Dispersion functions and factors that determine resolution for DNA sequencing by gel electrophoresis

John C. Sutherland, Kiley J. Reynolds and David J. Fisk

Biology Department, Brookhaven National Laboratory, Upton, NY 11973

ABSTRACT

The number of bases that can be read in a single run by a DNA sequencing instrument that detects fluorophore labeled DNA arriving at a “finish-line” located a fixed distance from the starting wells is influenced by numerous parameters. Strategies for improving the length-of-read of a DNA sequencer can be based on quantitative models of the separation of DNA by gel electrophoresis. The dispersion function of the electrophoretic system — the relationship between molecular contour length and time of arrival at the detector — is useful in characterizing the performance of a DNA sequencer. We adapted analytical representations of dispersion functions, originally developed for snapshot imaging of DNA gels, (samples electrophoresed for constant time), to finish-line imaging, and demonstrated that a logistic-type function with non-integral exponent is required to describe the experimental data. We use this dispersion function to determine the resolution length and resolving power of a LI-COR DNA sequencing system and a custom built capillary gel electrophoresis system, and discuss the factors that presently limit the number of bases that can be determined reliably in a single sequencing run.

Keywords: DNA sequencing, infrared laser scanner, fluorescence, gel electrophoresis, capillary gel electrophoresis, electrophoretic resolution, dispersion function

1. INTRODUCTION

The ability to determine the sequence of the bases in DNA is critical in biological and medical research and in applied fields including agronomy and pharmacology. Since the development of general methods to determine DNA sequence, there have been numerous advances in the technology, particularly the development of instruments that integrate electrophoretic separation, detection of single stranded DNA and analysis of the resulting data. While numerous other technologies for automated DNA sequencing have been explored, methods based on gel electrophoresis, including both the slab gel systems cited above and capillary gel electrophoresis, are still the methods of choice. Numerous incremental improvements in electrophoresis-based DNA sequencing have enhanced the performance of automated sequencers and hence raised the standards by which potential competing technologies are judged. Among the goals of continued development are improving the rate of sequencing, i.e., throughput, and reducing the need for labor-intensive steps, such as producing new gels and loading samples. One approach to these objectives is to increase the number of bases that can be determined during a single electrophoretic experiment either by increasing the number of samples or by increasing the number of bases determined in each sample, i.e., the length of read. This report addresses methods of quantifying some of the parameters that determine the number of bases that can be correctly identified in a single electrophoretic experiment.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
2. THEORY

2.1 Defining resolution

For purposes of sequencing DNA, adjacent bands in an electrophoretic gel are considered resolved if their correct number and order vis-à-vis bands in other lanes or decorated with other labels can be determined accurately. Both human observers and computational algorithms can call closely spaced bands. In developing improved electrophoretic procedures, we need a simple but robust quantitative measure of resolution with which to judge the performance of various protocols. The standard definition is that two bands are resolved if the separation between their peaks is greater than the average of their widths, i.e., \(|x - x_b| \geq |W_x - W_b|/2\), where \(x\) is whatever coordinate system is used to measure separation. Of course, peak locations and widths must be measured in the same units for this comparison to be valid. How do we define the width of a band? Some authors have used the width of the band measured at the baseline, but this is a poor choice as the point at which the trace just rises above the baseline is often not clearly defined. The usual measure of band width is the full width of the band measured at one half of the peak or maximum value. Frequently, the symbol \(\Gamma\) is used to denote the full width at half maximum. In Figure 1 we show a lane profile for three identical adjacent bands that are separated by their common full width at half maximum values. The axes are labeled to indicate the coordinates typical in automated DNA sequencing, i.e., fluorescence vs. time of arrival at the detector. Using \(\Gamma\)'s rather than "baseline" widths is a less conservative definition of resolution, but is still more conservative than is necessary, as both human readers and computer programs can correctly "call" bands that are more closely spaced. As a measure of resolution, however, this criterion is still useful for evaluating the performance of electrophoretic systems, as measured improvements will always indicate better separations.

A limitation of resolution as defined above is that two or more closely spaced bands are required for a determination, hence restricting the experimental situations in which resolution can be evaluated. For example, using this criterion for evaluating the resolution of a capillary gel system for reading DNA sequence, Luckey et al. were limited to evaluating resolution at only three points in a DNA sequence containing more than 350 bases. An alternate method, which permits more facile determinations of resolution is to compare a well resolved band with virtual copies of itself. This is the situation shown schematically in Figure 1. The implementation of this strategy is somewhat different for the analysis of DNA restriction fragments and other long DNA molecules compared to DNA sequencing. For longer DNA molecules, we determine the length of the DNA in the virtual bands that would just be resolved from the target band. In the case of DNA sequencing, however, we usually make the comparison with virtual bands migrating as if they are exactly one base longer or shorter than the experimental target band, as the goal in sequencing is always to separate bands that differ in length by only one base. In either case, we can determine the locations of the peak of the adjacent virtual bands using the equation

\[
\Delta x = \frac{dL}{dx} \Delta L = \frac{\Delta L}{dL/dx},
\]

where \(x\) is the (generalized) separation coordinate, \(\Delta x\) is the separation between the centers of the real and virtual bands, \(\Delta L\) is the difference in DNA length between the real band and the adjacent virtual band, \(L(x)\) is the relationship between the contour length of DNA and the separation coordinate, \(x\), and \(dL/dx\) is the derivative of \(L\) with respect to \(x\). \(L(x)\) is referred to as the dispersion function of the electrophoretic system. For the remainder of this report, we shall use "\(x\)" as the distance of electrophoretic migration, rather than as a "generalized separation coordinate".

2.2 Resolution length and resolving power

Lerman and Sinha\(^3\) defined the resolution length, \(\Delta L_r\), associated with a particular DNA band as the difference in lengths corresponding to a difference in band centers equal to \(\Gamma\). That is,

\[
\Delta L_r = \Gamma \frac{dL}{dx},
\]

where \(\Gamma\) is the full width at half maximum of the band and \(dL/dx\) is the derivative of the dispersion function with respect to its location determined at the center of this band. (Lerman and Sinha\(^3\) were interested in the distribution of DNA...
fragments recorded after a constant time of electrophoresis \textit{i.e.} \textit{snapshot imaging}, hence $x$ is the distance of migration in the gel. The minus sign is needed because in snapshot images longer distances of migration usually correspond to shorter DNA molecules, so $\frac{dL}{dx}$ is usually a negative number. In automated DNA sequencing, images of the DNA in the gel are usually acquired as a function of the time of arrival at a particular position, usually near the end of the gel, \textit{i.e.} \textit{finish-line imaging}, and the resolution length is given by the equation \cite{ribeiro1987}

$$
\Delta L_r = \Gamma \frac{dL}{dt}.
$$

Equation 2

A negative sign is not required because in finish-line images, a longer time of arrival usually corresponds to a longer molecule, so $\frac{dL}{dt}$ is usually positive. $\Gamma$ is always measured in the same units as the corresponding independent variable. It is sometimes useful to measure the resolution relative to the length of the molecules being studied. Thus, Ribeiro and Sutherland\cite{ribeiro1987} introduced \textit{resolving power} which is the ratio of the length of the DNA molecules in a band to the resolution length of that band. For snapshot images, resolving power is defined by the equation

$$
R = \frac{-L}{\Gamma \frac{dL}{dx}}.
$$

Equation 3

while for finish-line images, resolving power is defined by the equation

\begin{align*}
\end{align*}
Resolving power is a dimensionless quantity, and, perhaps surprisingly, some of the analytical expressions for resolving power are simpler than the corresponding expressions for resolution length, *vide infra*. Larger numerical values of resolving power indicate better resolution, in contrast to resolution length, where smaller numerical values indicate better resolution. For sequencing gels, we require that the separation of bands differing by only one base in molecular length be greater than the full width at half maximum of the given band. This is equivalent to the requirement that the resolution length is less than one base, or that the resolving power is greater than the numerical value of the length of the molecule, measured in bases. In both cases, we are using the derivative of the dispersion function to determine the location of virtual bands differing in length by one base from the selected band.

### 2.3 Dispersion functions

The dispersion function and its derivatives are determined by fitting a smooth curve to a data set consisting of the lengths of a set of DNA molecules and the centers of their bands recorded in either snapshot or finish-line images. This can be achieved either using local fitting procedure, or by use of some global function. We prefer the latter, because it is easier to perform extrapolations and the global dispersion function may provide more insight into the nature of the electrophoretic process. Exponential dispersion functions have been used extensively to describe the separation of proteins by gel electrophoresis, and there is a solid theoretical foundation for their use.\(^1\) Exponential dispersion functions are also used to describe the separation of DNA by gel electrophoresis. While such functions provide an accurate representation of the separatory process over a range of molecular lengths, they fail for both the largest and smallest DNA that can be separated under a given set of experimental conditions.\(^2\) Southern\(^3\) showed that the separation of DNA over a wider range of molecular sizes was better described by what he described as a "reciprocal model", which can also be characterized as an offset hyperbola. Whatever the name, this model requires determination of three constants from the experimental data. As presented by Southern, these were arbitrary fitting parameters,\(^4\) but the hyperbolic equation can be rearranged so that all three constants can be assigned physically meaningful names.\(^5\) In this case the dispersion function relating velocity, \(v\), to molecular length, \(L\), is given by the equation

\[
\frac{L}{L_{sm}} = \frac{v_0 - v}{v - v_m},
\]

where \(v_0\) is the average velocity of a DNA molecule of "zero length", \(v_m\) is the average velocity of a molecule of "infinite" length, and \(L_{sm}\) is the length of the molecules that travel with average velocity \(v_m = (v_0 + v_m)/2\). A snapshot image records the location of all DNA species after a constant time of electrophoresis, so the dispersion function for snapshot images is obtained by replacing all of the velocities in the above equation with corresponding distances, which we shall label as \(x\).

For finish-line images, the corresponding equation is

\[
\frac{L}{L_{tm}} = \frac{t - t_0}{t_m - t},
\]

where \(t_0\) is the time of arrival of a DNA molecule of "zero length" at the detector located a distance \(x_0\) from the well, \(t_m\) is the time of arrival of a molecule of "infinite" length, and \(L_{tm}\) is the length of the molecules that arrive at \(t_m = (t_0 + t_m)/2\). However, \(t_m\) does not correspond to the arrival of molecules traveling at \(v_m\). Indeed, \(L_{tm} = (x_0/v_0)I_{sm} = (x_0/v_m)I_{sm}\) (If very long DNAs do not reptate in a particular electrophoretic system, *i.e.*, \(v_0 = 0\), then different expressions apply, *vide infra*).

The hyperbolic dispersion function provides a reasonable approximations for many experimental data sets, and has the advantage that all three constants can be determined by a linear least-squares procedure.\(^6\) However, several workers\(^7\) have shown that better fits to the experimental data are obtained if one of the experimental parameters in the hyperbolic function is raised to a non-integral power. No physical significance was attached to the exponent,\(^8\) and increasing the number of fitting parameters is always expected to improve the quality of a fit. We shall show, *vide infra*, that some data sets are qualitatively incompatible with the three parameter hyperbolic function, but can be fit with a four parameter function in which one of the parameters is an exponent. A serious problem with most of the attempts to add an exponential term to the hyperbolic dispersion function is that the exponent is applied to a dimensioned quantity. When, as usually happens, the exponent is not an integer, the resulting equation is dimensionally inconsistent. The dispersion

\[
R = \frac{L}{\Gamma \frac{dL}{dt}}.
\]

\(R\) is a resolution parameter, which has a physical meaning in terms of velocity and length. For a given set of experimental conditions, the ratio of the resolution parameter to the dispersion parameter is expected to be constant, i.e.,

\[
\frac{R}{\Gamma} = \frac{\frac{L}{L_{sm}}}{\frac{dL}{dt}} = \frac{v_0 - v}{v - v_m}.
\]

Equation 4

This equation is dimensionally consistent, but the usual procedure is to use a smooth curve to a data set consisting of the lengths of a set of DNA molecules and the centers of their bands recorded in either snapshot or finish-line images.
function introduced by Rodbard and his colleagues\textsuperscript{24} which is a member of the class of logistic equations, is an exception, as the exponential term is applied to a dimensionless ratio. Following this approach, and introducing the physically meaningful parameters $v_0$, $v_\infty$, and $L_{vm}$, as defined above, results in the equation

$$\left( \frac{L}{L_{vm}} \right)^p = \frac{v_0 - v}{v - v_\infty}. \quad \text{Equation 5}$$

If electrophoretic conditions are maintained constant during the entire separation, Equation 5 can be converted to the form appropriate for snapshot imaging by replacing every occurrence of $v$ by $x$, the corresponding distance. For constant electrophoretic conditions, the corresponding form for finish-line imaging is given by the equation

$$\left( \frac{L}{L_{zm}} \right)^p = \frac{t - t_0}{t_\infty - t} . \quad \text{Equation 6}$$

where $t_0$, $t_\infty$, and $t_m$ are as defined above, and $L_{zm} = \left( \frac{v_0}{v_\infty} \right)^{\frac{1}{p}} L_{vm} = \left( \frac{x_0}{x_\infty} \right)^{\frac{1}{p}} L_{zm}$. These equations all revert to the corresponding forms obtained for the hyperbolic function as $p$ approaches unity. Dispersion functions based on Equation 6 for a finish-line image for the hyperbolic case ($p = 1.0$) and a typical logistic case ($p = 1.5$) are shown in Figure 2. (If very long DNAs did not reptate in a particular electrophoretic system, i.e., $v_\infty = 0$, then $\left( \frac{L}{L_{vm}} \right)^p = \frac{v_0 - 1}{v} = \frac{t}{t_0}$, and $L_{zm}$ is not defined).

Expressing the dispersion function as ratios in Equation 6 indicates that we can use any units to measure both $L$ and $t$ so long as the corresponding constants are measured in the same units. Also, the differences in both numerator and denominator on the right hand side of Equation 6 implies that we can choose the origin of the time axis at any convenient location rather than measuring the total duration of electrophoresis. Typically, we will only begin recording fluorescence a short time before the arrival at the detector of the smallest DNA is expected. However, we must use absolute coordinates in equations converting between $L_{zm}$ and $L_{zm}$ and certain other situations.

Using the logistic approximations for the dispersion function, the resolving power for snapshot imaging is given by the equation

$$R_i = \frac{p(x_0 - x_i)(x_i - x_\infty)}{\Gamma_i(x_0 - x_\infty)} , \quad \text{Equation 7}$$

where $x_i$ is the location of the center, $\Gamma_i$ is the full width at half maximum, and $R_i$ is the resolving power for the $i$th band in the image. The corresponding expression for the resolving power of a finish-line image is given by the equation

$$R_i = \frac{p(t_i - t_0)(t_\infty - t_i)}{\Gamma_i(t_\infty - t_0)} , \quad \text{Equation 8}$$
Figure 2 Dispersion functions derived from Equation 6 for a finish-line image showing the relationship between the \( t_0, t_m, t_n \) and \( L_m \) for \( p = 1 \) (---) and \( p = 1.5 \) (----). The dispersion function will pass through the point \((t_m, L_m)\) for all values of \( p \), but values of \( p > 1 \) will reduce the slope at \( t = t_m \), and throughout most of the range of \( t \) values except near \( t_0 \) and \( t_n \). Note the break in the \( L \) axis and the dashed vertical line indicating that as \( t \rightarrow t_m, L \rightarrow \infty \).

where \( t_i \) is the time of arrival of the center, \( \Gamma_i \) is the full width at half maximum, and \( R_i \) is the resolving power for the \( i^{th} \) band. The values of the band widths, \( \Gamma_i \) in these two equations must be measured in the same units as the independent variable, \( x \), or \( t \), respectively. If band widths were independent of the length of the DNA, resolving power for DNA distributions that are reasonably described by the logistic dispersion function would be parabolic with their maximum values at \( x_m \) or \( t_m \) respectively. In practice, band widths always vary as a function of length with the result that the peak in the dispersion function shifts in the direction corresponding to smaller values of \( L \). The shift is larger in the case of finish-line imaging than for snapshot imaging.\(^9\) The values of molecular length and resolving power for the three cardinal locations in snapshot and finish-line images are given in Table 1. The values of \( \Gamma \) in Table 1 refer to the full width at half maximum of a band located at the corresponding distance of electrophoresis or time of arrival, and are measured in the same units as that variable. There are singularities in the expressions for \( dL/dx \) and \( dL/dt \) at the points corresponding to molecules of "zero" length, i.e., \( x_0 \) and \( t_0 \). For the hyperbolic cases \( (p = 1) \) the extrapolated dispersion functions intersect the axis with a finite slope, [the values are \( L_m/(x_0 - x_n) \) and \( L_m/(t_m - t_n) \), respectively], while for the general logistic case, the extrapolated slopes either become infinite if \( p > 1 \), or zero if \( p < 1 \). The singularities disappear in the expressions for resolving power.

### 3. EXPERIMENTAL

#### 3.1 Slab gel electrophoresis

Gels were run in a LI-COR DNA Sequencer model 4000L (LI-COR Corp., Lincoln, NB) that uses near-infrared laser detection. All gels were 4% polyacrylamide (Long Ranger, FMS Bioproducts, Rockland, ME) in Tris (90 mM)
borate (90 mM) EDTA (2 mM) buffer and were 250 μm thick. Detection was 56 cm from the start of the gel. The static voltage across the gel was 2250 volts and temperature was maintained at 48° C. The DNA sample was produced from M13 by a Sequitherm Long-Read Cycle Sequencing Kit-LC (Epicentre Technologies, Madison, WI) using M13 forward (-29) fluorochrome labeled primer (LI-COR cat. no. 4000-20A) DNA sequencing capillary gel electrophoresis.

### Table 1

<table>
<thead>
<tr>
<th>Distance, $x$</th>
<th>Molecular Length, $L$</th>
<th>Resolution Length, $\Delta L_r$</th>
<th>Resolving Power, $R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_m$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>0</td>
</tr>
<tr>
<td>$x_m = (x_{\infty} + x_0)/2$</td>
<td>$L_{sm}$</td>
<td>$4 \Gamma L_{sm}/[p(x_0 - x_m)]$</td>
<td>$p(x_0 - x_m)/4\Gamma$</td>
</tr>
</tbody>
</table>
| $x_0$         | 0                     | $\Gamma L_{sm}/(x_0 - x_m)$     | 0                   if $p > 1$
|               |                       |                                 | 0                   if $p = 1$
|               |                       |                                 | 0                   if $p < 1$

### 3.2 Capillary gel electrophoresis

Data were collected from a bespoke four color detection apparatus similar in design to that described by Smith and his colleagues and equipped with a 10 mW argon ion laser emitting 514 nm light. The capillary was 100 μm in diameter and 34 cm long and operated with a static field of 220 V/cm. The M13 (-21) primer was used on a pGEM-32f(+) cloning vector using the Taq polymerase dideoxy termination cycle sequencing kit (Perkin Elmer Corp., Redwood City, CA) for samples terminating in thymine residues that were labeled with 6-tamra dye. The gel matrix was 4% linear polyacrylamide in 1 X TBE (3.5 M urea, 30% formamide), 1:1 AP, TEMED.
3.3 Data analysis

The LI-COR sequencer data collection program COL records the entire two-dimensional finish-line image which is subsequently analyzed by the LI-COR data analysis program ANL which extracts lane profiles and calls the sequence of the bases. The output from ANL is used by the program MKSCF that generates lane traces in Staden Curve Format. The resulting files were converted to ASCII format by a program EXPMKSCF supplied by LI-COR. Capillary gel data were stored in comma delineated ASCII format. Time-fluorescence lane profiles from both systems were imported into the peak fitting module of the data analysis program ORIGIN (Microcal Software, Inc., Northampton, MA) which was used to extract gaussian peak parameters. ORIGIN was also used to fit logistic functions to length-time data using the Levenberg-Marquardt algorithm, and for the preparation of all figures. In the fitting procedure, DNA length was treated as the independent variable and time of arrival was the dependent variable, because the length values are known accurately while the times are subject to experimental uncertainties.

4. RESULTS AND DISCUSSION

4.1 Dispersion functions for finish-line imaging of DNA sequencing gels

Figure 3 shows plots of DNA length as a function of time of arrival at the detector for static field electrophoresis in a slab gel and a capillary gel system. The data fall close to a straight line over much of the range. A straight line is what is predicted for the simple approximation that velocity is inversely proportional to molecular length, i.e. \( v \propto 1/L \), which implies that the time of arrival is directly proportional to molecular length. This hypothesis is untenable, however, as it implies that the extrapolated line should intersect the origin, which is not the case\(^1\). The hyperbolic dispersion function allows for a non-zero intercept with a constant slope \( (dL/dt = L_{\text{ref}}/t_0) \) if reptation does not occur, i.e. \( v_0 \approx 0 \). If reptation is significant, however, the hyperbolic model predicts that the slope of the dispersion function will not be constant, but will increase monotonically as a function of time after \( t_0 \), as shown in Figure 2. But the hyperbolic model cannot account for the experimental points of the smallest DNA's that fall below the straight lines fit to the data for longer DNA molecules. For the capillary gel data, an upward curvature above the straight reference line is also apparent for the longest DNA molecules in the sample. Comparison of the experimental points with the theoretical curves shown in Figure 2 indicates that a logistic function with \( p > 1 \) can reproduce the observed curvature in the data, the near linear fit over a limited range between \( t_0 \) and \( t_m \) and the increase in slope for the longer molecules. Thus, we shall use the logistic function to describe electrophoretic dispersion for these data.

4.2 Resolution of DNA bands: capillary gels with finish-line imaging

The ability of the logistic function presented in Equation 6 to describe the dispersion function for a static field capillary gel is shown in Figure 4. The top panel shows the trace of a lane in which the samples are terminated in thymine residues. The length of the DNA molecules in each band, which is known from the sequence of the DNA, is plotted as a function of the observed time of arrival of the band in the center panel of Figure 4. The solid line passing through the data points in the center panel is the logistic dispersion function determined by a non-linear least squares fitting procedure which returned \( t_0 = 17.6 \text{ min}, t_m = 71.3 \text{ min}, t_\alpha = 125 \text{ min}, L_{\text{ref}} = 726 \text{ bases}, \) and \( p = 1.43 \). Thus, Equation 6 provides an excellent approximation of the shape of the experimental dispersion function for these data. The open squares in the lower panel are the values of the full width at half maximum of each band plotted as a function of the time of arrival of that band. Ignoring the first few poorly resolved bands that reached the detector, the increase in band width as a function of arrival time is close to linear, with an increase of slightly more than a factor of two over the range of times recorded. Resolution length mirrors the band width function, as the derivative of the dispersion function is approximately constant for most of the data. Resolving power, which is plotted in the lower panel, peaks between 40 and 45 min, well before the median time, \( t_\alpha \) of 71 min, reflecting the increase in \( \Gamma \) as a function of increasing \( t \). Resolving power is greater than \( L \) over most of the range of times (note the change in vertical scales between the lower and middle panels). This comparison is seen more clearly in Figure 6. As a function of increasing time of arrival and DNA length,

\(^1\)Note that for the slab gel, data collection began \( \approx 4 \) hr after the start of electrophoresis, so the true origin is at about \( -4 \) hr.
the scatter observed for band widths increases, hence leading to corresponding variations in resolving power. Such apparently random variations are of considerable importance in DNA sequencing as the length of read is effectively limited by the first occurrence of unresolvable bands.

4.3 Resolution of DNA bands: slab gels with finish-line imaging

Figure 5 summarizes data recorded from a LI-COR automated DNA sequencer. The upper panel is the lane trace for a sample terminated at adenine residues. The lengths of the resolved bands are plotted against their times of arrival of their peaks at the detector in the center panel. Data collection began shortly before the arrival of the first band, about four hours after the beginning of electrophoresis, and continued for about 14 hours thereafter. We were able to resolve bands of known length arriving at the detector for about 9 hours, corresponding to DNA contour lengths of more than 1,000 bases. The software normally used with the LI-COR sequencer called the sequence to about 825 bases, and with high accuracy to 600 bases. This program had to integrate data from four lane traces, not just the single lane shown in Figure 5. The dispersion function, full width at half maximum, and resolving powers for these bands are shown in the center and lower panels of Figure 5. Using an extrapolation of the dispersion function described below, we find that the bands arriving at the detector at the end of the trace correspond to DNA lengths of about 1600 bases. As shown in Figure 3, the data drop below an extrapolated straight line for small values of $L$. The non-linear fitting procedure returned $\tau_0 = 0.31$ h,
Figure 4  DNA sequencing by capillary gel electrophoresis with visible fluorescence, finish-line detection. Fluorescence profile (upper panel), dispersion function (center panel), band widths, Γ, (lower panel, right axis, □ ) and resolving power, R, (lower panel, left axis, • ) for a sample of thymine labeled bases for pGEM-32f(+) DNA.

\[ t_w = 37.3 \text{ h, } L_{on} = 2460 \text{ bases, and } p = 1.34. \]  The resulting logistic approximation to the dispersion function is shown as a solid curve in the center panel. Data collection was continued for another 5 h, but the data were more difficult to resolve and are not included in the center and lower panels. The increase in band width is, as for the capillary gel data, approximately linear. Apparently random variations in band widths are responsible for the scatter in the resolving power.

4.4 Comparing resolution of slab and capillary gels

Resolving power and resolution length permit direct comparisons of electrophoretic systems that differ greatly in other experimental parameters. For example, we can compare the performance of snapshot and finish-line imaging systems using these methods."14 In Figure 6, we compare the resolving power of the capillary gel data from Figure 4 with the slab gel data from Figure 5. Up to about 400 bases, the capillary system gives higher resolving powers, with values well above the classical threshold for resolution equal to the corresponding DNA length, which is shown as the dashed line in Figure 6. The data from the slab gel, on average, never get above this line, but the peak in resolving power occurs at a far larger value of DNA length, about 800 bases.

4.5 Extrapolation of dispersion functions

The dispersion functions for the capillary and slab electrophoresis shown above are close to linear over much of their range, so we could easily measure a single slope to use in the determination of resolving power in Equation 4. In addition to perhaps providing some insight into the process of electrophoretic separations, the development of global analytical expressions for the dispersion function lets us extrapolate well past the limits of the experimental data, as shown in Figure 7. The display of the data in Figure 7 also indicate how well the logistic equation represents the
4.6 Factors affecting performance of DNA sequencing instruments

We have investigated two factors that affect the performance of DNA sequencers: the slope of the dispersion function, and the widths of individual bands. For sequencers that acquire finish-line images, the slope of the dispersion function is nearly constant over the useful range of the instrument, except for the smaller DNA molecules, for which resolution is usually not a problem. The length of read is limited by two features of band width. First, bands tend to increase (linearly) in width as a function of increasing DNA length and, hence, time of arrival. Second, there are significant, apparently random, variations in band width, which in some circumstances may result in the first uncallable site in the image.

There is another factor that may influence the length and accuracy of base identification that is not included in the formalism of resolving power, but which can be demonstrated from the same data sets. In determining resolving power, we have assumed that the center of each band arrives at the detector at exactly the time expected for a molecule of its length. Variations in the time of arrival of a band compared to its expected time of arrival as judged by the other bands in the data set may lead to errors or ambiguities in base identification. By analogy with other systems in which events (in this case the arrival of a band at the detector) are expected at predictable times, we can refer to this effect as phase jitter in the electrophoretic system. We could describe phase jitter in terms of the differences in the expected and actual times of arrival, but a better approach is to reflect the time differences off of the dispersion function and report the differences in apparent length of the DNA in the band associated with the difference in time of arrival. We define the length-time variation as the difference in molecular length of the DNA molecules in the $i^{th}$ band and the
Figure 6  Resolving power for the capillary gel data from Figure 4 (□) and the slab gel data from Figure 5 (●) plotted as a function of DNA contour length. The dashed line is the classical limit of resolution for which \( R \) equals the numerical value of \( L \). The solid lines through the two experimental data sets are parabolas determined by the method of least squares. In a typical DNA sequencing situation, length of read will tend to be limited by random scatter about these average values. Such variations result mainly from variations in band widths.

expected (interpolated) length of molecules in the band arriving at the observed time of arrival of the \( i^{th} \) band, i.e.,

\[
L_i - L(t_i) \equiv \left[ t_i - t(L_i) \right] \frac{dL(t_i)}{dt} \equiv \frac{t_i - t(L_i)}{t(L_i) - t(L_{i-1})}.
\]

Note that \( L_i \) is an integer, but \( L(t_i) \) is a real number. For this calculation, the small but systematic errors in the logistic dispersion function become significant, so we must use a local averaging procedure to determine the dispersion function, \( L(i) \). Describing this effect in terms of molecular length rather than time of arrival facilitates comparisons between different separation systems, as shown in Figure 8 for some of the slab and capillary gel data described above. In an ideal system there would be no phase jitter, while any workable DNA sequencer must exhibit values much less than unity. The data shown in Figure 8 fulfill this condition, although some of the extreme values approach unity for the longer DNAs separated in the slab gel system. There is no obvious correlation between the magnitude of the band width and the phase shift in the time of band arrival (data not shown).

5. SUMMARY

We have shown that hyperbolic functions are incapable of describing the separation of single stranded DNA in sequencing gels under some circumstances, but logistic functions, which are a superset of the hyperbolic functions, can reproduce all of the observable features of the experimental dispersion functions. Accurate dispersion functions are useful in quantitating electrophoretic resolution, and global models have the advantage of facilitating extrapolations to DNA lengths outside the experimental range. An important goal of future research is relating the parameters obtained
from model dispersion functions to properties of the electrophoretic (gel density, electric field strength, viscosity, ionic strengths, etc.) and to the molecular properties of the DNA.

Our results also demonstrate quantitatively an observation that has been made by other workers, that the main limitation of DNA sequencing for finish-line detection is band spreading and not loss of band separation. Average band widths appear to increase in an approximately linear manner over the range of our data, but apparently random scatter about average band widths may be the factor that determines the length of read in a particular experiment, although apparently random variations in band arrival times (phase jitter) may also play a role.

6. ACKNOWLEDGMENTS

We thank Drs. Betsy Sutherland, John Dunn, and Mark Quasada, Biology Department, BNL, Lyle Middendorf, Daniel Grone, and David Sorenson, LI-COR Corp. for helpful discussions and Denise Monteleone and John Trunk, for assistance with computer systems. Research supported by the Office of Health and Environmental Research, US Department of Energy and a Cooperative Research And Development Agreement with the LI-COR Inc., Lincoln, NB.
Figure 8 Plot of the difference between actual length of the DNA in a band arriving at the detector at time $t$, and the length of the DNA expected in a band arriving at the detector at that time for the capillary gel data shown in Figure 4(b) and the slab gel data from Figure 5 (●).

7. REFERENCES


Freeman, S.E., Blackett, A.D., Monteleone, D.C., Setlow, R.B., Sutherland, B.M. & Sutherland, J.C., "Quantitation of Radiation-, Chemical-, or Enzyme-Induced Single Strand Breaks in Nonradioactive DNA by Alkaline Gel Electrophoresis: Application to Pyrimidine Dimers", Analytical Biochemistry 158, 119-129 1986.


