Nucleic Acids Research, 1994, Vol. 22, No. 20 4321–4325

A program for selecting DNA fragments to detect mutations by denaturing gel electrophoresis methods

Stephen Brossette and Roger M.Wartell*
School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received April 25, 1994; Revised and Accepted August 22, 1994

ABSTRACT

A computer program was developed to automate the selection of DNA fragments for detecting mutations within a long DNA sequence by denaturing gel electrophoresis methods. The program, MELTSCAN, scans through a user specified DNA sequence calculating the melting behavior of overlapping DNA fragments covering the sequence. Melting characteristics of the fragments are analyzed to determine the best fragment for detecting mutations at each base pair position in the sequence. The calculation also determines the optimal fragment for detecting mutations within a user specified mutational hot spot region. The program is built around the statistical mechanical model of the DNA melting transition. The optimal fragment for a given position is selected using the criteria that its melting curve has at least two steps, the base pair position is in the fragment's lowest melting domain, and the melting domain has the smallest number of base pairs among fragments that meet the first two criteria. The program predicted fragments for detecting mutations in the cDNA and genomic DNA of the human p53 gene.

INTRODUCTION

Denaturing gel electrophoresis methods are widely used to detect point mutations or polymorphisms in DNA fragments. The methods include denaturing gradient gel electrophoresis (DGGE; 1,2), temperature gradient gel electrophoresis (TGGE) in both horizontal (3,4) and vertical formats (5,6), constant gradient gel electrophoresis (CDGE; 7) and temperature sweep gel electrophoresis (TSGE; 8).

The above methods utilize DNA melting properties to separate fragments differing by as little as 1 bp substitution or mismatch. A duplex DNA migrates in a polyacrylamide gel until it reaches a temperature or denaturant concentration which induces the least stable domain to unwind. The branched structure of a partially denatured DNA results in a large decrease in gel mobility. Homologous DNAs with different stabilities in their first melting domain unwind at different depths in the gel. Base pair changes are detectable if they are in the first melting domain and the domain unwinds at a temperature or denaturant concentration well below the last stage of strand unwinding. Although mutations may also be detected in all but the last temperature domain (9), we have focused attention on first domain melting (6).

Computer simulations of DNA melting behavior are frequently employed to predict if a DNA fragment can be used to detect mutations. DNA melting maps (1) or melting profiles (6) can determine if a designated base pair or base pair region is in an early melting domain, and derivative melting curves determine if a melting domain is separated from the final melting step. A G+C rich segment (a GC clamp) is often added to one end of a DNA to stabilize it and promote a two-step melting process (10,11). Computer simulations provide a way to select a fragment for detecting mutations with a minimum GC clamp length. Simulations will also be useful if psoralen chemi-clamps are employed (12). In the latter case, modifications to the theory are required to accommodate an effective T_m for the chemi-clamp, and the elimination of strand dissociation.

The current approach for selecting a DNA fragment involves calculating the melting behavior of trial fragments surrounding the region of interest. This procedure is straightforward when applied to one or a few regions. It can be laborious if one wishes to determine a large set of DNA fragments to detect mutations throughout a gene. The program described in this paper scans through a long DNA sequence and selects optimum fragments for detecting mutations at each base pair position. The program also selects the optimum fragment for detecting mutations within a user specified region in the sequence. The program, called MELTSCAN™, should be useful to workers applying DGGE, TGGE and related methods to detect mutations in DNA.

METHODS AND RESULTS

MELTSCAN is based on the statistical mechanical model of the DNA helix–coil transition (1,6,13,21). In this model the stability of each base pair depends on its base pair type (AT or GC) and stacking interactions with neighboring base pairs. The model includes a loop entropy term for opening an internal segment of base pairs, and the dissociation equilibrium of partial duplex to single strands. Thermodynamic parameters employed were from recent DNA melting studies; stacking interactions from Delcourt and Blake (14), dissociation parameters and loop entropy from McCampbell *et al.* (15). DNA melting transitions were

*To whom correspondence should be addressed

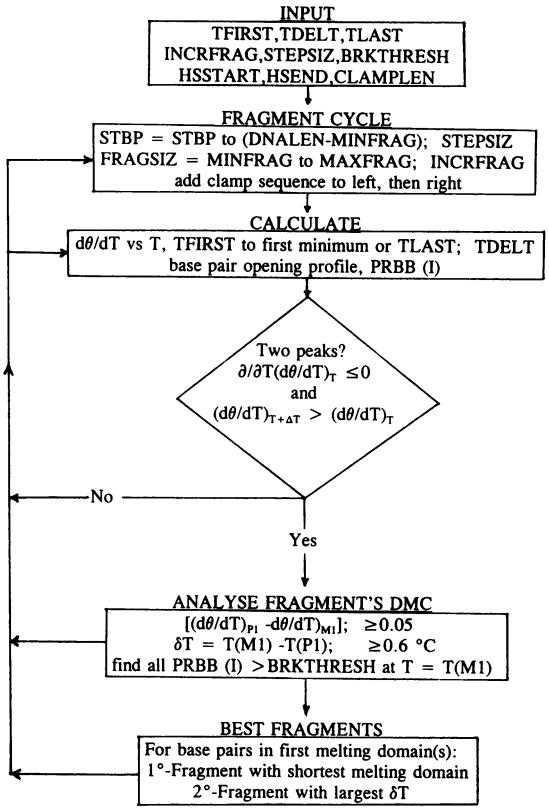

Figure 1. Flow chart outlining the parameters and procedures used in the computer program MELTSCAN. See text for definition of parameters and description of program.

calculated for a solution of 0.1 M Na^+. To estimate the mid-temperature in an experimental temperature gradient gel with 60 % denaturant (4.2 M urea and 24% formamide), 46°C was subtracted from the predicted T_m of the first melting domain (6). This empirical shift is consistent with observations that urea and formamide lower DNA melting temperatures by an amount that is, to a first approximation, independent of base composition (1,9).

Computer algorithm

Figure 1 shows the flow diagram of MELTSCAN. The program was written in FORTRAN-77 and compiled using the Microsoft Corp. Powerstation software. The hardware required is an IBM-PC or IBM-PC compatible personal computer with a 80386 or higher processor running an MS-DOS version 3.3 or later operating system. The Microsoft Corp. Powerstation run-time file DOSXMSF.EXE is also required. Program execution initiates from a batch file with the command 'MELTSCAN SEQNAME'. SEQNAME is the user-defined DNA sequence file. The first line of this file is for descriptive information of the sequence. The remaining lines list the DNA sequence to be analyzed. Blank spaces, numbers, or characters other than A, T, C, and G in the second and subsequent lines of SEQNAME are ignored. The following paragraphs describe how the program works. A user-oriented description of an interaction with the program with sample input parameters and acceptable parameter limits is shown in Table 1.

Table 1. Typical user application of MELTSCAN. User input underlined. Comments within (...)

```
C:>MELTSCAN p53cdna <CR>   (The sequence file here is called p53cdna)
                           (Sequence file must not have an extension in name)

MELTSCAN DNA SEQUENCE ANALYSIS PROGRAM
          .          .
          .          .
TYPE IN VALUES FOR
TFIRST,TDELT,TLAST
78,0.5,94 <CR>         (Values for TDELT; 0.1 to 0.5)
                       (Values for TFIRST/TLAST; 75-80 °C/93-99°C)
                       (TFIRST & TLAST may be adjusted to suit %GC of fragments)
                       (78 to 94 °C works for 46 < %GC < 68)
                       (Extremes of TFIRST AND TLAST may cause program to crash)
TYPE IN VALUES FOR
STEPSIZ,INCRFRAG,BRKTHRESH
20,20,0.5<CR>          (Values for STEPSIZ or INCRFRAG; 2 to 40)
                       (Values for BRKTHRESH; .3 to .7)

TYPE SEQ NUMBER OF FIRST BP IN HOTSPOT (HSSTART)
654<CR>                (Range for HSSTART; 1 to DNALEN-1)
                       (If no hotspot examined, use 100)

TYPE SEQ NUMBER OF LAST BP IN HOTSPOT (HSEND)
669<CR>                (Range for HSEND; {HSSTART +1} to DNALEN)
                       (If no hotspot examined, use 101)
TYPE GC CLAMP LENGTH
20<CR>                 (Range for GC clamp; 0 to 60)

 -----ANALYSIS PARAMETERS------
(The program prints the input parameters to the screen. It then prints)

   ---SCREEN PROGRESS INDICATOR---
          1          90
(The first number indicates the program is at bp position 1 and is starting to calculate the
properties of the shortest fragment, 90 bp. As the run proceeds, new lines are printed. They
indicate the program has reached the next position to calculate properties of a 90 bp fragment.
When the run ends, 'stop terminated'is printed. The output file listing the results is called
SEQNAME.OF where the user's sequence name replaces SEQNAME.)
```

The program prompts the user for the input parameters listed in the top box of Figure 1. The algorithm analyzes the DNA sequence file from its first base pair to its last base pair, DNALEN. DNALEN is evaluated when the sequence file is read in. TFIRST, TDELT, and TLAST are the first temperature, the temperature increment, and the last temperature used in calculating the derivative melting curves of the fragments. Values of 78, 0.5, and 94°C were used. The temperature range (78−94°C) adequately evaluated the melting behavior of DNA fragments while minimizing calculation time. Similarly, a value of 0.5°C for TDELT gave sufficient resolution for the derivative melting curves while minimizing calculation time. INCRFRAG is the increment in the DNA fragment length used in the fragment cycle, and STEPSIZ the step size between new starting base pairs in the DNA fragment cycle (described below). Values of 10 or 20 bp were used for the latter two parameters. BRKTHRESH is a threshhold value that is compared with the calculated probability of being open for each base pair in a fragment. It defines when a base pair is considered broken, i.e. non-hydrogen bonded and unstacked. A value of 0.5 was used. The apparent melting temperature for a base pair, T'_m, is the temperature at which the probability of being open exceeds BRKTHRESH. HSSTART and HSEND are the base pair positions in the DNA sequence that start and end the user designated 'hot spot' region. A hot spot region is where mutations/polymorphisms are anticipated. CLAMPLEN is the length of the GC clamp which may be added to the fragment ends. A GC clamp segment may range from 0 to 60 bp and is selected from the 60 bp sequence used by Abrams *et al.* (2). MELTSCAN seeks DNA fragments that contain the entire hot spot region in the first melting domain.

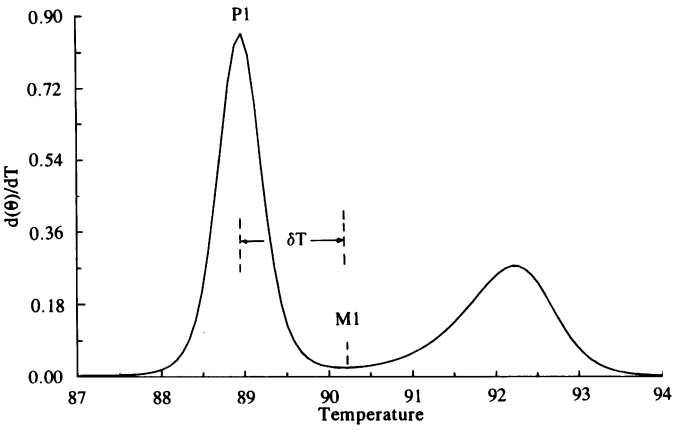

Figure 2. The derivative melting curve, DMC, of the 310 bp DNA sequence from position 311 to 600 in the p53 cDNA (16) with a 20 bp GC clamp left of 311. P1 is the peak of the first melting domain and δT is the temperature interval from P1 to the first minimum M1.

Table 2. Example output; analysis of cDNA p53 sequence and hot spot B. Using 20 bp GC lamp

```
>>>>>>>>>> HOT SPOT ANALYSIS<<<<<<<<<<<<<<<
HOT SPOT:
    STARTING BP:          654
    ENDING BP  :          669

OPTIMAL FRAGMENT FOR HOT SPOT :
          581 --->        690
    LENGTH (W/OUT G-C CLAMP) =          110
    G-C CLAMP ON LEFT
    TEMP AT DTHETA MAXIMUM =          91.000000
    DELT T =                           2.000000
    NUMBER OF OPEN BASE
    PAIRS AT DTHETA MIN    =           78
```

BASE PAIR NO.'S		--- OPTIMAL FRAGMENT --- START BP	END BP	LENGTH	G-C CLAMP ON	DELTA T	NO. BP OPEN AT TMIN
1 -->	115	0	0	1		.0	600
116 -->	135	101	490	390	RIGHT	2.5	232
136 -->	155	121	430	310	RIGHT	2.5	212
156 -->	174	141	390	250	RIGHT	3.0	192
175 -->	175	1	250	250	LEFT	3.0	76
176 -->	177	1	230	230	LEFT	5.0	55
178 -->	178	101	230	130	LEFT	4.0	53
179 -->	215	121	230	110	LEFT	3.5	52
216 -->	235	61	250	190	LEFT	2.5	75
236 -->	255	1	270	270	LEFT	3.0	95
256 -->	275	241	390	150	RIGHT	4.0	92
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1236 -->	1255	941	1270	330	LEFT	.5	202
1256 -->	1275	941	1290	350	LEFT	.5	222
1276 -->	1295	941	1310	370	LEFT	.5	242
1296 -->	1317	0	0	1		.0	600

```
Stop
```

We have applied MELTSCAN to sequences assuming hot spot regions 50 bp or less.

The second box of the flow diagram of Figure 1, FRAGMENT CYCLE, outlines how the program cycles through DNA fragments. The first DNA fragment contains the sequence from the starting base pair, STBP = 1, to (MINFRAG-CLAMPLEN) base pairs to the right. MINFRAG is the minimum fragment length examined. It was set at 110 bp. The GC clamp segment is added to the left end of the first fragment and a melting curve is calculated. The calculation is then repeated after moving the GC clamp to the right end of the fragment. The fragment length is then increased by adding to the right end the next INCRFRAG base pairs of the parent sequence. Melting calculations are again made with the GC clamp on each end. The length of the fragment is continuously incremented by INCRFRAG base pairs and two melting curves are calculated for each length until the fragment reaches or exceeds (MAXFRAG-CLAMPLEN). MAXFRAG was set at 520 bp. The starting base pair, STBP, is then moved to the left by STEPSIZ base pairs and the next series of fragments from 110 to 520 bp long are examined.

The core of the program is described from the CALCULATE box of Figure 1 to the end of the flow diagram. Derivative melting curves are calculated for all DNA fragments generated by the fragment cycle. The melting characteristics of the fragments are analyzed to find the best fragment for detecting mutations at each base pair position in the sequence. The program also determines the optimal fragment for the user specified segment of contiguous base pairs, i.e. the hot spot region.

The initial screening of DNA fragments looks for fragments with derivative melting curves (DMC) that have two peaks. A DMC is a plot of the temperature derivative of the fraction of intact base pairs, $d\theta/dT$, vs temperature, T. If the slope of $d\theta/dT$ at temperature T is negative or zero and then becomes positive at a higher temperature, there are at least two peaks in a DMC (Figure 2).

If the above analysis indicates two peaks in a DMC, the next step is to look for a clear separation between the first and second melting peak. This selection is described in the box ANALYSE FRAGMENT's DMC. A DNA must have a DMC with a difference between $d\theta/dT$ at the first peak (P1) and the first minimum (M1) of at least 0.05, and a temperature interval, δT, between the first peak and first minimum of at least 0.5°C (Figure 2). The values of 0.05 and 0.5°C are arbitrary, but were chosen based on observations of experimental DMCs of DNAs with two or more steps (13,15). We note that the above conditions do not actually determine the temperature difference between the first two peaks in a DMC. The selection conditions are, however, strongly correlated with well separated peaks. A test of these values was made by plotting DMCs of 33 DNA fragments selected by the algorithm (from Table 2). All show two (or more) peaks separated by at least 0.8°C. Since the selection decision is made at the first minimum of a DMC rather than the second peak, execution time is reduced.

Base pairs contributing to the first melting peak are determined from the calculated probability of each base pair being melted, PRBB. The file PRBB(I) contains the probability that each base pair (I = 1, 2, ...*n*) is melted at the calculated temperatures. A base pair is considered to be within the first melting peak if PRBB(I) is greater than BRKTHRESH at temperature M1. DNA fragments with two or more melting peaks are stored and indexed by the base pairs in their first melting domains. Figure 3a shows a melting map for the 310 bp DNA sequence used in the example of Figure 2. The apparent melting temperature, T'_m, for each base pair is shown. For the 310 bp DNA, the temperature M1 is 90.2°C. Base pairs considered to be in the first melting domain are, from Figure 3a, between positions 120 and 310. It may be worth noting that the calculation used for Figure 3a does not consider the duplex-to-single-strands dissociation (1,6). One cannot rely on melting maps alone to accurately predict the melting behavior of late melting domains (6). This is indicated by the observation that base pairs unwinding under the second melting peak in Figure 2 melt at ≈92.2°C, while T'_m values for these base pairs as indicated by Figure 3a are ≈94°C.

After the melting characteristics of all DNA fragments are calculated, the program scans through each base pair position and compares the characteristics of DNA fragments that have that base pair in their first melting domain. The 'best' fragment for a given position is selected using one of two characteristics. The first characteristic sought is the DNA fragment with the smallest number of base pairs in its first melting domain. This

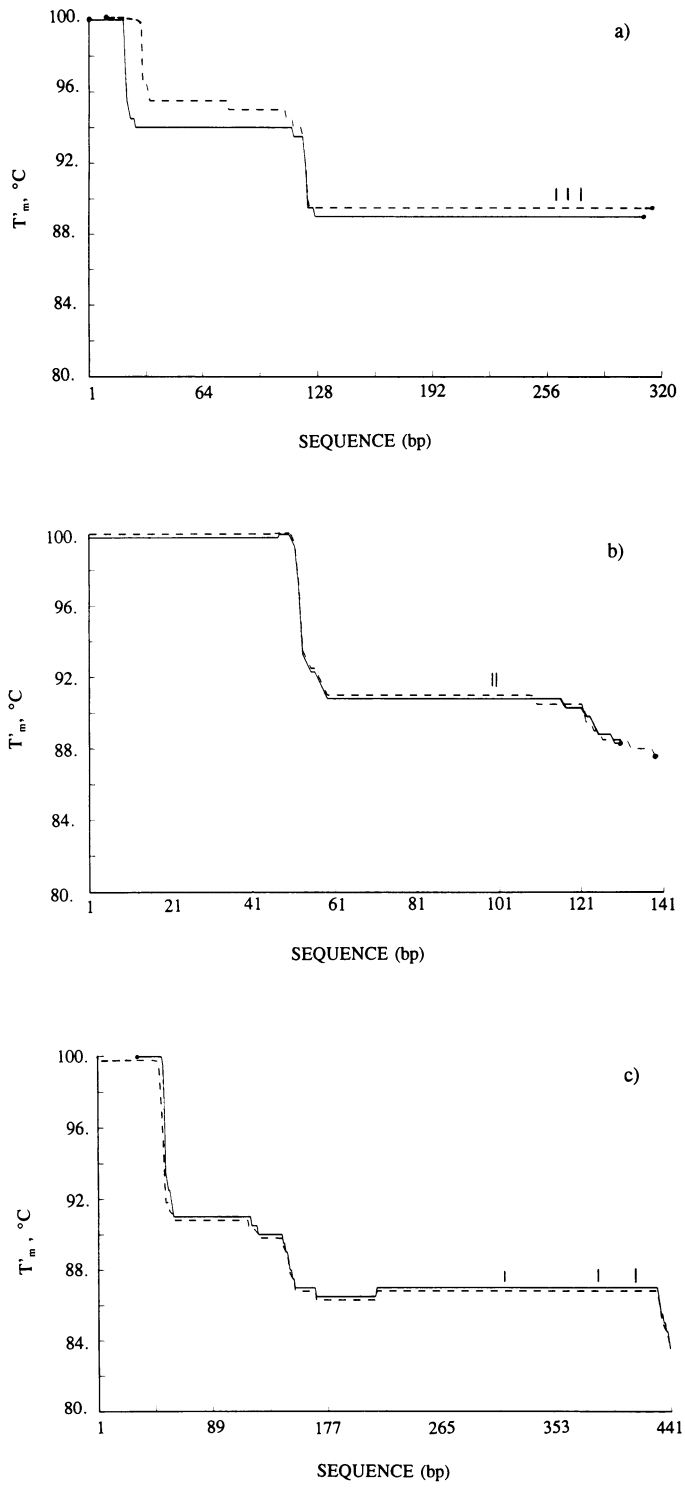

Figure 3. Melting maps of three DNA fragments determined by MELTSCAN for p53 cDNA hot spots and related DNA fragments used in ref. 6. The apparent melting temperature, T'_m, is plotted for each base pair. T'_m values in excess of 100°C are represented by 100°C. Short vertical lines indicate location of detected point mutations. (**a**) A-hot spot fragments: 310 bp DNA, 20 GC's and sequence 311–600 (—); 304 bp DNA, 20 GC's and sequence 321–604 (- - -). (**b**) B-hot spot fragments: 130 bp DNA, 20 GC's and sequence 581–690 (—); 140 bp DNA, 20 GC's and sequence 581–700 (- - -); (**c**) C/D-hot spot fragments: 410 bp DNA, 20 GC's and sequence 611–1000 (—), 440 bp DNA, 20 GC's and sequence 581–1000 (- - -).

enhances the ability to detect a base pair change at a given position. A base pair change in a small melting domain will have a larger effect on DNA melting than the same base pair change in a large melting domain. A second characteristic is used if the first melting domains of two DNA fragments are the same length. The DNA fragment with the largest δT is chosen.

p53 cDNA sequence

We examined the cDNA sequence of the human p53 tumor suppressor gene as an initial sample case. The cDNA sequence is 1317 bp long (16). A previous study showed by iterative calculations and viewing of derivative melting curves and melting profiles that three DNA fragments enable one to detect mutations in the four hot spot regions of the p53 cDNA (6). The four hot spot regions were designated A, B, C, and D. Their locations in terms of base pair position were A, 529–580 (codons 132–149); B, 654–669 (codons 174–179); C, 841–877 (codons 236–248); and D, 949–976 (codons 272–281) (17). The cDNA sequence region 841–976, C/D, was considered to be one hot spot region. Position numbers are based on the sequence from reference (16).

Table 1 shows an example of a user's application of the program with typical input parameter values. The DNA sequence file contained the 1317 bp p53 gene cDNA sequence. Acceptable parameter values and comments are in italics and between parentheses. Table 2 shows the output of this application of MELTSCAN. The hot spot region B described above was used in this example. Table 2 lists the optimal fragment for the specified hot spot region and the optimal fragment for each base pair or contiguous group of base pairs. Except for the first 105 bp, last 21 bp, and one 21 bp internal segment, the algorithm found fragments for mutation detection at all base pair positions. Sixty-six percent of the DNA fragments had δT values of 1.0°C or larger. The run time for the program employing a PC-compatible computer with a 80486 processor running at 66 MHz was 25 min.

The optimal fragments predicted for the three p53 hot spot regions A, B, and C/D were from base pair positions 311–600, 581–690 and 611–1000 respectively. A 20 bp GC clamp was on the left end in each case. These DNAs are essentially identical to the three DNA fragments previously shown to detect eight point mutations (6). The latter DNAs encompassed positions 321–604, 581–700, and 581–1000 with the 20 bp GC clamp also on the left ends. The agreement between the two sets of DNA fragments is not surprising since the criteria used to select DNA fragments in reference 6 is the same as that employed by MELTSCAN. The difference is that the previous method required the user to arbitrarily select DNA endpoints and iteratively calculate DNA fragment melting properties until a fragment with appropriate properties was found. Figure 3 shows the melting maps of the three DNA fragments predicted by MELTSCAN and the fragments employed in reference 6. The similarity of the melting maps indicates that one may move the endpoints of a predicted DNA fragment by up to ±10 bp without compromising the fragment's usefulness for detecting mutations. This can be helpful in optimizing primers to produce DNA fragments by the polymerase chain reaction (PCR).

p53 genomic sequence

Borrensen *et al.* (7) employed constant denaturant gel electrophoresis to detect p53 mutations in PCR-amplified fragments from human genomic DNA. Theoretical melting map calculations were

made to predict which DNA fragments would detect mutations in the four hot spot regions. The fragments employed were amplified using primers with 40 nucleotide long GC clamps attached to 20 nucleotide complementary sequences.

MELTSCAN was applied to the genomic p53 sequence to determine if shorter GC clamp primers would also allow mutation detection. The 1000 bp genomic sequence surrounding each hot spot was analyzed. For this calculation we assumed hot spot regions: A, codons 128–153; B, codons 161–185; C, codons 237–253; and D, codons 265–301 (7). Fragments were obtained for three of the four hot spot regions using 20 or 25 bp GC clamps. The fragment for hot spot A required a 40 bp GC clamp. The predicted fragments for the hot spot regions were: A, 13061–13140 with a 40 GC clamp on the left (120 bp); B, 13131–13245 with a 25 bp GC clamp on the left (140 bp); C, 13926–14110 with a 25 bp GC clamp on the left (210 bp); and D, 14021–14520 with a 20 GC clamp on the left (520 bp). Position numbers are based on the sequence presented in reference 18.

The DNA fragments predicted for hot spots B and C were tested by TGGE using genomic DNAs (19). Earlier work has shown that GC to AT mutations occur in codons 175 (hot spot B) and codon 248 (hot spot C) in the genomic DNA of cell line CEM (20). A different mutation in codon 248, CG to TA, was also characterized in cell line Namalwa (20). The 140 bp B-fragment and 210 bp C-fragment were amplified by PCR from genomic DNA isolated from the CEM and Namalwa cell lines, 11 other tumor cell samples, and peripheral blood cells with no p53 mutations (19). TGGE experiments verified the above mutations and found five mutations in hot spot regions B and C from the other tumor cells (19).

The utility of DNA melting theory for selecting DNA fragments for denaturing gel electrophoresis methods has been previously demonstrated (1,2,4,6,7,11). The new feature of MELTSCAN is its ability to automatically scan through a DNA sequence selecting an appropriate fragment for a hot spot region and for base pairs in the sequence. The results described above, although limited in scope, verify that MELTSCAN selects DNA fragments for detecting mutations. The program should simplify and broaden applications of denaturing gel methods. MELTSCAN is available at no cost to academic investigators. For further information readers may contact the corresponding author (internet e-mail address, roger.wartell@biology.gatech.edu.). MELTSCAN is proprietary to the Georgia Tech Research Corporation. MELTSCAN is a service mark of the Georgia Tech Research Corporation.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support provided by grant GM38045 from the National Institutes of Health. We also thank S.H.Hosseini for providing information on TGGE results of PCR-amplified fragments from genomic p53 DNA samples, and T. Maier for programming advise.

REFERENCES

1. Lerman,L.S., Silverstein,K. (1987) *Methods Enzymol.*, **155**, 482–501.
2. Abrams,E.S., Murdaugh,S.E., and Lerman,L.S., (1990) *Genomics*, **7**, 463–475.
3. Rosenbaum,V. and Riesner,D. (1987) *Biophys. Chem.*, **26**, 235–246.
4. Riesner,D., Henco,K., and Steger,G. (1991) In Chrambach,A., Dunn,M.J., and Radola,B.J. (eds), *Advances in Electrophoresis*. VCH Pub., New York, Vol. 4, pp.171–250.
5. Wartell,R.M., Hosseini,S.H., and Moran,J.D. (1990) *Nucleic Acids Res.*, **18**, 2699–2705.
6. Ke,S-H., Kelly,P.J., Wartell,R.M., Hunter,S.H., and Varma,V.A. (1993) *Electrophoresis*, **14**, 561–565.
7. Borresen, A-L., Hovig,E., Smith-Sorensen,B., Malkin,D., Lystrad,S., Andersen T.I., Nesland,J.M., Isselbacher, K.J., and Friend, S.H. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 8405–8409.
8. Yoshino,K., Nishigaki,K., and Husimi,Y. (1991) *Nucleic Acids Res.*, **19**, 3153.
9. Lerman,L.S., Fischer,S.G., Hurley,I., Silverstein,K., and Lumelsky,N. (1984) *Annu. Rev. Biophys. Bioeng.*, **13**, 399–423.
10. Myers,R.M., Lerman,L.S., and Maniatis,T. (1985) *Science*, **229**, 242–246.
11. Sheffield,V. C., Cox,D.R., Lerman,L.S., and Myers,R.M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 232–236.
12. Costes,B., Girodon,E., Ghanem,N., Chassignol,M., Thong,N.T., Dupret,D., and Goossens,M. (1993) *Hum. Mol. Genet.*, **2**, 393–397.
13. Wartell,R.M. and Benight,A.S. (1985) *Physics Rep.*, **126**, 67–107.
14. Delcourt,S.G. and Blake,R.D. (1991) *J. Biol. Chem.*, **266**, 15160–15169.
15. McCampbell,C.R., Wartell, R.M., and Plaskon, R.R. (1989) *Biopolymer*, **28**, 1745–1758.
16. Zakout-Houri,R. Bienz-Tadmor,B., Givol,D., and Oren,M. (1985) *EMBO J.*, **4**, 1251–1255.
17. Nigro,J.M., Baker,S.J., Preisinger,A.C., Jessup,J.M., Hostetter,R., Cleary,K., Bigner,S.H., Davidson,N., Baylin,S., Devilee,P., Glover,T., Collins,F.S., Weston,A., Modali,R., Harris,C.C., and Vogelstein,B. (1989) *Nature*, **342**, 705–708.
18. Chumakov,P.M., Almazov,V.P. And Jenkins,J.R. (unpublished). GenBank X54156.
19. Hosseini,S.H. (1994) Ph.D. thesis, Georgia Institute of Technology.
20. Cheng,J. and Haas,M. (1990) *Mol. Cell. Biol.*, **10**, 5502–5509.
21. Fixman,M. and Friere,J. (1977) *Biopolymers*, **16**, 2693–2704.