605

707800

COMPUTER ASSISTED CHROMOSOME ANALYSIS

Michael A. Bender

Vanderbilt University School of Medicine

Nashville, Tennessee

REPOSITORY Oak Ridge Operations Ofc.
COLLECTION Records Holding Area Documents 1944-1994
BOX No. A-42.3 Bldg 2714-H
FOLDER 4065 Vanderbilt c 6-8-70

1034371

The cytogenetic analysis of human somatic chromosomes has become increasingly important in both clinical diagnosis and research. Many clinical conditions, particularly congenital malformations, have been shown to arise from numerical of structural alterations of the normal chromosomal constitution. Thus Downs' syndorme, or mongolism, is associated with the presence of an extra one of a particular chromosome, and chronic myelogeneous leukemia is associated with a structural abnormality of one of the normal chromosomes in the leukemic cells. In many cases, the detection of a chromosomal abnormality in an otherwise normal person makes it possible to predict a high risk of producing abnormal offspring, facilitating genetic counceling. Techniques that are under development will allow the normality and even the sex of an unborn infant to be determined by means of chromosomal analysis. In addition, the analysis of somatic chromosomes for various types of structural aberrations offers a sensitive test for damage to the genetic material, and allows assessment of the genetic hazards of radiation, drugs, food additives, and various environmental factors. Aberration analyses actually offer a means of "biological dosimetry" in certain cases of exposure to radiation and possibly to other mutagens as well.

These applications have created a large demand for human chromosome analyses and chromosome aberration scoring. Unfortunately, conventional analysis and scoring are very time-consuming processes, and require personnel with a high level of training and experience. The scarcity of experienced cytogeneticists, together with the time and expense required for each analysis, puts a real limitation on the

1034372

number that can currently be carried out even for clinical diagnosis, and has severely curtailed large-scale population surveys. Any means of speeding up the process of analysis, or of utilizing less highly skilled personnel would thus be very valuable. Furthermore, the process of analysis as it is presently carried out is quite subjective, and anything that would make it more objective is also desirable. Progress toward alleviating these problems is being made through the use of digital computers for various steps in the process of chromosome analysis.

A number of groups are working toward automation of chromosome analysis, and the development of both the specialized hardware and the specialized pattern recognition computer programs required has been quite rapid. It is not the purpose of this review to detail all the technical features of any particular system; these are changing rapidly in any case. The purpose is rather to present a description of the biological problem itself, and an outline of the various approaches that have been taken toward solving it. Needless to say, the ultimate goal of complete automation has not yet been achieved. It furthermore seems unlikely that a completely automatic system that is satisfactory both economically and in performance will be forthcoming in the near future. As will be seen, however, there appears to be a good possibility that a practical compromise, a semi-automatic device that uses a human to solve the more difficult pattern recognition problems and a small computer for the actual analysis, can be achieved in the interem.

1034373

THE BIOLOGICAL PROBLEM

Most of the time, the chromosomes within each cell are diffuse, and cannot be distinguished under the microscope. Only when the cell divides, by a process known in somatic cells as mitosis, do the

chromosomes become condensed and easily visualized. A fixed number of chromosomes is characteristic of the cells of each species; in man this number, called the diploid number, is 46. This complement is made up, at least in the female, of 23 genetically identical (or nearly identical) pairs of chromosomes. Because sex is determined by a "pair" of genetically different chromosomes, which in man are also morphologically different, the normal male chromosome complement consists of 22 pairs of chromosomes known as the autosomes, plus the two non-identical sex chromosomes, called the X and Y. Thus there are, in the genetic sense, 24 different kinds of chromosomes in somatic cells from men, while in those from women, who have two X chromosomes rather than an X and a Y, there are only 23 kinds.

Figure 1 shows the chromosome completment of a white blood cell from a normal man. The cell underwent mitosis in a short-term tissue culture (a common source of material for cytogenetic analysis) and was fixed and stained by standard techniques[1]. Each chromosome is seen to consist of two more or less parallel strands, called chromatids, attached to each other at a relatively unstained area which is known as the centromere. The centromere divides each chromatid into two arms, which may be of anywhere from equal to very different lengths, depending on the position of the centromere.

Using, basically, the two shape parameters, the whole complement of 46 chromosomes may be sorted into different types. The two parameters are the total length of the chromosome (relative to the others in the cell) and the centromere position (or some other permutation of the arm lengths). The sorting leads to the production of what is known as a karyotype. Normally the karyotype is produced

1034374

by cutting the chromosomes from an enlarged photograph and arranging them in pairs according to a numbering convention generally agreed on by cytogeneticists. Figure 2 shows such a "cut and Paste" karyotype prepared from the photograph of Figure 1. The autosomes are arranged generally in order of decreasing total length; pair assignments are made on the basis of centromere position. Thus the pairs numbered 1 and 2 are of approximately the same length, but the centromeres of the number 1 pair are nearer the middle of the chromosomes than are those of the number 2 pair.

Ideally, all of the chromosome pairs, known to be different from each other in the genetic sense, would be different enough morphologically to be easily identifiable, one from the other. For several reasons, unfortunately, this is not true of the human complement. First, as can be seen from Figure 2, there is variability of lengths within pairs. Since the chromosomes as seen at cell division are coiled, somewhat elastic structures, and may actually be distorted by the fixing, spreading, and mounting techniques necessary for microscopic examination, this variability is not surprising. Second, although some of the pairs are quite obviously different from all the others, some are not. It is for this reason that groups of pairs, given alphabetical designations, are indicated in Figure 2. Thus though the first three pairs are separable with some degree of certainty, the next two pairs, designated the B group, are very similar to each other. Though an attempt to pair them has been made, the identifications are not certain, and others could probably just as well have been assigned. The same is true of the other groups.

The process of analysis of the chromosome complement of an individual subject is simply that of establishing that his cells contain the right number and morphological types of chromosomes. Because the chromosome identifications are uncertain in any individual cell, and also because individual cells are found that are missing one or more chromosomes (probably because of loss of a portion of the cell during preparation), the diagnosis must be made from a number of individual karyotypes prepared from a sample of cells. In this way it may be determined whether there are any consistent differences in numbers, lengths, or centromere positions from the normal mean values. In addition, it is possible for a person to have two or more populations of cells with different chromosome complements. In cases where this condition (known as mosaicism) is suspected, a larger number of individual karyotypes must be prepared.

In the case of scoring for the effects of mutagenic agents such as radiation, an even greater number of cells, usually 100 or more, must be analysed. This is because each cell is individually affected by the mutagen. Determinations of the degree to which the cells have been affected must be based on the average effect on the population of cells, and a statistically satisfactory sample must be obtained. Mutagens cause breaks in the continuity of chromosomes, and when more than one break occurs in the same cell, the broken chromosome ends may recombine with each other to form new and sometimes bizarre chromosome shapes and sizes. The types of chromosome aberrations usually scored include single-break deletions and two-break rings and

1034376

F-3

dicentrics. Figure 3 illustrates these aberrations. The presence of chromosome aberrations is usually determined by visual inspection and counting of the chromosomes. As will be seen, however, they can also be determined by karyotype analysis. Such an analysis offers the advantage that the individual chromosomes, or at least the chromosome groups involved can be identified, and furthermore that more subtle rearrangements (inversions and translocations) may sometimes be detected.

The use of digital computers for chromosome analysis and scoring depends on the idea that the chromosome complement may be described metrically by a series of numbers, such as a list of all the arm lengths in a given cell, and that a computer may be programmed to do the necessary operations, normally carried out by a cytogeneticist, through acquisition of such a list of numbers and by performing arithmatical and logical operations on it.

CHROMOSOME PATTERN RECOGNITION

The first steps in the analysis of chromosome complements are the location of suitable metaphase cells, the location of each chromosome in the spread, and the determination of the centromere positions and chromosome axes in order that arm measurements (length, area, density, etc.) can be made. These are essentially problems in pattern recognition. Considerable effort has been devoted to the construction of devices (or the application of existing devices) to acquire optical chromosome images and to carry out the required pattern recognition by means of computers. The problem is complicated by the very small size of chromosomes. The largest pair of human chromosomes averages some 7-8 microns in length,

1034377

and a whole chromosome spread rarely exceeds 50 or 60 microns in diameter. Direct observation must thus be done with the highest powers of the compound microscope, and image acquisition must either be through such an optical system or from photomicrographs. In either case serious optical and mechanical problems are encountered.

Identification of suitable chromosome spreads from among the collection of non-dividing cell nuclei, cell fragments, and other debris on the microscope slide is necessary before any analysis can be attempted. Though this is quite a difficult pattern recognition task, automation of the chromosome spread locating process appears feasible. This is, however, a job that can be relatively economically carried out by a human operator with only a minimum of experience, which perhaps explains why only one group has attempted to automate it. Their device[2] employs a mechanical stage to scan a large slide area with standard microscope optics illuminated with coherent light from a laser. Spatial filters are used to "recognize" the diffraction patterns characteristic of metaphase spreads. Though testing is not complete, the system is reported to find 80-90% of the metaphases on a slide. What percentage of all the objects it finds will turn out to be analysable metaphases remains to be seen.

Once a suitable metaphase chromosome spread has been found, the image may be disected and digitized by one of several types of devices. There are a number of constraints on such devices, however. First, they must possess sufficient resolution to permit accurate recognition and measurement. Since the chromosomes are very small, resolution approaching the theoretical optical limit is required for direct

1034378

scanning through the microscope. Second, the system must distinguish an adequate number of grey levels, or else have some means of controlling the black-white discriminator threshold, in order to be able to connect all parts of a single chromosome without joining areas that belong to different chromosomes. Third, of course, distortion over the scanned field must be kept low enough so that measurements are adequately reproduceable. Finally, the device must scan rapidly enough to digitize the image within a reasonable time.

Several approaches to the problem have used mechanical movement of the microscope slide or a photograph of it to accomplish scanning. A simple spot-reading photometer is then used to provide analog transmission or absorption data which can be fed through an analog-to-digital converter and thence to the computer. Drum type devices similar to wirephoto units have been used, as well as microscopes with rectilinear automated scanning stages such as the Carl Zeiss CYTOSCAN unit. Unfortunately, such mechanical systems are relatively slow in operation, and most instruments have used electron-optical scanning devices instead. These include image orthicon, vidicon, and image dissector tubes to scan the chromosome spread image, as well as flying spot scanners in which a cathode ray tube spot is swept in a raster over the chromosome spread and the intensity of the transmitted beam measured by a photomultiplier. A variation of the flying spot, image plane scanner uses a point source of light that is swept in a raster by a pair of oscillating mirrors. All of these systems encounter difficulties in the form of distortion, sensitivity, and signal-to-noise ratio problems, as well as some specific to each particular device, but all have nevertheless been employed with some degree of success to produce what amount to

1034379

digitized video signals either from the chromosome spread or from photographic transparencies.

Formidable as some of the technical problems with scanning devices are, those associated with storage and manipulation of the digitized image are even greater. To obtain the spatial resolution necessary for accurate measurement may require a raster of points perhaps 1024 x 1024, and even a 3 bit grey level discrimination would thus require over 3 million bits of information to be handled. To store this much information in core memory and have enough room for programs and data manipulation left over requires a large computer. Some means of reducing this mass of information to manageable proportions is clearly necessary.

Several approaches to this problem have been taken. Obviously, it should really not be necessary to store all of the points in the raster, since the vast majority of them are not a part of any chromosome. It should thus be possible to store only the location of each chromosome point and neglect all background points. Unfortunately, though, chromosomes vary in density, not only from cell to cell, but also within cells, and even between different regions of the same chromosome. Depending on the grey level selected for discrimination between chromosomes and background, adjacent chromosomes may be artifically joined together, parts of the same chromosome may become separated, or even whole chromosomes may disappear because they happen to be a little less dense than the others. For this reason it is not practical to discard much of the image information content before the pattern recognition process has been initiated.

1034380

Two general methods have been used to avoid this difficulty. One solution is to scan (or at least to hold in core memory) only a small portion of the image at a time, and to carry out the chromosome recognition and measurement steps on that image segment before moving on to the next. This method may employ mass storage devices such as magnetic tapes or discs to advantage, especially if care is taken to store the data in blocks suitable for later rapid readout of suitable image segments. The other method is to actually put the scanning operation under computer control, so that the computer may search for and measure only image areas of possible interest. Frequently this is implemented by actually tracing out the edge of any object encountered. Either approach reduces the data volume that must be kept in core memory sufficiently so that computers of moderate size are practical.

No matter what systems of image aquisition and handling are used, the computer must be programmed to recognize chromosomes, to determine features of interest, such as centromeres and arm ends for example, and to obtain the required measurements. In the case of storage of a video raster of density information, which is essentially a one-dimensional array, this requires that the image be somehow reconstructed in two dimensions in such a way that the computer is able to connect high density areas encountered on successive scans and recognize that they belong to the same object. This in itself is a less than trivial problem because the shapes of chromosomes are complicated and a given scan line may intersect the same chromosome a number of times. This problem is circumvented by the edge-tracing approach used to resolve the image data storage problem. Each object that might be

1034381

a chromosome is considered as a unit, and since chromosomes have only a limited number of possible shapes, a syntactical method, based on the sequence of turns of the tracing point in going around the object, may be used to recognize chromosomes and reject non-chromosomal objects.

Recognition of chromosome ends and centromeres may be accomplished in several ways. When an edge-tracing routine is employed for chromosome recognition these features may be recognized (ideally, at least) syntactically from the sequence and types of turns in the boundry. Thus a chromatid arm end is essentially a 180° outside turn, and a centromere is defined by the two "notches" where the whole chromosome is narrowest. Another approach is that of constructing the minimum rectangle that will contain the chromosome, and defining the centromere as the minor axis having the least dense material across it (i.e., the narrowest point). A third method that has been used is to "skeletonize" the chromosome by making the image of each arm progressively narrower until but a single line remains. Centromeres and ends may then be determined by graph theory from the matrix of all points on the lines representing one chromosome, and the lengths of the line segments (arms) determined.

Once the chromosomes are recognized, and depending somewhat on the method used for this recognition, a number of their physical characteristics may be measured and listed for use as input for scoring and analysis programs. As already noted, the two arm lengths (or some permutation such as total length and arm ratio) are commonly thought to be the important quantitative chracteristics used by human analysists in aberration scoring and karyotype preparation. There is no reason,

1034382

though, for a computer system to be limited to only these parameters, and some systems have made use of such measurements as perimeter, area, integrated density, or have derived rather abstract shape parameters in addition to or instead of lengths.

CHROMOSOME SCANNING HARDWARE

A number of hardware systems have actually been built or adapted for chromosome image scanning and pattern recognition. A brief summary will be given of several of the more important of them. More complete details are available through the references cited. The list is by no means complete, but is intended only to serve to indicate the general types of hardware in use.

CHLOE[3]. This machine scans film frames with a flying spot scanner and produces digital output on magnetic tape for later analysis on a large computer. It was originally built for scanning nuclear track photographs and only secondarily used for chromosome and other biomedical images. The device compacts the scan information by storing only the beginning and end of each scan line segment exceeding an externally-set threshold density. Programs have been developed specifically to recognize and pair chromosomes from "central moments" and integrated densities calculated from the CHLOE output data.

FIDAC[4]. This is a computer-controlled flying spot film scanner designed specifically for processing biomedical images on line with a fairly large digital computer. The software provided by its developer accomplishes chromosome recognition by the boundry tracing method. Chromosome arm ends and centromere are recognized by a syntactical boundry segment approach.

1034383

FIDAC devices have been installed, with some modification, in several laboratories that are developing their own chromosome recognition, measurement, and analysis programs. One group[5] has added automatic grey level setting to the device, and has combined the line segment storage system used for CHLOE with skeletonization and graph theory approaches to recognition and measurement. Another has circumvented the pattern recognition and grey level problems by tracing chromosome spreads in black on white, without overlaps, and then using FICAC to analyse these stylized images[6].

CYDAC[7]. This device is a direct "through the microscope" flying spot scanning microdensitometer. It produces image records on magnetic tape for later input to a large digital computer for chromosome recognition, measurement, and analysis. CYCAC is an accurate cytophotometric device, providing 256 digital grey levels, and has been used mainly for determination of stain content of whole chromosome arms (i.e. nucleic acid content or "mass") rather than for length measurements. As with other devices, however, the pattern recognition programs determine centromere positions and major chromosome axes.

PIQUANT[8]. This instrument is a film-reading flying spot scanner with 16 grey level discrimination. It operates on line with a large digital computer and provides a monitor-light pen system for active operator-intervention during chromosome recognition and measurement. Mass storage of the entire picture matrix is provided, and core processing is done sequentially on small picture segments. A boundry tracing method is used to outline each chromosome and a minimum circumscribed rectangle constructed to define the major axis. The ends and centromeres are found from the number of intersections minor axis scan lines make with the chromosome boundry.

1034384

In addition to these systems, several more general purpose, commercially available scanners are in use or are planned for use in chromosome analysis automation. These include, among others, the IBM 2282 35 mm film scanner that was developed for graphic input as part of the IBM System 360 family of computers, the Perkin-Elmer Corporation's CELLSCAN device, and the Zeiss CYTOSCAN machine.

KARYOTYPING

The image processing devices being applied to the chromosome scoring problem are all designed to produce a digital list of two or more parameters for each of the chromosomes found in the cell being analysed. In its simplest form the list for the normal human cell might consist of a set of 46 pairs of numbers, each representing, perhaps, the absolute lengths of the long and the short arms of each chromosome. The problem presented to the computer, then, is to determine whether the input list is that of a normal cell, and if not, to determine in what way the cell deviates from normal. This is, in fact, exactly the same job now done by human operators by cutting out chromosomes from a photograph and arranging them in pairs to prepare a karyotype. Hopefully, the only attributes of the chromosomes necessary for the human operator to successfully carry out the process are the length measurements. Certainly the human karyotyper does actually use other features (such as the constrictions that sometimes appear in certain chromosomes) when these are available, but they do not appear to be required, only helpful. It seems reasonable, then, that a computer might be programmed to carry out the process using only the arm length information. It also seems possible, however, that the machine could make use of other parameters as well, perhaps measurements

1034385

not readily apparent or quantifiable by the human operator, such as arm area or density or local density distribution within arms. One might even hope that these measures might allow greater discrimination of the different chromosomes than in possible with only arm lengths. But since the usefullness of these other measures has not been established, only karyotyping from arm length lists will be considered in detail. The logic involved is, of course, readily adaptable to other pairs of measurements, such as the arm stain densities that are measured by the CYDAC instrument.

Forced pairing. The simplest approach to karyotyping is simply to ask the computer to make the best possible pairs, based on some measure of similarity between the different possible pairwise combinations of measurements in the input list, and to order them with respect to total length. This scheme has the inherent limitation, though, that it assumes that there are, indeed, only pairs of chromosomes. The method is consequently unable to deal with cases of missing or extra chromosomes, or cases (such as the human male cell that has an XY "pair") where a morphological pair does not exist. Thus if a forced pairing routine were used to attempt to karyotype a cell with a missing chromosome, for example, it would inevitably pair the single representative of the pair to which the missing chromosome belonged with one from another group, and then mismatch all of the remaining ones. Depending on the order in which the chromosomes were paired, then, the computer would always conclude that the missing chromosome was from the last pair, a result that would only occasionally be correct by chance.

1034386

Pairing routines have been used with some degree of success, however, by adding the calculation of some measure of the similarity of the prospective pairmates[9]. Some empirical value of this measure is used as a threshold, and if no pairing can be made that gives a value greater than the threshold, the chromosome for which a pairmate is being sought is left unpaired. This method will in theory deal with both cases of missing chromosomes and of morphologically dissimilar "pairs".

Scaling. Because pairing routines operate only on the measurements made within one cell, no allowance need be made for the rather large variability in absolute measurements that exists between different cells, even from the same individual. All of the other karyotyping schemes that have been developed, however, depend in some way on comparisons of the unknown chromosome set with the normal human chromosome set. Because of the variability of length among cells, some normalization of the absolute measurements is required. Arm lengths may be recalculated and expressed, for example, as percentages of the total chromosome length for the cell. Selection of the proper reference against which to normalize presents a problem, however. For example, there is a difference in the total lengths of the chromosomes between male and female cells because the Y chromosome is considerably shorter than the X. It appears that this particular difficulty will probably be unimportant in practice, because the difference is only a few percent of the total chromosome length, and the variability between pairmates is also of this order. Selection of the proper scaling factor will certainly present a real problem though in cases where the chromosome complement is grossly abnormal, as for example in the case of a cell with a number of extra chromosomes. Probably this will have to be dealt with by some scheme that uses the total chromosome number

1034387

to modify the scaling factor, perhaps by subtracting (or adding) the average absolute human chromosome length from the total for each case of an extra (or missing) chromosome.

Plane slicing. The two measurements for each chromosome define its position in a two-dimensional space. What is known as a karyogram may thus be constructed by plotting total length against arm ratio for each chromosome in the cell. In theory, such a diagram may be sliced up into segments that include the average values for each chromosome pair, but not those of other pairs. Unfortunately, the average measurements for human chromosomes are variable enough so that there is considerable overlap, and only the major chromosome groups may be thus defined with reasonable accuracy[10,11]. The partitioned plane may be used as a "template" and form the basis of a karyotyping routine that simply counts the number of measurement pairs (chromosomes) that occupy each region and assigns those it finds there to the appropriate group[12]. Unfortunately, the chromosome measurement variability that is found in practice appears to lead to considerable misclassification. Also, the scheme is quite sensitive to scaling errors. Such errors tend to push chromosomes from one group to another, resulting in their misclassification. The template approach may be somewhat improved by defining actual tolerance areas round the normal mean values for each chromosome instead of simply slicing up the two dimensional space[11]. This allows rejection of measurement pairs that clearly do not belong to any normal chromosome, but the extremely large size of the tolerance areas at best allows assignment of chromosomes only to groups, not pairs.

Discriminant functions. Attempts have been made to accomplish karyotype analysis by a somewhat different technique, using more or

1034388

less arbitrary discriminant functions derived from chromosome measurements. The usual pair of chromosome measurements do in fact consititue a two-dimensional linear discriminant function. The multi-dimensional linear discriminant function method is essentially simple extension of the plane slicing technique to a many-dimensional space. Since the discriminant functions are linear, the space is cut up by hyperplanes, making the problem of calculation relatively simple. However, since only two measurements are actually available, the other functions must be derived from them arithmatically, and are thus not independent. The usefullness of some of the functions that have been used is quite understandable, but that of others is rather obscure. For example, including the square of the total chromosome length simply gives greater weight to length for long chromosomes than for short ones. On the other hand, the meaning of the cube root of the difference between the long and short arm lengths is more difficult to understand. In practice, the functions with the best capacity to discriminate are actually selected empirically. In tests of the discriminant function method[13] a 10-dimensional discriminant was derived empirically from tests with a series of actual arm length measurements and then used to assign a test set of measurements to eight chromosome groups (the usual groups designated B-G in Figure 2, plus two groups consisting of the number one pair by itself and of the second and third pairs together). The discriminant function successfully assigned 91% of the test chromosomes to their correct group. This was increased to 93% by adding a "reassignment" routine that actually forced the selection of the proper number of chromosomes for each group, i.e. two first chromosomes, four group B chromosomes, and so

1034389

forth. Such a routine is obviously not practical for actual scoring, though, because it can only be used in cases where the karyotype can already be assumed to be normal before the analysis is begun.

Although the approach has not yet been tried, it seems likely that a non-linear discriminant function could be derived that might be more powerful than any linear discriminant function. The separations between groups or pairs could thus be curved, or even closed hypersurfaces, rather than hyperplanes. In anology with the two-dimensional tolerance area approach mentioned above, it should be possible to reject chromosomes as belonging to any group or pair on the basis of their falling outside of any of the hyperspaces for normal chromosomes. By some calculation of "distance" from the center of a given tolerance hyperspace it should be possible as well to make selections among possible candidates on the basis of how well a given candidate "fitted" the space for a particular pair or group. Ultimately, however, the success of any discriminant function method for karyotyping must depend on finding functions that allow absolute discrimination between pairs or groups. And it seems optomistic to hope that such a function can be derived from only the two independent parameters, such as arm lengths, that have been tried so far.

The Bayesian statistic. While the human operator may use a multidimensional discrimination when preparing a karyotype, it seems certain that he must unconciously make use of probability judgements in deciding which chromosomes are which, and whether a candidate for a particular pair "fits" well enough to be considered normal. Such judgements must be made on the basis of posterior, rather than prior probability. Knowing the statistical distribution of arm measurements

1034390

for each chromosome pair in the normal population, for example, the probability that a given pair of arm measurements came from the population of each chromosome type may be estimated. This is the Bayesian method, and it is being incorporated into automated karyotvpe analysis routines[11]. The method requires that a statistical description, in the form of bivariate vectors of means and their variances, be available to the computer. After the raw measurements of the chromosome spread to be analysed are normalized, the probability (actually predictive density) that each of the measurement pairs is that

of a member of each of the normal human chromosome pairs is calculated.

The routine then sorts the measurement pairs into those that have the highest probability of coming from number one chromosomes, from number two chromosomes, and so forth. If two, and only two, chromosomes are found that have their highest probability of being number one chromosomes, and if the probabilities are acceptably high (this level is set empirically, and offers a means of checking "fit" similar to that used by human operators) these candidates are accepted, and an attempt to find the next chromosome pair is begun. If more than two acceptable candidates are found for a given pair, which happens particularly often with the pairs in the C group, then the two with the highest probabilities may be selected. Such a routine frequently succeeds in making the correct pair (not just group!) assignments, but suffers from the difficulty that an incorrect choice sometimes leads to a progressive error that is propagated through the rest of the analysis. This difficulty is shared by other schemes as well, of course. Thus selection of what is really a number 10 chromosome as one of the number 8 pair, for example, can lead to failure to assign the "real" number 8 chromosome at all.

1034391

Because the chromsomes within groups are of similar sizes and shapes such an error is quite possible in spite of the ability of the program to reject obviously poor candidates on the basis of their having insufficiently large probabilities. By the time the routine gets to the end of the group, however, a chromosome left out near the beginning of the group is likely to have too low a probability of being a member of the last pair in the group (or any other unsatisfied" pair) to be used at all, and the computer concludes that the karyotype is abnormal, when in fact it is not.

More than other chromosome analysis schemes, the Bayesian approach offers the possibility of solving this problem through the application of logical routines similar to those used by a human operator when preparing a karyotype. When the cytogeneticist finds himself in a situation like that outlined above, he attempts to re-pair the chromosomes in the group in such a way that all candidates are used, and that all pair requirements are satisfied with candidates that are acceptable, even if this means assigning a chromosome to a pair other than that to which it seemed at first to most probably belong. In fact, information gained during pairing attempts is used as feedback to alter the decision process. The key requirement is the ability to accept an assignment as "good enough", even though it may not seem the best possible assignment. A computer program may easily be given the same capability, provided some test of acceptability is available. The Bayesian probability forms the basis for such a test. Though the ultimate logical reassignment programs have not yet been written, limited testing of partial reassignment subroutines, used in conjunction

1034392

with the basic Bayesian karyotyping approach, have given encouraging results, in many cells assigning every candidate to its correct pair. It seems likely that only a modest amount of additional development will be required to make this method perform karyotyping at least as acceptably as do human operators.

ABERRATION SCORING

As already mentioned, scoring for chromosomal aberrations is generally done directly through the microscope. It would thus appear at first glance to be a pattern recognition problem. Pattern recognition is actually not necessary, however, except in the case of one rather infrequent aberration type, the ring chromosome. The other aberration types, and all combinations of them, may be recognized very simply from the structure of the chromosome arm length measurement list used as input for many automatic karyotyping routines. This greatly simplifies the automation of aberration scoring.[14]

The chromosome arm measurement list for a normal human cell must consist of 46 pairs of numbers (arm lengths). Numerical abnormalities will be reflected in the addition or deletion of pairs of arm measurements. Aberrations, on the other hand, result in single unpaired measurements or in measurement sets with three or more members. A chromosome deletion (Figure 3) for example, produces a shortened chromosome which nevertheless has two arms (except in the extremely rare case of break through the centromere), plus a fragment that does not have a centromere. The shortened chromosome will still generate the usual pair of arm measurements, but the acentric fragment can only be measured from end to end, and will thus generate but a single measurement. A cell with one chromosome deletion will, then, be characterized by a list containing

1034393

46 pairs of measurements plus a single measurement, and such a list, furthermore, can come only from a cell with a single deletion. Similarly, a dicentric chromosome with its accompanying fragment (Figure 3) will result in a set of three lengths (end to centromere to centromere, centromere to end) plus a single measurement for the fragment. The list will then contain 44 pairs, one set of three, and one single measurement. Again, such a list is unique for this aberration type.

Computer programs that use these unique features of the input arm measurement list for scoring aberrations have been written. Because a ring chromosome presents a real pattern recognition problem, however, such type of scoring program rings were recognized by a human operator during the preparation of the arm measurement list, and a numerical flag added to the single diameter measurement made for the ring. The program was used to score a group of 150 x-ray-treated human chromosome spreads for aberrations. It succeeded in giving exactly the same result, cell by cell, as was obtained independently by a cytogeneticist scoring the same sample by the usual microscopic method.

Although it has not yet been attempted, further sophistication of aberration scoring is clearly possible once karyotyping routines have been perfected. Such a program would carry out the karyotype analysis on only the non-aberrant chromosomes identified through the scoring routine, and could then go back and identify the chromosome from which a fragment was deleted, for example, or determine the chromosomes most likely involved in the formation of a dicentric. Furthermore, use of a karyotyping routine would also allow automatic, objective identification of aberration types not usually scored for

1034394

at present, such as translocations or inversions that change chromosome lengths, or produce chromosomes of unique arm ratios. Though such aberrations may be determined by the present manual karyotyping procedures, the time and labor involved in the analysis of a meaningful sample of cells usually precludes such scoring.

SEMI-AUTOMATIC SYSTEMS

As desirable as complete automation of the chromosome scoring and karyotyping operations is, it seems practical to consider as well partial automation with systems that, for example, use a human operator for the more difficult pattern recognition operations. Such compromises must be considered for at least two reasons. First in spite of the progress that has been made in the development of completely automatic chromosome recognizing and measuring devices such as FIDAC and CYDAC, no such device has been demonstrated to do these jobs as well as a human operator even on near-perfect chromosome spreads, and it seems unlikely that solutions will be found in the very near future to the problems of the touching or overlapping chromosomes encountered in the less perfect spreads with which the cytogeneticist commonly must deal. Second, it seems that even when they have been brought to the point of satisfactory routine operation, such completely automated chromosome scanning devices will require fairly large computers and substantial running time just to prepare the measurement list that must be the starting point for any analysis or scoring routines. Such systems will thus be very expensive both to purchase and to operate, and it seems unlikely that they will be economically competitive with human operators. The PIQUANT system, with its provision for human intervention during the chromosome recognition and measurement steps does represent a compromise solution to the first problem, but not to the second, since

1034395

it operates on line with a fairly large computer.

If one assumes that programs for scoring and karyotyping that require only an arm measurement list for input will be brought to a satisfactory level of development for practical use in the very near future, then very practical semi-automatic devices are possible. Human operators can be trained in a matter of a few hours to perform the chromosome recognition and measurement tasks required. Using a simple measuring device, such a pair of calipers connected to a digital voltmeter, an otherwise inexperienced operator can complete the measurement list for a cell in a few minutes. This compares very favorably with the half hour or more required by a highly skilled cytogeneticist to construct a karyotype. Furthermore, the analysis of each cell must wait until the input list of measurements has been completed. This is an enormous length of time in terms of computer calculation speed, and there thus seems no reason why very small computers could not be used for on-line analysis. A small core memory can be utilized even for complicated analysis routines, provided the routines are broken up into small sub-programs that are shuffled in and out of an auxillary disc or tape memory. Though this results in very long running times by the usual computer standards, they are still short with respect to the rate at which input measurements are acquired, and are thus quite acceptable for an inexpensive machine.

A small computer provided with an auxillary memory device and with some measurement input source such as the digital dividers mentioned above, would, together with suitable software, constitute a practical semi-automatic chromosome analysis machine. Such a device could perform all of the tasks of a fully automatic system except for chromosome recognition and measurement. Since the initial cost of the

1034396

hardware should be comparable with that of many instruments now in routine use in clinical and research laboratories, it seems likely that such "first generation" machines will be built in the relatively near future and will be used extensively in cytogenetics laboratories.

1034397

FIGURE LEGENDS

Figure 1. Metaphase chromosome spread in a peripheral leukocyte from a normal man. CTD-chromatids; C-centromere; A-chromosome arms.

Figure 2. The chromosomes from the cell shown in Figure 1 paired and arranged according to cytogenetic convention to form a karyotype.

Figure 3. Metaphase chromosome spread from an irradiated cell. A-acentric fragment; D-dicentric chromosome.

1034398

1034399

1/1

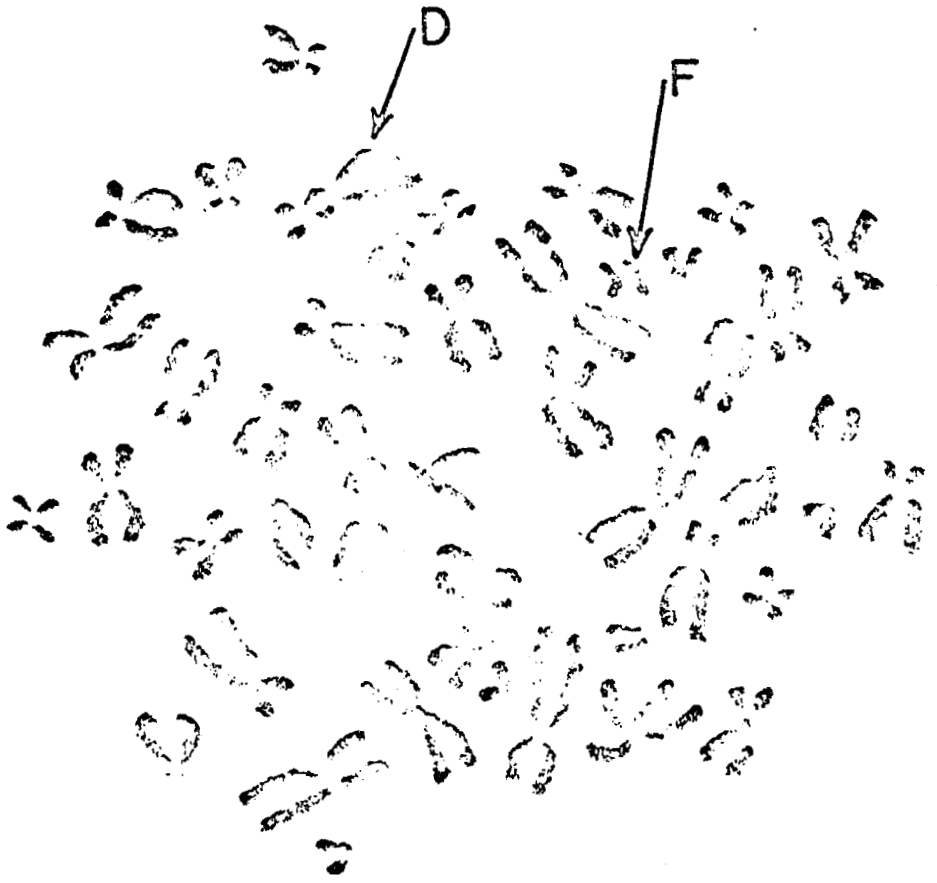
1 2 3 4-5(B) (XY)

6-12+X(C) 13-15(D)

16-18 (E) 19-20(F) 21-22(G)

1034400

42


D
F