Anal. Chem. **2000,** *72,* 1045−1052

DNA Sequencing up to 1300 Bases in Two Hours by Capillary Electrophoresis with Mixed Replaceable Linear Polyacrylamide Solutions

Haihong Zhou,[†] Arthur W. Miller,[†] Zoran Sosic,[†] Brett Buchholz,[‡] Annelise E. Barron,[‡] Lev Kotler,[†] and Barry L. Karger*,[†]

Barnett Institute and Department of Chemistry, Northeastern University, Boston, Massachusetts 02115, and Department of Chemical Engineering, Northwestern University, Evanston, Illinois 60208

This paper presents results on ultralong read DNA sequencing with relatively short separation times using capillary electrophoresis with replaceable polymer matrixes. In previous work, the effectiveness of mixed replaceable solutions of linear polyacrylamide (LPA) was demonstrated, and 1000 bases were routinely obtained in less than 1 h. Substantially longer read lengths have now been achieved by a combination of improved formulation of LPA mixtures, optimization of temperature and electric field, adjustment of the sequencing reaction, and refinement of the base-caller. The average molar masses of LPA used as DNA separation matrixes were measured by gel permeation chromatography and multiangle laser light scattering. Newly formulated matrixes comprising 0.5% (w/w) 270 kDa and 2% (w/w) 10 or 17 MDa LPA raised the optimum column temperature from 60 to 70 °C, increasing the selectivity for large DNA fragments, while maintaining high selectivity for small fragments as well. This improved resolution was further enhanced by reducing the electric field strength from 200 to 125 V/cm. In addition, because sequencing accuracy beyond 1000 bases was diminished by the low signal from G-terminated fragments when the standard reaction protocol for a commercial dye primer kit was used, the amount of these fragments was doubled. Augmenting the base-calling expert system with rules specific for low peak resolution also had a significant effect, contributing slightly less than half of the total increase in read length. With full optimization, this read length reached up to 1300 bases (average 1250) with 98.5% accuracy in 2 h for a single-stranded M13 template.

The Human Genome Project is now in an accelerated phase to obtain a draft DNA sequence as early as the spring 2000[1,2] and to complete the project by 2003.[3] This acceleration has been considerably assisted by the recent introduction of automated instrumentation based on capillary array electrophoresis with replaceable polymer solutions. DNA sequencing remains the gold standard for accuracy; thus, improvements in capillary electrophoresis will result in large benefits. One such improvement would be to achieve consistently long reads of more than 1000 bases in a single run.[4] This would substantially reduce sequencing cost for de novo DNA sequencing and for other applications, such as screening for multiple mutations. To achieve these benefits without sacrificing the high throughput critical to present-day applications, long DNA sequences must also be generated in a reasonably short time, e.g., not exceeding 1−3 h.

Currently, commercial sequencing instruments based on cross-linked slab gel electrophoresis with limited automation typically produce ∼700 bases/sample[5] and only provide 1000 or more bases at the expense of run times of 10−12 h.[6] A faster sequencer using an ultrathin gel cast in a disposable cassette[7] has a version that can generate 800−1000 bases in 3−4 h, but is not currently suitable for high throughput. In contrast, capillary electrophoresis enables high throughput by permitting unattended operation (automatic polymer matrix replacement and sample injection) and by multiplexing 96 or more capillaries.

Capillary electrophoresis for DNA sequencing has undergone rapid development in the past few years in both discrete capillary[8−12] and micromachined device[13,14] formats. One area of advance has been the adoption of replaceable polymer solutions for the separation matrix[15−19] as well as polymer mixtures.[9,20−22] To date, linear polyacrylamide (LPA) has provided the longest read

* To whom reprint requests should be sent: (e-mail) bakarger@lynx.neu.edu.
[†] Northeastern University.
[‡] Northwestern University.

(1) Pennisi, E. *Science* **1999**, *283*, 1822−1823.
(2) Marshall, E. *Science* **1999**, *284*, 1439−1441.
(3) Mullikin, J. C.; McMurray, A. A. *Science* **1999**, *283*, 1867−1868.
(4) Green, P. *Science* **1998**, *279*, 1115−1116.
(5) PE Biosystems, 377 Analyzer, http://www2.perkin-elmer.com/ab/about/dna/377/377a1a.html.
(6) Roemer, S. C.; Brumbaugh, K. A.; Boveia, V.; Jensen, M.; Gardner, J. 9th Int. Genome Sequencing Anal. Conf., Department of Energy, Hilton Head, SC, 1997.
(7) Yager, T. D.; Baron, L.; Batra, R.; Bouevitch, A.; Chan, D.; Chan, K.; Darasch, S.; Gilchrist, R.; Izmailov, A.; Lacroix, J. M.; Marchelleta, K.; Renfrew, J.; Rushlow, D.; Steinbach, E.; Ton, C.; Waterhouse, P.; Zaleski, H.; Dunn, J. M.; Stevens, J. *Electrophoresis* **1999**, *20*, 1280−1300.
(8) Ruiz-Martinez, M. C.; Berka, J.; Belenkii, A.; Foret, F.; Miller, A. W.; Karger, B. L. *Anal. Chem.* **1993**, *65*, 2851−2858.
(9) Salas-Solano, O.; Carrilho, E.; Kotler, L.; Miller, A.; Goetzinger, W.; Sosic, Z.; Karger, B. L. *Anal. Chem.* **1998**, *70*, 3996−4003.
(10) Tan, H.; Yeung, E. S. *Anal. Chem.* **1997**, *69*, 664−674.
(11) Kheterpal, I.; Mathies, R. A. *Anal. Chem.* **1999**, *71*, 31A−37A.
(12) Dovichi, N. J. *Electrophoresis* **1997**, *18*, 2393−2399.
(13) Liu, S.; Shi, Y.; Ja, W. W.; Mathies, R. A. *Anal. Chem.* **1999**, *71*, 566−573.
(14) Schmalzing, D.; Adourian, A.; Koutny, L.; Ziaugra, L.; Matsudaira, P.; Ehrlich, D. *Anal. Chem.* **1998**, *70*, 2303−2310.

10.1021/ac991117c CCC: $19.00
Published on Web 02/04/2000

lengths.[9,23] There have also been numerous approaches to the design of sensitive detection systems. Using confocal scanning[24] and different modes of on-column[25−29] and off-column monitoring,[30,31] there has been a trend toward greater multiplexing of capillaries. The current commercial capillary array sequencers are based on these prototypes. In addition, improvements have been obtained by switching to new energy-transfer dyes with greater emission fluorescence than earlier dyes.[32,33]

Our work has centered on using capillary electrophoresis to obtain long reads in DNA sequencing with relatively short analysis times. It was shown that increasing the molecular mass of the LPA but decreasing its concentration extended read length, while maintaining a short analysis time.[23] This trend was recently confirmed by others using a different polymer matrix.[34] Taking an integrated approach that optimized several parameters concurrently, among which were field strength, column temperature, and base-calling software, we routinely achieved DNA sequencing of 1000 bases in less than 1 h with a mixed separation matrix.[9] We have found that desalting the sample and removing the sequencing template is effective for long reads and for high reproducibility of electrokinetic injection[35,36] and overall robustness in high-throughput DNA sequencing. We report here on further progress in long-read DNA sequencing.

SEPARATION PRINCIPLES FOR LONG DNA READS

Many factors influence read length in DNA sequencing, such as sample quality and base-caller, but the most important is electrophoretic resolution. We summarize here how various factors affect resolution for large DNA molecules. It will be seen that the migration behavior of these highly charged polyelectrolytes through polymer solution is very complex, and a full understanding of the processes involved is not yet available.

Briefly, the resolution, R_s, of two adjacent peaks with mobilities μ_1 and μ_2 (inversely proportional to migration time) can be expressed as[37]

$$R_s = \left(\frac{\mu_1 - \mu_2}{\bar{\mu}}\right)\frac{N^{1/2}}{4} \qquad (1)$$

where $\bar{\mu}$ is the mean mobility of the two species. The mobility term in eq 1 is selectivity, which is normalized relative distance between peaks, and N is the efficiency of separation given in theoretical plates and equal to $(\sigma_{tot}^2/l)^{-1}$, where σ_{tot}^2 is total variance of the peak and l is effective column length (injection to detection point). Letting $\alpha = \mu_1/\mu_2$ and considering only large fragment sizes where α is very close to one, eq 1 can be rewritten as[38]

$$R_s \cong {}^1/_4(\alpha - 1)N^{1/2} \qquad (2)$$

Because $(\alpha - 1)$ has a small absolute value for long DNA fragments, resolution of such fragments demands efficiencies of millions of theoretical plates, which fortunately can be obtained with capillary columns filled with replaceable polymer solutions. For example, in a typical run in the current work, at 60 °C, $\alpha = 1.000\,55$ and $N = 5.22 \times 10^6$ plates were observed for DNA fragments 955 and 956 bases long, providing a resolution of 0.31 (acceptable for base-calling).

Efficiency N is related to the total variance σ_{tot}^2, which results from several linearly independent variances:[37]

$$\sigma_{tot}^2 = \sigma_D^2 + \sigma_p^2 + \sigma_{\Delta T}^2 + \sigma_{eof}^2 + \sigma_{ec}^2 + \ldots \qquad (3)$$

where σ_D^2 is variance due to molecular diffusion, σ_p^2 is variance due to the matrix network dynamic dissociation and polymer−DNA interactions, $\sigma_{\Delta T}^2$ is variance due to the temperature profile across the column, σ_{eof}^2 is variance due to electroosmotic flow, and σ_{ec}^2 is variance due to extracolumn effects related to injection and detection. Other band-broadening factors exist but are in general less important.

In our work, σ_{eof}^2 and $\sigma_{\Delta T}^2$ are negligible because electroosmotic flow is suppressed by the capillary coating, and thermal gradients across the column are minimized both by effective thermostating and by limiting electric current with a low-conductivity buffer.[9,23] As a result, the electric current was linearly proportional to the electric field over the range of studied fields. The terms σ_D^2, σ_{ec}^2, and σ_p^2 could be significant, however,[39] but σ_{ec}^2 and σ_p^2 are subject to considerable experimental control (see below). In our experiments, a contribution of the detector window length to σ_{tot}^2 was less than 1% for all DNA sizes. The width of the injection plug can be minimized by optimizing injection conditions.[36] We experimentally injected as much sample as possible while avoiding sample overloading.

(15) Bashkin, J.; Marsh, M.; Barker, D.; Johnston, R. *Appl. Theor. Electrophor.* **1996**, *4*, 39−41.
(16) Gao, Q.; Yeung, E. S. *Anal. Chem.* **1998**, *70*, 1382−1388.
(17) Fung, E. N.; Yeung, E. S. *Anal. Chem.* **1995**, *67*, 1913−1919.
(18) Goetzinger, W.; Kotler, L.; Carrilho, E.; Ruiz-Martinez, M. C.; Salas-Solano, O.; Karger, B. L. *Electrophoresis* **1998**, *19*, 242−248.
(19) Madabhushi, R. S. *Electrophoresis* **1998**, *19*, 224−230.
(20) Barron, A. E.; Sunada, W. M.; Blanch, H. W. *Biotechnol. Bioeng.* **1996**, *52*, 259−270.
(21) Kim, Y.; Yeung, E. S. *J. Chromatogr., A* **1997**, *781*, 315−325.
(22) Wu, C.; Quesada, M. A.; Schneider, D. K.; Farinato, R.; Studier, F. W.; Chu, B. *Electrophoresis* **1996**, *17*, 1103−1109.
(23) Carrilho, E.; Ruiz-Martinez, M. C.; Berka, J.; Smirnov, I.; Goetzinger, W.; Miller, A. W.; Brady, D.; Karger, B. L. *Anal. Chem.* **1996**, *68*, 3305−3313.
(24) Huang, X. C.; Quesada, M. A.; Mathies, R. A. *Anal. Chem.* **1992**, *64*, 2149−2154.
(25) Yeung, E. S.; Li, Q. In *High-Performance Capillary Electrophoresis*; Khaledi, M. G., Ed.; John Wiley & Sons: New York, 1998; pp 767−789.
(26) Quesada, M. A.; Zhang, S. *Electrophoresis* **1996**, *17*, 1841−1851.
(27) Behr, S.; Matzig, M.; Levin, A.; Eickhoff, H.; Heller, C. *Electrophoresis* **1999**, *20*, 1492−1507.
(28) Anazawa, T.; Takahashi, S.; Kambara, H. *Electrophoresis* **1999**, *20*, 539−546.
(29) Sosic, Z.; Salas-Solano, O.; Yongwu, Y.; Zhou, H.; Miller, A. W.; Karger, B. L. in preparation.
(30) Kambara, H.; Takahashi, S. *Nature* **1993**, *361*, 565−566.
(31) Dovichi, N. J. In *Handbook of Capillary Electrophoresis*; Landers, J. P., Ed.; CRC Press: Boca Raton, FL, 1994; pp 369−387.
(32) Lee, L. G.; Spurgeon, S. L.; Heiner, C. R.; Benson, S. C.; Rosenblum, B. B.; Menchen, S. M.; Graham, R. J.; Constantinescu, A.; Upadhya, K. G.; Cassel, J. M. *Nucleic Acids Res.* **1997**, *25*, 2816−2822.
(33) Ju, J.; Ruan, C.; Fuller, C. W.; Glazer, A. N.; Mathies, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 4347−4351.
(34) Heller, C. *Electrophoresis* **1999**, *20*, 1978−1986.
(35) Ruiz-Martinez, M. C.; Salas-Solano, O.; Carrilho, E.; Kotler, L.; Karger, B. L. *Anal. Chem.* **1998**, *70*, 1516−1527.
(36) Salas-Solano, O.; Ruiz-Martinez, M. C.; Carrilho, E.; Kotler, L.; Karger, B. L. *Anal. Chem.* **1998**, *70*, 1528−1535.
(37) Giddings, J. C. *Sep. Sci.* **1969**, *4*, 181−189.
(38) Karger, B. L.; Snyder, L. R.; Horvath, C. *An Introduction to Separation Science*; J. Wiley & Sons: New York, 1973.
(39) Tinland, B. *Electrophoresis* **1996**, *17*, 1519−23.

Band broadening related to hydrophobic polymer−DNA interactions and to polymer network dynamics (σ_p^2) can be controlled through the use of very hydrophilic polymers such as polyacrylamide. Note that hydrophobicity of single-stranded DNA increases with base number. Network dynamics, reflected in the rate of dissociation of entanglements (constraint release[40]) and in the lifetime of the pores (i.e., blobs[41]), can be expressed through the network relaxation time t_r. For a given polymer in a "good" solvent, t_r scales as[41]

$$t_r \sim C^{1.5} M_w^{\ 3}/kT \tag{4}$$

where C is concentration, M_w is polymer molecular mass, k is Boltzmann's constant, and T is temperature. Equation 4 shows that t_r is inversely proportional to the column temperature, and thus blob lifetime diminishes at elevated temperatures. Among other factors, decrease in blob lifetime can lower efficiency by leading to enhanced molecular diffusion. In addition, selectivity can be affected, since the blob lifetime should be greater than the residence time of the DNA molecule segment in the blob[41,42] in order to maintain a uniform migrational behavior (i.e., mobility) for DNA fragment molecules throughout the polymer solution matrix. On the other hand, eq 4 also shows that network stability and, consequently, its resolving power can be improved by increasing the polymer molecular mass or concentration. Concentrated polymer solutions, however, are not optimal for separation of long DNA fragments.[9,23,34,43] Therefore, the use of solutions of high molecular mass polymer at relatively dilute concentration is important for obtaining long read lengths.

In addition to the relationship of blob lifetime to solute residence time, selectivity is obviously dependent on DNA migration processes. Depending on DNA size, the molecule can either be sieved through the polymer network (Ogston model[44]) or reptate in a virtual tube (biased reptation models[40,45−47]). The Ogston model satisfactorily described the separation of DNA fragments smaller than the pore size. For larger DNA, biased reptation with fluctuations (BRF) describes polymer networks in which DNA fluctuates in effective length during migration, affecting mobility at moderate and high electric field strength.[40,47]

The BRF model considers two regimes of DNA migration, (1) "reptation without orientation", where the mobility of a DNA fragment is inversely proportional to its size, and (2) "reptation with orientation", where DNA mobility is size independent:[40,47]

$$\frac{\mu}{\mu_0} \sim \begin{cases} N_k^{-1}, N_k < N_k^* \\ \epsilon_0, N_k > N_k^* \end{cases} \tag{5}$$

in which ϵ_0 is a dimensionless variable of a reduced field intensity defined by

$$\epsilon_0 = q_0 E b/kT \tag{6}$$

and N_k is the DNA length in Kuhn segments (smallest independent piece of a reptating molecule[48]) with segment length b, N_k* is the size threshold beyond which DNA molecules become fully oriented in the direction of the electric field, with all DNA migrating at the same rate ($\alpha = 1$), q_0 is the charge on DNA per Kuhn length, and E is the electric field.

The important observation with respect to experimental optimization of DNA separations is that resolution gradually approaches zero as N_k approaches N_k^*. Since N_k* should increase with decrease of ϵ_0,[40,47,49] the position of zero resolution may be shifted to a higher base number by increasing temperature and decreasing electric field. Because elevated temperature additionally helps to resolve compressions and increase separation speed, the approach we have taken is to find an optimal column temperature and only then adjust the electric field strength for the longest possible read.[9,23] As a final note, blob size also affects the value of N_k*,[49] and therefore, reducing polymer concentration will improve selectivity for large fragments.

Here we have briefly tried to convey interlinked phenomena that affect long DNA fragment separation in entangled polymer solutions. Increasing column temperature will increase selectivity but beyond a certain point will greatly diminish efficiency, resulting in poor separation. More concentrated polymer solutions will improve efficiency but only at the cost of reducing selectivity at high base numbers. Optimization entails simultaneous consideration of numerous parameters and tradeoffs. In addition, other factors must be taken into account, including dye and reaction chemistry, base-caller, template, and composition of the injected sample.

EXPERIMENTAL SECTION

Many details of the experimental work have been published previously, such as descriptions of reagents, sample preparation techniques, sequencing instrumentation, and polymer synthesis.[9,18] A complete description is available on the Web as Supporting Information.

Polymer Characterization. Weight-average molecular mass of LPA was determined by multiangle laser light scattering (MALLS) using the DAWN-Optilab system (Wyatt Technology, Santa Barbara, CA). Stock solutions of polymer samples were prepared at concentrations of $\sim 1 \times 10^{-4}$ g/mL in deionized water and passed through 0.02 μm filters. All samples were made using a high-precision balance to estimate final concentration accurately.

For analysis of a new LMM LPA, the polymer sample was fractionated by gel permeation chromatography (GPC) prior to on-line MALLS detection, using a 2690 Separations Module (Waters, Milford, MA) with Shodex (New York, NY) OHpak columns SB-806 HQ, SB-804 HQ, and SB-802.5 HQ connected in series. Sample aliquots of 100 μL were injected into the tandem GPC-LS system (mobile phase: 0.1 M NaCl, 50 mM NaH_2PO_4, and 200 ppm NaN_3; flow rate 0.2 mL/min). ASTRA software (Wyatt

(40) Duke, T.; Viovy, J. L. *Phys. Rev. E* **1994**, *49*, 2408−2416.
(41) Cottet, H.; Gareil, P.; Viovy, J. L. *Electrophoresis* **1998**, *19*, 2151−2162.
(42) Bae, Y. C.; Soane, D. *J. Chromatogr., A* **1993**, *652*, 17−22.
(43) Mitnik, L.; Salome, L.; Viovy, J. L.; Heller, C. *J. Chromatogr., A* **1995**, *710*, 309−321.
(44) Ogston, A. G. *Trans. Faraday Soc.* **1958**, *54*, 1754−1757.
(45) Lumpkin, O. J.; Dejardin, P.; Zimm, B. H. *Biopolymers* **1985**, *24*, 1573−1593.
(46) Slater, G. W.; Noolandi, J. *Biopolymers* **1989**, *28*, 1781−1791.
(47) Semenov, A. N.; Duke, T. A. J.; Viovy, J.-L. *Phys. Rev. E* **1995**, *51*, 1520−1537.
(48) Carreau, P.; De Kee, D. C. R.; Chhabra, R. P. *Rheology of Polymer Systems: Principles and Applications*; Hanser Publishers: New York, 1997.
(49) Heller, C. *Electrophoresis* **1999**, *20*, 1962−1977.

Technology, Santa Barbara, CA) was used to analyze the data (e.g., derive Zimm plots allowing calculation of the weight-average molecular mass and root-mean-square radius, i.e., radius of gyration, of the polymer[50]) and to generate plots of differential weight fraction versus molecular mass.

Samples of high molecular mass LPAs were injected from a syringe pump into the DAWN-Optilab system in microbatch mode directly without GPC. Multiple known concentrations of a given polymer sample were required for the analysis. For each polymer sample (i.e., at each concentration), the DAWN-Optilab instrument measured the intensity of the scattered light as a function of angle for 18 different fixed angles. After all data were collected, the known concentration was assigned to each plateau region, and a Zimm plot[50] was constructed. A subset of the 18 angles was chosen for data fitting to minimize the noise, with no fewer than 9 chosen for the final results. All analyses were repeated three times to ensure accuracy and reproducibility.

RESULTS AND DISCUSSION

The goal of this work was to extend the read length significantly beyond 1000 bases/run while maintaining high base-calling accuracy and rapid separation time. As in previous work, an integrated approach was followed that emphasized optimization of several parameters. Starting with LPAs with higher molecular masses than previously, a new mixed replaceable matrix with advanced thermostability was developed. Column temperature and electric field were optimized for maximum read length. The sequencing reaction was also adjusted to enhance the signal at high base number, and the base-caller was fine-tuned. Up to 30% improvement in read length was achieved.

Characterization of Linear Polyacrylamide. MALLS is one of the few absolute methods that can be employed to determine the weight-average molecular mass, M_w, and coil radius (radius of gyration) of HMM polymers in dilute solution.[51] MALLS was used to measure M_w for the LPAs synthesized for this study, as well as the HMM LPA from the previous work,[9] with tandem GPC−MALLS being utilized to provide size distributions for LMM LPAs. GPC was not employed for HMM LPAs because of lack of separation at 10 MDa and above and the absence of LPA size standards above 9 MDa.

Using the Berry formalism,[52] the M_w of the previously prepared HMM LPA[18] was found to be 10.4 ± 0.4 MDa and the root-mean-square radius of gyration 164 ± 4 nm. For the larger HMM LPA, the M_w was determined to be 17.1 ± 0.5 MDa and the root-mean-square radius of gyration was 189 ± 3 nm. LMM LPA was found to have a M_w of 268 ± 2 kDa (hereafter rounded to 270 kDa) with a polydispersity of 2.1 (ratio of M_w to number-average molecular mass). The root-mean-square radius of gyration was calculated to be 20.8 ± 0.1 nm.

Adjustment of Sequencing Reaction. Commercial sequencing reaction kits are formulated to provide adequate dye signal up to 700 bases,[53] but fluorescence levels continue to decline beyond this point, with disproportionate changes for some dyes. While it is always possible to intensify the signal by injecting more sample, longer injections can reduce the attainable resolution.[36] Since our focus was on long reads requiring the maximum possible resolution, an experiment was performed to test the effect of adjusting the standard BigDye primer reaction protocol to augment signal at high base number. This strategy has already been used by one reagent supplier to reoptimize its kits for reads of 1000 bases or more.[54]

Table 1. Progress in Software for Long-Read Lengths[a]

	limiting resolution	read length[b] at 98.5% accuracy
graph-theoretic base-caller[c]	0.40	928
first version of expert system[d]	0.25	1116
current version of expert system	0.25	1249

[a] Results of different base-caller versions on a set of five M13 electropherograms run at 125 V/cm and 70 °C for ∼2 h. [b] Read length was defined as the longest stretch of called sequence having this overall accuracy. [c] From ref 23. [d] From ref 9.

Analyzing the fluorescence intensity profile at 700 bases and beyond that resulted from preparing the reaction according to the manufacturer's instructions, we found that the signal from the G-terminated reaction was particularly low at 1000 bases and beyond (data not shown). It was by far the lowest of the four bases, and caused base-calling errors. The addition of an extra G-terminated reaction (mixing five termination reactions to give the sample) caused the G-terminated signal to rise to a level comparable to that from the T-terminated reaction. Read length increased by 50−100 bases. We are currently investigating reformulation of the sequencing reaction to further boost signal at a high base number. We are also exploring modification of injection procedures and apparatus to permit more long fragments to enter the column without loss in resolution.[13,14,55−58]

Software. Table 1 shows the read lengths obtained by different versions of this laboratory's software for long-read sequencing on a set of M13 runs performed under the new conditions of 70 °C and 125 V/cm. The effect of reducing the limiting resolution of the base-caller from 0.40 to 0.25 is reflected in the ∼200 bases added between the graph-theoretic base-caller employed in 1996 and the first version of the expert system developed in 1998. Resolution is not the only parameter affecting base-caller accuracy, however, and while the original expert system was accurate to a resolution of 0.25 on data obtained at 200 V/cm, it was not as good on the new 125 V/cm data summarized in Table 1. Lowering field strength from 200 to 125 V/cm commonly resulted in a 2−3-fold smaller signal-to-noise ratio at fixed resolution. To improve accuracy at 125 V/cm, the number of times that the rules were cycled through to refine base-call assignments was increased, and rules relating to noise level were added. At 125 V/cm, the improvement in read length, compared to the previous version of the base-caller,[9] was ∼130 bases at 98.5% accuracy. Changes to

(50) Zimm, B. H. *J. Chem. Phys.* **1945**, *13*, 141−145.
(51) Wyatt, P. J. *Anal. Chim. Acta* **1993**, *272*, 1−40.
(52) Berry, G. C. *J. Chem. Phys.* **1966**, *44*, 4550−4564.
(53) PEBiosystems *ABI PRISM BigDye Primer Cycle Sequencing Ready Reaction Kit Protocol*; 1997; p 3.
(54) Epicentre Technologies *SequiTherm EXCEL II Long Read DNA Sequencing Kits -LC.* Product Information, 1998, p 1.
(55) Chien, R.-L.; Burgi, D. S. *Anal. Chem.* **1992**, *64*, 489A−496A.
(56) Shultz-Lockyear, L. L.; Colyer, C. L.; Fan, Z. H.; Roy, K. I.; Harrison, D. J. *Electrophoresis* **1999**, *20*, 529−38.
(57) Butler, J. M.; McCord, B. R.; Jung, J. M.; Wilson, M. R.; Budowle, B.; Allen, R. O. *J. Chromatogr., B* **1994**, *658*, 271−280.
(58) Ulfelder, K. J.; Schwartz, H. E.; Hall, J. M.; Sunzeri, F. J. *Anal. Biochem.* **1992**, *200*, 260−267.

the base-caller were always validated on human genomic templates to ensure that they improved performance on data that would be observed in production sequencing. Details of the software are presented elsewhere.[59]

Thermostabilization of Entanglements in the Separation Matrix by LMM LPA. Elevated column temperature will be beneficial for long reads (higher resolution, fewer compressions, faster separation), and this has been verified in our laboratory.[9,23,60] Faster separation speed and a reduction in compressions at elevated temperature also have been noted by others.[61] However, each matrix must have some optimum temperature above which the combined effects of greater thermal diffusion and shortened relaxation time of the polymer network diminish the separation of large DNA fragments. Polymer matrixes with more stabilized entanglements provide optimum separation at higher temperature than others, and in this paper we characterize them as being more thermally stable.

In recent work, we showed that the addition of 0.5% (w/w) 50 kDa LPA substantially increased the thermostability of the network in a 2% (w/w) 10 MDa LPA matrix without significant increasing viscosity and thereby migration time.[9] For this mixed matrix, an optimum temperature of 60 °C for a 1-h separation was found, 10 °C higher than for a matrix containing 2% (w/w) HMM LPA alone.[23] To test for a dependence on molecular size of LMM LPA, 50 kDa LPA was substituted by 270 kDa LPA. For DNA fragments longer than ~500 bases, selectivity increased significantly above 65 °C (data not shown). Efficiency was greatest at the lowest temperature (60 °C), as expected,[37] but did not decrease drastically until 75 °C. The greatest overall resolution occurred at 70 °C, where the added selectivity more than compensated for the loss in efficiency. Raising the temperature to 70 °C reduced migration time as well.

The greater stability achieved by increasing the molecular mass of the LMM component from 50 to 270 kDa is surprising given the fact that there should be no change in mesh size.[22,62] To examine the stabilizing effect in more detail, selectivity and efficiency were plotted as a function of DNA size using different mixed matrixes and run at different temperatures (Figure 1). Substituting 270 kDa LPA for 50 kDa LPA at 60 °C yielded an unexpected increase in selectivity at the beginning of the electropherogram, though the difference in selectivity relative to the 50 kDa LPA diminished as the DNA size increased (Figure 1A). By comparison, using 270 kDa LPA at 70 °C, the greater selectivity was maintained throughout the run. Efficiency was identical for both matrixes in the 60 °C runs (Figure 1B), but decreased slightly when the temperature was raised to 70 °C. The enhanced selectivity at large fragment sizes permitted greater read lengths, confirming that thermostabilization was a main feature of LMM LPA. Preliminary studies with 500 kDa LPA at 0.25 (w/w) and 0.5% (w/w) do not appear to show any significant improvement in read length beyond that provided by 270 kDa LPA. These empirical results indicate that additional research is needed to explain the phenomena of polymer mixtures.

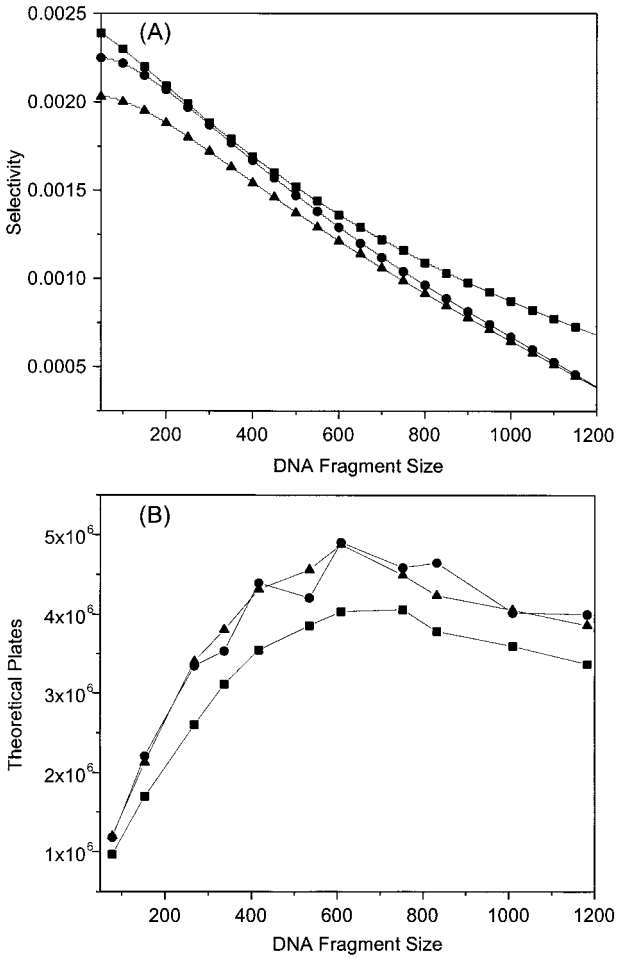

Figure 1. Effect of replacing 50 kDa LPA with 270 kDa LPA in a mixed separation matrix on (A) separation selectivity ($\Delta\mu/\bar{\mu}$) and (B) efficiency at different temperatures. Matrixes and temperatures: 2% (w/w) 10 MDa/0.5% (w/w) 50 kDa, 60 °C (▲); 2% (w/w) 10 MDa/0.5% (w/w) 270 kDa, 60 °C (●); 2% (w/w) 10 MDa/0.5% (w/w) 270 kDa, 70 °C (■). Samples were prepared using Universal BigDye-labeled (−21) primer cycle sequencing with AmpliTaq-FS on ssM13mp18 template. An additional G-terminated reaction was added as discussed in the text. Electrophoretic conditions: 75-μm-i.d., 365-μm-o.d., poly(vinyl alcohol)-coated capillary with effective length 30 cm, total length 45 cm; both running buffers were 50 mM Tris/50 mM TAPS/2 mM EDTA. The cathode running buffer and separation matrix also contained 7 M urea. Samples were injected at a constant electric field of 9 V/cm ($I = 0.7$ μA) for 10 s and electrophoresed at 125 V/cm.

Electric Field Optimization. The read length obtained at 200 V/cm and 70 °C with the mixed matrix containing 2% (w/w) 10 MDa and 0.5% 270 kDa LPA was ~1000 bases, the same as for the optimum conditions determined by the previous work.[9] However, it was expected that the read length could be improved up to a point by decreasing the electric field and thereby reducing the field-parallel orientation of the long DNA (eq 6). The results of field optimization at constant temperature 70 °C are shown in Table 2. This table shows that lowering the field from 250 to 100 V/cm indeed increased the DNA size at which a resolution of 0.3 (an arbitrarily chosen low value, selected to focus on the separation of long DNA fragments) was achieved, from 871 to 1236 bases. The greatest benefit occurred below 150 V/cm; the increment of 108 bases between 150 and 125 V/cm was 60% higher than the 65-base difference between 200 and 150 V/cm, despite the smaller field change.

(59) Miller, A. W.; Karger, B. L. Int. Pat. Appl. WO 99/53423, 1999.
(60) Kleparnik, K.; Foret, F.; Berka, J.; Goetzinger, W.; Miller, A. W.; Karger, B. L. *Electrophoresis* **1996**, *17*, 1860−1866.
(61) Zhang, J.; Fang, Y.; Hou, J. Y.; Ren, H.; Jiang, R.; Roos, K. P.; Dovichi, N. J. *Anal. Chem.* **1995**, *67*, 4589−4593.
(62) deGennes, P. G. *Scaling Concepts in Polymer Physics*; Cornell University Press: Ithaca, NY, 1979.

Table 2. Effect of Electric Field Strength on DNA Sequencing Results with LPA 2% (w/w) 10 MDa/0.5% (w/w) 270 KDa at 70 °C[a]

electric field (V/cm)	migration time for base 1019 (min)	fragment size at resolution 0.3	read length[b] at 98.5% accuracy
250	44.0	871	927
200	55.6	995	1042
150	80.5	1060	1127
125	101.0	1168	1190
100	131.0	1236	1172

[a] Samples and other electrophoretic conditions were as in Figure 1. Values in the table are averages of three to five experiments. [b] Read length was defined as in footnote *b* in Table 1.

Lowering the field from 150 to 125 V/cm experimentally increased read length less than it did the fragment size at resolution 0.3, and further field decrease to 100 V/cm even reduced read length slightly, despite the higher selectivity (data not shown) and higher resolution in the region of high base numbers. The major parameter counteracting higher resolution appeared to be the decline in signal intensity with decreasing field. As expected, the time required for separation, shown in Table 2 by the migration time for a fragment of 1019 bases, increased as the field was lowered.

Effect of Mixture Composition on Sequencing. Table 3 summarizes results from different LPA mixtures under comparable conditions of polymer concentration, temperature, and field. Increasing the molecular mass of either the LMM or HMM component increased both resolution and read length. Read length was strongly correlated with the fragment size at which a resolution of 0.3 was achieved, indicating that improved resolution was the source of greater read length under these conditions of temperature and field. It should be emphasized that increasing molecular mass of either LPA component had negligible effect on migration time.

The longest read length, an average ~1250 bases with 98.5% accuracy, was found by combining the 270 kDa LPA with the largest HMM LPA. Figure 2 shows an electropherogram generated with this matrix that had a read length of 1300 bases with 98.5% accuracy. A few miscalls at the beginning were followed by over 800 correctly called bases, and errors did not begin to accumulate until base 1240. The cause of base-calling inaccuracy beyond this point was low resolution, whereas at field strengths less than 125 V/cm, problems occurred earlier in the sequence due to low signal. For the runs summarized in Tables 2 and 3, read lengths for 99% accuracy were approximately 30−60 bases less than given for 98.5% accuracy.

Under optimum conditions, the combined effects of mixing 0.5% (w/w) 270 kDa and 2% (w/w) 17 MDa LPA produced a dramatic increase in read length over 2% (w/w) 10 MDa LPA alone. The 17 MDa polymer may create a more rigid network with superior strength of entanglements at high temperature, since the relaxation time of the network depends strongly on polymer molecular mass (eq 4). However, the additional stabilization of entanglements in the LPA network by the 270 kDa LPA was also clearly important.

While M13 is a model template, it does indicate the potential results with production templates. Samples prepared from cloned fragments of human genomic DNA, primarily from chromosome 17, were sequenced using the mixed matrix of 2% (w/w) 17 MDa and 0.5% (w/w) 270 kDa LPA under run conditions optimized using M13 template. Read lengths up to 1220 bases at 98.5% accuracy were obtained. The average read length of ~1100 bases appeared to be less than that observed for M13 primarily because of low signal resulting from the fact that these templates were prepared for commercial sequencers optimized for shorter sequences. Improving the signal could produce read lengths approaching those achieved with M13.

CONCLUSIONS

Many DNA sequencing applications would benefit from routinely longer read lengths. In this paper, significant progress has been described, based on the development of new separation matrixes and on the optimization of run conditions for capillary electrophoresis. Starting from fundamental considerations, a matrix containing 2% (w/w) 17 MDa LPA and 0.5% (w/w) 270 kDa LPA was employed in concert with temperature and electric field adjustments, with updated base-calling software, and with a modified formulation for the dye primer sequencing reaction. The combined optimizations resulted in read lengths of up to 1300 bases (average ~1250) for single-stranded M13 template and up to 1220 bases (average ~1100) for human genomic DNA. Implementing these features in a production setting would considerably increase throughput and reduce the costs of large-scale sequencing efforts.

Further improvements in read length should be possible by pursuing the strategies outlined here, an important aspect of which is to balance changes in matrix composition with appropriate changes in column temperature and electric field. Another factor in this balance is the level of fluorescent signal at high base

Table 3. Effect of Polymer Composition on DNA Sequencing Results[a]

LPA	temp (°C)	electric field (V/cm)	migration time for base 1019 (min)	fragment size at resolution 0.3	read length[b] at 98.5% accuracy
2% 17 MDa + 0.5% 270 kDa	70	125	105.4	1215	1249
2% 10 MDa + 0.5% 270 kDa	70	125	101.0	1168	1190
2% 17 MDa + 0.5% 50 kDa	70	125	100.0	1060	1083
2% 10 MDa + 0.5% 50 kDa	70	125	99.5	923	965
2% 10 MDa + 0.5% 50 kDa[c]	60	200	55.6	927	1013
2% 10 MDa[d]	50	150	81.0	925	951

[a] Samples and other electrophoretic conditions were as in Figure 1. Values are averages of three to five experiments. Temperature and electric field were optimized for each polymer matrix. [b] Read length was defined as in footnote *b* in Table 1. [c] Conditions same as optimum conditions determined in ref 9. [d] Conditions same as in ref 18.

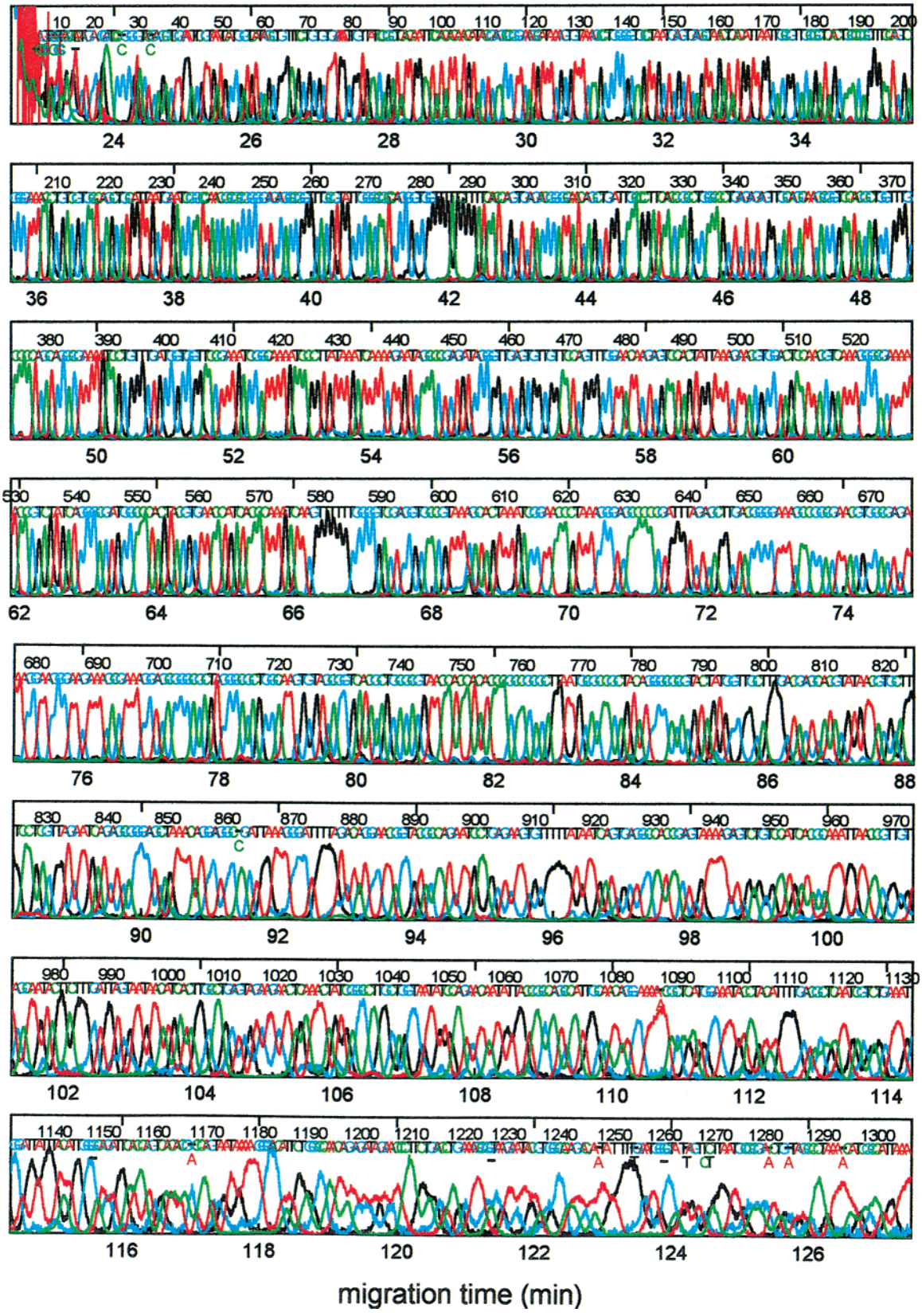

Figure 2. Read length of 1300 bases using the separation matrix LPA 2.0% (w/w) 17 MDa/0.5% (w/w) 270 kDa at 125 V/cm and 70 °C. Sample was prepared as in Figure 1.

number. Future benefits should accrue from reoptimizing the sequencing reaction to amplify signal at high base numbers and from enhancing base-caller accuracy when signal is low, and these benefits would also be applicable to microchips. Other possibilities to improve signal may include modifying the injection procedure or apparatus. We are currently pursuing these avenues.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of this work by DOE under Human Genome Project Grant DE-FG02-90ER-60985 (B.L.K.), by NIH under Grant 1R01HG01970-01 (A.E.B.), and by the Arnold and Mabel Beckman Foundation (A.E.B.). Support by DOE does not constitute an endorsement of views expressed in this paper. The authors thank the Whitehead Institute/MIT Center for Genome Research for providing cloned sequencing templates and helpful discussions. This is Contribution No. 772 from the Barnett Institute.

SUPPORTING INFORMATION AVAILABLE

A detailed description of materials, sequencing protocols, instrumentation, and experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review September 29, 1999. Accepted December 11, 1999.

AC991117C