Analysis of near-infrared dye-labeled Sanger sequencing fragments with gel electrophoresis using the time-resolved fluorescence lifetime identification methods

Suzanne Jeanel Lassiter
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations
Part of the Chemistry Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/2957
ANALYSIS OF NEAR-INFRARED DYE-LABELED SANGER SEQUENCING FRAGMENTS WITH GEL ELECTROPHORESIS USING TIME-RESOLVED FLUORESCENCE LIFETIME IDENTIFICATION METHODS

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Department of Chemistry

by

Suzanne Jeanel Lassiter
B.S., Georgia Southwestern State University, 1996
December, 2001
Acknowledgements

Mommy and Dad: Thanks for all the restful weekends on the beach and the patience that you showed through all of the trying times. Your love and support helped get me where I am today.

Jim: Without you I would not still have my sanity. I can’t tell you how much the IMs and e-mails got me through each day. I love you and can’t wait to enjoy the rest of our lives!

Michelle: The best roommate a sister can have. Glad we got to make up for all the times we couldn’t get along when we were younger.

Dr. Soper: Thank you for the encouragement and knowledge you bestowed upon me. I may have given up long ago if you had not been my advisor.

Alyssa: Thanks for the funny stories and useful conversations. You can’t imagine how our bitch sessions helped me relieve the frustrations.

Wieslaw: Without you I would still be going through my data point-by-point. Thank-you for all of the intellectual conversations and helping me solve problems.

The “Super” Soper Group: I will remember you all. From the beginning to the end there was always one of you there to help lift my spirits when things got rough, especially Christie, Shawn, Shelby, Musundi, Michelle, Angela and also the adopted members and post-docs, Michele and Seok-Kwan.
Table of Contents

Acknowledgements ........................................................................................................... ii

Table of Contents ........................................................................................................ iii

List of Tables .................................................................................................................. vi

List of Figures ................................................................................................................ vii

Abstract ........................................................................................................................... ix

Chapter 1. DNA Sequencing Using Fluorescence Detection Methods ......................... 1
  1.1. Genome Structure .............................................................................................. 1
  1.2. Methods for Determining the Primary Structure of DNA .............................. 3
    1.2.1. Maxam-Gilbert Sequencing ........................................................................... 6
    1.2.2. Sanger Chain Termination Method ............................................................... 9
  1.3 Modes of Electrophoresis .................................................................................... 12
    1.3.1. Slab Gel Electrophoresis ........................................................................... 15
    1.3.2. Capillary Electrophoresis ........................................................................ 18
  1.4. DNA Migration ................................................................................................. 23
    1.4.1. Ogston Model .............................................................................................. 23
    1.4.2. Reptation Model ......................................................................................... 24
    1.4.3. Biased Reptation Model ............................................................................ 24
  1.5. Detection Methods for DNA Sequencing ....................................................... 25
    1.5.1. Audioradiographic Methods ....................................................................... 25
    1.5.2. Fluorescence Detection ............................................................................ 27
      1.5.2.1. Fluorescent Dyes for DNA Labeling and Sequencing .................... 33
      1.5.2.2. Visible Fluorescent Dyes for DNA Labeling .................................... 35
      1.5.2.3. ET Dyes for DNA Sequencing ......................................................... 39
  1.6. Instrumental Formats for Fluorescence Detection in DNA Sequencing .............. 43
    1.6.1. Fluorescence Scanning Instruments ............................................................ 46
    1.6.2. Fluorescence Imaging Systems for DNA Sequencing ............................. 48
    1.6.3. Fluorescence Detection Formats ............................................................... 50
  1.7. Sequencing Strategies ....................................................................................... 54
    1.7.1. Single-color/four-lane ............................................................................ 54
    1.7.2. Single-color/single-lane ........................................................................... 56
    1.7.3. Two-color/single-lane ............................................................................ 57
    1.7.4. Four-color/single-lane ........................................................................... 60
  1.8. References .............................................................................................................. 62
Chapter 5. DNA Sequencing Using Time-Resolved Identification of Terminal Bases with Near-IR Fluorescence Detection

5.1 Introduction ......................................................................................... 164
5.2 Experimental ..................................................................................... 168
  5.2.1 Instrumentation ........................................................................... 168
  5.2.2 DNA Sequencing ........................................................................ 170
  5.2.3 Data Acquisition ....................................................................... 173
5.3 Data Analysis ...................................................................................... 174
5.4 Base Identification in Sequencing Trace ............................................. 175
5.5 Results and Discussion ..................................................................... 179
5.6 Conclusions ....................................................................................... 183
5.7 References ........................................................................................... 184

Chapter 6. Conclusions and Future Work ............................................ 186

6.1 Document Summary ............................................................................ 186
6.2 Future Work ....................................................................................... 187
  6.2.1 New Dyes .................................................................................... 187
  6.2.2 Microchip Analysis ..................................................................... 189
  6.2.3 Two-Color Detection .................................................................. 191
6.3 References ........................................................................................... 193

Vita .............................................................................................................. 194
List of Tables

1.1. Binary coding scheme for two-color DNA sequencing........................................59

3.1. Read accuracies and errors as a function of read length using a single-
dye/four-lane and two-lane/two-dye sequencing strategies .................................124

4.1. Calculated lifetime values of near-IR tricarbocyanine dyes in varying
concentrations of linear polyacrylamides with different concentrations of
denaturants, urea and/or formamide. (% linear polyacrylamide (w/v); M,
urea; %, formamide) .......................................................................................151

4.2. Calculated lifetime values of near-IR heavy-atom modified tricarbocyanine
dyes in various linear polyacrylamides and POP6. Flanagan acquired lifetime
measurements in a non-polymerized 6%T5%C acrylamide solution with
40%formamide [1, 10]. (% acrylamide; M urea; %, formamide) ......................155
List of Figures

1.1. Chemical structures of nucleotide base units: adenine, guanine, cytosine, and thymine; and dideoxy-phosphate unit (ddNTP or terminator) and deoxy-phosphate unit (dNTP).................................................................................................................. 2

1.2. Image of double helix and individual components of DNA molecule. ............ 4

1.3. Flow chart for ordered shotgun sequencing..................................................... 5

1.4. Cleavage of guanine in the Maxam-Gilbert sequencing method....................... 7

1.5. Steps involved in Maxam-Gilbert sequencing strategy...................................... 8

1.6. Cycle sequencing used to linearly amplify DNA............................................... 10

1.7. Steps in the Sanger sequencing method. ......................................................... 11

1.8. A) Free radical polymerization scheme of acrylamide. B) Cross-linking polymerization reaction of acrylamide and bis-acrylamide. Ammonium persulfate becomes a free radical when dissolved in water and when added to acrylamide solution in the presence of the catalyst tetra-methylenediamine, TEMED, the reaction proceeds........................................................................................................ 14

1.9. Dependence of pore size on acrylamide concentration ...................................... 15

1.10. Cartoon of sample wells formed in the separation matrix, during polymerization, by a plastic or paper comb displacing the matrix .......... 16

1.11. Simplified block diagrams of (A) a slab gel electrophoresis system and (B) a capillary gel electrophoresis system for DNA analysis. ......................... 17

1.12. EOF profile of fused silica at increasing pH values. ....................................... 19

1.13. Schematic of electroosmotic flow (EOF) produced in a fused silica capillary when an electric field is applied. Positively charged molecules migrate ahead of neutral molecules, which travel with the EOF, and negatively charged molecules migrate more slowly than the neutral molecules ......................... 19

1.14. Electrokinetic injection of DNAs into capillary gel column ......................... 22

1.15. Representation of the different mechanisms of DNA migration in an array of fixed obstacles .................................................................................... 25
1.16. Film of autoradiographic sequencing of an M13mp19 template with $^{32}$P labeling of deoxynucleotides .................................................................28

1.17. Structure of fluorescent dyes used for labeling primers for four-color DNA sequencing. Fluorescein and NBD are excited with the 488 nm lines of an Argon laser and Tetramethylrhodamine and Texas Red from the 514.5 nm lines..................................................................................................................30

1.18. Emission spectra of dyes used for labeling primers; F, fluorescein; N, NBD; T, tetramethylrhodamine; TR, Texas Red. .............................................................31

1.19. Experimental set-up used to excite and collect emission from a four-dye one-lane sequencing run. ..................................................................................32

1.20. Structure of common fluorescence dye set used for labeling primers in DNA sequencing applications. The functional group on each dye is a succinimidyl ester, which readily conjugates to primary amine groups.........................................36

1.21. Emission spectra and filter set used for detection of FAM, JOE, TAMRA, and ROX dyes ...............................................................................................37

1.22. Chemical structures of energy transfer (ET) labeled DNA sequencing primers. ...........................................................................................................41

1.23. Absorption (dark line) and emission spectra for both the ET primers and the single dye-labeled primers are shown for comparison. The number in parenthesis is the excitation wavelength used for collecting the emission profile. .............................................................................................................42

1.24. Structure of Big Dye terminator with energy transfer dyes
   Tetramethylrhodamine and Fluorescein attached to ddTTP..........................44

The fluorescence is sampled unidirectionally at 1500 Hz/channel with a resolution of 13.3mm/pixel ...................................................................................47

1.26. Schematic of the multiple sheath flow cell using gravity feed for the sheath flow. Twenty gel filled capillaries were aligned at a 0.35 mm pitch in an optical cell. ..................................................................................................................49

1.27. Schematic an imaging laser-induced fluorescence detector for reading four-color fluorescence from capillary gel arrays............................................50
1.28. Graphs of equation 1.8 for values n=1, 2, 3, 4. For increasing number of
detection channels an increase in the signal-to-noise ratio and intensity is
needed for the identification of a fluorescence signal ..............................52

1.29. Sequence of one-dye/four lane format of a M13mp18 template with a M13(-
29) primer. Each terminal base represented in a different color for sequence
reconstruction. Automated sequencer run at-1600 V with 1 µl of sample per
lane ..................................................................................................................55

2.1. Jablonski diagram. Block energy diagram showing the various radiative and
non-radiative processes that occur to a molecule upon the absorption of a
photon of light ..................................................................................................67

2.2. Block diagram of time correlated single photon counting instrument: CFD,
constant fraction discriminator; TAC, time-to-amplitude converter; MCA,
multichannel analyzer .......................................................................................70

2.3. Beam geometry in the Mira tunable-laser cavity as the light passes through
the Ti:Sapphire crystal ......................................................................................76

2.4. Part of Mira’s saturable absorber system: slit and beam cross section.........77

2.5. Schematic of silicon avalanche photodiode .................................................83

2.6. A) Electron energy level diagram illustrating the generation and subsequent
separation of an electron-hole pair by photon absorption within the depletion
region of a p-n junction. B) Illustration of the principle of operation of an
avalanche photodiode .......................................................................................84

2.7. Timing errors associated with leading edge discriminators. Both pulses are
initiated at the same time, however their arrival time is different due to the
threshold level and amplitude of the pulse .....................................................86

2.8. Schematic of pulse attenuation from a constant fraction discriminator .........87

2.9. Graph of monotonic non-decreasing function of two variables (m,τ) for
determining lifetime values of experimental data using the MLE ................90

2.10. Structure of near-IR heavy-atom modified tricarbocyanine dyes ..............95

2.11. Absorbance and emission profiles of heavy-atom modified tricarbocyanine
dye-primers measured in a 6%T5%C non-polymerized acrylamide solution
containing 1 X TBE and 40% formamide ......................................................96

3.1. Schematic of the time-resolved near-IR scanner .....................................104
3.2. Picture of near-IR time-resolved scanner mounted in the Li-COR 4000 automated DNA sequencer ................................................................. 107

3.3. Experimental set-up of near-IR time-resolved lifetime data acquisition system coupled to Li-COR automated DNA sequencer ........................................... 108

3.4. Front panel displays of the data analysis package for analyzing time-resolved data from sequencing gels ........................................................................... 109

3.5. Structure and spectroscopic properties of near-IR labeling dyes .................. 112

3.6. (A) Fluorescence decay profiles for IRD700- and Cy 5.5- labeled fragments. (B) Histograms showing the fluorescence lifetimes calculated for single dye tracts of G-labeled sequencing ladders electrophoresed in a denaturing slab gel (7 M urea) ........................................................................................................ 117

3.7. Monte Carlo simulations of two dyes with lifetimes of 750 and 850 ps. (A) Simulation with R = 1.0 and (B) R = 0.2 ......................................................... 121

3.8. Slab gel electropherogram of IRD700-labeled C-fragments and Cy 5.5-labeled T-fragments. The intensity electropherogram is shown in (A) along with the base assignments obtained from the calculated lifetime, using equation 3.1. The decay profiles were constructed from a single pixel at the top of peak, which contained the highest fluorescence intensity. (B) Expanded view of the band doublet shown in (A) with a resolution of 0.36 between these bands .................................................................................. 123

3.9. Electrophoretic trace with base calls for single-dye/four-lane strategy. Solid boxes indicate a mis-call from the automated sequencer, dashed-boxes indicate a miscall from lifetime base call and lines represent a deletion. The electrophoresis was performed at –1500 V with a loading volume of 1.2 µl of IRD700-labeled reactions. The average laser power was 1.0 mW and a laser repetition rate of 80 MHz ........................................................................................................ 125

4.1. Slab-gel image of near-IR tricarbocyanine labeled T-tracts run at –1500V in an 8% T crosslinked polyacrylamide gel. The bands in every lane travel at the same rate through the matrix, demonstrating uniform mobility within the dye set ........................................................................................................ 131

4.2. Near-infrared time-correlated single-photon counting instrument. Components: L, focusing lens; C, capillary; SF, spatial filter; BF, bandpass filter; SPAD, single photon avalanche diode detector; Amplifier, signal amplifier; Delay, time delay electronics; Disc., discriminator; PD, photodiode detector; TAC, time-to-amplitude converter ............................................................. 135
4.3. Structure of near-IR heavy-atom modified tricarbocyanine dyes with C6 amino linker and M13mp18 forward primer
4.4. Scheme of cleavage of tricarbocyanine dyes during reaction with dithiothreitol
4.5. Slab gel image of near-IR fluorine-modified tricarbocyanine dye T-tract before (2) and after (1) subjection to a cold ethanol precipitation
4.6. Slab gel image of sodium and ammonium acetate ethanol precipitation of near-IR bromine-labeled T-tract. Ammonium acetate precipitation shows lower fluorescence intensity due to the ion’s quenching effects on fluorescent molecules.
4.7. Difference in resolving power of a crosslinked gel capillary from the initial to the fourth electrophoresis run
4.8. Instrument response function of crosslinked gel column prior to electrophoresis and following four electrophoresis experiments. IRF from the final run was taken after the background value stabilized at the lowest point
4.9. Electropherogram of NIR-Cl G-tract injected for 1 minute at 150 V/cm and run in a 3T3C gel column at 170 V/cm, 50 minute delay before data acquisition
4.10. Decay profiles of NIR-I: A) 6% PA 7M urea, B) 6% PA 3.5M urea 30% formamide, and C) 6% PA 60% formamide
4.11. Structure of common denaturants, urea and formamide, used in electrophoresis
4.12. Separation of NIR-H G-tract in a commercial linear polyacrylamide with urea as the denaturant. Top panel was run the week the gel was received and bottom panel three weeks later. Degradation of the separation efficiency of the gel over a time-period is evident, under the same electrophoretic conditions, although the spectroscopic properties did not show any change
4.13. Structure of polydimethylacrylamide
4.14. Histogram of lifetime values calculated from labeled DNA fragments electrophoresed in POP6 with standard deviation and relative standard deviation
4.15. Electropherograms of near-IR labeled dye-primer (top) and dye-terminator (bottom) sequencing reactions separated in POP6. The background is noticeably lower and more stable in the dye-terminator separation .......... 157

4.16. Decay profiles of POP6 before and after the dye-primer and dye-terminator samples electrophorese through the detection zone ................................................. 158

4.17. Slab gel image of NIR-H G-tract dye-primer sequencing reactions with and without a size exclusion clean up ................................................................. 160

4.18. Electropherogram of dye-primer sequencing reactions with and without the size-exclusion step............................................................................................. 161

5.1. Structure of near-IR heavy-atom modified tricarbocyanine dyes developed for identification of terminal bases in DNA sequencing and IRD800 labeled with a M13mp18 (-29) sequencing primer.............................................................. 167

5.2. Near-Infrared time-correlated single-photon counting instrument. Components: L, focusing lens; C, capillary; SF, spatial filter; BF, bandpass filter; MO, objective; SPAD, single photon avalanche diode detector; DISC, discriminator; ADC, analog-to-digital converter; TAC, time-to-amplitude converter; MCA, multi-channel analyzer; PC, personal computer .................. 169

5.3. Structure of NIR-Cl labeled ddGTP and NIR-Br labeled ddCTP used in dye-terminator studies.......................................................... 172

5.4. Screen view of data collection and analysis program. The screen displays a single intensity line (that can be scaled) along a time axis. An intensity scale, that is color coded, and a description of the experiment are also displayed ... 173

5.5. Display used to set beginning and end time interval (in channel numbers) to search for peaks in electropherogram. Threshold and minimum slope values are also set for the peak recognition algorithm. Actual lifetime values can be displayed in the peak recognition or set values that are assigned to each base, including a value for unrecognized bases ........................................ 176

5.6. Base analysis of electropherogram of near-IR dye-labeled ddGTP and ddCTP fragments ................................................................. 176

5.7. Display screen used to chose MLE calculation process and to set the lifetime values and base associations used in peak recognition .................................... 177

5.8. Display screen for lifetime analysis of electrophoresis file ...................... 178
5.9. Instrument response function and decay profiles for IRD800, NIR-Br, ddGTP-NIR-Cl, and ddCTP-NIR-Br. Measurements taken from single base tracts run in POP6 ................................................................. 180

5.10. Comparison of slab-gel electropherogram of dye-labeled terminators to capillary two-dye-terminator run ................................................................. 181

5.11. Electropherogram of two-dye-terminator capillary run. Automated base calling was employed and followed by a pixel-by-pixel analyses in areas of poor electrophoretic resolution. Two miscalls resulting from poor electrophoretic resolution are noted by arrows. Run in POP6 at 135 V/cm. 183

6.1. Structure of water-soluble phthalocyanine and naphthalocyanine dyes for DNA sequencing. The metal center alters the lifetime of the dye .............. 188

6.2. PMMA microchips for separations are smaller than a ball-point pen. The walls of the PMMA chip are smooth and straight. The hot embossed PMMA microchip has a channel width of 20 μm, a channel depth of 100 μm, a separation channel length 4.2 cm, with an injection volume of 40 pL .......... 189

6.3. Near-IR microscope assembly for lifetime analysis of DNAs in PMMA microchips ................................................................................................. 190

6.4. Separation of IRD800 and IRD40 dye-primers in 1.0% hydroxyethyl cellulose (HEC) in TAE (pH = 8.2) ................................................................. 191

6.5. Schematic of a two-color detection instrument ......................................... 192
Abstract

The research presented in this dissertation involves the identification of sequencing fragments with time-resolved methods. For this application, near-infrared heavy-atom tricarbocyanine dyes were developed in our laboratory, which can be excited with a single laser and emission collected using a single detection channel. The dyes have four spectroscopically unique, but relatively short lifetimes that can be altered by the intramolecular heavy-atom they contain. The work described here involves the optimization of dye-primer chemistry for preparing Sanger sequencing reactions for longer reads and the optimization of the separation matrix for capillary gel electrophoresis that produces favorable statistical analysis of the aforementioned dyes’ lifetimes.

The performance of a two-lifetime experiment in which we modified an automated DNA sequencer to allow implementation of lifetime identification of DNA fragments labeled with near-IR fluorochromes and fractionation via slab-gel electrophoresis was investigated. A two-dye/two-lane sequencing experiment was carried out, in which two terminal bases, labeled with near-infrared dyes, were run in one lane and the other two bases in an adjacent lane. A lifetime evaluation of the resulting electropherogram on a pixel-by-pixel basis allowed the identification of the terminal nucleotide comprising a DNA band. The read accuracy was found to be better than a one-dye/four-lane approach using the software of the commercial instrument in spite of the fact that a spectroscopic call was implemented.
An automated peak recognition and base calling algorithm was also implemented and evaluated on two-tract dye-primer and dye-terminator capillary electrophoresis runs. The base calling accuracy was greater than 97% for both.
Chapter 1
DNA Sequencing Using Fluorescence Detection Methods

1.1. Genome Structure

The instructions for all-cellular structures and functions are encoded in the genome of any organism. The genome is made-up of deoxyribonucleic acid (DNA) that is tightly coiled into narrow threads. The threads of DNA are normally associated with many different types of proteins and are organized into structures called chromosomes that are housed within the nucleus of cells in the majority of eukaryotic organisms. The human genome consists of 23 pairs of chromosomes.

DNA is comprised of several different chemical units, which consist of a deoxyribose sugar unit, phosphate group and one of four different nucleotide bases adenine (A), guanine (G), cytosine (C), or thymine (T) (see Figure 1.1). It is the order of these bases that carries the code to build proteins that controls the function of various cells and also, determines the organism’s physical characteristics. In the human genome, 3 billion bases are contained within the 23 pairs of chromosomes, with the smallest chromosome containing 50 million bases and the largest 250 million bases. It is the primary function of DNA sequencing to determine the order of these nucleotide bases.

The bases are grouped into two different classes, the pyrimidines (T,C) and the purines (A,G). Since chromosomes are comprised of many of these individual nucleotide units strung together, the nucleotides are covalently attached via phosphodiester linkages, which occur between the phosphate group on the 5’ site of one sugar unit and the 3’ hydroxyl group on another nucleotide (see Figure 1.2). Therefore, DNA exists as a
Figure 1.1. Chemical structures of nucleotide base units: adenine, guanine, cytosine, and thymine; and dideoxy-phosphate unit (ddNTP or terminator) and deoxy-phosphate unit (dNTP).
biopolymer with repeating units consisting of deoxyribose and phosphate residues that are always linked together by the same type of linkage and form the backbone of the DNA molecule. Yet, the order of the bases along the biopolymer backbone can vary greatly and impart a high degree of individuality to any given DNA molecule.

James Watson and Francis Crick surmised that the two strands of the DNA helix (Figure 1.2) were held together via hydrogen bonds between a pair of bases on opposing strands [1]. From modeling, they found that A could only pair with T and G with C. Each of these base pairs possess a symmetry that permits it to be placed in the double helix in two different ways (A to T and T to A; C to G and G to C), giving rise to four possible permutations of sequence.

1.2. Methods for Determining the Primary Structure of DNA

DNA sequencing consists of three main steps including; mapping, sequencing, and assembly. Within these steps are a number of sub-steps, which include such processes as cloning and sub-cloning, template preparation, gel electrophoresis and computer algorithms required to assemble the small bits of sequencing data into contiguous strings that comprise the intact chromosome.

The important factor to consider when developing a sequencing strategy for whole genomes is that one can sequence only small sections (1,000-2,000 bp) of DNA due to instrument limitations and that entire chromosomes are comprised of well over 1 X 10^6 bp. The actual process of generating DNAs that can be handled by sequencing machines are typically involved and require a number of cloning and purification steps followed by actual sequencing and then assembling the pieces into contiguous regions of the target chromosome. Figure 1.3 gives a schematic diagram of a typical strategy that is
used in many sequencing laboratories. This strategy is termed an ordered shotgun approach and starts with the mechanical shearing (breaking apart) of intact chromosomes into pieces composed of 100,000 to 200,000 bp [2]. These sheared DNAs are then cloned into yeast artificial chromosomes (YACs). One sheared section per YAC allows the amplification of the number of copies of the insert. Cloning provides an unlimited amount of material for sequencing and serves as the basis for construction of libraries,
which are random sets of cloned DNA fragments. Genomic libraries are sets of overlapping fragments encompassing an entire genome. Once these libraries have been constructed, single inserts are extracted from the library and further sheared into fragments ranging from 1,000 – 2,000 bp. These fragments are then sub-cloned into M13 vectors to produce high quality single stranded DNAs appropriate for actual sequencing. Typically, M13 sub-clones are sequenced from both ends to allow construction of maps of the single YAC inserts, which are then used to provide a scaffold for the complete

Figure 1.3. Flow chart for ordered shotgun sequencing.
sequence analysis of the YAC insert. The important procedures that will be discussed here will be the procedures that are actually used to produce the sequence data of the M13 sub-clones. The two most common procedures used are the Maxam-Gilbert chemical degradation method and the Sanger dideoxy-chain termination method [3, 4]. The commonality in both of these methods is the production of a nested set of fragments that are all terminated (or cleaved) at a common base.

1.2.1. Maxam-Gilbert Sequencing

The Maxam-Gilbert sequencing method uses chemical cleavage methods to break single-stranded DNA molecules at either one or two bases, which is followed by a size fractionation step to sort the cleaved products. The cleavage reaction involves two different reactions; one that cleaves at G and A (purine) residues and the other that cleaves at C and T (pyrimidine) residues. The first reaction can be slightly modified to cleave at G only and the second at T only. Therefore, one can run four separate cleavage reactions, G only, A + G, T only, and T + C, from which the sequence can be deduced. In each cleavage reaction, the general process involves chemically modifying a single base, removing the modified base from its sugar and finally, breaking the bond of the exposed sugar in the DNA backbone.

The chemical steps involved in G cleavage are shown in Figure 1.4. In this step, dimethylsulfate is used to methylate G. After eviction of the modified base via heating, the strand is broken at the exposed sugar by subjecting the DNA to alkali conditions. To cleave at both A and G residues, the procedure is identical to the G cleavage reaction except that a dilute acid is added after the methylation step. Subjecting the DNA to hydrazine to remove the base and piperidine to cleave the sugar-phosphate backbone
Figure 1.4. Cleavage of guanine in the Maxam-Gilbert sequencing method.

...carries out the reaction that cleaves at either a C or C and T residue. The extent of each reaction can be carefully limited so that each strand is cleaved at only one site.

The entire Maxam-Gilbert process is depicted in Figure 1.5. As can be seen, four cleavage reactions are run, G, G + A, T, T + C. Prior to chemical cleavage, the intact DNA strand is labeled (typically with a $^{32}$P radiolabel for detection at the 5' end). Following chemical cleavage, the reactions are run in an electrophoresis gel...
Label many copies of template DNA at 5' ends

\[
\begin{array}{c|c|c}
5' & 32\text{PATGACCGATTTCGC} & 3' \\
3' & TACTGGCTAAACG32P & 5'
\end{array}
\]

Separate strands

\[
\begin{array}{c|c|c}
5' & 32\text{PATGACCGATTGC} & 3' \\
\end{array}
\]

Divide copies into four batches

\[
\begin{array}{c|c|c|c}
G & G+A & T+C & C \\
\end{array}
\]

Cleavage reaction mixture

\[
\begin{array}{c|c|c|c}
5' & 32\text{PATGACCGATTTCGC} & 3' \\
\end{array}
\]

Original strand

\[
\begin{array}{c|c|c|c|c}
32\text{PATGA} & CGATTTCGC & 32\text{PATGAC} & GATTTTCGC & 32\text{PATGACCGATTTCGC} \\
\end{array}
\]

Products from cleavage at C

Perform electrophoresis

Create autoradiogram

\[
\begin{array}{c|c|c|c|c}
32\text{PATGACCGATTTCGC} & 32\text{PATGACCGATTTCGC} & 32\text{PATGACCGATTTCGC} & 32\text{PATGACCGATTTCGC} \\
\end{array}
\]

Fragments cleaved at

Sequence of fragments cleaved at G

Figure 1.5. Steps involved in Maxam-Gilbert sequencing strategy.
(polyacrylamide) and separated based on size. The actual sequence of the strand is then deduced from the generated gel pattern.

1.2.2. Sanger Chain Termination Method

The Sanger procedure is an enzymatic method and involves construction of a DNA complement to the template whose sequence is to be determined. The complement is a strand of DNA that is constructed with a polymerase enzyme, which incorporates single nucleotide bases according to Watson-Crick base pairing rules (A-T; G-C) (see Figure 1.2). The nested set of fragments is produced by interrupting the polymerization by inserting into the reaction cocktail a base that has a structural modification, which consists of the lack of a hydroxyl group at the 3' position of the deoxyribose sugar (dideoxy nucleotide, ddNTP). Figure 1.1 shows the chemical structure of a ddNTP. When mixed with the deoxynucleotides (dNTP), the polymerization proceeds until a ddNTP is incorporated.

The Sanger sequencing process requires labeling and amplification of the DNA before analysis and uses a thermally cycled reaction (cycle sequencing) to increase the number of nested fragments for analysis (see Figure 1.6). The process is typically carried out by adding to the template DNA, a primer, a short oligonucleotide that has a known sequence and anneals (binds) to a complementary site on the unknown template. This primer can carry some type of label, for example a covalently attached fluorochrome. However, situating the fluorescent label on the ddNTP can be done as well. Following annealing of the primer to the template, the reaction progresses by the addition of a polymerase enzyme, all four dNTPs and one particular ddNTP. Therefore, four reactions are carried out, each containing a particular ddNTP or a single reaction if labeled terminators are
Figure 1.6. Cycle sequencing used to linearly amplify DNA.
Figure 1.7. Steps in the Sanger sequencing method.
Following polymerization, the reactions are loaded onto an electrophoresis gel and size fractionated. The final step involves reading the sequence from the gel. The entire Sanger, chain-termination protocol is shown in Figure 1.7. The Sanger method is the preferred sequencing method for most large-scale sequencing projects due to its ease of preparing the nested set of fragments. Many times reactions can be run under standard conditions without the need for chemical additions at timed intervals. In addition, the process is very conducive to automation.

1.3. Modes of Electrophoresis

Whatever detection method is chosen for identifying (calling) bases, the important analytical technique that is necessary is a fractionation step, in which the DNA fragments are sorted by size. There are three major gel electrophoresis platforms used for size separation of Sanger products; slab, capillary or microchip gel electrophoresis.

The common feature in all electrophoresis formats is the use of an electric field to shuttle the DNAs through a polymer media, which possesses pores of a definitive size. In all electrophoresis formats, the mobility of the molecule ($\mu$, cm$^2$/Vs, defined as the steady state velocity per unit electric field strength) in an electric field is determined by;

$$\mu = \frac{q}{f}$$  \hspace{1cm} (1.2)

where $q$ is the net charge on the molecule and $f$ is the frictional property of the molecule and is related to its conformational state as well as the molecular weight (MW) of the molecule. In the case of DNAs, either single stranded or double stranded, $f \sim (MW)^1$ because the DNA acts as a free-draining coil. Since $q$ is also related to the length of the
molecule, one finds that the mobility is independent of the number of bases comprising the DNA molecule for oligomers > 20 base-pairs (in free solution) [5]. Because of these properties of DNA, a polymer network is necessary for the separation of DNA fragments.

For any type of analytical separation, resolution is a key parameter that is optimized to improve the performance of the separation. The resolution (R) for electrophoretic separations can be calculated from the simple relationship:

\[
R = \frac{1}{4} \frac{\Delta \mu_{\text{app}}}{\mu_{\text{app,avg}}} N^{\frac{1}{2}}
\]  
(1.3)

where \(\Delta \mu_{\text{app}}\) is the difference in mobility between two neighboring bands, \(\mu_{\text{app,avg}}\) is the average electrophoretic mobility of the bands and N is the plate number that represents the efficiency (bandwidth) for the electrophoresis. The plate numbers are calculated from retention time of the peak, \(t_R\), and the width at half height of the peak, \(w_{1/2}\), given by:

\[
N_{\text{eff}} = \frac{L}{H} = 5.54 \left( \frac{t_R}{w_{1/2}} \right)^2
\]  
(1.4)

Resolution for capillary gel electrophoresis can also be calculated using the standard chromatography equation:

\[
R = \frac{t_2 - t_1}{\frac{1}{2}(w_2 + w_1)}
\]  
(1.5)

where \(t\) is the migration time of the electrophoretic band and \(w\) is the width at half height of the peak. As can be seen from equation 1.5, the resolution can be improved by increasing the difference in the mobility of the two bands (increased selectivity, gel properties). When \(R=1.5\) two bands are baseline resolved. For DNA sequencing, the accuracy in the base call depends greatly on the resolution obtained during gel fractionation.
Common sieving matrices used for DNA sequencing are linear or cross-linked polyacrylamides or polydimethylacrylamides, both of which possess the appropriate pore size distribution for sorting single stranded DNAs. Polyacrylamides are prepared from the acrylamide monomer unit (CH$_2$=CHCONH$_2$), Figure 1.8, which is copolymerized with a certain percentage of cross-linker, N,N'-methylenebisacrylamide (CH$_2$(NHCOCH=CH$_2$)$_2$), in the presence of a catalyst accelerator-chain initiator mixture for cross-linked gels. Acrylamide monomer can be polymerized in the absence of

![Figure 1.8](image.png)

**Figure 1.8.** A) Free radical polymerization scheme of acrylamide. B) Cross-linking polymerization reaction of acrylamide and bis-acrylamide. Ammonium persulfate becomes a free radical when dissolved in water and when added to an acrylamide solution in the presence of the catalyst tetra-methylenethyldiamine, TEMED, the reaction proceeds [6].
Figure 1.9. Dependence of pore size on acrylamide concentration [7].

bisacrylamide to form linear gels. The relative proportion of acrylamide monomer to crosslinking agent or total acrylamide concentration determines the porosity of the gel, Figure 1.9. For reproducible fractionation of DNAs, a single strand conformation must be maintained, which is accomplished by adding denaturants, such as urea or formamide, to the sieving gel.

1.3.1. Slab Gel Electrophoresis

DNA sequencing was originally performed in slab gels and it continues to be a widely used technique. In slab gel electrophoresis (block diagram of system shown in Figure 1.12A) polymerization of acrylamide as well as the electrophoresis is conducted in a mold formed by two glass plates with thin spacers ranging from 1 to 0.2 mm. When analytical electrophoresis is performed, several samples, usually 24 to 96 lanes are run simultaneously in the same slab. Since all samples are present in the gel, the conditions
of electrophoresis are quite constant from sample to sample. In slab gel electrophoresis, the sample is loaded using a pipette into wells formed during polymerization (see Figure 1.10) with a typical loading volume of 1-10 µl. The field strength that can be applied to these types of gels range from 50-80 V/cm, with the upper limit determined by Joule

**Figure 1.10.** Cartoon of sample wells formed in the separation matrix during polymerization by a plastic or paper comb displacing the matrix.
Figure 1.11. Simplified block diagrams of (A) a slab gel electrophoresis system and (B) a capillary gel electrophoresis system for DNA analysis.
heating caused by current flow through the gel. Due to the thick nature of the gel, the heat is not dissipated effectively causing convective mixing and thus, zone broadening (where bands of the analyte widen and the resolution is significantly decreased and many times all analytes elute in a continuous band), and thus limits the upper level on the electric field strength that can be effectively used.

1.3.2. Capillary Gel Electrophoresis

In order to reduce the development time in the electrophoresis, many sequencing applications now utilize capillary gel electrophoresis, in which a fused silica tube (i.d. = 20-100 µm; length = 30-50 cm) contains the sieving matrix and an electric field is applied to this capillary column, (see Figure 1.11) [8]. Due to the high surface-to-volume ratio afforded by the thin glass capillary, large electric fields can be applied (~300 V/cm), which result in shorter electrophoresis development times (2-3 hrs) and also, enhanced plate numbers (millions) compared to slab gel electrophoresis.

The glass capillary contains silanol groups on its surface, which at high pH are deprotonated, making the capillary wall more negatively charged above pH = 7.0 (see Figure 1.12). When a voltage is applied across the capillary, the cations migrate and build up at the wall, producing an electrical double-layer and exert a force on the surrounding fluid causing a bulk flow of solution toward the cathode (negative terminal), (Figure 1.13). This electrically induced flow is called the electroosmotic flow and the mobility of molecules in the capillary is affected by this forced flow, defined by:

$$\mu_{app} = \mu_{EP} + \mu_{EOF}$$  \hspace{1cm} (1.6)

where $\mu_{app}$ is the apparent mobility of the analyte, $\mu_{EP}$ is the electrophoretic mobility of the analyte, and $\mu_{EOF}$ is the electroosmotic mobility. The EOF can interfere with the
Figure 1.12. EOF profile of fused silica at increasing pH values [9].

Figure 1.13. Schematic of electroosmotic flow (EOF) produced in a fused silica capillary when an electric field is applied. Positively charged molecules migrate ahead of neutral molecules, which travel with the EOF, and negatively charged molecules migrate more slowly than the neutral molecules.
electrophoretic separation of the DNAs, since the mass-to-charge ratio of DNA > 20 base pairs is almost identical and thus, they will travel in a single plug with the EOF [5]. Another hindrance to the separation of DNAs in a fused silica capillary is adsorption of the molecules to the wall through hydrogen bonding and ionic interactions. Therefore, in DNA separations using glass capillaries, the wall is coated with some type of polymer (linear polyacrylamide for example) to suppress the electroosmotic flow. After coating the wall, the capillary can then be filled with the sieving gel, which can be a linear polyacrylamide (no cross linking) or some other type of gel, such as hydroxy cellulose or poly(ethylene oxide) to size-separate the DNAs. In addition, crosslinked gels may also be used in these small capillaries, but since crosslinked gels are not free flowing like their non-crosslinked counterparts, they must be polymerized directly within the capillary tube and cannot be removed once polymerized. Most capillary-based DNA sequencers, such as the ABI 3700 (Perkin Elmer-Applied Biosystems) use linear gels since they can be easily replaced using high pressure pumping allowing the capillary to be used for multiple sequencing runs [10].

The electrophoresis is performed following the filling of the column with gel by inserting one end of the capillary into a sample containing the sequencing mix and applying an electric field to the capillary tube. The injection end is usually cathodic (negative), with the opposite end being anodic (positive). There are two types of injection for electrophoresis, hydrodynamic or electrokinetic injection. Hydrodynamic injection cannot be used for sequencing purposes since the sample must be pushed into the capillary with pressure, which results in the sieving matrix being displaced from the capillary. Electrokinetic injection allows for all charged molecules in the sample solution
to migrate into the capillary when an electric field is applied, Figure 1.14. The DNA molecules in the sample solution will migrate towards the anode until they encounter the gel matrix. The molecules then begin to migrate through the gel based upon their size, smaller molecules encountering less resistance. Because the smaller molecules are more readily injected into the gel matrix, they will be present in greater quantities than the larger DNA molecules, resulting in biased injection giving rise to decreased signal for larger fragment sizes. A large concentration of salt or template DNA can also affect the injection since upon increasing the salt concentration, DNA sequencing fragments become a smaller and smaller fraction of the total ionic strength of the sample solution [11].

During electrokinetic injection, the analytes can be “stacked” at the end of the capillary by maximizing the conductivity difference between the separation and sample matrix. When analytes reach the sample/matrix interface, they encounter a much lower electric field and their velocity is dramatically decreased, causing sample stacking at the interface.

By applying a fixed voltage for a certain period of time, a controlled amount of sample can be inserted into the column with the injection volume ranging from 1-10 nL. The amount of analyte injected, $Q$, into the capillary can be determined from [12];

$$Q = \frac{\pi d^2 c_s (\mu_{ep} + \mu_{w}) E t \lambda_b}{4 \lambda_s}$$  \hfill (1.7)

where $E$ is the field strength, $\lambda_s$ is the conductivity of the sample solution, $\lambda_b$ is the conductivity of the buffer in the capillary, $\mu_{ep}$ is the electrophoretic mobility of the
Figure 1.4. Electrokinetic injection of DNAs into a capillary gel column. A) DNAs of various sizes in solution with end of gel filled capillary. B) Once a voltage is applied, the DNA molecules move in solution at the same rate to the capillary interface where the molecules meet resistance from the gel. The smaller molecules are able to move through the gel more aptly and the larger molecules meet more resistance, giving rise to biased injection.
analyte molecules, and $\mu_{eo}$ is the electroosmotic flow, $t$ is the duration of the injection, and $c_g$ is the concentration of analyte in the sample matrix.

1.4. DNA Migration

The existence of a gel network produces the electrophoretic migration pattern (electrophoretic migration) observed in the sequencing process. Any DNA fragments to be fractionated will inevitably encounter the gel matrix of polymer threads and pores. There are a number of models that account for migration of DNAs in polymer matrices (see Figure 1.15).

1.4.1. Ogston Model

This is the oldest model developed in the 1950s for migration of proteins in a gel [5]. According to Ogston, the gel is a static, infinite network of long, inert, and randomly distributed linear fibers having a certain average pore size. A macro-ion, such as DNA, is assumed to electrophorese through this network as a spherical coil, which must diffuse laterally until it encounters a pore large enough in diameter to permit its passage. Mobility depends on the volume fraction available to the rigid, undeformed coil. Thus, when the radius of gyration is larger than the average pore size of the gel, the Ogston model predicts that the electrophoretic mobility of the solute will rapidly approach zero. In other words, if the uncoiled DNA is larger than an available pore, its migration will stop. Ogsten derived a simple expression for the mobility ($\mu$) of a globular particle of radius $R$ in a random array of fibers[13]:

$$\mu = \mu_0 \exp[-K_r C]$$

(1.1)
where $\mu_0$ is the mobility of the molecule in pure solvent and $C$ is the gel concentration. $K_R$ is the retardation coefficient, proportional to $(R + r)^2$ with $r$ being radius of the gel fibers.

1.4.2. Reptation Model

This model was developed to account for the fact that large DNAs, which if assumed to behave as rigid spheres, would simply be trapped at the injection site in a gel having pore sizes too small for the DNA, but instead were found to migrate despite all known physical facts. According to the reptation model, randomly coiled DNAs too large to fit through a pore while maintaining a coiled conformation will migrate head-first, snakelike, through “tubes” formed by the pore network of the gel. No lateral motion is allowed within the tube and the DNAs are thought to alternately stretch and relax as they slither through the tube. Thus, the reptation model assumes that large DNAs, instead of migrating as undeformable particles with a fixed radius of gyration, can deform and stretch according to local conditions including overly high voltages, too low of a pore size, or a combination of both.

1.3.3. Biased Reptation Model

According to this model, at high electric fields, and/or for DNAs larger than about 40 Kbp, electric field-induced orientation extends the stretching periods of DNA, causing their random walk to become strongly biased in the forward direction so that DNA is stretched to a rod-like conformation. In a fully biased reptation regime, the mobility increases to saturation and size-based separation is lost with all large fragments migrating at the same rate. Some effects of biased reptation can be reduced by pulsing, reversing
Figure 1.15. Representation of the different mechanisms of DNA migration in an array of fixed obstacles. A) Ogston sieving. B) Reptation without orientation. C) Reptation with orientation (alignment in electric field).

the electric field for a given period of time to relax the large DNA back into their original conformation before continuing the separation.

1.5. Detection Methods for DNA Sequencing

1.5.1. Autoradiographic Methods

Following electrophoresis, the individual DNA bands separated on the gel must be detected and analyzed. One of the earliest methods implemented to detect DNA bands in gels was audioradiography. In this mode, one of the phosphates of an individual
nucleotide is replaced with a radioisotope, typically $^{32}$P ($\tau_{1/2} = 14$ days) or $^{35}$S ($\tau_{1/2} = 87$ days), both of which are radioprobes that emit $\beta$-particles. The $^{32}$P is a better choice for a radionucleotide considering its similar structure to the phosphate in the DNA backbone and the shorter half-life, making it better for waste disposal issues. When Sanger methods are used to prepare the sequencing reactions either the primer or the dideoxynucleotide can contain the radiolabel. The labeling is done using an enzyme (T4 polynucleotide kinase), which catalyzes the transfer of a $\gamma$-phosphate group from ATP to the 5’-hydroxy terminus of a sequencing primer. After the electrophoresis has been run, the gel is dried and then situated on a X-ray film. The film is developed (exposure to radiation from radioprobes) and dark bands are produced on the film where the DNA was resident. The DNA bands are then manually read from this film to reconstruct the sequence (see Figure 1.16).

The primary advantage of this approach is the inexpensive nature of the equipment required to perform the measurement. It requires only a gel dryer, which removes the buffer/moisture from the polyacrylamide, a film holder and film. The difficulties associated with this approach are numerous. One important issue is the fact that radioisotopes are used, and therefore, waste disposal becomes a difficult problem to contend with. Throughput issues (data production rates) are also a concern since radiography can sometimes require several days to expose the film in order to get strong signals to read the gel, reducing throughput. Also, since there is no means to identify the individual bases using radioprobes, each base must be analyzed in a different lane of the gel. And finally, the bases must be called manually, which leads to frequent errors in sequence reconstruction. Therefore, the inability to obtain data production rates
sufficient to accommodate large sequencing projects has made radiographic detection obsolete for high throughput applications.

1.5.2. Fluorescence Detection

For most DNA sequencing applications, irrespective of the separation platform used, fluorescence is the accepted detection protocol for several important reasons. Fluorescence allows one to perform the base calling and detection in an automated fashion and alleviates the need for manual base calling. In addition, fluorescence can be carried out during the separation eliminating long film development times. More importantly, due to the ability to implement multiple probes possessing unique spectral properties, the four bases comprising the DNA molecule can be identified in a single gel lane potentially increasing throughput by a factor of four compared to radiographic detection. All of these important advantages associated with fluorescence allow for higher data production rates in DNA sequencing applications. As such, fluorescence can be considered as one of the most important technical innovations in DNA sequencing and has made it feasible to consider tackling large genome sequencing projects, such as the human genome.

The first demonstrations on the use of fluorescence in DNA sequencing came with the work of Smith et al. (1986), Probe et al. (1986) and Ansorge et al. (1987). In these examples, slab gels or large gel tubes were used to fractionate the DNA ladders produced during enzymatic polymerization using Sanger sequencing strategies. The fluorescence detection was accomplished using four spectroscopically unique probes, which allowed the DNA sequence reconstruction to be done in a single electrophoresis lane of the gel. The chemical structures of the dye labels used in the Smith et al. (1986)
Figure 1.16. Film of autoradiographic sequencing of an M13mp19 template with $^{32}$P labeling of deoxynucleotides. Electrophoresed on a slab gel instrument with a 9%T crosslinked polyacrylamide at –900 V/cm, 2 µl of each sequencing reaction injected into the sample wells. Expanded view of gel shown, demonstrating manual reading and base calling.
experiments are shown in Figure 1.17. The dyes were either attached (covalently) to the sequencing primer or to the dideoxynucleotides. The advantage of using dye-labeled dideoxynucleotides is that the sequencing reactions can be performed in a single reaction tube, whereas the dye-labeled primer reactions must be performed in 4 separate tubes and pooled prior to electrophoresis. In the case of the dye-labeled terminators, succinylfluorescein analogs were used with slight structural modifications to alter the absorption/emission maxima. The dyes were attached either to the 5 position of the pyrimidine bases or the 7 position of the 7-deazapurines, both of which are non-hydrogen bonding sites on the nucleotide base (see Figure 1.1). The linker structure is also important, which in this case was a propargylamine, since the presence of the dye onto the terminator radically affects its ability to be incorporated by the polymerase enzyme. For dye-labeled primers, the oligonucleotides possessing the appropriate sequence were prepared on a standard DNA synthesizer. In the Smith et al. (1986) work, a thymidine derivative was prepared, which contained a phosphoramidite at the 3’ carbon and a protected alkyl amino group at the 5’ carbon (typically a-6-carbon linker structure). During the final addition cycle of the oligonucleotide prepared via solid-phase synthesis using phosphoramidite chemistry, the thymidine residue is added and following deprotection of the alkyl amino group and cleavage from the support, a free primary amine group results, which can be reacted with any amino-reactive fluorescent dye to produce the oligonucleotide derivative.

Figure 1.18 shows the emission profiles for the dye set used in these experiments. The major attributes of the dye sets are that they can be efficiently excited with either 488 and/or 514.5 nm lines of the Ar ion laser. However, there is significant overlap in the
Figure 1.17. Structure of fluorescent dyes used for labeling primers for four-color DNA sequencing. Fluorescein and NBD are excited with the 488 nm lines of an Argon laser and Tetramethylrhodamine and Texas Red from the 514.5 nm lines [14].

The optical hardware similar to that used to process the four-color fluorescence for the Smith system is depicted in Figure 1.19. For this example, a single laser was used as well as a single detection channel, which in this case consisted of a multi-line Ar ion laser and a conventional photomultiplier tube (PMT). Placed in front of the laser and PMT were filter wheels to select the appropriate excitation wavelength (488 or 514.5 nm)
Figure 1.18. Emission spectra of dyes used for labeling primers; F, fluorescein; N, NBD; T, tetramethylrhodamine; TR, Texas Red [14].

and emission color. The filter pairs used during fluorescence readout were 488/520 nm; 488/550 nm; 514/580 nm; 514/610 nm. In the case of the Prober et al. experiment (1986), due to the narrow distribution between the excitation maxima of the dyes, only the 488 nm line from the Ar ion laser was required to excite all dyes and two PMT tubes were required to process the emission from the four dyes. Discrimination of the four colors was accomplished by monitoring the intensity of each dye on both detectors simultaneously. By histograming the ratio of the fluorescence intensity of each dye (produced from an electrophoresis band) on the two detection channels, a discrete value was obtained that allowed facile discrimination of the four different fluorescent dyes. In
order to determine the limit of detection of these four-color fluorescence systems, injections of known concentrations of dye labeled sequencing primers were electrophoresed. In both cases, the mass detection limit was estimated to be $10^{-17}-10^{-18}$ moles [15].

![Diagram of experimental set-up](image)

**Figure 1.19.** Experimental set-up used to excite and collect emission from a four-dye one lane sequencing run [16].

Unfortunately, reading the sequence directly from the raw gel data becomes problematic due to several non-idealities, including signal from a single dye appearing on multiple detection channels due to the broad and closely spaced emission bands, dye-dependent electrophoretic mobility shifts, and non-uniformity in the intensity of the electrophoresis bands due to the enzymatic reaction used to construct the individual DNA fragments. As such, several post electrophoresis processing steps were required to augment sequence reconstruction. In the case of the Smith et al. example, these steps
involved: 1) High frequency noise removal using a low-pass Fourier filter. 2) A
time-delay between measurements at different wavelengths corrected by linear
interpolation between successive measurements. 3) A multi-component analysis
performed on each set of four data points, which produced the amount of the four dyes
present in the detector as a function of time. 4) The peaks present in the data stream
located. 5) The mobility corrected for the dye attached to each DNA fragment. In this
case, it was empirically determined that fluorescein and rhodamine-labeled DNA
fragments moved as if they were 1 base longer than the NBD-labeled fragments and the
Texas Red fragments moved as if they were 1 1/4 bases longer.

The important performance criterion in any type of automated DNA sequencer is
its throughput, the number of bases it can process in a single gel read and its accuracy in
calling bases. In terms of base calling accuracy, these early instruments demonstrated an
error rate of approximately 1% with a read length approaching 500 bases. The throughput
of the instrument described by Prober et al. was estimated to give a raw throughput of
600 bases per hour (12 electrophoresis lanes). Interestingly, many present day
commercial automated sequencers still use similar technology in their machines and the
throughput can be as high as 16,000 bases per hour (96 electrophoresis lanes and shorter
run times).

1.5.2.1. Fluorescent Dyes for DNA Labeling and Sequencing

As stated, fluorescence detection has had a tremendous impact in the area of DNA
sequencing because of the speed of the readout process as well as the ability to
discriminate amongst the four nucleotide bases in a single gel lane due to the unique
spectral properties of the target dye molecules. There are a variety of dye sets (typically
four dyes per set, one for each nucleotide base) that have been developed for DNA sequencing applications and whichever dye set one wishes to consider, certain properties associated with the dyes for DNA sequencing applications are important. These properties are:

(1) *Each dye in the set must be spectroscopically distinct.* The ideal situation from a throughput point of view is to identify each base in a single gel lane instead of four gel lanes. In most cases, the discrimination is based on differences in the emission properties of each dye (distinct emission maxima), however, other fluorescence properties can be used as well, such as fluorescence lifetimes.

(2) *The dye set should preferably be excited by a single laser source.* It becomes instrumentally difficult to implement multiple excitation sources since lasers are typically used as the source of excitation to improve the limit of detection in the measurement and the upkeep on multiple lasers becomes problematic.

(3) *The dye set should possess high extinction at the excitation frequency and also large quantum yields in the gel matrix used to fractionate the DNAs.* Good photophysical properties are necessary in order to improve detectability. In addition, the dye set should show reasonable quantum yields in denaturing gels, which consist of high concentrations of urea and/or formamide.

(4) *The dye set should show favorable chemical stability at high temperatures.* A common procedure in most Sanger sequencing strategies is to implement a thermostable polymerase and then, subject the sequencing reactions to multiple temperature cycles (55°C – 95°C) in order to amplify the amount of product generated (cycle sequencing).
Therefore, the dyes must be able to withstand high temperature conditions for extended periods of time.

(5) *The dyes must induce minimal mobility shifts during the electrophoresis analysis of the sequencing ladders.* Many times, the mobility shifts from individual dyes can cause misordering of the bases during sequence reconstruction. As such, post-electrophoresis corrections are employed to correct for this perturbation.

(6) *The dyes must not significantly perturb the activity of the polymerase enzyme.* This is especially true in dye-terminator chemistry, since the proximity of the dye to the polymerase enzyme can dramatically influence the ability of the enzyme to incorporate the dye-ddNTP conjugate into the polymerized DNA molecule.

### 1.5.2.2. Visible Fluorescence Dyes for DNA Labeling

The common dye set that is frequently used in many automated, fluorescence-based DNA sequencers is the FAM, JOE, TAMRA, and ROX series (see Figure 1.20), which consist of fluorescein or rhodamine analogs containing a succinimidyl ester for facile conjugation to amine terminated sequencing primers or terminators. These dyes can be efficiently excited by the 488 or 514.5 nm lines from an Ar ion laser and also, possess emission profiles that are fairly well resolved.

Many of the Applied Biosystems automated DNA sequencers use this particular dye set. Shown in Figure 1.21 are emission spectra for this dye set as well as the filter set that is used to isolate the emission from the dyes onto the appropriate detection channel. While this dye set is fairly robust and works well with typical DNA cycle sequencing conditions, there are some difficulties in using this set, most noticeably the broad emission profiles, dye-dependent mobility shifts and inefficient excitation of TAMRA
Figure 1.20. Structure of common fluorescence dye set used for labeling primers in DNA sequencing applications. The functional group on each dye is a succinimidyl ester, which readily conjugates to primary amine groups.
**Figure 1.21.** Emission spectra and filter set used for detection of FAM, JOE, TAMRA, and ROX dyes.

and ROX with the 514.5 nm line from the Ar ion laser. Many of the dyes discussed have been employed in dye-primer sequencing applications, in which the primer used for selecting the DNA polymerization site on the unknown template is determined by Watson-Crick base pairing. An alternative is dye-terminator chemistry, where the fluorescent dyes are covalently attached to the ddNTP used in Sanger sequencing.
strategies. The advantages of using these conjugates are twofold: (1) The sequencing reactions can be carried out in a single tube. In dye-primer chemistry, the sequencing reactions are carried out in four separate tubes and then pooled prior to the electrophoresis step. If one implements four spectroscopically unique fluorescent probes attached to the ddNTP, the reactions can be carried out in a single tube, which reduces reagent consumption and also, minimizes sample transfer steps. (2) Primers of known sequence do not need to be synthesized. In primer sequencing, making oligonucleotides of 17-23 bases in length can be costly and time consuming, especially if the dye must be chemically tethered to the primer. The use of dye-terminator chemistry eliminates the need for synthesizing dye-labeled primers. Unfortunately, dye-terminators themselves can be quite expensive and also, many polymerase enzymes are very sensitive to the type of dye attached to the ddNTP. For example, fluorescein dyes are poor labels for terminators when using Taq polymerase due to its incompatibility with the binding pocket of Taq, while the rhodamine dyes are more hydrophobic and as such are more suitable for use with Taq polymerase. The result is that the peak heights for the electrophoresis bands can vary tremendously due to differences in incorporation of the dye-modified ddNTPs by the particular polymerase enzyme. In dye-primer chemistry, this disparity is absent due to the large displacement of the dye from the polymerization site [17]. It is interesting to note that several mutant forms of Taq polymerase have been prepared to allow more facile incorporation of dye-labeled terminators [18]. For example, Taq Pol I (AmpliTaq FS) has two modifications in it, a substitution which eliminates the 3′→5′ nuclease activity and also a substitution that improves 2′,3′ddNTP incorporation.
The nature of the dye on the terminator can also influence the mobility of the polymerized DNA fragment as well. For example, rhodamine dyes are typically zwitterionic and as such, appear to stabilize hairpin (secondary) structures in the DNA fragment causing compressions in the electrophoresis data, especially in GC-rich regions of the template [17]. On the other hand, the fluorescein dyes, which are negatively charged, due not produce such anomalies.

Slight structural modifications on the base chromophore can also influence its incorporation during DNA polymerization. In addition, the linker structure can influence the incorporation efficiency. A set of dye-labeled terminators have been found that give fairly even peak heights in sequencing patterns and produce minimal mobility shifts [19]. The linker structures chosen for this set were either the propargylamine linker developed by Prober et al. [15] or a propargyl ethoxyamino linker. The choice of linker was selected to accommodate the polymerase enzyme and to minimize the mobility differences within the dye set. For the rhodamine terminators, very weak G-peaks, which appeared after A-peaks, were observed. However, for the d-rhodamine terminator, this disparity in peak intensity was alleviated with the peak heights in the pattern being much more uniform [19].

1.5.2.3. ET Dyes for DNA Sequencing

One of the major problems associated with many of the dye sets mentioned previously is that their absorption spectra are dispersed over a relatively large spectral range, which provides poor excitation efficiency even for dual laser (488 and 514 or 543 nm) systems. As such, the red-dyes are used at higher concentrations during DNA polymerization to circumvent poor excitation. To overcome this problem without
sacrificing spectral dispersion in the emission profiles, the phenomena of Förster energy transfer has been used to design sequencing primers, which can be efficiently excited with a single laser line [20-23]. It is informative to briefly introduce these ET (energy transfer) primers at this time in the context of DNA sequencing.

The chemical structures of the ET primers developed by Richard Mathies and his group are shown in Figure 1.22 [20]. The donor dye in this case was FAM, which could be excited with the 488 nm line of an Ar ion laser. The acceptor dyes were either FAM, JOE, TAMRA or ROX. The donor (FAM) was attached to the sequencing primer on the 5' end during the solid-phase DNA synthetic preparation of the M13 (40) sequencing primer using phosphoramidite chemistry. The sequencing primer also contained a modified base (T*) that possessed a linker structure with a primary amine. The appropriate acceptor was conjugated to the primary amine group off the modified base following cleavage from the solid support via a succinimidyl ester functional group. The spacer distance between the donor and acceptor was selected by positioning T* within the M13 (40) primer sequence during solid phase synthesis. The naming of these ET primers followed the convention, donor-spacer (bp)-acceptor. In the case of the FAM and JOE ET primers, a 10 base spacer was used and for TAMRA and ROX, a 3 base spacer was selected. The choice of spacer size was primarily determined by producing ET primers, which showed uniform electrophoretic mobilities.

Shown in Figure 1.23 are the absorption and emission profiles of the ET primer series. While the absorption spectra show bands from both the acceptor and donor dyes, the emission profiles are dominated by fluorescence from the acceptors. In fact, the energy transfer efficiency has been determined to be 65% for F10J, 96% for F3R and
Figure 1.22. Chemical structures of energy transfer (ET) labeled DNA sequencing primers [20].

97% for F3T [20]. As can be seen from the emission profiles displayed in Figure 1.23, the emission intensity was found to be significantly higher for the ET primers compared to the single dye primers when excited with 488 nm laser light from the Ar ion laser due to higher efficiency in excitation. This translates into improved fluorescence sensitivity of these ET primers for single frequency excitation.
Figure 1.23. Absorption (dark line) and emission spectra for both the ET primers and the single dye-labeled primers are shown for comparison. The number in parenthesis is the excitation wavelength used for collecting the emission profile [20].
Using these ET primers does offer some advantages due to their improved detection sensitivities, namely eliminating the need for adjusting the concentrations of the dye primers during polymerization and also the need for smaller amounts of template in the sequence analysis. In fact, the use of ET primers required about 1/4 the amount of template compared to the single dye primers.

Energy transfer dye pairs can also be situated on terminators (Figure 1.24) [19]. In one example, d-rhodamine and rhodamine dyes are used with a propargyl ethoxyamino linker. This dye set has been called Big Dye terminators. Again the dyes were selected so as to provide fairly uniform peak heights in the electrophoresis and also, uniform mobility shifts within the series.

1.6. Instrumental Formats for Fluorescence Detection in DNA Sequencing

The ability to read the fluorescence during the electrophoresis and accurately identify the terminal base is a challenging task due to a number of technical issues. A variety of fluorescence readout devices have been developed for reading such data. When considering the designs of a fluorescence detector for sequencing applications several instrumental constraints must be included in the design. High sensitivity, as mentioned earlier, is a key issue since the amount of material loaded into the gel can be in the low attomole range (10^{-18} moles) and the detector must be able to read this fluorescence signal with a reasonably high signal-to-noise in order to accurately call the base. Therefore, in most cases, the excitation source is a laser that is well matched to the excitation maxima of the dye set used in the sequencing experiment.

Identifying the bases using a spectral marker is another constraint put on the instrument. One of the four bases terminating the sequence fragments must be identified
Figure 1.24. Structure of Big Dye terminator with energy transfer dyes Tetramethylrhodamine and Fluorescein attached to ddTTP [19].

accurately by processing the fluorescence via spectral discrimination or some other fluorescence property, such as the lifetime. For spectral discrimination the fluorescence is sorted by wavelength using either filters or gratings and analyzed in multiple detection channels.
In most sequencing instruments, the fluorescence must be read and processed from multiple gel lanes or capillaries. This can either be done by utilizing a scanning system in which the relay optic is rastered over the gel lanes or capillaries or an imaging system, in which the fluorescence from multiple gel lanes or capillaries are imaged onto some type of multichannel detector, for example a charge coupled device (CCD) or photodiode array.

Finally, sequencing instruments should be robust, considering the majority of end users are novice operators and the instrument is run over extended periods of time. The detector format also must be able to operate on a dependable basis and be a simple turn key operation.

The short list of requirements for all types of fluorescence readout devices appropriate for sequencing presents it with many challenges that are often non-complementary with the sequencing requirements. For example, high sensitivity is a particularly demanding requirement since the separation platforms used to fractionate the DNA are becoming smaller and consuming smaller amounts of material. In addition, detecting material directly in the gel matrix can be problematic due to the large number of scattering photons it produces. The signal is also transitory in that the DNA fragment resides within the probing volume of the laser for only a few seconds and many separation lanes must be interrogated for high throughput applications. In addition to these considerations, high signal-to-noise is required to obtain high accuracy in the base calling. As such, significant design considerations go into fabricating a fluorescence detection system for DNA sequencing applications.
1.6.1. Fluorescence Scanning Instruments

In a system used by Huang et. al., a confocal geometry with epi-illumination was used in which the objective used to collect the emitted fluorescence also served to focus the laser beam into individual capillaries (or lanes of the slab gel) used for the electrophoresis (see Figure 1.25) [24, 25]. Following collection, the emission was focused onto a spatial filter at the secondary image plane of the collection objective. The laser light (488 nm, Ar ion laser, 1 mW) was directed into the objective using a dichroic filter. The capillaries were held into a linear array and in this particular example, the capillary array was translated beneath the microscope objective. The laser only irradiates one capillary at a time, but since the beam was tightly focused (diameter = 10–12 µm) the electronic transition could be saturated at relatively low laser powers, improving the signal-to-noise ratio in the fluorescence measurement [26]. In addition, the noise can be significantly suppressed in this system, since a pinhole is used in the secondary image plane of the collection microscope objective preventing scattered out-of-focus light generated at the walls of the capillary from passing through the optical system. The capillary array was scanned at a rate of 20 mm/s with the fluorescence sampled at 1500 Hz/channel (color channel) resulting in a pixel image size of ~13.3 µm. The fluorescence was collected by a 32X microscope objective (numerical aperture = 0.4) resulting in a geometrical collection efficiency of approximately 12%. After the fluorescence was collected by the objective, it passed through a series of dichroics to sort the color and then processed on one of four different PMTs with each PMT sampling a different color (spectral discrimination). While the present system was configured with four-color channels, the system could easily be configured to do two-color processing as well by
Figure 1.25. Schematic of laser-excited, confocal-fluorescence capillary array scanner. The fluorescence is sampled unidirectionally at 1500 Hz/channel with a resolution of 13.3mm/pixel [24].

removing the last dichroic filter in the optical train and two of the PMTs. The concentration limit of detection of this system was estimated to be $2 \times 10^{-12}$ M (SNR = 3), which was determined by flowing a solution of fluorescein through an open capillary [25]. The detection limit would be expected to degrade in a gel filled capillary due to the higher background that would be generated by the gel matrix.
1.6.2. Fluorescence Imaging Systems for DNA Sequencing

An example of an imaging system for reading fluorescence from multichannel capillary systems for DNA sequencing is shown in Figures 1.26 and 1.27 [27, 28]. In this example, a sheath flow cell was used with the laser beam(s) traversing the sheath flow and the capillary output dumping into the sheath stream. This geometry allows simultaneous irradiation of all the material migrating from the capillaries without requiring the laser beam to travel through each capillary, which would cause significant scattering and reduce the intensity of the beam as it travels through the array. The sheath flow cell also caused contraction of the fluid output of the capillary since the sheath flow runs at a greater linear velocity compared to the sample (capillary) stream. A fluorescence image of the output from the capillary array indicated that the sample stream diameter at the probing point was ~0.18 mm and also demonstrated minimal cross-talk between individual capillaries in the array. The laser beams (488 nm from Ar ion, 6 mW; 532 nm from frequency doubled YAG, 6 mW) were positioned slightly below the exit end of the capillaries with the beams brought colinear using a dichroic filter. The collection optic and focusing optic produced a total magnification of 1 and resulted in a geometrical collection efficiency of 1%. In order to achieve multi-color processing capabilities, the collected radiation was sent through an image-splitting prism to produce four separated (spectrally) line images on the array detector (one for each dye used to label the terminal bases) (see Figure 1.27). In addition, a series of narrow bandpass filters were placed in front of the image-splitting prism to assist in isolating the appropriate colors for data processing.

The detector that was used for this system was a 2-dimensional CCD camera with
a cooled image intensifier. Interestingly, the detection limit reported for this system was found to be $2 \times 10^{-12}$ M when operated in the four-color mode, comparable to that seen for the scanning system discussed above. However, it should be pointed out that the presence or absence of the gel matrix will not affect the sensitivity of the fluorescence measurement in this case, since the fluorescence interrogation is done off column in the sheath flow. In fact, researchers have reported that the implementation of the sheath flow geometry in gel electrophoresis can offer a significant improvement in limits of detection by minimizing scattering contributions to the background [29].

![Figure 1.26. Schematic of the multiple sheath flow cell using gravity feed for the sheath flow. Twenty gel filled capillaries were aligned at a 0.35 mm pitch in an optical cell [28].](image)
When comparing these two fluorescence readout systems, several issues should be highlighted. One is the duty cycle, which takes into account the loss in signal due to multiple lane sampling. For any type of scanning system, the sampling of the electrophoresis lanes is done in a sequential fashion. For a scanning system sampling 96 capillaries, the duty cycle is approximately 1%. However, in the imaging system all capillaries are sampled continuously and the duty cycle is nearly 100%. Therefore, comparisons of detection limits for any system must include a term for the duty cycle, since lower duty cycles will degrade the limit of detection.

1.6.3. Fluorescence Detection Formats

Several different formats can be implemented to reconstruct the sequence of the template when using fluorescence detection. In the case of spectral determination, these
formats can vary in terms of the number of dyes used, the number of detection channels required, or the need for running 1 to 4 parallel electrophoresis lanes. For instance, if the sequencing instrument has no spectral discrimination capabilities, the electrophoresis must be run in four different lanes, one for each base comprising the DNA molecule. Yet, if four unique dyes are used, the electrophoresis can be reduced to one lane. As a consequence, the production rate of the instrument is increased by a factor of four. The fluorescence-based strategies that will be discussed here are (number of dyes/ number of electrophoresis lanes), the single dye/ four lane; single dye/single lane; two-dye/single lane; and finally four-dye/single lane.

The most pressing issue in any DNA sequencing format is the accuracy associated with the base call, which is related to a number of experimental details. Some of these specifics include the number of spectral channels used in the instrument as well as the signal-to-noise ratio in the measurement. The information content of a signal, I, can be determined from the simple relation introduced by Swerdlow and co-workers [16];

$$I = n \log_2(SNR)$$  \hspace{1cm} (1.8)

where n is the number of spectral channels and SNR is the signal-to-noise ratio associated with the measurement. The term I is expressed in bits and typically at least 2 bits are necessary to distinguish between four different signals if there is no significant overlap between the dyes used for identifying the bases. However, in the majority of multi-color systems, the spectra of the dyes used in the sequencing device show significant overlap and as such, many more bits are required to call bases during the sequencing run. Figure 1.28 shows idealized graphs of intensity vs. signal-to-noise plots with increasing detection channel number. As the number of spectral channels increase, the SNR of
Figure 1.28. Graphs of equation 1.8 for values n=1, 2, 3, 4. For increasing number of detection channels an increase in the signal-to-noise ratio and intensity is needed for the identification of a fluorescence signal.

the signal must increase also for identification of the fluorescence signal from the magnified noise created from multiple detection channels.

While equation 1.8 can provide information on how to improve the accuracy of the base call, it does not provide the sequencer with information on the identity of the individual electrophoretic peaks (base call) nor the quality of a base call within a single gel read. For example, if four color sequencing is used with dye-primer chemistry, how should one process the data and what is the confidence to which an electrophoretic peak is called an A, T, G or C? In order to provide such information, an algorithm has been
developed to not only correct for anomalies associated with fluorescence-based sequencing, but also assign a quality score to each called base. The algorithm that is used is called the *Phred* scale and it uses several steps to process the sequencing data obtained from fluorescence-based automated DNA sequencers [30].

The data input into *Phred* consist of a trace, which is electrophoretic data processed into four spectral channels one for each base. The algorithm consists of four basic steps:

1. Idealized electrophoretic peak locations (predicted peaks) are determined. This is based on the premise that most peaks are evenly spaced throughout the gel. In regions where this is not the case, typically during the early and late phases of the electrophoresis, predictions are made as to the number of correct bases and their idealized locations. This step is carried out using Fourier methods as well as the peak spacing criterion and helps to discriminate noise peaks from true peaks.

2. Observed peaks are identified in the trace. Summing trace values to estimate the area in regions that satisfy certain criterion identifies peaks. If the peak area exceeds 10% of the average area of the preceding 10 accepted peaks and 5% of the area of the immediate preceding peak, it is accepted as a true peak.

3. Observed peaks are then matched to the predicted peak locations, omitting some peaks and splitting others. In this phase of the algorithm, the observed peak arises from one of four spectral channels, and thus, can be associated with one of the four bases. It is this ordered list of matched observed peaks which determines a base sequence for the DNA template in question.

4. Observed peaks that are uncalled (unmatched to predicted peaks) are
processed. In this step, an observed peak that did not have a complement in the predicted trace is called and assigned a base and finally inserted into the read sequence.

As can be seen, this algorithm mainly deals with sorting out difficulties associated with the electrophoresis by identifying peaks in the gel traces, especially in areas where the peaks are compressed (poor resolving power) or where multiple peaks are convolved or significantly overlapping due to significant band broadening produced by diffusional artifacts.

Many times, pre-processing of the traces is carried out prior to Phred analysis to correct for dye-dependent mobility shifts. In most cases, these mobility shifts are empirically determined by running an electropherogram of a single dye-labeled DNA ladder (for example, T-terminated ladder) and comparing the mobilities to the same ladder but labeled with another dye of the set. This type of analysis can be very complex and involved, since the mobility shift is not only dependent upon the dye and linker structure, but also upon the separation platform used. For example, dyes which show uniform mobility shifts in slab gel electrophoresis may not show the same effect in capillary gel electrophoresis. In addition, these mobility shifts can be dependent upon the length of the DNA to which the dye is attached [31].

1.7. Sequencing Strategies

1.7.1. Single-color/four-lane

In this processing format, only a single fluorescence detection channel is required to analyze the signal from the labeling dye since only a single dye is used to detect the sequencing fragments produced following chain extension. However, since no color discrimination is implemented, the electrophoresis must be run in four lanes, one for each
base similar to the format used in traditional autoradiographic detection. While this is a reasonable approach for slab gel separations, it is not a viable strategy in capillary gel applications due to the poor run-to-run reproducibility in the migration rates of the fragments traveling through the different capillaries. This is due to differences in the gel from capillary to capillary as well as differences in the integrity of the wall coatings used to suppress the electroosmotic flow. In the slab gel format, reproducibility in the migration times becomes less of a problem since all of the lanes are run in the same gel matrix.

Figure 1.29. Sequence of one-dye/four lane format of a M13mp18 template with a M13(-29) primer. Each terminal base is represented by a different color for sequence reconstruction. Automated sequencer run at-1600 V with 1 µl of sample per lane.
The output of a typical, single color/four lane sequencing device is shown in Figure 1.29 along with the called bases. In this example, each different color trace represents an electropherogram from an individual lane of the slab gel, which are overlaid to allow reconstruction of the sequence of the template. In this experiment, the device used a single microscope head containing the collection optics, a diode laser, filters and avalanche photodiode to read the fluorescence from the gel [32]. The microscope scanner was rastered over the gel at a rate of ~0.15 cm/s and monitors fluorescence along a single axis of the gel. The gel was approximately 20 cm in width and 42 cm in length and could accommodate 48 separate lanes. The time required to secure this data was 6 hours, with the extended time due primarily to the limited electric field that could be applied to the slab gel.

1.7.2. Single-color/single-lane

In this sequencing approach, only a single fluor is used and as such, a single laser is required to excite the fluorescence and only a single detection channel is needed to process the fluorescence. The advantage of this approach is that instrumentally it is very simple since the hardware required for detection is simple. In addition, since the sequence is reconstructed from a single electrophoresis lane and not four, the throughput can be substantially higher compared to a single fluor/four lane method.

The bases are identified by adjusting the concentration ratio of the terminators used during DNA enzymatic polymerization to alter the intensity of the resulting electrophoretic bands [16, 33-35]. If the concentration of the terminators used during DNA polymerization was 4:2:1:0 (A:C:G:T), a series of fluorescence peaks would be generated following electrophoretic sizing with an intensity ratio of 4:2:1:0 and the
identification of the terminal bases would be carried out by categorizing the peaks according to their heights. In order to accomplish this with some degree of accuracy in the base calling, the ability of the DNA polymerase enzyme to incorporate the terminators must be nearly uniform. This can be achieved using a special DNA polymerase, which in this case is a modified T7 DNA polymerase [36, 37]. This enzyme has been modified so as to remove its proof reading capabilities by eliminating its 3'→5' exonuclease activity. Since this method requires uniform incorporation of the terminators, it is restricted to the use of dye-primer chemistry. In addition, since the T7 enzyme is not a thermostable enzyme, cycle sequencing cannot be used.

Williams and Soper accumulated data using this base-calling strategy and the terminator concentration was adjusted at a concentration ratio of 4:2:1:0 (A:C:G:T) [33]. The accuracy in calling bases was estimated to be 84% up to 250 bases from the primer annealing site, with readable bases up to 400 but the accuracy deteriorating to 60%. This is a common artifact in this approach, that being poor base calling accuracy. The poor base calling results from variations in the activity of the T7 DNA polymerase and in addition, the null signal used to identify Ts. Ambiguities are present when one must identify multiple null signals, which can lead to insertions or deletions.

1.7.3. Two-color/single-lane

In order to improve on the base calling accuracy associated with the single color/single lane strategy without having to increase the instrumental complexity of the fluorescence readout device associated with sequencing instruments, one may employ a two color format to identify the four terminal bases in sequencing applications. In this approach, one or two lasers are used to excite one of two spectrally distinct dyes used for
labeling the sequencing primers and the fluorescence is processed on one of two detection channels consisting of bandpass filters and photon transducers.

A two-color scanning instrument was used to excite the fluorescence of the labeling dyes, FAM and JOE. Due to their similar absorption maxima, a single laser (488 nm) could efficiently excite the fluorophores. In addition, this dye-pair was selected since they produced sequencing fragments that co-migrated and thus, no mobility correction was required. The bases were identified using a binary coding scheme, which is shown in Table 1.1 [24]. During DNA polymerization, four separate reactions were run with the A-reaction containing an equimolar mixture of the FAM and JOE labeled sequencing primers. In the case of G, only the JOE labeled primer was present, while for T, only the FAM labeled primer was present and for C, no dye labeled primer was used. By ratioing the signal in the red (JOE) to green (FAM) channel, a value was obtained which could be used to identify the terminal base. The attractive feature associated with this protocol is that while the absolute intensity of the bands present in the electropherogram may vary by a factor of 20 due to sequence dependent termination, the ratio only varies by a factor of 1.7. The read length using this binary coding system was approximately 350 bases, with the number of errors ~15 (accuracy = 95.7%). The majority of the errors were attributed to C determinations, since a null signal was used to indicate the presence of this base.

In order to alleviate the errors in the base calling associated with identifying bases using a null signal, a two-dye, two-level approach can be implemented [38, 39]. In this method, the bases using a common dye-label have the concentration of the ddNTPs adjusted during chain extension to alter the intensity of the fluorescence peaks developed.
Table 1.1. Binary coding scheme for two-color DNA sequencing. The (1) indicates the presence of the dye-labeled primer during DNA polymerization and the (0) represents the absence of the dye-label [24].

<table>
<thead>
<tr>
<th></th>
<th>F A M</th>
<th>J O E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

during the electrophoresis. Also required in this approach are uniform peak heights, requiring the use of the modified T7 DNA polymerase in the presence of manganese ions. For example, Chen et al. used a FAM-labeled primer for marking Ts and Gs, with the concentration ratio of the ddNTPs adjusted to 2:1 (T:G). Likewise, the As and Cs were identified using a 2:1 concentration ratio of the terminators with the labeling dye in this case being TAMRA. Sequencing data produced an effective read length of 350 bases, with an accuracy of 97.5%. When the concentration ratios of the terminators sharing a common labeling dye was increased to 3:1, the read length was extended to ~400 bases with a base-calling accuracy >97%. As can be seen, the elimination of null signals to identify bases can improve the base calling accuracy in these types of sequence determinations.

While most fluorescence labeling strategies for DNA sequencing, which depend on differences in intensities of the electrophoretic peaks to identify bases use dye-labeled primers, internal labeling, where the fluorescent dye situated on the dNTP, can also be used [40]. The advantages associated with using dye-labeled dNTPs are the ability to use a wide range of primers since no dye-labeled primer is required, the incorporation of the
dye-labeled dNTP can be much more uniform than dye-labeled ddNTPs, and the dye-labeled dNTPs are much less expensive compared to dye-labeled primers or terminators. Using a tetramethylrhodamine-labeled dATP and a fluorescein labeled dATP, a two-color/single-lane sequencing assay has been reported [40]. For internal labeling, a two step polymerization reaction was used in which the template was annealed to the sequencing primer (unlabeled) along with the dye-dATP and the four unlabeled dNTPs as well as the polymerase enzyme. The extension reaction was incubated at 37°C for 10 minutes after which the appropriate terminator was added and the reaction allowed to proceed for an additional 10 minutes. The initial extension reaction extended 6-8 nucleotides to a quartet of As, with 80-90% of the fragments containing a single dye-labeled dATP. Since only two dyes were used in this particular example, the concentration ratio for a pair of terminators sharing a common dye was adjusted (3: 1) so as to allow discrimination based upon the intensity of the resulting electrophoretic peaks. Analysis of the sequencing data indicated that the read length was found to be 500 bases with an accuracy of 97%.

1.7.4. Four-color/single-lane

The commonly used approach in most commercial DNA sequencing instruments using fluorescence detection is a four-color/single-lane strategy for identifying the terminal bases. The primary reasons for using a four-color/single-lane approach is that it provides high accuracy in the base calling, especially for long reads and also, the throughput can be high due to the fact that all bases comprising the template DNA can be called in a single gel tract. Unfortunately, a four-color detector requires extensive optical components to sort the fluorescence and also, in some cases, multiple excitation sources.
are needed in order to efficiently excite the fluorophores used to label the individual sequencing ladders. In addition, post-electrophoresis software corrections may be required to account for spectral leakage into detection channels.

Most dye-terminator reads are used with this four color strategy, since the data analysis (base calling) does not depend on uniform incorporation efficiencies, which are hard to achieve using dye-labeled terminators. The same type of instrumentation that is used for four-color/dye-primer reads can also be used for four-color/dye-terminator reads as well. The only difference one finds is in terms of the sample preparation protocols and also, the software corrections in the sequencing data, such as different mobility correction factors. In most cases, a size exclusion step is used following DNA polymerization to remove excess dye-labeled terminators, since they are negatively charged and can mask the sequencing data due to the presence of a large dye-terminator band in the gel.

An example of a four-color detector for capillary gel electrophoresis is shown in Figure 1.19, in which two laser sources (Ar ion laser, 488 nm and green helium-neon laser, 543 nm) were used to excite the dye set, FAM, JOE, TAMRA and ROX. In order to process the emission on a single detection channel, a four-stage filter wheel was used, which was synchronized to a sector wheel situated in front of the two lasers. The synchronization was set so as to pass 488 nm excitation for FAM and JOE and simultaneously place the band pass filters for FAM and JOE in the optical path. Following this, the 543 nm laser light was passed and the filters for TAMRA and ROX were situated within the optical path. A read length exceeding 550 bases was obtained at an accuracy of 97%.
1.8. References


Chapter 2

Time-Correlated Single Photon Counting

2.1. Fluorescence

Fluorescence emission involves the absorption of a photon of light and eventual depopulation of the excited state by release of a photon illustrated by the Jablonski diagram (see Figure 2.1). Few molecules stay in the excited vibrational level for an extended period of time and most rapidly relax to the lowest vibrational level of $S_1$. This process is called internal conversion and generally occurs in $10^{-12}$ s. Since fluorescence lifetimes are typically near $10^{-8}$ s, internal conversion is usually completed before emission. Therefore, fluorescence emission normally occurs from a thermally equilibrated electronic excited state [1]. Fluorescence is observed as the electronically excited molecule returns to the ground state and emits a photon.

Molecules in the excited $S_1$ state can also undergo a conversion to the triplet state ($T_1$), called intersystem crossing. Radiative emission from $T_1$ is called phosphorescence, and is generally shifted to longer wavelengths compared to the fluorescence and has a longer time duration. Other processes can also affect fluorescence emission, such as solvent effects by competing with the fluorescence emission.

The fluorescence quantum yield is the ratio of the number of photons emitted to the number of photons absorbed. The emissive rate of the fluorophore ($\Gamma$) and the rate of radiationless decay to $S_0$ ($k_{nr}$) both depopulate the excited state. The fraction of fluorophores that decay by emission, defined by the quantum yield, is given by:
Figure 2.1. Jablonski diagram. Block energy diagram showing the various radiative and non-radiative processes that occur to a molecule upon the absorption of a photon of light.

\[ Q = \frac{\Gamma}{\Gamma + k_{nr}} \]  \hspace{1cm} (2.1)

The lifetime of the excited state represents the average time the molecule stays in the excited state prior to return to the ground state. The expression for the time-dependent fluorescence emission \((I(t))\) can be described as [1, 2];

\[ I(t) = \sum_{i=1}^{n} A_i e^{-t/\tau_f(i)} \]  \hspace{1cm} (2.2)

where \(n\) represents the number of components in the decay, \(A\) is a preexponential factor, \(t\) is the time, and \(\tau_f\) is the fluorescence lifetime. Because the fluorescence lifetime is
proportional to the fluorescence quantum yield, the fluorescence lifetime can be determined experimentally by measuring the time it takes the fluorescence intensity to fall to 1/e of its initial value following excitation.

The value of the fluorescence lifetime of a molecule depends on the radiative and nonradiative decay processes that are involved and can be expressed through \([1, 2]\):

\[ \tau_f = \frac{1}{k_r + k_{nr}} = \frac{1}{k_m} \]  \hspace{1cm} (2.3)

where \(k_r\) is the radiative rate constant \((s^{-1})\), \(k_{nr}\) is the nonradiative rate constant \((s^{-1})\), and \(k_m\) is the total decay rate \((s^{-1})\).

### 2.2 Time-Resolved Spectroscopy

Time-resolved spectroscopy involves the excitation of a molecule with a narrow pulse of light and measurement of the time of relaxation to the ground state by monitoring the photons emitted. In time-correlated single photon counting, the time evolution of individual photon events is processed and after a number of photon-events, a histogram is constructed that represents the decay kinetics of the excited state. The resolution and the precision of the lifetime determination is determined by the width of the excitation pulse, the reproducibility of the pulse, and the instrumental response of the photodetector.

A typical time-correlated single photon-counting (TCSPC) instrument, shown in Figure 2.2, consists of a pulsed light source, which generates an excitation pulse train that stimulates absorption of the sample molecule. At low rates of excitation, each sample molecule absorbs at most one photon, on a time scale that is effectively instantaneous. The subsequent relaxation of the molecules from the excited state to the ground state via emission of fluorescence photons occurs with a distribution of time
delays. The single photon counting technique records this distribution by measuring the time delays of the individual fluorescence photons with respect to the arrival time of the excitation pulse. The delay time between the excitation pulse and fluorescence photon from the sample reaching the detector, typically a photomultiplier tube (PMT) or a single photon avalanche diode (SPAD), is measured by a time-to-amplitude converter (TAC). The excitation pulse serves as a “start” signal for charging a capacitor within the TAC. The voltage on the capacitor increases linearly until a “stop” fluorescence photon is detected. The “start”-“stop” time interval generates a proportional voltage across this capacitor. This voltage is stored and converted to a digital signal using an analog-to-digital converter (ADC) and stored within a multichannel analyzer (MCA) and placed into the appropriate time bin. After repeating the “start”-“stop” cycle many times, a histogram is formed that represents the fluorescence decay.

Discriminators are used to aid in removing random noise pulses resulting from background photons and to ensure that the timing definition of the “start” and “stop” pulses is largely independent of the signal pulse height.

TAC operation requires the ability to register the first “stop” pulse detected after a “start” pulse. Accordingly, the “stop” pulse rate must be low enough so that the probability of detecting more than one photon is negligible. In effect, the “start” pulse rate must exceed the “stop” pulse rate. Failure to adhere to this means that the TAC will preferentially detect photons which occur at shorter times, and the fluorescence decay time distribution will be biased, resulting in decay times that will appear to be shorter than they really are. This photon pileup effect makes the single photon-counting technique inefficient when the fluorescence signal is intense due to the fact that the
available signal has to be wasted in order to ensure the absence of photon pileup.

The time-resolved mode of measuring fluorescence lifetime is a digital or photon-counting method that shows good signal-to-noise. Advantages of TCSPC include a high dynamic range and the measured decay profile is independent of fluctuations in the excitation pulse intensity. Additionally, single photon detection theory is based upon well-documented statistics for which the precision, data weights, goodness-of-fit, and other parameters can be easily calculated.
If, following one excitation pulse, the recording of a coincident event in channel $i$ corresponds to an average number of photons, $\bar{n}_i$, reaching the photodetector and liberating an average number of $\bar{x}_i$ photoelectrons, then [1, 2];

$$\bar{x}_i = \bar{n}_i q \tag{2.4}$$

where $q$ is the photodetector quantum efficiency. If there is a large number of excitation pulses for every count registered in channel $i$, the probability, $P_x(i)$, of liberating other than $\bar{x}_i$ photoelectrons is given by the Poisson distribution [1, 2];

$$P_x(i) = \frac{(\bar{x}_i)^x e^{-\bar{x}_i}}{x!} \tag{2.5}$$

with

$$\sum_{x=0}^{\infty} P_x(i) = 1. \tag{2.6}$$

The probability of at least one photoelectron pulse detected per excitation pulse can be determined via [1, 2];

$$P_{x\geq1}(i) = 1 - e^{-\bar{x}_i} = (1 - \bar{x}_i + \frac{\bar{x}_i^2}{2} + ...) \tag{2.7}$$

when, at low excitation powers, $\bar{x}_i >> \bar{x}_i^2 / 2$, which gives the following relationship;

$$P_{x\geq2}(i) \approx \bar{x}_i \approx \bar{n}_i q. \tag{2.8}$$

Therefore, under these conditions the probability of detecting a coincident fluorescence photon in channel $i$ is proportional to the fluorescence intensity at a delay time $t$. This is the general condition for accurate single photon counting measurements. Moreover, the probability of detecting two fluorescence photons per excitation pulse is further diminished provided that;
\[ \frac{S_p}{S_t} \leq 0.01 \] (2.9)

where \( S_p \) is the photodetector pulse due to the fluorescence photons over all delay times and \( S_t \) is the repetition rate of the excitation source. For pulsed excitation, the number of counts \( Y_i \), accumulated in channel \( i \) after excitation in a specified measurement time \( T \) for a single exponential decay time \( \tau_f \), can be determined by;

\[ Y_i = \alpha S_T \frac{\Delta t}{\tau_f} e^{-\frac{i \Delta t}{\tau_f}} \] (2.10)

where \( \Delta t \) is the time difference between the detected photon and the excitation pulse. Equation (2.10) provides another inherent advantage of the single photon-counting method, the measurement precision can be enhanced by increasing the run time of the experiment, thus allowing more photons to accumulate within the decay profile.

2.3. TCSPC Instrumentation

The main criteria for determining the overall system performance of the TCSPC system is the timing resolution and the sensitivity of the device. These characteristics vary from system to system and are based largely upon the performance of the individual components of the device, with the choice of detector dictating the overall response of the system in most cases. The TCSPC system usually includes; a pulsed light source, a photodetector, and the counting electronics, including the CFD, TAC, and the multichannel analyzer with an ADC.

2.3.1. Light Sources

There are basically three choices for pulsed excitation in TCSPC measurements. The choices are flashlamps, synchrotron sources, or mode-locked lasers. Flashlamps
and mode-locked lasers are the most commonly used. Since the experiments in this document utilize a mode-locked laser as the excitation source, a detailed description of these lasers is given here.

Laser is an acronym for Light Amplification by the Simulated Emission of Radiation, and is an optical oscillator that creates a very highly directed beam of light at a very precise wavelength or frequency. There are three important components associated with all lasers, which include the high reflector, the gain medium, and the output coupler.

For lasing to occur, the light passing through the gain medium must be amplified. The high reflector at one end of the laser and the output coupler define the laser cavity, in which the amplified light will return through the gain medium for further amplification. The output of the laser occurs when a fraction of the light is transmitted through the output coupler.

In order for lasing to occur, there are three types of energy exchanges needed: absorption, spontaneous emission and stimulated emission [2]. All three rely on the transitions from one energy level to another within the gain medium.

Absorption of a photon involves the promotion of the molecule of the gain medium from the ground state to the excited state. Upon excitation, the molecule can relax back to the ground state via the release of a photon by either spontaneous or stimulated emission. Spontaneous emission is required for initiating lasing, but this process does not have the directional properties of stimulated emission and therefore represents a loss in the laser cavity. In stimulated emission, an incident photon interacts with a molecule in the excited state and induces the emission of a photon with a
transition to the ground state, therefore the photon has an energy equal to the difference of the excited and ground state. Stimulated emission is required for lasing to occur.

The photons generated by stimulated emission have two important properties; the first is that photons have the same direction as the incident radiation and the second is that the photons produced by stimulated emission are in phase with the excitation photon, which is responsible for the coherent nature of the laser light.

It was mentioned that the formation of a pulse for TCSPC by a laser was produced by a process known as mode-locking. Laser cavities can support numerous longitudinal modes, each at different frequencies. For a mode-locked laser, a periodic loss is introduced into the cavity by some modulator. This modulator acts as a shutter, opening once every round trip for the light in the resonator. When the modulator is fully open, the light can pass through and aligns the phases of the different longitudinal modes with one another. When the amplitudes align, constructive interference reinforces the intensity of the pulse. Between the pulses, destructive interference eliminates the light. There are two ways to mode-lock a laser, either by passive or active mode-locking.

Passive mode-locking involves the addition of some material or mechanism within the laser cavity that automatically opens to allow light pulses through and subsequently closes. This type of mode-locking is usually done by using a dye whose absorption decreases with increasing irradiance to alternately block and admit light into the laser cavity.

Active mode-locking involves the addition of an optical shutter within the laser cavity. In order to initiate the pulse, an optical shutter is opened, closed and then
reopened at precisely the same rate to allow a pulse of light to pass through the shutter as it passes back and forth between the output coupler and high reflector. Only light that arrives at precisely the correct time is allowed to pass through without being blocked and thus will be amplified. In other words, the modulator frequency must be precisely equal to the reciprocal of the round trip time in the laser cavity.

The excitation source used in the TCSPC measurements reported in this document were produced by a passively mode-locked Ti:sapphire laser pumped by the all lines output of an Argon ion laser and an actively mode-locked Gallium-Aluminum-Arsenide (GaAlAs) semiconductor laser.

2.3.1.1. Ti:sapphire Laser

The Ti:sapphire laser (Coherent Model 900 Mira) pumped by an Argon ion laser (Coherent Innova 300) is a mode-locked ultrafast laser system that uses a Ti:sapphire crystal (Ti:SiO₃) as the gain medium for the near-IR and is tunable from 720 to 990 nm. The repetition rate of the laser is 76 MHz, producing a round trip transit time of light pulses within the cavity of 13.2 ns [3].

For the Ti:sapphire laser system to operate in TCSPC experiments, short light pulses must be generated. The Mira Ti:sapphire accomplishes this through passive mode-locking, in which the optical properties of light are used to form narrow mode-locked pulses.

The Mira cavity has been designed so that the beam diameter within the cavity changes by a small amount as the intensity of the light changes. More specifically, the beam diameter at certain locations within the cavity is large when the laser is operating continuously (CW) but becomes smaller when the laser is producing high intensity
Figure 2.3. Beam geometry in the Mira tunable-laser cavity as the light passes through the Ti:Sapphire crystal.

A simple slot or "slit" is now placed at the appropriate location and its width is adjusted so that the large diameter laser beam associated with continuous operation will be interrupted at its edges (Figure 2.4). A high intensity pulse, however, will pass uninterrupted through the slit, since the beam is smaller.

We refer to the mode-locking device in Mira as a saturable absorber system, since in reality it consists of two parts. Firstly, there is a material that decreases the laser beam size in the presence of high intensity pulses. Secondly, a slit which introduces losses for large beams.

The phenomenon utilized to change the beam diameter is quite interesting. The
properties of light passing through any material depend, among other things on a property referred to as the index of refraction, or \( n \). The higher \( n \), the lower the velocity of light. If the velocity of light is different for different parts of the light beam, the beam will bend or otherwise be reshaped, since different parts of the beam are travelling at different speeds. This is known as refraction. A common refractive element is a lens (e.g. a biconvex lens), which is thicker in the middle than at the edges, so that the center of the beam is slowed down more than the edges. This causes the light to bend toward the center. In the case of the lens, the index is the same everywhere, but still, since there is more glass in the middle than the edges, the edges are slowed down less. Making the

**Figure 2.4.** Part of Mira’s saturable absorber system: slit and beam cross section.
index of refraction at the center of the material larger than the index at the edges can also form a lens. This will also bend light and is referred to as a gradient index lens.

The most common way to change the index of a material is to change its chemical composition. In the Mira laser, the index is changed by the light itself. At sufficiently high intensity, the electric fields associated with the light can actually distort the atoms of the material and alter its index. This effect is known as the optical Kerr effect since the beam is less intense at its edges as compared to the center, the index at the center will be different, and a gradient index lens is formed. Since it is the optical Kerr effect which alters the index, the lens thus formed is referred to as a Kerr lens. The Kerr lens is formed only when the intensity of the light is extremely high. The instantaneous intensity of mode-locked light pulses is sufficient to form this lens, but the weak intensity of the laser, which is operating CW is not. Hence, the lens is only formed upon the arrival of a mode-locked pulse. It is this lens which narrows the laser beam and, consequently, a mechanism has now been created which narrows the beam only for mode-locked pulses. The addition of a slit to allow only narrow beams to pass unattenuated now forms the complete saturable absorber system, which provides a real driving force for mode-locking. The Ti:sapphire has a bandwidth of 11 nm and the pulse width is 150 fs (full width at half maximum, FWHM).

2.3.1.2. Semiconductor Laser

Semiconductor diode lasers offer the advantages of low cost, relatively high powers, stable output, and long lifetimes, usually more than 40,000 hours. These lasers have also been shown to be adequate excitation sources for TCSPC measurements [4].
Imasaki and coworkers were the first to implement diode lasers for single photon-counting studies of polymethine dyes [5].

Diode lasers use a semiconductor material as the gain medium for light production. A common semiconductor material being GaAlAs. The light production from semiconductor lasers occurs when electron-hole pairs recombine across the band gap of the semiconductor with the energy of the emitted photon ($E_g$) dictated by the band gap.

The electronic band structure of semiconductors determines the optical properties of the material. The energy diagrams of semiconductor materials predict that an electron can be promoted from an electron rich valence band to the electron depleted conduction band by absorbing a photon of at least the band gap energy ($h\nu > E_g$), leaving behind a hole in the valence band [6]. Each photon absorbed creates one electron-hole pair.

Light is emitted from semiconductors when these electron-hole pairs recombine across the band gap. The wavelength of light emitted is determined by $\lambda = hc / E_g$. The use of solid solutions or alloy semiconductors, such as Ga$_{1-x}$Al$_x$As, with the variable band gap energies provides a means for choosing the wavelength of the light emitted. The recombination of electron-hole pairs explains simple emission from a semiconductor, but it does not provide a means of obtaining population inversion needed for laser action. Semiconductor junctions are needed to confine the spatial distribution of the electron-hole pairs to obtain the necessary population inversion.

When a junction is formed between a n-type and p-type semiconductor, electrons flow from the electron rich n-type semiconductor to the p-type semiconductor.
until equilibrium is reached between the two materials. The transfer of the charge at the junction leaves a small area of net positive charge at the n-type side and a small region of negative charge at the p-type side of the junction.

The application of an external voltage source, which is negative relative to the voltage applied to the p-type side induces a current called forward biasing, in the semiconductor and subsequently disrupts the equilibrium condition. Thus, electrons must be provided with enough energy to move over the potential energy barrier from the n-type side to the p-type side, while the holes must be forced to move under the barrier from the p-type side to the n-type side during current flow in order for light production.

Under the influence of a forward bias voltage, electrons and holes recombine in the junction emitting light in the process. The wavelength of the light is dictated by the band gap energy, which is determined by the composition of the semiconductor material. A p-n junction used in this manner is called a light emitting diode and emits continuous wave light. Stimulated emission is not involved in this process; therefore lasing does not occur.

In order for lasing to occur, the addition of a laser cavity for amplification and a means of producing a population inversion are required. The diode laser cavity is typically constructed by having at least one pair of opposite faces that are flat and parallel to one another. Parallel faces can be constructed by mechanical cleaving that occurs naturally in a particular crystallographic direction. Due to the change in the index of refraction at the boundary between the crystal and the surrounding air, these parallel surfaces act as mirrors that bound the diode laser cavity. A p-n junction
constructed in this manner can support lasing when the cavity length is equal to a half-integral number of wavelengths.

For optical confinement, semiconductor layers with a lower index of refraction are situated adjacent to the active region of the junction and improve the efficiency of the recombination of electron-hole pairs. The probability of stimulated emission is increased due to the increased confinement of the light through the active region.

Placing materials with higher band gap energies next to the lasing region can improve the current confinement. The higher band gap energy can cause the formation of potential energy barriers between the materials. The electrons located in the lasing region are confined by this potential energy barrier, causing a local buildup of the electron and hole concentrations in the active region, increasing the chances for stimulated emission.

A disadvantage of diode lasers is that the beam-shape is elliptical due to the longitudinal separation between the diode laser’s emission points parallel and perpendicular to the junction. The separation of emission points is caused by a directional dependence on the refractive index of the lasing cavity. The amount of astigmatism varies with the diode lasers of different types as well as those of the same type.

The diode head of the PicoQuant diode laser used in the studies reported in this document was actively mode-locked and driven by an electrical short-pulse generator, which supplied high repetition rate current pulses (80 MHz) with a FWHM of 500 ps, and the FWHM of the output optical pulses was approximately 100 ps.
2.3.2. Detectors

All photodetectors have some degree of jitter or variations in the arrival time of the anode pulse relative to generation of the primary photoelectron event and the rate at which the electron cascades pass through the multiplier. The spread in transit times of arriving photoelectrons depends primarily upon the energy of the primary photoelectron, which is directly proportional to the wavelength of the incident photon, and the coating of the photocathode that is wavelength sensitive.

For TCSPC measurements, a detector is needed that has a low timing dependence on wavelength, a low timing jitter, high amplification, low noise, and a wide spectral range. Ultimately, it is these characteristics of the detector, which dictate the overall timing response of the TCSPC instrument. There are three general classes of detectors available for TCSPC experiments, including photomultiplier tubes, microchannel plate photomultipliers, and avalanche photodiodes. Since avalanche diodes were used in the experiments in this document, we will concentrate on them.

2.3.2.1. Single Photon Avalanche Diodes

Single photon avalanche diodes (SPAD) are semiconductors that function similarly to diode lasers, except that SPADs are run under reverse bias and above their breakdown voltage. Operation above the breakdown voltage is called Geiger operation. At this voltage, there is a strong statistical probability that the detector current will fluctuate to zero in the multiplication region and the avalanche photo diode will remain in the off state until a photo-generated carrier, such as a fluorescence photon, reaches the active layer of the p-n junction to trigger a cascade of electrons in which many electron-hole pairs are formed [7]. The number of electron-hole pairs that are formed
represents the gain of the photodetector, which is determined by the energy of the incident photon’s wavelength. Photons with higher energies traverse farther into the active layer of the semiconductor, therefore more electron-hole pairs are formed (see Figures 2.5 and 2.6). When the first electron-hole pair is formed within the semiconductor material, the leading edge of the avalanche current marks the arrival time of the photon. Once the avalanche is triggered, an output pulse is generated from the detector and sent to the counting electronics.

After the output pulse is formed, the SPAD voltage drops below the breakdown voltage. Following this, the SPAD bias is restored to its initial value. Therefore, the dead time of the detector is determined by the recovery time and typically sets an upper limit on the dynamic range of the detector to photocurrent rates of ~500 KHz for passive quenching.

**Figure 2.5.** Schematic of silicon avalanche photodiode [8].
Figure 2.6. A) Electron energy level diagram illustrating the generation and subsequent separation of an electron-hole pair by photon absorption within the depletion region of a p-n junction. B) Illustration of the principle of operation of an avalanche photodiode. An electron having reached point A has sufficient energy above the conduction band bottom to enable it to excite an electron from the valence band into the conduction band (C→D). In doing so it falls from A to B [7].
The SPAD, like other photodetectors, can be triggered by photons and also by carriers due to thermal effects inside the semiconductor. Theses processes cause a self-triggering of the cascade of the electron-hole pairs within the semiconductor material, which is called the dark count of the detector, $R$. The statistical fluctuations of these events, which compete with photons in triggering the detector, reduce the detector sensitivity. The dark count can also increase if the semiconductor material begins to heat up; therefore, cooling systems are employed in all photodiodes and can reduce the dark count exponentially [9].

2.4. TCSPC Electronics

2.4.1. Discriminators

The output of the fluorescence detector consists of a broad distribution of pulse heights, some pulses generated by dark noise, some generated by single-photon events, and some by multiple photon events. It is important to send these pulses through a discriminator in order to improve the signal-to-noise ratio and to furnish the TAC with constant amplitude pulses that are independent of the photodetector pulse shapes and heights. Discriminators provide a timing definition that is independent of pulse height and also discriminate against low-amplitude noise.

There are two types of discriminators: leading edge (LED) and constant fraction discriminators (CFD). The CFD is particularly suited to single photon timing studies compared to LED, since with the LED, photodetector pulses of different amplitudes will not cross the discriminator level at the same time, introducing error into the measurement (see Figure 2.7).
Figure 2.7. Timing errors associated with leading edge discriminators. Both pulses are initiated at the same time, however their arrival time is different due to the threshold level and amplitude of the pulse.

In order to eliminate the timing errors in single photon counting experiments discrimination using a constant fraction approach is employed. With constant fraction discrimination, the pulses are timed from a point on the leading edge that is a fixed fraction of the pulse height. For pulses of similar shape but different amplitudes, the timing point is constant, Figure 2.8.

The discriminator level for the excitation pulses is usually set to reject dark noise caused by signal pulses. The setting of the discriminator level for the fluorescence pulses is much more critical. Setting the discriminator level too low allows for the accumulation of background events in the fluorescence decay. Setting the discriminator level too high allows for multiphoton events to be counted. Fluorescence photodetector pulses have a wide pulse height distribution, therefore there is no clear dividing line between dark noise and pulses arising from fluorescence photons. Increasing the discriminator threshold seems to increase the signal-to-noise ratio. Consequently, the chosen discriminator level will be a compromise between acceptance of relatively more single photon pulses and rejection of dark noise.
2.4.2. Time-to-Amplitude Converter (TAC)

The TAC functions to determine the time interval between the excitation pulse and the subsequent fluorescence photon arriving at the detector. Upon receipt of a "start" pulse, and after a certain fixed delay, a timing capacitor is charged linearly from a constant current source. The charge on the capacitor is discontinued upon arrival of a "stop" pulse and an output pulse is generated with an amplitude proportional to the time between "start" and "stop" pulses. If no "stop" pulse is received after a certain time called the TAC range, charging is automatically stopped. Again, a fixed time elapses before the capacitor is reset, at which time the instrument is ready to accept another "start" pulse.

2.4.3. Multichannel Analyzer

The basic multichannel analyzer (MCA) consists of an analog to digital converter (ADC), a memory comprised of channels for storing data, and data input and output facilities. A standard instrument generally incorporates lower and upper discriminator levels and two modes of data collection: pulse height analysis mode for
the display of fluorescence decay profiles and multichannel scaling mode that bins the data into certain time increments. The data is usually displayed on a computer terminal or on an oscilloscope.

For pulse height analysis experiments, between 200 and 600 channels are sufficient for a decay curve; however, increasing the number of channels in the decay and subsequent reduction in the time per channel improves the timing resolution of the instrument. A drawback is the increased data accumulation time needed for an increased number of channels.

2.5. Data Analysis

Probably the most important aspect associated with lifetime determinations is consideration to the processing or calculation algorithms used to extract the lifetime value from the resulting decay. The accuracy in the base call depends directly on the lifetime differences between the fluors in the dye set and the relative precision in the measurement. Algorithms that deal with these parameters are required as well as those that can be performed on-line during the electrophoresis.

2.5.1. Non-linear Least Squares

The goal of NLLS is to obtain estimates of the parameters describing the decay and having the highest probability of being correct based on a starting model. The parameters which provide the best match between the data \( N(t_k) \) and the calculated decay \( N_c(t_k) \) are accepted based on the minimization of the goodness of fit parameter \( \chi^2 \), where \( \chi^2 \) is calculated from [10]:

\[
\chi^2 = \sum_{k=1}^{n} \frac{1}{\sigma_k^2} [N(t_k) - N_c(t_k)]^2 = \sum_{k=1}^{n} \left( \frac{N(t_k) - N_c(t_k)}{N(t_k)} \right)^2
\]  

(2.11)
The sum extends over the number \( n \) of channels or data points used in the analysis and \( \sigma_k \) is the standard deviation associated with each data point. In most NLLS analyses, the measured data are compared with values predicted from a model and the parameters of the model are varied to yield the minimum deviations from the data. Two advantages of NLLS analysis includes its ability to remove the contribution of the IRF through deconvolution, increasing the accuracy in the determination, and the ability to analyze multi-exponential decays.

### 2.5.2. Rapid Determination Method (RLD)

Fluorescence lifetimes calculated using the RLD method are performed by integrating the number of counts within the decay profile over a specified time interval and using the following relationship [11];

\[
\tau_f = -\frac{\Delta t}{\ln\left(\frac{D_1}{D_0}\right)}
\]

(2.12)

where \( \Delta t \) is the time range over which the counts are integrated and \( D_0 \) is the integrated counts in the early time interval of the decay spectrum, while \( D_1 \) represents the integrated number of counts in the later time interval. The RLD method can extract only a single lifetime value from the decay, which in cases of multi-exponential profiles would represent a weighted average of the various components comprising the decay.

### 2.5.3. Maximum Likelihood Estimator (MLE)

Another algorithm for on-the-fly fluorescence lifetime determinations has been used, namely the maximum likelihood estimator. The MLE calculates the lifetime via the following relation [12, 13];
\[ 1 + \left( e^{T/\tau} - 1 \right)^{-1} - m \left( e^{mnT/\tau} - 1 \right)^{-1} = N^{-1} \sum_{i=1}^{m} N_i, \] (2.13)

where \( T \) is the width of each channel, \( m \) the number of time channels, \( N \) is the total number of photon counts, \( N_i \) is the number of photon counts in time channel \( i \), and \( \tau \) is

**Figure 2.9.** Graph of monotonic, non-decreasing function of two variables (\( m, \tau \)) for determining lifetime values of experimental data using the MLE. The curve was plotted in Surface Graphics from the equation:

\[ 1 + \left( e^{T/\tau} - 1 \right)^{-1} - m \left( e^{mnT/\tau} - 1 \right)^{-1} \]

over \{m, 1, 4095\}, \{t, 300, 2000\}. 

90
the calculated lifetime. The right hand side of the equation is calculated from experimental values and the left-hand side describes a monotonic, non-decreasing function of two variables \((m, \tau)\) (see Figure 2.9). For the chosen number of time bins, \(m\), a binary search for the calculation of \(\tau\) in the left-hand side of the equation is performed. The binary search for \(\tau\) begins by selecting a region of possible \(\tau\)s as wide as 10 to 40,000 ps. The left-hand side of the equation is calculated at the ends of the region as well as in the middle of the region. Depending on the relation to the right hand side value, the appropriate half of the region is chosen and the calculations are repeated. In each step, the region of possible values of \(\tau\) satisfying the equation is narrowed by a factor of two. Computation continues until the search region of lifetimes is less than a predefined value, typically ~0.1 ps.

### 2.6. Single-color/four-lifetime Sequencing

While most sequencing applications using fluorescence require spectral discrimination to identify the terminal base during electrophoretic sizing, an alternative approach is to use the fluorescence lifetime of the labeling dye to identify the terminal base. In this method, either time-resolved or phase-resolved techniques can be used to measure the fluorescence lifetime of the labeling dye during the gel electrophoresis separation.

The monitoring and identification of multiple dyes by lifetime discrimination during a gel separation can allow for improved identification efficiency when compared to that of spectral wavelength discrimination since additional information regarding the identity of the dye is present. When the identity of the terminal nucleotide base is accomplished through differences in spectral emission wavelengths, errors in the base
call can arise from broad, overlapping emission profiles, which results in cross talk between detection channels. Lifetime discrimination eliminates the problem of cross talk between detection channels and also, can potentially allow processing of the data on a single readout channel. Several other advantages are associated with fluorescence lifetime identification protocols, including [14]: the calculated lifetime is immune to concentration differences, the fluorescence lifetime can be determined with higher precision than fluorescence intensities, and only one excitation source is required to efficiently excite the fluorescent probes and only one detection channel needed to process the fluorescence for appropriately selected dye sets.

One potential difficulty associated with this approach is the poor photon statistics (limited number of photocounts), which can result when making such a measurement. This results from the need to make a dynamic measurement (the chromophore is resident in the excitation beam for only 1-5 s) and the low mass loading levels associated with many DNA electrophoresis formats. Basically, the low number of photocounts acquired to construct the decay profile from which the lifetime is extracted can produce low precision in the measurement, which would affect the accuracy in the base call. In addition, the high scattering medium in which the fluorescence is measured (polyacrylamide gel) can produce large backgrounds again lowering the precision in the measurement. An additional concern with lifetime measurements for calling bases in DNA sequencing applications is the heavy demand on the instrumentation required for such a measurement. However, the increased availability of pulsed diode lasers and simple avalanche photodiode detectors has had a
tremendous impact on the ability to assemble a time-resolved instrument appropriate a sequencing applications.

There are two different formats for measuring fluorescence lifetimes, time-resolved [4, 14-19] or frequency-resolved [20-24]. Since the time-resolved mode is a digital (photon counting) method, it typically shows better signal-to-noise than a frequency-resolved measurement making it more attractive for separation platforms that deal with minute amounts of sample. In addition, the use of time-resolved methods allow for the use of time-gated detection in which background photons, which are coincident with the laser pulse (scattered photons) can be gated out electronically improving the signal-to-noise ratio in the measurement.

A device that has been used for making time-resolved measurements during capillary gel electrophoresis consisted of an actively pulsed solid-state GaAlAs diode laser with a repetition rate of 80 MHz and an average power of 5.0 mW at a lasing wavelength of 780 nm [4]. The pulse width of the laser was determined to be ~50 ps (FWHM). The detector selected for this instrument was a single photon avalanche diode (SPAD), which has an active area of 150 µm and is actively cooled. In addition, the SPAD has a high single photon detection efficiency (>60% above 700 nm). The counting electronics (constant fraction discriminator (CFD), analog-to-digital converter (ADC), time-to-amplitude converter (TAC), and pulse-height analyzer) were situated on a single TCSPC board. The board plugged directly into a PC-bus and exhibited a dead time of <260 ns, allowing efficient processing of single photon events at counting rates exceeding 2 x 10⁶ counts/s. This set of electronics allowed for the collection of 128 sequential decay profiles with a timing resolution of 9.77 ps per channel. The
instrument possessed a response function of approximately 275 ps (FWHM), adequate for measuring fluorescence lifetimes in the sub-nanosecond regime.

To demonstrate the feasibility of acquiring lifetimes on-the-fly during the capillary gel electrophoresis (CGE) separation of sequencing ladders, C-terminated fragments produced from Sanger chain-terminating protocols and labeled with a near-IR fluorophore on the 5' end of a sequencing primer were electrophoresed and the lifetimes of various components within the electropherogram determined [14]. The average lifetime determined using the MLE method was found to be 843 ps, with a standard deviation of ±9 ps (RSD=1.9%). The lifetime values calculated here compared favorably to a static measurement performed on the same dye.

Since the base calling is done with lifetime discrimination as opposed to wavelength discrimination, new type of dye sets can and need to be used that suit the identification method. For example, it is not necessary to use dyes with discrete emission maxima and as such, structural variations in the dye set can be relaxed. A dye set developed for lifetime discrimination has been prepared and consists of a near-IR chromophore, which have unique fluorescence lifetimes with the lifetime altered via the addition of an intramolecular heavy atom (see Figure 2.10) [25]. Each of these dyes possesses the same absorbance maximum and fluorescence emission maximum (see Figure 2.11), but different fluorescence lifetimes. The lifetimes of these dyes were found to be < 1.0 ns, with the lifetimes for the dye set ranging from 947 ps to 843 ps when measured in a polyacrylamide gel solution containing urea.

The attractive feature associated with fluorescence in the near-IR ($\lambda_{em} > 700$ nm) includes smaller backgrounds observed during signal collection and the rather
simple instrumentation required for carrying out ultra-sensitive detection. In most cases, the limit of detection for fluorescence measurements is determined primarily by the magnitude of the background produced from scattering and/or impurity fluorescence. This is particularly true in DNA sequencing since detection occurs within the gel matrix, which can be a significant contributor of scattering photons. In addition, the use of denaturants in the gel matrix, such as urea (7 M) or formamide, can produce large amounts of background fluorescence. The lower background that is typically observed in the near-IR can be attributed to the fact that few species fluoresce in the near-IR. In addition, the $1/\lambda^4$ dependence of the Raman cross section also provides a lower scattering contribution at these longer excitation wavelengths.

Figure 2.10. Structure of near-IR heavy-atom modified tricarbocyanine dyes [25].
Figure 2.11. Absorbance and emission profiles of heavy-atom modified tricarbocyanine dye-primers measured in a 6% T5% C non-polymerized acrylamide solution containing 1 x TBE and 40% formamide [25].

An added advantage of near-IR fluorescence is the fact that the instrumentation required for detection can be rather simple and easy to use. A typical near-IR fluorescence detection apparatus can consist of an inexpensive diode laser and single photon avalanche diode (SPAD). These components are solid-state allowing the detector to be run for extended periods of time requiring little maintenance or operator expertise.

Near-IR fluorescence can be a very attractive detection strategy in gel sequencing because of the highly scattering medium that the separation must be performed in. Due to the intrinsically lower backgrounds that are expected in the
near-IR compared to the visible, on-column detection can be performed without sacrificing detection sensitivity. To highlight the intrinsic advantages associated with the use of near-IR fluorescence detection in capillary gel DNA sequencing applications, a direct comparison between laser-induced fluorescence detection at 488 nm excitation and 780 nm excitation has been reported [26]. In this study, a sequencing primer labeled with FAM or a near-IR dye was electrophoresed in a capillary gel column and the detection limits calculated for both systems. The results indicated that the limits of detection for the near-IR case were found to be $3.4 \times 10^{-20}$ moles, while for 488 nm excitation, the limit of detection was $1.5 \times 10^{-18}$ moles. The improvement in the limit of detection for the near-IR case was observed in spite of the fact that the fluorescence quantum yield associated with the near-IR dye was only 0.07, while the quantum yield for the FAM dye was ~0.9. The improved detection limit resulted primarily from the significantly lower background observed in the near-IR. Near-IR has also been demonstrated in sequencing applications using slab gel electrophoresis where the detection sensitivity has been reported to be 2000 molecules [27, 28].

2.7. Research Focus

The focus of this work was directed toward the optimization of near-IR labeled sequencing Sanger fragments in slab gel and capillary gel electrophoresis to yield long (>600 bases) and accurate (>98%) base-calling reads using lifetime identification protocols. Pre-electrophoresis conditions such as cycle sequencing and sample clean-up, were investigated to determine effects on the fluorescence intensity and lifetime of near-IR labeled sequencing fragments. The separation matrix was also investigated to determine separation efficiency for DNA sequencing ladders and any effects the matrix
may have on the lifetime of the dyes. Preliminary characterization and sequencing data utilizing heavy-atom modified dyes attached to dideoxynucleotides was also performed.

2.8. References


Chapter 3

Time-Resolved Fluorescence Imaging of Slab Gels for Lifetime Base-Calling in DNA Sequencing Applications

3.1. Introduction

Fluorescence multiplexing allows the ability to detect several targets in a single assay and typically uses spectral discrimination methods, in which each reporter has unique emission properties. For DNA sequencing applications, four reporters are used, one for each nucleotide base to allow processing of the sequencing data in one electrophoresis lane, thereby increasing throughput. While the fluorescence identification of terminal bases in sequencing applications has been traditionally achieved using spectral discrimination and automated sequencers that possess four different color channels, lifetime discrimination can potentially offer increased multiplexing capabilities by utilizing color discrimination in conjunction with lifetime discrimination.

Several groups have been investigating the potential use of fluorescence lifetime discrimination techniques for identifying nucleotide bases in DNA sequencing [1-9]. Unfortunately, many of these groups using time-resolved fluorescence detection for DNA sequencing have implemented only a single separation lane of the gel. For high-throughput applications, it is necessary to perform the fluorescence measurement on many separation lanes whether in a slab gel or multiple capillaries in an array machine. Thus, if lifetime methods are to become a viable addition to spectral discrimination techniques for multiplexed applications, the readout hardware must be configured to obtain data from multiple separation lanes.
Recently, a scanning-type time-resolved fluorescence microscope that consisted of a pulsed diode laser and a large photoactive area avalanche photodiode operated above its breakdown voltage for single-photon-counting operation was reported [10]. All of the optical components were mounted on a single microscope body that could be scanned over surfaces to obtain lifetime images. However, the avalanche photodiode was passively quenched and possessed a large stray capacitance value resulting in a limited dynamic range (350-71000 photocounts/s) due to the slow recovery time of the diode generated from detection of the primary photoelectron event. Also, this time-resolved scanner was not integrated into an actual DNA sequencing instrument.

A second group described a time-resolved scanner for reading capillary arrays [11]. In their work, a confocal imager configured in a time-correlated single-photon-counting arrangement was kept stationary as the capillary array was linearly translated through the detection zone. Unfortunately, in many applications it is difficult to scan the separation medium and in this case, it is necessary to consider translating the entire detection apparatus and allow the separation platform to remain stationary.

We report on the integration of a near-IR time-resolved fluorescence scanner to an automated DNA sequencer that uses slab gel electrophoresis as the separation platform for lifetime-based identification of terminal nucleotides generated using standard Sanger chain termination methods. Due to the size of the slab gel and gel plates, the medium could not be translated beneath the relay optic of the detector. Therefore, it was necessary to consider a device in which the entire detection optics were moved in a linear fashion across the gel plates. The scanner was constructed from a microscope situated in a Li-COR automated slab gel electrophoresis instrument,
which contained a pulsed diode laser and a large photoactive area avalanche photodiode that was actively quenched to produce a large dynamic range. Also, the electronics for performing time-correlated single-photon counting (TCSPC) were situated on a single PC-based card inserted into a computer for processing the time-resolved data. The performance characteristics of the microscope head will be reported as well as the integration of the scanner to an automated DNA sequencer to obtain time-resolved images of slab gels. Finally, a two-lifetime base call of Sanger sequencing tracts will be demonstrated using this time-resolved near-IR scanner in slab gel electrophoresis.

3.2. Experimental Section

3.2.1. Instrumentation

A schematic of the time-resolved near-IR scanner head is shown in Figure 3.1. The head consisted of a modified microscope taken from a Li-COR 4000 automated DNA sequencer [12]. A PicoQuant GmbH model PDL 800 laser was mounted on the microscope head at an angle of 56° (Brewster’s angle) with respect to the plane of the gel plates to minimize reflected radiation from being coupled into the optical system. The laser was operated at a lasing wavelength of 680 nm and a repetition rate of 80 MHz (pulse width 100 ps, fwhm) with the output beam vertically polarized. The laser diode was powered by an electrical short-pulse generator, which supplied high-repetition-rate picosecond current pulses to the diode head. The driver consisted of a radio frequency pulse generator, a fast switch stage, a coax line driver, and a pulse shaper stage [13].

The laser radiation was focused onto the surface of the gel plates using a f/1.4 lens that produced a spot on the gel of approximately 20 µm x 30 µm (elliptical beam
shape of the diode laser). The emission generated from the gel was collected by f/1.2 optics mounted on the microscope head and filtered with a single band-pass filter (center bandwidth 710 nm, half bandwidth 20 nm). The emission was then focused onto the face of an actively quenched, large photoactive area (i.d. = 500 µm) avalanche photodiode operated in a Geiger mode for photon counting (PicoQuant SPM 200). The detector was also mounted on the microscope body at the secondary image plane of the relay optic set situated in the microscope. The dark count rate of the detector was determined to be 3500 counts/s, and the count rate was linear with the light intensity up
to ~2.5 Mcounts/s (dynamic range). The large photoactive single-photon avalanche diode was needed to minimize the changes in the instrument response function produced by minor defocusing when operated in a scanning mode [10, 14].

The single-photon pulse from the avalanche photodiode was directed to a constant-fraction discriminator (CFD), which shaped the input pulse and subsequently produced a NIM pulse. A window discriminator in the CFD was used to reject background noise from the detector. The synchronization pulse required for time-correlated single-photon counting was provided by the current source used to seed the laser diode. A time-to-amplitude converter (TAC) was used to determine the temporal position of the avalanche pulse with respect to the synchronization pulse. The TAC output was then sent to a programmable gain amplifier, and subsequently the amplified TAC signal was converted to a digital value using a flash analog-to-digital converter (ADC) and registered into the appropriate address of a multichannel analyzer. All of TCSPC electronic components were situated on a single PC card resident on the bus of the PC.

The assembled microscope head was mounted on a linear stepper motor in the Li-COR 4000 automated DNA sequencer as shown in Figure 3.2. The image processing and operational control systems for the Li-COR 4000 DNA sequencer were located in an IBM-compatible computer running the OS/2 operating system. These systems synchronized the data acquisition with the fluorescence scanner and also, constructed a fluorescence intensity image of the gel during readout. In our application, this system was not modified to allow image processing and base calling using the standard protocols developed by the manufacturer. The time-resolved data were
accumulated on a second IBM-compatible computer, which required digital pulses generated by the SPAD to construct intensity and lifetime images (see Figure 3.3). In order to accumulate data on both operating systems, an in-house-constructed linear rate-to-amplitude converter was built supplying 0-10 V to the analog circuitry of the data acquisition software in response to photons rates from 0 to 500 000 count/s. Also, a standard NIM pulse was generated for each photoelectron event as well to feed the time-correlated single-photon-counting board. Within the second PC a PicoQuant SPC-430 time-correlated single-photon-counting module and National Instruments AT-DIO-32HS digital input/output board was housed. The SPC-430 provided up to 4096 time channels for each of the 128 separately measured decay curves, and the channel resolution could be as small as 813 fs. Two onboard memory banks allowed simultaneous transfer/collection of the data. The National Instruments DIO was part of the circuitry required for synchronizing the position of the fluorescence scanner situated in the DNA sequencer with acquisition of the fluorescence decays for each pixel in the 2D image of the gel. Using quadratic coded signals generated by the electronics driving the scanner across the gel, the position of the head was detected along with the direction of the movements and turnarounds. Additionally, the pixel in the middle of each scan line for both directions of the head movement was marked to allow for precise alignment during image construction. There were 400 pixels in each image line with the acquisition time of each pixel set to about 10 ms. For the image, each pixel represented a size of ~35 µm across the gel and stored 256 points (8-bit time resolution, 97 ps/time channel) of 32-bit words for the decay data. The microscope head was rastered across the 21 cm gel at an average rate of 3.5 cm/s requiring 6 s to fully scan a
Figure 3.2. Picture of near-IR time-resolved scanner mounted in the Li-COR 4000 automated DNA sequencer. The scanner moved in a linear fashion across the gel (finish line detection) ~18 cm from the top of the glass plates.
**Figure 3.3.** Experimental set-up of near-IR time-resolved lifetime data acquisition system coupled to Li-COR automated DNA sequencer. The computer running the OS2 software, controlling the intensity data acquisition for the Li-COR sequencer was unable to handle the data load for time-resolved measurements. An IBM based computer was setup to simultaneously collect intensity and time-resolved data and was synchronized to the head movements of the sequencer with positioning hardware and a program written in Visual Basic.

A single line across the gel giving a sampling frequency of 0.17 Hz. To prevent signal aliasing during electrophoresis due to the low sampling frequency, the applied voltage used for electrophoresis was kept at –1500 V, acquiring approximately 5 to 6 data points per electrophoretic band.

The data acquisition software was written in Visual Basic (Microsoft, Seattle, WA) and consisted of several control and data acquisition functions, such as recording the position of the scanning head, streaming intensity and time-resolved data to the hard
Figure 3.4. Front panel displays of the data analysis package for analyzing time-resolved data from sequencing gels. (A) Gel intensity image obtained from the time resolved scanner. In this particular display, four lanes of sequencing data were imaged. The vertical and horizontal cursors (double line) defined a region of the gel from which the decay profile was constructed. In this region no dye-labeled DNA was present and the resulting decay represents the instrument response function of the equipment. (B) Single-base tract (intensity electropherogram) from the gel image shown in (A). The region in the electropherogram selected for constructing the decay is marked with the cursor (single line). In this case, the displayed decay represents counts from a single pixel (acquisition time 10 ms).
drive, and providing real time visualization of the acquired data. During a typical experiment, 4 to 12 GB of data was generated and stored in 2 GB chunks. After electrophoresis was finished, the data was compressed one line at a time and then assembled into one contiguous file. Additionally, the images of the intensity and calculated on-the-fly lifetimes for each pixel were stored in a separate file with no compression. During the data analysis, each line was read and decompressed separately for the generation of the 2-D image. This approach allowed manageable data file sizes at the expense of slightly longer times to access the information. Data analysis was performed by reading the compressed, binary files into an extractor program viewed in Microsoft Windows Explorer (see Figure 3.4). The data could be displayed as a 2-D gel image, along one horizontal line across the gel or along one tract of the gel, the position of which could be selected with cursors. In the single-lane mode, the normal intensity electropherogram was displayed. The decay profiles for each pixel or a series of pixels were selected by a pair of cursors from either the 2-D gel image or the single-lane electropherogram. After construction of the decay profile, the lifetime was calculated using the maximum likelihood estimator.

3.2.2. DNA Sequencing

The gel plates were boro-float glass that measured 21 cm x 47 cm. The sieving matrix was an 8%T (w/v) cross-linked gel (FMC Bioproducts Long Ranger, Rockland, ME) that contained 7.0 M urea as the denaturant and 1x TBE (pH 8.0). Polymerization of 30 ml of the 8% gel was initiated by adding 200 µl of 10%(w/v) ammonium persulfate and 20 µl TEMED, and a comb was inserted between the glass plates to form wells. After 1 1/2 hrs, the comb was removed and the plates placed in the sequencing
instrument and the buffer reservoirs filled with 1x TBE. The gel was heated to 50°C and prerun at -1500 V for 30 minutes. One µl of each sample was loaded into the sample wells and the electrophoresis was run at –1500 V for 12 to 24 hours.

The two labeling dyes used for the experiments consisted of IRD700 (Li-COR Biotechnology, Lincoln, NE) and Cy 5.5 (Synthagen, Houston, TX), both of which were covalently attached to the 5’ end of a 17mer, M13 forward (-29) sequencing primer through a C6 amino linker. The structures of the near-IR dyes along with their absorbance and emission maxima and other fluorescence properties are shown in Figure 3.5. The acid form of the dyes was converted to succinimidyl esters and then covalently linked to the amino terminus of the sequencing primers. As can be seen from the spectroscopic data, the absorption and emission maximums are very similar, allowing efficient processing of the emission on a single detection channel and excitation with a single source. However, differences in their charge resulted in electrophoretic mobility differences requiring postelectrophoresis corrections during sequence assembly.

Sanger sequencing reactions were prepared using a modified procedure for the Amersham 7-deaza dye primer cycle sequencing kit employing an M13mp18 single-stranded DNA template. The sequencing cocktail consisted of 5 pmol of the single-stranded DNA template, 20 µL of TE buffer, 1.0 µL of the appropriate dye-labeled primer, 5 µL of double-distilled H2O, and 24 µL of the A, C, G, or T extension mixture (Amersham Pharmacia Biotechnology, Piscataway, NJ).

Cycle sequencing was accomplished in a Genius series 96 well peltier thermal cycler (Techne, Minneapolis, MN) using the following conditions (30 cycles): (i) 92°C for 2 s; (ii) 55°C for 30 s; (iii) 72°C for 60 s, followed by a final extension step at 72°C
Figure 3.5. Structure and spectroscopic properties of near-IR labeling dyes.

IRD 700
\[ \lambda_{\text{abs}} = 685\text{nm} \]
\[ \lambda_{\text{em}} = 705\text{nm} \]
\[ \phi_f = 0.50 \]

Cy 5.5
\[ \lambda_{\text{abs}} = 675\text{nm} \]
\[ \lambda_{\text{em}} = 694\text{nm} \]
\[ \phi_f = 0.50 \]
for 7 min. The primer concentration was reduced compared to the manufacturer’s protocol and the cycle number increased to reduce the amount of unextended primer remaining in the sequencing cocktail, which minimized smearing of unextended primer in the gel tract. The reaction cocktails were then subjected to a cold ethanol precipitation that consisted of the addition of 7 µL of 7 M NaOAc and 100 µL of 100% cold ethanol. The solution was vortexed and placed in the freezer (4°C) for 30 min, followed by centrifuging for 30 min at 10 000 rpm at 4°C. The supernatent was removed by tapping the reaction tube on a counter top and then the samples were thoroughly dried in a Centro-vap for 1 h. Finally the DNA pellet was reconstituted in a formamide-loading buffer, vortexed for 1 min, and placed in the freezer until needed for gel loading.

3.2.3 Data Analysis

The lifetimes of the labeled sequencing fragments were calculated using maximum likelihood estimators from the following equation [15, 16],

$$1 + (e^{T/\tau} - 1)^{-1} - m(e^{mT/\tau} - 1) = \sum_{i=1}^{m} i N_i$$  \hspace{1cm} (3.1)

where $N_i$ is the total number of photocounts in the decay profile, $T$ is the time width of each bin (97 ps), $m$ is the time interval over which the lifetime was calculated (25 ns), $i$ is the time bin number and $N_i$ is the number of counts in the $i$th time bin. The standard deviation in the measurement using this relationship is given simply by $\tau f N_i^{-1/2}$, where $\tau f$ is the calculated fluorescence lifetime.

To evaluate the effects of poor electrophoretic resolution on the extracted lifetime using equation 3.1, which can only extract a single lifetime even for
multiexponential decays, a Monte Carlo simulation program was written in Pascal to model the expected responses under different electrophoretic resolution conditions. The program constructed Gaussian intensity profiles representing the electrophoretic bands with input parameters of the total intensity for each band and the standard deviation in the Gaussian profile as well as the resolution between bands. After the band positions were set, Gaussian noise was added to the intensity response, and followed by construction of a decay profile at each time interval within the electropherogram. The decay profile was comprised of the appropriate number of counts from each component with a time response that was determined by the input lifetimes of the two components using the equation:

$$I_{f(t)} = \sum_{i}^{n} A_i e^{-t/\tau_{f(i)}}$$  \hspace{1cm} (3.2)

where $I_{f(t)}$ is the fluorescence intensity at time, $t$, $A_i$ is the fractional contribution of component $i$ to the decay, and $\tau_{f(i)}$ is the input lifetime of component $i$. After construction of the decay using equation 3.2, it was convolved with the experimentally acquired instrument response function obtained using the equipment described above on the gel matrix only.

3.3. Results and Discussion

For reading multiple lanes in sequencing applications, there are basically two different strategies may be considered, sequentially scanning the detector over multiple lanes [17, 18] or imaging the lanes onto an array detector [19, 20]. The imaging system can offer some unique advantages, such as high signal-to-noise ratio due to the fact that the duty cycle is much higher than in scanning-type systems. Unfortunately, array
detectors that possess timing responses appropriate for subnanosecond lifetime measurements and high single-photon-detection efficiencies in the near-IR are not available. Therefore, we selected a large photactive area SPAD and scanned the entire detection optics over the stationary gel fractionating the DNA fragments. As with any scanning system, the duty cycle is a primary concern due to signal-to-noise considerations. In our system, the sampling rate was set at 0.17 Hz with an integration time at each pixel of 10 ms, producing a duty cycle of 0.0017. The short data acquisition time was necessary to prevent aliasing of the signal to preserve electrophoretic efficiency in the separation. While the duty cycle is slow and the acquisition time is short, it has been demonstrated that highly precise and accurate fluorescence lifetimes can be determined in single molecule experiments with data accumulation times as short as 1 ms using maximum likelihood estimators to calculate the lifetime [21].

To assess the stability of the time-resolved scanner and the relative precision in the lifetime measurements, single base tracts of G, guanine, labeled with both dyes were analyzed in separate lanes of the gel on the automated DNA sequencer. In Figure 3.6, the decay profiles for both IRD700- and Cy 5.5-labeled sequencing fragments are shown as well as the instrument response function (IRF), which was collected over a spot on the gel with no fluorescent dye. The decay profiles and prompt function were constructed by integrating the total number of counts over equal numbers of pixels obtained from the intensity image of the gel. The amounts of scattering photons integrated over the pixel number were significantly less than the photocounts that were accumulated in the case where DNA bands were present.
Inspection of the prompt function showed the timing response of the device to be 450 ps (± 10 ps, fwhm) with the prompt function stable during the scanning stages of data acquisition as indicated by the low standard deviation in the fwhm of the IRF (RSD = 2.1%). Due to the speed of the scanning stage and the inhomogeneity of the gel matrix, the image size on the avalanche photodiode would be expected to change slightly during scanning. The difficulties that arise from this artifact are that the response function for avalanche photodiodes when operated in a Geiger mode and using a small photoactive area (i.d. < 200 µm), can change dramatically [14]. However, the use of a large photoactive area diode significantly reduced the timing response dependence on minor defocusing but at the expense of an increase in the dark count associated with the detector [10]. The significance associated with changes in timing behavior of the TCSPC system is that the precision in the calculated lifetime depends greatly on the stability of the time response of the instrument, especially when using maximum likelihood estimators [15].

Using equation 3.1, the lifetimes of both the IRD700- and Cy 5.5-labeled G-tracts were determined with the results shown in Figure 3.6. The average lifetime value and standard deviation was calculated by making a histogram of the lifetime values and fitting the data to a Gaussian function. The average lifetime value for the IRD700-labeled fragments was found to be 718 ps with a standard deviation of 5 ps (RSD = 0.7%), while the value for the Cy 5.5-labeled fragments was 983 ps with a standard deviation of 13 ps (RSD = 1.3%). By integrating the area under the decay profiles, the average numbers of photocounts, background-corrected, included in the lifetime calculations was found to be 56 500 counts for the IRD70-labeled fragments and 45 750
Figure 3.6. (A) Fluorescence decay profiles for IRD700- and Cy 5.5-labeled fragments. The individual dye-labeled fragments were run in separate lanes of the gel. Also shown is the resulting instrument response function. The decays were constructed by integrating the counts over 100 pixels from the gel image. (B) Histograms showing the fluorescence lifetimes calculated for single dye tracts of G-labeled sequencing ladders electrophoresed in a denaturing slab gel (7 M urea). The lifetimes were calculated over 65 bands for the IRD700 and 74 bands for Cy 5.5. The resulting histograms were fit to Gaussian functions from which the standard deviations were calculated. The sequencing ladders were loaded into wells of the gel (1.2 µl) and the electrophoresis was run at an applied voltage of –1500 V. The laser power was set at ~1.0 mW with a lasing wavelength of 680 nm and a repetition rate of 80 MHz.
counts for the Cy 5.5-labeled fragments. For maximum likelihood estimation of the lifetime, the standard deviation can be determined from $\tau_{f(\text{av})} N_t^{-1/2}$, where $N_t$ is the total number of photocounts included in the calculation and $\tau_{f(\text{av})}$ is the average calculated lifetime. Calculation of the standard deviation based on the photocount numbers was 3 ps for IRD700 and 5 ps for Cy 5.5, in close agreement with the measured standard deviations obtained from actual measurements. This indicates that photon statistics and not instrumental or background artifacts primarily determine the variances in the lifetime measurements, a direct consequence of using near-IR fluorescence monitoring.

Using the means and standard deviations determined from data presented in Figure 3.6B, a Student’s $t$ test indicated that we could successfully discriminate between these two dyes at a >99% confidence interval ($n_{\text{obs}} = 141$; $t_{\text{calc}} = 1.65$; $t_{0.99} = 2.58$).

A problem we noticed was that a significant amount of unextended primer produced smearing in the gel tract, which was evident from the gel intensity image. If the prompt function was measured prior to the migration of the primer peak through the detection zone it showed a fwhm of the gel that was similar to a spot on the gel where no dye passed. However, after the migration of the primer through the detection zone, residual dye remained in the gel, which was evident from the appearance in the decay profile. To minimize this gel smearing produced by unextended primer, we reduced the dye-primer concentration during polymerization and also increased the number of thermal cycles carried out during chain extension to reduce the amount of unextended primer.

The ability to identify a particular labeling dye comprising an electrophoretic band, when more than one dye is used in a single tract, will inevitably depend not only
on the total number of photoelectron counts but also on the resolution between bands
since the algorithm we are utilizing can only extract a single value even in the case of
multiexponential decays. Therefore, the sequencing read length will depend on the
signal intensity in the band and also securing a lifetime along the band where minimal
overlap occurs between neighboring bands. An example of the behavior of the MLE
algorithm in the case of high (R = 1.0) and low (R = 0.2) electrophoretic resolution can
be seen from the Monte Carlo results shown in Figure 3.7 for dyes with lifetimes of 750
and 850 ps. For R = 1.0, the calculated lifetime converges to the actual lifetime of the
dye comprising the band using photocounts to construct the decay near the center of the
electrophoretic band where the signal is dominated by fluorescence photons originating
from the labeled DNA fragment and not the background photons. A large contribution
from background photons causes the extracted lifetime to be biased toward lower
values, as is the case on the rising and falling edges of the band. When R = 0.2, the
calculated lifetime becomes a weighted average of the two components comprising the
decay. However, judicious choice of sampling intervals along the electrophoretic band
can improve the accuracy in the determination. For example, selecting a time interval
that has minimal overlap with the neighboring component minimizes the number of
photocounts included in the calculation from this neighbor, but at the expense of
increasing the uncertainty in the measurement due to the inclusion of fewer counts.
Running a two-dye tract in a sequencing gel on the automated sequencer, with IRD700-
labeled C-fragments and Cy 5.5-labeled T-fragments, validated our Monte Carlo results.
In Figure 3.8A is shown a section of an electropherogram in which the base assignment
was determined from the center pixel of each band when the band was baseline
resolved. In the case of poor resolution, the pixels across the electrophoretic bands were used to identify terminal nucleotides from the resulting lifetime pattern. In Figure 3.8B, the peak pair depicted possessed a resolution of ~0.36. As can be seen from Figure 3.8, the calculated lifetime was found to increase when moving up the band front and then slightly leveled near the center of the first band at a lifetime value close to that expected for the IRD700-labeled C-fragments. Moving further along this first band and into the second band caused a dramatic increase in the observed fluorescence lifetime value, consistent with the second band composed of a Cy 5.5-labeled T-fragment. In the case of band doublets composed only of IRD700- or Cy 5.5-labeled fragments, the lifetime values across the doublet were found to level off at the expected lifetime of the appropriate dye. Therefore, the identity of individual components in a series of bands with poor electrophoretic resolution can be determined by tracking the lifetime values across the bands and comparing the resulting values to known lifetime values of the dye reporters.

To evaluate the efficiency in base calling using our lifetime approach, we performed a two-dye/two-lane sequencing experiment, in which A (IRD700 label) and T (Cy 5.5 label) tracts were run in one lane and the C (IRD700 label) and G (Cy 5.5 label) in an adjacent lane. The efficiency of our base calling was compared to a single-dye sequencing run, in which the four terminal bases were run in separate lanes of the gel using a single dye-labeled primer (Cy 5.5-labeled sequencing primer). Since it was found that the Cy 5.5-labeled fragments migrated at a slightly higher rate than that of the IRD700-labeled fragments, the identified IRD700 bands were frame shifted across the entire tract. To assist in identifying the number of components comprising a band
Figure 3.7. Monte Carlo simulations of two dyes with lifetimes of 750 and 850 ps. (A) Simulation with R = 1.0 and (B) R = 0.2.
and securing proper overlap between the two tracts, the base calling algorithms associated with this machine were first implemented, which identified peaks as either A’s or C’s in each tract and then overlaying the data from each tract. Following this process, A’s were then identified as either A or T and the C’s as either G or C using lifetime discrimination. In addition, the terminal base identification for a series of bands with poor electrophoretic resolution was carefully evaluated by obtaining lifetime patterns across these bands on a pixel-by-pixel basis.

A single-dye/four lane format was used to construct the sequence of an M13mp18 template with the called bases compared directly to the known sequence of the M13mp18 phage and known priming site (Figure 3.9). The single-dye/four lane method served as a standard for comparing the efficiency of our lifetime base call approach, since the single-dye/four lane method uses no spectroscopic identification process for calling the bases and the read accuracy depends primarily on the resolution in the electrophoresis. For the single-dye/four-lane sequencing run, the bases were called using the automated algorithms developed for this machine by the manufacturer. For a read length that included 670 bases, the total number of errors were found to be 29, resulting in a read accuracy of 95.7%. As can be seen from the data in Table 3.1, the majority of the errors were insertion errors, typically resulting from spacing artifacts due to poor electrophoretic efficiency for the later migrating fragments. If the read was dropped to 600 bases, the accuracy improved to 97.3%, while a read length of 550 bases produced a read accuracy of 99.3%, primarily due to a reduction of insertion/deletion errors.
Figure 3.8. Slab gel electropherogram of IRD700-labeled C-fragments and Cy 5.5-labeled T-fragments. The intensity electropherogram is shown in (A) along with the base assignments obtained from the calculated lifetime using equation 3.1. The decay profiles were constructed from a single pixel at the top of the peak, which contained the highest fluorescence intensity. (B) Expanded view of the band doublet shown in (A) with a resolution of 0.36 between these bands. The lifetimes were calculated over a 10 ms time interval at a sampling rate of 0.17 Hz. The dashed lines represent the expected lifetimes for IRD700 (718 ps) and Cy 5.5 (983 ps). The electrophoresis was performed single tract called using lifetime discrimination methods. The terminal bases A/T at an applied voltage of ~1500 V with a loading volume of 1.2 µl. The laser power was set at ~1.0 mW with a repetition rate of 80 MHz.
The bases were then called from a two-dye/two-lane format with the two bases in the run were in one lane and C/G in another. The results on the number and type of errors as a function of read length are presented in Table 3.1. The bases were called by selecting the center pixel of a band, where \( t = 10 \) ms/pixel and constructing a decay profile over this integration period from which the lifetime was calculated using the maximum likelihood estimator. When the resolution was < 0.5 for a series of bands, the lifetimes were calculated over 10 ms pixels across the band and the resulting pattern was used to identify the terminal bases. The acquisition of the lifetime pattern determined pixel-by-pixel across the band series not only assisted in identifying the terminal nucleotide base but also discerning the number of components comprising the region to minimize insertion or deletion errors. As can be seen in Table 3.1, the read accuracy was found to be better than the single-dye/four-lane strategy due to the fact that fewer insertion and deletion errors were found during sequence reconstruction. We found that the number of miscalled bases from the primer-annealing site was equal to 2

**Table 3.1.** Read accuracies and errors as a function of read length using a single-dye/four-lane and two-lane/two-dye sequencing strategies.

<table>
<thead>
<tr>
<th>read length (bp)</th>
<th>insertions</th>
<th>deletions</th>
<th>miscalls</th>
<th>accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-Dye/Four-Lane</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>550</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>99.3</td>
</tr>
<tr>
<td>600</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>97.3</td>
</tr>
<tr>
<td>670</td>
<td>21</td>
<td>4</td>
<td>4</td>
<td>95.7</td>
</tr>
<tr>
<td><strong>Two-Lifetime/Two-Lane</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>550</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>99.7</td>
</tr>
<tr>
<td>600</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>99.7</td>
</tr>
<tr>
<td>670</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>99.7</td>
</tr>
</tbody>
</table>
Figure 3.9. Electrophoretic trace with base calls for single-dye/four-lane strategy. Solid boxes indicate a mis-call from the automated sequencer, dashed-boxes indicate a miscall from lifetime base call and lines represent a deletion. The electrophoresis was
performed at –1500 V with a loading volume of 1.2 µl of IRD700-labeled reactions. The average laser power was 1.0 mW and a laser repetition rate of 80MHz. for a read length of 670 bases, similar to that found for the number of miscalls obtained in the single-dye/ four-lane approach. The number of miscalls would be expected to increase based on the fact that, in the lifetime call, a spectroscopic measurement is required to identify one of two terminal bases, which should add errors to the identification process. These results indicate that the lifetime approach as implemented here produces very high efficiency in the spectroscopic identification process. The reduction of insertion and deletion errors using lifetime patterning was a consequence of increased information content in the electrophoretic bands experiencing poor electrophoretic resolution.

3.4. Conclusions

We have modified a microscope head in an automated DNA sequencer to allow implementation of lifetime identification of DNA fragments labeled with near-IR fluorochromes and fractionated via slab gel electrophoresis. The simple instrumental modifications that were required indicate that many existing machines using steady-state fluorescence can easily be configured to do time-resolved fluorescence as well to increase system capabilities. In addition, scanning fluorescence detectors that possess color-discrimination capabilities can be modified to also carry out lifetime discriminations as well, potentially increasing the information content in a single electrophoresis run. For example, using a two-color approach with four-level lifetime discrimination will permit simultaneous forward and reverse sequencing reads off of a double-stranded template not only increasing throughput, but also assisting in gap closure in shotgun sequencing projects. Additionally, color/lifetime multiplexing will
be particularly attractive in fragment analysis, where it may be necessary to screen for several different targets in a single electrophoresis run. Our data indicates that the accuracy in the lifetime identification was high when compared to a single-dye/four lane strategy due to the use of near-IR fluorescence excitation, which minimizes impurity fluorescence generated from the sample matrix, minimizing variations in the calculated lifetimes. The high sensitivity of near-IR fluorescence also allows short integration times for construction of decay profiles across poorly resolved bands to construct a lifetime pattern, which can assist in elucidating components comprising the bands.

### 3.5. References


Chapter 4

Optimization of Sequencing Conditions for Near-IR Dye-Primers and Dye Terminators

4.1 Near-Infrared Dyes for DNA Sequencing

Any dye that is used for steady-state fluorescence detection can be used for time-resolved detection as well. However, it is important to use fluorophores that possess long lifetimes, at least as long or longer than the instrument response function (IRF), to allow efficient calculation of the lifetime. In multiplexed applications, where multiple dyes need to be identified in a single CE run, it becomes necessary to develop dyes with sufficiently different lifetimes to allow for efficient identification. The difference in lifetimes required for identification must be greater than the standard deviation in the measurement. In addition, a set of chromophores where there is a minimal difference in the excitation and emission profiles would allow efficient processing on a single detection channel.

A recently prepared series of near-IR fluorescent tricarbocyanine dyes, possess the capability to be used in time-resolved identification of multiple dyes in CE applications. These dyes, introduced in Chapter 2, are structurally similar, therefore show similar absorption/emission maxima, 765/794 nm, and also demonstrate uniform mobility in electrophoresis (see Figure 4.1). As can be seen from the slab gel image, the labeled T-tracts travel with the same apparent mobility through the crosslinked gel matrix. Across the entire tract, the mobilities are similar between the four dyes and no apparent shift appears due to the length of the DNA fragments.
Figure 4.1. Slab-gel image of near-IR tricarbocyanine labeled T-tracts run at –1500V in an 8%T crosslinked polyacrylamide gel. The bands in every lane travel at the same rate through the matrix, demonstrating uniform mobility within the dye set.

Incorporating a heavy atom modification into the molecular framework of the base chromophore altered the fluorescence lifetime of the dye series. This phenomenon is caused by two processes, the intramolecular heavy-atom effect, and internal conversion which results in an increase in $\tau_f$ with the increasing molecular weight of the
substituted heavy-atom. By strategically incorporating a single halide (I, Br, Cl, or F) into the molecular structure, the lifetime of the dye set varied from 889 ps to 688 ps [1]. While these differences are somewhat small, previous studies have shown the calculated lifetime obtained during capillary gel electrophoresis of near-IR labeled oligonucleotides can be determined with high precision using maximum likelihood estimators [2]. The high precision is due in part by utilizing near-IR excitation, which minimizes the background contributions into the estimated lifetime, improving measurement precision.

The work presented here investigates the preparation and separation environment of near-IR dyes to identify which factors may be important for optimization of multiplex fluorescence lifetime detection. In order to achieve this, a number of factors had to be jointly considered including sequencing chemistry, separation matrix, and electrophoretic column conditions. McGown and co-workers determined that the lifetimes of the primer-attached dyes they studied remained constant in gels of different composition [3]. We determined the dependence of fluorescence lifetime on gel matrix and on experimental conditions for the set of near-IR tricarbocyanine dyes employed for lifetime base calling.

4.2. Experimental Section

4.2.1. Capillary Preparation

Capillaries that are filled with a crosslinked or linear polyacrylamide must be coated to prevent the gel matrix from being extruded from the capillary by the forces exerted by the electroosmotic flow. In our studies, a 70-cm-long capillary, 75 µm-inner diameter, 365 µm-outer diameter fused silica capillary (Polymicro Technologies,
Phoenix, Arizona) was rinsed with 1M NaOH followed by 1M HCl with intermediate washings of distilled water. Next, the capillary was treated with a [γ-(methacryloxy) propyl]trimethoxysilane (γ-MAPS) solution adjusted to 2.5 pH with glacial acetic acid. One end of the capillary was connected to a water aspirator and the other end was placed into the γ-MAPS solution. Once the capillary was filled with the γ-MAPS solution, it was allowed to react for six hours after which vacuum was used to remove the solution. Distilled water was rinsed through the capillary to remove the residual γ-MAPS solution. One milliliter of an aqueous 4% (w/v) acrylamide solution was prepared and polymerization initiated by adding 1 µl of TEMED and 1 µl of 10% (w/v) ammonium persulfate. The solution was vortexed and introduced into the capillary by application of vacuum aspiration and allowed to react for 10 minutes before the excess gel was removed by applying a vacuum. The capillary was subjected to a final rinse with distilled water, dried in an oven at 40°C for 1 hour, and stored until needed.

Crosslinked gel capillaries containing a 3T3C or 5T5C concentration and 7M urea were commercially obtained (J&W Scientific, Palo Alto, California).

The synthesis of high molecular weight linear polyacrylamide (non-crosslinked) solutions consisted of preparing a 6% (w/v) acrylamide in water and purging for at least 30 min at 0°C with high-purity He to minimize dissolved oxygen, according to Karger and co-workers [4]. Addition of 1 µl of 10% (w/v) ammonium persulfate and 1µl TEMED/mL of acrylamide solution initiated the polymerization. The polymerization was allowed to proceed to completion over 24 hours at 0°C. Denaturants and buffer solutions were used to dilute the gel solution to produce the final separation matrix.
Commercial linear polyacrylamide, CEQ Separation Gel I, was obtained from
Beckman Coulter (Fullerton, CA). The denaturant was urea, but the concentration was
unknown.

The PDMA used for our experiments was acquired from Applied Biosystems
(Foster City, CA) under the name of POP6. The exact gel and denaturent
concentrations are not known, but are assumed to be around 6% PDMA and 7M urea.
POP6 is a self-coating matrix and did not require a coated capillary to reduce the
electroosmotic flow.

4.2.2. Instrumentation

Figure 4.5 shows a representation of the time-correlated single photon counting
(TCSPC) experimental set up for the electrophoresis experiments described here. The
instrumentation used to acquire the fluorescence lifetimes on-line during gel
fractionation consisted of a conventional time-correlated single photon counting
apparatus. The excitation source was a Ti:sapphire laser pumped by the all-lines output
of a small frame Ar ion laser. The Ti:sapphire laser was tuned to 765 nm to match the
excitation maxima of the four dye set used for sequencing. The laser light was directed
into a single mode fiber, which transported the laser light to the capillary. The use of
the fiber optic minimized beam walk, which required constant realignment of the
system. The light emanating from the fiber was collimated by a 10X microscope
objective and focused onto the capillary tube using a single laser focusing lens (Melles
Griot, Irvine, CA). The fluorescence emission was collected in a conventional 90°
format using a 60X (NA = 0.85) microscope objective (Nikon). The collected radiation
was imaged onto a slit, spectrally filtered (interference band pass filter, center
wavelength = 830 nm; half band width =30 nm, Omega Optical, Brattleboro, Vermont) and then reimaged onto the face of a photo detector. The detector consisted of a passively quenched single photon avalanche diode SPAD, SPCM-PQ-200 (EG&G optoelectronics, Vandrieul, Canada). The output of the SPAD was amplified 20-fold.

**Figure 4.2.** Near-infrared time-correlated single-photon counting instrument. Components: L, focusing lens; C, capillary; SF, spatial filter; BF, bandpass filter; SPAD, single photon avalanche diode detector; Amplifier, signal amplifier; Delay, time delay electronics; Disc., discriminator; PD, photodiode detector; TAC, time-to-amplitude converter.
with an in-house fabricated fast amplifier and conditioned by a constant fraction discriminator (CFD). The synchronization pulse obtained from the pulsed-laser was obtained from an intracavity photodiode, which was also directed into the CFD. The time-to-amplitude converter (TAC) was operated in a gated mode, with the gate pulse generated by the photocounts \( g \) from the SPAD. The start pulse was supplied by the SPAD output while the laser synchronization pulse was used as the stop pulse.

A Li-COR 4000 automated DNA sequencer (Li-COR, Lincoln, NE) was used for analysis of some samples. The gel plates were boro-float glass that measured 21 cm x 47 cm. The sieving matrix was an 8\% T (w/v) cross-linked gel (FMC Bioproducts Long Ranger, Rockland, ME) that contained 7.0 M urea as the denaturant and 1x TBE (pH 8.0). Polymerization of 30 ml of the 8\% gel was initiated by adding 200 \( \mu l \) of 10\%(w/v) ammonium persulfate and 20 \( \mu l \) TEMED, and a comb was inserted between the glass plates to form wells. After 1 1/2 hrs, the comb was removed and the plates placed in the sequencing instrument and the buffer reservoirs filled with 1x TBE. The gel was heated to 50\(^\circ\)C and prerun at -1500 V for 30 minutes. One \( \mu l \) of each sample was loaded into the sample wells and the electrophoresis was run at –1500 V for 12 to 24 hours.

### 4.2.3. DNA Sequencing

The labeling dyes used were near-IR heavy-atom modified tricarbocyanine dyes covalently attached to the 5’ end of a 17 mer, M13 forward sequencing primer through a C6 amino linker (see Figure 4.3).

Sanger sequencing reactions were prepared using a modified procedure for the Amersham 7-deaza dye primer cycle sequencing kit employing an M13mp18 single-
stranded DNA template. The sequencing cocktail consisted of 4 pmol of the single-stranded DNA template, 1.0 µL of the appropriate dye-labeled primer, 15 µL of double-distilled H₂O, and 2 µL of A, C, G, or T extension mixture (Amersham Pharmacia Biotechnology, Piscataway, NJ).

Cycle sequencing was accomplished in a Genius series 96 well Peltier thermal cycler (Techne, Minneapolis, MN) using the following conditions (54 cycles): (i) 92°C for 2 s; (ii) 55°C for 30 s; (iii) 72°C for 60 s, followed by a final extension step at 72°C for 7 min. The primer concentration was reduced compared to the manufacturer’s protocol and the cycle number increased to reduce the amount of unextended primer remaining in the sequencing cocktail, which minimized smearing of unextended primer

\[ \text{X} = \text{H, Cl, Br, F, I} \]

**Figure 4.3.** Structure of near-IR heavy-atom modified tricarbocyanine dyes with C6 amino linker and M13mp18 forward primer.
in the gel tract. A cold ethanol precipitation consisted of the addition of 7 µL of 7 M NaOAc and 100 µL of 100% cold ethanol. The solution was vortexed and placed in the freezer (4°C) for 30 min followed by centrifuging for 30 min at 10,000 rpm at 4°C. The supernatent was removed by tapping the reaction tube on a counter top and then the samples were thoroughly dried in a Centro-vap (Brinkman Instruments, Westbury, NY) for 1 h. Finally the DNA pellet was reconstituted in distilled water, vortexed for 1 min, and placed in the freezer until needed for injection.

For sequencing sample clean-up Microspin-25 Sephadex based size exclusion columns (Amersham, Piscataway, NJ) were employed. The columns trap small molecules, <15 bases in length, and allow larger molecules to pass through. For removal of the dye-primer the columns were vortexed to resuspend the column contents, placed in centrifuge tubes, and centrifuged at 3000 rpm for 1 minute. The columns were placed in a new centrifuge tube and 40 µl of the sequencing reaction placed in the center of the packed column bed. The column was placed back into the centrifuge for 2 min at 3000 rpm. The resulting solution that had passed through the column was subjected to a cold ethanol precipitation.

4.3. Results and Discussion

4.3.1. Sample Preparation

The cycle sequencing reaction that replicates the DNA template in a linear amplification format requires three steps: denaturing, annealing, and extension. The typical temperatures utilized for the sequencing conditions are; an initial denaturing step at 95°C for 2-5 minutes followed by 95°C for 5-120 sec, 55°C for 30-120 sec, and 72°C for 30-300 sec. In some cases, a final extension step ranging from 2-10 minutes is used.
Many different temperature schemes were investigated to determine the optimum conditions that should be employed for the near-IR labeled primers used. Due to the temperature sensitivity of the near-IR dyes [5], the sequencing conditions decided upon were as follows: 92°C for 3 sec, 55°C for 30 sec, and 72°C for 60 sec, with a final extension at 72°C for 7 min. An initial denaturing step was removed from our protocol since a single stranded template was utilized that does not need to be initially denatured. The elimination of the initial denaturing and extended periods of heat for denaturing the DNA during cycle sequencing increased the fluorescence signal seen on slab gel images and in capillary electropherograms. From experimentations with the temperatures used for the denaturing steps, it was observed that with a shorter duration of denaturing, there was an increase in the intensity of the sequencing fragments and a decrease in the amount of free dye seen in the sequencing run. The temperature sensitivity of these conjugates is thought to be caused by multiple factors including, heat decomposition of the tricarbocyanine dye and/or a chemical cleavage of the oligimer from the dye structure [5].

Strekowski has shown that aryl thiols efficiently displace phenols, which are at the meso position on bridged tricarbocyanine dyes [6]. Most extension reaction enzymes require the use of dithiothreitol (DTT) to activate the sequencing enzyme through the breaking of disulfide bonds that aid in efficiently incorporating deoxy- or dideoxynucleotides in DNA extension protocols. The action of DTT on the activity of sequencing enzymes is not completely understood, but its presence may stabilize the tertiary structure of the enzyme. Although DTT is not an aryl thiol, it does contain free alkyl thiol groups that could displace the phenol substituent at the meso position.
Flannagan et al. confirmed the substitution of DTT on the meso position of these type of tricarbocyanines via FAB-MS [5].

The cleavage of the functional group linking the dye to the primer is represented in Figure 4.4 and causes free dye, dye unattached to the primer, to be present in the sequencing mix that can be injected into the gel matrix for size separation. This dye artifact causes smearing in the sequencing tract of the DNA ladder (see Figure 4.5), since the free dye migrates more slowly than the dye primer. The excess dye-primer contained in the sequencing reaction that does not extend, shows up as a large intense spot on the sequencing gel and can cover up a portion of the sequence trace close to the priming site on the template. Because the free dye migrates through the gel matrix more slowly than the dye primer, it covers up the sequence further away from the
priming site on the template. In addition, two fluorescent species of the free dye were observed in the trace, and therefore will mask a large portion of the readable sequence.

To remove this free dye from the reaction mixture, a cold ethanol precipitation was found to be effective and was performed on the post sequencing reaction. This procedure precipitates the DNA out of the sequencing buffer mix using a salt, usually ammonium acetate (NH₄Oac) or sodium acetate (NaOAc) and cold ethanol. To the sequencing reaction, 5 µl of 3 M NaOAc solution and 200 µl 100% cold ethanol was added and placed at 4°C for 30 minutes. The DNA, which is insoluble in ethanol, will precipitate out of the solution. The solution is then centrifuged at 10,000 rpm for 30 minutes at 4°C to precipitate the DNA into a small pellet at the bottom of the tube. The supernatent, consisting of ethanol, free dye and other components from the sequencing mix is removed and the tube with the DNA pellet at the bottom is placed in a concentrator (Brinkman Instruments, Westbury, NY) at 45°C for 1 hour to evaporate any excess ethanol. The DNA is then reconstituted in water or a formamide-loading buffer and injected into the gel matrix. The precipitation not only allows for the removal of free dye but for concentration of the DNA extension fragments.

The acetate solution used to perform the ethanol precipitation was also investigated since ammonium ions are known fluorescence quenchers [7]. As can be seen in Figure 4.6, the precipitation performed with ammonium acetate produced a DNA ladder of lower intensity than the sodium acetate ethanol precipitation tract for equal amounts of sequencing cocktail inserted into the gel. Since our base identification is dependent upon the fluorescence lifetime of the dyes and quenchers may perturb the fluorescence decay, sodium acetate was used for further ethanol precipitations.
Figure 4.5. Slab gel image of near-IR fluorine-modified tricarbocyanine dye T-tract before (2) and after (1) subjection to a cold ethanol precipitation. Gel was an 8%T cross-linked polyacrylamide with 1.2 µl of NIR-F T-tract electrophoresed on a Li-COR 4000 automated sequencer at –1500 V for 10 hours.
**Figure 4.6.** Slab gel image of sodium and ammonium acetate ethanol precipitation of near-IR bromine-labeled T-tract. Ammonium acetate precipitation shows lower fluorescence intensity due to the ion’s quenching effects on the fluorescence molecules. One µl of NIR-Br T-tract injected and run at $-1600$ V in an 8% T cross-linked polyacrylamide gel on a Li-COR 4000 automated sequencer.
4.3.2. Separation Matrices

Polyacrylamide is one of the most common gel matrices used and can be prepared from the acrylamide monomer \((CH_2=CHCONH_2)\) in varying manners. The most common method of polymerizing acrylamide is to copolymerize it with a percentage of crosslinker, \(N,N'-\text{methylenebisacrylamide (CH}_2\text{(CHCOCH=CH}_2\text{)})\) in the presence of a free-radical initiator, ammonium persulfate. The \(\%T\), the total amount of acrylamide and \(\%C\), the amount of crosslinker, designate the gel composition. The \(\%T\) is calculated from;

\[
\%T = \frac{\text{acrylamide(g)} + \text{bisacrylamde(g)}}{100ml} \times 100\% \quad (4.1)
\]

and \(\%C\) is calculated from;

\[
\%C = \frac{\text{bisacrylamide(g)}}{\text{bisacrylamide(g)} + \text{acrylamide(g)}} \times 100\%. \quad (4.2)
\]

The acrylamide monomer can also be polymerized alone to form an entangled linear polymer solution [4, 8]. To ensure that the DNA maintains a single strand conformation during fractionation, a denaturant, such as formamide or urea, must be added to the polymer matrix as well.

4.3.3. Gel Studies

The separation matrix was next optimized to balance the needs of the electrophoretic separation parameters and the spectroscopic requirements for base identification using fluorescence lifetime methods and our near-IR dyes. In order to call bases efficiently, high electrophoretic resolution is needed so that every base can be identified and not obscured by overlapping with adjacent bands. Also, the matrix can
significantly alter the microenvironment of the dyes, affecting their apparent lifetime values. The gel matrix can enhance or degrade the differences between the lifetimes of the near-IR dyes, and for lifetime identification, the $\tau_f$ of the dyes must be consistent and well-resolved from the other $\tau_f$ values in the dye set to ensure proper identification. The target read length was at least 600 bases for a four-dye sequencing run, some a longer read length, makes it easier for sections of DNA to be reassembled [9].

4.3.4. Crosslinked Polyacrylamide

The dye-primers’ fluorescence intensities and lifetimes were characterized in various matrices with different concentrations of total acrylamide and denaturants, urea and/or formamide, to optimize the sequencing read length, which will depend on; intensity, mobility, and fluorescence lifetimes.

A standard crosslinked polyacrylamide column with urea as the denaturant was initially tested. Many difficulties could arise from the use of crosslinked gels. These potential difficulties could include; lengthy column conditioning between runs, degradation of the column during runs, large standard deviations of lifetimes values, gel memory effects, and/or nonlinear fluorescence decays.

Crosslinked capillary columns require extensive conditioning between subsequent analysis. For example, once a sample is injected onto the column, the electrophoresis must continue until the background reaches a value similar to that before the run, indicating the entire sample had completely migrated from the column. Unfortunately, the constant electric field placed across the capillary caused the gel columns to degrade due to the formation of voids at the ends of the capillary and within the matrix. Inspection of the capillary under a microscope revealed the presence of
these voids and continuation of the applied electric field caused the voids to get larger. To eliminate the rapid degradation of the gel matrix, the ends of the capillary could be clipped off before a new experiment begins. If the void at the injection end of the column is not removed before a new sample is injected, the sample will not be electrokinetically injected properly resulting in poor electrophoretic resolution.

The chromatographic efficiency of the cross-linked gel columns was found to decrease with increasing number of electrophoretic runs. Figure 4.7 shows two electrophoresis runs of a near-IR Br-labeled T-tract injected into a 3T3C capillary with 7 M urea for one minute at 150 V/cm and running at 170 V/cm. The initial run shows good separation of the sequencing fragments throughout the run, while the fourth run showed significant degradation in the separation efficiency. The inset of Figure 4.7 compares a portion of the M13mp18 T-tract for a T quintet and quartet for these two separations. The resolution of the fragments was significantly decreased in the fourth run, where the average R was 0.86 in the first run and reduced to 0.40 in the fourth run. Another observation from these electropherograms was an increase in the background from the first to fourth run and an increase in the average $\tau_f$ from run to run. The increase in background was attributed to a memory effect. Memory effects are caused when residues of the dye or unextended dye-primer remain in the matrix or adhere to the walls of the capillary and leave a “fingerprint”. These fingerprints can be observed by monitoring the “apparent” decay of the gel matrix (no DNA band).

An unused crosslinked capillary will accurately reflect the IRF of the system. After electrophoresing a dye-labeled primer or sequencing ladder through the capillary, the background does not return to the original value and the IRF appears to have a
larger width. The residual dye in the matrix can complicate further analysis by biasing lifetime results by creating multiexponential decays.

An example of this phenomenon is shown in Figure 4.8. The first panel is of a decay in a new cross-linked gel. The second panel follows four sequencing runs and was acquired after the background had stabilized to its lowest value. Even though there were no DNA bands migrating within the detection zone, there is clear evidence of fluorescence in the decay.

**Figure 4.7.** Difference in resolving power of a crosslinked gel capillary from the initial to the fourth electrophoresis run. Injection of NIR-Br labeled T-tracts for 1 minute at 150 V/cm and run at 170 V/cm. The first run had a delay of 45 minutes before data collection began; the fourth run had a delay of 25 minutes.
Evidence of the presence of fluorescence is the resultant wider full-width at half maximum (FWHM), which is used to describe the IRF. The method selected to calculate the lifetime values of the dye-labels is a single component method and as such, will effectively produce a single value, even for multi-component decays. In order to achieve accurate lifetime values, the IRF should be as narrow as possible to contribute little signal to the lifetime calculation.

A series of sequencing runs were next carried out using a single base NIR-Cl labeled tract in a 3T3C crosslinked gel column to investigate the variance in the calculated lifetime. In the first run, the average $\tau_f$ was $660 \pm 7$ ps, the second run, $783 \pm 11$ ps, the third run $901 \pm 12$ ps. In each run the calculated lifetime

![Figure 4.8](image)

**Figure 4.8.** Instrument response function of crosslinked gel column prior to electrophoresis and following four electrophoresis experiments. IRF from the final run was taken after the background value stabilized at the lowest point.
increased, resulting most likely from the fluorescence remaining from the previous run and “adding” to the decay in subsequent runs. When the sequencing samples were electrophoresed and analyzed in a new gel column, the average lifetime values were closer from run-to-run. A set of NIR-Fl labeled sequencing samples were next analyzed as the initial run on two different 3T3C crosslinked capillaries, with $\tau_f$ equal to 752 ps $\pm$ 6 ps and 763 ps $\pm$ 7 ps. The changing lifetime values of the dyes are attributed to changes in the dyes’ microenvironments due to urea or gel decomposition or the presence of radical initiators remaining from gel polymerization. The change in lifetime values demonstrates the need to use new gel matrices for subsequent electrophoretic runs.

![Electropherogram](image)

**Figure 4.9.** Electropherogram of NIR-Cl G-tract injected for 1 minute at 150 V/cm and run in a 3T3C gel column at 170 V/cm, 50 minute delay before data acquisition. Second electrophoresis run in the column.
Figure 4.9 shows a different type of variance seen in the cross-linked columns. The electropherogram shows a decreasing signal as the run proceeds due to biased injection. As the fluorescence intensity decreases the background contributes more to the decay biasing the calculated lifetime to shorter values as seen in this data.

Another problem identified with the cross-linked gels was the non-linearity in the decay profiles. As can be seen in Figure 4.8, the decay of the NIR-Br dye in a 3T3C crosslinked capillary demonstrates a multi-exponential behavior. The calculated lifetime of the dye changed from $\tau_f = 930$ ps in the crosslinked matrix to $\tau_f = 899$ ps in the linear polyacrylamide.

4.3.5. Linear Polyacrylamides

Linear polyacrylamides were next investigated in the lifetime measurements to potentially avoid the problems associated with the use of cross-linked gel matrices. Linear gels can be replaced between runs, greatly reducing memory effects seen in cross-linked gels. Also, since each run contains a new matrix, degradation and standard deviations due to memory effects seen in cross-linked gels can be eliminated.

The linear polyacrylamide matrices used for these experiments were made with a medium molecular weight (~700,000 Da) polymer with differing concentrations of urea and/or formamide and buffer. The dye concentration in all gel solutions remained constant at ~10 nM. The gels were continuously pushed through the capillary at a constant rate of 0.2 ml/min to prevent photobleaching of the dyes.

It was observed (see Table 4.1) that for a gel with 7M urea as the denaturant, the higher the concentration of gel in the matrix the higher the lifetime values were for the
Table 4.1. Calculated lifetime values of near-IR tricarbocyanine dyes in varying concentrations of linear polyacrylamides with different concentrations of denaturants, urea and/or formamide. (%, linear polyacrylamide (w/v); M, urea; %, formamide).

<table>
<thead>
<tr>
<th></th>
<th>2% 7M</th>
<th>4% 7M</th>
<th>6% 7M</th>
<th>2% 3.5M 30%</th>
<th>4% 3.5M 30%</th>
<th>6% 3.5M 30%</th>
<th>2% 60%</th>
<th>4% 60%</th>
<th>6% 60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIR-I</td>
<td>881</td>
<td>967</td>
<td>1126</td>
<td>783</td>
<td>799</td>
<td>794</td>
<td>904</td>
<td>871</td>
<td>837</td>
</tr>
<tr>
<td>NIR-Br</td>
<td>775</td>
<td>899</td>
<td>1066</td>
<td>732</td>
<td>766</td>
<td>780</td>
<td>863</td>
<td>825</td>
<td>802</td>
</tr>
<tr>
<td>NIR-Cl</td>
<td>796</td>
<td>927</td>
<td>1051</td>
<td>715</td>
<td>753</td>
<td>751</td>
<td>851</td>
<td>827</td>
<td>791</td>
</tr>
<tr>
<td>NIR-F</td>
<td>714</td>
<td>952</td>
<td>1059</td>
<td>698</td>
<td>736</td>
<td>733</td>
<td>867</td>
<td>783</td>
<td>756</td>
</tr>
</tbody>
</table>

Figure 4.10. Decay profiles of NIR-I in: A) 6% PA 7M urea, B) 6% PA 3.5M urea 30% formamide, and C) 6% PA 60% formamide. Buffer used for dilution was 1 X TBE with pH ~ 7.8 and a dye concentration ~10 µM for all solutions. λ<sub>ex</sub> = 760 nm, power = 1.35 mW and flow rate = 0.2 mL/min.
entire series. We also observed that urea gels produced a larger variation within the dye series in the lifetime values and more non-linear decay profiles compared to formamide gels (Figure 4.10). The exact reasoning for this phenomenon is not known, but it is postulated that the amine groups on the urea or formamide (see Figure 4.11) may be forming ammonium ions [10] that could be quenching the fluorescence, or producing complex decay kinetics.

From this set of experiments, we concluded that it was best to use formamide-based gels with high polyacrylamide concentrations. Also, since the lifetime values of NIR-Br and NIR-Cl were similar, the dye primer with hydrogen at the site for halogen modification was used since it gave a larger lifetime difference within the set.

Commercial gels obtained from Beckman and Polysciences exhibited a definitive shelf life, in which the separation efficiency of the gels decreased dramatically (Figure 4.12) over time. The second separation shown was run three weeks after the gel was originally opened and used. In a direct comparison with the initial separation using the same sequencing ladder, injection, and run conditions, the separation efficiency was noticeably decreased, especially for the larger fragments. We hypothesize that the entanglement threshold of the gel matrices increases, increasing the
pore size of the gel and decreasing the resolving power of the gel [9]. There does not appear to be a chemical degradation of the gel based on the calculated lifetimes obtained from analyzing the labeled sequencing fragments. The $\tau_f$ of the original run was $982 \pm 6$ ps and the $\tau_f$ of the separation run three weeks later was $979 \pm 8$ ps demonstrating a similar environment for the dye molecules.

![Graphs showing intensity over time](image)

**Figure 4.12.** Separation of NIR-H G-tract in a commercial linear polyacrylamide with urea as the denaturant. Top panel was run the week the gel was received and bottom panel three weeks later. Degradation of the separation efficiency of the gel over a time-period is evident, under the same electrophoretic conditions, although the spectroscopic properties did not show any change.
4.3.6 Linear Polydimethylacrylamides

We next looked at a modified polyacrylamide that has been developed and is commercially available. The matrix is a linear polydimethylacrylamide (see Figure 4.13) and has many properties that are ideal for DNA separations in a capillary.

Polydimethylacrylamide (PDMA) is a self-coating polymer that reduces the electrosmotic flow in silica-based capillaries so that the columns do not need a cross-linked gel coating on the wall. This property of PDMA allows the life of the capillary to be extended, since refilling the matrix between electrophoresis runs regenerates the capillary. Also, the resolving power of PDMA is appropriate for separating ssDNA.

In Table 4.2 are the $\tau_f$ values obtained from calculating the lifetime of the individual dyes in sequencing runs compared to earlier values obtained in other matrices. The lifetime values calculated in the POP6 were comparable to initial lifetime values obtained by Flanagan [1]. The calculated lifetimes in the polyacrylamide matrices show a small difference between lifetime values for the dye series and the NIR-Br and NIR-Cl dye lifetimes are stastically the same. POP6 demonstrated a trend

![Figure 4.13. Structure of polydimethylacrylamide.](image)
of increasing lifetime values for increasing molecular weight of the heavy-atom modification, except in the case of the protonated dye. There is also a definitive separation between the lifetime values, especially for the Br and Cl modified dyes.

Based on these studies, more extensive experiments were carried out to determine the standard deviations of the lifetime values obtained for the tricarbocyanine dyes in POP6. Sequencing samples were prepared as described earlier and electrophoresed in POP6. The calculated lifetime values for each dye, taken from multiple peaks within a run and between electrophoretic runs, the average lifetime and the standard deviation were calculated. A histogram of the lifetime values and standard deviations for the dye series are shown in Figure 4.14. The calculated values show an appropriate separation, the smallest difference found to be 40 ps. POP6 was decided upon for the separation matrix for sequencing DNA labeled with near-IR heavy-atom modified tricarbocyanine dyes. Using the means and standard deviations determined from data presented in Figure 4.14, a Student’s $t$ test indicated that we could successfully discriminate between these two dyes at a 95% confidence interval ($n_{obs} = 41$; $t_{calc} = 1.68$; $t_{95\%} = 2.02$).

Table 4.2. Calculated lifetime values of near-IR heavy-atom modified tricarbocyanine dyes in various linear polyacrylamides and POP6. Flanagan acquired lifetime measurements in a non-polymerized 6%T5%C acrylamide solution with 40%formamide [1, 11]. (%, acrylamide; M, urea; %, formamide).

<table>
<thead>
<tr>
<th></th>
<th>Flanagan</th>
<th>2% 7M 0%</th>
<th>2% 3.5M 30%</th>
<th>4% 0M 60%</th>
<th>POP6</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIR-I</td>
<td>889</td>
<td>783</td>
<td>837</td>
<td>871</td>
<td></td>
</tr>
<tr>
<td>NIR-Br</td>
<td>821</td>
<td>732</td>
<td>802</td>
<td>825</td>
<td>850</td>
</tr>
<tr>
<td>NIR-Cl</td>
<td>759</td>
<td>715</td>
<td>791</td>
<td>827</td>
<td>796</td>
</tr>
<tr>
<td>NIR-F</td>
<td>735</td>
<td>698</td>
<td>756</td>
<td>783</td>
<td>757</td>
</tr>
<tr>
<td>NIR-H</td>
<td>906</td>
<td></td>
<td></td>
<td></td>
<td>977</td>
</tr>
</tbody>
</table>
Figure 4.14. Histogram of lifetime values calculated from labeled DNA fragments electrophoresed in POP6 with standard deviation and relative standard deviation. A Gaussian was fit to the calculated lifetime values.

4.3.7. Sequencing Studies in POP6

Single-dye tracts of dye-labeled primers were compared to tracts using dye-labeled terminators in POP 6. A significant difference in the quality of the sequencing runs was noticed between the dye-primer and dye-terminator sequencing runs. The high background seen in Figure 4.15 for the dye-primer electropherogram later into the run was due to a high concentration of unextended dye-primer traversing through the gel. In the lower panel of Figure 4.15, a dye-terminator tract is shown and as can be seen, the background level was nearly three times lower. The decrease in background resulted from the lower dye-terminator concentration used in the sequencing sample reaction mix.
Figure 4.15. Electropherograms of near-IR labeled dye-primer (top) and dye-terminator (bottom) sequencing reactions separated in POP6. The background is noticeably lower and more stable in the dye-terminator run.
The memory effect visible in the gel matrix from the dye-primers was confirmed after analyzing the decay profiles taken before and after electrophoresis of the samples (Figure 4.16). Notice the higher background in the dye-primer decay compared to the dye-terminator. The initial background for the dye-primer sample is higher than the dye-terminator because the sampling was done before the fluorescence intensity zone. The observation of the high background levels led us to believe that separation and identification of the dye-terminator samples would be more amenable for lifetime analysis, unless the large dye-primer peak could be reduced or removed from the sequencing reaction.

**Figure 4.16.** Decay profiles of POP6 before and after the dye-primer and dye-terminator samples electrophorese through the detection zone. The initial small fluorescence decay and higher background in the dye-primer sample is due to small amounts of dye-primer in the detection zone before the bulk of the sample arrives.
In an attempt to remove the excess dye-primer from the sequencing reaction, we employed a Microspin-25 Sephadex-based size exclusion column that traps small molecules, <15 bases in length, and allows larger molecules to pass. The samples were analyzed on a slab gel automated sequencer and in a capillary format with POP6 as the separation matrix. The results indicated that the excess dye-primer was almost completely eliminated (see Figures 4.15 and 4.16). The slab gel image shows a decreased amount of primer present in the sample compared to a sample with no exclusion performed. The capillary electropherogram verified the elimination of the majority of the unextended dye-primer from the sequencing reaction. Analysis of the capillary results, shown as insets of Figure 4.18, demonstrates the reduction in the background values in the sequence trace during the electrophoresis and a decrease in the amount of fluorescence present in the matrix.

One concern was the size exclusion procedure variability from run-to-run in the amount of dye-primer removed from the sequencing reaction. Therefore, the samples must be monitored to ensure a strong signal and removal of the excess dye-primer. Other brands of size exclusion columns were evaluated to find a more reproducible product, but none were found. A desired by-product of the size exclusion clean up is the slight removal of smaller sequencing fragments present in the sample as seen in Figure 4.17. This reduction in the smaller fragments helps to reduce the effects of biased injection seen in electrokinetic injection for capillary electrophoresis, resulting in a more even peak height distribution in the capillary electropherogram (Figure 4.18).
Figure 4.17. Slab gel image of NIR-H G-tract dye-primer sequencing reactions with and without a size exclusion clean up. The sample that has not undergone the size exclusion step, the second lane, has a large amount of excess primer present. 8%T cross-linked polyacrylamide gel run at –1600V/cm, 1.2 μl of sample per lane.
Figure 4.18. Electropherogram of dye-primer sequencing reactions with and without the size-exclusion step. Decay profiles of background values during the electrophoresis are also shown, with more fluorescence signal present in the electropherogram where the size exclusion procedure was not performed.

4.4. Conclusions

POP6 appears to be the most suitable matrix found for the separation and identification of near-IR heavy-atom modified dye-labeled sequencing fragments using fluorescence lifetime methods. POP6 reduces the amount of capillary preparation since no coating is required to reduce the electroosmotic flow. Also, memory effects from run-to-run can be eliminated because the matrix is replaced prior to each run. We found
a high degree of reproducibility between electrophoretic runs in terms of resolution and lifetime values for these near-IR dyes. Sample clean-up, including a size-exclusion step and cold ethanol precipitation, is necessary for the accurate lifetime identification using dye-primer reactions for base calling. The dye-terminator reactions demonstrate low background values for sequencing runs and do not require extra sample preparation.

4.5. References


Chapter 5

DNA Sequencing Using Time-Resolved Identification of Terminal Bases with Near-IR Fluorescence Detection

5.1. Introduction

Several groups have been investigating the potential use of fluorescence lifetime discrimination techniques for identifying terminal nucleotide bases in DNA sequencing using both time-resolved and phase-resolved methods [1-10]. While the fluorescence identification of the terminal bases has been traditionally achieved using spectral discrimination by developing dyes with unique emission maxima, lifetime discrimination offers some potentially unique benefits, including: (1) The elimination of cross-talk between detection channels. The emission profiles of fluorophores in spectral discrimination techniques are typically broad and as a result, spectral leakage on the four detection channels occurs, which must be software corrected. For lifetime discrimination methods, cross talk does not occur, with the accuracy of the base call determined by the precision in the measurement. (2) The calculated lifetime is immune to concentration differences present within the electrophoretic peak. Therefore, dye terminators can be employed with a variety of polymerase enzymes to accommodate the given sequencing application. Intensity differences in spectral discrimination methods have been shown to dramatically affect the read length and as such, equalization of electrophoretic peak heights by altering the dNTP to ddNTP ratio during DNA polymerization becomes necessary [11]. (3) The ability to excite the fluorescence with one source and process the emission in a single detection channel. If the dye set possesses similar absorption and emission maxima but distinct lifetimes, the hardware
needed for the readout can be simplified. In addition, the duty cycle can be improved during the fluorescence readout by eliminating filter wheels, improving the signal to noise ratio. (4) Using the appropriate dye set, uniform electrophoretic mobilities can be obtained, eliminating the need for post-electrophoretic correction. While most dyes do result in a mobility shift of the oligonucleotides to which they are attached, the shift may not only depend upon the identity of the dye, but also on the number of bases present in the oligonucleotide complicating data processing or base calling. For spectral discrimination, significant modifications in dye structures are necessary to alter the emission properties, thereby imparting large mobility shifts that are dye-dependent. With lifetime discrimination methods, subtle changes in dye structures can be used, minimizing dye-dependent mobility shifts. (5) When using time-resolved methods, time filtering can be simultaneously implemented to improve the signal-to-noise ratio in the measurement. By processing late arriving photons with respect to the excitation pulse, scattering photons and photons produced by components with short lifetimes can be effectively removed from the fluorescence signal. This feature will be particularly attractive in gel electrophoresis applications, where the gel matrix can produce high scattering backgrounds.

Wolfrum and co-workers have demonstrated the use of a four-lifetime approach to calling bases in DNA sequencing applications [1]. In their work, a series of rhodamine derivatives were prepared which possessed absorption maxima at ~630 nm and fluorescence lifetimes that varied between 1.6 and 3.7 ns. Using appropriate linker structures, dye dependent mobility shifts were minimized, eliminating the need for post-electrophoresis processing. In a single lane, four-lifetime format with capillary
electrophoresis for the separation platform, these researchers were able to demonstrate a read length of 660 bases with a probability of correct identification of the called-bases greater than 90%.

A series of near-IR fluorescent dyes (see Figure 5.1), which possess the capacity to be utilized in time-resolved identification of terminal bases in Sanger sequencing strategies using either slab gel or capillary electrophoresis were evaluated [2]. These dyes are structurally very similar (Figure 5.1), thereby, exhibiting identical absorption (765 nm) and emission (794 nm) maxima and also, uniform mobilities in gel electrophoresis, irrespective of the dye-primer linker structure. Incorporating a heavy-atom modification into the molecular framework of the base chromophore, the fluorescence lifetime of the dye series could be altered. With the strategic placement of the heavy atom with-in the dye’s structure and the ability to change the identity of the heavy-atom modification, the fluorescence lifetime could be systematically changed [12]. Incorporating a single halide (I, Br, Cl, or F) into the molecular structure, the lifetimes varied from 889 ps to 688 ps when measured in a non-polymerized acrylamide solution containing 40% formamide (see results in Chapter 4). While these differences were found to be somewhat small ($\Delta \tau_f = 70$ ps, ~8% relative difference), our data has shown that the calculated lifetime obtained during capillary gel electrophoresis of near-IR labeled oligonucleotides can be determined with high precision using simple maximum likelihood estimators [13]. The high-precision is due, in part, to the use of near-IR excitation, which minimizes background contributions into the measurement.

We will report on the use of these heavy atom modified near-IR fluorescent dyes in conjunction with a capillary electrophoresis separation platform for highly efficient
Figure 5.1. Structure of near-IR heavy-atom modified tricarbocyanine dyes developed for identification of terminal bases in DNA sequencing and IRD800 labeled with a M13mp18 (-29) sequencing primer.
base calling in DNA sequencing. Sequencing of a control M13mp18 template will be demonstrated using dye-primer and dye-terminator chemistries. Cycling conditions utilizing Taq polymerase and sample preparation methods for this dye set have been optimized and were utilized to yield intense fluorescence signatures from the electrophoresis bands to provide high precision in the measurements.

5.2. Experimental

5.2.1. Instrumentation

The instrumentation used to acquire the fluorescence lifetimes on-line during gel fractionation consisted of a conventional time-correlated single photon counting apparatus that has been previously described (see Chapter 4). The instrument (see Figure 5.2) consisted of a Ti:sapphire laser pumped by the all-lines output of a small frame Ar ion laser. The Ti:sapphire laser was tuned to 765 nm to match the excitation maxima of the dye set used for sequencing. The laser light was directed into a single mode fiber, which transported the laser light to the gel column. The use of the fiber optic minimized beam walk, which required constant realignment of the system. The light emanating from the fiber was collimated by a 10X microscope objective and focused onto the capillary tube using a single laser focusing lens (Melles Griot, Irvne, Ca). The fluorescence emission was collected in a conventional 90° format using a 60X (NA = 0.85) microscope objective (Nikon). The collected radiation was imaged onto a slit, spectrally filtered (interference band pass filter, center wavelength = 830 nm; half band width = 30 nm (Omega Optical, Brattleboro, VT) and then reimaged onto the face of a photo detector. The detector consisted of a passively quenched single photon avalanche diode SPAD, SPCM-PQ-200 (EG&G optoelectronics, Vandrieul, Canada).
The output of the SPAD was amplified 20-fold with an in-house fabricated fast amplifier and conditioned by a constant fraction discriminator (CFD). The synchronization pulse obtained from the pulsed-laser was obtained from an intracavity photodiode, which was also directed into the CFD. The time-to-amplitude converter

Figure 5.2. Near-Infrared time-correlated single-photon counting instrument. Components: L, focusing lens; C, capillary; SF, spatial filter; BF, bandpass filter; MO, objective; SPAD, single photon avalanche diode detector; DISC, discriminator; ADC, analog-to-digital converter; TAC, time-to-amplitude converter; MCA, multi-channel analyzer; PC, personal computer.
(TAC) was operated in a gated mode, with the gate pulse generated by a photocount generated from the SPAD. The start pulse was supplied by the SPAD output while the laser synchronization pulse was used as the stop pulse.

The PC contained a PicoQuant SPC-430 time-correlated single-photon-counting module. The SPC-430 provides up to 4096 time channels for each of the 128 separately measured decay curves and the channel resolution can be as small as 813 fs.

5.2.2. DNA Sequencing

A 50 µm-i.d.fused silica capillary (Polymicro, CA) with a total length of 72 cm and 42 cm effective length was used for capillary-based separations. The separation matrix was a self-coating POP6 matrix (Applied Biosystems) with urea as the denaturant run at an electric field strength of 132 V/cm. A Li-COR 4200 automated sequencer (Li-COR Biotechnology, Lincoln, NE) was used with an 8 %T (w/v) cross-linked gel (FMC Bioproducts Long Ranger, Rockland, ME) that contained 7 M urea as the denaturant and 1 X TBE (pH 8.0). The electrophoresis was typically run at –1500 V.

The two labeling dyes used for the dye-primer experiments consisted of IRD800 (Li-COR Biotechnology, Lincoln, NE) and a near-IR Br-modified tricarbocyanine dye. The NIR-Br dye was covalently attached to the 5’ end of a 17 mer, M13 forward (-29) sequencing primer through a C6 amino linker. The dye-terminators consisted of the near-IR Br-modified tricarbocyanine dye attached to a ddGTP and the near-IR Cl-modified tricarbocyanine dye attached to a ddCTP.

Sanger sequencing reactions were prepared using a modified procedure for the Amersham 7-deaza dye primer cycle sequencing kit employing an M13mp18 single-
stranded DNA template. The sequencing cocktail consisted of 4 pmol of the single-stranded DNA template, 1.0 µL of the appropriate dye-labeled primer, 15 µL of double-distilled H₂O, and 2 µL of the C or T extension mixture (Amersham Pharmacia Biotechnology, Piscataway, NJ).

Cycle sequencing was carried out in a Genius series 96-well Peltier thermal cycler (Techne, Minneapolis, MN) using the following conditions (54 cycles): (i) 92°C for 2 s; (ii) 55°C for 30 s; (iii) 72°C for 60 s, followed by a final extension step at 72°C for 7 min. The primer concentration was reduced compared to the manufacturer’s protocol and the cycle number increased to reduce the amount of unextended primer remaining in the sequencing cocktail, which minimized smearing of unextended primer in the gel tract. The reaction cocktails were then subjected to a cold ethanol precipitation that consisted of the addition of 7 µL of 7 M NaOAc and 100 µL of 100% cold ethanol. The solution was vortexed and placed in the freezer (4°C) for 30 min, followed by centrifuging for 30 min at 10,000 rpm at 4°C. The supernatant was removed by tapping the reaction tube on a counter top and then, the samples were thoroughly dried in a Centro-vap (Brinkman Instruments, Westbury, NY) for 1 h. Finally, the DNA pellet was reconstituted in distilled water, vortexed for 1 min and placed in the freezer until needed for injection.

The dye-terminator sequencing fragments were prepared with an ABI Prism dye-terminator cycle sequencing kit (PE Biosystems, Warrington, England) using a M13mp18 (-40) primer and M13mp18 double-stranded template. The sequencing cocktail consisted of 4 µl sequencing buffer, 1.25 µl template, 0.75 µl primer, 1.0 µl dNTP mix, 1.0 µl Taq polymerase, 11.25 µl of distilled water and 0.75 µl of the NIR-Cl
Figure 5.3. Structure of NIR-Cl labeled ddGTP and NIR-Br labeled ddCTP used in dye-terminator studies.
labeled ddGTP or 10 µl of distilled water and 2.0 µl of NIR-Br labeled ddCTP. The structures of these dye-terminators are shown in Figure 5.3.

Cycle sequencing was completed in a Genius series 96-well Peltier thermal cycler (Techne, Minneapolis, MN) using the following conditions (34 cycles): (i) 95°C for 30 s; (ii) 55°C for 15 s; (iii) 60°C for 4 min.

5.2.3. Data Acquisition

The data acquisition software was written in Visual Basic (Microsoft, Seattle, WA) and consisted of several control and data acquisition functions, such as streaming data to the hard drive (both intensity and time-resolved data) and providing for real-time

![Figure 5.4.](image.png)

**Figure 5.4.** Screen view of data collection and analysis program. The screen displays a single intensity line (that can be scaled) along a time axis. An intensity scale, that is color coded, and a description of the experiment are also displayed.
visualization of the acquired data. The data shown in Figure 5.4 is a false color image (intensity) for the electropherogram as a function of time. After the electrophoresis run was completed, the data was compressed and assembled into one contiguous file using zlib.dll, a freeware multiplatform compression library written in C by Mark Adler and Jean-loup Gailly. zlib.dll is a created set of library routines that provide implementation of the deflate compression algorithm and can be used as the compression engine for Zip compatible archivers.

Data analysis was directly integrated into the controlling software. The data in the compressed binary file with the pixel decay information could be displayed as an entire 2D image of both the fluorescence intensity or lifetime along the y-axis as a function of time (x-axis). After construction of the decay profile, the lifetime was calculated using equation 5.1 (see below).

### 5.3. Data Analysis

The algorithms utilized to calculate the lifetime of the molecule from the histogram employed the maximum likelihood estimator (MLE) using the following equation [13, 14]:

\[
1 + \left( e^{\frac{T}{\tau}} - 1 \right)^{-1} - m\left( e^{\frac{mT}{\tau}} - 1 \right) - 1 = N^{-1}_T \sum_{i=1}^{m} i N_i \tag{5.1}
\]

where \( m \) is the number of time bins, \( N \) is the number of photocounts in the decay spectrum, \( N_i \) is the number of photocounts in time bin \( i \), \( T \) is the time width of each bin, and \( \tau \) is the lifetime.

The right hand side of the equation is calculated from experimental values and the left-hand side describes a monotonic, non-decreasing function of two variables (\( m, \tau \)).
(see Figure 2.9). For the chosen number of time bins, m, a binary search for the calculation of $\tau$ in the left-hand side of the equation was performed. The binary search for $\tau$ begins by selecting a region of possible $\tau$'s from 10 to 40,000 ps. The left-hand side of the equation is calculated at the ends of the region as well as in the middle of the region. Depending on the relation to the right hand side value, the appropriate half of the region is chosen and the calculations are repeated. In each step, the region of possible values of t satisfying the equation is narrowed by a factor of two. Computation continues until the search region of lifetimes is less than a predefined value, typically ~0.1 ps.

5.4. Automated Base Identification in Sequencing Trace

For the number of peaks acquired in a run, automatic peak recognition is required due to the large amount of data. Our algorithm for peak recognition is not based upon phred scoring, since all bases are not analyzed simultaneously and we cannot assume evenly spaced peaks [15]. For the given electrophoretic data, the detection at the beginning and end of each peak is based upon a defined threshold level and value of a slope (see Figure 5.5). The algorithm for automatic recognition makes the decision about the beginning and end of the peak based on two criteria derived from the intensity profile. After visual inspection of the electropherogram, a predefined value of the threshold and slope are set. To qualify as a beginning or end, the intensity has to exceed the threshold value and at the same time, the slope of the peak has to be greater than a predefined value. When the start and stop of a peak are identified, an additional analysis of the sign of the slope is made and based upon the analysis, possible sub-peaks can be identified (see Figure 5.6). A sub-peak occurs when the peaks are not
Figure 5.5. Display used to set beginning and end time interval (in channel numbers) to search for peaks in electropherogram. Threshold and minimum slope values are also set for the peak recognition algorithm. Actual lifetime values can be displayed in the peak recognition or set values that are assigned to each base, including a value for unrecognized bases.

Figure 5.6. Base analysis of electropherogram of near-IR dye-labeled ddGTP and ddCTP fragments. The slope was set for 1000 and the threshold set at 20,000 for peak picking analysis. The analysis produced two peaks with three and two sub-peaks, respectively. The details of the peaks and sub-peaks are listed as well as the calculated lifetime of the points within each peak and the base association of the lifetime value.
baseline resolved according to the preset threshold value. For each electrophoretic sub-
peak, the decays are constructed and MLE analysis performed. If the calculated $\tau$ falls
within $3\sigma$ of a predefined $\tau$, the sub-peak is classified and called as a base: A, C, G, or
T. If the calculated $\tau$ falls outside of these ranges, the peak is marked with “???” so that
further analysis can be carried out, mainly a pixel-by-pixel analysis of the decays to
assist in dye identification via pattern recognition of the resulting lifetime versus
electrophoresis time plot.

**Figure 5.7.** Display screen used to choose MLE calculation process and to set the
lifetime values and base associations used in peak recognition. The screen is set for
values used by the peak recognition for the analysis of the two-lifetime dye-terminator
electropherogram. While the program is currently built for two-dye runs, it can easily be
expanded to four-dye runs.
Figure 5.8. Display screen for lifetime analysis of electrophoresis file. The electropherogram is shown in the top portion of the screen. Using the computer mouse, a single pixel or multiple pixels from the electropherogram can be selected to form a decay profile that is displayed in the lower panel. The decay profile can be analyzed by selecting the channels over which the MLE calculation is to be performed. The calculated value is shown in the bottom left corner.

The reference $\tau$ values for the peak recognition are set in advance of the peak recognition and base identification along with the association of a particular lifetime value to the appropriate base (see Figure 5.7). The program can also be set to calculate the lifetime of all points in the data file using the MLE with a set number of bins (channel numbers) or with different confidence levels to detect the last bin where relevant lifetime data is accumulated.
A pixel-by-pixel analysis of the electropherogram can also be done with the analysis program. From a separate screen (Figure 5.8), sections of the electropherogram can be enlarged and selected for lifetime analysis. Single pixels or multiple pixels may be selected from the electropherogram to construct a decay profile. The time bins over which the MLE calculates the lifetime are selected and the MLE value displayed. The pixel-by-pixel analysis is necessary for peaks that overlap (i.e. poor electrophoretic resolution), since the MLE can only calculate a single value, even for multi-component decays.

5.5. Results and Discussion

To assess the relative precision in the lifetime measurements, single base tracts of the two dye-primers (NIR-Br and IRD800) labeled with T and the dye-terminators (NIR-BR and NIR-CI) were separately electrophoresed in a capillary with POP6 as the separation medium. Using equation 5.1, the lifetimes of several bands in the dye-primer and dye-terminator tracts were determined. For the dye-primer runs, the average lifetime value of the NIR-Br T-tract was found to be 810 ps with a standard deviation of 15 ps (RSD = 1.8%), while IRD800 had a calculated lifetime value of 598 ps with a standard deviation of 8 ps (RSD = 1.3%). For the dye-terminators, the average lifetime value of NIR-Br ddCTP was 780 ps with a standard deviation of 10 ps (RSD = 1.3%) and for NIR-CI ddGTP, \( \tau_f \) was 736 ps with a standard deviation of 6 ps (RSD = 0.8%). The decay profiles and instrument response functions for the dye-primer and dye-terminators are shown in Figure 5.10.

The dye-primer run was carried out using NIR-BR-T and IRD800-C tracts to evaluate the ability to accurately identify the calculated lifetime values. Using the
Figure 5.9 Instrument response function and decay profiles for IRD800, NIR-Br, ddGTP-NIR-Cl, and ddCTP-NIR-Br. Measurements taken from single base tracts run in POP6.
Figure 5.10. Graph of IRD800-C and NIR-Br-T tracts injected for 1 min at 110 V/cm and run at 138V/cm in POP6. (A) Calculated lifetime values determined by peak identification software. (B) Assigned values associated to bases by peak identification software: 1 = IRD800-C, 2 = NIR-Br-T, 2.5 = unknown. After a pixel-by-pixel analysis of the third peak shown which falls into the unidentified category, it was found that the peak was two unresolved bands. The base call for this region was determined to be C C C T T.
two-dye electropherogram, the peak identification algorithm correctly identified the majority of bases in the run and gave appropriate identification for peaks that did not fall within $3\sigma$ of the range (see Figure 5.10). The peak identification analysis calculates the actual lifetime for the peaks and assigns a constant value to each base for facile calling of the bases. The assigned value also alerts the end user to problem areas (ie. poor electrophoretic resolution regions) for further analysis to be performed. For the 201 T and C bases the lifetime based identification accuracy was 97.5%, with 5 miscalls. All miscalls were attributed to complete overlap of the two base tracts from mobility differences between the dye-primers.

The dye-terminator capillary electrophoresis run was compared to a slab gel image of a one-color/four-lane detection protocol of the M13mp18 template. The sequence obtained from the slab-gel automated call was confirmed from the known sequence of the M13mp18 phage and the known priming site. The two-dye terminator read from the capillary was analyzed with our peak recognition algorithm, but many of the bands were poorly resolved and recognized as a single peak, biasing the calculated lifetime result. The averaged lifetime resulting from overlapped peaks found by the peak recognition process highlighted problem areas, because it produced a $\tau_f$ greater than $3\sigma$, from $\tau_f$ for both the ddCTP and ddGTP tracts. Therefore, a pixel-by-pixel analysis needed to be completed to assist in base identification.

The resulting read-length of Cs and Gs was 276. Lifetime identification of the ddCTP/ddGTP tract run in the capillary resulted in a total of 9 miscalls, most of which resulted from overlapping electrophoretic peaks. The resulting read accuracy for the two-dye capillary run was 96.8%. Figure 5.11 shows a section of the electropherogram
with the bases and miscalls identified. Improving the resolution of the capillary-based dye-terminator run would improve the read accuracy.

Figure 5.11. Electropherogram of two-dye-terminator capillary run. Automated base calling was employed and followed by a pixel-by-pixel analyses in areas of poor electrophoretic resolution. Two miscalls resulting from poor electrophoretic resolution are noted by arrows. Run in POP6 at 135 V/cm.

5.6. Conclusions

We have been able to use dye-labeled primers for accurate lifetime identification of sequencing fragments in a capillary-based system. Also, a two dye-labeled terminator read has been accurately called with lifetime identification methods with a relatively small difference between \( \tau_f \) values. A peak recognition algorithm has been developed to aid in identification of bases with lifetime analysis.
5.7. References


Chapter 6
Conclusions and Future Work

6.1. Document Summary

The focus of this work was to prepare near-IR dye-labeled DNA fragments for sequencing with a time-correlated single photon counting detection system and subsequent analysis of the dyes to reconstruct the DNA sequence with lifetime identification procedures using the MLE method.

In Chapter 1, the background literature and history of DNA sequencing procedures were outlined. The principles of electrophoretic separations in slab and capillary electrophoresis were discussed as well as the matrices and models used for size fractionating DNAs. Finally, several different detection methods and sequencing formats were reviewed.

In Chapter 2, a thorough explanation of time-correlated single photon counting was given. This chapter included a description of the instrumentation necessary for making TCSPC measurements along with an overview of lifetime calculation methods. Lifetime-based sequencing was introduced and the advantages of detecting in the near-infrared region were discussed.

In Chapter 3, a slab-gel automated sequencer was fitted with a TCSPC scanning and detection system for lifetime analysis of a two-dye/two-lane sequencing format. Using two near-IR dye-primers, the bases were identified with an accuracy of 99.7% with 2 miscalls in 670 bases, compared to a 95.7% accuracy for a single-color/four lane approach. Monte Carlo results were shown to be an accurate representation of lifetime
distributions for poorly resolved peaks. By performing a pixel-by-pixel analysis, bases could be identified in these cases.

In Chapter 4, the sequencing of DNA and identification of near-infrared tricarbocyanaine dyes were investigated. Sequencing sample preparation methods were optimized to yield intense signals. Separation matrices were investigated to give good resolution of sequencing fragments and large differences between lifetime values of near-IR dyes.

In Chapter 5, a capillary-based separation of near-IR sequencing fragments in a two-dye format was completed. A peak recognition algorithm was introduced to aid in identification of bases using the MLE for lifetime calculations. A two-lifetime dye-primer separation of 201 bases was done and analyzed with the automated peak recognition with an accuracy of 97.5\% with all miscalls resulting from completely overlapped peaks. Near-IR dye-terminators were used in a two-dye experiment and yielded 96.8\% accuracy for 276 bases with automated lifetime identification.

6.2. Future Work

6.2.1. New Dyes

The heavy-atom modified tricarbocyanine dyes developed for the four lifetime sequencing strategy have many weaknesses, such as poor photochemical stability, heat sensitivity, relatively short lifetimes, and closely spaced lifetime values. Two new sets of near-IR dyes for lifetime-based sequencing are being prepared that are chemically robust, phthalocyanine and naphthalocyanine water-soluble dyes with metal centers to alter the lifetime values (see Figure 6.1). These dyes can be exposed to light for long periods of time with no sign of degradation and are also heat stable. The lifetime value
of these dyes can range from 1.5 to 4 ns, depending on the metal center. The phthalocyanine dye series can be excited at 683 nm and detected at 700 nm. Detection of the naphthalocyanine is at 780 nm and excitation is 735 nm. Just as in the heavy-atom modified tricarbocyanine dye series, the phthalocyanine and naphthalocyanine dyes will have identical mobilities to eliminate post electrophoresis mobility shift corrections. The higher level of chemical stability and the relatively longer lifetimes of these dye sets will provide a stable base for single-lane/four-lifetime sequencing in the near-infrared.

Figure 6.1. Structure of water-soluble phthalocyanine and naphthalocyanine dyes for DNA sequencing. The metal center alters the lifetime of the dye.
6.2.2. Microchip Analysis

Although capillary electrophoresis is the “gold standard” in sequencing technology for high-throughput analysis, microchip technology can speed up the sequencing process and consume lower amounts of reagents needed for sequencing, reducing the cost [1]. Capillaries must either be covalently coated with a polymer or the separation matrix must be a self-coating polymer to reduce the EOF in DNA sequencing applications. Utilizing polymethylmethacralate (PMMA) microchips for separations, the EOF is not a factor since the mobility of DNA exceeds the EOF created by the PMMA in an electric field [2].

![PMMA microchips for separations are smaller than a ball-point pen. The walls of the PMMA chip are smooth and straight. The hot embossed PMMA microchip has a channel width of 20 \( \mu \)m, a channel depth of 100 \( \mu \)m, a separation channel length 4.2 cm, with an injection volume of 40 pL.]

Another advantage of PMMA over glass, the traditional material for microchips, as a substrate is that the devices can be “stamped” into the polymer with straight channels (Figure 6.2) using hot embossing techniques instead of wet-chemical etching.
of glass, which produces slanted channels. Hot embossing of microchips allows for automation of chip production, reducing the cost of micro-devices.

![Near-IR microscope assembly](image)

**Figure 6.3.** Near-IR microscope assembly for lifetime analysis of DNAs in PMMA microchips. The single photon avalanche diode has an active area of 500 μm with a dynamic range from 3 to 900 K cps. The laser, which can be 680 nm or 780 nm excitation, has a spot size of 25 x 40 μm and a repetition rate of 80MHz.

Using the microscope assembly constructed for slab gel analysis documented in Chapter 3 (Figure 6.3), lifetime measurements of dye-primers and sequencing ladders in PMMA microchips have begun. Preliminary studies indicate that differences in lifetimes values of common dyes can be distinguished in the plastic microchips, see Figure 6.4. Two dye-primers were injected, IRD800 and IRD40, into the separation channel and electrophoresed at a field strength of 150 V/cm. The primers were separated and the resulting lifetime values could be discriminated.
Figure 6.4. Separation of IRD800 and IRD40 dye-primers in 1.0% hydroxyethyl cellulose (HEC) in TAE (pH = 8.2). Injection volume of 40 pL was electrophoresed at a field strength of 150 V/cm.

6.2.3. Two-Color Detection

Finally, the lifetime base calling can be combined with color discrimination to implement a two-color, four-lifetime/ single-lane format. This multiple-color/multiple-lifetime identification technique will allow eight dyes to be run and identified in a single separation lane, which will double the throughput of any separation platform. A two-color set-up is shown in Figure 6.5. Two diode lasers, one lasing at 680 nm and the
Figure 6.5. Schematic of a two-color detection instrument.

other at 780 nm are coupled through a fiber optic and focused onto the separation platform (capillary or microchip) and the fluorescence emission collected and spectrally separated by dichroic mirrors before being focused onto a SPAD. Some of the useful applications include screening for multiple mutations in a single run or sequencing both directions on double-stranded DNA, which can aid in gap closing for shotgun sequencing projects.
6.3. References


Vita

Suzanne Jeanel Lassiter was born to George E. Lassiter and Judy A. Lassiter on October 31, 1975, in Thomasville, Georgia. She has one sister, Michelle. From 1979-1982, she attended St. Thomas Episcopal School in Thomasville, Georgia. From 1982-1990, she attended elementary and middle schools in Thomasville, Georgia. From 1990-1994, she attended Thomasville High School in Thomasville, Georgia.

Upon graduation, she enrolled at Georgia Southwestern State University in June, 1994, on a Alumni Scholarship. While attending Georgia Southwestern, she was a member of Gamma Beta Phi and Chemistry Club. She graduated cum laude with a bachelor of science degree in chemistry in December 1996.

In January 1997, she began graduate school at Louisiana State University in the Department of Chemistry. In March 2001 she received the James Robinson award for excellence in Analytical Chemistry.