An Automated Film Reader for DNA Sequencing Based on Homomorphic Deconvolution

Jeffrey T. Ives, Raymond F. Gesteland, and Thomas G. Stockham, Jr., *Fellow, IEEE*

***Abstract*—An automated reader for electrophoresis based DNA sequencing methods is described that provides fast and accurate sequence determination. Digitized sequencing lanes are processed with homomorphic blind deconvolution in preparation for peak detection, interlane alignment, peak refinement and base calling. Initial reads from direct blot sequencing films have error rates of about 1% at the rate of 5 nucleotides/s. Typical read lengths are 500–600 nucleotides. The described reader is a significant improvement over existing readers and could be an essential component in the sequencing efforts of the Human Genome Project.**

I. Introduction

THE HUMAN genome project is an international effort to map and sequence the entire genetic composition of humans. as well as several model organisms [1]. Successful completion of this task should provide a large database of biomedical information relating to diseases, inherited traits, genetic expression, genome organization, and evolution. However, the size of the genome and the capabilities of conventional methods make substantial technological improvements necessary to achieve this goal in the next 10–20 years. The human genome contains approximately 100 000 genes coded within approximately 3 billion deoxyribonucleic acid (DNA) base pairs. Under ideal, error-free conditions with existing technology, sequencing 3 billion base pairs would require over 10 000 man-years and would cost approximately $1 per nucleotide. Although novel, potentially high speed methods to sequence DNA have recently been proposed [2]–[5], the most practical and successful approaches for the foreseeable future use automation, improved technology, or process multiplexing to increase the throughput of the standard steps in conventional sequencing [6]–[9]. The most common conventional sequencing procedure enzymatically creates complementary strands from a "master" DNA strand, often called the template [10]. All complementary strands begin at the same nucleotide sequence on the template strands and add nucleotides sequentially. Random addition of sequence-terminating nucleotides, that are nearly identical to the four standard nucleotides in every other respect, create complementary strands that are fractions of the length of the complete template. Doing the enzymatic reaction in four different vessels with a specific terminating nucleotide per vessel creates complementary strands with all the lengths that terminate in that nucleotide. The resulting strand sets are separated by size with polyacrylamide gel electrophoresis to create fragment positions that look like a four rung wide ladder (Fig. 1) with all the fragments in any given lane terminating in the same nucleotide. Reading the DNA sequence is a conceptually simple ordering of size with reference to the corresponding lane. Beginning at any fragment position, often called a band, the next band in any of the four lanes is only one nucleotide longer or smaller depending on the direction. The nucleotide strand is read by stepping from band to band and noting the order of the lanes where the bands are located. The short sequence shown in Fig. 1 is GCTCACTCATTAGGCA.

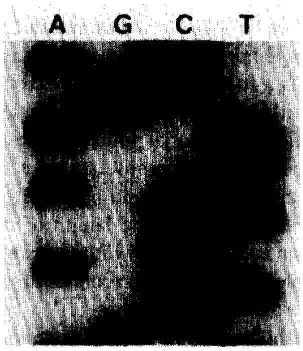


Fig. 1. DNA sequencing film illustrating four lanes and bands. Reading from the bottom to the top, the DNA sequence is GCTCACTCATTAGGCA.

Conventional sequence images are obtained by labelling the DNA fragments with radioisotopes and exposing 14″ × 17″ X-ray films with similar sized polyacrylamide gels. With approximately 300 nucleotides of readable sequence per lane set, where a lane set is the collection of A, G, C, and T terminating fragments from the same template, and 12–24 lanes sets per gel, about 5000 bands can be read per film. The entire human genome will therefore require reading on the order of one million films. The number of films quickly rises with redundancy and other practical considerations. For an expert human reader reading a film in around two hours, two million man-hours becomes overwhelming. The previously mentioned improvements to conventional methods generally increase the rate of separating fragments or detecting their relative positions, but the separated fragments must still be read quickly and with high accuracy. High accuracy reads

Manuscript received November 3, 1992; revised February 24, 1994. This work was supported by the Department of Energy under Grant DEFG02-88ER60700.

J. T. Ives was with the University of Utah, Department of Human Genetics. He is now with Molecular Tool, Incorporated, Alpha Center, Hopkins Bayview Research Campus, Baltimore, MD 21224 USA.

R. F. Gesteland is with Howard Hughes Medical Institute, Salt Lake City, UT 84112 USA.

T. G. Stockham, Jr. is with the University of Utah, Department of Human Genetics, Salt Lake City, UT 84112 USA.

IEEE Log Number 9401294.

0018–9294/94$04.00 © 1994 IEEE

are critical, particularly in regions of DNA that code for genes where single nucleotide changes can have significant functional effects. The importance of individual nucleotides for biology and medicine is a principal motivation of sequencing the human genome.

Although the task of converting relative position into nucleotide sequence is conceptually simple, the 1–3% error rate of human readers indicates that reading is more complex in practice. Band amplitudes and positions vary due to enzyme behavior and other biological factors and due to instrumentation and handling factors like uneven temperature distributions. Band positions as a function of fragment size typically follow either quasilogarithmic or constant spacing rules depending on the instrumentation, but spatial jitter and position anomalies can be large enough to superimpose adjacent band fragments. Interlane band amplitudes vary, and intralane band amplitudes change both locally and along the length of the read. Across a given film, bands change in width, as well as appearing tilted or in complex shapes. Backgrounds can also vary due to nonspecific adsorption and scratches.

The first published report of a complete automated sequence reader was by Elder, Green, and Southern [11]. They constructed a digitizing scanner and wrote software to read the scanned films. Pattern recognition was performed to identify the bands and lanes, and the combination of band heights and positions determined the base calls. Development of this system continued at Bio-Rad Laboratories, Inc. [12]. Commercial software packages designed to scan and read sequencing films have been developed by Bio-Rad, Milligen/Bioimage, Intelligenetics, pdi, Genomic and US Biochemicals. Pattern recognition is probably the most common band identification method in these packages although exact details are not published. The reading algorithms of the automated, fluorescence-based sequencing instruments (ABI, Pharmacia, Milligen, Hitachi, and LiCor) have also not been published in depth. The commercial readers are comparable to or only slightly faster than human readers and generally require regular parameter tuning for long and accurate reads. The apparent lack of speed may be due to the complex processing required for pattern recognition. To improve reading accuracy and possibly extend the number of resolved bands, several groups have concentrated on better band resolution. Image reconstruction using the maximum entropy method has been investigated by Elder [13] and Xu, Tso, and Martin [14]. These methods may be restricted to fairly small regions of DNA because they require an accurate description of the point spread function. The blurring function usually varies along the length of a sequencing lane, as well as between different lanes or films. Sanders and coworkers analyzed the images of fluorescently labelled bands in a commercial automated sequencer and corrected distortions in band shape [15]. The resulting traces down the lane centers indicate bands and spacings more precisely. In addition to resolution improvements, heuristic rules for interpreting lane traces have been programmed by Overbeek [16] and Ehrhardt, Englisch, and Neuhoff [17]. Both groups have emphasized that successful rule-based band calling is very dependent on the prior steps of lane tracking, distortion correction and band resolution.

In this article, we describe an automated reader based on homomorphic blind deconvolution [18]. Homomorphic blind deconvolution matches the problem of DNA sequence reading very well because precise knowledge of the blurring signal is not required, the speed of linear digital signal processing is utilized and interlane band amplitudes are automatically normalized. After deconvolution, the following steps; peak detection, interlane alignment, peak refinement and base calling, result in a called sequence. Due to the high resolution and signal to noise ratio of the deconvolved lane traces, the succeeding steps are relatively simple and provide very accurate sequence reads. Although the automated read of an example sequence is discussed, this article is intended to serve as a technique description establishing the fundamentals of our reader. A future article will concentrate on the results of applying the automated reader to sequencing films.

II. Theory and Application of Homomorphic Blind Deconvolution

Signal processing is often presented in terms of linear systems that must therefore satisfy the property of superposition. Superposition requires that the transformation of added input functions or input functions multiplied by a scalar is equal to adding the transformation of the individual inputs or multiplying the transformed input by the same scalar. However, many common systems process signals that are more than the addition of input vectors or scalar products. Amplitude-modulated transmission where the input is a product of two signal functions is one example. Homomorphic signal processing applies a generalized superposition principle that considers operations like vector multiplication and convolution, as well as vector addition and scalar multiplication. Homomorphic systems will often convert nonlinear inputs to signals suitable for linear processing. For example, multiplicative homomorphic systems compute the logarithm of two multiplied signals, resulting in added signals that can be linearly filtered. The reader is referred to the references by Oppenheim and coworkers for more thorough discussions of homomorphic deconvolution [18]–[20].

The work described in this paper utilizes blind deconvolution in addition to homomorphic processing. In contrast to the more common scenario where the impulse response of the system is known and precisely characterized, blind deconvolution refers to the process of separating two convolved signals when both are unknown, at least unknown beyond some general description. Blind deconvolution is designed for systems like sequencing lanes where neither the signal, i.e., the band positions, or the convolver, i.e., the blurring function, are well characterized and consistent over time and space. Digital restoration of acoustic recordings is an example of successful blind deconvolution [21].

For the present topic of DNA sequencing bands, the image of sequenced DNA bands can be represented as the convolution of a blurring function with varying amplitude pulses. The pulses occur at the central positions of the bands. The blurring function is caused by the emission pattern of the isotope labels and the diffusion width of the bands. Additive noise is also distributed throughout the image. A point source, or

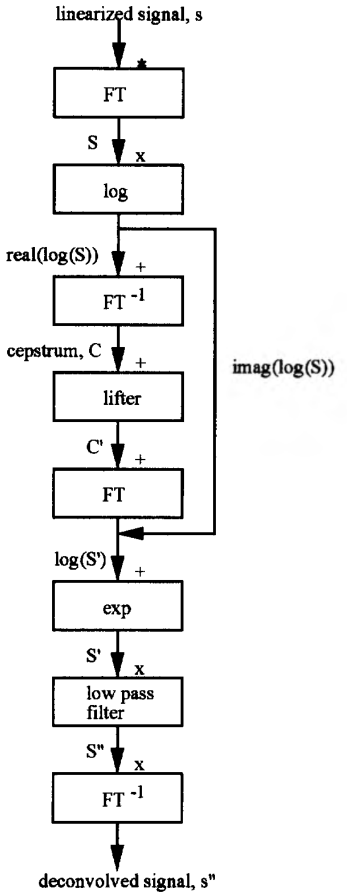


Fig. 2. Processing steps in homomorphic deconvolution.

Lorentzian, pattern of emission was assumed for analysis and provides a reasonably accurate match to the actual band shapes (the term, blurring function rather than point spread function, will be used in this article because the spatial distribution of a sequencing band is not necessarily due to a point source). The process of deconvolution, i.e., separating the peaks from the system blur, is a linear process that can be performed by inverse linear filtering, but the spatial variance of the blurring function makes inverse filtering or Wiener filtering prone to error. Deconvolution can more practically be achieved by homomorphic blind deconvolution.

The processing steps are shown in Fig. 2. Each sequencing lane is a convolution of the peaks and the blurring function, so the computed spectrum of each sequencing lane is the product of the Fourier transforms of each individual signal (convolution and multiplication ,are Fourier transform pairs). The simplest approach at inverse filtering would be applied at this point by dividing by the Fourier transform of the blurring function. Noise and uncertainties in the blurring function make this approach likely to fail. Homomorphic processing begins by computing the complex logarithm of the spectrum, thereby converting the product of Fourier transforms into a sum of log-spectra. As noted earlier in discussions on superposition, addition allows linear filtering to be applied to remove or minimize the blur component. Again, the precise characteristics of the blurring function are not well known, but the blind deconvolution method used in this work relies on the general difference in shape between the log-spectra of the signal and the blurring process. The log-spectrum of a Lorentzian point spread function is a straight line with negative slope. Different width bands simply adjust the slope of the line, and most of the energy in straight lines remains contained in the low frequencies near zero, unlike the widely scattered frequency distribution of the peaks. Therefore, applying a generalized high-pass linear filter should significantly attenuate the blurring function, have only a slight affect on the widely distributed peak frequencies and operate successfully in spite of spatially variant blurring functions. The ability to apply the same high-pass filter regardless of the absolute band size makes this process blind deconvolution.

This process is relatively straightforward, although some nomenclature has been introduced to reduce the confusion of computing multiple Fourier transforms. Computing the inverse Fourier transform of the real part of the log-spectrum results in a cepstrum (the word cepstrum is a variant of spectrum to recognize that a second Fourier transform has been computed [22]). The cepstrum of the blurring function is predominant in the low quefrencies (variant of frequency) as discussed above, and the band position cepstrum is distributed over the full bandwidth. A high-pass lifter (variant of filter) then multiplies the cepstrum to reduce the low quefrencies, and the 0 quefrency amplitude of the lifter also normalizes the energy across all deconvolved sequencing lanes. Reversing the Fourier transform-log-Fourier transform processing path creates a lane with sharpened bands at the band centers, but high frequency noise causes significant degradation. Therefore, a low pass filter is inserted as shown in Fig. 2.

Two additional comments about the processing shown in Fig. 2 are probably helpful. Subtracting the straight line component of the log-spectrum that is due to the blurring function is one method of deconvolution and would result in sharpened bands. However, accurately estimating the component to be subtracted was difficult and not generally useful. The second comment addresses the imaginary component of the log-spectrum that is routed around the liftering process. The Lorentzian model of the blurring function is real and even, as are most other physically reasonable profiles. Therefore, only the real parts of the spectrum and log-spectrum of the blur are nonzero, and the imaginary component of the log-spectrum is entirely due to the pulses representing the DNA band centers. Liftering this component would serve no useful purpose and would double the processing time for these steps.

III. METHODS

The following sections expand on the series of processes shown in Fig. 3 for automated DNA sequence reading.

A. DNA Sequencing and Film Digitization

Sequences of M13mp19 were prepared following standard protocols for dideoxynucleotide terminators, SequenaseII DNA polymerase and manganese buffer [23]. DNA was labelled with ^{32}P isotopes. A 60-cm long, 18-cm wide, 60-μm thick, 4% polyacrylamide gel (34:1 acrylamide:bis acrylamide ratio) was used for electrophoresis at approximately 88 volts/cm (60–62 W). A sharkstooth comb with about 3 mm spacing between teeth defined the lanes. DNA bands were deposited on a nylon membrane (Biodyne B, Pall) using a locally made direct blotting apparatus based on the original work of Pohl and Beck [24]. Direct blotting refers to the transfer of DNA from the bottom of the electrophoresis gel to a moving membrane. The process is similar to collecting fractions in other separation methods. Electrophoresis and blotting proceeded for approximately five hours with a membrane speed of 3.0 mm/min. Films were obtained by exposing X-ray film (Kodak XAR) to the membranes for about 12 hours.

Films were scanned using a Truvel TZ-3X, 8-bit densitometer (Vidar). Scanner resolution was approximately 300 dots, or samples, per inch. Digitized files were analyzed on a Decstation 3100 workstation. Programs were written using the matrix analysis package, MATLAB (The MathWorks, Inc.). For the reading example in this article, a single lane set containing 566 bands and 5844 samples per lane was scanned and analyzed. The smallest and largest DNA fragments in the scanned lane set were 122 and 688 nucleotides long, respectively. Smaller fragments were not scanned due to problems during direct blotting not associated with the fragment size. Larger fragments were not included because film quality and reading accuracy were quickly decreasing.

B. Image Conversion and Signal Linearization

Each lane set of the sequencing film images was reduced to four one-dimensional vectors corresponding to the average of a 15 sample wide line through the center of each lane (Fig. 4(a)). Lane tracking and vectorization were performed by a MATLAB program that sampled every twentieth line across the four lanes (perpendicular to the direction of band migration), smoothed the sampled lines by a low pass filter, and identified minima indicating band centers. Minima were then organized with other minima having progressively similar lateral positions. A least squares fit through associated minima established a line through the band centers of each lane in the direction of electrophoretic migration. This lane tracking approach was very preliminary and is probably inadequate for routine sequencing. More advanced lane tracking and distortion correction [15] will probably be an important complement to accurate, general purpose reading. This project to date has concentrated more on the steps following lane tracking. Complete lane traces along an entire film are 5000–7000 samples long. For processing speed and local accuracy, subsequent processing operated on overlapping, 2048 sample subdivisions of the full lane traces. Future versions of the reader will likely include automated segmentation to achieve the same goals.

Two compensation steps were performed on the four vectors. The values read out from the scanner had been adjusted by the manufacturer for improved viewing on the nonlinear displays of standard monitors. The first compensation step corrected for this adjustment, and was based on the scanner circuitry. The compensated values are then proportional to the percent transmission through the film. Second, the transmission values were converted to numbers proportional to the emission strength, or intensity, of the isotopes in the bands (Fig. 4(b)). Different emission strengths are linearly related and therefore well suited for linear systems analysis. Converting the transmitted values requires compensation for the nonlinear exposure-density relationship of film. Currently film exposures are assumed to be in the approximately linear region of the D-log (E) film curve where the film density, D, is linearly related to the log of the exposure, E. Exposure is the product of intensity and time. The slope, also called the film gamma, is assumed to be 2.2 ($D = 2.2\log(E)$). The transmitted values are related to the film density by Beer's Law ($D = -\log(I_t/I_0)$ where I_t is the light transmitted and I_0 is the light transmitted without the film. The log is base 10.) Combining these factors states that the emission strength of the isotope, I_i, is inversely proportional to the transmitted value raised to the 1/2.2 power ($I_i = k/(I_t)^{1/2.2}$ where k is a proportionality constant). Continuing work is generalizing the linearization to a larger region of the film curve. The linearization process may not be necessary for detection schemes that do not use film such as direct beta detection [25], [26] or optical detection using fluorescence or chemiluminescence [27], [28].

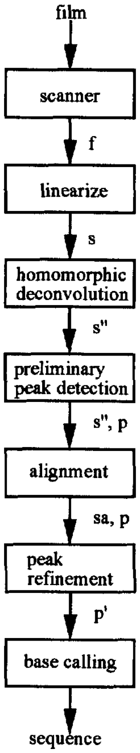


Fig. 3. Processing steps incomplete reader.

C. Homomorphic Blind Deconvolution and Low Pass Filter

Each lane proceeds independently through the operations shown in Fig. 2. Example plots after some of these steps on a

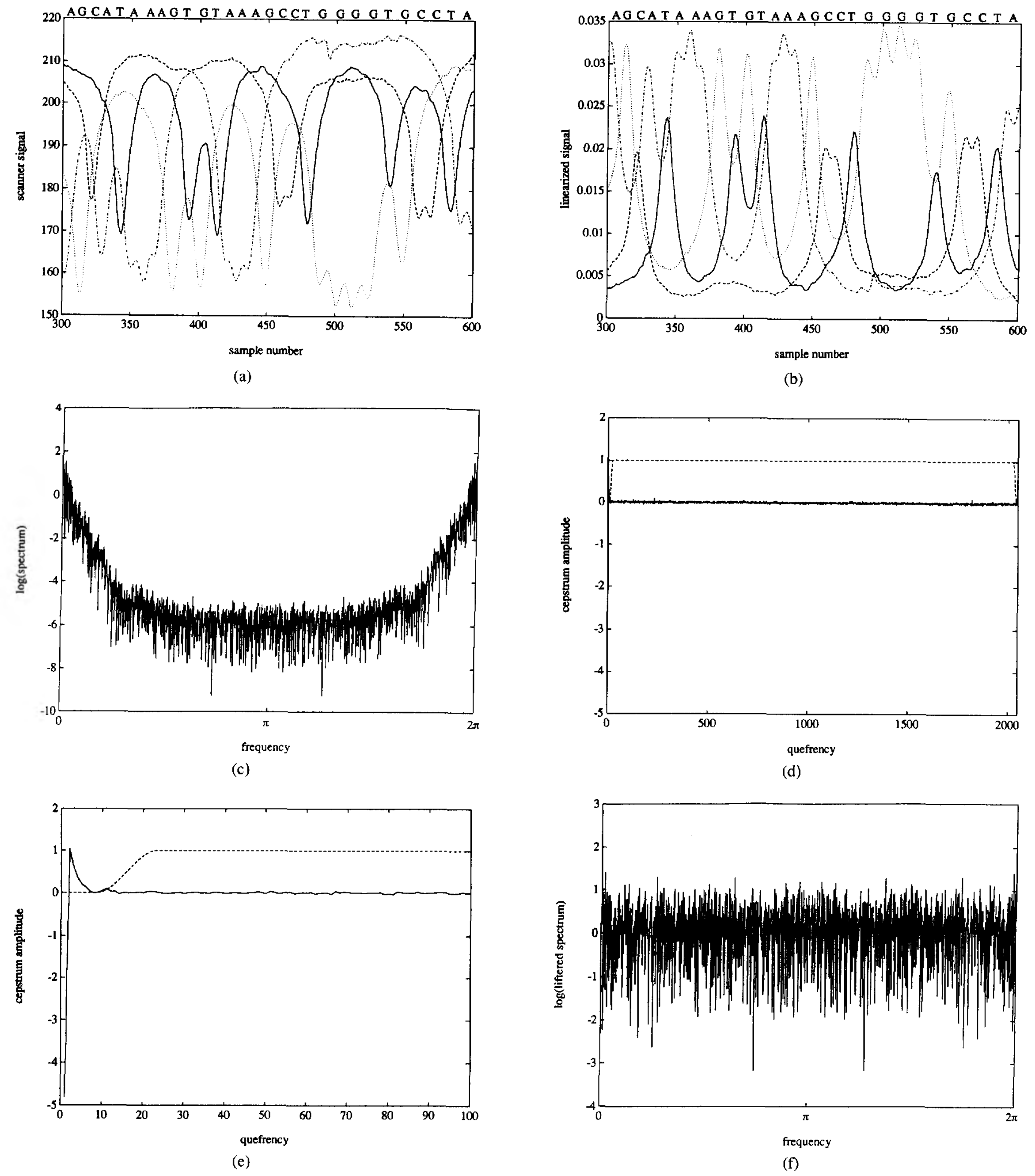


Fig. 4. Example plots of data prior to and through homomorphic blind deconvolution. All plots are from the first 2048 sample subdivision of the entire scanned lane set (5844 samples). Sample numbers on the abscissa of (a), (b), and (g) refer to the same 300 sample segment within the 2048 samples. Also shown in (a), (b), and (g) is the automated read of the plotted data. The spikes at the approximate quefrencies of 300 and 1800 in (d) are unique to this sequencing lane and are probably due to an anomalous noise pattern, rather than some common periodicity in the signals. (a) scanned sequencing lanes, (b) linearized data, (c) log (spectrum) of A lane, (d) cepstrum (solid line) for A lane and lifter (dashed line), (e) higher resolution view of cepstrum and lifter near origin, (f) log (liftered spectrum) of A lane, (g) deconvolved and aligned lanes. The line patterns are: T (solid), C (dashed), G (dotted), and A (dash-dot). All successive figures follow the same line pattern designations.

2048 sample section of a DNA lane are shown in Fig. 4. The log-spectrum of a Lorentzian is a straight line with negative slope, and this feature will be evident in later figures from actual data. The linear envelope of the log-spectrum at low frequencies (Fig. 4(c)) is consistent with the assumption of a Lorentzian blurring function. Examination of the cepstrum around the origin (Fig. 4(d) and shown in more detail in Fig. 4(e)) reveals a low quefrency curve that is primarily

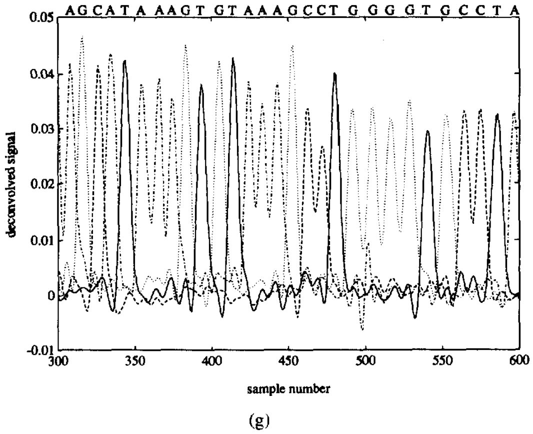


Fig. 4. (Continued)

caused by the blurring function. The noise and band position signal are distributed throughout the quefrency window. The high pass lifter simultaneously deamplifies the low quefrency blurring curve and, with zero amplitude at the origin, normalizes the signal in each lane. The lifter typically follows a raised cosine shape (shape factor = 0.5) around the origin and has an amplitude of 0.5 at the fifteenth sample (Fig. 4(e), dashed line). After liftering, the log-spectrum (Fig. 4(f)) is essentially flat indicating that the blurring function has been removed to a large degree. Reversing the steps as shown in Fig. 2 results in a deconvolved signal.

The deconvolved signal generally contains too much high frequency noise for accurate peak detection. Consequently, a Gaussian low pass filter is applied to the liftered spectrum, S' (Fig. 2). The optimal bandwidth of the low pass filter depends on the noise and the desired resolution between adjacent bands. For the peak detection and base calling used in this work, the highest accuracy reads were obtained when the filter bandwidth was chosen to obtain relatively broad sequencing peaks (Fig. 4(g)). When the four sequencing lanes are superimposed, filtered peaks are broad enough to present a near continuous envelope that moves from peak to peak and obscures the noisy baselines of the lanes without a local peak. For consistent sequencing conditions, the filter bandwidth is constant between films. The filter bandwidth we used was about 100 frequency samples (where π is 1024 samples). The improved resolution after deconvolution and filtering is demonstrated by comparing Fig. 4(b) and (g). Methods to automatically adjust the filter bandwidth in response to varying sequence and film conditions should be incorporated in future improvements. Alternative approaches to converting deconvolved bands into DNA sequence may tune the filter bandwidth to emphasize other features such as narrow peak resolution.

D. Preliminary Peak Detection

Due to the clarity of the resolved peaks, peak detection is relatively simple and fast. Peaks are detected by simply identifying samples that have larger amplitudes than the immediate neighbors and larger than a varying threshold. The threshold is based on a least squares exponential fit to the prominent peaks within each lane.

E. Interlane Alignment

The lanes in a given set are not necessarily in correct relative registry due to several factors such as lanes misaligned relative to the axes of the scanner, temperature variations across the gel and gel inhomogeneities. Scanned misalignments can be so severe that the bands are not in the correct order, much less evenly spaced. Since accurate reading is very dependent on the relative order of the bands, correct registry of the lanes is important.

The alignment process generally relies on the relative distance between neighboring peaks for every lane-lane combination, and then shifts the lanes by the number of samples necessary to produce peaks separated by equal relative distances. As a simple example, if the A-T separation (peaks in the A lane immediately followed by T lane peaks) is N samples and the T-A separation is $3N$, then advancing the T peaks by N samples will produce a regular periodic spacing of $2N$ samples between all A and T peaks. Expanding on this process, all possible lane-lane combinations of average interpeak distances can be organized in a 4×4 matrix with the rows and columns ordered as T, C, G, and A lane peaks. The row–column combination of every matrix element is a unique lane–lane pair. The difference between the measured interpeak distances and the ideal interpeak distance of aligned lanes is the alignment shift. Sufficiently accurate estimates of the ideal peak–peak separation can be made from the total number of samples divided by the number of detected peaks or by considering the sum of lane-lane pairs located in transposed elements of the matrix, i.e., A-T and T-A.

In actual practice, the alignment process follows a two step, coarse then fine, alignment process. To compensate for incorrect relative order of the peaks, the first step in the alignment process coarsely aligns the four lanes. The coarse alignment is basically an iterative version of the shifting process. The spacing matrix is generated, and an expected spacing is determined by dividing 2048 samples by the total number of preliminary peaks. The T lane is used as the reference lane, and the spacing matrix elements for T-C, T-G, and T-A are compared to the expected spacing. Spacing values that deviate significantly from the expected spacing are noted, and the lane with the maximum deviation is shifted relative to the other lanes. Significant deviation is an absolute difference between the actual and expected spacing that is greater than one-fourth of the expected spacing. Shifts begin with a delay of one-half the expected spacing and incrementally advance the shift by one-quarter of an expected spacing. For each shift position, a new spacing matrix is generated. Shifting stops for an individual lane when the spacing is acceptably close to the expected spacing or the absolute error is larger for another lane. The coarse alignment is entirely finished when the spacing of all lanes is sufficiently small. Larger shifts and finer resolution are of course possible. With carefully prepared sequence lanes, coarse alignment is usually unnecessary.

The second, more precise alignment step again considers the T lane as the reference lane and fits a straight line through the separation distances as a function of peak position for each lane pair; T-C, T-G, and T-A. The lines are biased by the expected spacing. The intercept is used for the alignment shift at one end of the sequence, and the slope difference between lanes is used to add interpolated samples and compensate for a varying shift along the sequence. The straight line assumption is probably reasonable for the nearly constant band spacing of direct blotting, but alternative models may be more accurate, particularly with nondirect blot methods.

F. Peak Refinement

Generally, a small percentage of the peaks that were detected earlier along each lane are incorrect. Excess peaks that occur are primarily due to noise, anomalous background like scratches and oscillations caused by deconvolution. Most incorrect peaks are low in amplitude, particularly compared to the correct peak in one of the other lanes. Superimposing the normalized and aligned lanes and selecting only those peaks that are maximum in amplitude eliminates most of the erroneous peaks. An average interpeak spacing is then determined from the total number of samples divided by the number of maximal peaks.

The error rate based solely on the maximal peaks is usually too large; 2–10% depending on the sequence quality. In addition to excess peaks, unresolved peaks occur where bands are highly blurred and/or severely overlapped and appear as broad plateaus or shoulders even after deconvolution. These undetected peaks are often resolved by considering interpeak distances. If the interpeak distance is greater than 1.7 times the average peak separation, an intermediate peak is added. Some excess, but maximal, peaks are also removed by the interpeak distance. If the separation between adjacent peaks in the same lane is less than one-half the average distance, only the peak with larger amplitude is used. The scalars, 1.7 and 0.5, are empirical and may not be optimal.

G. Base Calling

High quality deconvolution, alignment and peak assignments make determining the nucleotide sequence straightforward. The sequence of DNA bases is simply the correspondence between peak order and the lane associated with each peak.

IV. Results and Discussion

Fig. 5 shows a full length view of the sequencing film read by the automated reader. As shown in Fig. 6, the first error occurred after reading 393 nucleotides and, coincident with decreasing sequence quality, six more errors occurred before finishing the read at 566 nucleotides. Automated reads with such low error rates for sequences 500–600 nucleotides long have not previously been reported.

Six of the seven errors are deletions, and an example of a deleted C is shown on the right side of Fig. 7 near sample number 4960. The deletion is observed as a slight shoulder followed by two resolved C peaks. The separation between the deleted C band and the next C band is reduced relative to the typical band spacing. Therefore, the C band is not well resolved and merges with the neighboring peak creating a shoulder. The isolated C band in Fig. 7 is an extreme example of reduced band separation where three adjacent bands have been merged into a single broad and tall peak. The reduced band separation in regions of DNA that have a large percentage of G and C nucleotides is a commonly recognized sequencing artifact. This phenomenon is often called band compression. Compressions have been reported to be the predominant error in some large sequencing projects [29], and are also the major problem in the automated read discussed here; five of the six deletion errors are deleted C bands. The single insertion error, an extra A band, is due to the narrow second peak observed near sample number 4960

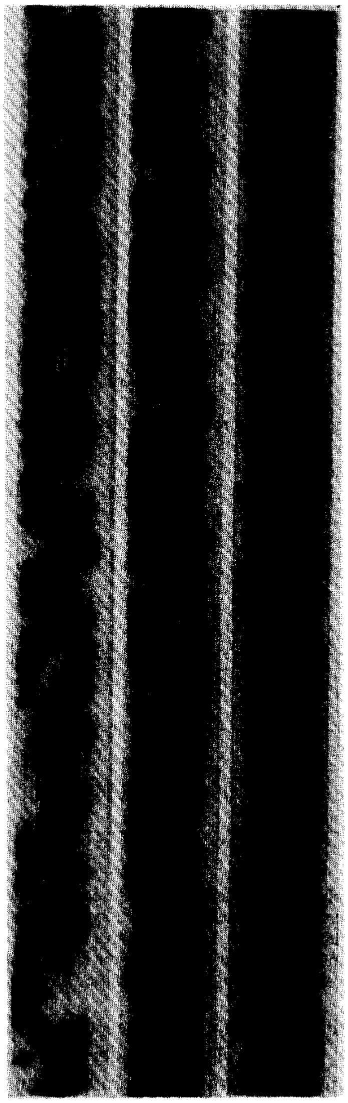


Fig. 5. Full length image of the sequencing film as initially presented to the digitizer. The film is approximately 60 cm long, and has been manipulated to appear as three parallel sections of about 20 cm each. The first section is shown on the left with the smallest fragments at the top, and successive sections are raised and shifted to the right. The lane sets analyzed in Fig. 4 were obtained from the left, small fragment section.

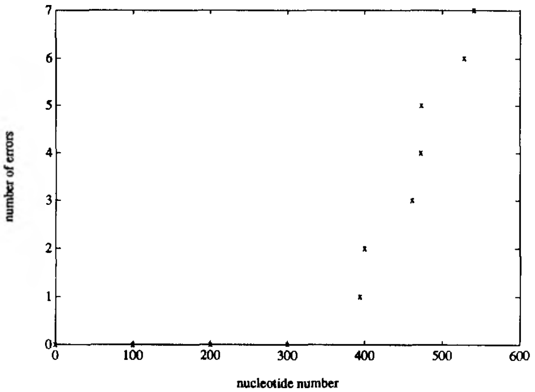


Fig. 6. Location of reading errors and cumulative error count.

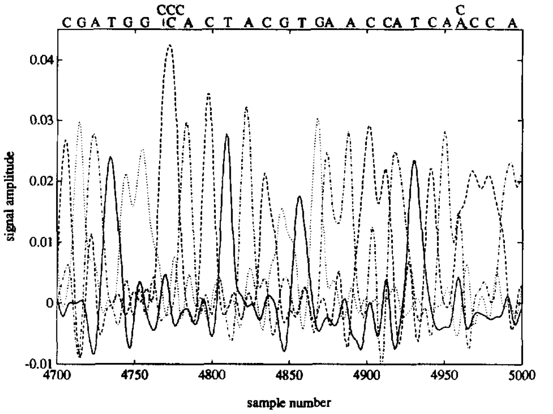
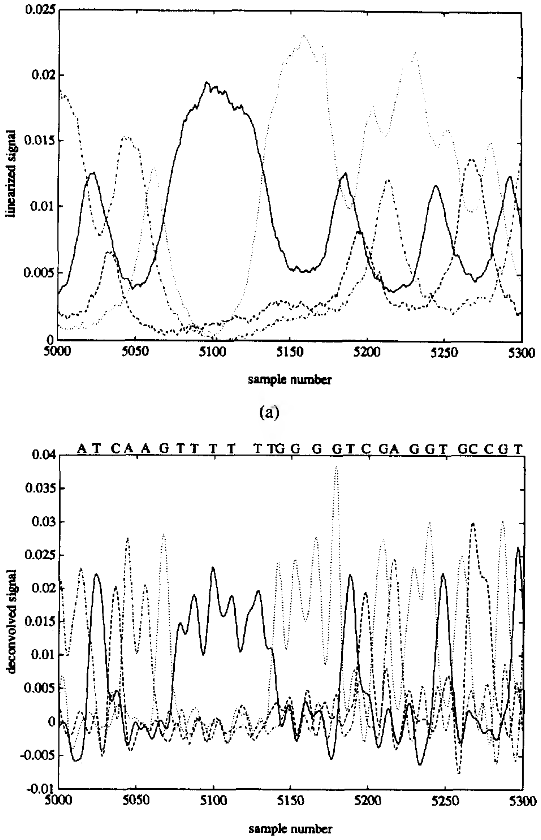


Fig. 7. Deconvolved lanes around example errors. The nucleotide sequence determined by our reader is listed across the top of the plot. The two errors, C at about sample 4770 and A at about sample 4960, are identified by the correct base assignment(s) listed above the error. Sample numbers shown on the abscissa refer to the position within the entire 5844 sample scan.

(Fig. 7). Close examination indicates that a secondary peak is present on the film. Therefore, the error is primarily a sequence or film preparation error, rather than a mistake by the automated reader.

An increasing error rate as DNA fragment size increases is common for both manual and automated sequencing instruments [26]. The fragment-dependent change in band resolution and signal strength is apparent in the film image shown in Fig. 5. The series of processing plots shown earlier in Fig. 4 were obtained from the relatively high resolution section of Fig. 5 containing DNA fragments about 200 nucleotides long. Fig. 8(a) plots the linearized traces from larger, approximately 600 nucleotide long DNA fragments near the end of Fig. 5. Comparison between the linearized traces of smaller (Fig. 4(b)) and larger fragments (Fig. 8(a)) illustrates the substantial decrease in resolution with larger fragments. This loss of resolution is partially due to broader, i.e., spatially variant, band shapes along a lane set. The signal amplitudes in Fig. 8(a) also indicate reduced emission intensities and decreased signal-to-noise ratios associated with larger fragments. These changes are reflected in the deconvolved signals which are not as well resolved or as large in amplitude with increased fragment size (Figs. 4(f) and 8(b)). It is interesting to note that the cepstra for these two fragment sizes are very similar (Fig. 9) in spite of the substantial change in width of the blurring function. This feature is convenient for blind deconvolution with a single generalized lifter, and also indicates that the decreased resolution and increased error rate associated with reading larger fragments is caused more by the loss in signal-to-noise ratio, than by band broadening.



(a)

(b)

Fig. 8. Data plots from a sequence region with larger DNA fragments than in Fig. 4: (a) linearized traces, (b) deconvolved and aligned traces. These curves are comparable to Fig. 4(b) and (g).

The change in band resolution with fragment size is particularly significant when our reading process relies on peak detection. In alternative versions of our reader, we have considered additional parameters such as the width of each band or band series and peak resolution. Generally, the overall error rates were not improved by the additional parameters. While specific errors might be resolved, new errors would often balance out the improvements. Although our efforts with alternative parameters were not exhaustive, the computation aspect of our general approach to improved sequence accuracy is concentrating on more fundamental improvements in the processing, like peak resolution.

The automated read in this article has an error rate of 1.06% out to almost 600 bases. For about the first 400 bases, no errors were made, and the high resolution of the smaller

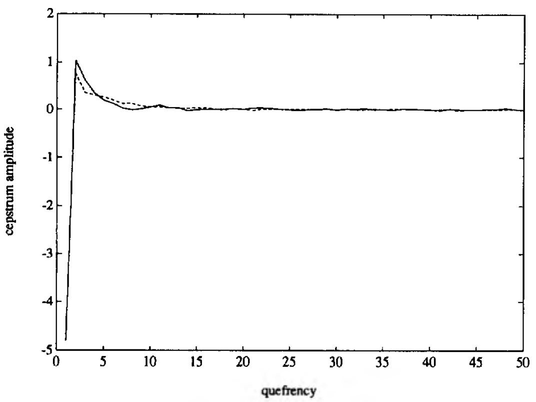


Fig. 9. Cepstra near the origin for smaller (solid line, refer to Fig. 4) and larger DNA fragments (dashed line, refer to Fig. 8).

fragments indicates that accurate reads could be pushed toward smaller fragments another 50–100 nucleotides. Although a direct comparison with the same film or DNA sequence has not been done, these numbers are significantly better than reports from fluorescence-based sequencing instruments [9], [26] or commercial film-based readers, particularly beyond about 300 nucleotides. Furthermore, the error rates reported by other publications are usually after some errors have been eliminated by the 7–10 fold redundancy of standard random sequencing strategies and some manual editing. Even an experienced human reader will have an error rate of 1–3%. A recent sequencing project at the University of Utah used manual reading for about 10 kb (kilobases) and had an error rate of 2–3%. About two-thirds of the errors were data entry errors like mistyped keys, errors that an instrument would not make, and reading was slower than the five bases/s of our prototype reader.

Perhaps the best plan for error reduction is an active collaboration between sequence preparation and instrument development. This collaboration will ultimately lead to a complete, automated and reproducible system. Rather than developing more complex automated readers, some improvements in reading will be more reliable and fundamentally sound when done at the initial level of sequence preparation and film exposure. In addition to band resolution and signal-to-noise factors, errors associated with large DNA fragments are often caused by distortions in band shape that become more prevalent around large fragments with long electrophoresis times. Band compressions can be reduced and band amplitudes made more uniform by adjusting the buffer composition, electrophoresis temperature and incorporating modified nucleotides [30], [20]. Improved repeatibility through automation should also improve the reader's accuracy by providing more consistent spacing, shape, amplitude and film background. In the development stage of improved sequence conditions, our reader can potentially record the quantitative effects of modified conditions much more easily and quickly than manual measurements.

Currently, our reader calls about five bases per second after an individual lane set in the scanned image has been reduced to four one-dimensional vectors. The total time to read a film is about 3.5 min to scan the 65-cm long direct blot films used in this work, about 2 min to convert the digitized image into one-dimensional vectors averaged along the center of each lane and then the read time. If each lane set were read out to 500 nucleotides, the total time for a direct blot film would be about 15 min. Correspondingly, a standard 14″ x 17″ film with 12 lane sets and 300 nucleotides per lane set is digitized in about 2.5 min, and would require about 30 min for a completed read. This time compares very favorably with 10 or more hours of electrophoresis, four to 12 hours of film exposure and two hours of human reading. With the relative speed of the reader, a single reader could support several sequence generation stations, and the number of supportable stations increases as the computing speed goes up. The speed is expected to increase substantially as the Matlab script files are converted to a compiled language like C or Fortran, programs are optimized and processing hardware continues to increase in speed.

The reader as it presently exists demonstrates the fundamental operations and handles individual lane sets. Future improvements include developing a convenient user interface and scaling up for multiple lane sets over an entire film. At that point, direct comparisons should be made between our reader and commercial or academic alternatives. An additional capability should also be added to the reader. As sequencing methods advance, they are often extending the read length with longer electrophoresis times. The resulting band patterns generally degrade in quality and error rates increase with read length. Manual post-read editing and/or preselected scan limits could restrict the reads to accurate sequence. However, an automated limiter based on error estimation within the reader would probably be preferred, or even required, when the reader has the potential to exceed human performance.

The reader performance cited in this article came from sequencing films that were exposed to direct blot membranes. Direct blotting results in band deposition in approximately even intervals. The same regular spacing is obtained by on-line detection methods with fluorescently labelled DNA, and the reader could be helpful in these instruments as well. The general procedures discussed in this article are also applicable to conventional, fixed time detection where the film is essentially a contact print directly from the electrophoresis gel. The band positions along the length of the gel are separated in a quasilogarithmic pattern with the smaller fragments at one end being relatively far apart and the separation between bands decreasing toward larger fragments. Band shapes in fixed time detection are more constant over the length of the sequence than bands in on-line detection. The reduced variance of the blurring function should allow higher resolution deconvolution.

V. Conclusion

The lack of an adequate method to read DNA sequences automatically has substantially hindered sequence acquisition, particularly large scale sequencing where fragments with 300–600 bases are merged into sequences of 10 000 or more bases. Sequencing these large DNA pieces is important because the majority of genes have such sizes. This article concentrates on the application of homomorphic blind deconvolution to DNA sequence ladders and the subsequent conversion of deconvolved waveforms into nucleotide sequence. Homomorphic blind deconvolution is a novel approach to the problem of reading sequencing films and provides several benefits in signal normalization, rapid processing and user-independent operation. Based on the high resolution and signal-to-noise ratio of the deconvolved data, subsequent processing steps can be done accurately and often more simply than with alternative methods. Initial results of speed and accuracy are very encouraging, and indicate that additional reductions in error rates depend more on improved sequence preparation than reader development. A more extensive reading project is in progress and will provide more significant measures of error rates and read lengths [31].

ACKNOWLEDGMENT

The authors thank Bob Weiss and coworkers for sequence preparation and general guidance in molecular biology, and Peter Cartwright, Paul Leo, and coworkers for their technical expertise and helpful discussions.

REFERENCES

[1] J. D. Watson, "The human genome project: past, present and future," *Science*, vol. 248, pp. 44–51, 1990.

[2] R. J. Driscoll, M. G. Youngquist, and J. D. Baldeschwieler, "Atomic scale imaging of DNA using scanning tunnelling microscope," *Nature*, vol. 346, pp. 294–296, 1990.

[3] G. R. Parr, M. C. Fitzgerald, and L. M. Smith, "Matrix assisted laser desorption/ionization mass spectrometry of synthetic oligodeoxyribonucleotides," submitted for publication.

[4] R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov, "Sequencing of megabase plus DNA by hybridization: Theory of the method," *Genomics*, vol. 4, 114–128, 1989.

[5] J. H. Jett, R. A. Keller, J. C. Martin, B. L. Marrone, R. K. Moyzis, R. L. Ratliff, N. K. Seitzinger, E. B. Shera, and C. C. Stewart, "High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules," *J. Biomol. Struct. Dynam.*, vol. 7, pp. 301–309, 1989.

[6] G. M. Church and S. Kieffer-Higgins, "Multiplex DNA sequencing," *Science*, vol. 240, pp. 185–188, 1988.

[7] H. Swerdlow and R. F. Gesteland, "Capillary gel electrophoresis for rapid, high resolution DNA sequencing," *Nucleic Acids Res.*, vol. 18, pp. 1415–1419, 1989.

[8] A. J. Kostichka, M. L. Marcbanks, R. L. Brumley, Jr., H. Drossman, and L. M. Smith, "High speed automated DNA sequencing in ultrathin slab gels," *Bio/Technology*, vol. 10, pp. 78–81, 1992.

[9] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, "Large scale and automated DNA sequence determination," *Science*, vol. 254, pp. 59–67, 1991.

[10] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci.*, vol. 74, pp. 5463–5467, 1977.

[11] J. K. Elder, D. K. Green, and E. M. Southern, "Automatic reading of DNA sequencing gel autoradiographs using a large format digital scanner," *Nucleic Acids Res.*, vol. 14, pp. 417–424, 1986.

[12] J. West, "Automated sequence reading and analysis," *Nucleic Acids Res.*, vol. 16, pp. 1847–1856, 1988.

[13] J. K. Elder, "Maximum entropy image reconstruction of DNA sequencing gel autoradiographs," *Electrophoresis*, vol. 11, pp. 440–444, 1990.

[14] D. Q. Xu, M. K-S. Tsao, and W. J. Martin, "Automatic interpretation of digital autoradiograph of DNA sequencing gels," in *Image Analysis and Processing II*. New York: Plenum, 1988, pp. 501–509.

[15] J. Z. Sanders, A. A. Petterson, P. J. Hughes, C. R. Connell, M. Raff, S. Menchen, L. E. Hood, and D. B. Teplow, "Imaging as a tool for improving length and accuracy of sequence analysis in automated fluorescence-based DNA sequencing," *Electrophoresis*, vol. 12, pp. 3–11, 1991.

[16] R. Overbeek, "Application of logic programming techniques to DNA sequence gel-reading," *Abstracts of Papers of the Amer. Chem. Soc.*, vol. 198, p. COMP10, 1989.

[17] W. Ehrhardt, U. Englisch, and V. Neuhoff, "Automatic evaluation of nucleic acid sequencing gel autoradiographs," *Electrophoresis*, vol. 10, pp. 265–266, 1989.

[18] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, 1968.

[19] A. V. Oppenheim and R. W. Schafer, "Homomorphic signal processing," in *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975, ch. 10, pp. 480–531.

[20] A. V. Oppenheim and R. W. Schafer, "Cepstrum analysis and homomorphic deconvolution," in *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989, ch. 12, pp. 768–834.

[21] T. G. Stockham, Jr., T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proc. IEEE*, vol. 63, pp. 678–692, 1975.

[22] B. Bogert, M. Healy, and J. Tukey, "The quefrency analysis of time series for echoes," in *Proc. Symp. on Time Series Analysis*, 1963, ch. 15, pp. 209–243.

[23] S. Tabor and C. C. Richardson, "Effect of manganese ions on the incorporation of dideoxynucleotides by bacteriophage T7 DNA polymerase and Escherichia coli DNA polymerase," *Proc. Natl. Acad. Sci.*, vol. 86, pp. 4076–4080, 1989.

[24] F. M. Pohl and S. Beck, "Direct transfer electrophoresis used for DNA sequencing," *Methods Enzymol.*, vol. 155, pp. 250–259, 1987.

[25] R. F. Johnston, S. C. Pickett, and D. L. Barker, "Autoradiography using storage phosphor technology," *Electrophoresis*, vol. 11, pp. 355–360, 1990.

[26] S. Burbeck, "Direct digital imaging of radio-labeled two-dimensional gel beta emissions using micro-channel plate image enhancement," *Electrophoresis*, vol. 4, pp. 127–133, 1983.

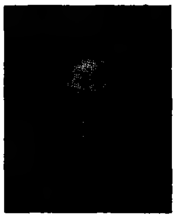
[27] Z. A. M. Boniszewski, J. S. Comley, B. Hughes, and C. A. Read, "The use of charge-coupled devices in the quantitative evaluation of images, on photographic film or membranes, obtained following electrophoretic separation of DNA fragments," *Electrophoresis*, vol. 11, pp. 432–440, 1990.

[28] A. Karger, J. T. Ives, R. B. Weiss, J. M. Harris, and R. F. Gesteland, "Imaging of fluorescent and chemiluminescent DNA hybrids using a 2-D CCD camera," in *Proc. SPIE*, 1990, pp. 78–89.

[29] J. Sulston, Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, S. Dear, A. Coulson, M. Craxton, R. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough, and R. Waterston, "The C. elegans genome sequencing project: A beginning," *Nature*, vol. 356, pp. 37–41, 1992.

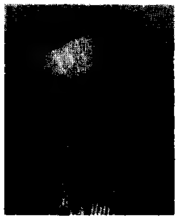
[30] S. Mizusawa, S. Nishimura, and F. Seela, "Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP," *Nucleic Acids Res.*, vol. 14, pp. 1319–1325, 1986.

[31] J. T. Ives, R. Weiss, T. G. Stockham, Jr. and R. F. Gesteland, submitted for publication.



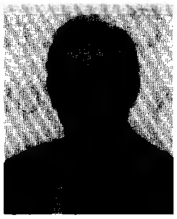
Jeffrey T. Ives received the Ph.D. degree in bioengineering from the University of Utah in 1990.

His thesis work investigated integrated optic polymer slab waveguides as biomedical surface sensors. From 1988–1992, he worked as a postdoctoral fellow in the Department of Human Genetics at the University of Utah where he was introduced to molecular biology. During this time, his research involved nonisotopic detection of DNA and automated base calling of DNA sequences. After one year in the Department of Chemistry at the University of Utah studying surface enhancement techniques, he moved to Molecular Tool, Inc. (Baltimore, MD), where he is currently developing improvements in optical detection, instrumentation and information processing for genetic identification.



Raymond F. Gesteland received the B.S. degree in chemistry (1960) and the M.S. degree in biochemistry (1961) from the University of Wisconsin. In 1965 he received the Ph.D. degree in biochemistry from Harvard University, where he studied with J. D. Watson.

In 1966, he became an NIH postdoctoral fellow at the Institute de Biologie Moleculaire, Geneva, Switzerland. He was an assistant director for research at Cold Spring Harbor Laboratory, before joining the University of Utah as professor of biology in 1978. That year he also became a Howard Hughes Medical Institute investigator. He has been a co-chairman and professor of human genetics at the University of Utah since 1984. In 1991 he became director of the Utah Center for Human Genome Research (an NIH sponsored project). His current research includes reprogrammed genetic decoding (recoding) and DNA sequencing technology.



Thomas G. Stockham, Jr., (S'55–M'60–SM'69–F'77) received the S.B., S.M., and Sc.D. degrees in 1955, 1956, and 1959, respectively, from the Massachusetts Institute of Technology, Cambridge.

From 1955 to 1959, concurrent with his graduate studies, he was a Teaching Assistant from 1955 to 1959, receiving the Goodwin Teaching Award in 1957. In 1959, he was appointed Assistant Professor in the Department of Electrical Engineering, Massachusetts Institute of Technology. Since 1958, he has centered his research in the field of information processing. From 1966 to 1968, he was a Staff Member of the M.I.T. Lincoln Laboratory, Lexington, Mass., where he developed nonlinear systems for image enhancement and audio compression using the principle of generalized superposition. In July 1968, he was appointed to the Computer Science faculty at the University of Utah, Salt Lake City, where he and his students developed methods for modeling human vision; graphics image shading; image deblurring and compression; sound dereverberation; and color image enhancement. In 1973–1974 he served on a panel of experts for Chief Judge John J. Sirica to examine the Watergate tapes. In 1975 he founded Soundstream, Inc. Under his direction the company developed digital commercial sound recording and editing. These developments led to the establishment of these practices industry wide. The resulting accumulation of more than 100 digitally recorded albums by early 1981 triggered the advent of the compact disc (CD). In 1983 he rejoined the University of Utah faculty in the Department of Electrical Engineering where he has since been developing improved visual models and applying them to image compression methods using vector quantization. He is also an active consultant in digital image and sound processing technologies.

Dr. Stockham is a member of Tau Beta Pi, Sigma Xi, Eta Kappa Nu, and the Association for Computing Machinery. He is a Fellow of the IEEE and the Audio Engineering Society and served as president of the latter in 1982–1983. He received the IEEE Group on Audio and Electroacoustics Senior Award in 1969, the IEEE Utah Chapter Award in 1972, 1978, and 1985, the SMPTE Alexander M. Poniatoff Gold Medal for Technical Excellence in 1985, the Utah Engineers Council Engineer of the Year Award and the University of Utah College of Engineering Outstanding Teacher Award in 1986, the Audio Engineering Society Gold Medal in 1987, a National Academy of Television Arts and Sciences "Emmy" award in 1988, and a National Academy of Recording Arts and Sciences "Grammy" award in 1993.