

Robust Estimation of Standard Curves for Protein Molecular Weight and Linear-Duplex DNA Base-Pair Number after Gel Electrophoresis

BRIAN D. PLIKAYTIS,* GEORGE M. CARLONE,†¹ PAUL EDMONDS,†
AND LEONARD W. MAYER†

*Statistical Services Activity, and †Molecular Biology Laboratory, Meningitis and Special Pathogens Branch, Division of Bacterial Diseases, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333

Received June 10, 1985

An accurate procedure for estimating linear-duplex DNA base-pair numbers and protein molecular weights after electrophoresis in single concentration gels is presented. A robust modified hyperbola was found to be superior for determining molecular weights and base-pair numbers for a set of known standards when compared with the conventional log transformation and a similar hyperbolic model. We describe the use of a soft laser-scanning densitometer to measure band-migration distances of wet, stained polyacrylamide gels for proteins and photographic negatives of agarose gels containing DNA stained with ethidium bromide. This automated densitometric method was more accurate than existing methods. A BASIC computer program detailing the procedure is included. © 1986 Academic Press, Inc.

KEY WORDS: standard curve; molecular weights; protein; DNA; gel electrophoresis; densitometry.

One of the first quantitative gel electrophoresis methods for determining optimal separation, resolution, and resultant extrapolation of apparent molecular weights was described by Ferguson (1). The Ferguson plot analysis, whose theoretical basis was defined by Ogston (2) and studied by others (3,4), requires electrophoresis using several different gel concentrations since no single concentration is optimal for all electrophoretic separations (5,6). This is the only electrophoretic technique currently available for determining molecular weights of proteins with unequal surface charge densities. The analysis is usually applied to sample mixtures containing only a few components of interest (5), and, although extremely reliable, some investigators have reported nonlinearity in the plots (3), emphasizing the fact that no one method for molecular weight determination is universally ideal.

Another method for determining molecular weights of proteins with equal surface charge

densities after gel electrophoresis is to compare band-migration distances with a set of known standards in a single concentration gel (3,6-8). Standard sets, which include a broad range of protein molecular weights or DNA base pairs, frequently exhibit a curvilinear relationship between the migration distances of the bands and their molecular weights or base-pair numbers in single concentration gels (9-11). One way to examine these data is to form a standard curve by fitting a simple linear model relating the logarithm of the molecular weights or base-pair numbers to the migration distances of the various bands (7,8,12,13). When band-migration distances are measured for materials in a sample run, their respective molecular weights or base-pair numbers are then interpolated from this curve.

This approach is subject to errors from a variety of sources. Numerous tools such as rulers, calipers, and magnifying glasses used to measure the migration distances are inherently imprecise (5). Potentiometric (5) and densitometric (14) measurements have been

¹ To whom reprint requests should be addressed.

used to increase the precision of these measurements and relative mobilities (R_f) are generally used to reduce measurement inaccuracies within a series of gels (8,15). The log transform, however, often does not fully linearize the data over the entire range of the observations (5,8,15,16). Although measures of fit, such as the Pearson correlation coefficient (17), may indicate that the log transformation works well, comparisons of the known molecular weights or base-pair numbers with their predicted values from the standard curve may reveal the percentage error to be unacceptably large. The inadequacy of the log transformation cannot be clearly visualized without considering graphs of the data with the resultant fitted line and a table of known and predicted values of molecular weights or base-pair numbers derived from the standard curve. When the data are clearly nonlinear over the full range of observations, investigators will commonly fit the simple linear model to just that portion of the data that are strictly linear and extend the line through the remaining points by hand (8). This method suffers in that it is difficult to accurately predict molecular weights or base-pair numbers outside the range of linearity and those predictions are inevitably subjective and unreliable.

Investigators have introduced several models to describe the inherent curvature of protein and DNA gel electrophoresis data. Rodbard postulated a sigmoidal or logistic relationship relating the relative mobility of proteins to their molecular weight (3,6). In addition to the logistic function, Rodbard outlined a number of alternative models to describe these data including a class of models relating log molecular weight to simple linear and polynomial functions in relative mobility and another class of models relating molecular weight to simple linear and polynomial functions in log relative mobility. Duggelby *et al.* (18) used a polynomial function in log base-pair number to describe the relative mobility of DNA fragments and Parker *et al.* (19) used an exponential polynomial in relative mobility to describe DNA base-pair numbers. Schaffer

and Sederoff (8,20) implemented a program to fit a hyperbola described by Southern (21) to determine DNA fragment length. These procedures account, to varying degrees, for the inherent curvature of the data and achieve greater degrees of accuracy than the simple log transformation when dealing with protein molecular weights and the base-pair number of linear-duplex DNA.

In any experimental system there are occasional readings which do not fit the general trend of the data. These aberrant observations, or outliers, can severely affect the outcome of the log transformation and any other procedure which uses the method of least squares to describe the data (6). Tiede and Pagano (22) addressed these difficulties when dealing with the logit model used to form calibration curves related to radioimmunoassay data. They fit a more general modified hyperbola to their data, which resembles the one proposed by Southern, but is more accurate over a wider range of applications. They also implemented a robust fitting technique which is not sensitive to outliers and minimizes the effect of these aberrant observations on the overall fit of the model.

We propose the application of a robust modified hyperbola to develop standard curves of gel electrophoresis data from single concentration gels for proteins and linear-duplex DNA after precisely measuring band-migration distances with a soft laser-scanning densitometer.

DERIVATION

Standard curves for proteins and DNA have traditionally been formed by fitting a simple linear model relating the logarithm of the molecular weights or base-pair numbers to the migration distances of the various bands in the standard

$$y_i = a + b(x_i) + e_i$$

$$i = 1, \dots, n \text{ (the number of data pairs)}$$

where y_i = log of the molecular weights or DNA base-pair numbers, x_i = migration dis-

tances of the respective bands, and e_i = error term associated with the fit. After band-migration distances in sample runs are measured, their respective molecular weights or base-pair numbers are interpolated from this curve, given the calculated slope (b) and y -intercept (a) terms. However, the following nonlinear function describes the inherent curvature of the data more accurately than the log transformation:

$$f(x_i, \mathbf{O}) = y_i = a + b/(1 + c(x_i)^d) + e_i \quad [1]$$

$$i = 1, \dots, n$$

where y_i is the known molecular weight or number of DNA base pairs; x_i is the migration distance of the proteins or DNA fragments in the gel measured by a soft laser-scanning densitometer; e_i is the error term associated with the fit; and \mathbf{O} is the vector of parameters, a , b , c , and d of the model to be estimated.

When parameterized in this fashion, the model coefficients may be interpreted as follows: the first coefficient, a , is an estimate of the horizontal asymptote of the curve or that point on the y axis (molecular weight or DNA base-pair number) where the hyperbola would become horizontal if the curve were extended indefinitely. The coefficient b is the point on the y axis where the hyperbola would become vertical, again, if the curve were extended indefinitely. The coefficients c and d are measures of curvature, determining the sharpness of the curve, etc.; d was generally between 0.25 and 1.5 in our trials.

The band-migration distances were normalized to eliminate the variability that exists with each experimental run. Variability associated with terminating the electrophoretic run at a fixed distance causes the bands to be offset by a small distance. After each run the densitometer scaled the migration distances to range from 0 to some maximum value set by the operator (the maximum value is the length of the graph produced by the densitometer in inches). The migration distances were preprocessed by first dividing each of the distances by the graph length so that they would range from 0.0 to 1.0:

$$x_{\text{standard revised}}$$

$$= \text{migration distance/graph length}$$

where $x_{\text{standard revised}}$ = revised calculated migration distances of the standard. This technique is also applicable to data collected by other means if the R_f value (the migration distance of the protein or DNA fragment divided by the migration distance of a tracking dye) is used in place of $x_{\text{standard revised}}$ in the following derivation.

The shortest distance (corresponding to the highest molecular weight or base-pair number in the standard) was then subtracted from all the values to correct for the problem of band-distance offset. A correction factor of 0.01 was added to each of the readings to avoid a 0 migration distance for the band associated with the heaviest molecular weight or longest DNA fragment. Each value was then multiplied by 100 to give readings which ranged from 1.00 to 101.0:

$$x_{\text{normalized}} = ((x_{\text{standard revised}}$$

$$- \text{minimum}(x_{\text{standard revised}})) + .01) \times 100)$$

where $\text{minimum}(x_{\text{standard revised}})$ is the shortest migration distance of the standard after having been divided by the graph length, and $x_{\text{normalized}}$ is the final normalized migration distance values used in Eq. [1] to solve for the four parameters a , b , c , and d . If the beginning of the gel is thought of as the position of the first band (i.e., the heaviest molecular weight or the longest DNA fragment), then each normalized value can be thought of as the approximate percentage of the distance traveled by the respective band to the total distance of the gel.

Once the parameters a , b , c , and d have been estimated, molecular weights or base-pair numbers may be interpolated from the standard curves after the sample run data have been preprocessed in the same way as the standard data. Since the sample runs are obtained from the same gel as the standard run, the offset for the sample runs must still be the minimum distance of the standard run. After

the sample run migration distances have been divided by the graph length, the final values are obtained as

$$x_{\text{sample final}} = (((x_{\text{sample revised}} - \text{minimum}(x_{\text{standard revised}})) + .01) \times 100)$$

where $x_{\text{sample revised}}$ = revised calculated migration distances of the sample run (i.e., the original readings divided by the graph length), and $x_{\text{sample final}}$ is the final calculated migration distance values used in Eq. [1] to find the molecular weights or DNA base-pair numbers after the parameters a , b , c , and d have been estimated.

Ordinary least squares may be unduly influenced by outliers in the data while the robust procedure produces estimates which minimize the effect of these outlying observations. In the absence of outliers, the robust fit will be similar to the least squares results.

Given the function $y_i = f(x_i, \mathbf{O})$, \mathbf{O} , the vector of parameters, a , b , c , and d , may be estimated using the Taylor series expansion or linearization method (23). It can be shown that the vector of parameters \mathbf{O} may be expressed in the form of a pseudo-weighted least-squares estimate

$$\mathbf{O}_{j+1} = \mathbf{O}_j + (\mathbf{A}'_j \mathbf{W}_j \mathbf{A}_j)^{-1} \mathbf{A}'_j \mathbf{W}_j \mathbf{R}_j \quad [2]$$

where \mathbf{O}_j is the former estimate and \mathbf{O}_{j+1} is the revised estimate for the parameter vector; \mathbf{A} is the matrix of partial derivatives of the function with respect to each of the parameters, evaluated at \mathbf{O}_j .

Setting A_{ip} equal to the partial derivative of $f(x, \mathbf{O})$ with respect to O_p , where $p = 1, \dots, 4$ for a, b, c , and d , evaluated at the normalized migration distance x_i , then the matrix \mathbf{A} is

$$\begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} \\ \vdots & \vdots & \vdots & \vdots \\ A_{i1} & A_{i2} & A_{i3} & A_{i4} \\ \vdots & \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & A_{n4} \end{matrix} \quad \begin{matrix} i = 1, \dots, n \\ \text{observations} \end{matrix}$$

It follows that \mathbf{A}_j equals the matrix \mathbf{A} evaluated at the former estimates for $\mathbf{O} = \mathbf{O}_j$ (i.e., $a, b,$

c , and $d = a_j, b_j, c_j$, and d_j). The partial derivatives are defined as

$$A_{i1} = 1$$

$$A_{i2} = (1 + cx_i^d)^{-1}$$

$$A_{i3} = -bx_i^d(1 + cx_i^d)^{-2}$$

$$A_{i4} = -bx_i^d c (\log(x_i))(1 + cx_i^d)^{-2}$$

Examining the fourth partial derivative, it is clear why the small correction factor was added to the migration distances to avoid a 0 normalized distance for the band associated with the heaviest molecular weight or longest DNA fragment; the log of 0 is undefined.

\mathbf{R}_j are the residuals of the function (i.e., the predicted weights or DNA base-pair numbers derived from the function subtracted from the actual molecular weights or DNA base-pair numbers), again evaluated at $\mathbf{O} = \mathbf{O}_j$:

$$\begin{matrix} R_{1j} \\ \vdots \\ R_{nj} \end{matrix}$$

The summation of the squares of the weighted residuals, that is, the summation of the product of the squared residuals with the j th weights associated with each point, is one measure of how well the present set of parameters, \mathbf{O}_j , describe the data; the best fit will tend to minimize this sum. When this statistic is computed for the nonrobust and unweighted fitting procedures, the weights are all assumed to be equal to 1.0.

Weighting functions, as described by Rodbard, correct for nonuniformity in the variances of the molecular weights (3,6), whereas the weighting function, \mathbf{W} , incorporated in the present robust procedure is used for the detection and elimination of outlying observations (22,24). These weights determine the "robustness" of the procedure. This factor would simply be excluded from expression [2] for the least-squares case. For the robust fit, we chose the SINE estimator (24) which results in an $n \times n$ weight matrix composed of 0's with the following factors along the diagonal elements:

$$sc/R_{ij}[\sin(R_{ij}/sc)] \quad \text{if } |(R_{ij}/sc)| \leq \pi$$

or

$$0.0 \quad \text{if } |(R_{ij}/sc)| > \pi$$

The sine function is performed in radian mode with π equal to 3.141593, $| \quad |$ indicates absolute value, s is a robust estimate of scale, and c is a constant which determines how sensitive the fit will be to the outlying observations. This sensitivity increases as c decreases. The value for c is set equal to 2.1 and s is defined as

$$\text{median}(\text{largest}(n-3)|\text{residuals}|).$$

The weights range from 0.0 to 1.0 corresponding to large or small residuals, respectively. Data points which are close to the fitted line (i.e., points with small residuals) will have weights close to 1.0, whereas outlying observations will typically have weights near 0.0.

The solution for \mathbf{O} can be obtained using the standard Gauss-Newton least-squares algorithm (23) which is detailed in the BASIC program listing and documentation in the appendix. Initial starting values for the parameter vector \mathbf{O} are substituted for \mathbf{O}_j in expression [2]. The matrices \mathbf{A}_j , \mathbf{W}_j , and \mathbf{R}_j are then calculated using these initial estimates. The new, revised estimates for \mathbf{O} (\mathbf{O}_{j+1}) are then calculated and used in the next round of calculations as \mathbf{O}_j . This iterative process continues until the solution converges, that is until \mathbf{O}_j and \mathbf{O}_{j+1} are extremely close:

$$|(O_{i(j+1)} - O_{ij})/O_{ij}| < e$$

where e is set to some small tolerance level (e.g., 0.00001).

As with all nonlinear estimation procedures initial parameter estimates are required to start the process. Schaffer and Sederoff (20) implemented an algorithm to find the solutions to the hyperbola described by Southern (21):

$$(x_i - m_0)(y_i - l_0) = h \quad [3]$$

where x_i is related to migration distance; y_i to DNA base-pair number, and, in our applications, protein molecular weight; and h is a constant of proportionality. Their solution for

the parameters m_0 , l_0 , and h provides good starting values for the iterative procedure. Fundamental algebra leads to the following relationship between the parameters a , b , c , and d of [1] and m_0 , l_0 , and h of [3]. If d is set equal to 1.0, then

$$a = l_0$$

$$b = -h/m_0$$

$$c = -1.0/m_0.$$

The present strategy is to find the solution to [3] using Schaffer and Sederoff's algorithm; solve for a , b , and c ; set $d = 1.0$; and use these results as starting values for the iterative process.

MATERIALS AND METHODS

*Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE).*² SDS-PAGE of proteins was carried out with a modified Laemmli gel system (25). A discontinuous buffer system was used with a 4% stacking and 11% separating gel (pH 6.8 and 8.8, respectively). The gels (1.5 × 180 × 160 mm) were cast with 12.0 ml of distilled water, 10 ml of acrylamide solution (30.0 g acrylamide, 0.95 g *N,N'*-methylenebisacrylamide, in 100 ml water), 7.5 ml of 1.5 M Tris, pH 8.8, and 300 μ l of 10% SDS; the gel solution was degassed for 10 min. Polymerization was initiated with 10 μ l of *N,N,N',N'*-tetramethylethylenediamine (TEMED) and 150 μ l of 10% ammonium persulfate. A 10-mm stacking gel was prepared by mixing 1.4 ml of acrylamide solution, 2.5 ml of 0.5 M Tris (pH 6.8), 100 μ l of 10% SDS, 5.9 ml of distilled water, 5 μ l of TEMED, and 100 μ l of 10% ammonium persulfate. Known molecular weight standards were diluted in 25 μ l of distilled water, and an equal volume of treatment buffer was added [20% glycerol, 10% 2-mercaptoethanol, 4% SDS, in 0.125 M Tris (pH 6.8), and 10 μ l of

² Abbreviations used: SDS-PAGE, sodium dodecyl sulfate-polyacrylamide gel electrophoresis; TEMED, *N,N,N',N'*-tetramethylethylenediamine; bp, base pairs.

0.05% bromphenol blue per milliliter of buffer]. Preparations were heated to 100°C and held for 5 min, cooled, and 25 μ l was loaded into each well of the gel and subjected to electrophoresis. A constant voltage of 150 V and initial current of 50 mA were used for electrophoresis. Each run took approximately 4.5 h for the tracking dye (bromphenol blue) to migrate within 1 cm of the bottom of the gel. The upper and lower chamber buffer consisted of 0.025 M Tris, 0.192 M glycine, and 0.1% SDS, pH 8.3. The buffer was cooled to approximately 5°C. Gels were stained overnight at room temperature with 0.125% Coomassie brilliant blue R-250 in a solution of 50% methanol and 10% acetic acid. After destaining, the wet gels were analyzed with a soft laser-scanning densitometer (Biomed Instruments, Fullerton, Calif.).

The following protein molecular weight standards were used: phosphorylase *b* (92.5K), bovine serum albumin (66.2K), ovalbumin (45.0K), carbonic anhydrase (31.0K), soybean trypsin inhibitor (21.5K), and lysozyme (14.4K). Glycerol, methanol, and acetic acid were obtained from Fisher Scientific, Fair Lawn, New Jersey, and all other reagents used were electrophoresis grade and obtained from Bio-Rad Laboratories, Richmond, California.

Gel electrophoresis of DNA. Agarose powder (FMC corporation, Rockland, Maine) was used to prepare horizontal gels (11 \times 13 \times 4 cm) of varying concentrations (0.85 to 1.0%) in Tris-acetate buffer (40 mM Tris, 40 mM sodium acetate, 2 mM EDTA, pH 8.2). One microliter (about 0.5 μ g) of a predigested linear-duplex DNA molecular weight marker (210 and 12216 base pairs; Bethesda Research Laboratories, Inc., Gaithersburg, Md.) was suspended in 25 μ l of buffer (50 mM Tris-hydrochloride, 1 mM EDTA, pH 8.0) and 7 μ l of stop mix (5% sodium dodecyl sulfate, 0.025% bromphenol blue). Then, 30 μ l of each sample was added to separate wells of agarose gels and electrophoresed using varying currents (30 to 70 mA) and for varying periods (2 to 3 h). After electrophoresis, gels were stained in ethidium bromide (about 2 μ g/ml)

for 20 min, destained in distilled water overnight, and photographed using a 302-nm uv-transilluminator (Fotodyne, Inc., New Berlin, Wisc.) and a Polaroid MP-4 Land Camera (Cambridge, Mass.) equipped with a 23A Kodak Wratten Filter (Eastman Kodak Co., Rochester, N. Y.) and film type 665. Photographic negatives were then analyzed with a densitometer.

Densitometry. Band-migration distances for SDS-polyacrylamide and agarose gels were determined with a Model SL-2DUV two-dimensional soft laser-scanning densitometer (Biomed Instruments). The densitometer was fitted with a high-intensity, nonslit, monochromatic red (633 nm) laser (3- μ m bandwidth and 10-mm band length) and neutral density filters (0.3 and 0.6). Binary data were collected using an Apple IIc personal computer operated with an auto-steppover program (Biomed Instruments) for data capture and Videophoresis II program (Biomed Instruments) for data analysis (26).

RESULTS AND DISCUSSION

A robust modified hyperbola was compared with the standard log transformation, the nonrobust modified hyperbola (regular least squares), and the Southern procedure as implemented by Schaffer and Sederoff (8,20). Various experimental conditions were used to evaluate the robust fitting technique. Six protein standards of known molecular weights ranging from 14.4 to 92.5K and linear-duplex DNA ranging from 1018 to 12216 bp were separated by SDS-polyacrylamide and agarose gel electrophoresis, respectively, and analyzed by a soft laser-scanning densitometer.

The log transform did not adequately describe the data over the entire range of observations in any data set we examined. The modified hyperbola accounted for the curvature of the data more completely and predicted molecular weights and base-pair numbers more accurately throughout the range of the observations, eliminating any subjectivity that may be inherent with the log transformation.

The results from a typical linear-duplex DNA experiment are shown in Fig. 1A. Although the robust modified hyperbola fits the data with little error, the log fit appears adequate as evidenced by the large correlation coefficient of -0.9934 (Table 1). However, the data in Table 1 indicate that the percentage error for each observation of the log fit is as high as 13.52% while the error of the robust fit does not exceed 3.08% and is usually less than 1.0%. The log transform compresses the base-pair number scale in such a fashion that the curvature of the data is suppressed, making the observations appear linear enough for the straight line (log transform) to describe the data well, giving the high correlation coefficient. This statistic is deceptive because comparisons of the actual and predicted base-pair numbers from the standard curve may indicate the error rate to be excessive when the results of such a fit are examined on the original scale. The error associated with the log fit and the accuracy of the robust modified hyperbola are seen more clearly when the data are displayed on the original scale (Fig. 1B).

The robust and regular least-squares fit to the same data in the presence of an artificially

generated outlier are illustrated in Table 2. This outlying observation was created by arbitrarily selecting the migration distance 2.116 and increasing it 25% to 2.650. For the robust fit the weight of each point determines that point's influence on calculations of the model coefficients in succeeding iterations. Observations close to the line with weights near 1.0 exert greater influence on the resultant calculations than distant observations with near 0 weights. In this manner outliers are deemphasized or ignored and the procedure attempts to fit the line through the main body of the data. The regular least-squares procedure ignores this information and permits all observations to influence the fit with a weight of 1.0. This allows outliers to adversely affect the outcome of the overall fit. The robust fit is unaffected by the outlier in the data presented as demonstrated by the zero weight associated with the point (Table 2). The error for the outlying point is large (20.79%) while the error associated with the rest of the observations remains extremely low (mostly less than 1%), indicating that the robust procedure ignored this aberrant observation and fit the line through the main trend of the data. The

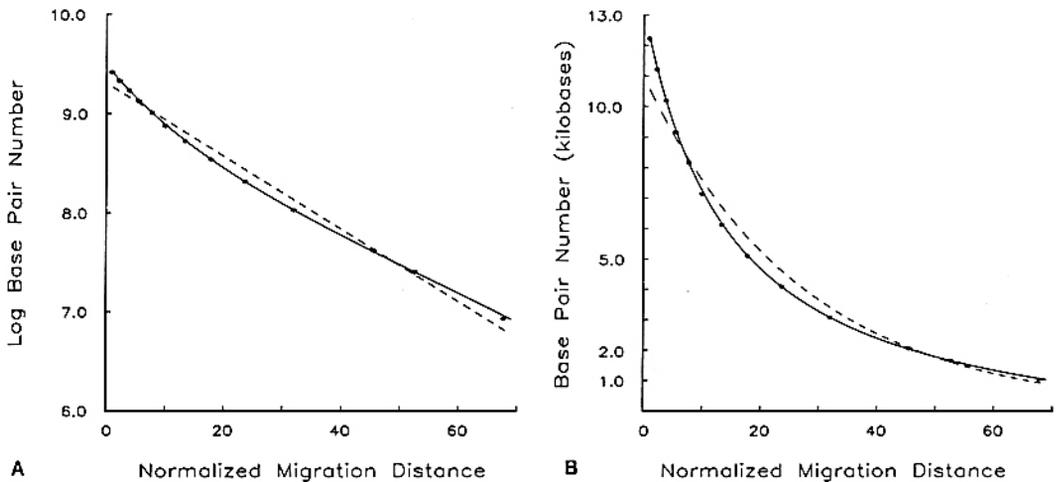


FIG. 1. Comparison of the robust modified hyperbola (—) and log transformation (---) fit to linear-duplex DNA base-pair numbers. Migration distances were normalized as described under Derivation. Electrophoresis was performed in agarose gel as described under Materials and Methods. The DNA fragments contained from 1018 to 12,216 base pairs (see Table 1).

TABLE 1
RESULTS FROM A STANDARD CURVE FOR LINEAR-DUPLEX DNA

Actual migration distance ^a	Normalized migration distance ^b	Known base-pair number ^c	Predicted base-pair number ^d	Residual ^e	% Error ^f
Robust modified hyperbola					
1.204	1.00	12216	12226.95	-10.95	0.09
1.335	2.31	11198	11207.51	-9.51	0.08
1.497	3.93	10180	10118.43	61.57	0.60
1.660	5.56	9162	9185.05	-23.05	0.25
1.888	7.84	8144	8095.91	48.09	0.59
2.116	10.12	7126	7201.73	-75.73	1.06
2.454	13.50	6108	6139.11	-31.11	0.51
2.897	17.93	5090	5079.19	10.81	0.21
3.483	23.79	4072	4054.16	17.84	0.44
4.310	32.06	3054	3049.55	4.45	0.15
5.664	45.60	2036	2008.03	27.97	1.37
6.380	52.76	1635	1629.21	5.79	0.35
7.891	67.87	1018	1049.32	-31.32	3.08
Model statistics					
	<i>a</i> ^g	<i>b</i>	<i>c</i>	<i>d</i>	<i>SSq</i> ^h
	-1422.28	14500.54	0.062	1.033	1.353×10^4
Log transformation					
1.204	1.00	12216	10563.80	1652.20	13.52
1.335	2.31	11198	10068.04	1129.96	10.09
1.497	3.93	10180	9487.01	692.99	6.81
1.660	5.56	9162	8936.23	225.77	2.46
1.888	7.84	8144	8219.04	-75.04	0.92
2.116	10.12	7126	7559.41	-433.41	6.08
2.454	13.50	6108	6677.69	-569.69	9.33
2.897	17.93	5090	5675.86	-585.86	11.51
3.483	23.79	4072	4577.72	-505.72	12.42
4.310	32.06	3054	3379.58	-325.58	10.66
5.664	45.60	2036	2056.34	-20.34	1.00
6.380	52.76	1635	1581.23	53.77	3.29
7.891	67.87	1018	908.26	109.74	10.78
Model statistics					
	<i>y</i> Intercept	Slope	<i>SSq</i> ^h	<i>r</i> ⁱ	
	9.302	-0.0367	5.776×10^6	-0.9934	

^a Actual migration distances measured with a soft laser-scanning densitometer scaled to a graph length of 10.

^b Distances normalized as described under Derivation.

^c Values published by manufacturer (see Materials and Methods).

^d Predicted values from standard curves.

^e Predicted base-pair number subtracted from known base-pair number.

^f Absolute value of residual divided by known base-pair number multiplied by 100.

^g *a*, *b*, *c*, *d* are parameter estimates for the robust fit of the modified hyperbola (see Derivation).

^h *SSq*, weighted sum of squares of the residuals.

ⁱ *r*, Pearson's correlation coefficient (17).

TABLE 2
COMPARISON OF THE ROBUST AND LEAST-SQUARES FIT FOR THE MODIFIED HYPERBOLA FOR DUPLEX DNA
WITH AN ARTIFICIALLY GENERATED OUTLYING OBSERVATION

Actual migration distance ^a	Normalized migration distance ^b	Known base-pair number ^c	Predicted base-pair number ^d	Residual ^e	% Error ^f	Weight ^g
Robust fit						
1.204	1.00	12216	12219.87	-3.87	0.03	0.9994
1.335	2.31	11198	11211.68	-13.68	0.12	0.9922
1.497	3.93	10180	10129.86	50.14	0.49	0.8983
1.660	5.56	9162	9199.96	-37.96	0.41	0.9409
1.888	7.84	8144	8112.40	31.60	0.39	0.9588
2.650 ^h	15.46	7126	5644.66	1481.34	20.79	0.0000
2.454	13.50	6108	6153.46	-45.46	0.74	0.9160
2.897	17.93	5090	5090.48	-0.48	0.01	1.0000
3.483	23.79	4072	4061.72	10.28	0.25	0.9956
4.310	32.06	3054	3053.07	0.93	0.03	1.0000
5.664	45.60	2036	2007.42	28.58	1.40	0.9663
6.380	52.76	1635	1627.21	7.79	0.48	0.9975
7.891	67.87	1018	1045.47	-27.47	2.70	0.9688
Model statistics						
	a^i	b	c	d	SSq^j	
	-1422.74	14477.52	0.061	1.038	8.351×10^3	
Least-squares fit						
1.204	1.00	12216	12174.96	41.04	0.34	
1.335	2.31	11198	11185.99	12.01	0.11	
1.497	3.93	10180	10167.70	12.30	0.12	
1.660	5.56	9162	9303.64	-141.64	1.55	
1.888	7.84	8144	8291.42	-147.42	1.81	
2.650 ^h	15.46	7126	5930.91	1195.09	16.77	
2.454	13.50	6108	6428.74	-320.74	5.25	
2.897	17.93	5090	5379.81	-289.81	5.69	
3.483	23.79	4072	4327.28	-255.28	6.27	
4.310	32.06	3054	3249.22	-195.22	6.39	
5.664	45.60	2036	2069.22	-33.22	1.63	
6.380	52.76	1635	1620.61	14.39	0.88	
7.891	67.87	1018	909.48	108.52	10.66	
Model statistics						
	a^i	b	c	d	SSq^j	
	-2768.85	15884.13	0.063	0.940	1.775×10^6	

^a Actual migration distances measured with a soft laser-scanning densitometer scaled to a graph length of 10.

^b Distances normalized as described under Derivation.

^c Values published by manufacturer (see Materials and Methods).

^d Predicted values from standard curves.

^e Predicted base-pair number subtracted from known base-pair number.

^f Absolute value of residual divided by known base-pair number multiplied by 100.

^g Final weights assigned to each point for the robust fit (see Derivation).

^h Artificial outlying observation created by increasing original data point by 25%.

ⁱ a , b , c , d are parameter estimates for the robust and regular least-squares fit for the modified hyperbola parameters (see Derivation).

^j SSq , weighted sum of squares of the residuals.

least-squares fit, on the other hand, is affected by the outlier to the extent that the error associated with this fit is distributed over more of the observations, indicating the influence the point exerted in the final outcome of the procedure; this is clearly illustrated in Fig. 2. Without the outlier the least-squares fit would be extremely close to the robust fit shown in Table 1 and Figs. 1A and B. This robust treatment of outliers has many advantages over the techniques described by Rodbard (6). The present methods identify aberrant observations and corrects for their presence automatically without additional intervention by the investigator.

The hyperbola described by Southern (Eq. [3]) fits these data with the same degree of accuracy as the nonrobust least-squares fit. This is not surprising as Schaffer and Sederoff estimate the coefficients of this hyperbola using a least-squares method. Due to this estimation procedure their technique is as severely af-

ected by the outlying observation as the least-squares fit shown in Table 2 and Fig. 2.

The modified hyperbola is more accurate than Southern's model when dealing with protein standard data ranging from 14.4 to 92.5 kDa. The analysis of a representative set of data using these two procedures is presented in Table 3, and the log fit is also seen to be highly inaccurate.

Schaffer and Sederoff's strategy for estimating the parameters of [3] was to use least squares to find the solutions for m_0 and l_0 which minimize the corrected sum of squares of the predicted values for h and then calculate an average value for h for use in their predictive equation. Although this solution is not exact, it does provide a method for obtaining good starting values for the iterative procedure which solves for all the coefficients simultaneously and results in a solution which minimizes the sum of squares of the residuals. We implemented the same iterative robust and least-squares procedures to Southern's model to see if this hyperbola would be more accurate after obtaining the exact solutions for the parameters. We found that the differences between the two models still existed. They were similar in accuracy when dealing with the DNA data, whereas the modified hyperbola was consistently more accurate when analyzing the protein standard data.

The modified hyperbola performed optimally when used for analyzing data sets with a fair number of observations (12–13 standards, as in the linear-duplex DNA in agarose gels). We found the procedure to work well with protein molecular weight data sets with as few as five observations. However, trying to estimate four parameters with as little as five observations may occasionally lead to nonconvergence; this occurred with 2 of the 35 data sets examined. Good starting values are essential for the success of any iterative fitting procedure and we found that values based on Schaffer and Sederoff's solution to Southern's model to be quite adequate. However, if outliers are present in the data, Schaffer and Sederoff's solution will be influenced by

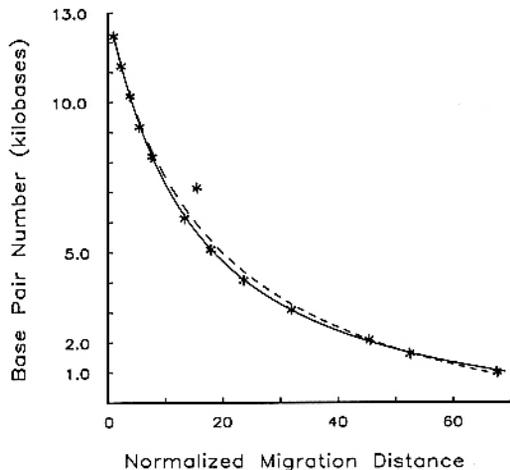


FIG. 2. Comparison of the robust (—) fit and the nonrobust or least-squares fit (---) for the modified hyperbola in the presence of an artificially generated outlier for linear-duplex DNA base-pair numbers. Migration distances were normalized as described under Derivation. The actual migration distance of 2.116 was increased by 25% to 2.650 to generate the outlying point. Electrophoresis was performed in agarose gel as described under Materials and Methods. The DNA fragments contained from 1018 to 12,216 base pairs (see Table 2).

TABLE 3
RESULTS FROM A STANDARD CURVE FOR PROTEIN MOLECULAR WEIGHT DETERMINATION

Actual migration distance ^a	Normalized migration distance ^b	Known M_r^c	Predicted M_r^d	Residual ^e	% Error ^f
Robust modified hyperbola					
0.913	1.00	92500	92560.61	-60.61	0.07
1.918	11.05	66200	65710.33	489.67	0.74
3.736	29.23	45000	46145.75	-1145.75	2.55
6.204	53.91	31000	30405.25	594.75	1.92
8.352	75.39	21500	20576.83	923.17	4.29
9.755	89.42	14400	15244.26	-844.26	5.86
Model statistics					
	a^g	b	c	d	SSq^h
	-287121.94	395166.44	0.041	0.449	3.338×10^6
Southern hyperbola ⁱ					
0.913	1.00	92500	91009.94	1490.06	1.61
1.918	11.05	66200	69010.95	-2810.95	4.25
3.736	29.23	45000	45803.21	-803.21	1.78
6.204	53.91	31000	28910.61	2089.39	6.74
8.352	75.39	21500	20167.36	1332.64	6.20
9.755	89.42	14400	16041.89	-1641.89	11.40
Model statistics					
	m_0	l_0	average h	SSq^h	
	-38.497	-17446.137	4283686.83	1.960×10^7	
Log transformation					
0.913	1.00	92500	85077.39	7422.61	8.02
1.918	11.05	66200	69901.69	-3701.69	5.59
3.736	29.23	45000	48993.21	-3993.21	8.87
6.204	53.91	31000	30241.12	758.88	2.45
8.352	75.39	21500	19871.40	1628.60	7.57
9.755	89.42	14400	15104.66	-704.66	4.89
Model statistics					
	y Intercept	Slope	SSq^h	r^j	
	11.371	-0.0196	8.847×10^7	-0.9946	

^a Actual migration distances measured with a soft laser-scanning densitometer scaled to a graph length of 10.

^b Distances normalized as described under Derivation.

^c Values published by manufacturer (see Materials and Methods).

^d Predicted values from standard curves.

^e Predicted M_r subtracted from known M_r .

^f Absolute value of residual divided by known M_r multiplied by 100.

^g a , b , c , d are parameter estimates for the robust fit of the modified hyperbola (see Derivation).

^h SSq , weighted sum of squares of the residuals.

ⁱ Figures obtained by using Schaffer and Sedroff's least-squares algorithm.

^j r , Pearson's correlation coefficient (17).

these points and may lead to poor starting values. This may result in the algorithm failing to converge to a solution. In this case coefficients from a previously successful run with similar data may be substituted for the calculated starting values and the iterative process retried. Other suggestions are offered in the Appendix in the description of the program.

The robust fitting techniques worked successfully in a variety of applications. The procedures were applied to protein standards in both linear and gradient polyacrylamide gels with weights ranging from 14.4 to 92.5K and with other data sets with weights ranging from 14.4 to 200.0K (data not shown). It should also be emphasized that the successful application of these techniques is not restricted to protein molecular weight or DNA fragment length data. They may be extended to other types of data which exhibit a curvilinear relationship similar to the data presented in this paper. Tiede and Pagano's (22) application of the techniques to radioimmunoassay data is an example. Also, these methods are not restricted to data collected solely by a densitometer. Although measurement of migration distances by a ruler or calipers is not as accurate as a densitometer, these methods can work satisfactorily as long as the data retain the same basic curvilinear structure.

Although the robust modified hyperbola described the standard protein and linear-duplex DNA data sets well, anomalous migration behavior, due to compositional or physical characteristics of proteins and DNA fragments, has been observed (27,28) and these types of proteins and DNA fragments do not lend themselves to this analysis. In addition, limitations of the electrophoresis process must be considered in order to assure molecular weight and base-pair number accuracy.

The robust modified hyperbola was consistently more accurate when compared with the log transformation and a similar hyperbola proposed by Southern. Therefore, the present modified hyperbola with the robust fitting technique is recommended for general use, due to the importance of determining accurate

standard curves under a wide variety of experimental conditions and data types.

APPENDIX

The BASIC program listed in Table A1 was written, compiled, and tested on an IBM personal computer and, with minor modifications, will run on any personal or home computer with a BASIC compiler or interpreter. The speed of execution of the program will increase substantially if it is compiled before running. The data are entered at the end of the program, as described by the comments at the beginning of the listing.

The program may occasionally fail to converge to a solution with some data sets. Outliers will severely affect the outcome of Schaffer and Sederoff's algorithm, resulting in the computation of poor starting values for the iterative procedure. Additionally, because Southern's hyperbola is less accurate than the modified hyperbola when dealing with protein standards ranging from 14.4 to 92.5 kDa, even in the absence of outliers, weak starting values may result for this particular application, preventing convergence to a solution. In this case coefficients from a previously successful run with similar data may be substituted for the calculated starting values and the iterative process retried. Another method for finding a solution when the procedure does not converge is to reexpress Eq. [2] as

$$\mathbf{O}_{j+1} = \mathbf{O}_j + \mathbf{B}_j \quad [4]$$

where $\mathbf{B}_j = (\mathbf{A}_j' \mathbf{W}_j \mathbf{A}_j)^{-1} \mathbf{A}_j' \mathbf{W}_j \mathbf{R}_j$.

The most recent estimate for the parameter vector $\mathbf{O} = \mathbf{O}_{j+1}$ is equal to the last estimate, \mathbf{O}_j , plus an amount equal to \mathbf{B}_j . If \mathbf{B}_j is very large then the solution may circle around some value without actually converging or may rapidly expand and diverge to the point of causing a numeric overflow in the program. To avoid this the magnitude of \mathbf{B}_j can be decreased. The program prompts the user to enter the proportion for \mathbf{B} update—usually 1.0. To start the process, enter 1.0 which results in [4]. If the solution does not converge or if it diverges

TABLE A1

```

10 ' Iterative robust and approximate least squares solutions for the
20 ' modified hyperbola.
30 ' For input the data should be entered in the following order:
40 ' The graph length or total length of the gel; the number of
50 ' readings for the standard curve; the individual data pairs -
60 ' the migration distance followed by the known molecular weight or
70 ' DNA fragment length. The observation corresponding to the
80 ' heaviest or longest band MUST be entered first. This is followed
90 ' by the number of unknown migration distances from which molecular
100 ' weights or DNA fragment lengths will be calculated. Use 0
110 ' if there are none. Finally enter the migration distances for the
120 ' unknowns.
130 ' # after a variable name or number indicates double precision.
140 OPTION BASE 1
150 DIM X$(50), X1$(50), Y$(50), B$(4), B1$(4), BB$(4), W$(50),
    XTWX$(4,4)
160 DIM XTWXINV$(4,4), XTWR$(4), YHAT$(50), RESIDUAL$(50), XX$(50,4),
    RES$(50)
170 DIM RESIDUAL1$(50)
180 DIM A$(16), L(4), M(4)
190 DIM PROD(50), DWT(50), DDIST(50), DPROD(50), C(50), D(50)
200 INPUT "Enter proportion for B update - usually 1.0";PROP#
210 INPUT "Enter 1 for Robust fit; 2 for approximate least squares
    fit";FIT
220 IF (FIT <> 1 AND FIT <> 2) THEN GOTO 210
230 READ GRAPH#
240 READ N1
250 ' Compute starting values for the iterative process.
260 LET SWT=0
270 LET SDIST = 0
280 LET SPROD = 0
290 FOR I = 1 TO N1
300 READ X1$(I), Y$(I)
310 LET X$(I)=X1$(I)/GRAPH#
320 IF (I=1) THEN LET TEMP1#=X$(1)
330 LET X$(I)=(((X$(I)-TEMP1#)+.01#)*(100#))
340 LET SWT = SWT + Y$(I)
350 LET SDIST = SDIST + X$(I)
360 LET PROD(I) = Y$(I)*X$(I)
370 LET SPROD = SPROD + PROD(I)
380 ' Initialize the weight vector W.
390 LET W$(I) = 1#
400 NEXT I
410 LET MWT = SWT/N1
420 LET MDIST = SDIST/N1
430 LET MPROD = SPROD/N1
440 FOR I = 1 TO N1
450 LET DWT(I) = Y$(I) - MWT
460 LET DDIST(I) = X$(I) - MDIST
470 LET DPROD(I) = PROD(I) - MPROD
480 NEXT I
490 LET CSSL = 0
500 LET CSSM = 0
510 LET CSCPML = 0
520 LET CSPMLL = 0
530 LET CSPMLM = 0
540 FOR I = 1 TO N1
550 LET CSSL = CSSL + DWT(I)^ 2
560 LET CSSM = CSSM + DDIST(I)^ 2
570 LET CSCPML = CSCPML + DWT(I)*DDIST(I)
580 LET CSPMLL = CSPMLL + DPROD(I)*DWT(I)
590 LET CSPMLM = CSPMLM + DPROD(I)*DDIST(I)
600 NEXT I
610 LET DET = CSSL*CSSM - CSCPML^ 2
620 LET MO = (CSSM*CSPMLL - CSCPML*CSPMLM)/DET
630 LET LO = (-CSCPML*CSPMLL + CSSL*CSPMLM)/DET
640 LET SC = 0
650 FOR I = 1 TO N1
660 LET C(I) = (Y$(I) - LO)*(X$(I) - MO)

```

TABLE A1—Continued

```

670 LET SC = SC + (C(I) - C(1))
680 NEXT I
690 LET CBAR = SC/N1 + C(1)
700 ' Compute starting values
710 LET B#(1) = LO
720 LET B#(2) = -(CBAR/MO)
730 LET B#(3) = -(1/MO)
740 LET B#(4) = 1#
750 ' Start the iteration process
760 LET II=0
770 LET II = II + 1
780 PRINT USING "###";II;
790 PRINT USING "#####.#####";B#(1), B#(2), B#(3), B#(4)
800 ' Calculate the matrix of partial derivatives evaluated at X and
810 ' the current estimate of B.
820 LET SSQ1# = 0#
830 LET SSQ2# = 0#
840 ' Initialize summation variable used for checking convergence of
      solution
850 LET SUMB# = 0#
860 FOR II = 1 TO N1
870 ' Calculate predicted Y's (YHAT) and RESIDUALS
880 LET YHAT#(II) = B#(1) + (B#(2) / (1# + (B#(3) * (X#(II) ^ B#(4))))
890 LET RESIDUAL#(II) = Y#(II) - YHAT#(II)
900 LET RESIDUAL1#(II)=ABS(RESIDUAL#(II))
910 ' Calculate the sums of squares of the residuals
920 LET SSQ1# = SSQ1# + (RESIDUAL#(II) ^ 2) * W#(II)
930 LET XX#(II,1) = 1#
940 LET XX#(II,2) = 1#/(1# + (B#(3) * (X#(II) ^ B#(4))))
950 LET XX#(II,3) = (-1#) * B#(2) * (X#(II) ^ B#(4)) * (XX#(II,2) ^ 2)
960 LET XX#(II,4) = XX#(II,3) * B#(3) * LOG(X#(II))
970 NEXT II
980 IF (FIT = 1) THEN GOSUB 3150
990 ' initialize X'WX and X'W[Y-f(X)] and update BB
1000 FOR J=1 TO 4
1010 FOR K=J TO 4
1020 LET XTWX#(J,K) = 0#
1030 ' Take advantage of the symmetry of the X'WX matrix
1040 LET XTWX#(K,J) = 0#
1050 NEXT K
1060 ' initialize X'W[Y-f(X)]
1070 LET XTWR#(J) = 0#
1080 ' update BB
1090 LET BB#(J) = B#(J)
1100 NEXT J
1110 FOR II = 1 TO N1
1120 ' Calculate X'WX
1130 FOR J=1 TO 4
1140 FOR K=J TO 4
1150 LET XTWX#(J,K) = XTWX#(J,K) + W#(II) * XX#(II,J) * XX#(II,K)
1160 ' Take advantage of the symmetry of the X'WX matrix
1170 LET XTWX#(K,J)=XTWX#(J,K)
1180 NEXT K
1190 ' Calculate X'W[Y-f(X)]
1200 LET XTWR#(J) = XTWR#(J) + W#(II)*XX#(II,J)*RESIDUAL#(II)
1210 NEXT J
1220 NEXT II
1230 ' compute INV[X'WX]
1240 GOSUB 1950
1250 ' Compute new Beta vector (B)
1260 FOR I=1 TO 4
1270 ' Initialize the temporary B vector
1280 LET B1#(I) = 0#
1290 FOR J=1 TO 4
1300 LET B1#(I) = B1#(I) + XTWXINV#(I,J) * XTWR#(J)
1310 NEXT J
1320 ' Finally update the last estimate of B with this new calculation.
1330 ' Make sure that the proportion of change, inputted by the user,
1340 ' is incorporated.
1350 LET B#(I) = B#(I) + PROP# * B1#(I)
1360 ' Now compute the difference between the old B vector (BB) and the
1370 ' new B vector (B)
1380 LET B1#(I)=B#(I)-BB#(I)

```

TABLE A1—Continued

```

1390 ' Add up these differences to check for convergence later on
1400 LET SUMB# = SUMB# + ABS(B1#(I)/BB#(I))
1410 NEXT I
1420 ' Now compute a second sums of squares of residuals. This
1430 ' statistic is what the least squares procedure is really
1440 ' minimizing: Y-f(X)-
      [B1(1)XX(1)+B1(2)XX(2)+B1(3)XX(3)+B1(4)XX(4)]
1450 ' where B1=B-BB (or the new estimate-the old estimate)
1460 FOR I = 1 TO N1
1470 ' Initialize the temporary variable RES.
1480 LET RES#(I) = 0#
1490 FOR J = 1 TO 4
1500 ' now compute [B1(1)XX(1)+B1(2)XX(2)+B1(3)XX(3)+B1(4)XX(4)]
1510 LET RES#(I) = RES#(I) + B1#(J) * XX#(I,J)
1520 NEXT J
1530 ' Now compute the second modified sum of squares.
1540 LET SSQ2# = SSQ2# + ((RESIDUAL#(I)-RES#(I)) ^ 2) * W#(I)
1550 NEXT I
1560 PRINT "Regular and adjusted sums of squares = "; : PRINT USING
      "###.##### ^^^^"; SSQ1#, SSQ2#
1570 PRINT " "
1580 ' Check for convergence of former and present B estimate
1590 IF ((SUMB# <= .00001#) OR (I1 >100)) THEN GOTO 1610
1600 GOTO 770
1610 PRINT " "
1620 IF (FIT = 1) THEN PRINT "                               Results of
      Robust fit"; : PRINT " "
1630 IF (FIT = 2) THEN PRINT "                               Results of approximate
      least squares fit"; : PRINT " "
1640 PRINT " " ; PRINT " "
1650 PRINT " Migr.          Rel.          Known          Pred."
1660 PRINT " Dist.          Positn       M.W.           M.W.           Residual
      % Error   Weight"
1670 PRINT " "
1680 FOR I = 1 TO N1
1690 PRINT USING "###.###"; X1#(I);
1700 PRINT USING "#####.##"; X#(I);
1710 PRINT USING "#####"; Y#(I);
1720 PRINT USING "#####.##"; YHAT#(I);
1730 PRINT USING "#####.##"; RESIDUAL#(I);
1740 PRINT USING "#####.##"; (ABS(RESIDUAL#(I))/Y#(I))*100#;
1750 PRINT USING "#####.##"; W#(I)
1760 NEXT I
1770 PRINT " " : PRINT " The final coefficients are:" : PRINT " "
1780 PRINT USING "#####.#####"; BB#(1), BB#(2), BB#(3), BB#(4) :
      PRINT " " ; PRINT " "
1790 PRINT "Regular and adjusted sums of squares = "; : PRINT USING
      "###.##### ^^^^"; SSQ1#, SSQ2#
1800 PRINT " "
1810 ' Compute molecular weights or DNA fragment lengths for unknowns.
1820 READ N1
1830 IF (N1 < 1) THEN GOTO 1940
1840 PRINT "Calculation of molecular weights or DNA fragment lengths" :
      PRINT " "
1850 PRINT " Migr.          Rel.          Pred."
1860 PRINT " Dist.          Positn       M.W. " ; PRINT " "
1870 FOR I = 1 TO N1
1880 READ XUNKN#
1890 LET XUNKN1# = XUNKN#/GRAPH#
1900 LET XUNKN1# = (((XUNKN1#-TEMP1#)+.01#)*(100#))
1910 LET YUNKN# = BB#(1) + (BB#(2) / (1# + (BB#(3) * (XUNKN1# ^ BB#(4))))))
1920 PRINT USING "###.###"; XUNKN#; : PRINT USING "#####.##"; XUNKN1#;
      : PRINT USING "#####.##"; YUNKN#
1930 NEXT I
1940 END
1950 '***** SUBROUTINE INVERT *****
1960 ' Subroutine transcribed from the IBM FORTRAN SSB routine MINV
1970 '
1980 ' Method
1990 ' The standard Gauss-Jordan method is used. The determinant
2000 ' is also calculated. A determinant of zero indicates that the
2010 ' matrix is singular.
2020 '

```

TABLE A1—Continued

```

2030 LET N=4
2040 FOR I= 1 TO N
2050 FOR J= 1 TO N
2060 LET IJ=N*(J-1) + I
2070 LET A(IJ)=XTWX(I,J)
2080 NEXT J
2090 NEXT I
2100 ' Search for the largest element
2110 LET D#=1#
2120 LET NK=-N
2130 FOR K=1 TO N
2140 LET NK=NK+N
2150 LET L(K)=K
2160 LET M(K)=K
2170 LET KK=NK+K
2180 LET BIGA#=A(KK)
2190 FOR J=K TO N
2200 LET IZ=N*(J-1)
2210 FOR I=K TO N
2220 LET IJ=IZ+I
2230 IF (ABS(BIGA#) < ABS(A(IJ))) THEN LET BIGA#=A(IJ) : LET L(K)=I :
    LET M(K)=J
2240 NEXT I
2250 NEXT J
2260 ' Interchange rows
2270 LET J=L(K)
2280 IF (J <= K) THEN GOTO 2380
2290 LET KI=K-N
2300 FOR I=1 TO N
2310 LET KI = KI+N
2320 LET HOLD#=-A(KI)
2330 LET JI=KI-K+J
2340 LET A(KI)=A(JI)
2350 LET A(JI)=HOLD#
2360 NEXT I
2370 ' Interchange columns
2380 LET I=M(K)
2390 IF (I <= K) THEN GOTO 2490
2400 LET JP=N*(I-1)
2410 FOR J=1 TO N
2420 LET JK=NK+J
2430 LET JI=JP+J
2440 LET HOLD#=-A(JK)
2450 LET A(JK)=A(JI)
2460 LET A(JI)=HOLD#
2470 NEXT J
2480 'divide column by minus pivot (value of pivot element is contained
    in BIGA)
2490 IF (ABS(BIGA#) <= .00000000000001#) THEN LET D#=0# : PRINT
    "determinant=0 - the matrix is singular" : END
2500 FOR I=1 TO N
2510 IF (I = K) THEN GOTO 2540
2520 LET IK=NK+I
2530 LET A(IK)=A(IK)/(-BIGA#)
2540 NEXT I
2550 ' Reduce matrix
2560 FOR I=1 TO N
2570 LET IK=NK+I
2580 LET HOLD#=A(IK)
2590 LET IJ=I-N
2600 FOR J=1 TO N
2610 LET IJ=IJ+N
2620 IF (I=K) THEN GOTO 2660
2630 IF (J=K) THEN GOTO 2660
2640 LET KJ=IJ-I+K
2650 LET A(IJ)=HOLD#*A(KJ)+A(IJ)
2660 NEXT J
2670 NEXT I
2680 ' Divide row by pivot
2690 LET KJ=K-N
2700 FOR J=1 TO N
2710 LET KJ=KJ+N
2720 IF (J=K) THEN GOTO 2740

```

TABLE A1—Continued

```

2730 LET A#(KJ)=A#(KJ)/BIGA#
2740 NEXT J
2750 ' Product of pivots
2760 LET D#=D#*BIGA#
2770 ' Replace pivot by reciprocal
2780 LET A#(KK)=1#/BIGA#
2790 NEXT K
2800 ' Final row and column interchange
2810 LET K=N
2820 LET K=K-1
2830 IF (K <= 0) THEN GOTO 3070
2840 LET I=L(K)
2850 IF (I <= K) THEN GOTO 2950
2860 LET JQ=N*(K-1)
2870 LET JR=N*(I-1)
2880 FOR J=1 TO N
2890 LET JK=JQ+J
2900 LET HOLD#=A#(JK)
2910 LET JI=JR+J
2920 LET A#(JK)=-A#(JI)
2930 LET A#(JI)=HOLD#
2940 NEXT J
2950 LET J=M(K)
2960 IF (J <= K) THEN GOTO 2820
2970 LET KI=K-N
2980 FOR I=1 TO N
2990 LET KI=KI+N
3000 LET HOLD#=A#(KI)
3010 LET JI=KI-K+J
3020 LET A#(KI)=-A#(JI)
3030 LET A#(JI)=HOLD#
3040 NEXT I
3050 GOTO 2820
3060 ' convert vector back to matrix array
3070 FOR I= 1 TO N
3080 FOR J= 1 TO N
3090 LET IJ=N*(J-1) + I
3100 LET XTWXINV#(I,J)=A#(IJ)
3110 NEXT J
3120 NEXT I
3130 RETURN
3140 END
3150 ' ***** SUBROUTINE WEIGHT *****
3160 ' This subroutine computes the weight vector for the robust fit.
3170 LET IE=N1
3180 LET J1 = 1
3190 FOR I5=1 TO N1
3200 ' Set the lower relative index of the array to be sorted.
3210 LET IB=I5
3220 GOSUB 3380
3230 NEXT I5
3240 ' Compute SD for those times N1 is even
3250 LET IND=3 + INT((N1-2)/2)
3260 LET SD#=RESIDUAL1#(IND)
3270 ' If N1 is odd then use the average of the two middle values of
3280 ' the last N1-3 elements of RESIDUAL1
3290 IF ((N1 MOD 2) <> 0) THEN LET
SD#=((RESIDUAL1#(IND)+RESIDUAL1#(IND+1))/2#)
3300 ' Now set up the weight vector
3310 FOR I=1 TO N1
3320 LET W#(I) = 0#
3330 LET TEMP#=(RESIDUAL1#(I)/(SD#*2.1#))
3340 IF ((-3.141592654# <= TEMP#) AND (TEMP# <= 3.141592654#)) THEN LET
W#(I)=SD#*2.1#*((SIN(TEMP#))/RESIDUAL1#(I))
3350 NEXT I
3360 RETURN
3370 END
3380 ' ***** SUBROUTINE ORDERS *****
3390 '
3400 ' This subroutine finds the i'th order statistic from an array of
3410 ' numbers. RESIDUAL1 is the array; J1 is the order statistic
3420 ' desired; IB is the lower index of the array to be examined;
3430 ' IE is the upper index of the array to be examined.

```

TABLE A1—Continued

```

3440 '
3450 LET JO = IB + J1-1
3460 LET NN = IE + 1
3470 LET IJ = IB - 1
3480 LET K=NN
3490 LET I=IJ
3500 LET T#=RESIDUAL1#(JO)
3510 LET K=K-1
3520 LET W1#=RESIDUAL1#(K)
3530 IF (W1# > T#) THEN GOTO 3510
3540 IF (W1# < T#) THEN GOTO 3570
3550 IF (K <> JO) THEN GOTO 3510
3560 IF (I = K) THEN RETURN
3570 LET I = I + 1
3580 LET Z# = RESIDUAL1#(I)
3590 IF (Z# < T#) THEN GOTO 3570
3600 IF (Z# > T#) THEN GOTO 3680
3610 IF (I <> JO) THEN GOTO 3570
3620 IF (I = K) THEN RETURN
3630 LET RESIDUAL1#(I) = W1#
3640 LET RESIDUAL1#(K) = Z#
3650 LET I = IJ
3660 LET NN = K
3670 GOTO 3500
3680 IF (I = K) THEN RETURN
3690 LET RESIDUAL1#(I) = W1#
3700 LET RESIDUAL1#(K) = Z#
3710 IF (K <> JO) THEN GOTO 3510
3720 LET K = NN
3730 LET IJ = I
3740 GOTO 3500
3750 END
3760 DATA 10, 13
3770 DATA 1.204, 12216
3780 DATA 1.335, 11198
3790 DATA 1.497, 10180
3800 DATA 1.660, 9162
3810 DATA 1.888, 8144
3820 DATA 2.116, 7126
3830 DATA 2.454, 6108
3840 DATA 2.897, 5090
3850 DATA 3.483, 4072
3860 DATA 4.310, 3054
3870 DATA 5.664, 2036
3880 DATA 6.380, 1635
3890 DATA 7.891, 1018
3900 DATA 3
3910 DATA 1.497, 1.888, 6.380

```

to numeric overflow, then try a smaller proportion, e.g., 0.5, which would modify Eq. [4] to

$$\mathbf{O}_{j+1} = \mathbf{O}_j + (0.5)\mathbf{B}_j.$$

The proportion may be altered at will to compute smaller and smaller increments of change for the \mathbf{O}_j vector.

If the solution does not diverge but circles around a solution without actually converging, then examine the sum of squares of the residuals that is returned after each iteration. The solution will tend to minimize this sum. The coefficients that return the smallest sum of squares can be used as starting values for an-

other attempt at finding a solution. Locate the statements in the program where the initial starting values are computed and assigned (program line numbers 710–740) and replace them with these revised starting values. When executing the program, select 0.5 or 0.25 for the proportion for \mathbf{B} update and monitor the iterative process. If the solution continues to circle around a set of values without actually converging, select those four values which return the smallest sum of squares of the residuals and test them by computing predicted values for each of the known standards. If the error rates are acceptable, then these coefficients may be used to form the standard curve.

The program with the attached data set should produce the results for the robust fit listed in Table 1. If the value 2.116 on the seventh DATA card is changed to 2.650, the program will return the results listed in Table 2.

To avoid transcription errors while entering the program listed in Table A1, the authors will supply a copy of the program formatted for an IBM PC after receiving a blank, 5¼-in. diskette.

REFERENCES

1. Ferguson, K. A. (1964) *Metabolism* **13**, 985-1002.
2. Ogston, A. G. (1958) *Trans. Faraday Soc.* **54**, 1754-1757.
3. Rodbard, D. (1976) in *Methods in Protein Separation* (Catsimpoalas, N., ed.), Vol 2, pp. 145-179, Plenum, New York.
4. Rodbard, D., and Chrambach, A. (1970) *Proc. Natl. Acad. Sci. USA* **65**, 970-977.
5. Chrambach, A., Jovin, T. M., Svendsen, P. J., and Rodbard, D. (1976) in *Methods in Protein Separation* (Catsimpoalas, N., ed.), Vol 2, pp. 27-144, Plenum, New York.
6. Rodbard, D. (1976) in *Methods in Protein Separation* (Catsimpoalas, N., ed.), Vol 2, pp. 181-218, Plenum, New York.
7. Shapiro, A. L., Vinuela, E., and Maizel, J. V., Jr. (1967) *Biochem. Biophys. Res. Commun.* **28**, 815-820.
8. Schaffer, H. E. (1983) in *Statistical Analysis of DNA Sequence Data* (Weir, B. S., ed.), pp. 1-14, Dekker, New York.
9. Frank, R. N., and Rodbard, D. (1975) *Arch. Biochem. Biophys.* **171**, 1-13.
10. Weber, K., and Osborn, M. (1969) *J. Biol. Chem.* **244**, 4406-4412.
11. Dunker, A. K., and Rueckert, R. R. (1969) *J. Biol. Chem.* **244**, 5074-5080.
12. Fisher, M. P., and Dingman, C. W. (1971) *Biochemistry* **10**, 1895-1899.
13. Fangman, W. L. (1978) *Nucleic Acids Res.* **5**, 653-665.
14. Kersters, K., and De Ley, J. (1975) *J. Gen. Microbiol.* **87**, 333-342.
15. Hames, B. D. (1981) in *An Introduction to Polyacrylamide Gel Electrophoresis* (Hames, B. D., and Rickwood, D., eds.), Vol. 2, pp. 1-91, Information Retrieval Limited Press, Oxford.
16. Helling, R. B., Goodman, H. M., and Boyer, H. W. (1974) *J. Virol.* **14**, 1235-1244.
17. Armitage, P. (1971). *Statistical Methods in Medical Research*, p. 147-166, Wiley, New York.
18. Duggelby, R. G., Kinns, H., and Rood, J. I. (1981) *Anal. Biochem.* **110**, 49-55.
19. Parker, R. C., Watson, R. M., and Vinograd, J. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 851-855.
20. Schaffer, H. E., and Sederoff, R. R. (1981) *Anal. Biochem.* **115**, 113-122.
21. Southern, E. M. (1979) *Anal. Biochem.* **100**, 319-323.
22. Tiede, J. J., and Pagano, M. (1979) *Biometrics* **35**, 567-574.
23. Draper, N. R., and Smith, H. (1981) *Applied Regression Analysis*, 2nd ed., p. 458-529, Wiley, New York.
24. Andrews, D. F. (1974) *Technometrics* **16**, 523-531.
25. Laemmli, U. K. (1970) *Nature (London)* **227**, 680-685.
26. Carlone, G. M., Sottnek, F. O., and Plikaytis, B. D. (1985) *J. Clin. Microbiol.* **22**, 708-713.
27. Sanger, F., Coulson, A. R., Hong, G. H., and Hill, D. F. (1982) *J. Mol. Biol.* **162**, 729-773.
28. Williams, J. G., and Gratzer, W. B. (1971) *J. Chromatogr.* **57**, 121-125.

ERRATUM

Volume 152, Number 2 (1986), in the article "Robust Estimation of Standard Curves for Protein Molecular Weight and Linear-Duplex DNA Base-Pair Number after Gel Electrophoresis," by Brian D. Plikaytis, George M. Carlone, Paul Edmonds, and Leonard W. Mayer, pages 346-364: On page 355, in Fig. 2 the outlying point described in the legend was inadvertently deleted during the printing process. For the reader's convenience, the correct Fig. 2 and its legend are reproduced below.

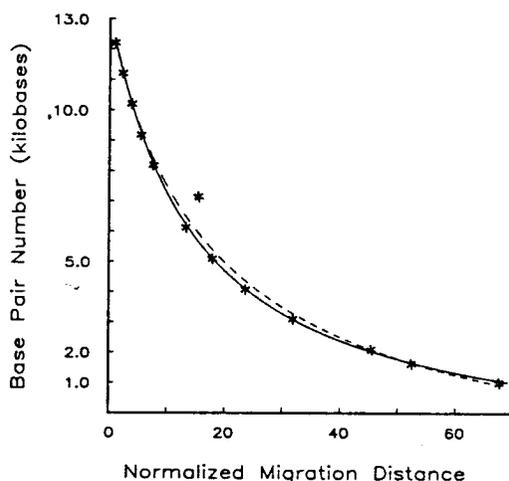


FIG. 2. Comparison of the robust (—) fit and the non-robust or least-squares fit (---) for the modified hyperbola in the presence of an artificially generated outlier for linear-duplex DNA base-pair numbers. Migration distances were normalized as described under Derivation. The actual migration distance of 2.116 was increased by 25% to 2.650 to generate the outlying point. Electrophoresis was performed in agarose gel as described under Materials and Methods. The DNA fragments contained from 1018 to 12,216 base pairs (see Table 2).