Short Technical Reports

Sequencing Homopolymer Tracts and Repetitive Elements

BioTechniques 20:862-868 (May 1996)

ABSTRACT

We investigated the use of Taq dye primer and Taq terminator sequencing chemistry to optimize the quality of sequence data obtained from templates containing homopolymer tracts and repetitive elements. In direct side-by-side comparisons using the Applied Biosystems Model 373A Fluorescent Sequencer, the Taq terminator sequencing chemistry gave much cleaner and more consistent results on long homopolymer tracts and dinucleotide repeats. We also investigated various thermal cycling conditions and determined that higher annealing temperatures and longer denaturation times improved the ability to sequence through these problem templates.

INTRODUCTION

The Nucleic Acid Core Facility (NACF) was established at the National Center for Human Genome Research (NCHGR) as a resource for routine sequencing of templates for the new intramural program. The NACF is using Applied Biosystems (ABI) Model 373A Automated DNA Sequencers, which are based on real-time detection of electrophoresed DNA bands by means of laser-induced fluorescence (1,4) to sequence a variety of double-stranded (ds) templates. Each DNA sample is labeled with four different fluorescent dyes, which is then run in one lane on the Model 373A Sequencer (2). The vast majority of these templates are derived from projects investigating microsatellite detection or exon trapping and thus contain various repetitive elements or homopolymer tracts. The difficulties associated with se-

quencing these templates on the Model 373A Fluorescent Sequencer are well documented (3) and can be very problematic. Furthermore, it is difficult to standardize the choice of a vector, methods for library construction, techniques for insert selection and template preparation in a core situation. Core sequencing facilities are now being established routinely (5), and many of these sites have encountered similar methodological problems as the NACF and have had the same difficulties in educating the local community on the use of the core and the technical limitations of the fluorescent sequencing technology. In this work, we describe the methodology used by the NACF to overcome many of the difficulties encountered in sequencing homopolymer tracts and repetitive elements, and we have presented examples of data from these problem templates on the World Wide Web (WWW) with the hope of

Short Technical Reports

helping to educate other users of similar services.

We initially used the ABI *Taq* dye primer sequencing chemistry with the standard ABI thermal cycling conditions for all reactions on ds templates, but observed problems when ds templates contained homopolymer tracts such as the poly T tracts found in the exon trapping vector system (pAMP-1). We subsequently investigated the use of the *Taq* terminator chemistry to alleviate noisy background and premature termination of the signal from these problem templates. Finally, we modified the thermal cycling conditions to accommodate very long repetitive elements, and now we routinely

use a higher annealing temperature and longer denaturation times for all fluorescent sequencing of ds templates done at the NACF.

MATERIALS AND METHODS

All of the fluorescent sequencing reagents were purchased from Perkin-Elmer/Applied Biosystems Division (PE/ABI, Foster City, CA, USA). Unlabeled primers were purchased from Genosys (The Woodlands, TX, USA). The thermal cycling protocols were conducted on a Model PTC-100-96™ Thermal Cycler (MJ Research, Watertown, MA, USA). Excess ddNTPs were removed from the *Taq* terminator reac-

tions using Centri-Sep™ columns purchased from Princeton Separations (Adelphia, NJ, USA).

DNA Preparation

Double-stranded plasmid templates having the host strains DH5α or XL1-Blue were prepared by picking a single colony from a streaked LB plate and incubated overnight in a shaker-incubator at 37°C in 5 mL of LB broth from Advanced Biotechnologies (Columbia, MD, USA) with appropriate antibiotic. The following day, DNA was prepared from these cultures by using either the Wizard™ Miniprep kit (Promega, Madison, WI, USA) or the QIAwell™-8 plasmid kit (Qiagen, Chatsworth, CA,

Figure 1. Plot of chromatographs. Plot of chromatograph of the dye primer data (A) and *Taq* terminator data (B) using the ABI thermal cycling conditions. These chromatographs show the sequence of a GT repeat 269 bases into the sequence. (C) Plot of chromatograph of *Taq* terminator data using the NACF's thermal cycling conditions.

Figure 2. Plot of chromatograph *Taq* terminator data using the NACF's thermal cycling conditions showing a 100-bp repeat sequence starting at position 150 into the sequence.

Short Technical Reports

USA). The DNA was accurately quantitated by checking the optical density (OD) value on a Model 650 DU®-Spectrophotometer (Beckman Instruments, Fullerton, CA, USA) using a capillary tube system to read the DNA solution directly without dilution. The control template (vector without insert) was incubated and prepared in parallel with the sample DNA as a quality control for the template DNA preparation.

***Taq* Dye Primer Sequencing Reactions**

In the fluorescently tagged primer A and C reactions, we used 200 ng plasmid DNA, 3.2 pmol labeled primer, 1 μ L dNTP mixture, 1 μ L 5 $\times$ cycle sequencing buffer (400 mM Tris-HCl, pH 9.0, 10 mM MgCl₂) and 1 μ L diluted *Taq* DNA polymerase (0.57 U/ μ L) in a final volume of 5 μ L. These amounts were doubled for the G and T reactions to accommodate the lower quantum yield from the TAMARA and ROX

fluorescent tags. The sequencing reactions were run with the recommended ABI thermal cycling conditions in a Model PTC-100-96 thermal cycler (initial denaturation at 95°C for 3 min, 95°C for 30 s, 55°C for 30 s, 70°C for 1 min and then back to the denaturation step for 14 more cycles). The reactions were then finished by 14 cycles of 95°C for 30 s and 70°C for 1 min. The reactions were then stored at 4°C until further processing. The sequencing reactions were then concentrated by adding 80 μ L of 95% ethanol with 1.5 μ L of 3 M sodium acetate (pH 5.3) in a clean tube, combining all four sequencing reactions into this mixture and precipitating the DNA fragments at -70°C for 1 h. The samples were then spun for 30 min at 15000 rpm (Eppendorf® Model 5145C Centrifuge; Brinkmann Instruments, Westbury, NY, USA), the supernatant was decanted, and the pellets were washed with 250 μ L of 70% ethanol. The samples were re-spun for

10 min, the supernatant was discarded and the pellets were dried in a Speed-Vac® (Savant Instruments, Holbrook, NY, USA) for 1–3 min. The samples were then resuspended in 4 μ L of DNA loading buffer (deionized formamide with 50 mM EDTA), heated at 90°C for 1 min and loaded onto the Model 373A Fluorescent Sequencer (PE/ABI).

***Taq* Terminator Reactions**

The *Taq* terminator reactions used 1 μ g of plasmid DNA, 5 pmol unlabeled primer and 9 μ L of *Taq* terminator mixture. We initially used the ABI standard thermal cycling conditions (96°C for 3 min, 96°C for 10 s, 50°C for 5 s, 60°C for 4 min and back to the denaturing step for 24 more cycles on the Model PTC-100-96 thermal cycler). We subsequently used a modified cycle that consisted of 93°C for 3 min, 94°C for 30 s, 55°C for 5 s, 60°C for 4 min and back to the 30-s denaturing step for 24 more cycles on the Model PTC-100-96.

The sequencing reactions were placed over the Centri-Sep columns to remove excess dye terminators (following the Centri-Sep column protocol), dried down on the SpeedVac and then denatured in 4 μ L of loading buffer as above. The reactions were loaded completely on the Model 373A Sequencer (PE/ABI) and run under standard electrophoretic conditions.

Quality Assessment

The quality of the data was determined by the number of ambiguous base calls (number of n's) between bases 50 and 350. Less than 5% ambiguous calls was considered a successful run. Furthermore, the quality of the chromatographic data coming after a homopolymer stretch or repeat was considered when optimizing the thermal sequencing conditions. The data were aligned using the "Sequencing Management Project" program (DNA Star Inc., Madison, WI, USA) to compare the analyzed data from the various conditions run on these samples. We did not manually edit any of the data. Large amounts of sequence data are generated by the Core Facility. These data were moved to a Sun SPARCstation[®] 10 (SUN Microsystems, Mountain View, CA, USA) for both analysis and archival purposes. Through the Genetic Data Environment (GDE) and custom programs, it was possible to take the ABI sequence data and calculate a success rate with a 5% error criteria. This information was then taken and published on the WWW (http://uranus.nchgr.nih.gov/NACF_html).

RESULTS

We sequenced a series of plasmid clones using both the ABI *Taq* dye primer and ABI *Taq* terminator chemistry and directly compared the chromatographic profiles as well as the analyzed sequence data. We specifically picked clones that contained problem regions such as homopolymer tracts and dinucleotide repeats. We enumerated the number of ambiguous base calls between bases 50 and 350 of the analyzed data and compared the clarity of chromatographic profile following the problem region in the clone. In all cases, we observed that the *Taq* terminator chem-

istry gave superior results (see below).

In Figures 1A, we show an example of a chromatographic profile from a region of a double-stranded clone sequenced using *Taq* dye primer chemistry that contains a (GT) repeat from position 269 to position 299 in the analyzed data. The chromatographic profile for the *Taq* dye primer reactions had low peak heights with much higher background noise, and the signal

dropped off after about 270 bp in the middle of the (GT) repeat. There were 30 ambiguous calls between 250 and 370 bp, and the data were noisy before the repeat as well. It may be that the template itself is difficult to denature due to the high GC content of the repeats, resulting in the mispriming of the primer and premature termination of the chain extension reaction.

In contrast, the terminator reactions

Short Technical Reports

produced a clear chromatographic profile with no ambiguous calls between positions 250 and 370 (Figure 1B). The peak heights were high throughout the first 350 bases with baseline peak resolution. In addition, the data analysis was clean through the (GT) repeat, and readable sequence was obtained out to 390 bases.

We also show the sequence from the same plasmid clone using the *Taq* terminator kit with the modified thermal cycling conditions. The analyzed results from the chromatographic profile (Figure 1C) had a 1.5-fold stronger average signal (G 248:161, A 161:93, T 97:72, C 83:75) than the corresponding *Taq* terminator chemistry. The peak height was much higher and distinct for the modified cycling conditions both through the repeat region and around the repeat regions. We saw baseline resolution of the peaks throughout the region and obtained a readable sequence out to 500 bases. Thus, the modified thermal cycling conditions greatly improves the overall sequence data through difficult homopolymer tracts and dinucleotide repeats.

Figure 2 shows the analyzed data from a clone that contains a (TCTG) repeat and a (TC) repeat in tandem sequenced with the *Taq* terminator chemistry and the modified thermal cycling protocol. We were able to sequence through two distinct repeats maintaining high signal strength, baseline peak resolution and no ambiguous base calls in the problem region from the first 50–250 bp. This sequencing run had 3% ambiguous calls over the readable sequence out to 360 bases.

DISCUSSION

It should be noted that about 20% of the templates submitted to the core contain homopolymer tracts and repetitive elements. The use of *Taq* terminator reactions is highly recommended over the use of the *Taq* dye primers for all sequencing in order to get through difficult templates, such as the homopolymer tracts. Furthermore, the use of the *Taq* terminator chemistry obviously facilitates the routine use of custom primers, which is crucial in the studies conducted at NCHGR as, in many instances, clones must be

“walked on” to obtain the sequence flanking dinucleotide repeats for micro-satellite analysis.

Our thermal cycling conditions gave better sequence data when compared to the thermal cycling conditions of ABI, though they were subtle on some clones. In our experience, our thermal cycling conditions have allowed us to get through some very long repeats and poly-T tails, as well as given us excellent data on routine clones without repetitive elements. We hypothesize that the higher annealing temperature used in our thermal cycling conditions allows for more primer specificity and less slippage in the repeat or homopolymer tract during the thermal cycling, which would theoretically assist in high-fidelity copying through the full repetitive element. In addition to using a higher annealing temperature, we also used a longer denaturation step at a lower temperature, which is less harsh on *Taq*, while allowing the higher annealing temperature to keep these problem templates denatured longer.

The methodology presented in this report should help other core facilities to overcome many of the systematic problems associated with sequencing ds templates obtained from many investigators, especially with respect to templates with inserts of human origin. Furthermore, it is hoped that the information presented will help educate the various local communities on the use of the core and the technical limitations of fluorescent sequencing technology. In light of the latter point, we have established a home page on the WWW (http://uranus.nchgr.nih.gov/NACF_html/NACF.Home.html) where we present many examples of problems routinely encountered in a core situation along with their resolution and a troubleshooting flow chart for core facilities.

REFERENCES

1. Ansorge, W., B. Sproat, J. Stegemann, C. Schwager and M. Zenke. 1987. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* 15:4593-4602.
2. Connell, C., S. Fung, C. Heiner, J. Bridgman, V. Chakarian, E. Heron, R. Jones, S. Menchen, W. Mordan, M. Raff, M. Recknor, L. Smith, J. Springer, S. Woo and M. Hunkapiller. 1987. Automated DNA sequence analysis. *BioTechniques* 5:342-348.
3. Perkin-Elmer/Applied Biosystems Division. 1991. Applied Biosystems User Manual 18, October. Foster City, CA.
4. Pon, R.T., G.A. Buck, R.L. Niece, M. Robertson, A.J. Smith and E. Spicer. 1994. A survey of nucleic acid services in core laboratories. *BioTechniques* 17:526-534.
5. Prober, J.M., G. Trainor, R. Dam, F.W. Hobbs, C.W. Robertson, R.J. Zagursky, A.J. Cocuzza, M.A. Jensen and K. Baumeister. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238:336-341.

Address correspondence to Patrick M. Gillevet, NIH/NCHGR, 49 Convent Drive MSC 4470 Rm. 3A76, Bethesda, MD 20892-4470, USA. Internet: gillevet@uranus.nchgr.nih.gov

Received 13 January 1995; accepted 20 November 1995.

Christiane M. Robbins, Elvin Hsu and Patrick M. Gillevet
*National Center for Human Genome Research
National Institutes of Health
Bethesda, MD, USA*