosen for each class-conditional HMM M_i. This number has been computed as $s_i = \frac{f_i}{k}$, where f_i is the average length of the sequence of vectors used to train M_i, and k is a design parameter measuring the average number of feature vectors modelled per state. This rule of setting up s_i attempts to balance modelling effort across states and also captures important discriminative information about the typical length of each chromosome class. Following this rule, a first series of experiments was carried out on a partition involving $2400 + 400$ training and testing cells. We used the 9p feature set and the values of k: $1.5, 2, 2.5, 3$ and 4. For each value of k, the number of Gaussian densities per state was varied as $1, 2, 4, ldots$ until reaching a minimum classification error rate. The results obtained, which are omitted here for brevity, showed a degradation of the error rate as the value of k increases. So, in accordance with these results, a value of $k = 1.5$ was fixed for the remaining experiments.

After deciding on the value of k, a second series of experiments was conducted on the same data partition to study the classifier performance as a function of the number of Gaussian densities per state, and for several feature sets. The results are shown in Fig. 2. From these results, it is clear that an appropriate feature set consist of using grey densities plus horizontal and vertical derivatives (9p+D+V). Also, 64 seems to be an adequate number of Gaussian densities per state.

Although the results obtained up to this point were satisfactory, we decided to do further experiments using windows of feature vectors instead of single vectors. This was done as as an attempt to reproduce the behaviour of the input layer of Recurrent Neural Networks, in which a small moving-window of feature frames is processed at each time and this seems to be very effective in improving classification results [5]. Fig. 3 shows the classification error rate (estimated as before) for varying Gaussian densities per state and several window sizes. As expected, the use of windows of feature vectors helps in improving classification error and, in fact, the best result (5.6%) was obtained with a window of 3 feature vectors (and 64 Gaussian densities per state).

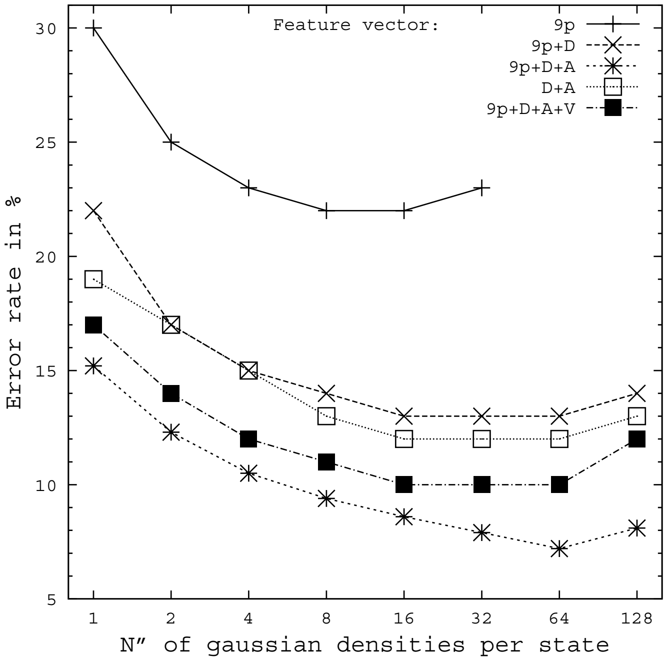

Fig. 2. Classification error rate as a function of the number of Gaussian densities per state, and for several feature sets

All the experiments reported so far were carried out using a single partition of the complete corpus. In order to obtain more precise results, the classification error rate was also estimated using a 7-fold cross-validation procedure in which the blocks were chosen to have 400 cells each. It is given in Table 3.1 as a function of the number of Gaussian densities per state (the remaining parameters were set to the values that provided a 5.6% of error). The best result for the cross-validation method, again obtained with 64 Gaussian densities, is 7.5%.

3.2 Context-Dependent Classification

The classification error rate can be reduced by taking into account the fact that the normal karyotype consists of 22 pairs of autosomes and a pair of sex chromosomes. This knowledge imposes a constraint that penalizes, e.g., the allocation of more than two chromosomes to one class and less than two chromosomes to other class. This is the called *context-dependent* classification.

An iterative algorithm was formulated to restrict the isolated chromosome classification by including the contextual cell information. The algorithm receives as input, for each chromosome of a cell, the output probabilities of each HMM. Then, in successive iterations, the algorithm classifies pairs of chromosomes for each class using *solved classes*; i.e., classes with only two chromosomes having

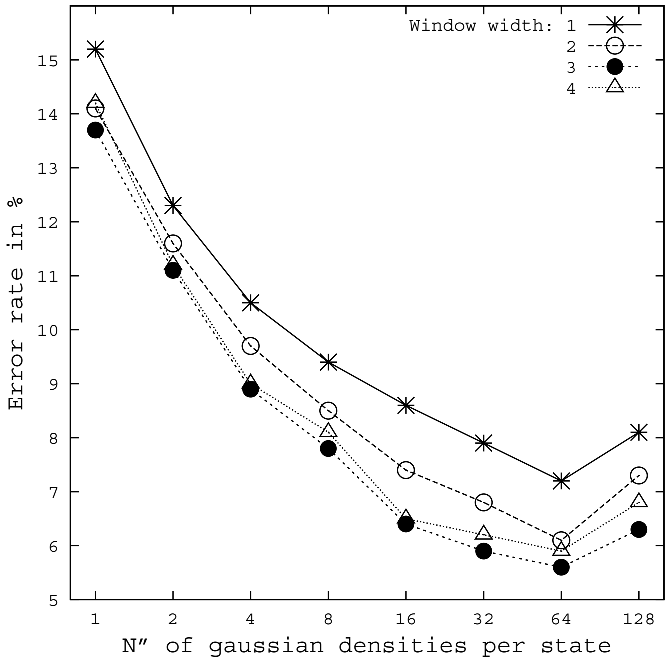

Fig. 3. Classification error rate as a function of the number of Gaussian densities per state, and for window widths

the highest probability for that class, and both probabilities greater than a lower bound. After that, the probabilities of solved classes are crossed-out for the remaining chromosomes, whose probabilities are renormalized and the process is repeated until the complete cell is classified.

In order to complete the experiments reported in the preceding subsection, the context-dependent classification algorithm discussed above was applied to the best classifier found under the context-free framework. The classification error rate was again estimated using the 7-fold cross-validation procedure and, as expected, it was reduced from 7.5% up to 4.6%. This figure is close to those provided by the most accurate classifiers [6].

4 Conclusions

We have provided up-to-date results on the application of Markov models to chromosome analysis. It has been shown that the use of *continuous Hidden Markov Models (HMMs)* allows to obtain error rates that are very close to those provided by the most accurate classifiers. As in the case of handwriting recognition, the main advantage of using HMMs is that they can be easily integrated into systems based on finite-state technology [4].

Table 1. Classification error rate estimated using a 7-fold cross-validation procedure

N° of G. D.	error rate
1	15.7
2	12.5
4	10.4
8	9.3
16	8.4
32	7.9
64	7.5
128	7.7

A number of improvements can be applied to the entire system. The skeleton technique used to obtain the longitudinal axis has the disadvantage of being non-parametric, so approximations using eigenvectors have to be used to calculate the slices. A parametric axis could help to reduce some errors introduced by the current method due to morphological filtering and the skeletonization algorithm (especially in short chromosomes). In this line, other methods are being studied: polynomial curve-fitting, implicit polynomials, etc. The iterative algorithm that implements the *context-dependent* classification could be improved by dynamic programming, allowing a more detailed analysis for finding the *solved classes*.

Acknowledgements

The authors wish to thank Dr. Gunter Ritter (*Fakultät für Mathematik und Informatik, Universität Passau, Passau, Germany*) for kindly providing the database of chromosome images Cpa used in this work.

References

[1] J. Doménech, A. H. Toselli, A. Juan, E. Vidal, and F. Casacuberta. An off-line HTK-based OCR, system for isolated handwritten lowercase letters. In *Proc. of the IX Spanish Symposium on Pattern Recognition and Image Analysis,* volume II, pages 49-54, Benicàssim (Spain), May 2001. 496

[2] H. García. Preproceso y extracción de características (sintáctica) para el diseño de clasificadores de cromosomas humanos. Master's thesis, Faculty of Computer Science, Polytechnic University of Valencia, 1999. 496

[3] J. Gregor and M. G. Thomason. A Disagreement Count Scheme for Inference of Constrained Markov Networks. In L. Miclet and C. de la Higuera, editors, *Grammatical Inference: Learning Syntax from Sentences,* volume 1147 of *Lecture Notes in Computer Science,* pages 168-178. Springer, 1996. 495

[4] A. Juan et al. Integrated Handwriting Recognition and Interpretation via FiniteState Models. Technical Report ITI-ITE-01/1, Institut Tecnològic d'Informàtica, Valencia (Spain), July 2001. 496, 499

[5] César Martínez, Alfons Juan, and Francisco Casacuberta. Using Recurrent Neural Networks for Automatic Chromosome Classification. In *Proc. of the Int. Conf. on Artificial Neural Networks ICANN 2002,* volume 2415 of *Lecture Notes in Computer Science,* pages 565-570, Madrid (Spain), August 2002. Springer-Verlag. 496, 497

[6] G. Bitter and G. Schreib. Using dominant points and variants for profile extraction from chromosomes. *Pattern Recognition,* 34:923-938, 2001. 495, 496, 497, 499

[7] M. G. Thomason and E. Granum. Dynamic Programming Inference of Markov Networks from Finite Sets of Sample Strings. *IEEE Trans. on PAMI,* PAMI-8(4):491501, 1986. 495

[8] S.J. Young et al. HTK: Hidden Markov Model Toolkit. Technical report, Entropic Research Laboratories Inc., 1997. 495, 496