Computer-assisted DNA length measurements from electron micrographs with special reference to partial denaturation mapping

Roland K. Littlewood and Ross B. Inman

Biophysics Laboratory and Department of Biochemistry, University of Wisconsin, Madison, WI 53706, USA

Received 14 September 1981; Revised 5 January 1982; Accepted 25 January 1982

ABSTRACT

A convenient semi-automatic procedure is described for the measurement, manipulation and display of electron microscopic partial denaturation data. Preliminary work on a completely automatic system is also discussed.

INTRODUCTION

Obtaining DNA length measurements from electron micrographs is an error prone and tedious task. It is advantageous to delegate as much of the procedure as possible to a computer without bypassing various subjective decisions that must be made in the analysis of electron micrographs. We wish to describe a semi-automatic procedure which requires that the operator actually traces each molecule (and thereby makes all pertinent decisions); apart from this, practically all other aspects of the process are handled automatically.

The technique of partial denaturation mapping (1,2) provides a useful frame of reference that allows the mapping of other features of interest on a DNA molecule. A typical partially denatured molecule may contain about 20 denatured sites and we would therefore need 41 segment lengths to define its denaturation pattern. If a particular experiment required analysis of 50 molecules, then over 2000 data would need to be acquired, stored and displayed. In addition, we would also need a record of molecule identification and positional information for other points of interest on each molecule. We have found that the semi-automatic system proves to be a very useful aid in answering many questions amenable to electron microscopic investigation and this system has been used for many years in this laboratory. However, there are also many types of investigation that require measurement of a 10-100 fold larger sample or longer molecules and here the manual tracing task becomes a prohibitive factor in such an undertaking. After a description of the semi-

automatic measurement system, we will describe preliminary attempts to completely automate length measurements and thereby eliminate the need to photograph and trace the molecules.

SEMI-AUTOMATIC METHOD

General outline

Projected images of micrographs are traced by hand using a digitizer which sends a stream of X-Y coordinates to the computer and a computer program continuously calculates distances moved by the digitizer pen. Distances between any two points on the molecule are associated with codes sent from a 16-key keyboard operated in conjunction with the digitizer. This length information can then be printed, stored and plotted to yield a linearized version of the molecule showing positions of denatured sites and other features of interest. The information can be manipulated in various ways, secondary data extracted and then final versions of the molecules plotted. After any given set of molecules has been processed, information about the sample average is output in the form of a histogram or printed material. The various components of the system are set out in diagrammatic form in Fig. 1.

Tracing and recording of length information

Micrographs are recorded on 70 mm or 35 mm roll film and projected onto the underside of a 59 cm diameter ground glass screen. A PRADO-UNIVERSAL (Ernst Leitz) projector containing three interchangeable lenses and a folded optical system is used to yield magnifications of 9X, 13X and 18X from film to screen. Provision is made for motor-driven X-Y film translation and focus adjustment and the controls for these are located at a convenient position to one side of the viewing screen.

The magnified image is then traced, along the top side of the viewing screen, using a digitizer. The digitizer used in this system is a NUMONICS GRAPHICS CALCULATOR which was chosen because it can be used with a back projected image, is of sufficiently high accuracy and of reasonable cost. This digitizer can resolve 0.025 cm over a working range of 59 cm x 59 cm. As the tracing head is moved along the image, a stream of X-Y coordinates is accepted by the computer at a rate of more than 100 coordinate pairs/sec via a BCD interface. A hexadecimal keyboard, located close to the operator, can also be used to input information through the BCD interface and to signal various facts about the stream of X-Y coordinates. The keyboard is used to indicate the start and end of a molecule, the start and end of denatured sites, the position of other points of interest and to identify various types of mole-

Fig. 1. Semi-automatic measurement system.

The computer (model 9825T, 61670 bytes read/write memory, system programming ROM and flexible disc ROM), flexible disc drive (9885) and X-Y plotter (9872) are from Hewlett-Packard (Colorado, USA). The printer (Model 40) is manufactured by Teletype Corp. (Illinois, USA) and the digitizer (Electronic Graphics Calculator) is from Numonics Corp. (Pennsylvania, USA). The Prado Universal projector (Ernst Leitz, Wetzlar, Germany), viewing screen and motor driven film, focus and translation controls were mounted in a cabinet constructed by Physical Sciences Lab (Wisconsin, USA). The keyboard (ER2296 Chomerics, Massachusetts, USA) logic circuit was constructed by ourselves.

cules. The keyboard is also used to restart measurement in case of an obvious error while tracing. The computer responds to these various keyboard signals with coded audible signals which confirm what has just been input to the computer. When the user signals the end of a molecule, the length data for the various segments, along with information about what they represent, are stored on a floppy disk (alternatively, the data can be stored on magnetic tape). The segmented length data for each molecule are accompanied by a molecule number and a code number (which can be used in various ways to cha-

racterize a molecule, i.e. circular, linear, branched, etc.). The length/base pair of double stranded DNA differs from the length/base of single stranded DNA; because of this, all denatured site segment lengths are multiplied by a factor prior to storage (under our spreading conditions this factor is found to be 0.95).

The computer program uses a modified version of the hypotenuse rule to calculate lengths, based on the incoming stream of X-Y coordinates. One modification is to apply a jitter test, whereby X-Y coordinates which are not at least 0.05 cm distant from the most recently accepted coordinates are rejected. This tends to remove the effects of small randomly-oriented movements generated during manual tracing. A second modification is to apply a "single exponential smoothing" procedure (3). Each incoming coordinate is transformed into a smoothed coordinate by the formula:

$$C_n = S \cdot U_n + (1-S) \cdot C_{n-1}$$

where U_n is the unsmoothed coordinate (either X or Y), C_n is the smoothed coordinate, C_{n-1} is the previous smoothed coordinate and S is the smoothing factor (a value between 0 and 1; typically 0.35). This tends to dampen the "stair-step" effect caused by the finite resolution of the coordinates transmitted by the digitizer. Finally, we apply a test to check that the keyboard signals are reliable; this is only necessary for keyboards that do not have debounce circuitry. It is important that the program code for the digitizer reading loop has a reasonably fast execution time. A slow reading loop (less than 30 readings/sec) will decrease the rate at which coordinates are acquired and will therefore require slower manual tracing velocities. The reading loop currently used is executed more than 100 times/sec and for tracing velocities between 0.3 and 0.8 cm/sec (see below) this will result in acquisition of coordinate pairs at approximately 0.005 cm length increments. This value is a reasonable match with the resolving power of the digitizer (0.025 cm). Tests with wire models show that lengths of either gently curved or highly convoluted lines (containing many bends with radii of curvatures down to 0.5 mm) can be determined with an accuracy of $\pm 1\%$. These tests were carried out at a tracing velocity which enabled convenient but accurate tracing along typical DNA images (0.3-0.8 cm/sec).

Alignment of partial denaturation patterns

The technique of partial denaturation mapping produces a frame of reference on an otherwise featureless DNA molecule. Such a frame of reference can then be used to answer questions about the position of some feature of interest on the molecule. The alignment procedure can be conveniently carried

out with computer assistance and results in high precision and a great saving of time.

As an example, Fig. 2(a) shows an electron micrograph of a partially denatured replicative intermediate of bacteriophage λ DNA. This particular molecule consists of a circle and a tail (sigma structure). The branch point is shown at (A) and the tail ends at (B). This molecule was measured in two parts; first the circle, starting in the direction of the arrow at (S), and then the tail segment, starting at the branch point and ending at (B).

The actual tracing procedure for this particular molecule involves the following steps. First the digitizer cursor is placed at the chosen starting position on the circle which corresponds to the cross-over position near (S) in Fig. 2(a) and a "start of molecule" signal is transmitted from the keyboard. As the circle is traced, the beginnings and ends of denatured sites and single stranded gaps are signalled with pre-assigned codes from the keyboard as is the position of the branch point (A). When the tracing of the circle is complete and the cursor has been returned to the starting position, an "end of molecule" signal is generated at the keyboard followed by a type code which can be used to identify the type of molecule (in this case a circle). At this stage the program will output a linearized version of the molecule at the X-Y plotter (Fig. 2b). Denatured sites are shown as rectangles drawn above the line while the single stranded gap is indicated by a rectangle below the line. The rectangles corresponding to denatured sites have been filled in for illustrative purposes; the actual plotted output yields open rectangles. The branch point is the perpendicular bar drawn below the line. Because this molecule is a circle, the line segment representing the molecule has common left and right ends. The tail segment is now measured, starting from the branch point (A) and ending at position (B). It is given a type code indicating that it is a tail segment. The plotted output for this tracing is shown in Fig. 2(f). As each molecule is traced it is automatically assigned an identification number by the program.

We now wish to align these segments so that (1) the denatured sites (in (b) and (f)) give a good match with those of mature linear λ DNA and (2) the branch-points recorded in the two representations also match. Fig. 2(k) shows the average position of sites in mature λ DNA, determined in an earlier experiment at a similar degree of denaturation. The standard molecule is represented with a length of 17.06 μ and because λ DNA is known to have a homogeneous length, the first part of the alignment procedure is to rescale the measured length of the circle to 17.06 μ ; this is then plotted out as in (c).

Fig. 2. (a) Electron micrograph of a λ DNA replicative intermediate. (b-e) Computer assisted manipulation of the denaturation map corresponding to the circular part of the molecule. (f-i) Manipulation of the branched portion of the molecule. (j) Final aligned molecule. (k) Denaturation pattern of mature linear λ DNA obtained at a similar degree of denaturation. Denatured sites are shown as black rectangles while single stranded gapped regions are indicated by the open rectangles.

A comparison of (c) with the standard map (k) shows that sites will have a similar relative order if the map (c) is reversed as shown in (d). If the circular map (d) is now permuted so that the position shown by the arrow is moved to the left end, a good match is obtained with the standard. The aligned molecule is then plotted as shown in (e). A similar procedure is used with the tail segment (f). It is first rescaled, using the same scaling factor used for the circle, to give (g). Reversal gives a good match for relative position of sites and is shown in (h) and finally the branch point is aligned with the common position on the aligned circle by shifting the map, as shown in (i). The final aligned segments are redrawn in (j) and show good alignment of denatured sites with the standard map (k). It can be seen that in this particular molecule the two single-stranded gapped regions align quite closely and one can be reasonably certain that these two regions broke apart at some stage of replication or sample preparation. This contention is strongly supported by the fact that the micrograph (a) shows a short broken single strand (T) protruding from the circle at exactly the position at which the tail should have joined the circle in the original structure. It should be noted that λ DNA usually replicates as a theta structure with two growing points (4,5) and that the above exceptional molecule was chosen simply as a good example of the alignment procedure.

In any given set of molecules there will usually be some that are broken. During the alignment procedure these molecules can be recognized and dealt with in one of two ways. Broken molecules can be rejected from the data set (by giving them a special code number). Alternatively they can be retained and allowance made for the missing end by shifting the fragment either to the left or to the right to produce a good fit with the standard denaturation map of the molecule.

Display of aligned molecules and extraction of secondary data

After a set of molecules have been aligned, as just described, a final version can be plotted. This program allows any given set of molecules to be plotted and any given type of molecule can be selected on the basis of its type code. The program will allow each molecule to be rescaled to any appropriate constant length or to be plotted out without further change in scaling. During this final plotting stage, two further types of information are extracted and stored for future use. First, a histogram is compiled which represents the average frequency and position of denatured sites (generation of histograms is described in the next section). Various other types of information can also be obtained at this time. For instance, a listing of

position of branch points, position or size of denatured sites, length of molecules or almost any other parameter of interest can easily be extracted from the data and stored for later use in programs which produce either graphs or histograms of such secondary data. The histogram averages are computed only for the segments recorded as denatured sites (such segments are drawn as the solid rectangles above the line in Fig. 2). Provision is also made for various alternatives in the final plotting and compilation of histogram averages. For instance, once molecules have been aligned on the basis of their denaturation patterns, it is often desirable to plot the molecules without showing the denatured sites. The program is arranged so that under these conditions the histogram average then refers to the position of other points of interest on the molecule.

Denaturation map histogram averages

In an experiment that rests on partial denaturation mapping as the means of alignment, it is important to have some assurance that the alignment is valid. It is therefore essential to be able to express the average denaturation pattern in some convenient way. This is accomplished by a histogram average, which represents a normalized frequency distribution of sites. Each molecule is divided up into a maximum of 1000 segments and a histogram constructed of rectangles for each of these segments. The height of each rectangle is proportional to the fraction of molecules that have a site covering the segment. Histograms therefore show the average denaturation pattern. The position of the peaks in such a histogram reflects the position of A+T rich regions. Fig. 3(c) shows such a histogram for mature linear λ c72 DNA and was generated by averaging 61 molecules. Because these molecules were linear, no alignment was required except that the molecules were reversed if necessary so that the left halves were always those with the fewest sites. The molecules were all normalized to 17.06 μ prior to plotting the histogram. Once a histogram has been generated, it can be manipulated in much the same way as described earlier for manipulation of individual molecule maps. Maps can be rescaled, reversed, permuted and translated. Additionally, they can be combined with other histograms (either by addition or subtraction); this allows one to also delete regions or insert segments from other histogram averages. Fig. 3 shows an example that involves many of these manipulations. Ch9oopS3 is a variant of phage λ which was constructed so that it contained two origins of replication (6). The relationship between λ and this variant is shown in (a) and (b). A large section from the central region of λ was excised and replaced by a small fragment, composed of two regions close to the

Fig. 3. Examples of manipulation of histogram average denaturation maps. (a,b) Relationship between λ DNA ($\lambda c72$) and a variant ($Ch9oopS3$) containing a substitution, insertion and a deletion. (c) Histogram average map of the parental molecule ($c72$). (d) Predicted map of the variant derived from (c) using the information shown in (a) and (b). The region at 11.3μ is blank because the short section of inserted material at this position has an unknown denaturation pattern. (e) The experimental histogram average of the variant ($Ch9oopS3$).

normal origin of replication (the two solid areas in (a)). Additionally, the variant has an insertion at the 13.15μ position and a deletion between the 14.20μ and 15.25μ positions in λ DNA. The histogram average for the parental DNA ($\lambda c72$) is shown in Fig. 3(c). The information shown in (a), (b) and

(c) was used along with the histogram manipulation program to construct a predicted histogram average of the variant molecule (Ch9oopS3); this construct is shown in (d). When the variant phage was actually made, then an experimental histogram average could be determined; this is shown in (e) and was generated from 48 individual denaturation maps after they had each been normalized to 14.6 μ . Most of the peaks agree very well except for the very sharp peak at 6.9 μ in the constructed histogram.

Computer requirements and Availability of Programs

The combined HPL program code described in the preceding sections occupies about 98 Kbytes and is stored in segmented form on a magnetic tape cartridge. The programs are operated as four separate working units. By suitable segmenting and automatic overlay techniques the combined program and data memory requirements for each working unit can be contained in 21 Kbytes of computer memory. These programs are available for the use of others and may be obtained on magnetic tape from the authors. The program will run on Hewlett-Packard 9825 and 9826 systems, by the use of the Hewlett-Packard HPL to BASIC translation program, they can be modified to run on the Hewlett-Packard 9835 and 9845 systems. Other programs dealing with analysis of heteroduplex molecules are also available. (Plans for the construction of the film projection unit, including film transport, X and Y movement and focus control, are available from Physical Sciences Laboratory, Stoughton, Wisconsin. Interface wiring between digitizer and computer can be obtained from Numonics Corporation, Lansdale, PA and the circuit for the hexadecimal keyboard can be obtained from the authors).

TOWARDS FULLY AUTOMATIC DNA LENGTH MEASUREMENTS

General concept

The procedure described so far has proven to be close to ideal for experiments involving about 50 molecules, each with about 20 denatured sites. Experiments that require 10X or 100X as much data would be too time consuming, even using this semi-automatic procedure. There are, therefore, many interesting experiments with bacterial DNA which cannot be undertaken at the present time. If we could automatically measure molecules, and make these measurements quickly, the experiments requiring acquisition of larger amounts of data would at once become attractive. We will now describe preliminary work along these lines but wish to note that the work that has so far been carried out is simply to find out if, in the near future, automatic line following is a practical notion to consider. Our experiments have not been

designed to be of practical use but rather to answer the above question.

The system that we have examined is shown in diagrammatic form in Fig. 4. It consists of a photosensitive detector (RETICON AR-32x32A) containing an array of 32×32 photosensitive diodes, this array has an overall dimension of 0.32×0.32 cm. The array is mounted on a simple optical system (containing a lens and several infrared emitting diodes) so that an image of a small segment of the DNA molecule is formed in the plane of the photo-diode array. The 1024 photo-diode signals are sent to the computer and a program is designed to work out the heading of the DNA segment viewed by the photosensing head. Once a heading is determined the program directs the movement of the photo-diode array a short distance along this heading. The computer then receives new data from the photo-diode array and the process is reiterated.

Fig. 4. Automatic measurement system.

The computer, plotter, flexible disc drive and printer have already been described in Fig. 1. The 32×32 photodiode array (RA-32x32A) was purchased from Reticon Corp. (California, USA) and the video conditioning and data conversion interface was made by the Physical Sciences Lab (Wisconsin, USA). The optical and I.R. illumination system was constructed in our laboratory.

There are two problems associated with this scheme. First, we have to be sure that the photo-diode information is accurate and presents sufficient information to enable a sophisticated program to determine line headings. Second, one would have to design a program to handle this information and to be able to determine the line heading with high reliability. We feel that the first problem has been solved and that good progress has been made with the second, more difficult, aspect of this project. Fig. 5(a) shows a section of an electron micrograph of a partially denatured DNA molecule and (b) is a visual representation of what is received by the computer when the array is set over position X in the micrograph. The complete montage (c) was accumulated after the array was made to track between the arrows shown in the micrograph. It can be seen that the information is sufficient for the human eye to follow the molecule; in principle, therefore, one should be able to design a computer program to do the same.

Line following programs

In order to find out how complex a computer program has to be in order to follow a DNA image with reasonable reliability, we have initially made three simplifications. First, rather than trace DNA images we have used hand-drawn black ink representations. Second, these tracings contained no crossovers or denatured sites but did have curvatures well below those corresponding to typical DNA images (radii of curvatures down to 0.1 cm). Third, because initially we are interested only in learning about the complexity of the problem, we have used a computer and a programming language which are designed more for ease of use than for speed of execution. The computer is a Hewlett-Packard 9825; the language is HPL, a highly interactive, interpreted derivative of BASIC developed by Hewlett-Packard specifically for use on its 9800 series of desktop computers. We estimate that the tracing rates, to be presented below, could be increased 10-100 fold by the use of a faster more appropriate machine and a faster more appropriate programming language.

Two programs have been designed which show a reasonable ability to follow curved lines. The first program uses the principle shown in Fig. 6. The 32 x 32 array of photo-diode information consists of either true or false data bits corresponding to the presence or absence of the line image at the position of each diode (a true value will be called a "hit"). The program constructs a histogram representing the proportion of hits in each column of diodes which is perpendicular to the selected baseline. Histograms are constructed for several baselines which are at different angles with respect to the fixed base of the diode matrix. For baselines which are neither parallel nor perpen-

Fig. 5. (a) Micrograph showing a short section of a partially denatured DNA molecule. (b) Example showing one complete frame of data sent to the computer when the photo-diode array was placed at position X on the micrograph. (c) Montage of many frames of data obtained when the array was moved along the molecule between the arrows shown in (a).

Fig. 6. Diagram showing one method used to deduce the heading angle from photo-diode array information obtained from a short section of a DNA image.

dicular to the fixed base of the diode matrix, the columns represented near each end of the histogram contain relatively few diodes, so the proportional data for these columns are less reliable than for columns represented near the center of the histogram; these outside columns are ignored for such baselines.

A histogram constructed along a 0° baseline for the molecule segment depicted in Fig. 6 is quite flat (6(a)). Similarly, a histogram constructed along a 270° baseline for the same molecule is relatively flat (6(b)). However, a histogram constructed along a 45° baseline has a quite sharp peak (6(c)) and therefore the heading of the line should be approximately perpendicular to this baseline. The program is designed to detect the sharpest peak from 4 baseline angles and to then move the diode matrix along the calculated best heading. In practice, the heading is derived by interpolation between the 4 histogram baselines and the actual distance moved along this angle is modulated by the sharpness of the best histogram. This program has been found to work quite reliably for most situations where only one line is covered by the diode matrix. Using a standard test image this program results

in a tracing rate which is 500X slower than manual tracing (in a fast computer this would presumably result in rates that were 5-50X slower).

The second program analyzes the hit data in a much less detailed manner. This program utilizes only 24 rows and 24 columns of the diode matrix -- these are partitioned into 9 blocks with 8 rows and 8 columns (64 diodes) each. The data analyzed are the hit counts for each of the blocks. These block hit counts can be obtained approximately 100 times faster than the histograms described above; therefore, this program runs considerably faster. If the heading of the molecule is as shown in Fig. 7(a) and the most recent diode matrix movement had placed the matrix center over block X, then block 2 would contain the highest number of hits (block X is used only for consistency checking; block 7 is ignored because the program keeps track of the general direction in which the molecule is heading, which is away from block 7). In practice, the next heading to use is calculated by interpolation between the block with the highest hit count and its two nearest neighbors; the extent of the movement is based on the magnitude of the highest hit count and on the ratio of the hit count in the highest-count-block to the combined hit counts of its two nearest neighbors. This program has a tracking speed 5-10X slower than manual tracing (up to 20X faster if a fast computer were used).

Both tracking programs can automatically trace and calculate distance moved along the ink-drawn test images and images of DNA molecules that do not contain crossovers or denatured sites.

Fig. 7. (a) Simplified method for obtaining heading angle information. (b,c) Proposed method for obtaining information about complex configurations. In (b) we show a crossover, which would lead to the detection of 4 regions in a peripheral band of diodes. In (c), the denatured site would be detected by 3 such zones (solid areas).

Further work is now in progress to deal with the complications due to crossovers and denatured sites (Fig. 7(b,c) respectively). One way to alert the program to these conditions is to use a peripheral band of diodes to detect how many line segments are actually being monitored by the array. A crossover (b) would yield a value of 4 in the peripheral band (shown as the black areas) while the start or end of a denatured site (c) would yield a value of 3. A simple DNA segment and the beginning or end of a molecule would yield values of 2 and 1, respectively. Such information could then be used to direct the tracking program to deal with these complications. For instance, when 4 or more segments are detected in the peripheral band (a cross-over or two separate molecule segments), the tracking program might strive to proceed slowly along the curve which is the best fit through the most recent 3-4 tracked positions, until the number of segments in the peripheral band decreases to less than 4. This would be analogous to the course an automobile driver might steer when the view of the road is temporarily obstructed in a storm. In the case of the detection of the start of a denatured region (a transition from 2 to 3 segments in the peripheral band), the program could calculate a heading which proceeded from the widest segment (which would be the undenatured region) to either of the other segments. In the case of the detection of the end of a denatured region (another transition from 2 segments to 3 segments, where the two denatured strands come back together after one of them has been outside the coverage of the matrix for several moves), the program could calculate a heading which would move toward the widest segment. Starts and ends of denatured regions would have to be distinguished from context.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health and the American Cancer Society.

REFERENCES

1. Inman, R. B. (1960) *J. Mol. Biol.* 18, 464-476.
2. Schnös, M. and Inman, R. B. (1970) *J. Mol. Biol.* 49, 93-98.
3. Hewlett-Packard 9864A Digitizer Peripheral manual (1972).
4. Ogawa, T., Tomizawa, J. and Fuke, M. (1968) *Proc. Natl. Acad. Sci.* 60, 861-865.
5. Schnös, M. and Inman, R. B. (1970) *J. Mol. Biol.* 51, 61-73.
6. Schnös, M., Denniston, K. J., Blattner, F. R. and Inman, R. B. (1981) submitted.