850–860 *Nucleic Acids Research, 1997, Vol. 25, No. 4*

Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein–DNA binding constant

Susan E. Shadle[+], Douglas F. Allen[§], Hong Guo, Wendy K. Pogozelski[‡], John S. Bashkin[¶] and Thomas D. Tullius*

Department of Chemistry, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

Received September 30, 1996; Revised and Accepted December 23, 1996

ABSTRACT

A computer program, GelExplorer, which uses a new methodology for obtaining quantitative information about electrophoresis has been developed. It provides a straightforward, easy-to-use graphical interface, and includes a number of features which offer significant advantages over existing methods for quantitative gel analysis. The method uses curve fitting with a non-linear least-squares optimization to deconvolute overlapping bands. Unlike most curve fitting approaches, the data is treated in two dimensions, fitting all the data across the entire width of the lane. This allows for accurate determination of the intensities of individual, overlapping bands, and in particular allows imperfectly shaped bands to be accurately modeled. Experiments described in this paper demonstrate empirically that the Lorentzian lineshape reproduces the contours of an individual gel band and provides a better model than the Gaussian function for curve fitting of electrophoresis bands. Results from several fitting applications are presented and a discussion of the sources and magnitudes of uncertainties in the results is included. Finally, the method is applied to the quantitative analysis of a hydroxyl radical footprint titration experiment to obtain the free energy of binding of the λ repressor protein to the O_R1 operator DNA sequence.

INTRODUCTION

Chromatographic techniques, such as HPLC and FPLC, are invaluable not only because of the relative ease with which separation of substances can be achieved, but also because of established methods for quantifying amounts of resolved material. In biochemistry and molecular biology, gel electrophoresis is employed for the separation of proteins and nucleic acids. Methods available for quantitation of electrophoresis data require a digital image of the gel, obtained *via* phosphorimagery or densitometry, followed by analysis of the digital image to obtain positions and intensities of bands in the gel. However, the approaches which have been used most often to quantify electrophoresis results have suffered from a variety of limitations. As a result, no standard for high-resolution, quantitative analysis of electrophoretograms has been adopted. In order to realize the full potential of the electrophoresis technique, an easy, reliable method for quantitative analysis is needed.

The simplest quantitative approach involves the integration of intensity in a spot or rectangle drawn on the gel image. While this technique has been successfully applied in numerous experiments [e.g., refs (1–3)], it can only be used to determine individual band intensity in cases for which bands are extremely well-resolved. As a result, this approach is often limited to the integration of a group of closely spaced bands.

A second type of approach is generally applied to an average peak profile, or linegraph, representation of the data. It involves simple integration of peak area between selected boundaries, chosen to be the minima between adjacent peaks (4–6). This approach is also limited to cases in which peaks are extremely well-separated, or when the area under several adjacent peaks is sought. For overlapping bands, the determination of minima to divide the peaks is unreliable and the technique is therefore unable to quantify individual band intensities with accuracy.

Because most electrophoresis data consist of a series of overlapping bands, meaningful quantitative results can only be obtained for individual bands if the contribution to the area of each peak by its neighbors is accounted for. To address this point, a number of programs have been described which model each band in an average linegraph peak profile with an analytic function and use least-squares optimization to fit the data (6–11). The fitted curves are then integrated to provide information about bands in the data. The Gaussian function has been the most commonly used function in this type of approach (12–15). However, the Gaussian, which is characterized by low intensity in the tailing regions of the function, may not be an accurate model for electrophoresis bandshapes (16). While the Gaussian may fit the top half of the band profile, in cases for which the tails of gel peaks are visible, the data are much wider than the Gaussian. Some authors have suggested that electrophoresis bands might be better modeled by asymmetric functions (17,18)

*To whom correspondence should be addressed. Tel: +1 410 516 7449; Fax: +1 410 516 8468; Email: tom@radical.chm.jhu.edu

Present addresses: [+]Department of Chemistry, Boise State University, Boise, ID 83725, USA, [§]Stratus Computer, Vienna, VA 22182, USA, [‡]Department of Chemistry, SUNY Geneseo, Geneseo NY 14454, USA and [¶]Molecular Dynamics, Sunnyvale, CA 94086, USA

and/or functions having broader tailing regions than the Gaussian (19).

Many of the approaches which have used analytic functions for band fitting require subtraction from the raw data of a background, which increases towards the top of a gel lane, in order to produce a good fit to the data (6,19). Such a procedure has the unfortunate effect of subtracting off the tailing regions of the bands. Other methods have also been used to determine an appropriate background subtraction of the raw data for analysis (17,18). However, in many cases the method for determination of a background, to allow for valid comparisons between lanes or peaks in a lane, is not clear.

We have developed a new computer program, GelExplorer, to accomplish quantitative analysis of electrophoresis data. It provides a straightforward, easy to use approach, and includes a number of features which offer significant advantages over existing methods for quantitative gel analysis. Our method uses curve fitting with a nonlinear least-squares optimization to deconvolute overlapping bands. Unlike most curve fitting approaches, however, we treat the data in two dimensions, fitting all the data across the entire width of the lane. This allows for accurate determination of the intensities of individual, overlapping bands, and in particular allows imperfectly shaped bands to be accurately modeled.

In this paper we describe experiments which demonstrate empirically that the Lorentzian lineshape reproduces the contours of an individual gel band and provides a better model than the Gaussian function for curve fitting of electrophoresis bands. We introduce the strategy employed by GelExplorer for curve fitting analysis. Results from several fitting applications are presented and a discussion of the sources and magnitudes of uncertainties in the results is included. Finally, the method is applied to the quantitative analysis of a hydroxyl radical footprint titration experiment to obtain the free energy of binding of the λ repressor protein to the O_R1 operator DNA sequence.

MATERIALS AND METHODS

Single-band electrophoresis data

A gel image with a single, isolated band was obtained using the following procedures. The plasmid pUC18 was amplified in the DH5α strain of *Escherichia coli*, isolated using the alkaline lysis method, and purified by ultracentrifugation through a cesium chloride gradient (20). The DNA was 3′-radiolabeled by standard methods (20) at the *Bam*HI site using [α-^{32}P]dGTP (Amersham, Arlington Heights, IL) and ddATP. After a second cut at the *Pvu*II site, the desired 123 bp fragment was separated from the 199 bp fragment on an 8% native polyacrylamide gel, and recovered using the 'crush and soak' method (21). A sample of the 123 bp fragment (100 d.p.m./lane) was ethanol precipitated, rinsed, and lyophilized. The pellet was dissolved in formamide loading dye, heated at 90°C for 5 min, and loaded onto a 6% denaturing polyacrylamide sequencing gel [acrylamide:bisacrylamide ratio of 19:1]. After electrophoresis, the gel was dried onto filter paper and exposed to an imaging phosphor plate for 36 h.

Hydroxyl radical treatment of A-tract DNA

The construction of the pUC18 plasmid containing an A_5N_5 insert has been described previously (22). A 260 bp *Acc*I–*Pvu*II restriction fragment, 3′-radiolabeled at the *Acc*I site by standard methods (20) with [α-^{32}P]dCTP (Amersham), was used for hydroxyl radical cleavage reactions (23,24). Each 100 µl hydroxyl radical cleavage reaction involved treatment of radiolabeled DNA (10 000 d.p.m.) and the following final reagent concentrations: 10 mM Tris–HCl (pH 8), 10 mM NaCl, 50/100 µM Fe(II)/EDTA, 0.3% H_2O_2, and 1 mM sodium ascorbate. The reaction was stopped after 2 min by the addition of 100 µl of a solution of 13.5 mM thiourea, 13.5 mM EDTA, and 0.6 M sodium ascorbate. DNA was precipitated, rinsed, and lyophilized. The pellet was dissolved in formamide loading dye, heated at 90°C for 5 min, and loaded onto an 8% denaturing polyacrylamide sequencing gel [acrylamide:bisacrylamide ratio of 19:1]. After electrophoresis, the gel was dried onto filter paper and exposed to an imaging phosphor plate for 48 h.

Protein footprint titration

The λ cI repressor used in these studies was the generous gift of Professor Gary Ackers and was prepared as previously described (1). A stock solution of 6.92 µM protein in storage buffer [10 mM Tris (pH 8.0), 0.2 M KCl, 2 mM $CaCl_2$, 0.1 mM DTT, 0.1 mM EDTA, 5% glycerol] was made and stored at –70°C. Dilutions in a 5:8 ratio were made starting with 10 µl stock plus 6 µl dilution buffer [10 mM Tris (pH 8.0), 200 mM KCl, 2 mM $CaCl_2$, 0.1 mM EDTA, 5% glycerol]. Subsequent 5:8 dilutions were made from each protein dilution. A DNA binding activity of 80% and a dimer dissociation constant of 27.7 nM (25) was used to calculate the concentration of λ repressor dimer present in solution. Total monomer concentrations were corrected for the reduced activity prior to calculation of the dimer concentration (26).

A 31 bp insert containing the O_R1 binding site was cloned into pUC18 at the *Pst*I restriction site. The plasmid was amplified in the DH5α strain of *E.coli*, isolated using the alkaline lysis method, purified by ultracentrifugation through a cesium chloride gradient (20), and stored in TE buffer at –20°C. A 231 bp *Eco*RI/*Bgl*II restriction fragment was 3′-radiolabeled at the *Eco*RI site according to published methods, using [α-^{32}P]dATP and [α-^{32}P]dTTP (Amersham) to 'fill in' the site, followed by a 'cold chase' of dNTPs (27). The desired labeled fragment was gel-purified and isolated by overnight 'crush and soak' treatment at 4°C (21).

The DNA-repressor binding reaction was performed as follows. Each 35 µl binding reaction mixture contained 3.5 µl binding buffer (0.1 M HEPES buffer (pH 7.0), 0.5 M KCl, 10 mM $CaCl_2$, and 0.1 mM EDTA), 2 µl calf thymus DNA (0.1 mg/ml), 14.5 µl TE buffer, 5 µl radiolabeled DNA (~27 000 c.p.m. total), and 5 µl λ repressor of appropriate concentration. The binding reactions were allowed to come to equilibrium in a water bath at 22°C for 30 min. Hydroxyl radical footprinting (28) involved addition of 5 µl each of 2/4 mM Fe(II)/EDTA, 10 mM sodium ascorbate, and 0.3% H_2O_2. The Fe(II)/EDTA and H_2O_2 solutions were made fresh; the ascorbate solution was stored at –20°C. The footprinting reaction was stopped after 1 min by addition of a stop solution to give final concentrations of 7.5 mM thiourea and 0.3 M NaOAc. DNA was precipitated, rinsed, and lyophilized. The pellet was dissolved in 3 µl of formamide loading dye, heated at 90°C for 5 min, and loaded on a 10% denaturing polyacrylamide sequencing gel [acrylamide:bisacrylamide ratio of 19:1]. Each 6 mm-wide well in the gel was separated by 3 mm to maximize separation of the lanes of data for fitting. After electrophoresis, the gel was transferred to Whatman filter paper,

dried, and exposed to an imaging phosphor plate for >9 days. The long exposure time was necessary to maximize the signal-to-noise ratio for curve fitting analysis.

Data analysis

For each of the above data sets, the exposed imaging phosphor plate (Molecular Dynamics, Sunnyvale, CA) was scanned with a Model 400E PhosphorImager (Molecular Dynamics). An image of each lane for curve fitting was cropped from the gel image using the ImageQuant™ software package.

Curve-fitting experiments

GelExplorer, the software package developed in our laboratory for quantitative analysis of electrophoresis data, utilizes the IRIS Explorer™ (version 3.0) programming environment (29) for data visualization and analysis. Nonlinear least-squares fits to the data utilize the Levenberg–Marquardt algorithm as coded in *Numerical Recipes in C* (30). The fitting routine has been adapted to output confidence limits (one standard deviation) and a correlation matrix of the parameters from the covariance matrix calculated in the fitting algorithm. The code for these additions was adapted from GnuPlot fit.c.

Fitting experiments were performed on a Silicon Graphics Indigo R3000 (33 MHz) or Indigo^2 R4400 (200 MHz) workstation; an average fit took ~12 h or 2 h, respectively (75–80 slices, 60–70 peaks/slice).

Determination of binding constants

Binding isotherms obtained from footprint titration data were fit using NONLIN, a program for non-linear least-squares analysis (31). The free energy of λ repressor dimer binding to the O_R1 site was determined according to published methods (27) with modifications described below.

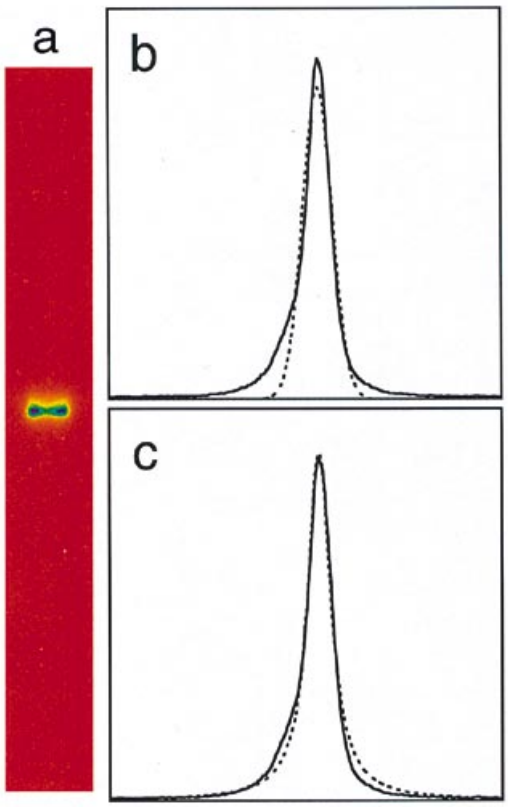


Figure 1. (**a**) Image of a lane of electrophoresis data containing a single, isolated band, as described in the text. The image is generated from an 8% denaturing polyacrylamide gel exposed to an imaging phosphor plate. (**b**) Linegraph comparison of average data (—) and GelExplorer Gaussian fit (----) to the gel data in (a). (**c**) Linegraph comparison of average data (—) and GelExplorer Lorentzian fit (----) to the gel data in (a).

CURVE FITTING OF ELECTROPHORESIS DATA

Lineshape of electrophoresis bands

The use of curve fitting to obtain reliable quantitative information about the intensities of bands in a lane of electrophoresis data requires that the modeling function be an accurate representation of the bandshape of the data. Because most electrophoresis data are a series of bands with overlapping intensities, the true shape of a single band can be difficult to determine. To overcome this difficulty, we produced a lane of electrophoresis data containing a single band. The image of this band is shown in Figure 1a. This band was modeled using our quantitative two-dimensional curve fitting approach (*vide infra*) with Gaussian and Lorentzian lineshapes. The fits are compared to the linegraph of the band in Figure 1b and c. The fitting results clearly show that the Lorentzian lineshape is a better model for the data than the Gaussian. Our results are consistent with other studies which have found that functions having greater intensity in the 'tailing' regions of the peak provide a better approximation of electrophoresis band intensity than the Gaussian (19). While the peak in Figure 1 demonstrates some asymmetry, the symmetric Lorentzian function is a very good approximation of the peak intensity. In tests using a wide variety of electrophoretic data, we have found that gel bands are reproduced very well by the Lorentzian function, without the inclusion of an additional parameter allowing for peak asymmetry.

In the Appendix to this paper Jeremy Berg provides a mathematical derivation which demonstrates that the lineshape of the image of a gel band is Lorentzian.

Curve-fitting: approach

The GelExplorer program uses curve fitting to obtain a quantitative description of gel electrophoresis data, one lane at a time. The bands in a gel lane are deconvoluted by simultaneously fitting a set of Lorentzian lineshapes to each band in the lane. The optimized Lorentzians provide accurate information about the integrated intensity and position of each band in the lane.

Each lane is treated as a two-dimensional image of pixels. The data are analyzed as a set of neighboring slices. Each slice is one pixel wide and extends the length of the lane, parallel to the direction of electrophoresis. Each peak in each slice is modeled with a Lorentzian function. Nonlinear least-squares optimization to the data is performed *separately* for each slice of data in a lane. Because each slice is optimized separately, variations in the bandshape across the lane are reproduced and a detailed description of *all* data present in the lane is obtained.

Curve-fitting: implementation

Here we briefly describe the steps involved in the use of GelExplorer for quantitative analysis of electrophoresis data. GelExplorer runs under the IRIS Explorer™ programming environment. IRIS Explorer™ was first implemented on Silicon Graphics workstations, and has since been ported to Sun, Hewlett Packard, IBM, DEC, and Cray computers. Individual modules in IRIS Explorer™ are linked together to perform specific program functions with an easy-to-use, graphical interface.

Quantitative analysis of electrophoresis data first requires a digital image of each lane of a gel. We have found that phosphorimager-generated data is superior to densitometer data because of the larger dynamic range and because long exposures of phosphorimager plates allow for increased signal-to-noise ratios of the image. A constant background is subtracted from the entire image to account for the plate (or film) background. The average pixel intensity from a gel region without any data serves as the background value.

Next, the region of a given lane to be fit is defined. The top and bottom boundaries define the least- and best-resolved bands in the lane which will be modeled, respectively. The left and right boundaries define how much of the width of the lane will be fit (how many pixel slices will be included). The criteria which determine the choice of these boundaries will be described in further detail below.

The Lorentzian function used to model each peak in a lane of electrophoresis data is given by equation **1**, where C = amplitude, x = position, and γ = full-width at half-height.

$$y(x) = \frac{C\gamma}{(x-x_o)^2 + \frac{\gamma^2}{4}} \qquad \mathbf{1}$$

Starting values for three parameters must be specified for each Lorentzian to be included in the fit. Peaks are specified and positions are chosen by clicking on the image of the data at each band position at which a Lorentzian lineshape should be modeled. Widths are given a default value of 20 (pixels) and starting amplitudes are guessed automatically *in each slice* so that the height of the starting Lorentzian matches the pixel intensity of the data at the center designated for the Lorentzian. Thus, the standard set of starting parameters defines a set of peaks; each peak has the same starting width and position in all slices, but has a different amplitude in each slice (to account for variability in the band over the width of the lane). All starting parameters can be conveniently edited. Nonlinear least-squares optimization to the data is performed separately for each slice of data in a lane. The criteria for convergence are defined by the user such that the χ^2 function of the optimization changes by less than a specified value (the tolerance) for n specified iterations.

Figure 2a highlights a single slice within an image of a gel lane to be modeled by a series of Lorentzian curves. Linegraphs depicting the total fit to the slice and the individual optimized Lorentzian contributions to the fit are shown in Figure 2b and c, respectively. In a full fitting analysis of the lane, each slice will be modeled by such a sum of Lorentzian contributions.

After optimization, the parameters for a particular band can be averaged or summed across all slices. An average position and average width, and an average or summed amplitude, are obtained. Importantly, the Lorentzian for a given band may have different positions over the width of the lane, but since each slice is fit separately, the true amplitude of the band over the width of the lane is reflected in the summed amplitude. Further, since the amplitude of a Lorentzian is directly proportional to the area (integrated Lorentzian area = $2\pi C$), the amplitude parameter can be used directly for comparisons which require integrated band intensities. Unlike approaches which aim to quantify electrophoresis data by fitting an average linegraph or several pixels in the center of a lane, our approach takes into account variations in bandshapes and band intensity across the width of a lane.

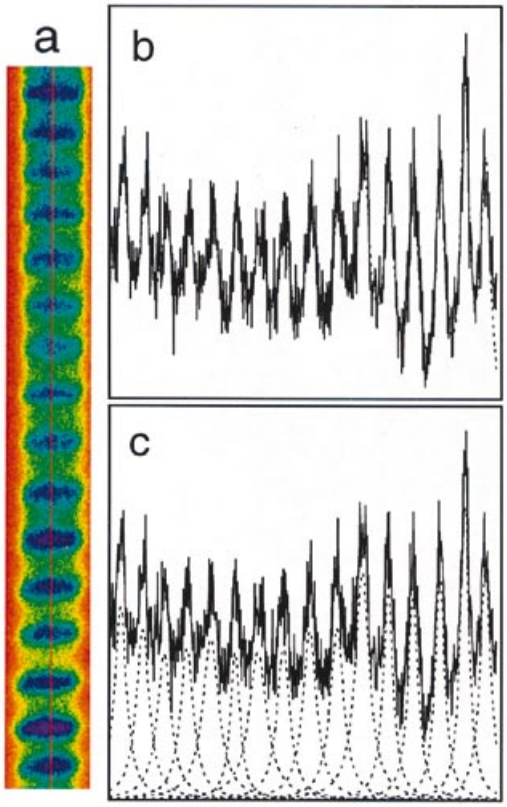


Figure 2. (**a**) Image of a lane of electrophoresis data containing multiple bands. The image is generated from an 8% denaturing polyacrylamide gel exposed to an imaging phosphor plate. The red vertical line in the center of the lane highlights a single slice of data. (**b**) Linegraph comparison of a single slice of data [highlighted in (a)] (—) and the GelExplorer fit to the data (----). (**c**) Linegraph comparison of a single slice of data [highlighted in (a)] (—) and the individual Lorentzian curves which have been optimized in the fit to the data (----).

Two-dimensional analysis is essential for accurate quantitation of the intensity of an electrophoresis band. An image of an irregularly-shaped electrophoresis band is shown in Figure 3a. One approach to modeling this band is to fit a single Lorentzian lineshape to the average data, equivalent to fitting a linegraph produced by a one dimensional scan of the lane. Figure 4a shows the comparison between the linegraph of the data and a *single* Lorentzian fit to the data. The fit deviates from the data because of asymmetry in the linegraph. While GelExplorer offers this option, we find that the data are better modeled by the alternative, two-dimensional, strategy for quantifying electrophoresis band intensity. The image of a multiple-slice fit to the data, generated from individual Lorentzians optimized to each slice of the data, is shown in Figure 3b. The individual Lorentzians can be averaged across all slices in the fit to obtain a linegraph of the fit.

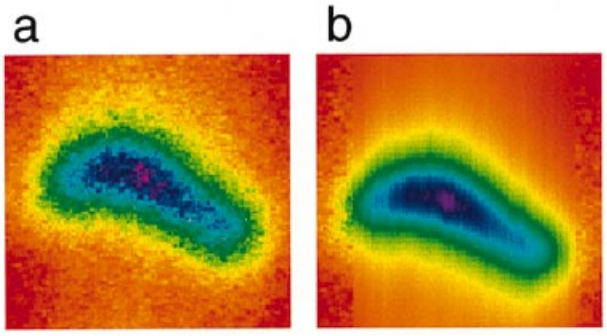


Figure 3. (**a**) Image of a single irregularly-shaped electrophoresis band. (**b**) Image of the sum of Lorentzian curves fit to the data.

This fit linegraph is compared to the linegraph of the data in Figure 4b. The multiple-slice fit clearly reproduces the contours of the average (linegraph) data better than the single Lorentzian fit. The reason for this is illustrated in Figure 4c, in which the individual, optimized Lorentzians are plotted in three dimensions, across the slices of the lane. Both the amplitudes and the positions of the Lorentzians vary across the width of the lane. When these Lorentzians are averaged, as in Figure 4b, an asymmetric peak results. If the optimized Lorentzians from the fit shown in Figure 4c are all assigned the same position and then averaged, the lineshape in Figure 4d is obtained. This is a representation of what would be observed for the average band intensity if the bandshape were perfectly perpendicular to the direction of electrophoresis.

The amplitude of the band obtained from the *single* Lorentzian fit to the linegraph of the data (Fig. 4a) is 1213 ± 13, with a width of 31.0 ± 0.4. In contrast, the average of the Lorentzian amplitudes across all slices of the multiple-slice fit (Fig. 4b) is 1150 ± 2, with an average width of 24.8 ± 0.1 pixels. The fit to the average data therefore overestimates the band amplitude by ~5%. It is not uncommon for electrophoresis bands to adopt an irregular shape. For a series of irregular, but experimentally reasonable, electrophoresis bands, we determined the intensity from fitted amplitudes. Depending on the degree of irregularity, the band intensity was overestimated by 2–6% by single Lorentzian fits to the average linegraph data, relative to average of fits to individual slices. This underscores the importance of fitting over the whole width of the lane to obtain an accurate description of band intensity.

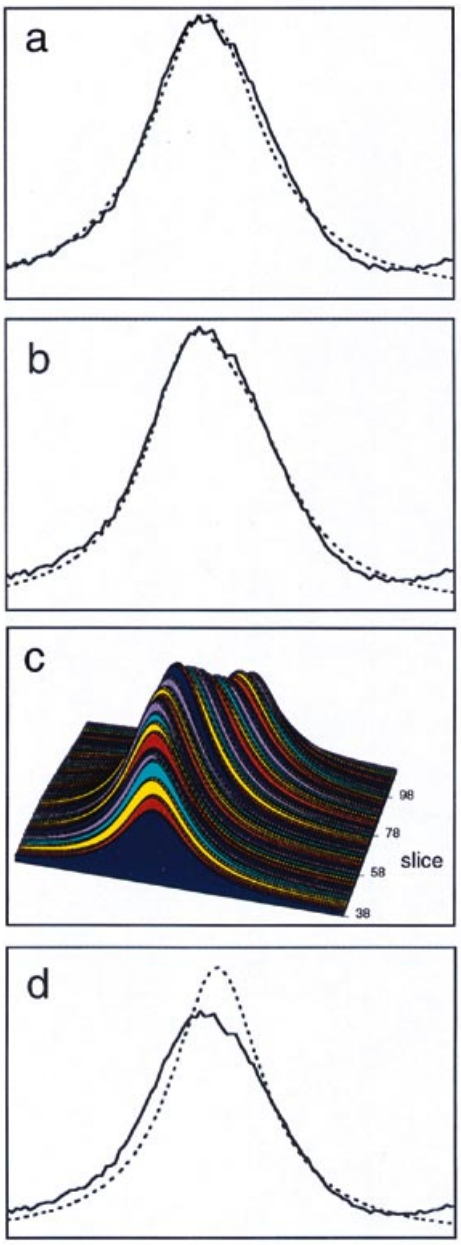


Figure 4. (**a**) Linegraph comparison of data averaged across the width of the lane (—) for band shown in Figure 3a and the Lorentzian fit to the average data (----). (**b**) Linegraph comparison of data averaged across the width of the lane (—) for band shown in Figure 3a and the average of Lorentzians fit to each slice of the data, across the width of the lane (----). (**c**) Three-dimensional plot of the individual Lorentzians, in even-numbered lanes, optimized to the band in Figure 3a. (**d**) Linegraph comparison of data averaged across the width of the lane (—) for band shown in Figure 3a and the average of Lorentzians fit to each slice of the data, across the width of the lane, modified to have the same position (----).

APPLICATIONS

Hydroxyl radical cleavage patterns of A-tract DNA

GelExplorer has been applied to quantify the intensities of bands produced by hydroxyl radical treatment of a restriction fragment containing four phased A tract [A_5TG_3C] sequences (22). We have chosen this DNA molecule as a test case because A tracts have characteristic hydroxyl radical cleavage patterns (32). Having four repeats of the same sequence in a DNA molecule allows for evaluation of the consistency and reproducibility of the curve fitting procedure to model the experimental cleavage pattern. Shown in Figure 5a is an image of the background-subtracted data. The hydroxyl radical cleavage pattern, which reflects structural variations in the A tracts (22), shows a repeating sinusoidal pattern. There is an apparent increase in overall intensity towards the top of the gel lane, which is often attributed to an unspecified 'background' in hydroxyl radical cleavage data. GelExplorer fitting was undertaken to determine the source of the signal variations over the length of the lane. Figure 5b shows the image of the fit, generated from the sum of 70 optimized Lorentzian curves. Linegraphs comparing the average data to the average fit (across the width of all slices in the fit) are shown in

Figure 6a. A subset of the Lorentzian contributions to the total fit are highlighted in Figure 6b. The fitting procedures reproduce the contours of the data extremely well. The amplitudes for the peaks in the four A_5N_5 sequences are fairly consistent over the length of the lane. This is demonstrated by a histogram plot of the fitted amplitudes in Figure 6c. While the individual Lorentzians have somewhat higher peak heights at the top of the lane, the widths of the peaks also decrease towards the top of the lane, and, as a result, the amplitudes are relatively invariant. Thus, the increase in total intensity towards the top of the lane (Fig. 6a) is a result of overlapping bands which are less well-resolved than those at the bottom of the lane, *not* because of a change in the background or in intensities of bands. This example shows how deconvolution of gel bands by curve fitting allows quantitative comparison of cleavage at nucleotides throughout a DNA molecule.

Uncertainties in the fit results

There are several sources of uncertainty in the fitting method. For example, the error bars in Figure 6c reflect the uncertainty in the amplitude parameter introduced by the fitting procedure itself. This uncertainty is generally ≤1% of the amplitude value for a given peak, with uncertainties up to ~2% in the least well-resolved peaks in the fit. However, the uncertainty in the fitted amplitudes determined by the fitting procedure is not the only source of uncertainty in the quantitative results. Other factors which contribute to uncertainty include the choice of baseline subtraction, the boundaries chosen for the fit, the dependence of the fit on starting parameters, and the convergence criteria. We have performed a series of fitting experiments on several sets of data to determine the magnitude of uncertainty introduced by the method.

The background subtracted from the raw data reflects only the imaging phosphor plate background. GelExplorer will, however, successfully fit both raw data and data from which more than the plate background has been subtracted. The fitting results vary as expected: lower amplitude values are obtained for fits to data with higher background subtractions. As a result, it is important to use the same criteria for background-subtraction for all fitting procedures. For comparisons of different lanes within the same gel, identical background values are subtracted and the background subtraction does not contribute uncertainty to comparisons between these lanes.

The top and bottom boundaries of the fit define which peaks will be included in the fit. The bottom boundary is chosen so that the fits include the best-resolved band in the lane. The top boundary of the fit is limited by the fact that, in a single slice of data, the valleys between poorly-resolved peaks are often not well-defined. Thus, at the top boundary of the fit, an artificial endpoint must be imposed. The top-most peaks defined by the boundary will have intensity contributions from peaks above them in the lane which are not modeled by the fit. As a result, the amplitudes of the top-most peaks in the fit are not an accurate reflection of the intensities of those bands. Comparison of a 70 peak fit to a 65 peak fit for the A-tract cleavage pattern in Figure 6a reveals that the top four peaks in the 65 peak fit have amplitudes which differ from the analogous peaks in the 70 peak fit by more than the uncertainties determined by the fits. All other differences between the two fits are less than the uncertainty in the fit. For this reason, one must always fit beyond the peaks in the lane which are of interest for quantitative analysis. It is our practice to fit *at least* five peaks beyond (preferably 10 peaks beyond) the region for which reliable fit parameters are sought. Further, peaks at the top of the fit are not always well behaved because of the artificially-defined boundary. Fixing the Lorentzian widths and positions at reasonable values for the top three peaks in a fit solves this problem.

The left and right boundaries of the fit define the width of the lane (in pixel slices) over which the fit will be performed. The pixel-width of images obtained for different lanes (even within the same gel) can differ because of variations in the shapes of wells or in the amount of salt in a lane. In order to compare all the data in one lane to all the data in another, then, it is necessary to fit from one edge of the lane to the other. The left and right boundaries of our fits were chosen on the basis that, in a single

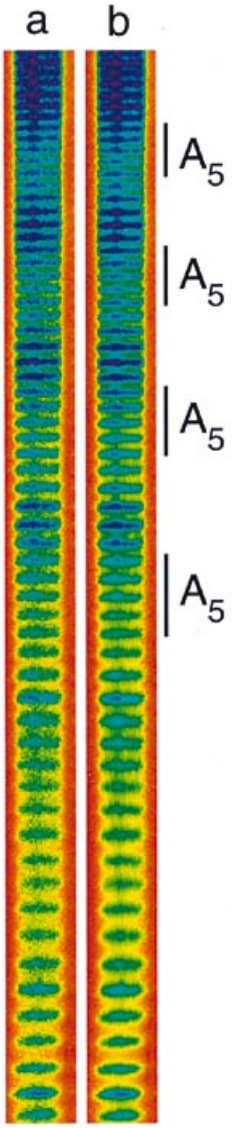


Figure 5. (**a**) Image of a lane of electrophoresis data showing the hydroxyl radical cleavage pattern of a restriction fragment containing a series of phased A-tracts. The image is generated from an 8% denaturing polyacrylamide gel exposed to an imaging phosphor plate. (**b**) Image of the sum of Lorentzian curves fit to the data. The four A_5 sequences, which show an attenuation of cleavage, are highlighted.

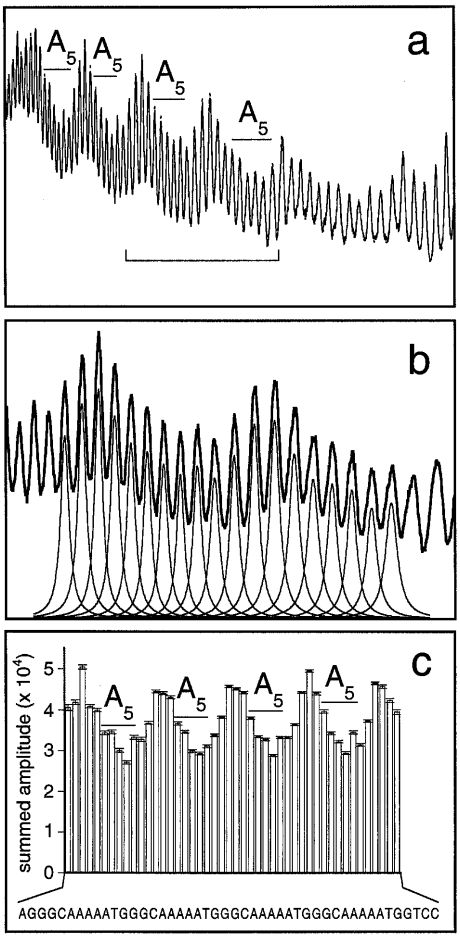


Figure 6. (**a**) Linegraph comparison of average data (—) and average GelExplorer fit (----) to the data for the A-tract hydroxyl radical cleavage data shown in Figure 4. The four A_5 sequences, which show an attenuation of cleavage, are highlighted. (**b**) A subset [designated in (a)] of the linegraph of the average data (heavy line) and the average Lorentzian contributions to the fit (light line). (**c**) Histogram of individual average peak amplitudes summed across the width of the lane for each peak. Error bars reflect the uncertainty calculated by the fitting procedure.

slice at the left or right extreme of the lane, the peak shapes in the data must be visibly discernible above the noise level. In practice, GelExplorer has difficulty converging if peaks are not obviously above the noise. At the edges of a lane, the intensities of bands drop off gradually, and at the boundaries chosen by our criteria, the fitted amplitudes of peaks are approximately one-half of the maximum observed for peaks in slices in the central part of the lane. To appropriately compare fitted amplitudes of bands in different lanes it is important to compare the sum of amplitudes for a given peak over all slices in the fit. Average amplitudes are less appropriate for comparison because the weight accorded to the data in a particular slice depends on how many slices are included.

For comparisons of peak amplitudes within a lane, it is important that the lane of data be approximately the same width over its entire length. While most lanes are somewhat wider at the top than in the most well-resolved region, the difference must be minimal to ensure that all of the data in each peak in the lane is being reflected in the fit results.

In some cases, the extreme right or left slice in the fit did not result in a well-behaved fit (e.g., peaks had unreasonable widths) and had to be excluded from the total summed result. Further, it may be possible to apply the above criteria for choice of the right and left fit boundaries and arrive at a slightly different choice of limits. We have found that omission of one slice on either the left or the right of the lane introduces an average variation in the summed amplitudes (over all peaks in a lane) of 0.6–1%, depending on the lane tested. A conservative estimate of the degree to which the choice of left and right pixel boundaries might be different is four slices (two on each edge). This introduced a variation in the resultant amplitudes of 2.6–3.8%, depending on the lane tested.

Different sets of reasonable starting parameters (e.g., positions chosen several times, different starting widths) were used for a series of fits and the fitted amplitudes were compared. Generally, the variation in amplitude was <0.5%, with a few peaks in each fit varying as much as 1–2%. The highest variations were generally observed for the least well-resolved peaks in the lane.

Convergence is defined such that the χ^2 function of the optimization changes by less than the tolerance for n successive iterations. Our fitting experiments have shown that the results are virtually independent of the tolerance, and that most large changes in parameter values occur within three iterations. All fits were thus performed with a tolerance of 0.1 and three iterations. Since identical criteria are applied, we have assumed the convergence criteria do not contribute significant uncertainty to comparisons between lanes.

In reporting optimized peak amplitudes we have added, to the uncertainty calculated by the fitting routine, additional uncertainties due to the choice of fitting boundaries of 3% and the choice of starting parameters of 1% to the fit uncertainties. These errors are carried through in analyses which utilize the fitting results.

The uncertainty in the optimized peak positions is <0.03% as calculated by the fit and varies by <0.05% for fits performed with different starting parameters. Values for peak widths are related to amplitudes and thus vary in a similar manner.

Quantitative determination of λ cI repressor binding to the O_R1 site

One of the most important applications requiring quantitative analysis of electrophoresis band intensities is the determination of a protein–DNA binding constant from a footprint titration experiment (1). In this experiment, a DNA-bound protein protects the backbone of the DNA from cleavage by hydroxyl radical (33), or other cleaving agent (1). A series of reactions are performed in which protein concentrations are varied systematically. The relative amount of cleavage at a particular nucleotide position is a measure of the fraction of DNA molecules that do not have protein bound. The protein concentration-dependent protection can be analyzed to obtain thermodynamic information for protein binding (1). We have generated a hydroxyl radical footprint titration of the λ cI repressor bound to the O_R1 operator sequence. This is a very well-understood system (1,26–28,34,35) and provides a clear demonstration of the utility of GelExplorer for quantitative analysis.

The titration experiment was conducted over 22 protein concentration points. Images of the data for the hydroxyl radical reference lane (reaction without protein) and the hydroxyl radical footprint lane containing the highest concentration of repressor are shown in Figure 7a and c, respectively. Figure 7b and d show the images of the 60-peak fits to these data, respectively. The average linegraphs of the data and fit (across the width of the lane) for the reference and footprint data are shown in Figure 7e and f, respectively. The footprint shows three regions of protection by the protein, labeled a′, b′ and c′ (28). The regions a′ and b′ correspond to the edges of the major groove within which the repressor is thought to make sequence specific contacts with the DNA bases of the operator (36–39). The exact nature of the protection in the c′ region, which is across the minor groove from the main binding region of the protein, has not been elucidated. The fitting procedure clearly reproduces the data very well. The peak numbers used in the fit, which correspond to sequences which show protections, are summarized in Table 1.

In a thermodynamic analysis of protein binding, the amplitudes of gel bands serve as a reflection of relative rates of cleavage (or relative degrees of protection). The fitted amplitudes in each lane must be normalized before reliable comparisons of peak intensities can be made. The amplitudes for a set of 18 peaks (10 peaks below the footprint and 8 peaks above the footprint) in each lane were summed. These peaks were chosen on the basis that they showed very little variation over the series of 22 lanes. For each lane, the amplitude of each peak was multiplied by a factor such that the summed amplitudes for the normalizing peaks had the same value as that for the reference lane.

The normalized amplitudes were converted to fractional protection (p_i) according to equation **2**, where $A_N(n,\text{site})$ is the normalized amplitude of nucleotide n from a lane containing protein, and $A_N(n, \text{ref})$ is the normalized amplitude of nucleotide n from the reference lane.

$$p_i = 1 - \frac{A_N(n, \text{site})}{A_N(n, \text{ref})} \qquad \mathbf{2}$$

Fractional protections were converted to fractional saturations Y as has been described for other footprint titration analyses (1,27).

The relationship between the fractional protection at a given nucleotide and the protein concentration is the binding isotherm. Protein binding constants are obtained by fitting the Langmuir expression, given in equation **3**, to each nucleotide's binding isotherm. The microscopic equilibrium binding constant is k and [P] is the concentration of unliganded protein, active to bind DNA.

$$Y = \frac{k[\text{P}]}{1 + k[\text{P}]} \qquad \mathbf{3}$$

In contrast to analyses which quantify protections by integrating the intensity of bands in a rectangle drawn around the entire binding site, thereby obtaining a single binding constant for the entire site, our approach provides quantitative binding isotherm curves for each nucleotide individually.

Representative fits to binding isotherms for footprinted nucleotide positions are shown in Figure 8. There are 12 positions, including all of the positions previously reported to show protection from hydroxyl radical cleavage (28), having binding isotherms which could be fit to obtain binding constants and free energy of binding (ΔG). The results for the different nucleotide positions range from $\Delta G = -11.4 \pm 0.3$ to -12.7 ± 0.4 kcal/mol and are summarized in Table 1. These results are in good agreement with the previously reported value of –12.6 kcal/mol for this system (1). An analysis was also performed for the sum of all the normalized amplitudes of peaks within the footprint region (peaks 23–45). Fits to the single isotherm gave $\Delta G = -11.6 \pm 0.2$ kcal/mol. For individual nucleotides, no systematic variations were observed in ΔG values for the different regions of the footprint in our data. In particular, nucleotides in footprint region c′ exhibit titration behavior similar to that of the main portion of the footprint (see Table 1).

Table 1. Nucleotides protected from hydroxyl radical cleavage by λ repressor

Sequence label[a]	Footprint region	Peak[b]	ΔG (kcal/mol)[c]
T*	c′	23	–11.7 ± 0.2
A*		24	–12.0 ± 0.2
G*	a′	32	–11.4 ± 0.3
A*		33	–12.5 ± 0.3
C*		34	–12.7 ± 0.4
C*		35	–11.5 ± 0.2
G		36	–11.8 ± 0.2
C		37	–11.6 ± 0.2
A*	b′	42	–11.5 ± 0.3
T*		43	–11.6 ± 0.2
T*		44	–11.7 ± 0.2
A		45	–12.3 ± 0.2

[a]An asterisk denotes positions for which protection from hydroxyl radical cleavage by λ repressor has been previously reported.
[b]Peak numbers used in fits to the data.
[c]Determined from the footprint titration data as described in the text.

DISCUSSION

In this paper, we have described GelExplorer, which uses a new methodology for obtaining quantitative information about electrophoretic band intensities. The program uses a novel two-dimensional curve fitting approach to deconvolute band intensities and to account for variations across the width of a lane of electrophoresis data. The Lorentzian lineshape has been demonstrated to successfully model electrophoresis bandshapes and is appropriate for use in curve fitting analysis. Because reasonably close initial parameter values are of vital importance for a successful nonlinear least-squares optimization, the program is designed to provide an excellent set of starting parameters.

High quality data are required for successful fitting by GelExplorer. In particular, the signal-to-noise ratio for the data must be very good because curves are optimized to a single slice of data for which no averaging or smoothing has been applied to reduce noise. As a result, faint bands, which can be difficult to fit, may require additional criteria in the choice of starting Lorentzian parameter values or fixing of some parameters. Further, lanes must be reasonably straight, as curved lanes are not easily treated. The fitting method is generally not limited by the resolution or separation of bands, and is not limited by variability in band shape or lane width.

Because GelExplorer consists of a set of modules linked together in maps, it is very flexible. It is currently equipped to read images generated by ImageQuant™ software. Expansion or adaptation of the program for use with images from other

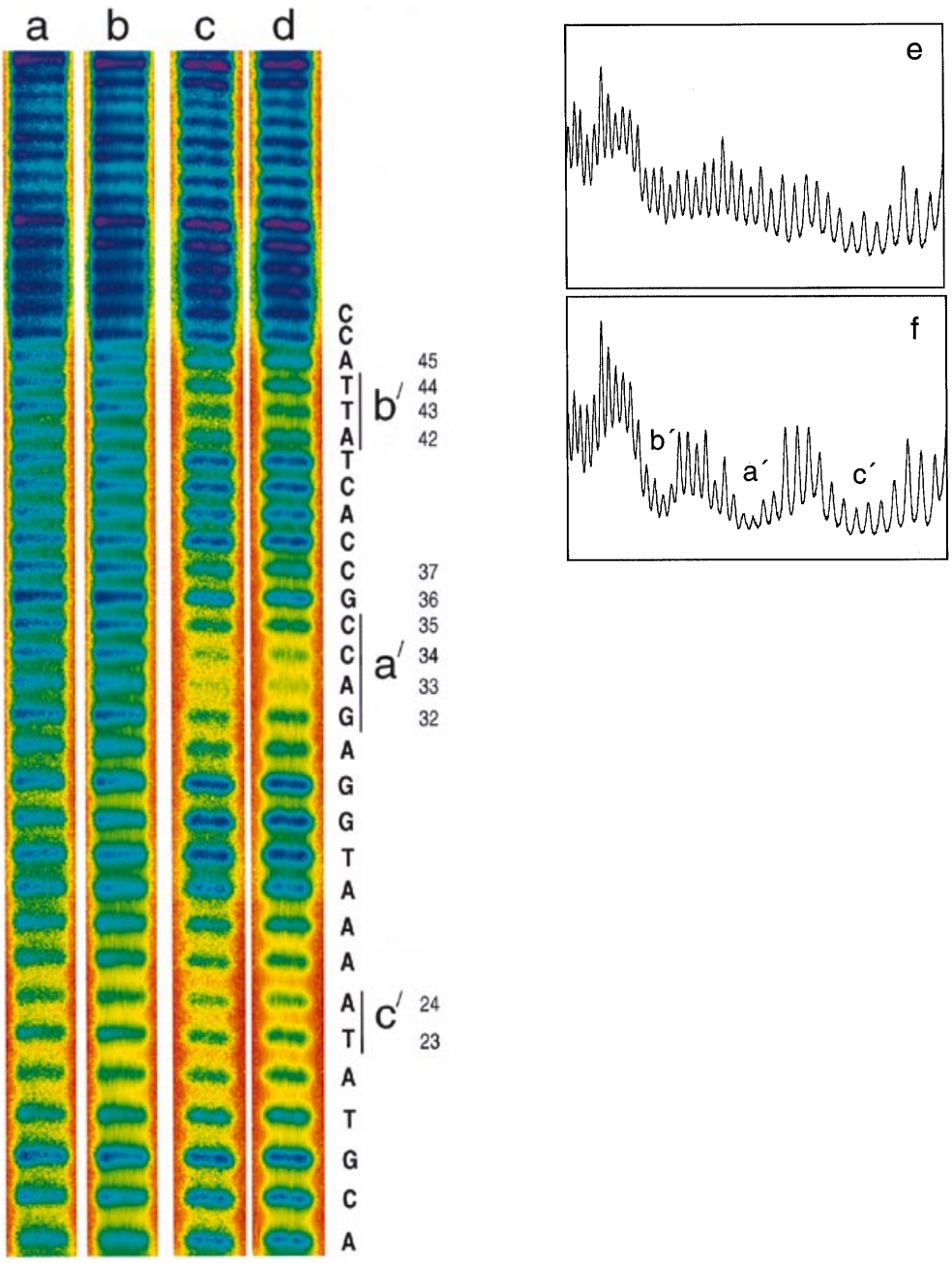


Figure 7. (**a**) Image of a subset of the lane of electrophoresis data showing the hydroxyl radical cleavage pattern of a restriction fragment containing the O_R1 binding site. The image is generated from an 10% denaturing polyacrylamide gel exposed to an imaging phosphor plate. (**b**) Image of the sum of Lorentzian curves fit to the data in (a) by GelExplorer. (**c**) Image of a subset of the lane of electrophoresis data showing the hydroxyl radical cleavage pattern of the restriction fragment containing the O_R1 binding site to which was bound λ repressor at saturating concentration. The image is generated from a 10% denaturing polyacrylamide gel exposed to an imaging phosphor plate. (**d**) Image of the sum of Lorentzian curves fit to the data in (c) by GelExplorer. The nucleotide sequence is shown at the right. The footprints of λ repressor are labeled a′, b′, and c′, corresponding to the original footprint designations in ref. 28. These footprints are indicated by vertical lines. Peak numbers shown at the right correspond to the numbering in Table 1. (**e**) Linegraph comparison of average data (—) and average GelExplorer fit (----) to the data for the data shown in (a) and (b). (**f**) Linegraph comparison of average data (—) and average GelExplorer fit (----) to the data for the data shown in (c) and (d).

programs, for special applications, or for additional analysis of fitting results, is easily accomplished by the user within the IRIS Explorer™ programming environment.

GelExplorer includes a calculation of uncertainties in the fit parameter outputs, allowing the reliability of the fitting results to be evaluated. In addition, other sources of uncertainty in the fitting results have been evaluated. The overall uncertainty in the fit results is very small, yielding well-determined values for peak intensities and positions. The uncertainties reported here have been evaluated for fitting of relatively low percent (6–10%) polyacrylamide sequencing gel data. Fits to other types of

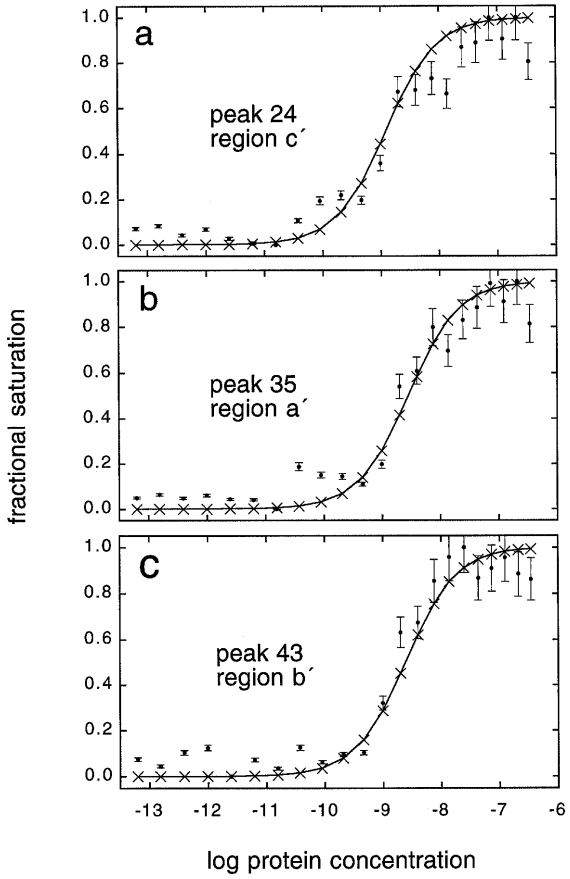


Figure 8. Fits (—×—) of the Langmuir equation to the binding isotherm data (●●●) of selected individual nucleotides in the λ repressor footprint titration. Peak numbers and calculated ΔG values are listed in Table 1. (**a**) Peak 24; (**b**) peak 35; (**c**) peak 43. Error bars reflect the uncertainty in the fractional saturation as calculated using uncertainties introduced by the GelExplorer methodology.

electrophoresis data will require additional experiments to evaluate uncertainties in the results.

The application of the program to the determination of free energy of binding of λ repressor to the O_R1 binding site has demonstrated the utility of GelExplorer for quantitative analysis. The values for ΔG obtained from our analysis are in very good agreement with those previously reported. The observed differences likely result from lower protein activity and/or concentration than those used in the analysis presented here, which is likely given the age of the protein sample (>6 years). For example, a decrease in the concentration of protein active to bind DNA would result in a more negative value of ΔG at each nucleotide. Importantly, the analysis demonstrates the level of detail which can be reliably obtained from this type of analysis. In particular, the protection pattern in region c′ of the footprint has titration behavior similar to the other regions of the footprint, which demonstrates that the observed c′ protections are related to the same protein binding event as causes protections in regions a′ and b′. These results are possible only as a result of high resolution, quantitative curve fitting analysis. Details such as these, at the level of individual nucleotide binding, will provide further insight into protein–DNA binding events. We expect that the methodology employed by GelExplorer will prove successful in the analysis of a wide variety of problems requiring quantitative analysis of electrophoresis data.

PROGRAM AVAILABILITY

GelExplorer software is available upon request from Prof. Tom Tullius by anonymous file transfer protocol (FTP). For users from academic institutions there is no charge to obtain the program. However, users must be licensed to use IRIS Explorer™ version 3.0 (29). Detailed instructions for the use of GelExplorer are described in an on-line manual, which is available at http://dna.chm.jhu.edu.

ACKNOWLEDGEMENTS

This research was supported by PHS National Research Service Award Grant 5 F32 GM 16828-02 (S.E.S.) and by PHS grant GM 40894 (T.D.T). We gratefully acknowledge the use of phosphorimagery instrumentation maintained by the Institute for Biophysical Research on Macromolecular Assemblies at Johns Hopkins, which was supported by an NSF Biological Research Centers Award (DIR-8721059) and by a grant from the W.M. Keck Foundation. We thank Ruth M. Ganunis for generation of the A tract hydroxyl radical cleavage electrophoresis data, Lori M. Ottinger for construction and purification of plasmid DNA containing the O_R1 binding site, Prof. Gary Ackers for the λ repressor protein, and Dr Michael Brenowitz for the NONLIN fitting program and instruction in its use. We also appreciate help from John Chandler, Computer Science Department, Oklahoma State University, in implementing the least-squares error analysis in GelExplorer.

REFERENCES

1 Brenowitz, M., Senear, D. F., Shea, M. A., and Ackers, G. K. (1986) *Methods Enzymol.* **130,** 132–181.
2 Morrison, T. B., and Parkison, J. S. (1994) *Biotechniques* **17,** 922–926.
3 Stankus, A., Goodisman, J., and Dabrowiak, J. C. (1992) *Biochemistry* **31,** 9310–9318.
4 Jacot-Descombes, A., Todorov, K., Hochstrasser, D. F., Pellegrini, C., and Pun, T. (1991) *Comput. Appl. Biosci.* **7,** 225–232.
5 Gray, A. J., Beecher, D. E., and Olson, M. V. (1984) *Nucleic Acids Res.* **12,** 473–491.
6 Smith, J. M., and Thomas, D. J. (1990) *Comput. Appl. Biosci.* **6,** 93–99.
7 Galat, A., Serafinowshi, P., and Koput, J. (1984) *Biochim. Biophys. Acta* **801,** 40–47.
8 Galat, A. (1989) *Electrophoresis* **10,** 659–667.
9 Haselgrove, J. C., Lyons, G., Rubenstein, N., and Kelly, A. (1985) *Anal. Biochem.* **150,** 449–456.
10 Pulleyblank, D. E., Shure, M., and Vinograd, J. (1977) *Nucleic Acids Res.* **4,** 1409–1418.
11 Vohradsky', J., Mares ová, H., Vec erek, B., and Kyslík, P. (1990) *Methods Mol. Cell Biol.* **1,** 223–233.
12 Horgan, G. W., and Glasbey, C. A. (1995) *Electrophoresis* **16,** 298–305.
13 Hansen, P. K., Christensen, J. H., Nyborg, J., Lillelund, O., and Thøgersen, H. C. (1993) *J. Mol. Biol.* **233,** 191–202.
14 Lutter, L. C. (1978) *J. Mol. Biol.* **124,** 391–420.
15 Ribeiro, E. A., and Sutherland, J. C. (1993) *Anal. Biochem.* **210,** 378–388.
16 Bashkin, J. S., and Tullius, T. D. (1993) in Revzin, A. (ed.), *Footprinting of Nucleic Acid–Protein Complexes* . Academic Press, San Diego, pp. 75–106.
17 Vohradksy′, J., and Pánek, J. (1993) *Electrophoresis* **14,** 601–612.
18 Galat, A., and Goldberg, I. (1987) *Comput. Appl. Biosci.* **3,** 333–338.
19 Smith, J., and Singh, M. (1996) *BioTechniques* **20,** 1082–1087.
20 Maniatis, T., Fritsch, E. F., and Sambrook, J. (1989) *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
21 Maxam, A. M., and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.

22 Price, M. A., and Tullius, T. D. (1993) *Biochemistry* **32,** 127–136.
23 Price, M. A., and Tullius, T. D. (1992) *Methods Enzymol.* **212,** 194–219.
24 Tullius, T. D., and Dombroski, B. A. (1985) *Science* **230,** 679–681.
25 Koblan, K. S., and Ackers, G. K. (1991) *Biochemistry* **30,** 7817–7821.
26 Senear, D. F., Brenowitz, M., Shea, M. A., and Ackers, G. K. (1986) *Biochemistry* **25,** 7344–7354.
27 Brenowitz, M., Senear, D., Jamison, E., and Dalma-Weiszhausz, D. (1993) In Revzin, A. (ed.), *Footprinting of Nucleic Acid–Protein Complexes.* Academic Press, San Diego, pp. 1–44.
28 Tullius, T. D., and Dombroski, B. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 5469–5473.
29 IRIS Explorer™ can be obtained from the Numerical Algorithms Group (http://www. nag.com).
30 Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1990) *Numerical Recipes in C*; Cambridge University Press, New York.
31 Johnson, M. L., and Frasier, S. G. (1985) *Methods Enzymol.* **117,** 301–342.
32 Burkhoff, A. M., and Tullius, T. D. (1987) *Cell* **48,** 935–943.
33 Dixon, W. J., Hayes, J. J., Levin, J. R., Weidner, M. F., Dombroski, B. A., and Tullius, T. D. (1991) *Methods Enzymol.* **208,** 380–413.
34 Ackers, G. K., Shea, M. A., and Smith, F. R. (1983) *J. Mol. Biol.* **170,** 223–242.
35 Senear, D. F., and Brenowitz, M. (1991) *J. Biol. Chem.* **266,** 13661–13671.
36 Beamer, L. J., and Pabo, C. O. (1992) *J. Mol. Biol.* **227,** 177–196.
37 Pabo, C. O., and Sauer, R. T. (1984) *Annu. Rev. Biochem.* **53,** 293–321.
38 Pabo, C. O., and Lewis, M. (1982) *Nature* **298,** 443–447.
39 Lewis, M., Jeffrey, A., Wang, J., Ladner, R., Ptashne, M., and Pabo, C. O. (1983) *Cold Spring Harbor Symp. Quant. Biol.* **47,** 435–440.

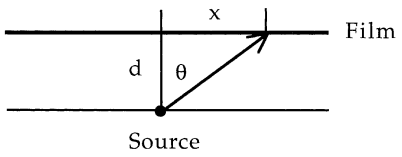
APPENDIX

Lorentzian lineshapes are intrinsic to autoradiographic detection

Jeremy M. Berg

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Consider the detection of a point radiation source using autoradiography. The source will emit radiation in all directions with equal probability. For the emission and detection of each photon, the geometrical arrangement shown below applies:



Scheme 1.

Here, d is the distance between the source and the image plate or film (hereafter, the term film will be used), θ is the angle between the normal to the film and the emitted radiation, and x is the distance from the point directly above the source to the point at which the radiation strikes the film. The quantities are related by

$$x = d\tan\theta.$$

Since emission at any angle θ is equally likely but x increases more rapidly at larger values of θ, the density of detected radiation as a function of x will be proportional to the inverse of the rate of change of x with respect to θ.

$$\frac{dx}{d\theta} = \frac{d}{\cos^2\theta}$$

$$\text{But, } \cos\theta = \frac{d}{\sqrt{x^2 + d^2}} \text{ so that } \cos^2\theta = \frac{d^2}{x^2 + d^2}$$

$$\text{Thus, } \frac{dx}{d\theta} = \frac{x^2 + d^2}{d}$$

and the density on the film will be give by

$$\rho + N\frac{1}{\frac{dx}{d\theta}} = N\frac{d}{x^2 + d^2} \text{ where N is a scale factor.}$$

Thus, ρ, the density of radiation detection as a function of x, will be Lorentzian with full width at half height of 2d, corresponding to C =N/2 and γ = 2d in equation (1) in the paper. Therefore, with a point radiation source, the lineshape observed by autoradiography will be Lorentzian. For detection of electrophoretically-generated bands, the observed lineshape will be generated by convoluting these Lorentzian lines with the distribution of radioactive material in the gel, most likely a Gaussian or skewed Gaussian. A simulation demonstrates that a Lorentzian gives a good fit to a Gaussian distribution of Lorentzian lines. Bandwidths will also be affected by the distance from the gel and the image plate or film.