Volume 6 Number 12 1979 Nucleic Acids Research

Computer programs for analysis of nucleic acid hybridization, thermal denaturation, and gel electrophoresis data

Robert F.Murphy, William R.Pearson* and James Bonner

Division of Biology, California Institute of Technology, Pasadena, CA 91125, and *Department of Microbiology, Johns Hopkins Medical School, Baltimore, MD 21205, USA

Received 17 April 1979

ABSTRACT

Computer programs for the analysis of data from techniques frequently used in nucleic acids research are described. In addition to calculating non-linear least-squares solutions to equations describing these systems, the programs allow for data editing, normalization, plotting and storage, and are flexible and simple to use. Typical applications of the programs are described.

INTRODUCTION

The increasing complexity and volume of data being generated in biochemical and biophysical experimentation, and the proliferation of mini-computer systems, has created a need for portable, interactive data storage and analysis programs. This paper describes non-linear least-squares fitting programs which have been used for the analysis of data from gel electrophoresis (1-4), DNA-DNA and RNA-DNA hybridization (5-7), and DNA, chromatin and protein-DNA thermal denaturation (8-10). These programs have been implemented on a PDP-11 mini-computer system, and are written in FORTRAN for ease of transfer to other computer systems. The programs require approximately 16,000 16-bit words of memory and a disk mass storage device (such as flexible disk). Mini-computers capable of running this data analysis system are currently available for less than $5,000.

In this paper we will first describe three different programs in the package, and then discuss the reliability and significance of the parameter estimates calculated by the programs.

PROGRAM DESCRIPTIONS

The three programs in the least-squares fitting package are

COTFIT, for analysis of nucleic acid hybridization and denaturation data, GELFIT, for determination of the positions and areas of Gaussian curves fit to gel electrophoresis data, and MELSMR, for removal of noise from closely-spaced data. These programs use a common data file format, which allows interaction of the various programs for uses other than those described above. For example, CsCl density gradient data could be entered with COTFIT, smoothed with MELSMR, and then analyzed with GELFIT.

COTFIT

COTFIT is a non-linear least-squares fitting program developed from the NNNBAT program described by Pearson, Davidson and Britten (11), which was in turn based on the FINGER program of Britten, Graham and Neufeld (12). The program accepts English commands to control data entry and fitting. Data may be read from disk files or entered at the terminal. In addition to least-squares fitting, COTFIT provides general facilities for entering and editing data, and offers a variety of options for plotting and printing the curves calculated from the data.

Initial parameter estimates are improved by a modification of the method of Marquardt (13). Parameter values which would produce undefined function values are detected without causing arithmetic errors. This enables the fitting routine to try a broad range of possible parameter values without causing program termination.

As an additional option, files containing function values (with or without specified errors in the parameters) over a given interval can be generated. This feature is useful for displaying the functions under various conditions.

The functions described by Pearson, Davidson and Britten (11) have been modified, and a new function has been added. The NNNBAT function names and the corresponding COTFIT names are listed below.

N	COTFIT	NNNBAT	DESCRIPTION
1	FINGER	FINGER	second order DNA renaturation.
2	DIGEST	WHATOR	variable order renaturation.
3	DRIVEN	NUFORM	tracer/driver reaction with different nucleation rates.
4	EXCESS	EXCESS	first order renaturation.
5	MELTFN	------	thermal denaturation function.

The MELTFN function is

$$H(T) = I + \sum_{i=1}^{m} Fi\{0.5 + \frac{1}{\sqrt{\pi}} \int_{Tmi}^{T} \exp\,[\frac{-(x-Tmi)^2}{2Si^2}]dx\} \qquad (1)$$

where H is the hyperchromicity at temperature T, I is the initial hyperchromicity, m is the number of components, Fi, Tmi, and Si are the hyperchromicity, transition midpoint and transition width (equivalent to the standard deviation of a Gaussian curve) of component i. The function is a normalized form of the error function, which is the integral of the Gaussian distribution. The program uses a polynomial approximation to this integral (14). The addition of this function allows direct fitting of thermal denaturation data, as opposed to the more common use of the Gaussian function to fit derivatized data. Although the two methods are theoretically identical, in practice the new approach eliminates errors introduced by the derivatization and smoothing process, and requires significantly fewer data points to determine component parameters. This is especially useful for analyzing DNA melts assayed by hydroxyapatite or filter binding (R.B. Wallace, G. Schaeffer, T. Hirose, K. Itakura, R.F. Murphy and J. Bonner, submitted to Nucleic Acids Res).

GELFIT

The GELFIT program uses a set of commands which is compatible with that of the COTFIT program, and fits Gaussian curves to integer data with evenly-incremented X values. The program can calculate the molecular weight of the species in a given band using a polynomial approximation to standards data specified by the user. The simplicity of the function being fitted allows the program to make initial estimates of the number and position of the bands in a given gel, and then proceed with fitting. The restriction to Y-only integer data was added to minimize data storage space and computation time in view of the large numbers of data points commonly collected during gel scanning in our laboratory. The fitting method is similar to that used by COTFIT. In order to reduce program size, a band matrix (15) is used in place of the symmetric triangular matrix COTFIT uses. The number of adjacent curves whose values are allowed to affect an individual curve's parameter estimates can be adjusted at run-time by the user.

MELSMR

The MELSMR program smoothes and/or derivatizes data having a constant X increment between points using the method of Savitzky and Golay (16-18). The degree to which the data are smoothed is controlled by the user. Digital data which has been collected from instruments such as spectrophotometers frequently contain fluctuations in the less significant digits. MELSMR provides a means of reducing or eliminating this noise.

RESULTS AND DISCUSSION

While these programs can significantly shorten the time required to analyze gel and melt data, and are essential for accurate measurement of nucleic acid hybridization rates and component amounts, the significance of the calculated parameter estimates must not be overestimated. This section addresses three issues encountered in fitting nucleic acid data with least squares programs: 1) The significance of the exponent in S1 nuclease-assayed reassociation data analysis; 2) the effect of using equation (1) to analyze melt data; and 3) the reproducibility of melt data parameter estimates.

Analysis of nuclease-assayed reassociation data

Morrow (19) and Smith, Britten and Davidson (20) have analyzed the kinetics of DNA-DNA reassociation assayed by the single-strand-specific S1 nuclease of *Aspergillus oryzae*. They concluded that data from S1 nuclease assayed renaturations were best fit using the equation

$$\frac{S}{Co} = (1 + kCot)^{-n} \qquad (2)$$

where S is the S1 nuclease sensitive (single stranded) DNA NT concentration, Co is the total DNA NT concentration, t is time, k is the rate constant which would be observed if the reaction were assayed on HAP, and n was found to be 0.44 (19) or 0.453 (20) for driver and tracer DNAs of the same size. The deviation from second order kinetics indicated by a value of n less than 1 was attributed to the lowered reactability of the single strand regions of partial duplexes relative to free single strands.

Table 1. Parameters for equation (3) for data of Sala-Trepat et al (6)

DRIVER*	c-DNA+	N++	U	F	K	n	%E+++
liver	RSA	16	-0.0482	1.0490	0.001104	0.440	3.021
			0.1154	0.8782	0.000466	0.965	2.257
kidney	RSA	14	-0.0389	1.0414	0.001160	0.440	3.760
			0.1240	0.8543	0.000472	0.976	3.033
hepatoma	RSA	17	-0.0713	1.0644	0.001109	0.440	2.404
			0.0885	0.8980	0.000506	0.903	1.748
liver	AFP	14	-0.0543	1.0510	0.001344	0.440	2.204
			0.0628	0.9280	0.000736	0.745	1.718
kidney	AFP	12	-0.0643	1.0542	0.001317	0.440	3.370
			0.0925	0.8884	0.000498	1.000	2.128
hepatoma	AFP	17	-0.0663	1.0593	0.001420	0.440	2.932
			0.0742	0.9110	0.000673	0.846	2.422

* sheared to 300-400 nucleotides. The first values are with n fixed, the second with n allowed to vary.

+ 1000-2200 nucleotides (7). RSA=rat serum albumin. AFP=alpha feto-protein.

++ N = number of points.

+++ %E = root mean square error (RMS) divided by data mean.

Smith, Britten and Davidson (20) observed that while no simple physical meaning can be associated with the exponent n, equation (2) is useful for data reduction.

For incomplete reactions, or reactions consisting of multiple components, COTFIT uses the DIGEST function

$$F(t) = U + \sum_{i=1}^{m} Fi(1 + KiCot)^{-n} \qquad (3)$$

where U is the fraction unreacted (single-stranded) at infinite time, m is the number of components, and Fi and Ki are the fraction and rate for component i. Table 1 contains the best-fit parameters of this function for the data of Sala-Trepat et al (6) when the exponent n is fixed at 0.44 or allowed to vary. The unfixed exponents vary from 0.745 to 1. This shift toward second-order kinetics may be due to either the different driver and tracer lengths used, or to the interrupted nature of the

Table 2. Comparison of melt fitting methods using data from Wallace et al (8)

Difference	Mean	StdDev	%Mean relative to A
$\lvert F(B) - F(A)\rvert$	0.0167	0.0285	42.0
$\lvert Tm(B) - Tm(A)\rvert$	1.1	2.9	1.99
$\lvert S(B) - S(A)\rvert$	2.1	1.1	33.0
Parameter	Mean	StdDev	Maximum
S(A)	6.2	2.4	11.8
S(B)	4.4	1.6	7.7
%E(A)	5.23	1.32	7.41
%E(B)	1.22	0.41	1.92

Initial absorbances were 0.9-1.3 A260 units at 25°C. Data was collected every 0.4° while heating at 0.25°/min.
A. Parameters for Gaussian curves fit to 15-point cubic-quartic first derivative of absorbance data (using GELFIT).
B. Parameters for equation (1) fit to the absorbance data normalized to fraction hyperchromicity (using COTFIT).

albumin and AFP genes in rat DNA (T.D. Sargent, J.R. Wu, J. Sala-Trepat, R.B. Wallace, T. Reyes, and J. Bonner, manuscript in preparation). In either case, the results demonstrate the need for careful determination of the exponent for individual S1 assayed experiments, since the calculated K values may vary by greater than 2.5 fold. Values of n significantly different from the expected value may indicate inaccurate tracer and driver length determinations or other systematic error.

Analysis of DNA and chromatin thermal denaturation data

Since many estimates of chromatin and DNA melting component parameters have been made by fitting Gaussian curves to derivatized data, we have compared this method to the use of the COTFIT MELTFN function. Data from our previously published melts of various chromatin and nucleosome samples (8), which had been derivatized with MELSMR and fit with GELFIT, were re-fit using COTFIT. As Table 2 shows, the two methods yield similar results, the Tm's differing by an average of only 2%. However, COTFIT yields an average error almost five times lower than that produced by GELFIT. Thus, although the MELSMR/GELFIT method has the advantage of ease of visual interpretation of plots, the COTFIT method produces more accurate parameter estimates.

Table 3. Rat liver chromatin melting transitions

	Sample 1	Sample 2	Sample 3	Mean	StdDev	%StdDev
I	0.0068	0.0033	0.0047	0.0049	0.0018	36
F1	0.0989	0.1034	0.1076	0.1033	0.0044	4
Tm1	65.5	60.7	56.4	60.9	4.5	7
S1	9.5	7.9	6.2	7.9	1.6	21
F2	0.0789	0.0978	0.0508	0.0758	0.0237	31
Tm2	76.9	76.8	69.1	74.3	4.5	6
S2	5.8	5.2	3.7	4.9	1.1	23
F3	0.0742	0.0811	0.0652	0.0735	0.0080	11
Tm3	77.9	77.4	76.4	77.2	0.8	1
S3	6.0	9.7	8.4	8.0	1.9	24
F4	0.0660	0.0431	0.1188	0.0760	0.0389	51
Tm4	82.4	82.6	79.8	81.6	1.6	2
S4	2.1	1.9	4.2	2.7	1.3	46
%E	1.67	1.09	1.62	1.46	0.30	22

Melts were performed in 0.25 mM EDTA pH 8. Initial absorbances were 0.7-1.5 A260 units at 25°C. Data were collected from 25-95°C every 0.4° while heating at 0.25°/min, and fit to equation (1) using COTFIT. The average percent standard deviations in F, Tm, and S are 24, 4, and 29, respectively.

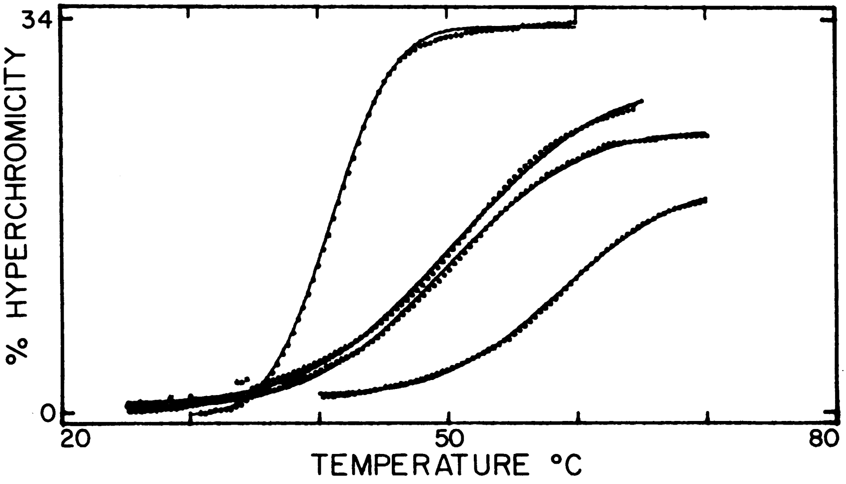

Figure 1. Fitting of DNA melting transitions using MELTFN. From left to right are rat liver DNA and oligomers A (every 4th point shown), C (every other point shown), and B (every other point shown) (see Table 4).

Table 4. DNA melting transitions

	I	F1	Tm1	S1	%E
I. Unsheared Rat liver DNA. 1.1-1.6 A260 units/ml, 0.25 °C/min.					
A. 0.25 mM EDTA pH 8 0.4 °C/pt, 30-60 °C.					
Mean (5)	-0.0011	0.3369	43.6	3.86	1.88
StdDev	0.0040	0.0108	3.9	0.47	0.81
%StdDev	378	3.2	8.9	13.1	43.4
B. 0.01 x SSC pH 7 0.4 °C/pt, 35-75 °C.					
Mean (2)	0.0064	0.3554	58.2	6.09	1.83
StdDev	0.0008	0.0069	0.1	0.28	0.01
%StdDev	12.6	2.0	0.2	4.6	0.4
II. Synthetic oligonucleotides. 1 M NaCl 0.01 M PB, 0.5 °C/min.					
A. CCGAATTCGG GGCTTAAGCC 0.77-0.78 A260 units/ml, 0.1 °C/pt, 25-65 °C.					
Mean (2)	0.0049	0.2808	50.3	9.28	3.18
StdDev	0.0019	0.0223	0.4	0.49	0.39
%StdDev	38.1	7.9	0.8	5.2	12
B. GGATCACCGCC CCTAGTGGCGG 0.72-0.77 A260 units/ml, 0.2 °C/pt, 40-70 °C.					
Mean (3)	0.0151	0.1817	58.6	7.26	2.06
StdDev	0.0012	0.0050	0.2	0.40	0.56
%StdDev	7.9	2.7	0.3	5.5	27
C. CATGAATTCATG GTACTTAAGTAC 0.55-0.57 A260 units/ml, 0.2 °C/pt, 25-70 °C.					
Mean (2)	0.0047	0.2396	49.7	8.34	2.29
StdDev	0.0006	0.0118	0.3	0.32	0.29
%StdDev	12.8	4.9	0.6	3.8	13

To test the reproducibility of this function, data from melts of rat liver chromatin, rat liver DNA, and synthetic oligonucleotides were fit using COTFIT (Tables 3 and 4 and Figure 1). The rat liver chromatin and DNA melts are of different sample preparations, and the oligonucleotide melts are of different samples from the same preparation. All melts were from separate runs. Some fluctuations in Tm resulting from differences in buffer concentration can be seen for the 0.25 mM EDTA melts, but the standard deviation in Tm is still less than 10%. The deviations for the other melts are much smaller.

The closeness of the data and fit is in accordance with the shape predicted by theoretical treatments of nucleic acid melting (21-24), although there is no immediate correlation between the parameters of equation (1) and the physical parameters of the system. The agreement of the rat DNA melt data with equation (1) is probably due to the variation in nucleotide composition of rat DNA, and a resulting combination of a Gaussian distribution of small transitions (25). Deviations from the fitted curve are more apparent for the oligonucleotide melts, as might be expected. In light of the ease of estimation of the parameters of equation (1), our results lend support to its use in fitting nucleic acid thermal denaturation data, especially for comparative purposes.

CONCLUSIONS

We have described a set of flexible, interactive programs for the analysis and storage of biochemical data. The ability of the programs to accept English commands and prompt the operator for needed information allows even an inexperienced computer user to analyze a reassociation curve or gel profile in under an hour. In addition to the analysis of data from nucleic acid hybridization, thermal denaturation, and gel electrophoresis (1-10), the programs may be used for a number of other applications, such as resolution of components in velocity and equilibrium density gradients and the determination of rate constants for enzyme reactions. The programs also provide a framework for the development of other data analysis systems.

The programs described in this paper are available from the authors on a variety of machine-readable media.

ACKNOWLEDGMENTS

We thank T.D. Sargent, J.W. Posakony and R.B. Wallace for helpful discussion and critical review of this manuscript. This research was supported by U.S. Public Health Service Grants GM-13762 and GM-00086.

REFERENCES

1. Garrard, W.T., Pearson, W.R., Wake, S.K., and Bonner, J. (1974) Biochem. Biophys. Res. Comm. 58, 50-57.

2. Murphy, R.F., and Bonner, J. (1975) Biochim. Biophys. Acta 405,62-66.

3. Gottesfeld, J.M., Murphy, R.F., and Bonner, J. (1975) Proc. Nat. Acad. Sci. USA 72, 4404-4408.

4. Murphy, R.F., Wallace, R.B., and Bonner, J. (1978) Proc. Nat. Acad. Sci. USA 75, 5903-5907.

5. Wallace, R.B., Dube, S.K., and Bonner, J. (1977) Science 198, 1166-1168.

6. Sala-Trepat, J.M., Sargent, T.D., Sell, S., and Bonner, J. (1979) Proc. Nat. Acad. Sci. USA 76, 695-699.

7. Sala-Trepat, J.M., Dever, J., Sargent, T.D., Thomas, K., Sell, S., and Bonner, J. (1979) Biochemistry, in press.

8. Wallace, R.B., Sargent, T.D., Murphy, R.F., and Bonner, J. (1977) Proc. Nat. Acad. Sci. USA 74, 3244-3248.

9. Bonner, J., Wallace, R.B., Sargent, T.D., Murphy, R.F., and Dube, S.K. (1978) Cold Spring Harbor Symp. Quant. Biol. 42, 851-857.

10. Bakke, A.C., Wu, J.R., and Bonner, J. (1978) Proc. Nat. Acad. Sci. USA 75, 705-709.

11. Pearson, W.R., Davidson, E.H., and Britten, R.J. (1977) Nucleic Acids Res. 4, 1727-1737.

12. Britten, R.J., Graham, D.E., and Neufeld. B.R. (1974) in Methods in enzymology (L. Grossman and K. Moldave, Eds.), Vol. 29E, pp. 363-418, Academic Press, New York.

13. Marquardt, D.W. (1963) J. Soc. Indust. Appl. Math. 11, 431-441.

14. Gautschi, W. (1972) in Handbook of Mathematical Functions (M. Abramowitz and I.A. Stegun, Eds.), pp. 295-329, Dover Publications, New York.

15. Martin, R.S., and Wilkinson, J.H. (1971) in Linear Algebra (J.H. Wilkinson and C. Reinsch, Eds.), pp. 50-56, Springer-Verlag, New York.

16. Savitzky, A., and Golay, M.J.E. (1964) Anal. Chem. 36, 1627-1639.

17. Steinier, J., Termonia, Y., and Deltour, J. (1972) Anal. Chem. 44, 1906-1909.

18. Madden, H.H. (1978) Anal. Chem. 50, 1383-1386.

19. Morrow, J. (1974) Ph. D. Thesis, Stanford University.

20. Smith, M.J., Britten, R.J., and Davidson, E.H. (1975) Proc. Nat. Acad. Sci. USA 72, 4805-4809.

21. Zimm, B.H. (1960) J. Chem. Phys. 33, 1349-1356.

22. Applequist, J., and Damle, V. (1963) J. Chem. Phys. 39, 2719-2721.

23. Applequist, J., and Damle, V. (1965) J. Am. Chem. Soc. 87, 1450-1458.

24. Abzel, M.Ya. (1973) Phys. Rev. Lett. 31, 589-593.

25. Abzel, M.Ya. (1979) Proc. Nat. Acad. Sci. USA 76, 101-105.