Hidden Markov Models for Chromosome Identification

John M. Conroy
Center for Computing Sciences
Institute for Defense Analyses
Bowie, Maryland

Robert L. Becker, Jr.
William Lefkowitz
Kewi L. Christopher
Rawatmal B. Surana
Timothy J. O'Leary
Armed Forces Institute of Pathology
Washington, DC

Dianne P. O'Leary
Computer Science Department
and Institute for Advanced Computer Studies
University of Maryland
College Park, MD

Tamara G. Kolda
Computational Science and Mathematics Research Department
Sandia National Laboratories
Livermore, CA

Abstract

In this talk we present a Hidden Markov Markov for automatic karyotyping. Previously, we demonstrated that this method is robust in the presence of different types of metaphase spreads, truncation of chromosomes, and minor chromosome abnormalities, and that it gives results superior to neural network on standard data sets. In this work we evaluate it on a data set consisting of a mix of chromosomes obtained from blood, amniotic fluid and bone marrow specimens. The method is shown to be robust on this mixed set of data as well as giving far superior results than that obtained by neural networks.

Technical areas: Signal and image processing in medicine; software systems in medicine.

0-7695-1004-3/01 $10.00 © 2001 IEEE

1. Introduction.

Automatic methods for chromosome karyotyping are of great interest in yielding a preliminary classification for G-banded chromosomes, making the final classification process much less tedious. The most popular automated techniques are based on neural networks (see, for example, [1, 2, 3]). In previous work [4] some of the authors have proposed the use of hidden Markov models (HMMs) for automatic karyotyping, and we demonstrated the effectiveness of the method on standard data sets: Philadelphia, with chromosomes taken from chorionic villus, and Edinburgh and Copenhagen, with chromosomes taken from blood. We ran three tests of the methods, training on half of the samples in each data set and then evaluating based on performance on the other half of the data. In each case, the HMM method gave higher accuracy than the neural network.

In this work, we continue our study of the HMM method, evaluating it on a data set collected at the Armed Forces Institute of Pathology, consisting of a mix of chromosomes obtained from blood, amniotic fluid and bone marrow specimens.

2. Data gathering.

Metaphase spreads were obtained from blood, amniotic fluid and bone marrow specimens that were sent for "routine" cytogenetic evaluation and appropriately cultured. Slides were stained using a trypsin-Giemsa technique [5]. Selected metaphases were then visualized using an Olympus Vanox microscope and digitized in gray scale using a Panasonic RS 170 digital camera. Although bent and overstained chromosomes were included in the image set, overlapping chromosomes were excluded. Following identification of each chromosome as autosome 1-22, X, or Y, axial densitometric traces of the chromosomes were obtained, giving sequences of gray levels.

The data were randomly spit into two pieces,, each containing about 2500 sample chromosomes. The HMMs and the neural network were trained on the first set and then tested on the second.

3. Methods.

Netmaker Professional for Windows and Brainmaker Professional for Windows (California Scientific Software, Nevada City, California) were used to generate **backpropagation neural networks** with an input layer of 15, 30 or 45 nodes, a single hidden layer of 200 nodes, and an output layer of 24 nodes. The input to the neural net was a sequence of 15 gray-level values for each chromosome, or the gray levels augmented by 15 first and 15 second differences.

The **hidden Markov model (HMM)** we used was identical to the one presented in [4]. We highlight a few features of the model here and refer the reader to the journal article for details.

For each chromosome type (the autosomes 1,2,...,22 and the sex chromosomes X and Y) we determine a HMM based on half of the available data. The model consists of hidden states whose number is determined by the median number of observations in the training data corresponding to that chromosome. Each state corresponds to the corresponding positions in the sequence of gray level values for an idealized chromosome of that type. Chromosomes to be karyotyped could be shorter or longer than the average length, or perhaps even truncated at either end, so the model is designed to allow states to be skipped or repeated in taking a walk though the chain. The observations, which are modelled as outputs of a probabilistic function of the hidden states, are the gray scale value, and possibly the first and second differences. The latter two give additional discrimination power, because, they indicate how quickly the gray scale value changes at a position, and the curvature of the gray scale sequence. A Gaussian mixture model is used to model this three-tuple of observations and we compute a mixture model for each position in the template for each chromosome type.

To classify an unknown chromosome we score it with each of the 24 HMMs and then interpret these 24 scores as a feature vector for input to a linear discriminant analysis classifier.

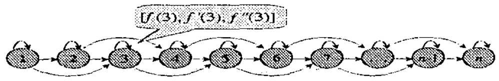

Figure 1: Markov Model for Karyotyping

4. Experimental Results.

After training the neural network and the HMM, we ran three experiments:

1. Evaluate using the chromosomes in the test set. (Recall that the test

	Normal	Truncated	Flipped
NN, no differences	.88	.51	.86
NN, 1 differences	.91	.51	.84
NN, 2 differences	.91	.51	.81
HMM, no differences	.90	.80	.81
HMM, 1 differences	.95	.91	.93
HMM, 2 differences	.96	.92	.93

Table 1: Comparing HMM and Neural Network

set was a random selection of half of the data and consisted of approximately 2500 chromosomes.) This experiment is designated by "Normal" in Table 1.

2. Truncate the last 10% of the data for each chromosome, simulating the identification of broken chromosomes or chromosome records that were improperly cut. This experiment is designated by "Truncated".

3. Take the gray level values in the middle 10% of the sequence for each chromosome and reverse their order, simulating an internal inversion of chromosomal material. This experiment is designated by "Flipped".

Table 1 gives the percent of the test chromosomes that were correctly identified by each method.

For further confidence the data were repartitioned into approximately equal pieces so that a patients chromosomes would never be split between the testing and training data. The HMM was run on the repartitioned data and the results agreed with the above within 2 digits of accuracy.

5. Conclusions.

We have demonstrated that the HMM method is a robust method for automatic karyotyping, robust in the presence of different types of metaphase spreads, truncation of chromosomes, and minor chromosome abnormalities, and that it gives results superior to neural networks.

References

[1] Graham J, Errington P, Jennings A. A neural network chromosome classifier. J.Radiat.Res.(Tokyo.) 1992;33 Suppl:250-7.

[2] Sweeney WPJ, Musavi MT, Guidi JN. Classification of chromosomes using a probabilistic neural network. Cytometry 1994;16(1):17-24.

[3] Errington PA, Graham J. Application of artificial neural networks to chromosome classification. Cytometry 1993;14(6):627-39.

[4] Conroy JM, Kolda TG, O'Leary DP, O'Leary TJ, Chromosome Identification Using Hidden Markov Modles: Comparison with Neural Network, Singular Value Decomposition, Principal Components Analysis, and Fisher Discriminant Analysis, Laboratory Investigation. 2000;80:1629-1641.

[5] Gustashaw, KM. Chromosome Stains, in The AGT Cytogenetics Laboratory Manual, Barch MJ, Knutsen T, and Spurbeck JL, Eds. Lippincott-Raven, Philadelphia, 1997:259-324.