SPLINT: a cubic spline interpolation program for the analysis of fragment sizes in one-dimensional electrophoresis gels*

Cheryl E. Garipey, Margaret I. Lomax⁺ and Lawrence I. Grossman[§]

Department of Cellular and Molecular Biology, Division of Biological Sciences, University of Michigan, Ann Arbor, MI 48109, USA

Received 13 May 1985

ABSTRACT

SPLINT (SPLine INTERpolation) is a BASIC microcomputer program that uses electrophoretic data on the migration distances of DNA fragments of known size to approximate the relationship between distance moved and DNA fragment size by a piecewise cubic spline function. This function places all standards exactly on the calibration curve and then determines from it the sizes of other fragments in the gel. Special advantages of this program include the use of a digitizer to enter the fragment positions directly from a gel photograph and the optional generation of a graph of the approximating function plus the placement on it of points from any specified gel lanes.

INTRODUCTION

Computers are increasingly being applied routinely to nucleic acids research. One of these routine applications is the analysis of one-dimensional electrophoretic gels used in the separation of DNA fragments generated by restriction endonuclease digestion. Several methods have been proposed for approximating the relationship between DNA migration distance and size. It has been observed empirically that, for well behaved combinations of DNA fragment sizes and gel concentrations, a straight line results from a plot of logarithm of fragment size versus relative mobility (e.g. 1) or of size versus reciprocal relative mobility (2,3). Often, however, there exists a positive curvature for large fragments and a negative curvature for small fragments (4). This paper describes a program that accepts entry of fragment positions through a digitizer, determines a piecewise cubic spline function to approximate the relationship between the logarithm of known fragment size and fragment migration distance,

* A listing of the SPLINT program will be provided upon request accompanied by a self-addressed mailing label. Alternatively, the source program can be transferred to a blank tape cartridge for the Hewlett-Packard Series 9800 computer, which should be accompanied by a mailing label and return postage.

optionally graphs this function, and determines the sizes of unknown DNA fragments in the gel, with an average error of about 2.5%. The program is also applicable to the migration of other macromolecules, such as proteins in SDS-polyacrylamide gels.

ALGORITHM

SPLINT is an adaptation of a FORTRAN program by de Boor (5). It uses a piecewise cubic spline for efficient and smooth approximation of the relationship between migration distance and logarithm of size. After reading in both the migration distances $[m_1, m_2, \dots, m_N]$ and the sizes of the standards $[g(m_1), g(m_2), \dots, g(m_N)]$, the program constructs a piecewise cubic function, f , approximating g , where g is the true relationship between size and band position.

On each interval $[m_i, m_{i+1}]$ the program makes f agree with a fourth order polynomial, P_i . That is, it sets $f(x) = P_i(x)$ for $m_i \leq x \leq m_{i+1}$ ($i=1, \dots, N-1$) where P_i is a fourth order polynomial. P_i is then selected such that

$$P_i(m_i) = g(m_i),$$

and

$$P_i(m_{i+1}) = g(m_{i+1}) \quad i = 1, \dots, N-1.$$

This demands that the approximation function must give the correct sizes of the standards. P' is then calculated such that

$$P'_i(m_i) = s_i,$$

and

$$P'_i(m_{i+1}) = s_{i+1} \quad i = 1, \dots, N-1.$$

The s_i ($i = 1, \dots, N$) are the slopes. The equations impose the condition that the approximating pieces have equal slope at the standard point between them. The s_i are determined from the condition that f (which is made up of the P_i 's) must be twice continuously differentiable. This results in enough conditions to determine the desired function provided only that s_1 and s_N are chosen. Given s_1 and s_N , the program has a system of $N-2$ equations for $N-2$ unknowns ($s_2, \dots, s_{N-1}$) from which it can determine the unique interpolating function, f .

SPLINT uses the "not-a-knot" condition to determine s_1 and s_N . The program chooses s_1 and s_N so that $P_1 = P_2$ and $P_{N-1} = P_N$. This means that there are, in fact, $N-3$ polynomial pieces making up f (rather than $N-1$), but that there is a specific point through which the first and last

FIGURE 1. Calibration curve for the data shown in Table 1. The curve was generated by SPLINT using the piecewise cubic spline program. Symbols (●) represent the standards described in the Table legend.

polynomial pieces must travel. The first and last polynomial pieces interpolate g (the true function relating logarithm of size to distance migrated) at an additional point that is not a break point.

The system specified above has only one solution and the resulting function proves to be a very good approximation to g .

USING SPLINT

SPLINT was written for the Hewlett-Packard (HP) 9825A computer in standard HP BASIC. It should readily be transferrable with minor modifications to other common laboratory microcomputers, such as the IBM PC. It is used here with a Numonics Corp. (Lansdale, PA) model 224 digitizer, an HP model 9871A printer, and an HP model 9872A plotter. The resolution of the digitizer is 0.25 mm.

A. Entering the standard data

The migration data for the standards are entered via the digitizer from a photograph of the gel by touching each band or lane origin with the cursor. The origin of the lane is entered first and then each fragment is entered in order of increasing distance from the origin. SPLINT calculates each distance traveled by taking the difference between the origin and the datum point. The sizes of the standards are then entered by specifying

Table 1. Accuracy of Fragment Size Estimation Using SPLINT

Fragment	Digest	g(m)	f(m)	g(m)-f(m)	error, %
1	Xba I/Xho I	15004	16713	+ 1713	11.4
2	Ava I	14677	13634	- 1043	7.1
3	Acc I	13070	13634	+ 564	4.3
4	Acc I	11828	10850	- 978	8.3
5	Xba I/Xho I	8990	9008	+ 18	2.0
6	Ava I	8614	8442	- 172	2.0
7	Acc I	6957	6741	- 216	3.1
8	Ava I	6888	6741	- 47	0.7
9	Acc I	5581	5491	- 82	1.5
10	Ava I	4718	4685	- 33	0.7
11	Ava I	3730	3715	- 15	0.4
12	Acc I	3574	3585	+ 11	0.3
13	Acc I	2720	2735	+ 15	0.5
14	Acc I	2190	2189	- 1	0.0
15	Ava I	1881	1872	- 9	0.5
16	Ava I	1674	676	+ 2	0.1
17	Ava I	1602	1580	- 22	1.4
18	Acc I	1444	1473	+ 29	2.0
19	Acc I	639	649	+ 10	1.6
mean error = 2.5% ± 3.1					

Fragments from the indicated digests of bacteriophage lambda DNA were displayed on 0.5% agarose gels run for 15 hr at 1.22 V/cm. Gels were photographed after staining with ethidium bromide and the photographs used for digitizing band positions. Standards were lambda DNA digested with Hind III plus phi-X 174 DNA digested with Hae III. g(m) indicates true fragment size (18) and f(m) is the size estimated by SPLINT, both in base pairs.

either one of the sets of sizes for commonly used standards stored in SPLINT, or each fragment size individually through the keyboard. The program allows the user to review the migration distances and sizes of the standards and correct the sizes if desired.

B. Generating the curve

When the user indicates that the standard sizes are entered correctly the program will print the standard migration distances and sizes and determine the cubic spline function. The user may then optionally generate

a graph of this function on the plotter (Figure 1).

C. Interpolation for unknowns

SPLINT asks the user to designate a name for the lane whose fragment sizes are to be determined. The user then enters the migration distances of the DNA fragments of that lane through the digitizer in the same manner as the standards were entered. The program will determine the fragment sizes and will print them along with the lane name and the fragment migration distances on the page already containing the standard data. If the migration distance of a fragment falls outside of the interpolation range, i.e. it is greater than that of the smallest standard fragment but less than that of the largest standard fragment, the program will indicate that there is a fragment outside the interpolating range and will give that fragment's size as zero in the lane data table.

If the user asked for the generation of a graph when the calibration curve was determined, the program will now offer the option of marking on that graph the interpolation points used in the current lane. The specific symbol used will be printed below the lane data for easy reference. When finished with a lane the user names another lane and repeats the process if desired. A maximum of 15 lanes with up to 20 fragments each may be interpolated in this manner.

D. Accuracy of results

We have used SPLINT to evaluate the sizes of restriction fragments in the range 0.64 to 15 kb generated from sequenced DNAs. Table 1 shows 19 fragments, generated from bacteriophage lambda DNA by various digests, whose sizes were determined by comparison to a Hind III digest of lambda DNA and a Hae III digest of phi-X 174 DNA (Figure 1). The overall mean percent error ± 1 standard deviation is seen to be 2.5 ± 3.1 . Not surprisingly, relatively large fragments are determined with the greatest error; for instance, the four fragments larger than 10 kb have a mean percent error of 7.8 ± 2.9 while the remaining fifteen yield 1.1 ± 0.9 . Similar results were seen when other standards were used.

DISCUSSION

We have presented a program in standard BASIC, SPLINT, for rapidly and accurately entering electrophoretic fragment positions through a digitizing tablet and interpolating their sizes relative to standards by an algorithm that places each standard fragment on the interpolating curve. The shape of the standard curve and the position on it of any fragment set can be

viewed through optional graphic output.

SPLINT has obvious advantages in speed and ease over interpolation of fragment sizes by hand. More importantly, it also provides greater accuracy and consistency. Although other algorithms have proven useful (6-10), SPLINT, unlike these programs, places every standard point on its interpolating function by the use of a piecewise cubic spline algorithm. The apparent impetus for allowing standard curves to be drawn that do not contain every standard point is to allow one to draw a graph of the desired shape. However, if a specific standard fragment shows an unexpected migration distance in a given gel, it seems likely that well-behaved fragments of similar size in the same gel would exhibit similarly unusual migration behavior. Since peculiarities in the gel are indicated by the positions of the standards, there seems little justification for not insisting that the interpolating function agree exactly with the standards.

Cases have been documented in both agarose and acrylamide gels in which fragment migration does not reflect true size. These include a processing intermediate in the excision of intron RNA from primary transcripts (i.e. lariats) (11), fragments containing various bound antibiotics (12), a kinetoplast DNA fragment containing a bent helical structure (13,14), and the migration of a bacteriophage lambda DNA Eco RI fragment in polyacrylamide gels (15), reflecting an effect of base composition (16). These cases have in common the fact that fragment migration is aberrant because the fragments themselves are not well-behaved. Indeed, in several cases it was novel migration behavior that first pointed to an unusual feature of nucleic acid structure. However, as long as the standards employed are known to be well-behaved, such cases do not detract from the desirability of generating an interpolating function that contains all of the standard points.

Several programs based on different standard curve algorithms for interpolating the sizes of DNA fragments have been written for a pocket calculator and systematically compared (17). These algorithms include a piecewise cubic spline function; a best fit function determined by least squares approximation using a sequence of orthogonal polynomials; and three functions obtained by joining the standard points with straight lines, where the standard points are plotted directly, as logarithm of size versus band position, and as size versus inverse band position. The algorithms were compared to each other and to interpolation by hand. The piecewise

cubic spline algorithm was found the most accurate, with an overall error of 1.8%. Our value of 2.5% appears comparable, particularly given the relative choices of data sets.

ACKNOWLEDGEMENTS

We thank Ronald J. Lomax for critically reading the mathematical portion of the manuscript and Marilyn Galler for typing. This work was supported by grants from NSF (PCM-8118915) and NIH (GM 26546).

§To whom correspondence should be sent at present address: Science, 1333 H Street NW, Washington, DC 20005, USA

+Present address: Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

REFERENCES

1. Helling, R.B., Goodman, H.M. and Boyer, H.W. (1974) *J. Virol.* 14, 1235-1244.
2. Southern, E. (1979) *Meth. Enzymol.* 68, 152-176.
3. Schaffer, H.E. and Sederoff, R.R. (1981) *Anal. Biochem.* 115, 113-122.
4. Moore, K.H., Johnson, P.H., Chandler, S.E.W. and Grossman, L.I. (1977) *Nucl. Acids Res.* 17, 1273-1289.
5. de Boor, C. (1978) *A Practical Guide to Splines*, pp. 57-59, Springer-Verlag, New York.
6. Gray, A.J., Beecher, D.E. and Olson, M.V. (1984) *Nucl. Acids. Res.* 12, 473-492.
7. Kieser, T. (1984) *Nucl. Acids Res.* 12, 679-688.
8. Rood, J.I. and Gawthorne, J.M. (1984) *Nucl. Acids Res.* 12, 689-694.
9. Maina, C.V., Nolan, G.P. and Szalay, A.A. (1984) *Nucl. Acids Res.* 12, 695-702.
10. Russell, P.J., Crandall, R.E. and Feinbaum, R. (1984) *Nucl. Acids Res.* 12, 493-498.
11. Grabowski, P.J., Padgett, R.A. and Sharp, P.A. (1984) *Cell* 37, 415-427.
12. Loucks, E., Chaconas, G., Blakesley, R.W., Wells, R.D. and van de Sande, J.H. (1979) *Nucl. Acids Res.* 6, 1869-1879.
13. Marini, J.C., Levene, S.D., Crothers, D.M. and Englund, P.T. (1982) *Proc. Natl. Acad. Sci. USA* 79, 7664-7668.
14. Wu, H.-M. and Crothers, D.M. (1984) *Nature* 308, 509-513.
15. Thomas, M. and Davis, R.W. (1975) *J. Mol. Biol.* 91, 315-328.
16. Zeiger, R.S., Salomon, R., Dingman, C.W. and Peacock, A.C. (1972) *Nature New Biol.* 238, 65-69.
17. Gough, E.J. and Gough, N.M. (1984) *Nucl. Acids Res.* 12, 845-853.
18. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Peterson, G.B. (1982) *J. Mol. Biol.* 162, 729-773.