Computer-based image analysis of one-dimensional electrophoretic gels used for the separation of DNA restriction fragments

Alexander J.Gray*, David E.Beecher and Maynard V.Olson⁺

Biomedical Computer Laboratory and ⁺Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

Received 15 August 1983

ABSTRACT

A stand-alone, interactive computer system has been developed that automates the analysis of ethidium bromide-stained agarose and acrylamide gels on which DNA restriction fragments have been separated by size. High-resolution digital images of the gels are obtained using a camera that contains a one-dimensional, 2048-pixel photodiode array that is mechanically translated through 2048 discrete steps in a direction perpendicular to the gel lanes. An automatic band-detection algorithm is used to establish the positions of the gel bands. A color-video graphics system, on which both the gel image and a variety of operator-controlled overlays are displayed, allows the operator to visualize and interact with critical stages of the analysis. The principal interactive steps involve defining the regions of the image that are to be analyzed and editing the results of the band-detection process. The system produces a machine-readable output file that contains the positions, intensities, and descriptive classifications of all the bands, as well as documentary information about the experiment. This file is normally further processed on a larger computer to obtain fragment-size assignments.

INTRODUCTION

One-dimensional gel electrophoresis has become established during the past decade as the central analytical technique of experimentation with both proteins and nucleic acids. Because a wealth of useful information can be extracted from these gels by simple hand measurements, augmented if necessary with lane-by-lane scanning densitometry, there has been little emphasis on automating the analysis of these gels. This situation contrasts with that for two-dimensional protein gels, in which case the high information content of typical individual gels has provided strong motivation for computerized analysis (1-5).

In principle, the same hardware that has been applied to two-dimensional protein gels could be adapted to the analysis of one-dimensional gels. However, recent advances in imaging technology have decreased the attractiveness of vidicon and rotating-drum scanners, which form the basis of most current two-dimensional gel-analysis systems, relative to solid-state

cameras, whose photo-sensors are partitioned into linear or rectangular arrays of uniformly spaced, independent detectors by integrated circuit technology. Furthermore, from the standpoint of software development, one- and two-dimensional gels pose quite different problems.

For these reasons, we decided to meet our need to analyze large numbers of photographs of ethidium bromide-stained agarose and acrylamide gels by the de novo development of a stand-alone, interactive computer system that would be dedicated to this specific task. As recombinant DNA techniques are used to investigate progressively larger gene families and chromosomal regions, there is likely to be an increasing demand for computer processing of photographs of DNA gels. Particularly when extensive inter-gel comparisons are required, as is the case in any project that involves the analysis of hundreds or thousands of clones, there is no practical alternative to computer processing. In this paper, we describe the hardware and software that make up our image-processing system and illustrate its application to the analysis of typical gels.

MATERIALS AND METHODS

Photographs of Stained Gels

The gels were stained in 0.1 $\mu\text{g/ml}$ ethidium bromide and photographed using 300 nm transillumination to stimulate the fluorescence. Sheets of 4x5-inch Eastman Kodak contrast process pan film (No. 4155) that had been lightly fogged by pre-flashing with white light were exposed for 5 min at f4.5 through an f=135 mm lens. Eastman Kodak Wratten No. 9 and No. 23A filters were placed in front of the camera lens, with the No. 9 filter towards the gel; the order of the filters is important because the No. 23A filter fluoresces if exposed to 300 nm light.

Photographs of Digitized Images

Our system provides no method of obtaining photographs of the displayed images other than directly photographing the color video monitor. Although photographs taken in this way are adequate for local documentation, they are unsatisfactory for publication. The figures in this paper were produced by transporting the digitized images to an image-processing system at the Laboratory of Neuro Imaging (Washington University School of Medicine) where they were displayed and photographed using a Model 632 Color Camera System (Dunn Instruments); this system has a built-in video monitor that is specifically designed to be photographed with a standard 35 mm camera.

HARDWARE CONFIGURATION

Overview

The principal hardware components of the imaging system are illustrated in Figure 1. They include a digitizing camera, an image processor with a color video monitor, standard peripheral devices (keyboard terminal, disk-storage units, and a printer), and a microcomputer, all of which are available from commercial vendors. Because the digitizing camera lacks sufficient sensitivity to detect the ethidium bromide fluorescence of a stained gel directly, the digital image is acquired by scanning a trans-illuminated 4x5-inch negative photographic transparency of the gel. The resolution of the displayed image, which can be seen on the video monitor in Figure 1B, is limited by the display system to one-quarter the linear resolution that is produced by the digitizing camera for internal use during automatic band detection. Nonetheless, the video image preserves nearly as much detail as can be discerned by eye on the original photograph.

In the following sections, the characteristics of the individual hardware components are described in more detail.

Hardware Components

Microcomputer. The system is organized around a dedicated LSI 11/23 microcomputer running the RT-11 (version 4) operating system (Digital Equipment Corporation, Maynard, MA). The LSI 11/23 was chosen primarily for its complete compatibility with the other major components of the system. For example, the manufacturers of the digitizing camera and the image processor both supply high-performance hardware interfaces for the LSI 11/23's Q-Bus, as well as software device handlers written for the RT-11 operating system. Although the LSI 11/23 can also function as an all-purpose microcomputer, our intention of carrying out such computationally intensive procedures as fragment-size assignments and inter-gel comparisons on a separate larger computer made it unnecessary to design the imaging system around a large central processor.

Digitizing Camera. The camera is an EC 78/99 image digitizer (Eikonix Corporation, Bedford, MA). This camera contains a linear 2048-pixel (picture element) photodiode array mounted on a motor-controlled stage. The linear array is aligned parallel to the gel lanes so that any given sampling line contains intensity readings at 2048 points along the direction of electrophoretic migration. The stepping motor can drive the array through 2048 discrete sampling positions to produce a 2048x2048-resolution digital image. The camera returns an independent intensity reading for each pixel;

A

B

Figure 1. Components of the gel-imaging system. A. A schematic drawing in which the arrows indicate directions of information flow during various phases of the system's operation. B. A photograph of the system in use. The dark band across the bottom of the video monitor is a photographic artefact.

in our applications, we recover this intensity reading in one byte (8 binary digits), only five or six bits of which are significant (ca. 2% precision). The camera is supported by an extensive electronics package, which allows the LSI 11/23 microcomputer to control automatically both the camera's stepping motor and the intensity of the trans-illuminator's light source. The camera is presently fitted with a fixed-focus $f=80$ mm lens mounted 35 cm above the gel photograph. This configuration provides an 11x11 cm field of view, which is suitable for our 4x5-inch negatives. Alternate film formats could be accommodated by changing the lens and re-positioning the camera.

Image processor and color video monitor. The image processor is a System 3400 Video Image and Graphics Processor (Lexidata Corporation, Billerica, MA), which drives an 8830 Analog Patriot Series high-resolution, 19-inch color monitor (Aydin Controls, Fort Washington, PA). The image processor's basic task is to store a digital image in its pixel memory and to use this image to refresh the display on the video monitor. The pixel memory is organized as a 640x512 array, each position of which can store 8 bits of data. The elements of the array are in 1:1 correspondence with the picture elements of the video image, which is formed from 512 raster lines, each of which contains 640 pixels.

In addition to its basic task of storing the video image and refreshing the image on the monitor, the System 3400 performs a number of more sophisticated functions. These functions are most easily understood with reference to the structure of the 8 bits of data that are stored in each address of the image processor's pixel memory. In our applications, the gray-scale gel image is normally stored in the first six bits, an arrangement that allows 64 gray-scale levels to be assigned to each pixel; this number of levels is more than sufficient to produce the illusion of a continuously-graded gray scale in the displayed image. The translation of these stored values into intensity levels on the video monitor is easily modified since on each refresh cycle the image processor translates them into intensity levels by employing a set of 64 scale factors that are stored in a "lookup" table. By simply re-writing the lookup table, it is possible to alter such image characteristics as the background gray level or the degree of contrast in the image. This capability allows the operator to switch rapidly between a variety of display modes, each of which optimizes the visibility of particular image features such as faint bands or overexposed regions.

Similarly, the use of lookup tables by the image processor facilitates the superposition of colored overlays on the gray-scale gel image. In most of

the gel-imaging software, the last two bits in each pixel memory address are used to control blue (bit 7) and red (bit 8) overlays that are used for such purposes as labeling lanes, indicating the positions of detected bands, and displaying a moveable cursor. As in the case of different gray-scale display modes, it is possible to toggle the overlays on and off almost instantaneously simply by changing the contents of the lookup table (rather than by the more time-consuming process of re-writing the entire pixel memory, which contains 0.3 Mbytes of data.)

The video monitor itself has two principal distinctive features that are important to the gel-imaging application. One of these is its 512-line resolution, which is twice that of the typical home television screen. The other is its long-persistence phosphor, a feature that sustains the image with minimal flicker. Long-persistence phosphors are unsuitable for displaying motion, but they provide superior displays of the static images that are created during the processing of a gel photograph.

Standard Peripheral Devices. The operator controls the system from a VT101 keyboard CRT terminal (Digital Equipment Corporation) and obtains printed output from an LA120 printer (Digital Equipment Corporation). The system was originally developed, and can still be operated, using an FD-211 Double Density Floppy Disk (Charles River Data Systems, Natick, MA) as its only mass-data-storage device. This dual-disk drive provides 1 Mbyte of on-line data storage capacity and also offers a convenient means of transferring files to and from other computer systems. In order to facilitate the future development of a more sophisticated software package, we recently added an FCV06 Winchester Disk Drive (Plessey Peripheral Systems, Irvine, CA) which provides 28 Mbytes of on-line data storage capacity.

SOFTWARE SYSTEM

Introduction

Our goal in designing the software was to produce a modular package of programs that would at the same time meet our immediate need for an operational system and also provide a convenient framework within which system enhancements could be introduced in the future. To avoid excessive dependence on a particular experimental format, we made a number of steps interactive, such as lane-finding and the definition of the precise regions of the gel that are to be analyzed, that could in the future be completely automated. Similarly, we put more emphasis on the convenience with which the raw results could be edited interactively than on optimizing the performance of the

automatic band-detection algorithm.

The software was developed under the RT-11 operating system using RATFOR ("RATional FORTRAN") as the high-level programming language (6). Although the LSI 11/23 will support more powerful operating systems, RT-11 (a standard Digital Equipment Corporation single-user, real-time operating system developed for the PDP-11 family of computers) offers a good compromise between performance and mass-storage requirements. While RT-11 can be implemented on systems that use only floppy disks for data storage, it provides all the software utilities that are required to support such sophisticated peripheral devices as the digitizing camera and the image processor. The high-level programming language that we used, RATFOR, combines the advantages of FORTRAN's universality with the structured programming features of more modern languages. RATFOR source programs are converted by a pre-processor to functionally equivalent FORTRAN versions, which can be compiled by standard FORTRAN compilers.

In addition to over one hundred RATFOR program modules, the software package includes a few short assembly-language modules that expedite communications with the digitizing camera and the image processor. These modules are exceptions to the general rule that most transactions between the LSI 11/23 and these devices are programed using FORTRAN-callable subroutines supplied by their manufacturers.

Organization into Subsystems

The overall software system is organized around a menu module from which any one of seven subsystems can be entered using the RT-11 chaining facility. These subsystems are completely independent, communicating when necessary by passing from one to another the name of the disk file that contains all the permanently stored information that has accumulated about the current gel. Capsule descriptions of the seven subsystems, in the most likely order of use, are:

1. Null Normalization. This step is the quickest and simplest method of preparing the digitizing camera for use when the system is turned on. It simply loads into a small memory in the camera's electronics package a set of default scale factors, which are used to translate the raw output from the photodiode array into meaningful light-intensity values. The default scale factors take no account of individual differences in the dark currents or sensitivities of individual pixels.
2. Full Normalization. This procedure, which is an alternative to null

normalization, individually calibrates the 2048 pixels in the photodiode array. Dark and uniform bright field scans are carried out, and scale factors are computed individually for each pixel, normalizing its response relative to that of the other members of the array. In our experience, there is no major improvement in system performance when full normalization, as opposed to null normalization, is employed.

3. Gel-File Creation. This subsystem allows the user to create a new gel data file. The same file that will ultimately contain all the data acquired from the gel, also contains a header into which documentary information is entered during gel-file creation. This information includes a gel-identification code, relevant dates and personnel names, lane and clone identifiers, and descriptive comments about the overall gel, the individual lanes, and the individual clones. All information is entered in a fill-in-the-blank format, and the operator has an opportunity to change any entry on a particular "page" before proceeding to the next page.
4. Gel Reading. All the activities from the initial scanning of a gel through the automatic band-detection procedure are programmed in this subsystem. Its operation is described and illustrated in more detail below.
5. Gel Editing. This module, in which the operator edits the results of the automatic band-detection procedure, is also described in greater detail below.
6. Report Generation. The programs in this section supervise the generation of a printed report about the gel. This report contains all the documentary information entered during gel-file creation, as well as data specifying the position, the peak intensity, and the descriptive classification of each band that was detected.
7. Header Alteration. This subsystem allows the editing of documentary information entered during gel-file creation. It is useful not only for correcting errors, but also for entering supplementary comments about features of the gel that may come to the operator's attention during the processing of the image.

Primary Processing of the Gel Image

The initial analysis of the gel image, which is managed by the Gel-Reading subsystem, occurs in two phases. During the first phase, a video image of the gel photograph is displayed on the monitor, and the operator

interactively specifies the regions of the image that are to be analyzed. During the second phase, which is entirely automatic, high-resolution scanning is followed by band detection on a lane-by-lane basis.

This process will be illustrated here using the agarose gel shown (as a conventional photographic print) in Figure 2A. The digitized, 512x512-pixel video image of this gel is shown in Figure 2B; acquisition and display of a gel image at this resolution requires only 60 sec. Once the image is displayed, the operator defines the outer boundaries of the region that is to be analyzed with the aid of a moveable red cursor. Using single-stroke keyboard entries, the operator can move the cursor cross to the position of a given boundary and then order the corresponding coordinates to be stored in the gel data file.

From the positions of these boundaries, approximate locations of the lane centers are calculated and indicated by colored overlays on the displayed image. The operator carries out interactive fine-tuning of these estimates by adjusting the phase and the spacing of the lane-center overlay grid through the use of further keyboard entries. Finally, the width of the zone within each lane that is to be analyzed is specified interactively in a similar manner. The final result of the series of steps that leads from the end of the original scan (Figure 2B) to the stage immediately preceding band detection is an image such as that shown in Figure 2C, in which the zones to be analyzed within each lane and the overall boundaries of the image are demarcated with broad red overlays. This sequence of steps can be performed by an experienced operator in less than 2 min.

At this stage, automatic band detection ensues on a lane-by-lane basis. First, a high-resolution scan of the appropriate region is carried out by a sampling strategy that features both temporal and spatial averaging: the output of the photodiode array is recorded eight successive times at sixteen different scan lines within the lane, and these 128 sets of readings are averaged on the fly to produce a single sample vector. This one-dimensional sample vector is analyzed for peaks using the band-detection algorithm. The high-resolution scanning requires about 15 sec per lane, while the peak-detection calculations require an additional 7 sec.

Once automatic band detection has been completed for all the lanes, the gel image is displayed with an overlay of blue lines at the positions where bands have been detected. Each band is also annotated with one of three descriptors: S (singlet), D (doublet), or ? (questionable). Presently, the classification of bands is based on a simple comparison of the peak intensity

Figure 2. A conventional photograph and two digital images of an ethidium bromide-stained agarose gel. A. A conventional photographic copy of the original gel photograph. The experimental samples (lanes 2-6 and 8-11) are *Eco*I/*Hind*III double digests of lambda clones containing yeast DNA. Lanes 1, 7, and 12 contain size markers, ranging in size from 0.4 to 22 kb. B. A photograph of a digitized image of the same gel; the resolution is 512x512 and the densitometric measurements are represented by 64 gray-scale levels. C. The digitized image after the overall boundaries of the region to be analyzed and the sampling regions within each lane have been demarcated with broad, red overlays.

of each band with the intensities of its immediate neighbors. The auxiliary descriptors T (possible triplet), * (an all-purpose flag), and M (size-marker band) are available for interactive assignment by the operator.

Editing the Raw Results

Because any practical band-detection algorithm is certain to produce both false positives and false negatives, we put considerable emphasis on developing a convenient procedure for editing the raw output of the automatic band detector. The editing programs, which are contained in the Gel-Editing subsystem, provide a number of utilities to assist the operator in interpreting the gel image. Initially, the operator is simply provided with a video display of the gel image, in which colored overlays indicate the positions and classifications of the detected bands; a red, rectangular cursor, which identifies a particular band as the current candidate for editing actions, is also displayed (see Figure 3 - the cursor is in the middle of the first lane). The operator can delete or re-classify the band specified by the cursor or add a new band to either side of it. These actions, as well as the movement of the cursor from band-to-band, are all implemented with single-stroke keyboard commands.

Figure 3. Edit mode. The positions of the detected bands are indicated by blue lines that are superimposed on the gray-scale image. The single-character descriptors (S,D,T,M,*,? - see text) appear immediately above the bands in red.

To assist the operator in discerning subtle features of the gel image, three auxiliary commands are also available. The first of these simply allows the overlays on the gel image to be toggled on and off so that the raw image and its current interpretation can be compared sequentially. The second command toggles the intensity scale of the displayed image between the normal linear mode and a logarithmic mode. This capability is particularly important in the case of gels that contain a wide range of band intensities. In these situations, which occur frequently, the visibility of the faint bands is optimized by scanning the photograph at such low levels of trans-illumination that regions of the image containing intense bands become too dark to interpret. The final auxiliary editing command, allows the operator to toggle back and forth between the gel image and an annotated plot of the sample vector for the current lane. An example of a sample-vector display produced by this command is shown in Figure 4. Such plots are particularly useful when closely-spaced multiplets must be interpreted. The sample vector has higher intrinsic resolution than the video display for two reasons: first, it is displayed at a linear resolution of 1024 pixels, compared to the 512 pixels of

Figure 4. Vector-display mode. The raw sample vector for lane 3 of the same gel illustrated in Figures 2 and 3 is plotted at 1024-pixel resolution. To achieve this resolution the image is folded: the top of the lane is to the left of the vertical line in the bottom panel and the bottom of the lane is to the right of the vertical line in the top panel. In both panels, bands appear as minima in a plot of light intensity vs. pixel number.

the video display, and secondly it benefits from the extensive temporal and spatial signal averaging that is carried out during its construction.

SYSTEM PERFORMANCE

Our first-generation software is oriented towards the measurement of band positions, from which fragment sizes are calculated. In this section, we comment on the reproducibility of these measurements, as well as on two other important system characteristics - the quality of the photometric measurements and the spatial resolution of the digital images.

Mobility Measurements

Band positions are measured by identifying, within the band's light-absorption profile, the pixel with the lowest light-intensity value. Mobilities are simply reported as pixel numbers, which are counted starting at an arbitrary point in the array; no effort is made to identify the positions of the sample wells. Since the gel is aligned with the lanes parallel to the photodiode array, the origins of the lane vectors form an arbitrary reference line that is perpendicular to the direction of electrophoretic migration.

Reproducibility of the raw mobility measurements is excellent. Relative band positions generally agree within one or two pixels when the same photograph is scanned repetitively, even when the scans are made over a period of months. Since one pixel corresponds to 0.05 mm on the gel photograph, this reproducibility is much better than can be achieved by hand measurements. We find that the precision of hand measurements, even with the assistance of a hair-line viewfinder mounted on a micrometer, is no better than ± 0.2 mm, largely because of the subjectivity involved in defining the "positions" of bands that are typically several tenths of a millimeter wide.

Surprisingly, the commercial, single-lane gel scanners that we have tested are inferior even to hand measurements for defining band positions. The reason for their poor performance is that the measurements depend on some method of recording the position of a mechanical device that must move through large distances while translocating either the gel photograph or the scanning head. A major advantage of imaging systems that employ photodiode arrays is that they can be designed so that there is no mechanical scanning in the most critical measurement direction: the positional reproducibility along the array axis is built into the physical structure of the device.

Given a set of mobility measurements, our method of making the fragment-size assignments requires that a group of experimental lanes be flanked by two lanes containing identical sets of size markers. The

interpolation algorithm involves constructing a set of lines that connect the size-marker bands in one of the flanking lanes with the corresponding bands in the other, thereby creating a set of "size contours"; the sizes of the fragments in a given experimental lane are then determined by polynomial interpolation between the points where the lane-center line intersects the size-contour lines. This procedure makes the fragment-size assignments relatively insensitive both to non-idealities in the running of the gel and to mis-alignment of the gel photograph on the camera's stage.

With this method of making fragment-size assignments, the small imprecision that we observe in the raw mobility measurements gives rise to variations of only one or two tenths of a percent in the fragment sizes when the same gel is analyzed repetitively. Since this degree of imprecision is an order of magnitude less than the experimental irreproducibility that we typically observe when making inter-gel comparisons, it is not an important limitation.

Band Intensity Measurements

Under ideal conditions, the integrated absorbance associated with a band on a scanned gel photograph is proportional to the amount of DNA present in the original gel band. This assumption can fail for a variety of reasons that are related to the staining or photographic conditions rather than to the densitometry, but even given a suitable photograph, scanning densitometry makes considerable demands on the quality of the photometric measurements. The best results are obtained if the bands that are being compared have similar intensities, and if these intensities give rise to intermediate absolute absorbances.

The extent to which these conditions are met when a typical agarose gel is analyzed on our system is illustrated in Figure 5. This figure shows the absorbance profile for lane 3 of the same gel that has already been illustrated in Figures 2-4. The region of the lane shown in Figure 5 contains 15 peaks that correspond to 16 restriction fragments derived from a typical lambda clone. There is a single unresolved doublet migrating at pixel number 900, which corresponds to the fourth largest band in the lane. It is futile to attempt to analyze the two bands at the top of the lane, which are derived from the vector arms, both because of obvious saturation in the first peak and because of depletion of the second due to inter-arm re-annealing. The remaining peaks, however, were integrated in five groups, corresponding to the single peak at pixel number 1240, the clusters of four peaks at 840 and 680, the cluster of three peaks at 520, and the single peak at 380. By considering

Figure 5. Absorbance profile for lane 3 of the same gel described in Figures 2-4. This graph was prepared by converting the contents of the raw sample vector (see Figure 4 and text) into absorbance values and plotting them vs. pixel number.

the fragment sizes assigned to the 14 non-vector bands, the integrated absorbance corresponding to the five groups could be converted to estimates of the number of fragments per group. When normalized relative to the middle group, the estimates were in the ratio (0.6 : 4.8 : 4.0 : 2.7 : 0.8), while the theoretical values are in the ratio (1 : 5 : 4 : 3 : 1). These results illustrate the typical effect that bands which give rise to high peak absorbances (above about 0.9 in our system) have less than the expected integrated intensities due to saturation effects, while faint bands have relatively high errors due to poor signal-to-noise ratios. As a second example, the seven largest non-vector bands in lane 10 were chosen for analysis since they represent a series of well-isolated bands, the middle one of which is an unresolved doublet; in this case, when normalized relative to the third band in the series, the estimates for the number of fragments per band were in the ratio (0.8 : 0.9 : 1.0 : 1.8 : 0.9 : 1.1 : 1.0) while the theoretical values are in the ratio (1 : 1 : 1 : 2 : 1 : 1 : 1).

Although scanning densitometry is useful for a variety of purposes, its main application to gels on which restriction-enzyme digests have been analyzed lies in the recognition of unresolved multiplets. From the above results, it is apparent that the data produced by our system allow reliable distinctions to be made between singlets and doublets throughout the major part of a gel lane, and under some circumstances, considerably finer distinctions can be made. We have proceeded with caution, however, in embedding this type of analysis in our standard software because of the complexity of dealing with saturation effects, multiple overlapping peaks and regions with poor signal-to-noise ratios.

Resolution of Closely-Spaced Bands

Because digital images are composed of a discrete number of picture elements, there are inherent limitations on their spatial resolution. In the most favorable case (infinitely sharp bands and an ideal detector), a trivial limit applies: two bands must be separated by twice the pixel-to-pixel spacing in order to be resolved into separate peaks. In our system, an 11 cm field of view is divided into 2048 pixels, giving a pixel-to-pixel spacing of 0.054 mm. Since most gel bands are at least a few tenths of a millimeter wide, it is rare for the practical resolution of the system to be limited by the discreteness of the image. However, it is of interest that given extremely sharp bands on a gel photograph, the system's resolution approaches the theoretical limit. This point is illustrated in Figure 6, which shows a plot of the absorbance profile for a 5 mm region that contains nine bands;

Figure 6. An illustration of the system's spatial resolution. The top portion of the figure is a greatly enlarged photographic print of a 5 mm region of a native acrylamide gel; nine fragments that are approximately 1 kb in size are migrating in this region. The bottom portion shows the absorbance profile derived from a raw sample vector produced by the gel-imaging system. The boxes indicate the positions of the actual points corresponding to each pixel.

these bands, which were produced by fragments of about 1 kb on a native acrylamide gel, are much sharper than the bands obtained on agarose gels. The square boxes indicate the actual positions of the data points that correspond to successive pixels. The most closely-spaced bands, which are clearly resolved into separate peaks, are separated on the original gel photograph by only 0.2 mm, a distance that corresponds to four pixel-to-pixel spacings.

DISCUSSION

In this paper, we describe an interactive computer system that was designed to facilitate the analysis of one-dimensional DNA gels. Given the overwhelming importance of restriction mapping in nucleic acids research, there is a large potential demand for more efficient methods of analyzing such

gels. Much of this demand can undoubtedly be met with simpler, and much less expensive, technology than that described here. For example, digitizing tablets (7, 8) are seeing increasing use as a means of improving on the traditional ruler and pencil. However, the immense complexity of the DNA present in cellular genomes provides this area of research with a built-in trend towards the analysis of progressively more complex collections of DNA molecules. Our own motivation for developing this system, for example, came from a project whose ultimate goal is to carry out the complete physical mapping of the yeast chromosomes, largely by analyzing restriction digests of several thousand lambda clones. Experience to date with a few hundred gel photographs, containing approximately 5×10^5 bands, suggests that the efficiency and orderliness that this system brings to the processing of large numbers of gel photographs will be indispensable to this undertaking. Although DNA sequencing projects commonly require the processing of even larger numbers of bands, restriction-mapping problems pose a unique requirement for precise, quantitative measurements of band mobilities.

It is apparent that our present system is software-limited. The digital images are of excellent quality and can be acquired rapidly enough for convenient interactive operation. Although we have implemented a software system that solves all the basic problems of managing the acquisition and interactive processing of the image, work has only begun on the problem of extracting the useful information from the image completely automatically. While several heuristic band-detection algorithms have been published (9), and we have experimented with a number that we developed ourselves, more work is required on this problem. Given relatively clean data, it is not difficult to find an algorithm that produces only a few percent false positives and false negatives, but in our experience, the available procedures are not sufficiently robust to cope well with the ordinary vagaries of gel photographs. In particular, none that we have examined is remotely competitive with a trained human interpreter. Given the high quality of the digital images and - at least by artificial-intelligence standards - the basic simplicity of the task, the goal of reducing the required human input to a negligible level does not appear unrealistic. Algorithms that approximate this goal, however, are certain to be computationally intensive since the amount of data present in a full digitized image (ca. 4 Mbytes) is substantial. Nonetheless, given continuing rapid increases in the power of the computing hardware that is available to researchers at practical costs, it would appear that future enhancements of this system should be designed around

the real needs of experimentalists rather than the constraints imposed by today's generation of microcomputers.

*Current address Hewlett Packard, Information Network Division, 19420 Homestead Rd., Cupertino, California 95014

ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of several members of the Biomedical Computer Laboratory staff: specifically, L.J. Thomas, Jr., and G.J. Blaine for their contributions to the design and development of this system, as well as J. Markham for her assistance with the development of the band-detection algorithm. We also appreciate the assistance of A.W. Toga, of the Laboratory of Neuro Imaging, in the preparation of photographs of the displayed images. Finally, J. Goodloe provided expert photographic assistance, and gel photographs were contributed by C. Helms and M. MacCollin. K. Webb's secretarial assistance was indispensable in meeting the deadline for submission of the manuscript. This work was supported in part by National Institutes of Health Grants RR01380 from the Division of Research Resources and GM28232 from the Institute of General Medical Science.

REFERENCES

1. Bossinger, J., Miller, M.J., Kiem-Phong, V., Geiduschek, E.P., and Xuong, N.-H. (1979) *J. Biol. Chem.* 254, 7986-7998.
2. Garrels, J.I. (1980) *Trends Biochem. Sci.* 5, 281-283.
3. Aycock, B.F., Weil, D.E., Sinicropi, D.V., and McIlwain, D.L. (1981) *Computers Biomed. Res.* 14, 314-326.
4. Anderson, N.L., Taylor, J., Scandora, A.E., Coulter, B.P., and Anderson, N.G. (1981) *Clin. Chem.* 27, 1807-1820.
5. Lemkin, P.F., Lipkin, L.E., and Lester, E.P. (1982) *Clin. Chem.* 28, 840-849.
6. Kernighan, B.W., and Plauger, P.J. (1976) *Software Tools*, Addison-Wesley, Reading, MA.
7. Gingeras, T.R., Rice, P., and Roberts, R.J. (1982) *Nucleic Acids Res.* 10, 103-114.
8. Lautenberger, J.A. (1982) *Nucleic Acids Res.* 10, 27-30.
9. Cole, B.R., Shapiro, M.B., and Rodbard, D. (1978) in *Electrophoresis '78*, Catsimpoolas, N. Ed., pp. 79-92, Elsevier/North-Holland, New York.