A SEQUENTIAL DECISION PROCEDURE FOR COMPUTER IDENTIFICATION OF HUMAN CHROMOSOMES

by

Everett Morris Sherwood

A Thesis Submitted to the Faculty of the

DEPARTMENT OF SYSTEMS AND INDUSTRIAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE
WITH A MAJOR IN SYSTEMS ENGINEERING

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 7 6

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Everett M Sherwood

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Peter H. Bartels

PETER H. BARTELS
Professor of Microbiology and Medical Technology, and Optical Sciences

Nov. 2nd 1976

Date

ACKNOWLEDGMENTS

The author is pleased to express appreciation for the support extended by the Arizona Chapter of the American Heart Association.

The continuing sponsorship of George L. Wied, M.D., at the Department of Obstetrics and Gynecology, Chicago Lying-In Hospital, University of Chicago, greatly facilitated the development of this project.

Special thanks must go to John Davis, M.D., at the Department of Pathology, The University of Arizona. Dr. Davis provided both biological materials and an ongoing collaboration which were most essential.

I am deeply indebted to Dr. Peter H. Bartels. His guidance, his supervision, and his encouragement were the framework of the investigation.

Most of all, the love and understanding of my wife, Judith Ann Sherwood, sustained this work.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT

This paper considers the feasibility of a low-cost screening system for normal human chromosome identification. The paper discusses the acquisition of chromosome banding patterns for computer analysis. Included are descriptions of both the hardware and software necessary to interface a computer with an optical scanning microscope. An algorithm for contour tracing of chromosomes with the scanning stages is demonstrated.

The body of the work describes techniques for the pairing of chromosomes and for the identification of the chromosome pairs. A new procedure for making matrix assignments is defined. A recursive method for pre-processing of the matrix is demonstrated. Two methods for testing the new techniques are implemented.

Computer programs which implemented these methods are included in the Appendix.

REVIEW OF THE LITERATURE

Since the classic paper by Tjio and Levan (1956) which definitely established the normal number of human chromosomes, cytogeneticists have focused their research on the identification of human chromosomes. Several good books are available to the nonbiologist for an informative review of chromosomes and their biology. One of the most lucid of these is Watson's (1970) Molecular Biology of the Gene. For details specific to cytology and an introduction to the nomenclature, Chapters 2 and 3 of Cytogenetics by Priest (1969) serve very well. The paper by Ledley, Lubs, and Ruddle (1972) and the paper by Neurath, Kess, and Low (1972) define the scope and constraints of the problem and acquaint the reader with modern identification systems. For a brief historical introduction and for a summation of the usefulness of automated analysis of chromosomes, the reader is referred to the article by Hsu and Hirschhorn (1972).

The first attempts at chromosome identification were based on visual inspection of morphology (Fig. 1). Best known of these methods is the karyogram which was formalized at the Chicago Conference (1966). Indeed, the karyogram is still the evaluation method of choice. Its standardization provides the starting point for all further analysis. However, the visual inspection which produces karyograms is,

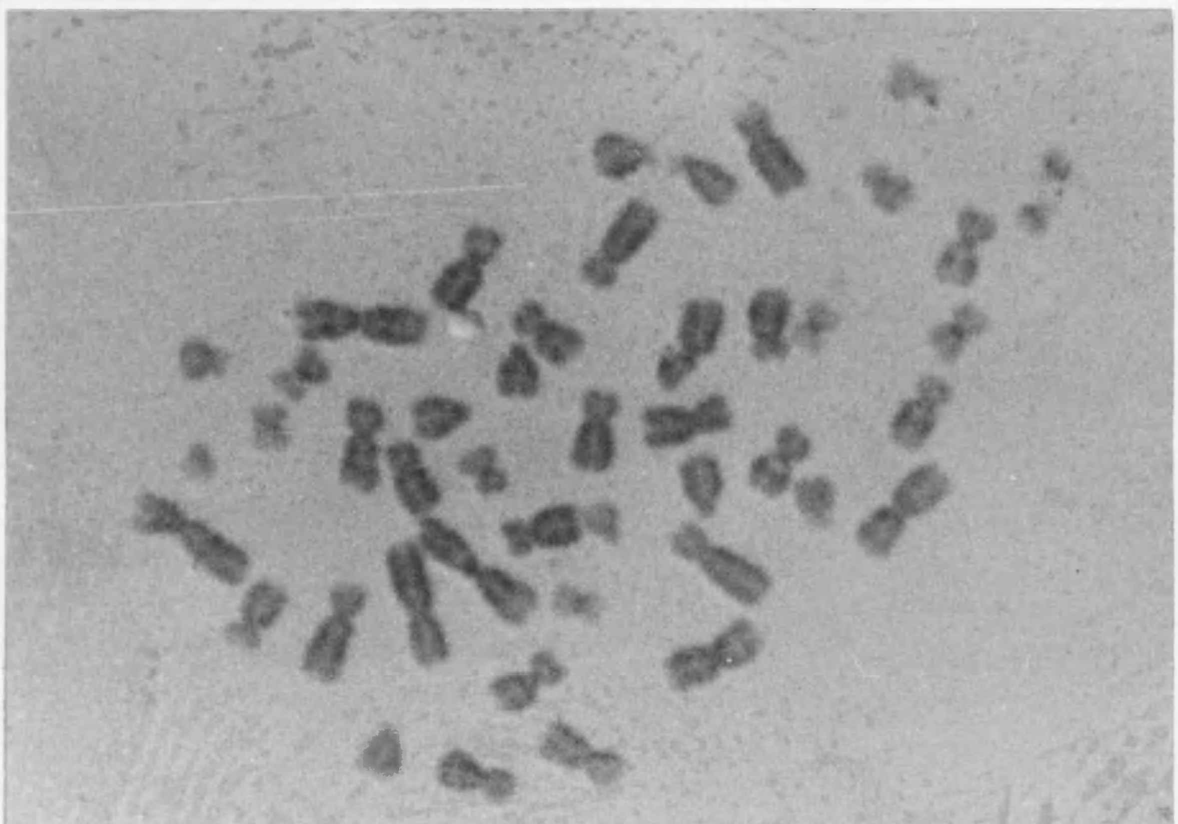


Fig. 1. Chromosome Spread -- This photograph of a human chromosome spread shows banding patterns brought out by Giemsa staining.

admittedly, imprecise and often open to interpretation. The karyogram is more a clustering by size and shape than a set of individual identifications. To provide better identifications researchers have turned to more precise instruments and sophisticated mathematics.

New electro-optical devices produce greater resolution at greater speeds for the cytogeneticist. Flying spot scanners, photomultiplier tubes, and image dissectors now provide the investigator with flexibility for trade-offs between data rates and data precision. Butler, Butler, and Stroud (1968) discuss the problems in instrumentation and in image processing. Butler et al. evaluate the use of photographic negatives. On a more comprehensive level Willard (1965) details the trade-offs between instrumental methods. Kazan and Knoll (1968) have produced a text which offers descriptions of both the theory and the application of modern optical-electrical energy transducers. As instrumentation has provided more information with greater precision, researchers have sought to evaluate this information with new algorithms.

Visual inspection has given way to measures of length and width. These measures, in turn, have been supplemented by measures of arm-length ratio and of centromeric index. "DNA Content and DNA Centromere Index of 24 Human Chromosomes" by McMorris et al. (1973) is among the more recent articles applying this description. This article is of particular interest because of the application to hybrid clones. A second application of these methods occurs in a paper by Enslein and Neurath (1969). Here the variables of index and length are similar, but the analysis is carried out by a version of the BMD programs developed by Dixon (1973). Further sophistication in the problem area of data acquisition is found in "Normalization of Chromosome Measurements" by Hilditch and Rutovitz (1972). This article contains an approach to the problem of uneven contraction of the DNA during preparation. On a more fundamental level this article supplies a geometric feature extraction scheme aimed specifically at chromosome analysis.

The contraction problem proves to be a great difficulty in the identification of chromosomes. An analogous problem is encountered in the analysis of electrocardiograms. It is worthwhile to read two articles concerned with the classification of electrocardiograms. The paper by Eddleman and Bancroft (1967) suggests a normalization technique. A second method by Klingeman and Pipberger (1967) points to the difficulty of forming classes.

Identification based on centromere indices, chromosome lengths, widths, and contours, employed increasingly complicated designs to deal with the difficulty posed by the contraction problem. A decision theory approach using constrained Bayesian classification was implemented by Low, Selles, and Neurath (1971). Two interesting articles by

Hilditch (1969a, 1969b) extend the geometrical approach offered by Neurath and others. This latter procedures reduces each chromosome to a linear "skeleton" and is based on work done in character recognition. Regrettably, the practical usefulness of this design has been found to be restricted by its sensitivity to the presence of noise.

While work continued on the identification of chromosomes on the basis of their geometrical properties, new staining techniques brought the use of banding patterns to the fore. Both quinacrine mustard and Giemsa stains allow the banding structure to be viewed. The quinacrine mustard stain has been used extensively in fluorescent microscopy (Caspersson, Lomakka, and Møller, 1971; Caspersson, Lomakka, and Zech, 1971; Caspersson and Zech, 1972). To analyze his data Caspersson uses harmonic analysis to characterize each chromosome in a set. These techniques have sparked interest in the use of banding techniques. The more recent utilization of the Giemsa stain has been employed by several research groups. The Giemsa stain permits viewing in a normal light microscope.

Two papers in particular point to the biological difficulties in chromosome analysis. "Localization of Heterochromatin in Human Chromosomes" by Arrighi and Hsu (1971) gives an introduction to staining problems of DNA. "Heterochromatin, Satellite DNA, and Cell Function" by Yunis and Yasmineh (1971) points to problems in defining structtural composition of DNA.

The automated identification of chromosomes involves problems in many different areas. A fully automatic system must not only identify individual chromosomes, but it also must find the metaphase spreads in which they occur. Since high-powered microscope optics are used, an automatic focusing device must be employed in the search for metaphase spreads. Both of these difficult problems have been solved. In 1972 Herron et al. developed a spectral method for detecting mitotic cell spreads. In the same year Mendelsohn and Mayall (1972) derived a successful focusing algorithm based on the summation of optical densities. These methods form the basis for automatic scanning devices.

While these researchers sought to automate chromosome identification, others included the human operator as a critical element in the identification process. The paper by Low et al. (1971) offers a valuable new approach by implementing its decision interactively with a human operator. In this same direction Bender et al. (1972) supply a method of combined Bayesian decisions and logic selections which compares favorably with the performance of trained cytogeneticists. Bender et al.'s choise of a small computer reflects the developments in instrumentation control by computer and the desire for a commercially realizable system. A system by Frey (1969) also is designed as an interactive system. Frey's report offers both an analysis and suggestions for the solution of problems that have hindered some of the larger systems designed to cope with chromosome identification. The MACBAC developed by Ledley et al. (1972) shows particular care in the implementation of the man-machine interface. Their choice of data, instrumentation, and system design make this paper significant reading.

The theoretical methods employed in chromosome identification are a subset of the methods described as pattern recognition. Many articles deal with this broad area of research. Several of these are specifically related to or form the theoretical basis for chromosome identification algorithms. Most specific are Prewitt's (1970) article "Object Enhancement and Extraction," and Rosenfeld's (1969) chapter on feature extraction which supplies a generic approach to the problems of localization and orientation. At a broader level are the texts by Andrews (1972), Duda and Hart (1973), and Fukunaga (1972). An excellent overview of image processing is given by Huang, Schreiber, and Tretiak (1971). The paper addresses itself to the question of choosing data precision. Kendall and Stuart's (1968) three-volume set provides an excellent source for the statistics that are involved in pattern recognition methods. Additional help is available from Liu (1968) in combinatorics and from Siegel (1956) in nonparametric statistics.

THE CONSTRAINED PROBLEM

The identification of human chromosomes is of considerable importance to medical research and patient counseling. Genetic counseling could be made available before pregnancy. High-risk patients could be screened and given special attention. Congenital diseases such as hemophilia could be determined. Suspected congenital links in heart disease, diabetes, and epilepsy could be identified. Populations could be screened for Klinefelter's syndrome and Down's syndrome. Mosaicism, the mixture of dissimilar cell lines in the same organism, could be studied. The changes in genetic structure due to chemical agents and to radiation could be followed. A population data base would permit the examination of genetic pools. Biological agents which give rise to chromosome damage such as the oncogenic viruses could be pursued with greater precision. The study of these areas will substantially affect the health and welfare of many people. Reliable chromosome identification would greatly aid these studies.

Currently, the majority of hospitals and labora tories doing chromosome identification use the karyogram. To prepare a karyogram, a slide with metaphase spreads is examined. The investigator may spend from one to two hours examining the slide of a single patient. When a suitable spread is found, it is photographed. The film is developed, and an enlargement is made. At this point, the pathologist cuts out chromosomes from the prints and arrays them as pairs. Similar chromosomes are placed in the autosome groups A through G, or they are labeled as sex chromosomes. The pairs are pasted under their labels on a paper sheet, and the result is called a "karyogram" (Fig. 2).

Karyograms are routinely used to check for an abnormal number of chromosomes. This abnormality is easily detected by visual inspection, but a significant improvement could be obtained by reducing the time necessary to find metaphase spreads. This automation has been accomplished by several groups, most notably Herron et al. (1972) at the University of Pittsburgh. Herron et al. report good selection of mitotic cells by use of width measures. Not only is the time reduced for metaphase selection, but also the laboratory technician is free for work elsewhere.

The time of the pathologist or cytogeneticist is an important cost factor. Since most individuals have normal chromosome structures, the pathologist can spend time best by investigating abnormal spreads. An identification of normal individuals would screen a great deal of data for the pathologist.

Cost considerations are important in the choice of equipment. Many algorithms require the use of a large-scale computer. While offering precision of computation, these

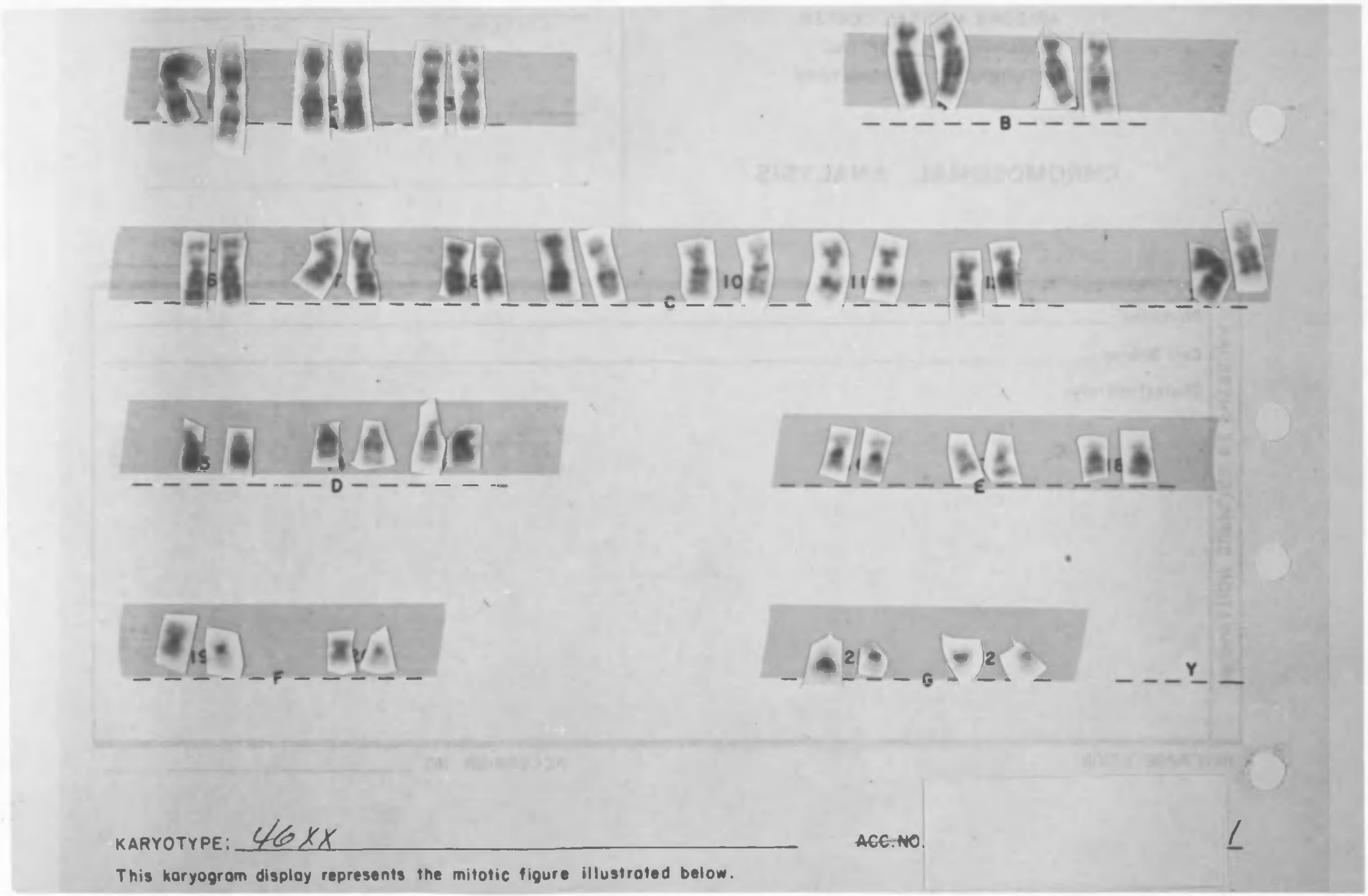


Fig. 2. Karyogram -- This karyogram of a human female shows the grouping and numbering nomenclatures. Note that the X chromosomes are grouped in the C class.

machines are frequently beyond the financial reach of small hospitals and laboratories. Several groups (e.g., Bender et al., 1972; Castleman, 1975), have demonstrated the use of small computers for data acquisition. Although traditionally slower than large machines, minicomputers are easily interfaced to electro-optical transducers and are particularly effective as controllers of mechanical apparatus. The minicomputer also lowers costs by requiring fewer personnel and almost no air conditioning. Finally, the uncomplicated software and the "hands-on" access make the minicomputer available to laboratory personnel untrained in computer technology.

Accurate and complete identifications are of first importance to the pathologist. It is not enough to identify only one or two chromosomes at a high level of precision; whole spreads must be identified in order for a screening algorithm to be viable. Nevertheless, it is a prime consideration that no chromosome be misidentified, even at the cost of not identifying the spread.

These considerations of accuracy, cost, and time give a constrained definition of the problem of developing an automated karyotyping device.

RATIONALE

Data

Numerous attempts have been made to solve the problem of automated chromosome identification. To date, there has been no clear solution, but the work of other researchers points to the likelihood of success with proper equipment and the use of proper analysis.

From several studies (Castleman, 1975; Low et al., 1971; Frey, 1969) it can be noted that an interactive approach significantly reduces both computation and error. A totally automated system would include location of metaphase spreads, location of chromosomes within those spreads, orientation of chromosomes if necessary, automatic focusing, path selections for the banding patterns, data input, comparison pairing, and identification. Of these processes the human operator reliably performs all but the last three. The location of metaphase spreads, although performed more quickly by a computer, is an easy pattern recognition problem that may be accomplished by any laboratory technician with a small amount of training. Similarly, location, orientation along an axis, and path selection may be accomplished easily by the human. These tasks would otherwise represent either a time or cost increase far over the human observer. Consequently, for the system proposed here, the human selects and orients the chromosome.

What data will best represent the chromosomes for possible identification? The banding pattern appears as the best choice. The banding pattern of the chromosome is a graph of the optical density plotted for the entire length of each chromosome. It is a function of the amoutn of DNA stained at each location. Ideally, each type of chromosome will have a unique banding pattern. Moreover, the banding pattern offers more detailed data than the geometrical-morphological methods. An added inducement is that the banding pattern gives a smooth function representation which makes it amenable to many types of mathematical comparison.

A data base from banded chromosomes could be stored for referral and analysis. Special interest occurs if the DNA base is cross-referenced to patient history.

The selection of banding patterns, the need for their storage and comparison, and the desire for a low-cost system all constrain the choice of equipment.

First, how can the banding patterns be stored? Handwritten output is slow, cumbersome, and offers no easy way for retrieval. Analog graphs such as strip chart recording have some advantage but offer no ready means of computation. The obvious choice for speed, ease of computation, and storage is a digital computer equipped with a mass storage medium. A minicomputer offers a ready way to convert banding patterns into computerized form through analog to digital channels, a storage medium, e.g., magnetic tape, and computational ease. Moreover, a minicomputer costing in the range of ten to twenty thousand dollars is within the reach of a clinical laboratory.

The problem now is to convert a chromosome image into a set of voltages representing its banding patterns. Instruments such as vidicons, photodigitizing cameras, and photomultipliers, are immediate considerations. The vidicon must be rejected for its lack of resolution. Digitizing cameras, although excellent for resolution, are not reasonable in price. A photomultiplier tube offers low-cost (under one thousand dollars) and adequate resolution.

However, the selection of a photomultiplier tube creates a problem. The photomultiplier tube produces a signal only for that part of the chromosome directly in a path of light. The PMT is focused over a small aperture, about 10 microns by 40 microns. The aperture determines the field of view and, therefore, selects the part of the chromosome that is to be digitized. In order to digitize other points along the chromosome, the chromosome must be moved under the aperture. This control of chromosome location is accomplished with stepping motors (Fig 3).

A scanning optical microscope is equipped with a motorized stage which, in this case, is controlled and programmable via the computer. The stage allows the

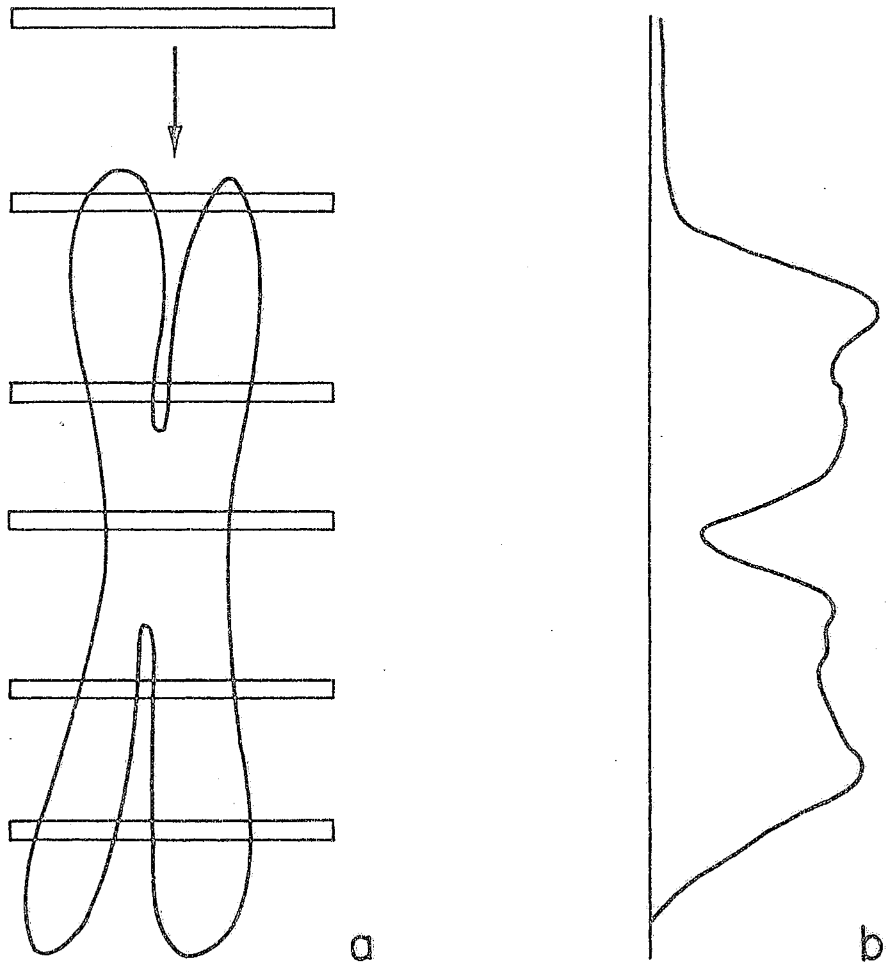


Fig. 3. Slit Scan -- The banding pattern is determined by the path of the slit aperture over the chromosome. The motor advances the chromosome under the scanning aperture. At each step of the motor, the photomultiplier converts the light energy to a signal, (a), resulting in a banding pattern, (b).

chromosomes to be placed in the path of a light beam, a measurement taken, and the next point of the chromosome positioned. There are financial and technical limitations on the incremental stepping motors of the scanning optical microscope. As the incremental step size becomes smaller and smaller, the price becomes larger and larger. Indeed, only recently have increments of one-tenth micron become available at any price.

Even more important are considerations of signal-to-noise ratio. As spot size decreases, the signal becomes lost in the normal amplification noise of the PMT. To circumvent this problem a second, inexpensive step was added to the data acquisition.

Photographs of metaphase spreads were made with an oil immersion lens at 1250x magnification. The negative of this film, white chromosomes and dark background, was then placed under the scanning optical microscope. This pre-magnification of the image sacrifices no resolution with well-chosen film. The disadvantage is that microphotographs require additional scanning time for the photographic step. However, the increase in time requirements for identifying is marginal. Good metaphase spreads can best be selected by the human operator. The time to take a picture is small compared to the time to search for the spread. The time required for the preparation of the sample is much greater than the development time that is necessary. Photographs have the additional advantage that they may be readily stored.

For the banding procedure to be effective, an appropriate biological stain must be used. Of the stains currently used by biologists for banding, Giemsa appears to offer the most consistent results and permits the use of ordinary microscopes. To further enhance the effects of Giemsa stain, phase contrast optics are sometimes employed. This method renders a high-contrast optical image.

Thus, equipment selection links the following devices: a scanning optical microscope with 10 micron incremental motors, a photomultiplier tube to transform light signals into electrical signals, analog-to-digital converters to change voltages into magnetic storage, a minicomputer equipped with a bulk storage medium, and, finally, a microscope-mounted camera to supply the microphotographic negatives of the chromosome spreads.

Mathematics

Intuitively, the problem of chromosome identification may be viewed as an assignment problem. The 46 chromosomes represented by centromere index, contours, banding patterns, or other features must be placed into one of 23 categories so that exactly two are placed in each category.

Andrews (1972) points out in his text that feature selection is the most important step in applied pattern recognition. To date, the banding patterns provide the most promise of meaningful results. There are four approaches to making assignments using banding patterns as the feature.

Caspersson and Zech (1972) at the Karolinska Institute in Sweden have used Fourier coefficients as chromosome features. Their analysis uses seven amplitude and seven phase coefficients to describe the individual chromosome. These 14 coefficients are assumed to be drawn from a Gaussian distribution. The probability that an individual chromosome belongs to a given class, e.g., "A1," becomes a summation of its probabilities for each of the coefficients. A large data base of 5,000 spreads gives this method support.

Caspersson's method is powerful but provides no solution to the problem of uneven contractions of the chromosomes which may change the Fourier coefficients. Ledley et al. (1972) report a possible correction of this problem by using an algorithm developed by Kimberling and Hosteller. This algorithm takes a linear function of a linear regression coefficient and several length measures of the chromosome to create a normalized length. Hilditch and Rutovitz (1972) also have developed a linear transform based on the statistical properties of the chromosomes within a specific spread.

An interesting approach has been developed at Vanderbilt University by Bender et al. (1972). The human operator is used to determine the centromere and the end points of the chromosome. Analysis is done by the calculation of *a posteriori* probabilities. Bender et al. feel that this process models the human thought process in making a "cut and paste" karyogram.

In the study by Low et al. (1971), chromosomes were assigned one of 10 new categories. With this exception, however, all other studies have attempted to make the assignment to the standard karyogram classes so as to provide a link with existing medical records.

By assuming only "normal" data (nonpathological spreads), some strong constraints may be added to the general problem. If no attempt is made to recognize abnormal conditions and if only females are considered, then the problem may be considered as the assignment problem outlined before. (Otherwise, there are 24 classes since the male carries both X and Y chromosomes.) These assumptions permit the problem to be split into two parts. The first part will be the pairing of the chromosomes. The second part will be the identification of the pairs. This separation makes the second part of the problem representable as a square matrix.

There are many matrix representations of assignment problems. Each method depends on its algorithm and on the measures of similarity forming the elements of the matrix. Caspersson and Zech's (1972) use of Fourier coefficients offers a seven-tuple as an encoding of the banding pattern. Geometrical methods use at least three-tuples (width, length, centromere index). To use an assignment matrix the measures of similarity must be reduced to a scalar quantity. Several functionals immediately come to mind to compare chromosome banding patterns, among which are the least squares and the Kolmogorov-Smirnov statistics.

Whichever of these statistics is chosen to compute the elements of the matrix, the technique for reducing the matrix of similarity measures to a matrix of particular assignments must be given careful consideration. Matrix diagonalization methods are well known. Unfortunately, although useful for many physical problems, these methods seek to minimize an error (usually least squares) which, however, does not lead to chromosome identification. Techniques based on these methods fail because the least squares error will arrive at a total solution even at the possible expense of a physically impossible assignment. The chromosome identification problem demands that the number of correct identifications be maximized. The acceptable error form is the number of <u>unidentified</u> chromosomes.

Dynamic programming methods are often a choice for problems which involve sequential procedures. Unfortunately, these methods are difficult to implement because of problems inffinding a suitable optimization strategy.

Chromosomes are in no sense divisible entities. Thus, the allocations in the identification problem make it appear similar to problems solvable by integer programming. Techniques based on the simplex method fail for the same reason as the diagonalization methods from which they are derived. Branch-and-bound and enumeration techniques fail because they demand complete identification.

To cope with the error form given, a sequential reduction in the matrix is employed. At each step this method selects the best assignment. If the best assignment is judged to be better than a criterion, the assignment is assumed correct. This assignment reduces the matrix by eliminating all elements in the row or column of the assignment position. The algorithm continues until all assignments are made or until the criterion has failed. It should be noted that the criterion may be a step-wise function of the matrix as well as a static value.

ACQUISITION OF DATA

Biological Materials

The clinical preparation is initiated with a peripheral blood sample drawn into a sodium heparin vacuum tube. A one-half to one-hour period allows the plasma to separate, and 1 cc of plasma is drawn off. A culture tube receives the plasma. After a waiting period to permit mitotic activity, a spindle poison, colchicine, is added to arrest cellular division in the lymphocytes. The tube is centrifuged, and the supernatant is removed. A hypotonic solution followed by an incubation period swells the cells. Fixation and washes followed by centrifugation complete cell preparation. The cells are suspended in solution and placed on a slide. A sharp blowing disperses the cells. Giemsa stain is added.

An operator then selects good metaphase spreads. A good spread is defined as having no overlapping chromosomes, a normal complement of chromosomes, good staining, no twisted chromosomes, and few bent chromosomes. Good spreads are photographed with a camera microscope using a 125x oil immersion objective. Shellburst film (type 2476, ASA 400) responding to a wavelength of 540 nanometers gives good results when exposed at f8 for 1/30 of a second and developed in D19 for five minutes. Phase contrast microscopy enhances the chromosome boundary and adds reliability to the tracing algorithm.

Operator Procedures

Of the many spreads on a slide, only a few will be suitable for use. Photographs are taken of the best spreads. In turn, the best photographic negative is placed on a glass slide, and this slide is placed on the stage of a scanning microscope which is controlled by the computer. The technician selects the first chromosome to be scanned and positions it slightly to the right of the cross hairs on the microscope. He then adjusts the light settings and starts the CHRMAS system. The program then follows the contour of the chromosome and displays it on a CRT. The technician may then use a small joystick to select the path where the banding patterns are to be taken. As he moves the joystick, a trace appears on the screen indicating the path (Fig. 4). As the joystick moves, the computer takes readings through the microscope. If the operator is not satisfied with the trace, a keyboard entry will permit a re-scan. If satisfied, he acts to continue the analysis. The display vanishes and a hard copy of the banding pattern is printed out. This output is followed by a graph of the banding pattern on the display unit. On command the display terminates, and the program is ready for the next chromosome. The trace is deleted from the screen so that the

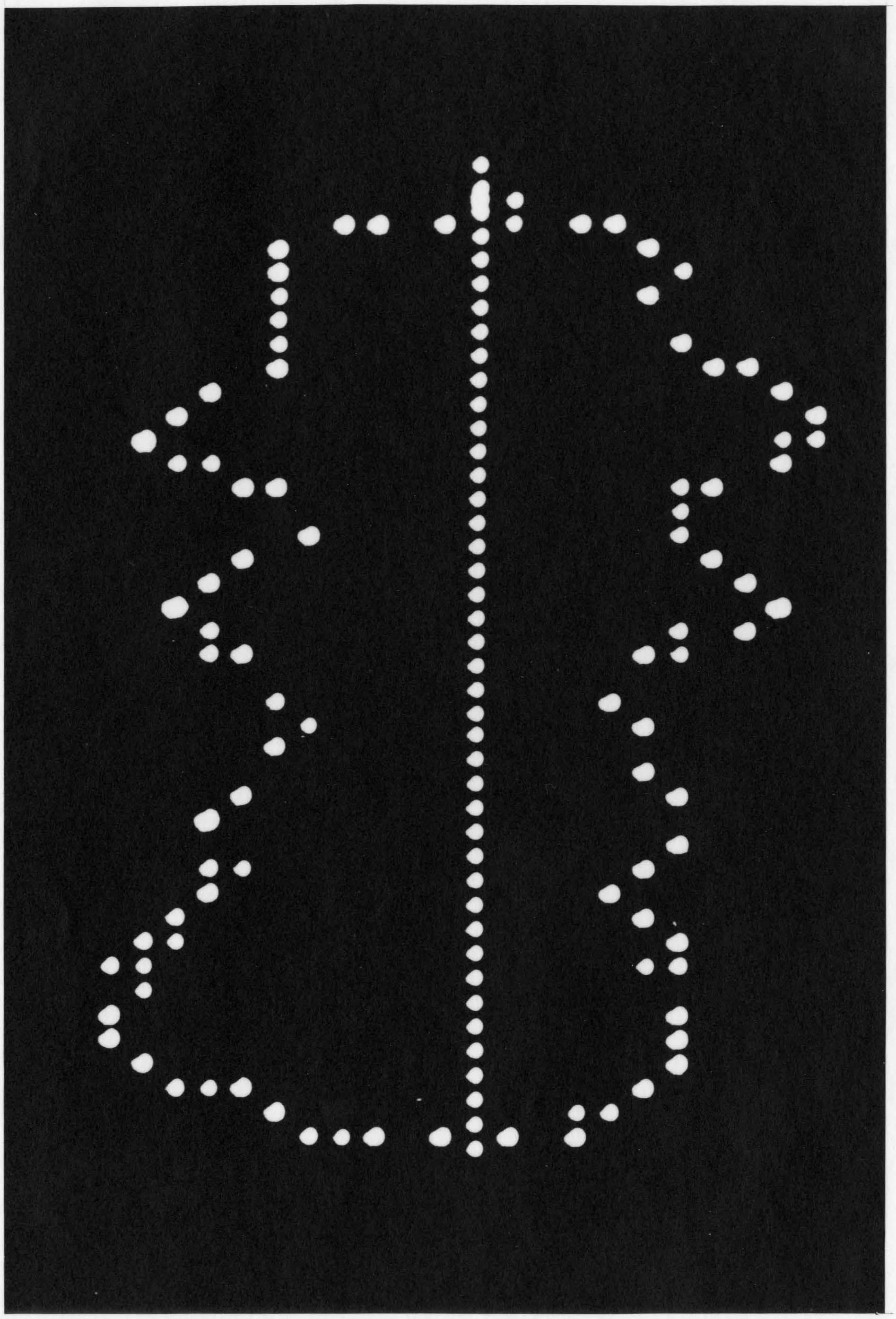


Fig. 4. Boundary Trace -- The results of the contour-following algorithm are displayed on the screen. The line through the center shows the path the operator selected for the trace of the banding pattern.

joystick may be used to make a new trace. When all 46 banding patterns have been traced in this manner, the program reduces these patterns to 23 pairs. The program then asks the operator how much reliability is desired. The operator may select a value between zero and 100. This number permits the program to be used as a tool for the pathologist or physician. By selecting zero level reliability, the machine will identify every chromosome; this gives the pathologist the "opinion" of the machine. For screening operations, however, the reliability should be set at a much higher value. In this fashion, control may be exerted so that spurious identification will be high improbable. The program compares the 23 banding patterns it now has with a set of standardized patterns in its memory. It will identify all those chromosomes that surpass the level of reliability asked for by the technician. Unidentified chromosomes and their counterparts will be listed separately. If a complete identification has been made, data may be stored for reference, or the results may be brought to the attention of the pathologist.

Hardware

Both visually and logically the hardware falls into two categories: (1) the microscope and analog instrumentation and (2) the computer and its peripheral devices.

A scanning optical microscope, the Leitz MPV, provides all of the functions delineated in the Rationale chapter of this paper. The microscope is equipped with a 3.5x objective and a 10x ocular. The tube factor is 1.25x. The microscope is supplied with light from a 100 watt DC power supply and a 30 watt bulb. The scanning stages of the microscope are controlled by 10-micron incremental stepping motors with a total travel of 1,000 steps. A photomultiplier tube is mounted on top of the microscope. The PMT has a rise time of 30 nanoseconds and is supplied from a 1,200 volt power supply. A digital volt meter supplies readout of the photomultiplier values. An adjustable aperture is located directly in front of the photomultiplier tube to control the field view.

A PDP-12 minicomputer manufactured by Digital Equipment Corporation was available. This machine was equipped with 8K of 12-bit words in its core memory. The operating system, OS-12, was supported by two linc tape drives. The linc tape drives are unique in the industry as they provide random access to a sequential medium. Communication to the PDP-12 was through an ASR 33 teletype. Conv rsions of voltages were accomplished with 12-bit A/D and D/A converters. The analog-to-digital converter received the output of the photomultiplier tube, and the digital-to-analog converters permitted output of control voltages to the stepping motors. To provide for interactive capability, a VR14 point plot display and a control panel were added to the above configurations. The control panel consisted of a joystick, four potentiometers, and two buttons. These inputs could be read by A/D channels and provided capability for efficient operator control. Their exact functions will be discussed later in this paper. Such a panel may easily be added to the computer system as a plug-in module. The equipment proved satisfactory for the problem discussed in this paper (Fig. 5).

Software

The task of completely automating the acquisition of banding patterns involves the following subtasks: (1) a mapping algorithm to locate all chromosomes on the slide, (2) a recognition algorithm to distinguish chromosomes from dirt and other noise patterns, (3) a focusing program to automatically focus the microscope, (4) a chromosome orientation program to position the chromosome in a standard orientation, (5) a tracing algorithm, and (6) a scanning program. Although a fully automated system would be of great economical interest, the problem demands control equipment not yet existing in commercial form and introduces greatly increased error problems. Moreover, by substituting a human operator for locating, orienting, and proper focusing of chromosomes, the problem became feasible for the project horizon. The human operator can quickly,

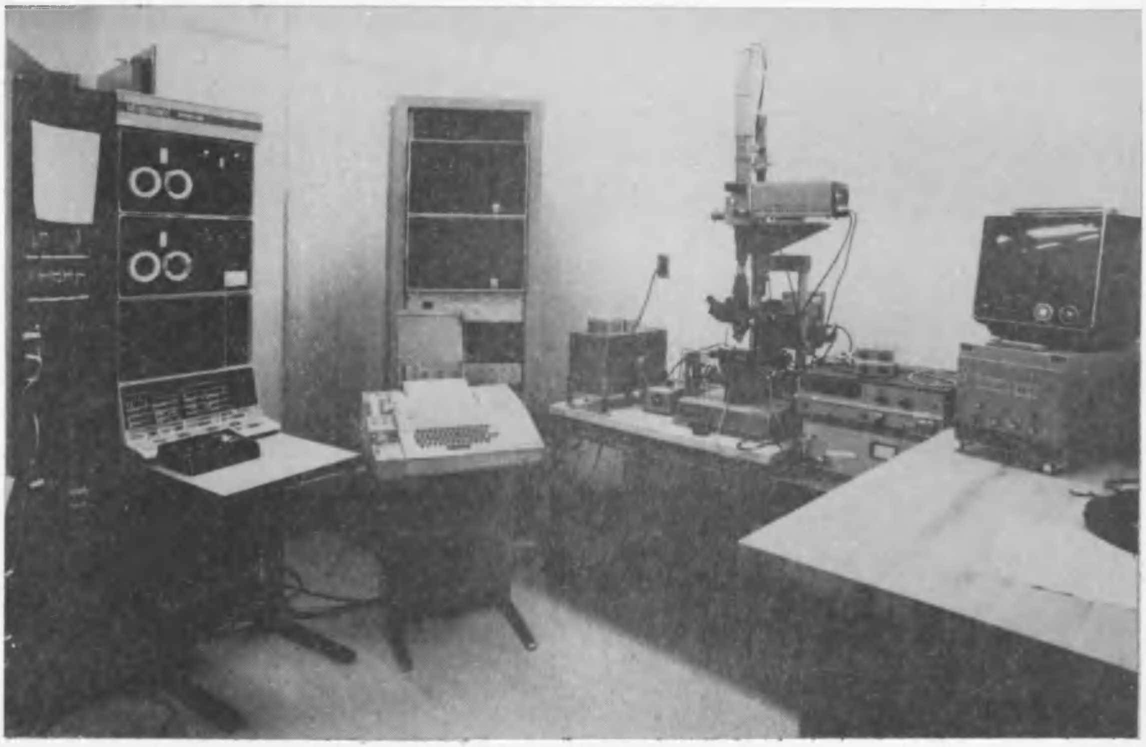


Fig. 5. Hardware -- The hardware configuration of the system includes the PDP-12 computer (far left), the operator's panel with joystick and pushbuttons, the ASR-33 teletype, and the microscope configuration. Seen on top of the microscope is the photomultiplier tube.

efficiently, and for a lower initial cost, focus and detect chromosome locations on the slide. Most importantly, the human is a reliable operator in this setting and may be substituted for a set of automatic equipment that is largely convenient rather than necessary.

Thus, we will assume that humans prepare the spreads, photograph them, and perform the selection, orientation, and focus necessary to prepare a given chromosome for input. These operations can easily be accomplished by manipulation of the appropriate microscope controls. However, the selection of a recording path is a more complicated process. By inputting a voltage from the joystick, the program can send a signal to the stepping motors and thus change the position of the chromosome relative to the measuring aperture. This joystick control allows the operator to move the chromosome under the measuring aperture and directly choose the path for recording the banding pattern. However, since the program was structured as a developmental program, multiple path selections became a useful feature. For this reason a point plot display of the chromosome perimeter was added to the program. The perimeter was first traced and then scaled to the size of the screen. The joystick still controls the position of the stage relative to the cross hairs. The position of the cross hairs is depicted on the screen as a point. The path of the cross hairs over the chromosome may be recorded by actuating a button on the control panel.

To investigate pathologies it may be necessary to look at banding patterns in specific, unusual locations, or to scan with more precision. With this goal in mind, a display is needed so that selected banding patterns may be specified. A convenient reference frame for this display is the outline of the chromosome. With this outline the human operator may meaningfully select different locations for banding patterns.

To place an outline on the screen, it was necessary to produce an algorithm which would locate and track the boundary of the chromosome. The human operator is unable to perform this task because the photographic grain in the film inbeds the image in a manner that makes detection of boundary points imprecise.

It may be well to review the operator procedure for taking data in the framework just described. After finding, focusing, and orienting the chromosome, the operator moves the cross hairs slightly to the right side of the chromosome. A button on the control panel actuates the program. First, the stage moves left a step at a time until it finds the edge of the chromosome. The edge may be detected in a variety of ways. The simplest is to look for a sharp decrease in the optical density which notes the change from background to chromosome image. A spot reading can be taken at the cross hairs. If the reading is less than a predetermined threshold value the stage is moved one step to the left. This procedure causes the machine to sample background points and to move progressively closer to the boundary. When a reading is taken that is greater than the threshold, the algorithm backs up one step and assumes the boundary has been reached.

Next, the trace algorithm is started. This algorithm is derived from a paper by Troxel (1971) entitled "Automated Reading of the Printed Page." Troxel's algorithm proceeds as follows: read a value from the photomultiplier. If the value is above a threhold previously established, turn right from the current position; if the value is below the threshold, turn left. Note that left and right turns are relative directions. They are directions only with respect to the previous movement (Fig. 6). The application to the problem discussed here demanded some modification. Since the algorithm often covers the same point twice in its boundary search, it becomes unstable when the same point produces a value first above the threshold and then below the threshold (Fig. 7). Such fluctuation may occur due to the presence of electrical and optical noise. The edge of the chromosome is fuzzy due to the gradual thinning of the DNA content. Often, values at the edge differ from the threshold by only one or two per cent. A three per cent noise obviously produces an impairment in an algorithm that

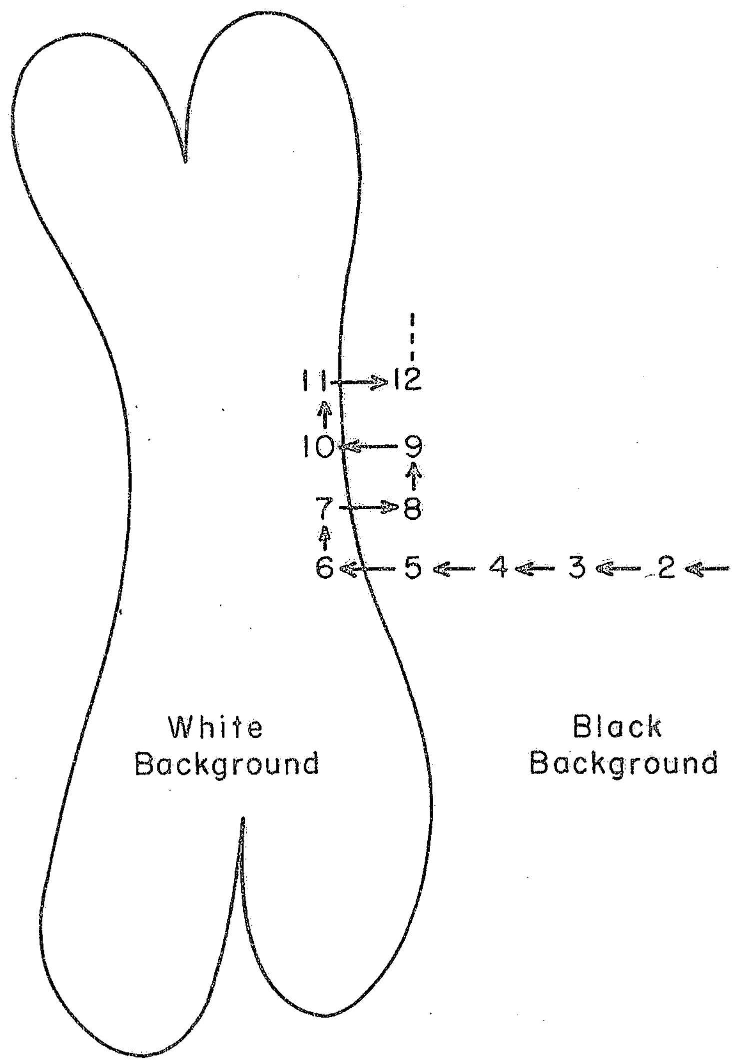


Fig. 6. Boundary Trace -- The trace algorithm records crossings from dark to light. After the boundary is detected (step 5-6), the algorithm turns "right" on detecting a "white" background and "left" on detecting a "black" background. (The background is dark due to the reversal of the photographic negative.)

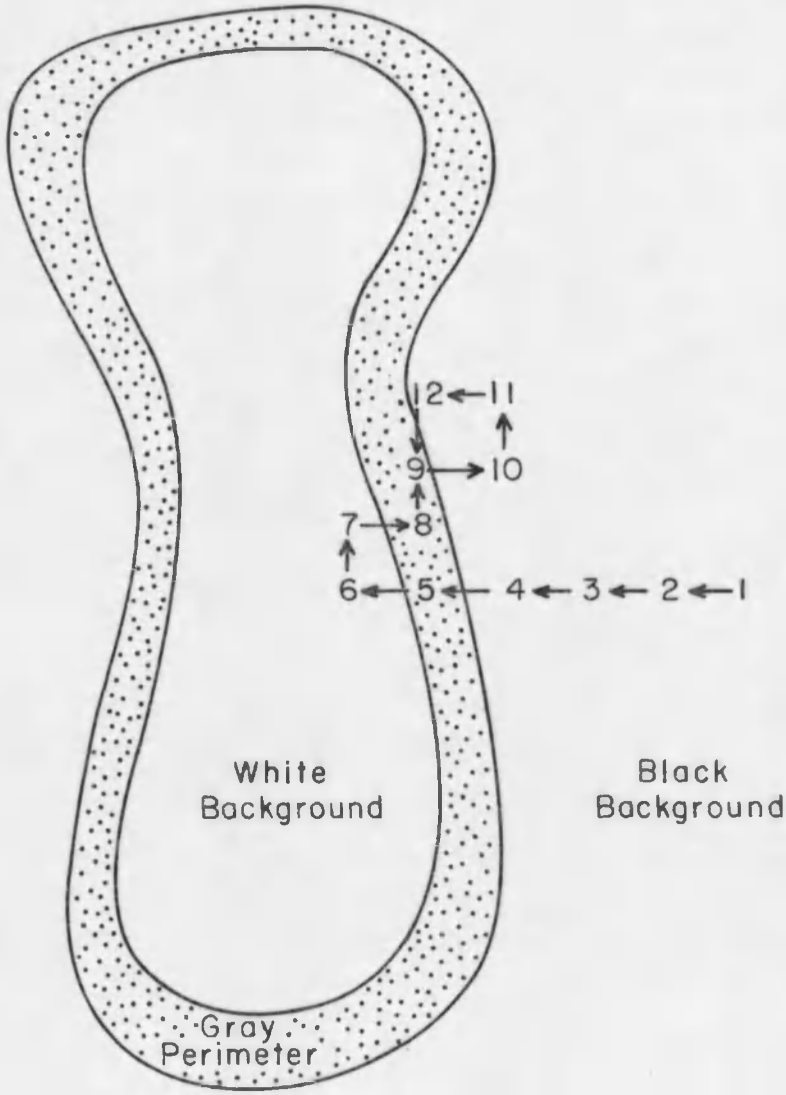


Fig. 7. Lost Boundary Trace -- This example of the boundary trace algorithm depicts how the algorithm may become lost. Suppose that at point 9 a value above threshold is erroneously recorded due to the presence of photometric noise. Upon return to point 9 the value is correctly recorded as below threshold. The algorithm is now lost and will continue to cycle through points 9, 10, 11, and 12 until point 9 is again recorded as above threshold, a rare event.

depends on repeated values of the same point. Although a dwell average is used to take photomultiplier readings, it cannot correct this difficulty. To correct this interference, one may increase the step size. This solves the problem by jumping over the fuzzy area and moving definitely into the chromosome image or definitely into the background at each step. Two disadvantages occur with this change. The large step size prevents any detailed look at the contour of the chromosome which is of great interest to the trained cytogeneticist. Second, the large step size has difficulty negotiating sharp turns (ninety degrees or more). Rather than increase the step size, two modifications were added to the algorithm. A push-down stack was constructed to record the last four turns. The algorithm will reread the same point only after passing through three other points. By examining the fourth element of the stack, instead of taking a reading, the algorithm will continue to track the boundary after four successive turns in the same direction. Enough random noise may still cause the algorithm to become lost. Suppose that a point is read as black (above threshold) as are the next three points. This produces four right turns followed by four right turns followed by four right turns--this should not happen at a boundary. The stack, which was included to produce consistent decisions for the same point, now prevents correcting any mistakes due to noise. To be exact, if the first of four points is stored above threshold when, except for noise, it should be read as below threshold, and if the following three points are also above threshold, then the algorithm becomes lost. Enough noise in the system can cause this seemingly improbable occurrence to be both annoying and frequent. A simple counter solves the problem. By counting the number of successive right (or left) turns and by reversing direction for two steps when the counter becomes equal to five, the occurrence of "lost" boundary traces drops to less than two per cent of all attempted traces. A simple restart button also provides abort control for the operator. To further aid the operator the trace is depicted on the screen while in progress.

After the trace is completed, a keyboard command either will start a new trace or will continue the program. (A restart button conveniently repositions the cross hairs to the earlier starting position.) If the program is continued, it scales the chromosome trace to screen size and gives joystick control to the operator.* By manipulation of the joystick, the operator can position any part of the chromosome under the cross hairs. Visual feedback is obtained through the screen display. To store the values at

*The author is indebted to James Puls for the assembly language drivers which permitted easy use of the joystick, display screen, stages, and photomultiplier used in this effort.

a series of data points, one button initiates the process, and a second button halts storage.

The typical procedure follows this outline: (1) the chromosome is displayed in outline form on the screen and the cursor is moved to the top of the chromosome outline; (2) the automatic scan is initiated, and a trace appears on the display as the photomultiplier reads in voltages; (3) the trace is terminated at the boundary, either automatically or by the operator.

Two conditions call for operator intervention. First, the ideal scan assumes that the chromosome is reasonably straight. If it is not straight, either joystick control or keyboard control of the cursor must be used to traverse the paths manually rather than automatically. Second, the operator must be sure that the photomultiplier is adjusted so that overflow does not occur in the analog-to-digital buffer. If it does occur, the chromosome must be rescanned.

Keyboard control permits selection of step size. Operationally, a step size of 40 microns proved to be the best compromise between precision and timing trade-offs.

Most chromosomes are not significantly bent in preparation. This fortunate circumstance of the preparation makes an autotrace feature useful. It assumes the long axis of the chromosome to be parallel to the vertical axis of the cross hairs. Button control initiates and stops vertical trace down the major axis.

Once the data are taken, the operator is queried if the trace is satisfactory. A keyboard reply will permit either a new trace or will move the program into the computation and display of the banding pattern.

The computation of the banding pattern considers the chromosome shape. Since the chromosome may be bent, it is necessary to form a bivariate linear interpolation for each of the coordinates along the banding trace. The resultant function is displayed on the screen (Fig. 8). As before, the operator may elect to continue the program or to repeat the trace.

The banding trace may be taken with different options in the acquisition of data. The aperture located in front of the photomultiplier permits selection of the measuring field. One option is to use a thin slit aperture perpendicular to the long axis of the chromosome. This permits a single signal to result from both arms of the chromosome. Since the DNA is expected to have only minor variations across the sides of the chromosome, this physical integration saves laboirous computation. If the chromosome is bent, however, the position of the slit becomes crucial. The microscope chosen provides that the aperture slit may be formed only along the X or Y axis; it does not rotate. As a result, a bent chromosome forces the operator to

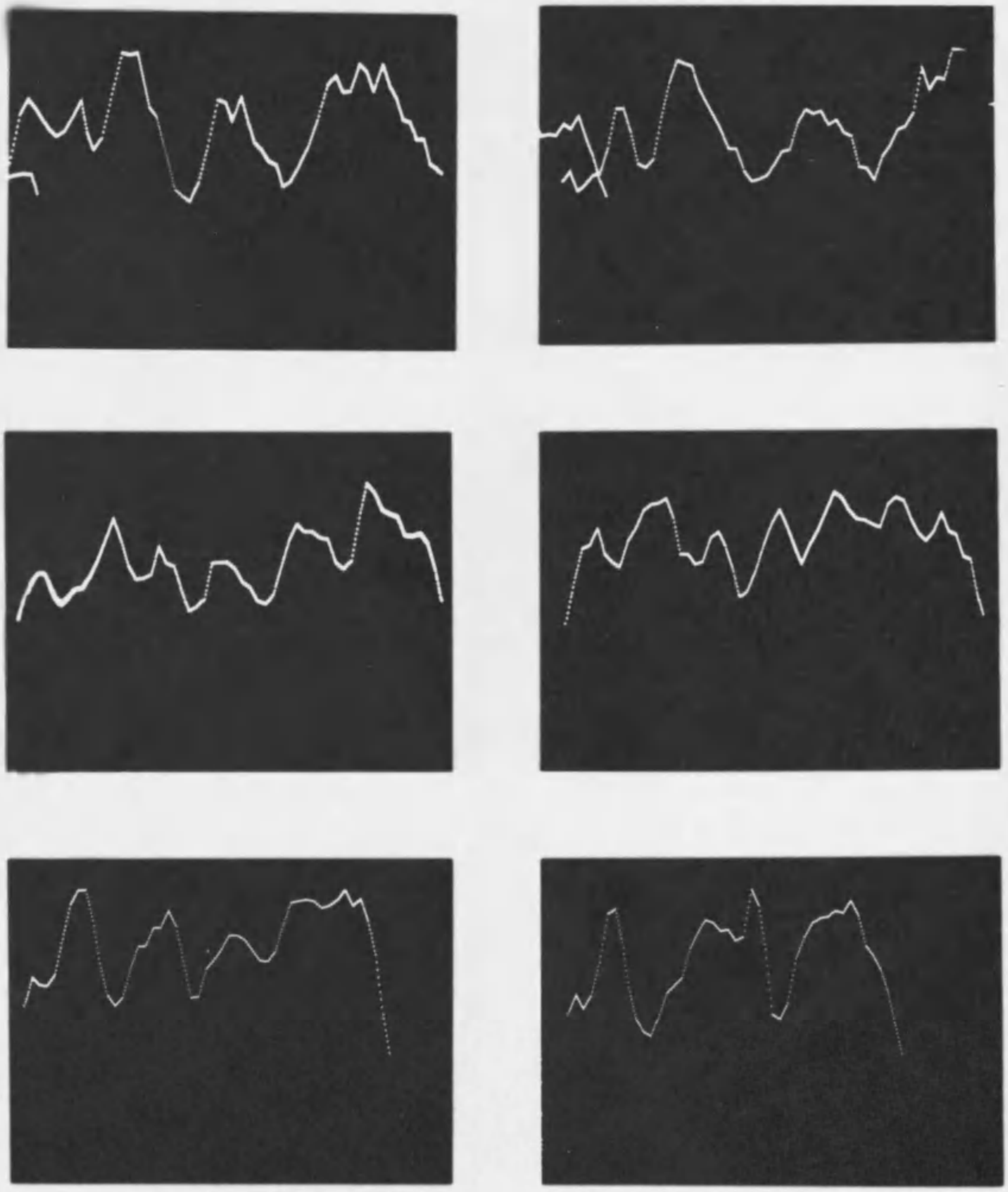


Fig. 8. Banding Patterns of the Normal Human Female -- The patterns are paired by their reference given in the pairing algorithm. Note that some dissimilar patterns achieve similarity when the pattern of one is reversed.

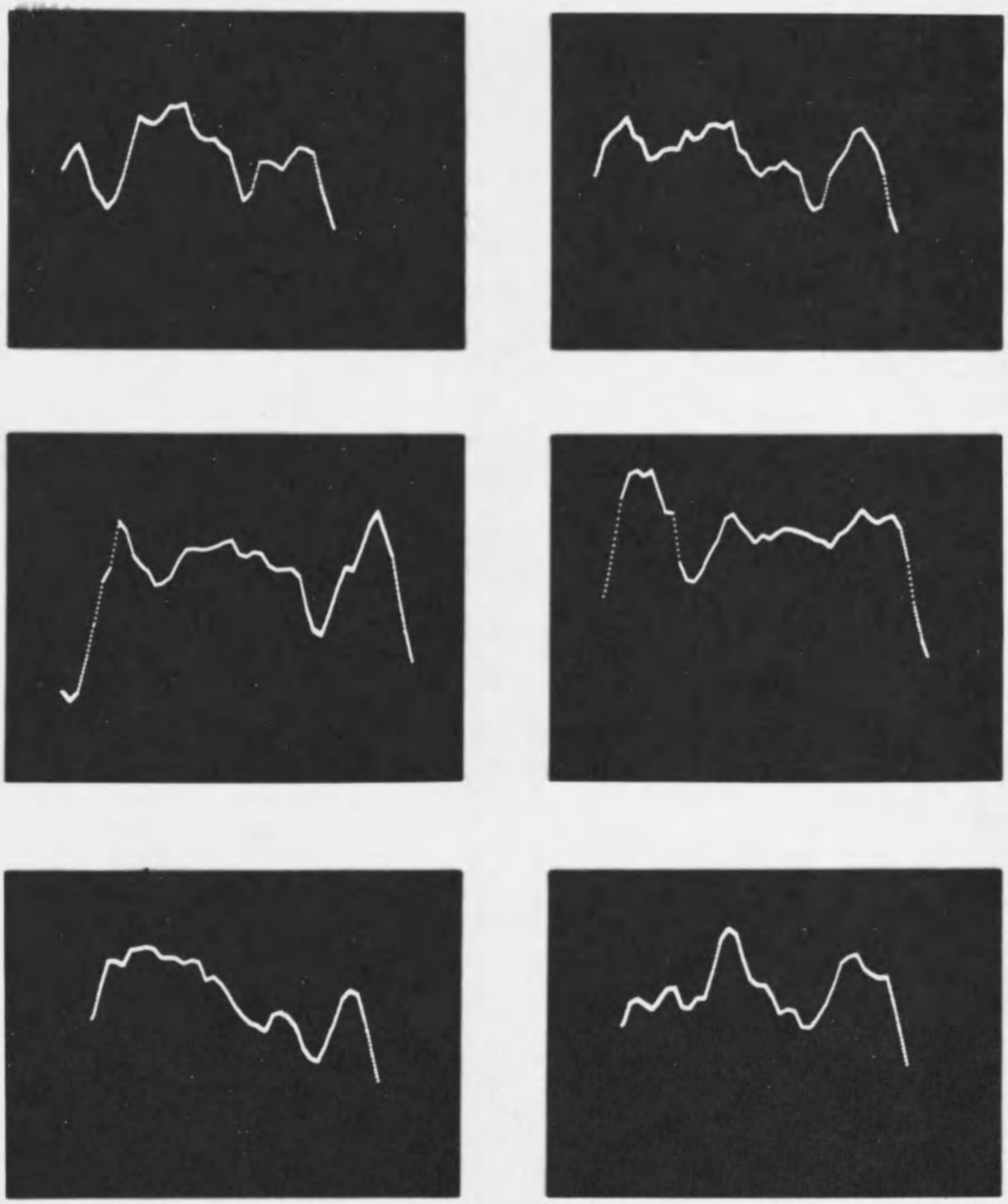


Fig. 8.--Continued Banding Patterns of the Normal Human Female

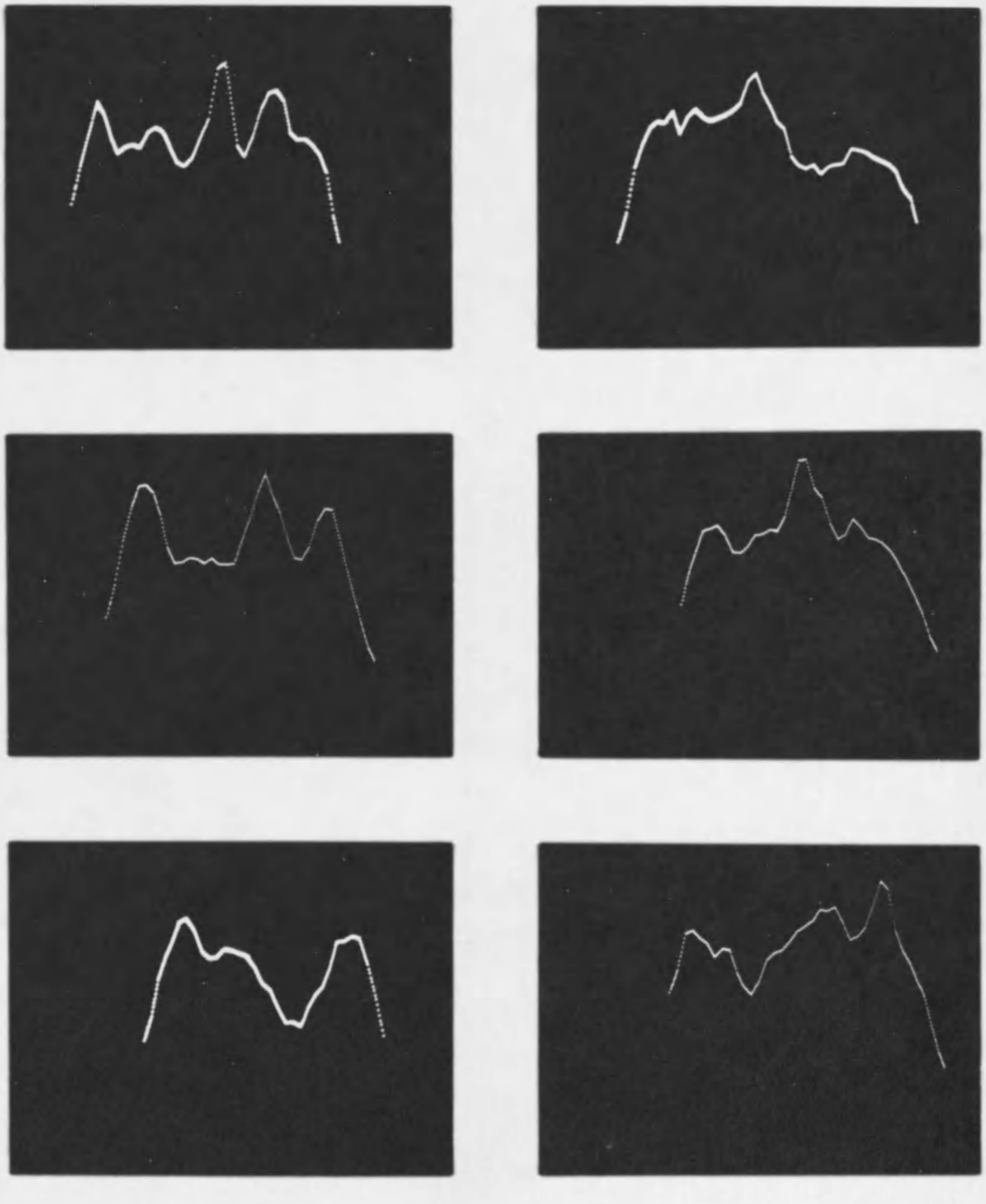


Fig. 8.--Continued Banding Patterns of the Normal Human Female

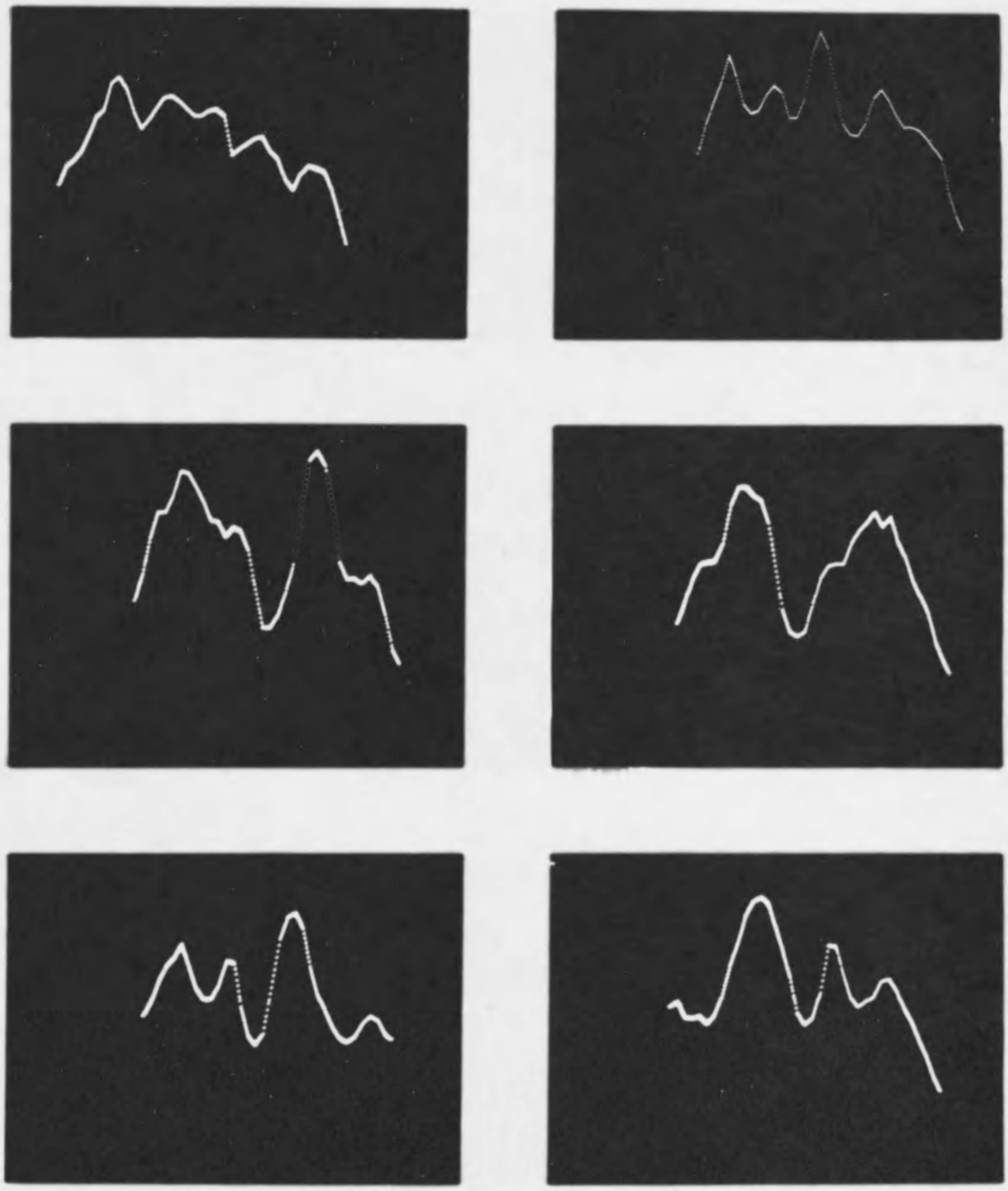


Fig. 8.--Continued Banding Patterns of the Normal Human Female

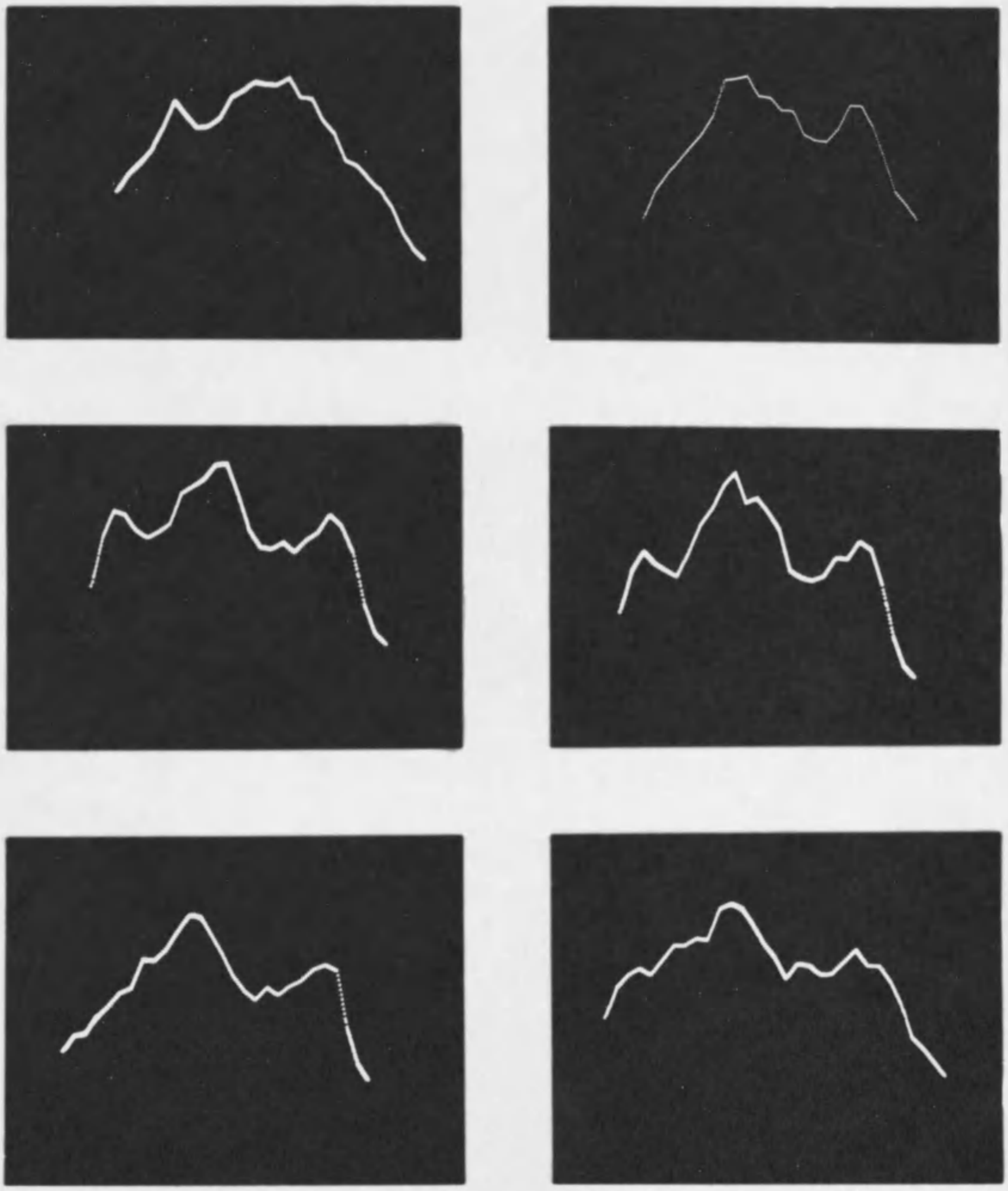


Fig. 8.--Continued Banding Patterns of the Normal Human Female

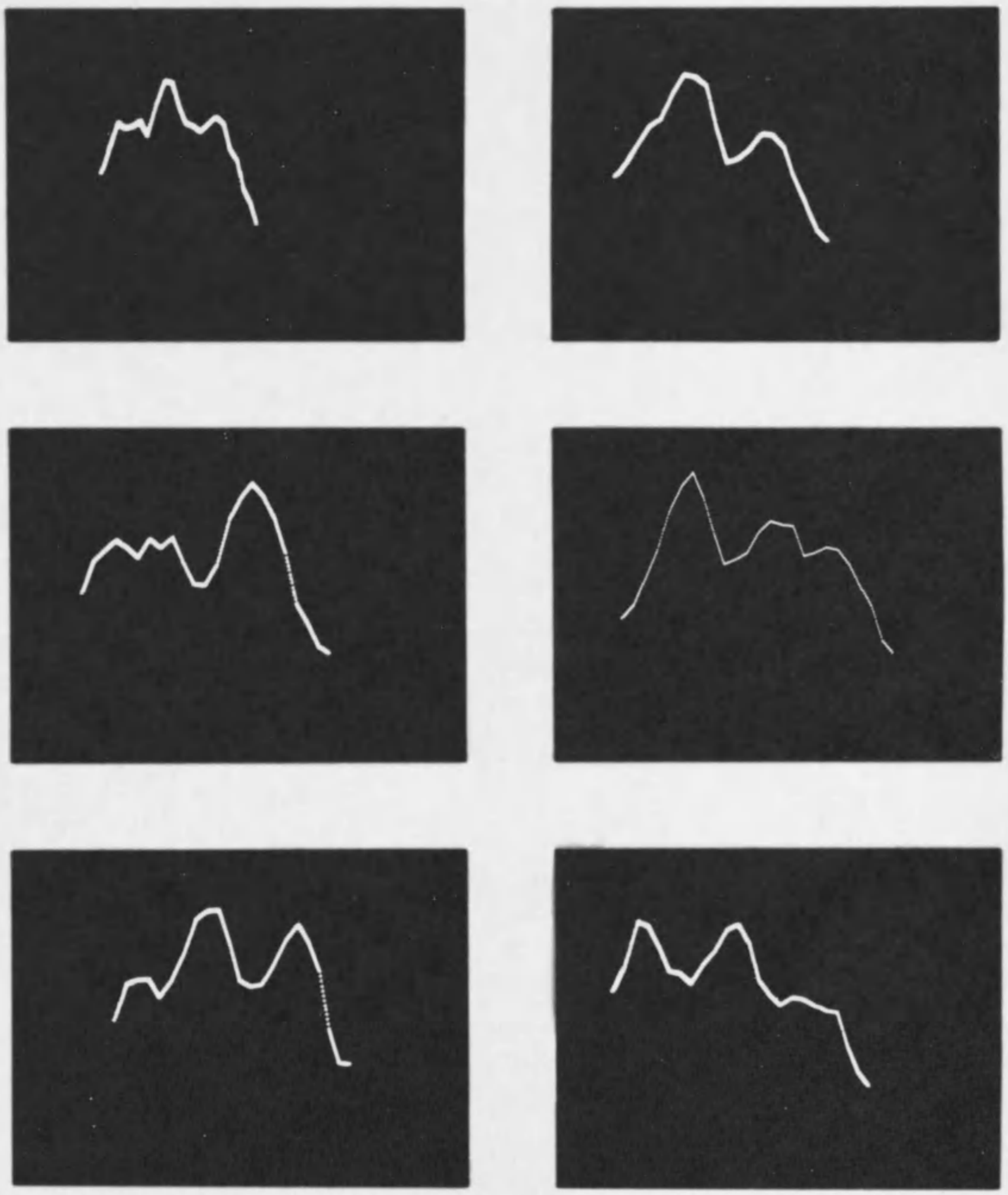


Fig. 8.--<u>Continued</u> Banding Patterns of the Normal Human Female

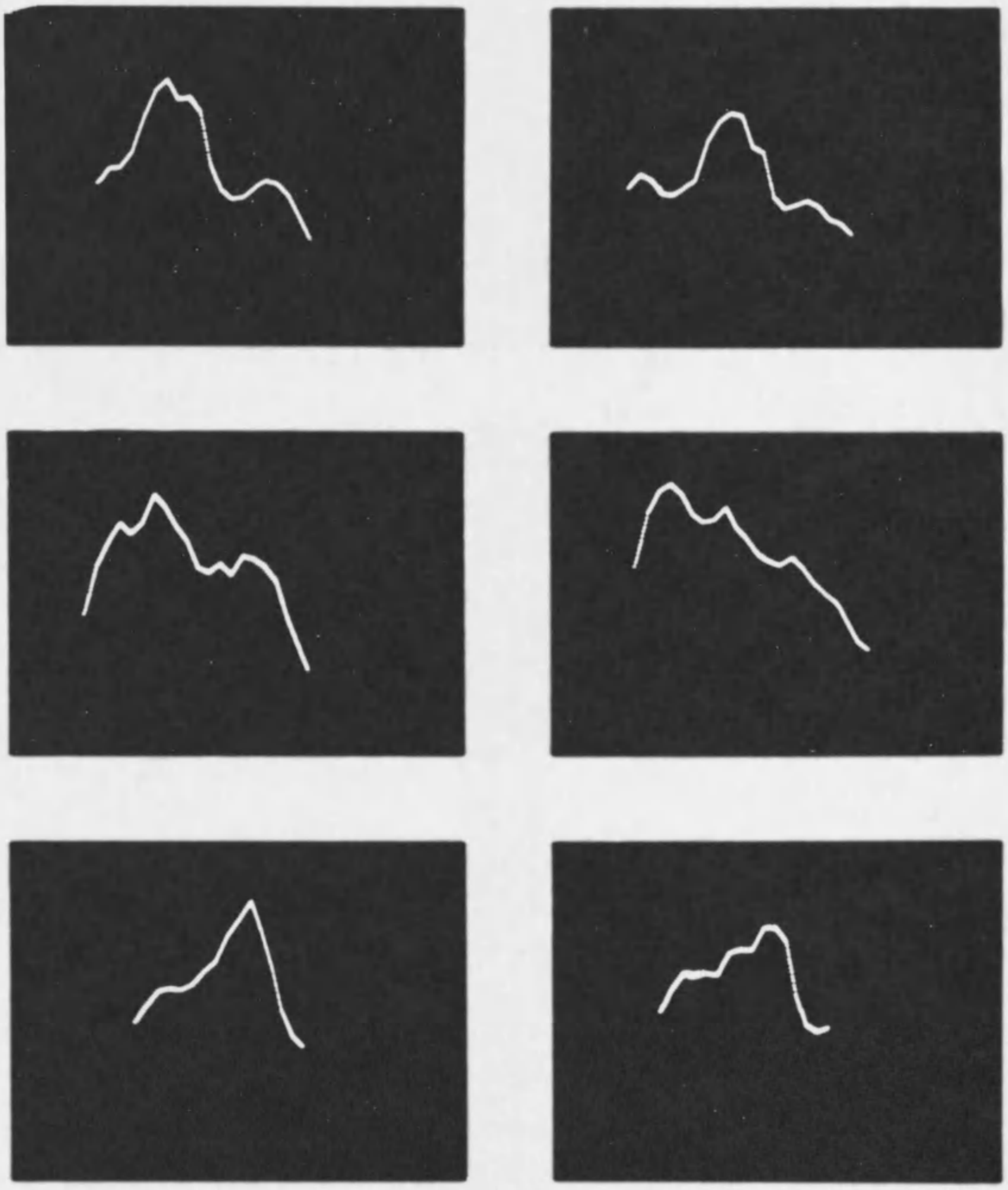


Fig. 8.--Continued Banding Patterns of the Normal Human Female

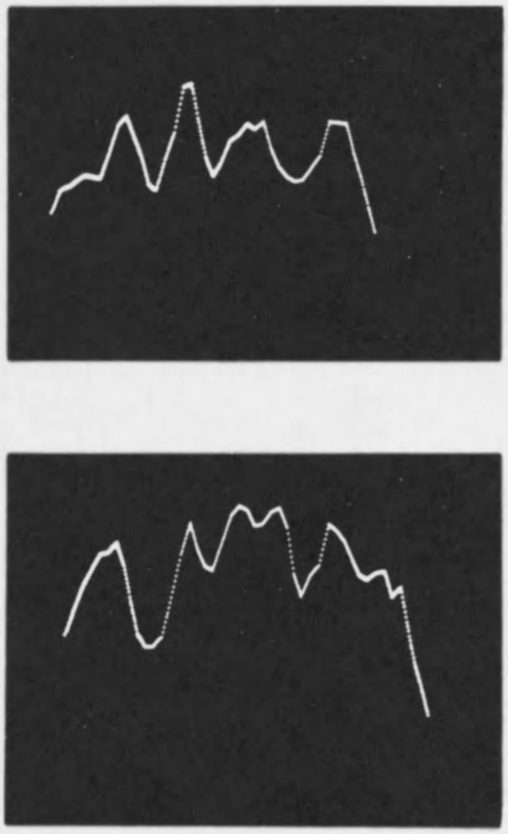


Fig. 8.--Continued Banding Patterns of the Normal Human Female

carefully reposition the stage which rotates only in a manual mode.

Before the banding pattern is stored on magnetic tape, a plot may be produced from the teletype at keyboard request. A user command will store the data, and the program will restart to trace the next chromosome.

As a further convenience, after each step the program queries the operator to insure human satisfaction with the overall system performance. Thus, the software remains interactive throughout data acquisition.

The foregoing discussion and the program commentary should enable easy examination of the coding. All programming, except for drivers to analog devices, was performed in FORTRAN. This selection was based on the author's opinion that the commonality of this language would allow program adaptation for other machines and ease of modification by other programmers.

THE ALGORITHMS

Measures

As discussed earlier in this paper, there are two parts in the identification. The chromosomes are first paired together reducing the identification process to the size of the haploid set. Next, the pairs are matched to prototype chromosomes. The measure used for both parts of this procedure is the Kolmogorov-Smirnov statistic.

There are many methods available with which to compare two distributions. The chromosome problem lacks any information about the distribution. Although an interval scale exists, an interval measure is not appropriate because of the previously mentioned contraction of the DNA. Three measures are commonly used for distribution-free, ordinal scale tests:

1. Pearson: $\chi^2(x)$
2. Cramer-von Mises: $\int_{\infty}^{\infty}[(F(x)-G(x)]^2 dF(x)$
3. Kolmogorov-Smirnov: $\underset{x}{\mathrm{MAX}}|F(x)-G(x)|$.

The disadvantage of the χ^2 statistic for both sampling and for choice of expected value eliminated it from consideration. A good case could be made for using either the Cramer-von Mises test or the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov is more attractive for its conservatism and for its common acceptance. The Kolmogorov-Smirnov statistic is sensitive to central tendency, skewness, and kurtosis.

From a computational standpoint it was necessary to process the data before computing the Kolmogorov-Smirnov measure. Since the operator may start the chromosome trace far within the dark area of the slide, patterns frequently start with a long, low amplitude lead (Fig. 9). The long lead in the curve must be eliminated before calculating any measure of similarity. To do this, the area under the curve from its beginning to a point, R, is assumed to be less than five per cent of the total area. All measures in this interval are then eliminated from the computation of the cumulative curve. Since banding patterns rise sharply, there is no concern that if a five per cent cutoff is applied to a trace without an amplitude lead, the Kolmogorov-Smirnov statistic will be affected. It is not necessary to perform the reverse procedure since the automatic cutoff in the acquisition program detects the boundary and eliminates the problem.

Pairing

After 46 chromosomes have been scanned and the digitized banding patterns placed on magnetic tape, the pairing algorithm is initiated (Fig. 10). Since 46 is the diploid number of chromosomes, the identification problem

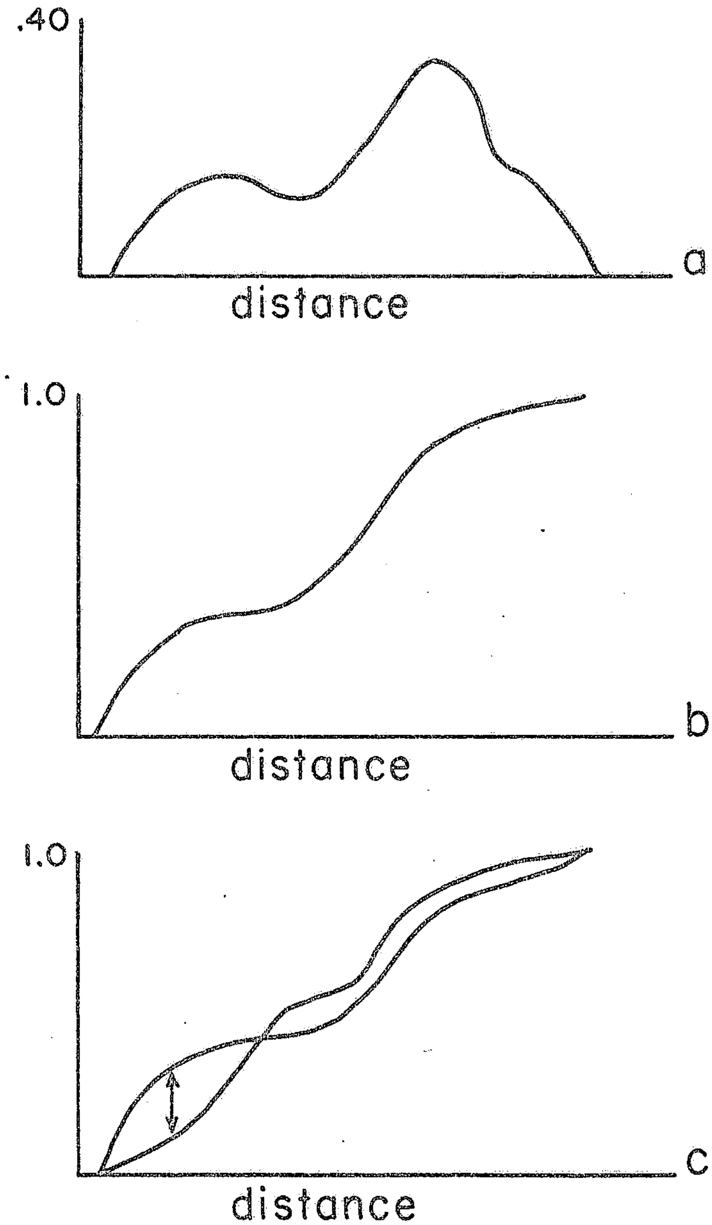


Fig. 9. Kolmogorov-Smirnov Measure -- To compute the Kolmogorov-Smirnov measure, each banding pattern is first normalized, (a), and its cumulative distribution determined, (b). The largest absolute difference between the cumulative distributions of two banding patterns becomes a measure of their similarity, (c).

Fig. 10. Pairing Algorithm for a Set of Four Pairs -- The original matrix, (a), contains "forward" comparisons in the upper triangle and "backward" comparisons in the lower triangle. The maximum value in the positions (i,j) and (j,i) replaces the value in (i,j), and the diagonal is ignored in (b). Row vectors and column vectors are created from a merit function in (c). The highest element of both these vectors points to the selected row or column of the matrix. The largest element in the indicated row or column is designated the "match" element. (Pre-processing of the matrix would occur immediately before this step when implemented.) In tableau (d) the selected match element has been flagged with a -1 and the corresponding rows and columns zeroed for ease of inspection. This procedure "deflates" the matrix, and the algorithm will ignore the values in these rows and columns.

	A	B	C	D	E	F	G	H
A	100	80	64	29	37	35	90	84
B	37	100	32	71	88	47	56	34
C	34	12	100	59	91	12	39	25
D	66	83	43	100	83	63	59	22
E	53	50	18	71	100	33	27	64
F	61	72	22	36	55	100	31	14
G	79	39	41	80	51	65	100	73
H	70	46	44	62	28	45	40	100

a

Fig. 10. Pairing Algorithm for a Set of Four Pairs

	A	B	C	D	E	F	G	H
A	–	80	64	66	53	61	90	84
B	80	–	32	83	88	72	56	46
C	64	32	–	59	91	22	41	44
D	66	83	59	–	83	63	80	62
E	53	88	91	83	–	55	51	64
F	61	72	22	63	55	–	65	45
G	90	56	41	80	51	65	–	73
H	84	46	44	62	64	45	73	–

b

Fig. 10.--Continued Pairing Algorithm for a Set of Four Pairs

	A	B	C	D	E	F	G	H	X
A	–	80	64	66	53	61	90	84	103
B	80	–	32	83	88	72	56	46	92
C	64	32	–	59	91	22	41	44	139
D	66	83	59	–	83	63	80	62	86
E	53	88	91	83	–	55	51	64	99
F	61	72	22	63	55	–	65	45	86
G	90	56	41	80	51	65	–	73	106
H	84	46	44	62	64	45	73	–	106
Y	103	92	139	86	99	86	106	106	

$$X_i = \frac{\sigma_i \max_j (M(i,j))}{\sigma_i'}$$

$$X_1 = \frac{13.59 \cdot 90}{11.78}$$

Pointer = MAX (X , Y)

$$Y_i = \frac{\sigma_j \max_i (M(i,j))}{\sigma_j'}$$

where σ' computes the standardized deviation of the vector without the highest element

C

Fig. 10.--Continued Pairing Algorithm for a Set of Four Pairs

	A	B	C	D	E	F	G	H
A	–	80	0	66	53	0	90	84
B	80	–	0	83	88	0	56	46
C	0	0	–	0	0	-1	0	0
D	66	83	0	–	83	0	80	62
E	53	88	0	83	–	0	51	64
F	0	0	0	0	0	–	0	0
G	90	56	0	80	51	0	–	73
H	84	46	0	62	64	0	73	–

d

Fig. 10.--Continued Pairing Algorithm for a Set of Four Pairs

can be simplified if the 46 banding patterns are reduced to 23 pairs. To find the 23 pairs it is necessary to compare each banding pattern with every other banding pattern. Here the computer provides a service that the human could not perform reliably in a reasonable time. Using a computer program a matrix of 46 x 46 comparisons can be formed. Each element of this matrix is the Kolmogorov-Smirnov measure between chromosome i and chromosome j. If it were necessary to make only one comparison between any two chromosomes, then the matrix would be symmetrical, that is position i,j in the matrix would have the same value as position j,i. However, the Kolmogorov-Smirnov measure is directional. Another way of expressing it is: we don't know the head of the chromosome from its tail. Thus, to make sure that both directions are considered, two comparisons are made: the first between patterns i and j, and a second comparison betweehn chromosome i and a reversed pattern j. Consequently, the lower triangle of the matrix contains the "forward" comparisons, the diagonal contains comparisons of chromosomes with themselves, and the upper triangle contains "backward" comparisons. The matrix now represents all possible measures between the chromosomes.

It is important to emphasize that this procedure is not intended to lead to an identification, only a pairing of one chromosome with a partner.

At the suggestion of the biological personnel who worked closely with the author, the matrix values were inverted and scaled to a maximum of one hundred. That is, a high measure of similarity was expressed as a value of one hundred; no similarity was reflected by a value of zero. This inversion in scaling aided discussion about the pairing process.

To reduce the number of comparisons, the matrix is first converted into upper triangular form. To do this, the value of each index pair, i,j, is replaced with a maximum value at either i,j or at j,i. This procedure chooses the maximum of the "forward" or "backward" values and makes this maximum the new entry in both positions. The justification for this is that either the chromosomes are a pair, or they are not. If they are not a pair, there should be low agreement in their comparison measure for both "forward" and "backward" values. If they are a pair, only one of the two values will reflect this information. By always choosing the highest value from every "forward-backward" pair, no correct identification will be omitted.

Since the matrix was symmetrical after choosing the maxima, the lower triangle and the diagonal of the matrix were zeroed. There are now (46 x 46)/2 - 46 elements in the matrix. Twenty-three of these elements must be selected to pair the chromosomes. The problem is further constrained. For any location, i,j, that is designated a "match," no other "match" may occur in row i, row j, column i, or column j. This constraint is imposed because the data assume that there are no trisomy syndromes and that the sex is female. With these constraints in mind, the following algorithm was used to identify the matrix.

The "best" element of the matrix was selected using an algorithm described below. The position, (i.j), of this "best" element was flagged with a -1. The remaining elements in the rows and columns indexed by i and j were zeroed. This process reduced the matrix by eliminating four "vectors" from the matrix. The method continues until the matrix is identified or until the "best" element falls below a critical value.

The search for a "best" value intuitively demands a merit function to have a very high selection value for one chromosome and to have low values for all other chromosomes. There is some physical information that can be added about chromosome structure to verify the choice. Chromosomes are grouped by letter and number. The letter indicates centromere location; the number is an index of length. These groupings by letter indicate that a chromosome should have several high values corresponding to other members of the same group. Suppose we now rank the j Kolmogorov-Smirnov values for a given chromosome, i. The highest value should match chromosome i. The second highest may be a member of the same group. If these two values are chosen, then it is difficult to decide if chromosome j_{high} or chromosome $j_{next\ high}$ should be selected for the match. If, however, these values show a great difference, then we can be more sure of our matching. Thus, one method is to consider a merit function comprised of the ratio of the maximum value divided by the second highest value for a given chromosome, i.

After first using this merit function, it was necessary to consider an index of its performance. If the matrix is constructed so that the identification by pathologists places the chromosomes in karyogram order ($A1_a, A1_b, A2_a, A2_b, \ldots, X_a, X_b$), then an identified matrix that coincided with human identification would have its flagged values -1 located just off the diagonal. The greater the disagreement between the algorithm and the human identification, the further the flagged values would be from the diagonal.

The first use of the algorithm was encouraging but placed too many flags (20%) outside the area of reasonable identification. Although other simple merit functions were tried, there was no marked improvement in the results. To cope with this difficulty further conditioning of the matrix was needed (Fig. 11).

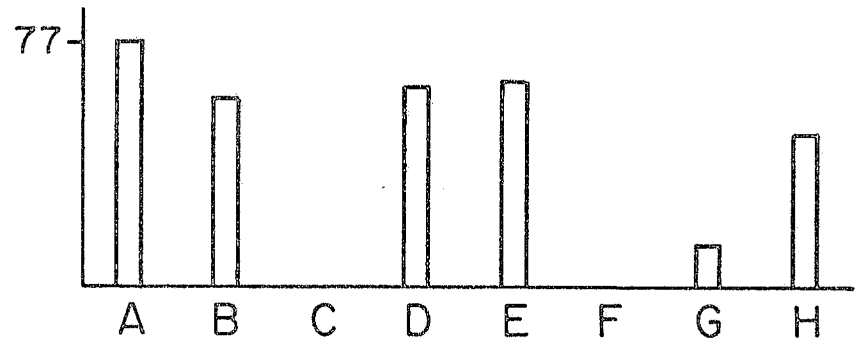
Examination suggested that the algorithm appeared sensitive to spurious high identification values caused by intra-group similarities. An example would be the pairing of an X chromosome with a member of the C class which it

	A	B	C	D	E	F	G	H
A	–	84	77	64	70	80	25	54
B	84	–	59	22	19	74	37	38
C	77	59	–	61	65	88	12	47
D	64	72	61	–	22	88	61	91
E	70	19	65	22	–	19	31	63
F	80	74	88	88	19	–	53	69
G	25	37	12	61	31	53	–	18
H	54	38	47	91	63	69	18	–

a

Fig. 11. Preprocessing of the Pair Matrix Using Squared Error Criterion -- The symmetrical matrix is taken from step (a) of the pair identification algorithm. To compute the element (C,F) of the new matrix, the C row and the F column are first converted to distributions (b). Elements constituting comparison values between C and F are excluded as are the diagonal values (C,C) and (F,F). The squared error term for the two generated distributions is computed in (c). The generated matrix is shown (d) after one complete step. The algorithm may be applied again to this new matrix using a different comparison measure. Note that the new matrix has a high-low reversal relative to the original matrix.

A	B	C	D	E	F	G	H
77	59	–	61	65	X	12	47



A	B	C	D	E	F	G	H
80	74	X	88	19	–	53	69

b

Fig. 11.--<u>Continued</u> Preprocessing of the Pair Matrix Using Squared Error Criterion

$$(C,F) = \left[(A_F - A_C)^2 + (B_F - B_C)^2 + \ldots (H_F - H_C)^2\right]^{1/2}$$

$$= \left[(80-77)^2 + (74-59)^2 + \ldots (69-47)^2\right]^{1/2}$$

$$= 72.42 = (F,C)$$

c

	A	B	C	D	E	F	G	H
A	—	71.3	30.6	88.5	99.8	66.4	92.2	43.0
B	71.3	—	67.7	61.9	62.59	80.1	90.5	96.4
C	30.6	67.7	—	87.8	92.4	72.4	80.0	47.65
D	88.5	61.9	87.8	—	80.6	65.1	103.0	66.6
E	99.8	62.5	92.4	80.6	—	92.3	99.2	91.4
F	66.4	80.1	72.4	65.1	92.3	—	116.0	83.2
G	92.2	90.5	80.0	103.0	99.2	116.0	—	65.1
H	43.0	96.4	47.6	66.6	91.4	83.2	65.1	—

d

Fig. 11.--<u>Continued</u> Preprocessing of the Pair Matrix Using Squared Error Criterion

closely resembles. A preprocessing of the matrix which would make the intra-group differences more pronounced would make the pairing algorithm stronger.

For each chromosome, i, there are 45 identification values. If these are ordered consistently, they may be thought of as a distribution. Since there are 46 chromosomes, there are 46 distributions. Suppose now that the distribution of elements for chromosome i and the distribution for chromosome j are compared for similarities with a statistic such as Kolmogorov-Smirnov or least squares. This new measure of similarity may be used as the new value of element i,j. Continuing in this fashion, a new set of values for the matrix may be generated. This new computation compares how a distribution of comparisons for a given chromosome compares with a distribution of comparisons for a second chromosome. This method should insure high values between real pairs by their similar response to other distributions. Non-pairs should have disparate similarities to other distributions resulting in a low value.

Several other developments make this method interesting. There is no reason that the matrix should not be recomputed several times. That is, the matrix may be computed recursively until identification is made by this process alone. Further, at each computational step a different statistic may be used. The introduction of a new statistic could be predetermined or based on the values in the preceding matrix. The selection of statistics in this recursion process may be made so that different types of error complement one another. For the problem at hand, a least squares measure and one recursion algorithm were sufficient to extract the X pair from the C group.

Identification

Complete identification requires that the chromosome pairs are matched to the 23 classes. Identification is constrained so that there may be one and only one match for each class. As discussed earlier in this paper, partial identification is preferrable to complete identification if it prevents the possibility of spurious matches.

A 23 x 23 matrix represents comparison values between the 23 classes and the 23 input pairs. The matrix identification is similar to the procedure used in the pairing of the chromosomes. The "best" element of the matrix is chosen by a merit function. If the value generated by the merit function is above a criterion, the element is considered to represent a match and is flagged with a -1. Since this element is a match between chromosome i and class j, chromosome i cannot be matched with any other class; further, class j cannot be matched with any other chromosome. These restrictions allow row i and column j of the matrix to be zeroed and reduce the dimensions of the matrix to 22 x 22. This procedure continues until the matrix is identified or until the value of the merit function drops below a threshold.

The criterion for deciding whether or not to halt the algorithm must come from the merit function. Consider any row or column in a matrix. Assume that the values in the matrix are generated as before: zero represents dissimilarities and 100 represents complete similarity. In a given row or column vector the largest value represents the best comparison. If a vector consisted of one value that was very high and of all other values that were very low, the choice would be easy since only one of the comparisons may be regarded as a candidate for the match. If this model is considered an ideal, the criterion value should be derived from it. The criterion should have a low value for either a vector with many low entries or a vector which more than one high entry. A merit function which reflects this criterion is

$$e = \frac{\sigma}{\sigma'} \text{MAX (vector)}$$

where σ is the standard deviation of the vector entries and σ' is the standard deviation of the vector entries after the highest entry has been removed.

If the vector is homogeneous, that is, if no single high element exists, the ratio of σ/σ' will be small since $\sigma' \cong \sigma$. If a single high element does exist, then $\sigma >> \sigma'$. The ratio is weighted by the value of the highest element to preclude a set of all low values from giving a spuriously high figure.

To insure low risk the criterion is based on two separate factors. First, the merit function must generate a large enough value to satisfy one criterion parameter. Second, the value of the highest element in the vector must be greater than a specified value of the second criterion parameter.

The merit function is computed for every row and for every column in the matrix. The vector with the highest merit function value is examined. If it fails either of the criteria, the next highest vector is considered. The vectors are considered in turn until one is found which passes both criteria. If all vectors fail the criteria, the algorithm halts. Partially identified matrices are brought to the attention of the pathologist (Fig. 12).

Like the pairing algorithm, this identification procedure is based on the use of similarity measures for the elements of the identification matrix. For the reasons outlined earlier, the Kolmogorov-Smirnov statistic was used.

Each standard chromosome must be compared with each pair of chromosome inputs. Since the pair is represented by two banding patterns, it is necessary to decide which of these, or their average, should be used for the comparison. An ensemble average over the pairs is an attractive answer. However, the stretching and bunching of the DNA within the

Fig. 12. Sequential Reduction Procedure for a 4x4 Matrix -- The row and column vectors are generated from the original matrix in (a). As in the pairing algorithm, the highest element in either vector points to the identifiable matrix element. In the next step, (b), the row and column of the first identified element have been zeroed, and the element has been flagged with a -1. Tableau (c) shows the resultant matrix.

a

	A	B	C	D	
E	87	42	93	27	87
F	38	24	91	47	219
G	57	39	56	68	80
H	63	80	81	95	123
	135	198	87	134	

Maximum element in both vectors points to highest element

$$e = \frac{\sigma_j \underset{i}{\mathrm{MAX}}\,(M(i,j))}{\sigma_j'}$$

where

σ_j = standard deviation in the marginal vector

σ_j' = standard deviation in the marginal vector computed without entry $\underset{i}{\mathrm{MAX}}\,(M(i,j))$

Fig. 12. Sequential Reduction Procedure for a 4x4 Matrix

	A	B	C	D	
A	87	42	0	27	256
B	0	0	1	0	–
C	57	39	0	68	78
D	63	80	0	95	126
	324	862	–	112	

b

	A	B	C	D
A	1	0	0	0
B	0	0	1	0
C	0	0	0	1
D	0	1	0	0

c

Fig. 12.--Continued

chromosome prohibits this because it produces functions with different numbers of values. (A great deal of research has dealt with this problem. A practical application is to be found in the analysis of ECG and EKG waveforms. To date no satisfactory method exists for averaging functions of these forms.) Still, the ith pair and the jth standard must be compared. Arbitrarily, one banding pattern from each pair is chosen at random, and the Kolmogorov-Smirnov statistic is computed between it and the prototype pattern. The error incurred by this choice will be discussed in following paragraphs.

ERROR EVALUATION

Sources of Error

Data errors may occur at many different places. Biological data are characterized by high inherent variability. The photographic recording and processing may introduce errors in the banding patterns. Optical and electronic noise may affect the values of the banding patterns. Operator error may introduce poor trace selection. Errors may be present in the sample chosen as a set of prototype chromosomes.

Gross morphological anomalies in the chromosomes will be assumed to be detected by one of the operators. Other abnormalities in the DNA structure cannot be detected with present technology, and this risk must be accepted. Photographic negatives that pass a careful visual inspection will be assumed to contain errors of low magnitude. Optical and electronic noise will be assumed to be random and to introduce an error of less than one per cent.

Operator tracing and adjustment errors were prevented in this feasiblity study by using two operators to check one another. The standard chromosome traces were assumed to be correct as received. Digitizing of these standard chromosomes was done with a Y-axis resolution of ten-bit precision over a ten inch distance. X-axis resolution was fifty points per inch. Apart from the physical and operator introduced errors are the theoretical errors in the algorithm.

The most probable source of error in the reduction of paired chromosomes is within-group switching. As described before, this error occurs when an $A1_a$ is matched with an $A2_b$ and the $A2_a$ is matched with the $A1_b$ or in other similar arrangements. This is a troublesome problem. At this point in the procedure no identification has yet taken place. Although two matches are switched, it is entirely possible that this matching is correct. Only secondary inspection of other measures such as centromere location can spot potential mis-identifications.

Errors outside of groups are more readily detected by their identification distance from the diagonal. Moreover, these identifications can point to the within-group pairing errors just described. Since both a correct match and a within-group mismatch will zero corresponding rows and columns, a mismatch will eliminate vital data from the matrix. These missing data further degrade the matrix. The algorithm stops very soon when the contour criteria are not met.

Before the identification procedure can start, yet another source of potential error is introduced. The pairs of chromosomes must be matched with one of the 23 classes. Since contraction of the DNA prevents any meaningful average between the two banding patterns in the pair, it is necessary to choose one of the pairs at random. This choice is not so indiscriminate as it may first appear. Either the members of a pair are greatly alike, or they are not. If they are greatly alike, it makes no difference which one is chosen since either is a good representative. If they are greatly different, then there are two further possibilities. The pattern chosen may be either a better or a worse match with its (as yet unknown) standard. If it promotes a high score of similarity between it and the standard, all is well. If it matches poorly with its standard, its value should not harm the algorithm, because a low value will produce a correspondingly low merit function value. At worst, this will halt the algorithm during identification. The danger occurs if the banding pattern chosen spuriously resembles another chromosome with a high value. The likelihood of this situation is rare. If the merit function is high, it is very probable that a correct match has occurred. If the merit function is low, the selection of that element will be correspondingly improbable. Finally, spurious similarity is unlikely from a physical standpoint. Examination of many banding patterns suggests that it is difficult to find high similarity values between nonpaired chromosomes and that a spurious similarity would be most rare.

The alpha error is of prime importance to the biologist during the identification algorithm. The probability of wrongly accepting the identification of a chromosome should be very low. For this stringent restriction of the alpha level the biologist is willing to accept substantial beta error even if that acceptance will cause the algorithm to halt after only partially identifying the chromosome spread.

The sequential reduction of the matrix selects the highest element in the best vector in the matrix. The choice of the best vector is determined by a merit function. Two criteria must be passed for the algorithm to make an identification. The highest element must have a value equal to or greater than the value specified. The vector selected must have a merit function value greater than or equal to a specified value.

At each reduction step N + N - 1 elements of the matrix are eliminated. This is almost ten per cent of the elements in the matrix. If the wrong element is selected, necessarily some elements which point to other identifications will be eliminated. Thus, if a vector passes the criteria and the wrong element is identified, eventually the criteria will halt the algorithm because key values for other identifications have been eliminated. These error conditions indicate that a partial identification should be carefully checked. This checking should be done as a matter of course since all partially identified spreads will be called to the attention of the cytogeneticist. A low frequency of occurrence of partially identified spreads is acceptable for a screening algorithm.

In addition to the checks described above, there is assupplementary error assessment which can be used to cross-check the results or can be used within the algorithm to enhance identification. Inspection by pathologists or by trained technicians can produce an a priori probability estimate of the correctness of each chromosome's label. These probabilities may be compared with the results of the identification to provide verification. The estimates may also be incorporated into the algorithm by premultiplying their values with each column of the identification matrix. Neither of these applications is suitable for screening, but the algorithms are flexible enought to incorporate them for investigatory purposes.

Whether used for investigation or for automated screening, the algorithm is sensitive to errors in the measure of similarity. Lindgren (1968) shows that for a given level of false rejection of the null hypothesis, α, a statistic, D, may be computed. The null hypothesis, in this case similarity, is rejected if

$$D > \{-1/2 \cdot (1/m+1/n) \cdot \log\alpha/2\}^{1/2},$$

where D is defined as

$$D = \max_{\text{all } x} |F_m(x) - G_n(x)|$$

and where m and n are the sample sizes of the respective distributions.

<u>Testing the Identification Algorithm</u>

The algorithm was tested by comparing a known set of chromosome data with a degraded version of the same set, Il. A set of banding curves was chosen at random from one of the measured sets. A copy of each curve was made. Two separate procedures were used to systematically degrade the second set.

To test the sensitivity of the algorithm to shifts in the higher frequencies, a linear approximation was used (Fig. 13). For a given banding pattern in the second set, a sampling distance, K, was chosen. The banding patterns were sampled at every Kth point. Points between the values at the n and n+1st points were approximated by a straight line. Obviously, as the value of K grew larger, fewer samples were taken and the second pattern became more and more linear. The end points of the pattern were always forced to be members of the set of colocation points. To do this the algorithm permits the distance between the last two points to be less than K if necessary. This procedure was applied to every member of the second set (Fig. 14).

After the second set had been degraded by a particular amount, the assign algorithm performed the identification. Every member of the original set was compared with

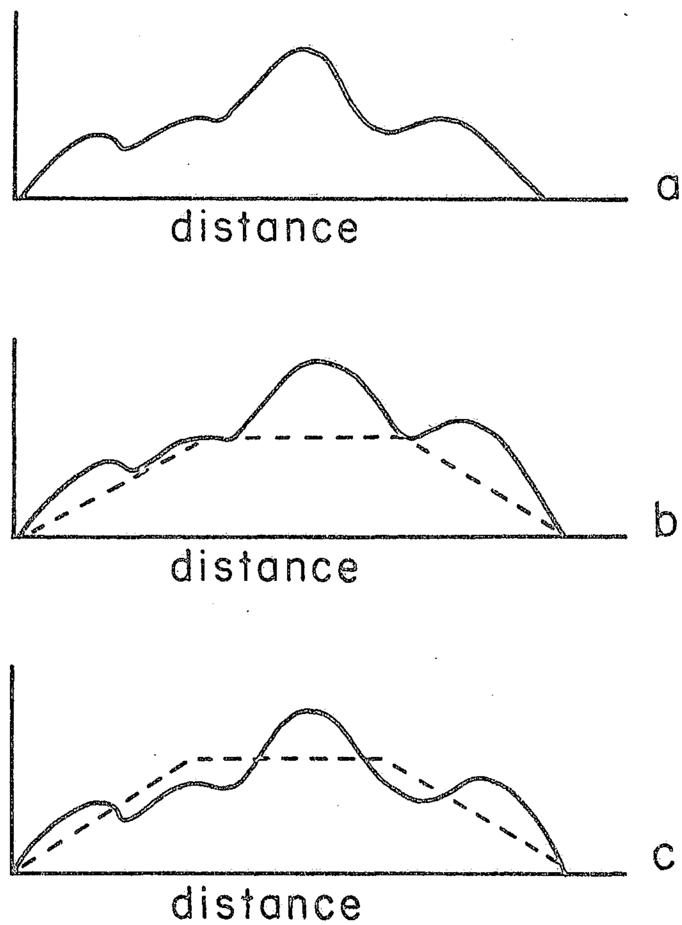


Fig. 13. Piecewise Linear Testing -- This method degrades an original banding pattern, (a), by sampling the pattern at an interval, K, and constructing linear approximations between the value of the function at those points, (b). Before comparisons are made, the degraded pattern is scaled so that it has the same area as the original pattern, (c).

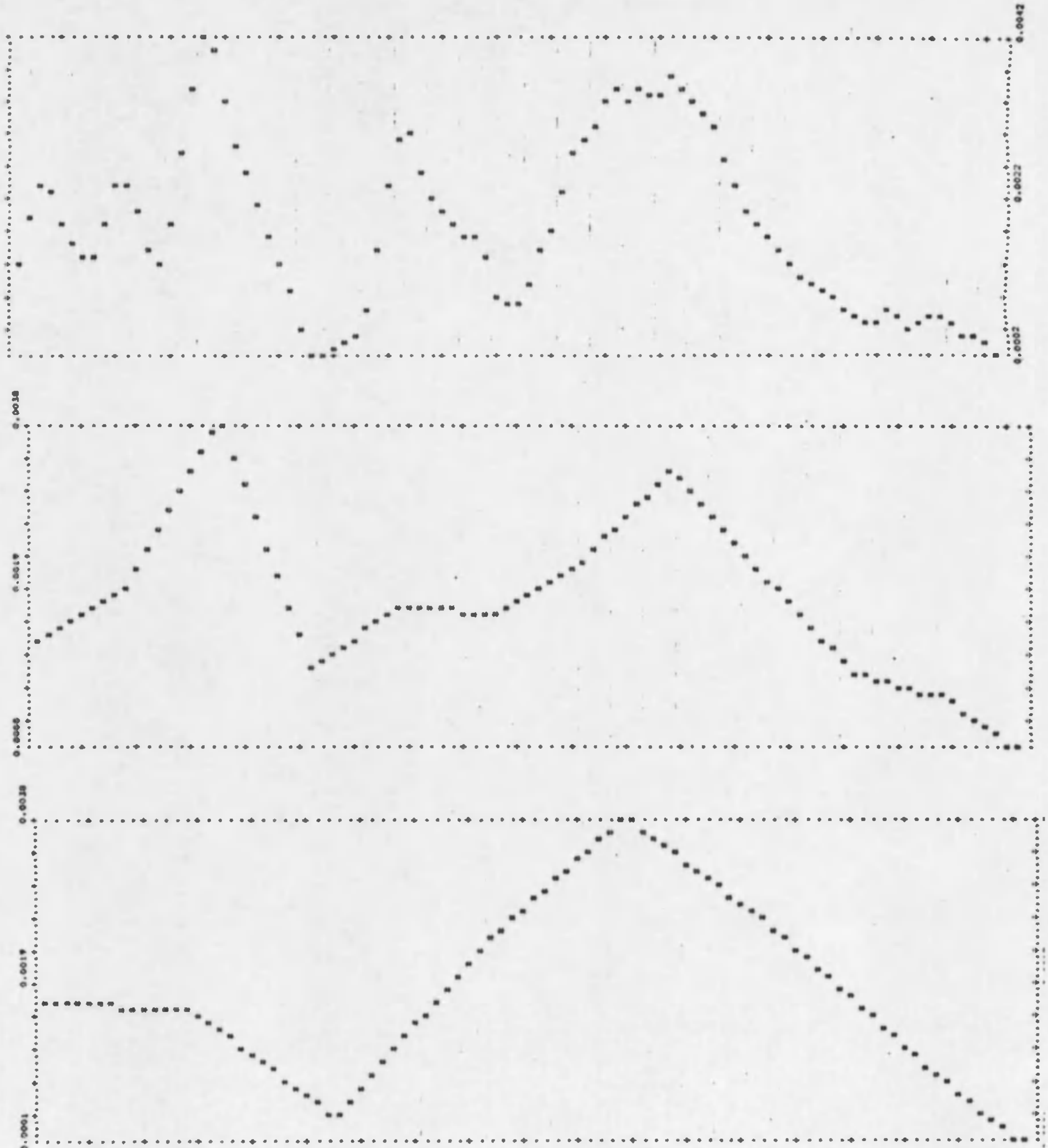


Fig. 14. Linear Degradation of Data -- As the sampling rate, K, increases, each banding pattern becomes more linear. An Al chromosome was sampled at distances of (a) 5, (b) 50, and (c) 100 using this method.

every member of the degraded set by computing the Kolmogorov-Smirnov value between the banding patterns and placing it in the identification matrix (Fig. 15). The matrix was identified with no constraints. The number of correct identifications was plotted as a function of the values of K (Fig. 16). The rapid breakdown in the curve reinforces the argument that misidentification can easily be spotted or controlled by reliability parameters.

To test the algorithm for sensitivity to low-frequency changes, such as shifts in the distribution of DNA, an orthogonal series method was used to degrade members on the second set of chromosomes. A Fourier series consisting of 14 sine coefficients, 14 cosine coefficients, and the DC term was used to describe each of the chromosomes. The fundamental coefficients for both the sine and cosine terms were altered by a percentage, P. The altered set of coefficients was then used to construct a degraded chromosome resembling the original (Fig. 17). As before, each member of the set was compared with its degraded copy using the identification matrix discussed earlier. Again, the number of correct responses was plotted as a function of the value of the percentage of degradation (Fig. 18). Note that large percentages must appear in the fundamental frequencies over all chromosomes before identification fails.

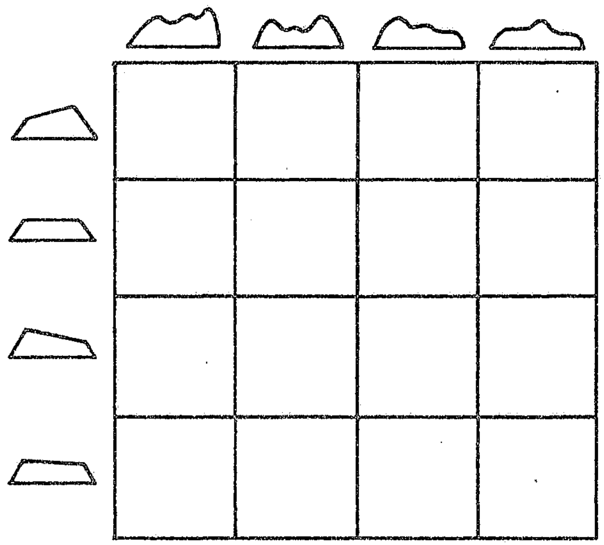


Fig. 15. Comparison of Degraded Chromosome Set with a Set of Chromosomes -- A set of chromosomes was used as paradigms. These were linearly degraded and matched with a degraded set. Corrected identifications would occur along the diagonal.

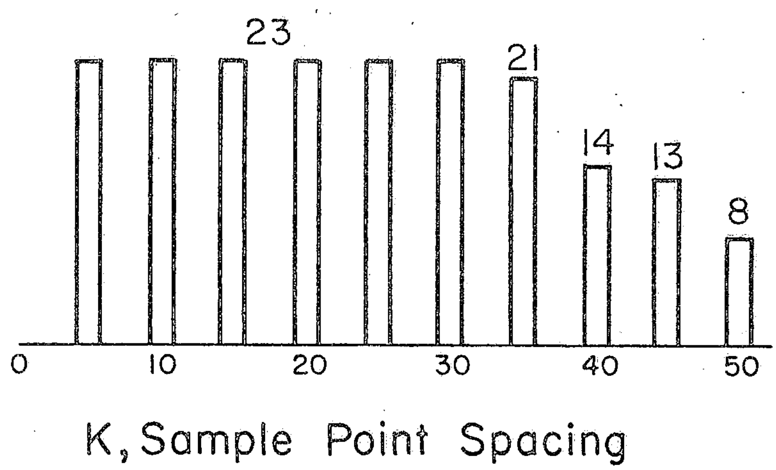


Fig. 16. K Graph for Linear Approximation Test -- Identification becomes more difficult as the sample point spacing increases. A spacing at every 35th sample point misidentifies only two of the 23 chromosomes. Larger spacings cause a rapid decrease in the number of correct identifications.

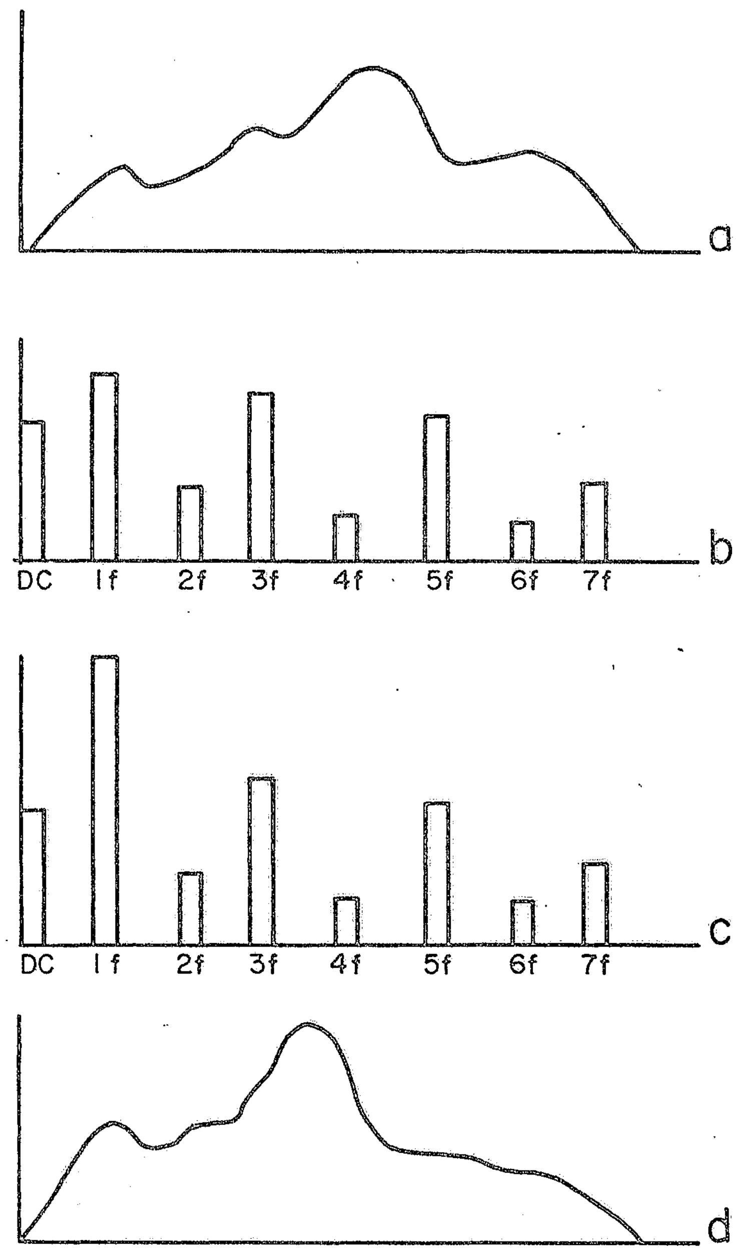


Fig. 17. Fourier Series Testing -- This method converts an original banding pattern, (a), into a Fourier series, (b), and then increments a percentage to the coefficient of the fundamental frequency, (c). The degraded pattern is then reconstructed from the degraded series representation, and the area is normalized to that of the original pattern, (d).

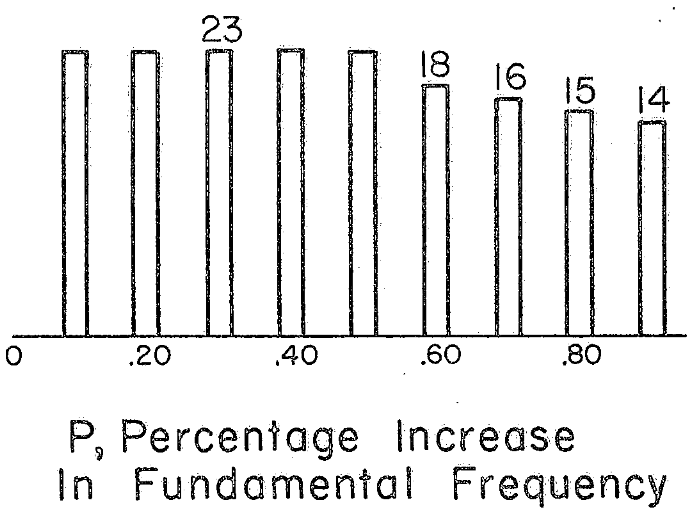


Fig. 18. P Graph for Fourier Analysis -- All 23 chromosomes were correctly identified with as much as a 50% change in both the sine and cosine coefficients at the fundamental frequency. With a 90% change the algorithm still identified 14 of the 23 banding patterns.

DISCUSSION AND CONCLUSIONS

An evaluation of the work presented here falls into three parts: (1) the assign algorithm, (2) the pair algorithm, and (3) the acquisition of data.

The assign algorithm is the basis for all identification. This algorithm has been evaluated with real chromosome data using two different methods to degrade the set of banding patterns systematically. Inspecting the matrix showing the first occurrence of failure in the linear degradation (see Fig. 16), one may note that in spite of the high similarity of values in the matrix, the algorithm made only two mistakes (Fig. 19). This level of performance suggests a robust algorithm which is insensitive to high-frequency noise such as the noise introduced by the data acquisition procedure.

The second test of the assign algorithm considered the sensitivity to shifts in DNA content. A 50% change in both the sine and cosine coefficients at the fundamental frequency had no effect on the algorithm's identification capability. These results suggest that the assign algorithm does provide the decision procedure necessary to the identification of chromosomes with standards. Further, the algorithm can be expected to improve as better measures of similarity are introduced or as more sensitive and specific

97.	90.	92.	88.	89.	87.	84.	88.	92.	89.	80.	90.	88.	92.	92.	93.	89.	86.	91.	91.	84.	90.	96.
91.	96.	94.	93.	94.	92.	85.	93.	95.	89.	86.	90.	91.	93.	90.	94.	92.	92.	91.	93.	88.	92.	91.
87.	90.	89.	94.	92.	88.	80.	92.	90.	84.	88.	86.	87.	89.	86.	89.	87.	92.	90.	92.	91.	88.	89.
90.	96.	94.	92.	96.	93.	87.	95.	95.	90.	86.	93.	93.	94.	90.	93.	93.	93.	91.	93.	88.	92.	89.
90.	92.	93.	92.	97.	95.	85.	95.	96.	91.	84.	89.	90.	94.	92.	93.	90.	91.	92.	93.	88.	94.	88.
90.	92.	93.	93.	93.	91.	86.	95.	92.	91.	86.	90.	91.	94.	90.	94.	91.	91.	94.	94.	88.	89.	89.
90.	89.	90.	85.	84.	91.	93.	88.	89.	93.	79.	94.	93.	90.	93.	90.	92.	84.	89.	87.	81.	91.	89.
89.	94.	91.	91.	96.	95.	88.	96.	94.	91.	87.	93.	92.	94.	90.	91.	92.	93.	91.	91.	88.	92.	87.
90.	93.	93.	90.	94.	97.	87.	93.	97.	94.	82.	92.	90.	93.	94.	93.	91.	88.	91.	93.	86.	96.	88.
94.	93.	95.	90.	93.	94.	88.	92.	95.	94.	82.	92.	91.	96.	94.	96.	92.	88.	95.	93.	86.	95.	92.
83.	88.	89.	95.	90.	86.	81.	92.	87.	83.	96.	86.	87.	89.	83.	89.	88.	92.	90.	91.	93.	84.	82.
93.	92.	92.	86.	90.	93.	87.	89.	94.	94.	80.	94.	91.	91.	95.	91.	92.	85.	90.	89.	82.	95.	91.
91.	94.	93.	90.	93.	94.	88.	94.	93.	94.	83.	96.	96.	94.	94.	94.	97.	89.	92.	93.	86.	95.	89.
92.	95.	96.	91.	93.	94.	91.	95.	93.	93.	84.	95.	97.	97.	92.	96.	96.	90.	95.	94.	87.	91.	91.
92.	93.	92.	88.	91.	94.	89.	91.	95.	96.	81.	95.	92.	91.	96.	92.	94.	87.	90.	90.	83.	97.	90.
92.	92.	93.	88.	91.	91.	92.	91.	93.	93.	80.	94.	94.	94.	94.	95.	94.	86.	93.	91.	84.	93.	93.
90.	94.	92.	90.	92.	92.	87.	93.	92.	96.	83.	97.	95.	93.	95.	92.	96.	89.	91.	92.	86.	94.	89.
86.	91.	87.	86.	93.	94.	82.	94.	90.	91.	85.	92.	90.	89.	91.	87.	92.	95.	86.	87.	84.	90.	85.
91.	93.	94.	95.	91.	90.	89.	91.	93.	91.	88.	92.	95.	97.	90.	96.	94.	91.	98.	96.	90.	90.	90.
91.	94.	96.	92.	92.	91.	89.	92.	94.	92.	85.	93.	92.	97.	91.	97.	92.	91.	97.	96.	89.	91.	90.
90.	95.	94.	91.	92.	91.	85.	92.	94.	90.	83.	90.	91.	92.	89.	93.	92.	90.	91.	93.	86.	91.	91.
91.	96.	94.	88.	94.	93.	87.	91.	94.	90.	81.	92.	93.	91.	91.	92.	94.	87.	90.	91.	84.	93.	90.
90.	86.	88.	89.	84.	83.	85.	85.	87.	85.	81.	86.	89.	90.	88.	91.	87.	85.	92.	91.	85.	85.	92.

-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.	0.
0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-1.

Fig. 19. Comparison and Identification Matrices -- Testing the algorithm with a piecewise linear approximation gave the first misidentification at a spacing of K=35. (a) shows the corresponding matrix of comparisons, and (b) shows the resulting identification matrix. Note the distribution of values in (a) has a narrow range.

gain functions are found. An example of an improved comparison would be the use of the Kolmogorov-Smirnov statistic without first normalizing the area under the distribution to unity. This method would introduce a sensitivity to scale which does not exist in the present algorithm.

The evaluation of the pairing algorithm must be considered deparately. The pairing algorithm successfully identified 42 out of 46 chromosomes or 21 out of 23 pairs. The performance of the pairing algorithm must be perfect for its results to be useful when passed to the assign algorithm. After detailed inspection the author feels that this level of performance can be obtained by the implementation of a better means of similarity such as that outlined above and by applying preprocessing techniques described earlier. The large size of the matrix makes it especially sensitive to any noise. The results of the preprocessing techniques may be best shown by a set of noisy data. Initial identification of the matrix of data so perturbed gave nine misidentified chromosomes, among these a misidentification of the X chromosome with members of the C group. After one recursion of the preprocessing algorithm, the number of errors was reduced to four, and the X chromosomes were properly identified (Fig. 20).

The author wishes to include in this paper a proposal for the development of a set of standards: let the standards be a set of curves, possibly orthogonal, or

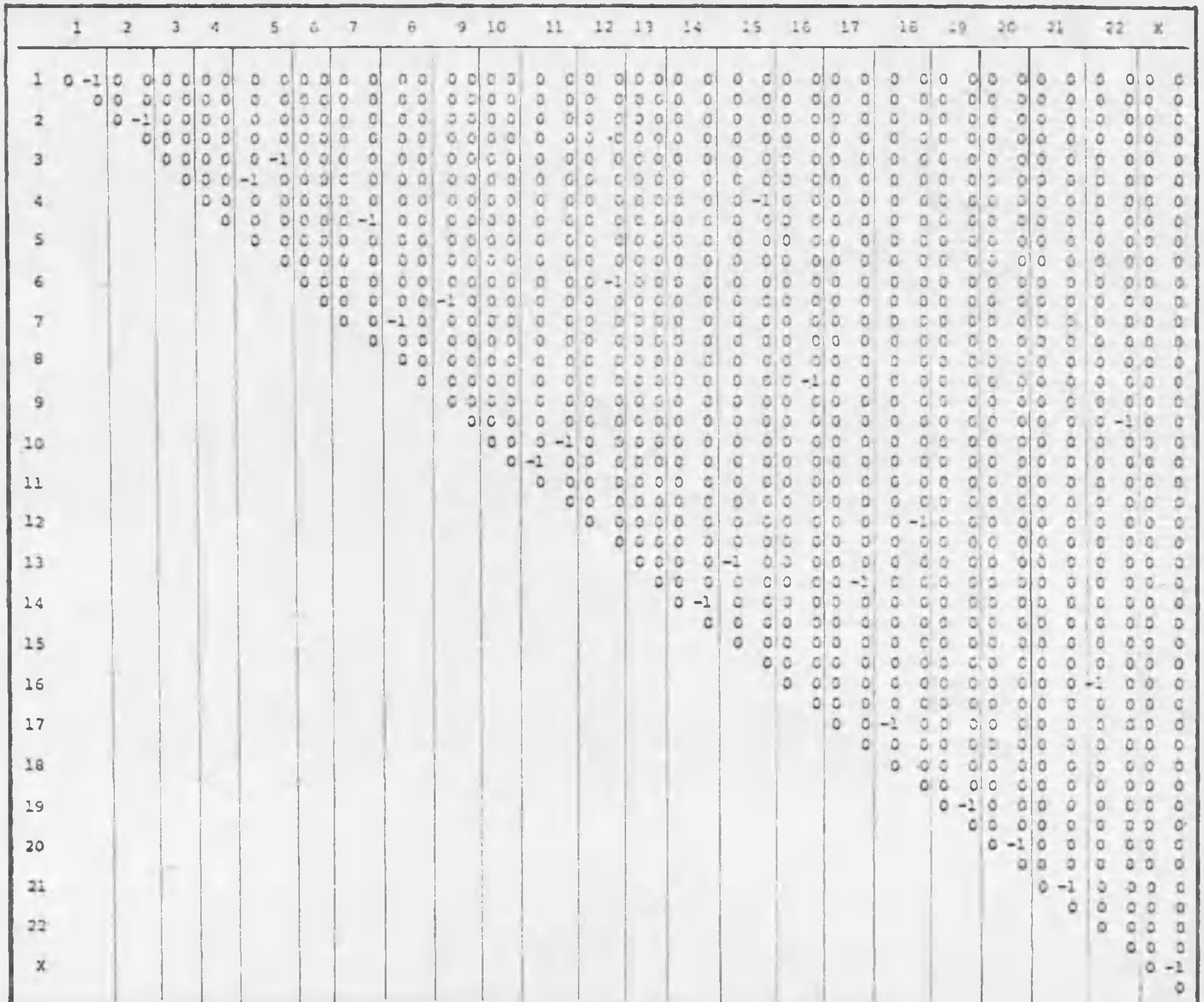


Fig. 20. Pair Matrix -- A matrix of identified pairs shows four misidentifications out of 46. The most "serious" misidentification was the classification of the X chromosomes with the C group. Application of the preprocessing step later paired the X chromosomes together.

perhaps disparate in shape. Compare a large number of chromosome sets with these standards. The standard curves should be chosen so that they discriminate between each of the chromosome types. This is essentially a clustering algorithm where the objective is to create 23 sets such that all banding patterns are members of one and only one set. Then, by agreement, the Al_a chromosome will be that chromosome which compares with the standards in a specific manner, e.g., a vector whose distance is within a certain value of the standard cluster. Note that it is not necessary to have 23 standards; indeed, for a given statistic, one standard curve might be sufficient.

Two auxiliary methods may help implement the identification of chromosomes using standards. These methods are beyond the scope of this paper, but a brief inspection of each of them may show additional flexibility of the algorithms.

Like the pair matrix, the identification matrix may be preprocessed to enhance the chances of successful identification. An obvious condition is the premultiplication of the matrix by the a priori probabilities due to a separate measure such as length or centromere index. An extension of this method is to partition the matrix into submatrices based on these probabilities. This gives n-1 degrees of freedom for each submatrix and reduces the possibilities of error.

A second technique is to use a modification of the recursion method discussed in the pairing matrix. Unlike the pair matrix, the identification matrix has separate sets of inputs and outputs. To use the recursion scheme an auxiliary matrix must be constructed. This matrix is composed of comparisons of standards with themselves. Returning to the original matrix, input i may be represented by the distribution of its values with each of the standards. This distribution may be compared with the distribution of standard j with each of the other standards in the auxiliary matrix. For a squared error comparison of the identification matrix, I, and the auxiliary matrix, A, the algorithm may be written:

$$I(m,n) = \sqrt{\sum_{i=1}^{N} [A(m,i)-A(n,i)]^2}$$

A different auxiliary matrix must be constructed for each statistic used. Implementation of this method more than once is also questionable since the auxiliary matrix will not change in the same manner as the identification matrix. However, implementation of the initial step might be fruitful.

The overall evaluation of this system is as follows: The data acquisition stage was entirely successful. The application of a human operator for several tasks gave good reliability and insured financial feasibility. The choice of the screen display and joystick control aided the operators in the acquisition of data.

The algorithms were successful from a methodological viewpoint. Based on the data acquired, they give promise of fulfilling the time constraints and the error criterion necessary for a screening operation. Further, the system showed an overall flexibility which would permit detailed inspections of chromosomes by the cytologist. The cost was well within normal expenditures for laboratory instrumentation.

Therefore, a screening system for the identification of human chromosomes is feasible under the constraints outlined and subject to the development of standards for comparison.

APPENDIX

COMPUTER PROGRAMS

```
      COMMON IC,B,MBLOK,NUM,INT,MINT,IP,ID
      DIMENSION IA(2,300),LK(4),B(170),ID(2,46),IC(2,500)
888   FORMAT(15)
165   FORMAT(10X'SATISFIED?')
26    FORMAT('BOUNDARY')
      NUM=1
      READ(1,888) MBLOK
1     PAUSE
      WRITE(1,92) NUM,MBLOK
92    FORMAT('CHROMOSOME NUMBER',I3,3X,'BLOCK NUMBER',I4)
91    CALL LFN(3)
      CALL LFN(1)
      IP=IXL(11)
      IF(IP) 1,93,93
93    Y=0.
      DO 33 I=1,20
33    Y=Y+FADC(12)
      Y=Y/20.
      IF(Y)91,25,25
25    CALL LFN(7)
      CALL LFN(1)
      WRITE(1,26)
2     MM=0
      NN=0
      INT=0
      KSW=0
      ISW=0
      JSW=0
      LL=201
      MAXX=0
      MAXY-0
      MINY=1000
      MINX=1000
      LK(1)=2
      LK(2)=3
      LK(3)=6
      LK(4)=7
      L=2
3     K=LK(L)
      CALL LFN(5,4)
      CALL LFN(K)
      CALL LFN(1)
      Y=0.
      I=0
      IP-IXL(11)
      IF(IP) 1,4,4
4     I=I+1
      Y=Y+FADC(12)
      IF(I-20) 4.5,5
5     KSW=0
```

```
67    Y=Y/20.
      IF(Y) 6,6,8
6     ISW=0
      LL=LL-1
      JSW=JSW+1
      IF(JSW-4) 12,23,23
23    JSW=0
      LL=LL-2
      GO TO 12
8     IF(ISW) 9,9,10
9     INT=INT+1
      IF(MM-MAXX)30,30,29
29    MAXX=MM
30    IF(MM-MINX)31,34,34
31    MINX=MM
34    IF(NN-MAXY) 36,36,35
35    MAXY=NN
36    IF(NN-MINY) 64,66,66
64    MINY=NN
66    IA(1,INT)=6*MM
      IA(2,INT)=6*NN
10    ISW=ISW+1
      JSW=0
      LL=LL+1
      IF(ISW-4) 12,11,11
11    LL=LL+2
      ISW=0
12    L=IREM(LL/4)+1
      IF(L-2)13,14,15
13    NN=NN+1
      GO TO 18
14    MM=MM+1
      GO TO 18
15    IF(1-3) 18,16,17
16    NN=NN-1
      GO TO 18
17    MM=MM-1
18    DO 22 I=1,5
22    CALL FDIS(IA,INT)
      IF(INT-40) 3,3,7
7     IF(MM)3,19,3
19    R=FLOAT(NN)
      R=ABS(R)
      IF(R-2.)20,20,3
20    WRITE(1,165)
864   IP=IXL(13)
      IF(IP) 44,44,839
44    CALL FDIS(IA,INT)
      GO TO 864
```

```
839   READ(1,888) J
      IF(J) 21,1,21
21    ISC=6
      DO 27 I=1,INT
      IA(1,I)=IA(1,I)/ISC
      IA(2,I)=IA(2,I)/ISC
27    CONTINUE
      MAXX=MAXX-MINX
      MAXY=MAXY-MINY
      IF(MAXX-MAXY) 65,65,48
48    ISC=MAXX
      GO TO 28
65    ISC=MAXY
28    IF(ISC-25) 85,85,37
37    ISC=8
      GO TO 24
85    ISC-12
24    DO 32 I=1,INT
      IA(1,I)=IA(1,I)*ISC
32    IA(2,I)=IA(2,I)*ISC
199   KSC=ISC/4
      K=20
      L=1
      INT=INT+1
      MINT=INT
      ISW=-1
      CALL LFN(5,K)
      CALL LFN(7)
      CALL LFN(1)
      MM=-20
      IA(2,INT)=NN*KSC
      IA(1,INT)=-20*KSC
200   IP=IXL(13)
      IF(IP)201,201,202
201   CALL FDIS(IA,INT)
      GO TO 200
202   READ(1,888)IY
      IF(IY-7)223,224,223
223   IF(IY-6)218,219,218
219   WRITE(1,210)
210   FORMAT('TYPE IN STEP SIZE')
      READ(1,888)K
      CALL LFN(5,K)
      GO TO 200
218   IF(IY-5)212,213,212
213   ISW=0
      GO TO 200
212   IF(IY)211,80,211
211   IF(IY-1)203,208,203
203   IF(IY-2)204,217,204
```

```
204   IF(IY-3)205,206,205
205   MM=MM+K
      CALL LFN(3)
      CALL LFN(1)
      GO TO 214
206   MM=MM-K
      CALL LFN(7)
      CALL LFN(1)
      GO TO 214
217   NN=NN-K
      CALL LFN(6)
      CALL LFN(1)
      GO TO 214
208   NN=NN+K
      CALL LFN(2)
      CALL LFN(1)
214   IF(ISW) 215,209,215
209   INT=INT+1
      Y=0.
      DO 216 J=1,20
216   Y=Y+FADC(12)
      B(L)=Y/20.
      L=L+1
215   IA(1,INT)=MM*KSC
      IA(2,INT)=NN*KSC
      GO TO 200
80    WRITE(1,165)
81    IP=IXL(13)
      IF(IP) 82,82,83
82    DO 84 I=1,20
84    CALL FDIS(IA,INT)
      GO TO 81
83    READ(1,888) J
      IF(J) 99,207,99
207   INT=MINT-1
      CALL LFN(4)
      IF(MM) 231,232,233
231   J=3
      MM=-MM
      GO TO 234
233   J=7
234   DO 50 LZ=1,MM
      CALL LFN(J)
      CALL LFN(1)
50    CONTINUE
232   IF(NN)235,199,236
235   J=2
      NN=-NN
      GO TO 237
236   J=6
```

```
237   MM=NN
      NN=0
      GO TO 234
99    IP=0
      XY=B(1)*200.
      K=INT-1
      L=MINT+1
      DO 101 I=L,K
      Y=FLOAT(IA(2,I)-IA(2,I-1))
      X=FLOAT(IA(1,I)-IA(1,I-1))
      D=SQRT(X*X+Y*Y)
      IG=IFIX(D)
      XX=B(I=MINT+1)*200.
      S=(XX-XY)/D
      DO 102 J=1,IG
      IP=IP+1
      IC(1,IP)=IP-200
      Z=XY+S*FLOAT(J-1)
102   IC(2,IP)=IFIX(Z)
101   XY=XX
      WRITE(1,165)
103   J=IXL(13)
      IF(J) 104,104,105
104   CALL FDIS(IC,IP)
      GO TO 103
105   READ(1,888) J
      IF(J) 207,207,106
106   CALL CHAIN('WRIT')
      GO TO 1
224   IP=IXL(11)
      IF(IP) 80,80,221
221   CALL LFN(6)
      CALL LFN(1)
      INT=INT+1
      IA(1,INT)=MM*KSC
      Y=0.
      NN=NN-K
      IA(2,INT)=NN*KSC
      DO 222 I=1,30
      Y=Y+FADC(12)
222   CALL FDIS(IA,INT)
      B(L)=Y/30.
      L=L+1
      GO TO 224
      END
```

```
      COMMON IC,B,MBLOK,NUM,INT,MINT,IP,ID
      DIMENSION B(170),ID(2.46),IC(2,500)
56    WRITE(1,30)
30    FORMAT('TYPE A 1 FOR A HARD COPY, OTHERWISE TYPE 0')
      READ(1,40) J
40    FORMAT(I5)
      IF(J) 24,24,53
53    DO 54 J=1,IP,3
      I=(IC(2,J)+15)/3
      RK=' '
      DO 55 L=1,I
55    WRITE(1,70) RK,
70    FORMAT(A1)
      RK='X'
54    WRITE(1,70) RK
24    L=0
      DO 9 I=1,IP,2
      L=L+1
      IC(1,L)=IC(2,I)
 9    IC(2,L)=IC(2,I+1)
      L=L+1
      IC(1,L)=999
      IC(2,1)=-999
      WRITE(1,20) L
20    FORMAT(I5)
      LL=2*L
      CALL WTAPE(1,MBLOK,LL,IC)
 2    ID(1,NUM)=MBLOK
      ID(2,NUM)=LL
      MBLOK=MBLOK+IREM(LL/128)+2
      NUM=NUM=1
      IF(NUM-46) 23,8,8
 8    CALL WTAPE(1,10,92,ID)
      STOP
23    CALL CHAIN('CCHAIN')
      END
```

```
      /  FDIS, A DISPLAY SUBROUTINE FOR THE PDP-12
      /  J H PULS, JUNE 1972
      /  FORTRAN: CALL FDIS (JARRAY,NUMBER), WHERE:
      /  JARRAY IS ARRAY OF OUTPUT POINTS, JX1,JY1,JX2,
                                   JY2,ETC.
      /  NUMBER IS THE NUMBER OF POINTS TO DISPLAY
         ENTRY FDIS
         DUMMY A
         DUMMY N
FDIS,    BLOCK 2
         TAD I FDIS
         INC FDIS#
         DCA A
         TAD I FDIS
         INC FDIS#
         DCA A#
         TAD I FDIS
         INC FDIS#
         DCA N
         TAD I FDIS
         INC FDIS#
         DCA N#
      /  ARGUMENTS NOW AVAILABLE
         OPDEF LINC 6141
         OPDEF CDF 6201
         OPDEF RDF 6214
         ABSYM PDP 2
         ABSYM DIS 150/ACTUALLY, DIS+10
         ABSYM IRN 10
         ABSYM ADD 2010/ACTUALLY, ADD+10
         ABSYM STC 4010/ACTUALLY, STC+10
         LINC SAVE IR 10
         ADD
         PDP
         DCA SAVIR
         TAD I N
         CIA/-WC
         DCA COUNT
LOOP,    TAD I A
         TAD P400
         LINC
         STC
         PDP
         INC A#
         TAD I A
         LINC
         DIS
         PDP
         INC A#
         CLA
```

```
        ISZ COUNT
        JMP LOOP
        TAD SAVIR
        LINC
        STC
        PDP
        RETRN FDIS
COUNT,  0
SAVIR,  0
P400,   400
A,      BLOCK 2
N,      BLOCK 2
        END
```

```
        / FUNCTION FADC TO READ PDP12 A/D
        / X= FACD (ICHANNEL), X WILL BE -1.0 TO +1.0
          ENTRY FADC
          DUMMY N
FADC,     BLOCK2
          TAD I FADC
          INC FADC#
          DCA N
          TAD I FADC
          INC FADC#
          DCA N#
        / ARGUMENT NOW AVAILABLE
          OPDEF LINC 6141
          ABSYM PDP 2
          TAD I N
          TAD (100
          DCA L2
          LINC
L2,       0  /NOW SAM + N
          PDP
          CALL 0,FLOT
          CALL 1,FMP
          ARG C1
          RETRN FADC
N,        BLOCK 2
C1,       1704 /21-9
          0
          0
        / FUNCTION IXL(N) TO READ EXTERNAL SENSE LINES
        / IXL=+1 IF SET, IXL=-1 IF NOT SET
          ENTRY IXL
IXL,      BLOCK 2
          TAD I IXL
          INC IXL#
          DCA N
          TAD I IXL
          INC IXL#
          DCA N#
        / ARGUMENT NOW AVAILABLE
          ABSYM COM 17
          TAD I N
          TAD C400
          DCA L3
          IAC /SET = -1
          LINC
L3,       0 /NOW SXL + N
          COM /NOW AC = +1 OR -1 (O.C.)
          PDP
          SPA
          IAC /NOW AC = +1 OR -1 (T.C.)
          RETRN IXL
```

```
        / SUBROUTINE TO CONTROL LEITZ SCANNING STAGE
        / CALL LFN(IOPTION) OR LFN(5,IDISTANCE)
          OPDEF LIO 6311
          SKPDF LSK 6311
          ENTRY LFN
          DUMMY X
LFN,      BLOCK 2
          TAD I LFN
          INC LFN#
          DCA N
          TAD I LFN
          INC LFN#
          DCA N#
        / ONE ARG AVAILABLE NOW
          TAD I N
          DCA L2 /L2 IS USED AS "I N"
          TAD L2
          TAD(-5
          SZA
          JMP SWITCH /WAS A 5
          TAD I LFN
          INC LFN#
          DCA X
          TAD I LFN
          INC LFN#
          DCA X#
        / BOTH ARGS NOW IN
SWITCH,   SKP /SKP OR 6311 FROM LAST INST
          JMP SWITCH
          CLA
          TAD KSKP
          DCA SWITCH /CLEAR OUT WAIT
          TAD L2 /TAD I N
          TAD (-1
          SNA
          JMP XIT /WAS A 1
          TAD KLIO /(INST-10)+6311=6310+INST
          DCA L1
          TAD I X
L1,       HLT /631N BY NOW
          CLA
          TAD L2 /TAD I N
          RTR /MOTION COMMAND?
          SNL CLA
          JMP XIT /NO
          TAD KLSK /SETUP WAIT
          DCA SWITCH
XIT,      RETRN LFN
KLIO,     LIO
```

```
KLSK,   LSK
KSKP,   SKP
X,      BLOCK 2
        END
```

```
C       INPUTS BANDING PATTERNS AND SCALES THE OUTPUTS
        DIMENSION IA(1000)
        CALL ASSIGN(1,'T100.CHM',8,IER)
        CALL ASSIGN(2,'T200.CHM',8,IER)
        DO 2 LP=1,46
        DO 1 I=1,1000
1       IA(I)=0
        J=-12
8       J=J+13
        K=J+12
        READ(1,10)(IA(I),I=J,K)
10      FORMAT(13I5)
        L=0
        DO 3 I=J,K
        L=L+1
3       IF(IA(I).EQ.999)GO TO 5
        GO TO 8
5       K=J+L-1
        M=32000
        DO 6 I=1,K
6       M=MINO(M,IA(I))
        IF(M.GT.0)M=-M
        N=0
        DO 7 I=1,K
        IA(I)=IA(I)-M
7       N=MAXO(N,IA(I))
        WRITE(2,10)K,N,M
        WRITE(2,10)(IA(I),I=1,K)
        DO 4 I=1,3
4       READ(1,10)IDUM
        WRITE(6,10)LP,K,N,M
2       CONTINUE
        END
```

```
      READ(6,30) K
30    FORMAT(A1)
      IF(K.EQ.'Y') GO TO 7
50    FORMAT(1X,'TOO MANY POINTS')
      END
```

```
      SUBROUTINE PAIR(P1,P2)
      COMMON/ONE/A(46,46),B(46)
      DO 100 LL=1,23
      HUGE=0.
      DO 1 I=1,46
      B(I)=GAIN(I,N,X)
      IF(HUGE.GT.B(I))GO TO 1
      HUGE=B(I)
      J=I
      L=N
1     CONTINUE
      IF(HUGE.LT.P1) GO TO 1000
      IF(X.LT.P2) CALL SPEC
      WRITE(6,20) J,L
20    FORMAT(3X,'ROW',I,2X,'COL',I,2X)
      CALL RDUCE(J,L)
100   CONTINUE
1000  RETURN
      END
```

```
      SUBROUTINE ASSIGN(P1,P2)
      COMMON/ONE/A(23,23),B(46)
      DO 100 LL=1,23
      WRITE(6,10) P1,P2
10    FORMAT(1X,23F3.0)
      ISW=0
      HUGE=0.
      DO 1 I=1,23
      B(I)=GAIN(I)
1     HUGE=AMAX1(HUGE,B(I))
      WRITE(6,10) HUGE
      CALL TRANS
      DO 2 I=1,23
      J=23+I
      B(J)=GAIN(I)
2     HUGE=AMAX1(HUGE,B(J))
      WRITE(6,10) HUGE
      WRITE(6,10) B
      IF(HUGE.LT.P1) GO TO 1000
      J=0
      DO 3 I=1,46
      J=J+1
      WRITE(6,30) J
30    FORMAT(10X,I)
3     IF(B(J).EQ.HUGE) GO TO 4
4     IF(J.GT.23) GO TO 5
      CALL TRANS
      J=J+23
      ISW=1
5     J=J-23
      X=0.
      DO 6 I=1,23
      IF(X.GT.A(J,I)) GO TO 6
      L=I
      X=A(J,I)
6     CONTINUE
      IF(X.LT.P2) CALL SPEC
      WRITE(6,20) J,L,ISW
20    FORMAT(3X,'ROW',I,2X,'COL',I,2X,'ISW',I)
      CALL REDUCE(J,L)
      IF(ISW.EQ.0) CALL TRANS
      WRITE(6,10) ((A(II,JJ),JJ=1,23),II=1,23)
100   CONTINUE
1000  CONTINUE
      RETURN
      END
```

```
      SUBROUTINE REDUCE(M,N)
      COMMON/ONE/A(23,23),B(46)
      DO 1 I=1,23
      A(M,I)=0.
1     A(I,N)=0.
      A(M,N)=-1.
      RETURN
      END


      SUBROUTINE IFIVE(IX,NUM)
      DIMENSION IX(700)
      SUM=0.
      DO 1 I=1,NUM
1     SUM=SUM+IX(I)
      R=SUM*.05
      L=0
      SUM=0.
      DO 2 I=1,NUM
      L=L+1
      SUM=SUM+IX(I)
2     IF(SUM.GE.R)GO TO 3
3     M=0
      DO 4 I=L,NUM
      M=M+1
4     IX(M)=IX(I)
      M=M+1
      DO 5 I=M,NUM
5     IX(I)=0.
      RETURN
      END


      FUNCTION GAIN(M,N,Y)
      COMMON/ONE/A(23,23),B(46)
      Y=0.
      DO 1 I=1,23
      IF(Y.GT.A(M,I))GO TO 1
      N=I
      Y=A(M,I)
1     CONTINUE
      Z=0
      DO 2 I=1,23
      IF(A(M,I).EQ.Y) GO TO 2
      Z=AMAX1(Z,A(M,I))
2     CONTINUE
      GAIN=Y*(Y-Z)
      RETURN
      END
```

```
C          CREATES 46X46 MATRIX FOR LATER IDENTIFICATION
C          ARRAY.FTN
           COMMON/ONE/A(46,46),Y(700)
           DIMENSION IX(700),Z(700),X(700)
           CALL ASSIGN(1,'T100.CHM',8,IER)
           CALL ASSIGN(2,'T200.CHM',8,IER)
           DO 100 I=1,46
           DO 100 J=1,46
100        A(I,J)=0.
           READ(2,10) MUM
           READ(2,10)(IX(LL),LL=1,MUM)
           MUM=MUM-1
           DO 3 I=1,45
           DO 200 IM=1,700
           X(IM)=0.
200        Z(IM)=0.
           READ(1,10)NUM,MAX
           READ(1,10) (IX(LJ),LJ=1,NUM)
           NUM=NUM-1
           DO 300 LK=1,NUM
300        X(LK)=IX(LK)
           DO 7 IR=1,700
7          Z(IR)=X(701-IR)
           CALL FIVE(X,700)
           CALL FIVE(Z,700)
           SUM=0.
           DO 11 IY=1,700
           X(IY)=X(IY)+SUM
11         SUM=X(IY)
80         FORMAT(10F7.0)
           SUM=0.
           DO 12 IY=1,700
           Z(IY)=Z(IY)+SUM
12         SUM=Z(IY)
           K=I+1
           DO 2 J=K,46
           DO 500 MR=1,700
500        Y(MR)=0.
           READ(2,10) MUM,MAXY
           READ(2,10) (IX(LL),LL=1,MUM)
           MUM=MUM-1
           DO 400 LH=1,MUM
400        Y(LH)=IX(LH)
           CALL FIVE(Y,700)
```

```
      SUM=0.
      DO 13 IY=1,700
      Y(IY)=Y(IY)+SUM
13    SUM=Y(IY)
      SMIR=0.
      S=0.
      DO 15 II=1,700
      D=ABS(Y(II)-X(II))
      DD=ABS(Y(II)-Z(II))
      S=AMAX1(S,DD)
15    SMIR=AMAX1(SMIR,D)
      A(I,J)=SMIR
      A(J,I)=S
      WRITE(6,2000) I,J,A(I,J),A(J,I)
2000  FORMAT(1X,2I,2(3X,F7.0))
2     CONTINUE
      REWIND 2
      DO 4 L=1,K
      READ(2,10) MUM
4     READ(2,10) (IX(LL),LL=1,MUM)
3     CONTINUE
      CALL ASSIGN(3,'CHRM.ARY',8,IER)
      WRITE(3,20) ((A(I,J),J=1,46),I=1,46)
20    FORMAT(10F7.0)
10    FORMAT(13I5)
      END
```

```
      FUNCTION SMIR(X)
      COMMON/ONE/A(46,46),Y(700)
      COMMON/TWO/ X(700)
      SMIR=0.
      DO 1 I=1,700
      D=ABS(Y(I)-X(I))
1     SMIR=AMAX1(SMIR,D)
      RETURN
      END
```

```
      SUBROUTINE CUM(X,NUM)
      COMMON/TWO/ X(700)
      SUM=0.
      DO 1 I=1,NUM
      X(I)=X(I)+SUM
1     SUM=X(I)
      RETURN
      END
```

```
      COMMON/ONE/A(46,46),B(46)
      CALL ASSIGN(1,'CHRM.ARY',8,IER)
      CALL ASSIGN(2,'LIST.CHM',8,IER)
      CALL ASSIGN(3,'PAIR.ARY',8,IER)
      READ(1,10)((A(I,J),J=1,46),I=1,46)
10    FORMAT(10F7.0)
      DO 2 I=1,46
      DO 2 J=1,46
      A(I,J)=AMIN1(A(I,J),A(J,I))
2     A(J,I)=A(I,J)
      X=1000000.
      Y=0.
      DO 1 I=2,46
      M=I+1
      DO 1 J=M,46
      X=AMIN1(X,A(I,J))
1     Y=AMAX1(Y,A(I,J))
      X=Y-X
      DO 3 I=1,46
      DO 3 J=1,46
      IF(I.EQ.J) GO TO 3
      A(I,J)=(Y-A(I,J))/X
3     CONTINUE
      CALL PAIR(0.,0.)
      DO 4 I=1,46
      DO 4 J=1,46
4     IF(A(I,J).EQ.-1.)WRITE(2,20)I
20    FORMAT(I)
      WRITE(3,10)((A(I,J),J=1,46),I=1,46)
      END
```

REFERENCES

Andrews, Harry C. 1972. Introduction to Mathematical Techniques in Pattern Recognition. John Wiley and Sons, Inc., New York.

Arrighi, Frances E., and Hsu, T. C. 1971. Localization of Heterochromatin in Human Chromosomes. Cytogenetics 10:81-86.

Bender, Michael A., Kastenbaum, Marvin A., Lever, Claudia S., and Pelster, Donald R. 1972. Combination Bayesian and Logical Approach to Analysis of Normal and Abnormal Chromosome Spreads. Computers in Biology and Medicine 2:151-166.

Butler, J. W., Butler, M. K., and Stroud, A. 1968. Automatic Classification of Chromosomes--III. In Data Acquisition and Processing in Biology and Medicine, Vol. 5, K. Eslein, ed., p. 21-38. Pergamon Press, New York.

Caspersson, Torbjorn, Lomakka, Gösta, and Møller, Aage. 1971. Computerized Chromosome Identification by and of the Quinacrine Mustard Flourescence. Hereditas 67:103.

Caspersson, Torbjorn, Lomakka, Gösta, and Zech, Lore. 1971. The 24 Flourescence Patterns of the Human Metaphase Chromosomes--Distinguishing Characters and Variability. Hereditas 67:89-102.

Caspersson, Torbjorn, and Zech, Lore. 1972. Chromosome Identification by Flourescence. Hospital Practice, September, p. 51.

Castleman, K. R. 1975. An Automated System for Chromosome Analysis. Vol. 1. Goals, Systems Design and Performance. Jet Propulsion Laboratories (1200-2400) (Rev. A), September 30, Final Report.

Chicago Conference. 1966. Standardization in Human Cytogenetics. National Foundation, New York.

Dixon, W. J. (ed.). 1973. Biomedical Computer Programs. University of California Press, Berkeley.

Duda, Richard O., and Hart, Peter E. 1973. Pattern Classification and Scene Analysis. John Wiley and Sons, Inc., New York.

Eddleman, E. E., and Bancroft, W. H., Jr. 1967. Quantification of the Kinetocardiogram. Computers and Biomedical Research 1:237-250.

Enslein, Kurt, and Neurath, Peter W. 1969. Augmented Stepwise Discriminant Analysis Applied to Two Classification Problems in the Biomedical Field. Computers and Biomedical Research 2:568-581.

Frey, Harvey S. 1969. An Interactive Computer Program for Chromosome Analysis. Computers and Biomedical Research 2:274-290.

Fukunaga, Keinosuke. 1972. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Herron, J., Ranshaw, R., Castle, J., and Wald, N. 1972. Automatic Microscopy for Mitotic Cell Location. Computers in Biology and Medicine 2:129-135.

Hilditch, C. Judith. 1969a. Linear Skeletons from Square Cupboards. In Machine Intelligence IV, B. Melzer and D. Hichie, eds., p. 403-426. Edinburgh University Press, Scotland.

Hilditch, C. Judith. 1969b. A System of Automatic Chromosome Analysis. In Automatic Interpretation and Classification of Images, A. Grasselli, ed., p. 363-387. Academic Press, New York.

Hilditch, C. Judith and Rutovitz, Denis. 1972. Normalization of Chromosome Measurements. Computers in Biology and Medicine 2:167-179.

Hsu, Lillian Y. F., and Hirschhorn, Kurt. 1972. The Potential Usefulness of Computerized Chromosome Analysis in Human Genetics. Computers in Biology and Medicine 2:99-106.

Huang, Thomas S., Schreiber, William G., and Tretiak, Oleh J. 1971. Image Processing. Proceedings of the IEEE 59:1586-1609.

Kazan, B., and Knoll, M. 1968. Electronic Image Storage. Academic Press, New York.

Kendall, M. G., and Stuart, A. 1968. The Advanced Theory of Statistics, Vol. 1 (3rd ed), Vol. 2 (3rd ed.), Vol. 3 (2nd ed.). Hafner, London.

Klingeman, Jack, and Pipberger, Herbert V. 1967. Computer Classification of Electrocardiograms. Computers and Biomedical Research 1:1-17.

Ledley, Robert S., Lubs, Herbert A., and Ruddle, Frank H. 1972. Introduction to Chromosome Analysis. Computers in Biology and Medicine 2:107-128.

Lindgren, B. W. 1968. Statistical Theory, 2nd ed. Macmillan, New York.

Liu, C. L. 1968. Introduction to Combinatorial Mathematics. McGraw-Hill, New York.

Low, David A., Selles, William D., and Neurath, Peter W. 1971. Bayes' Theorum Applied to Chromosome Classification. Computers and Biomedical Research 4:561-570.

McMorris, F. A., Chen, T. R., Ricciuti, F., Tischfield, J., Creagan, R., and Ruddle, F. 1973. DNA Content and DNA-based Centromeric Index of 24 Human Chromosomes. Science 179:1129-1131.

Mendelsohn, Mortimer L., and Mayall, Brian H. 1972. Computer-Oriented Analysis of Human Chromosomes--III. Focus. Computers in Biology and Medicine 2:137-150.

Neurath, Peter W., Kess, B., and Low, D. A. 1972. Individualized Human Karyotyping through Quantitative Analysis. Computers in Biology and Medicine 2:181-193.

Prewitt, Judith M. S. 1970. Object Enhancement and Extraction. In Picture Processing and Psychopictorics, B. Lipkin and Azriel Rosenfeld, eds. Academic Press, New York.

Priest, Jean H. 1969. Cytogenetics. Lea and Febiger, Philadelphia.

Rosenfeld, Azriel. 1969. Picture Processing by Computer. Academic Press, New York.

Siegel, Sidney. 1956. Nonparametric Statistics: for the Behavioral Sciences. McGraw-Hill, New York.

Tjio, Joe Hin, and Levan, Albert. 1956. The Chromosome Number of Man. Hereditas 42:1.

Troxel, Donald E. 1971. Automated Reading of the Printed Page. Visible Language 2:125-143.

Watson, J. D. 1970. Molecular Biology of the Gene, 2nd ed. W. A. Benjamin, New York.

Willard, Hobart H. 1965. Instrumental Method of Analysis. Van Nostrand, Princeton, N. J.

Yunis, Jorge J., and Yasmineh, Walid G. 1971. Heterochromatin, Satellite DNA, and Cell Function. Science 174:1200.