

8 Sequencing Technology

LYLE R. MIDDENDORF

PATRICK G. HUMPHREY

NARASIMHACHARI NARAYANAN

STEPHEN C. ROEMER

Lincoln, NE, USA

- 1 Introduction 185
- 2 Overview of Sanger Dideoxy Sequencing 185
- 3 Fluorescence Dye Chemistry 186
 - 3.1 Fluorophore Characteristics 187
 - 3.2 Commercial Dye Fluorophores 187
 - 3.3 Energy Transfer 189
 - 3.4 Fluorescence Lifetime 190
- 4 Biochemistry of DNA Sequencing 192
 - 4.1 Sequencing Applications and Strategies 193
 - 4.1.1 New Sequence Determination 193
 - 4.1.2 Confirmatory Sequencing 194
 - 4.2 DNA Template Preparation 195
 - 4.2.1 Single-Stranded DNA Template 195
 - 4.2.2 Double-Stranded DNA Template 195
 - 4.2.3 Vectors for Large-Insert DNA 195
 - 4.2.4 PCR Products 196
 - 4.3 Enzymatic Reactions 196
 - 4.3.1 DNA Polymerases 196
 - 4.3.2 Labeling Strategy 197
 - 4.3.3 The Template–Primer–Polymerase Complex 197
 - 4.3.4 Simultaneous Bidirectional Sequencing 198
- 5 Fluorescence DNA Sequencing Instrumentation 198
 - 5.1 Introduction 198
 - 5.1.1 Excitation Energy Sources 199
 - 5.1.2 Fluorescence Samples 199

- 5.1.3 Fluorescence Detection 199
- 5.1.4 Overview of Fluorescence Instrumentation Related to DNA Sequencing 200
- 5.2 Information Throughput 201
 - 5.2.1 Sample Channels (n) 201
 - 5.2.2 Information per Channel (d) 202
 - 5.2.3 Information Independence (I) 202
 - 5.2.4 Time per Sample (t) 202
- 5.3 Instrument Design Issues 203
 - 5.3.1 Laser Excitation and Fluorescence Emission 204
 - 5.3.2 Detector Signal 204
 - 5.3.3 System Noise 205
- 5.4 Commercial Electrophoresis Embodiments for Fluorescence DNA Sequencing 207
 - 5.4.1 Slab Gels 207
 - 5.4.2 Capillary Gels 207
 - 5.4.3 Micro-Grooved Channel Gel Electrophoresis 209
- 5.5 Non-Electrophoresis Embodiments for Fluorescence DNA Sequencing 209
- 5.6 Non-Fluorescence Embodiments for DNA Sequencing 210
- 6 DNA Sequence Analysis 210
 - 6.1 Introduction 210
 - 6.2 Lane Detection and Tracking 211
 - 6.3 Trace Generation and Base Calling 212
 - 6.4 Quality/Confidence Values 215
- 7 References 216

1 Introduction

DNA sequencing technology is a major component of the genomics discovery pipeline. The technology is rooted in the late 1960s and early 1970s when efforts to sequence RNA took place. The nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli* (BROWNLEE et al., 1967), 16S- and 23S-ribosomal RNA (FELLNER and SANGER, 1968), and R17 bacteriophage RNA coding for coat protein (ADAMS et al., 1969) are some of the early examples of RNA sequencing. A few years later SANGER reported on the sequencing of bacteriophage ϕ 1 DNA by primed synthesis with DNA polymerase (SANGER et al., 1973, 1974). In the same time period GILBERT and MAXAM (1973) reported on the DNA nucleotide sequence of the *lac* operator.

This pioneering work led to the plus/minus method reported by SANGER and COULSON (1975) which determined nucleotide sequence based on two approaches: (1) a “minus” system where four separate samples of partially double-stranded DNA fragments (containing a “full length” template “–” strand and random chain extension of an oligonucleotide primer for the “+” strand) are further incubated with DNA polymerase in the presence of only three deoxyribonucleoside triphosphates such that synthesis proceeds as far as it can until the polymerase needs to incorporate the missing nucleotide of the particular sample; and (2) a “plus” system where the four separate samples of partially double-stranded DNA fragments are further incubated in the presence of only one of the four triphosphates and then subjected to exonuclease which degrades the single-stranded overhang of the “–” strand. The DNA fragments for both approaches were then subjected to gel electrophoresis for length (and thus sequence) determination.

In 1977 SANGER reported on the use of modified nucleoside triphosphates (containing dideoxyribose sugar) in combination with the natural deoxyribonucleotides to terminate chain elongation (SANGER et al., 1977). In that same year MAXAM and GILBERT (1977) disclosed a method for sequencing DNA that utilized chemical cleavage of DNA preferentially at guanines, at adenines, at cytosines and thy-

mines equally, and at cytosines alone. These two methods accelerated manual sequencing based on electrophoretic separation of DNA fragments labeled with radioactive markers and subsequent detection via autoradiography.

The first reports of automation of DNA sequencing occurred in the mid-1980s due to novel techniques to fluorescently label DNA (SMITH et al., 1986; ANSORGE et al., 1986, 1987; PROBER et al., 1987; BRUMBAUGH et al., 1988; KAMBARA et al., 1988; MIDDENDORF et al., 1988). This automation, in conjunction with the commencement of the human genome initiative (DELISI, 1988), spurred the explosion in genomics research that is in existence today. DNA sequencing technology is now only one tool, albeit a very important and dynamic one, in the genomics toolbox along with other tools such as DNA array and lab-on-a-chip technologies as well as automated protein analysis.

This chapter illustrates the multi-disciplinary nature of DNA sequencing technology in that the chapter organization is delineated into chemistry, biology, instrumentation, and software components. It is intended to provide an exhaustive reference structure in order to allow further in-depth investigation of each of these components, and the reader is invited to take advantage of the reference list in order to capture the fuller essence of sequencing technologies.

2 Overview of Sanger Dideoxy Sequencing

DNA sequencing is the determination of the nucleotide sequence of a specific deoxyribonucleic acid (DNA) molecule. Knowing the sequence of a DNA molecule is pivotal for making predictions about its function and facilitating manipulation of the molecule. Originally, DNA was sequenced using one of two methods. MAXAM and GILBERT (1977) devised a method that chemically cleaves DNA selectively between specific bases. SANGER et al. (1977) developed an enzymatic method based on the use of chain-terminating dideoxynucleotides.

The Sanger dideoxy method is now by far the most widely used technique for sequencing DNA. Informative texts by ALPHEY (1997) and ANSORGE et al. (1997) review many variations made to this sequencing technique, but the principle remains the same. The method depends on the synthesis of a new strand of DNA starting from a specific priming site and ending with the incorporation of a chain terminating nucleotide.

Specifically, a DNA polymerase extends an oligonucleotide primer annealed to a unique location on a DNA template by incorporating deoxynucleotides (dNTPs) complementary to the template. Synthesis of the new DNA strand continues until the reaction is randomly terminated by the inclusion of a dideoxynucleotide (ddNTP). These nucleotide analogs are incapable of supporting further chain elongation since the ribose moiety of the ddNTP lacks the 3'-hydroxyl necessary for forming a phosphodiester bond with the next incoming dNTP. This results in a population of truncated sequencing fragments of varying length.

Typically, the identity of the chain-terminating nucleotide at each position is specified by running four separate base-specific reactions each of which contains a different dideoxynucleotide (ddATP, ddCTP, ddGTP, or ddTTP). The four such fragment sets are loaded in adjacent lanes of a polyacrylamide gel and separated by electrophoresis according to the fragment size (Fig. 1). Remarkably, DNA fragments differing in length by just one nucleotide can be resolved. If a radioactive label is introduced into the sequencing reaction products, then autoradiographic imaging of the DNA band pattern in the gel can be used to deduce the DNA sequence (SANGER et al., 1977; SMITH, 1989). If the reaction products are labeled with an appropriate fluorescent dye, then an automated DNA sequencing system is used for the real-time detection of DNA fragments as they move through a portion of the electrophoresis gel that is irradiated by a laser. The fluorescence emission is collected by a detector and the resultant signal produces a band or trace pattern which correlates to a DNA sequence.

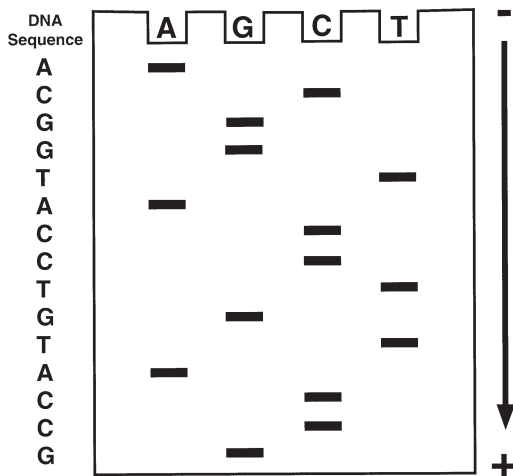


Fig. 1. DNA sequencing electrophoresis. The DNA fragments are prepared to terminate at one of four base types (A, G, C, T). A-type fragments of varying length are loaded in the "A" loading well at the top of the gel, and so forth for the G-, C-, and T-type fragments. Over time the shorter fragments in each lane migrate farther down the gel (toward the positive electrode). The DNA sequence is determined by noting the particular lane in which each succeeding band is spatially located in the vertical dimension (taken from MIDDENDORF et al., 1993).

3 Fluorescence Dye Chemistry

The original methods of DNA sequencing (MAXAM and GILBERT, 1977; SANGER et al., 1977), were implemented through the use of radioactive labels. High sensitivity and ease of labeling still make radioactive methods popular in thousands of biology laboratories around the world that practice manual radioactive DNA sequencing. However, the dangers associated with radioactivity such as health hazards and waste disposal regulations, along with the lack of automation, paved the way for the emergence of alternative non-radioactive labels (KESSLER, 1992). Most prominent among the sensitive, non-radioactive detection techniques are chemiluminescence and fluorescence. Despite excellent sensitivity, chemi-

luminescent methodology has not been viable for DNA sequencing due to its indirect detection limitation. Fluorescence detection (LAKOWICZ, 1999), on the other hand, employs direct detection methodology that is simple, sensitive, and easy to automate. Fluorescence methods and fluorescent dye labels have set a new standard in today's DNA sequencing community.

Several methods have been developed to sequence DNA using fluorescent labels (ANSORGE et al., 1986; SMITH et al., 1986; PROBER et al., 1987; BRUMBAUGH et al., 1988). Commercialized instruments employ one or more of the following methods for automated sequencing: four distinct dye-labeled primers with non-fluorescent terminators per DNA sample; one dye-labeled primer with non-fluorescent terminators per DNA sample; and one non-fluorescent primer with four distinct fluorescent terminators per DNA sample (see Sect. 4). This section provides a brief summary of important aspects on the advancement of chemistry of fluorescent dyes for DNA sequencing.

3.1 Fluorophore Characteristics

Fluorescence is the emission of light from electronically excited fluorophores. An electron of the fluorophore is energized into an excited orbital through the absorption of a photon where it is paired to a second electron that is in the ground-state orbital (LAKOWICZ, 1999). The excited orbital is one of several vibrational energy levels associated with one or more electronic energy states. The fluorophore is usually excited into a higher vibrational level of either the first or second electronic energy state. In a very fast process known as internal conversion the excited molecule first relaxes to the lowest vibrational level of the first electronic energy state. This is followed by a relaxation to a higher excited vibrational ground-state level with the emission of a photon. Because of the multiplicity of vibrational levels as well as electronic levels, the spectra of both absorption and emission are polychromatic and generally are mirror images of one another.

Both the absorption and emission spectra of the fluorophore depend on its chemical struc-

ture as well as the environment (solvent, pH, temperature, etc.) of the fluorophore. The spectral wavelength of fluorescence emission is generally independent of the excitation wavelength of the absorbed photons. However, because of the rapid initial non-radiative decay associated with internal conversion as well as the final decay to higher vibrational levels of the ground state, the energy of the emitted photon is less than that of the absorbed photon. This shifts the fluorescence spectra to longer wavelengths relative to the absorption spectra and is known as the Stokes' Shift (STOKES, 1852).

3.2 Commercial Dye Fluorophores

The physiological response of the human eye qualitatively defines the visible wavelength region (in nanometers or nm) of the electromagnetic spectrum. Wavelengths shorter than, but adjacent to that of the visible region, are identified as ultraviolet. Wavelengths longer than, but adjacent to that of the visible region, are identified as near infrared. The commercialized fluorescent labels that are currently in use in automated DNA sequencing are either visible dyes (450–600 nm absorption and fluorescence range) or near infrared dyes (650–860 nm absorption and fluorescence range).

The first commercialized near infrared dyes introduced for automated DNA sequencing were IRDye41 and IRDye40 (STREKOWSKI et al., 1992; NARAYANAN et al., 1995; SHEALY et al., 1995) (Fig. 2). These dyes are from the heptamethine carbocyanine dye family and nominally absorb and fluoresce near 800 nm. IRDye41 was attached to a DNA primer via a stable thiourea linkage formed by conjugating the dye to an amino linker located at the 5' end of the primer. A phosphoramidite version (IRDye800; Fig. 3) (NARAYANAN et al., 1998) provides for direct labeling of DNA primers using an automated DNA synthesizer. For dye labeled terminator chemistry, the IRDye800 is attached to bases which are linked to a triphosphate through an acyclo bridge (Fig. 4). The incorporation of this substrate terminates DNA chain elongation in a manner similar to that obtained by using dideoxynucleotides

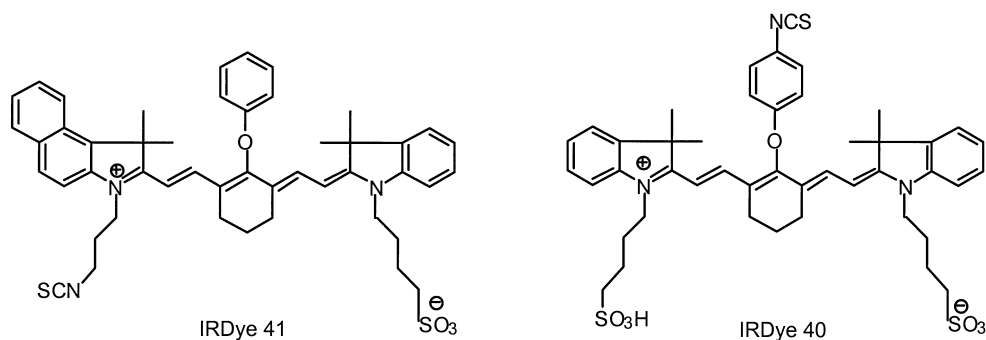


Fig. 2. Structures of IRDye41 and IRDye40. Both dyes are members of the polymethine carbocyanine dye family which is characterized by two heteroaromatic residues connected by a conjugation bridge of polyethylene units. The length of the conjugating bridge affects the absorbance and fluorescence maxima (MATSUOKA, 1990). IRDye41 and IRDye40 are heptamethine carbocyanine dyes which contain seven carbons in their conjugating bridge. The isothiocyanate (NCS) reactive functionality is used to couple the dye to a primary amine which results in a thiourea linkage.

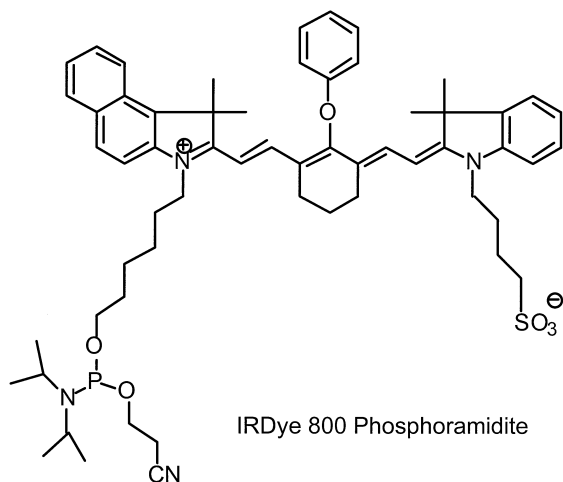


Fig. 3. Structure of IRDye800 Phosphoramidite. The amidite functionality is used to couple the dye to the 5'-OH of the 5' terminus nucleotide of an oligonucleotide via automated DNA synthesis. See Fig. 2 legend for additional information.

(see Sect. 4). Dye properties for IRDye40, IRDye 41, and IRDye800 are listed in Tab. 1.

Commercialized near-infrared dyes that absorb and fluoresce around 650–700 nm are from the pentamethine carbocyanine dye family. They include IRDye700 (NARAYANAN et al., 1998), Cy5 (MUJUMDAR et al., 1989, 1993; ZHU et al., 1994), and Cy5.5 (TU et al., 1998). Dye properties for IRDye700, Cy5, and Cy5.5 are listed in Tab. 1 and the structures are shown in Fig. 5.

Shown in Fig. 6 are two fluorescein dye derivatives (FAM, JOE) and two rhodamine dye

derivatives (TAMRA and ROX) first in use for four visible dye primer-based DNA sequencing. Fluorescein dye has also been used in single dye sequencers (ALF DNA Sequencer, Pharmacia Biotech). Shown in Fig. 7 are two rhodamine dyes (R110, R6G) which are combined with TAMRA and ROX for use in four visible color dye terminator-based DNA sequencing. Dye properties for FAM, JOE, TAMRA, ROX, R110, and R6G are listed in Tab. 1.

In order to give more even and narrower peak heights than the rhodamine dye termina-

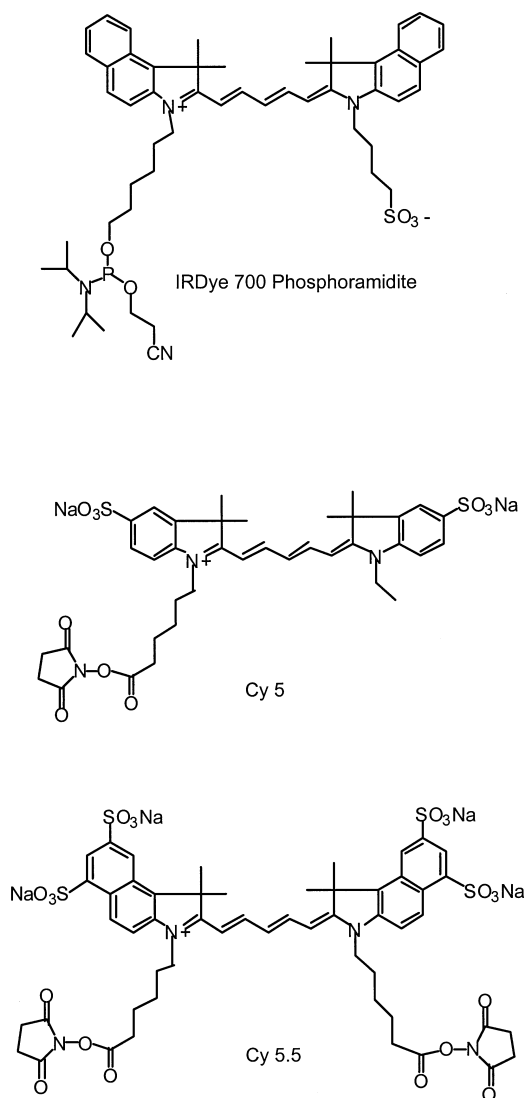


Fig. 5. Structures of IRD700 (phosphoramidite functionality), Cy5 (succinimidyl ester functionality), and Cy5.5 (bis-succinimidyl ester functionality). All three dyes are members of the pentamethine carbocyanine dye family which is characterized by five carbons in the conjugating bridge (see Fig. 2 legend).

An approach that involves energy transfer in labeled primer chemistry uses the oligonucleotide backbone to separate the donor and acceptor dyes (JU et al., 1995a, b, 1996;

HUNG et al., 1996a, b, 1997, 1998; METZKER et al., 1996). Another approach uses tethered donor and acceptor dyes for either labeled primers (LEE et al., 1997) or labeled terminators (ROSENBLUM et al., 1997). These tethered dyes use fluorescein as a donor dye and one of the four dRhodamine dyes (see Sect. 3.2) as an acceptor dye and are linked through 4-amino-methyl benzoic acid. The structure of a tethered fluorescein/dR110 is shown in Fig. 9.

3.4 Fluorescence Lifetime

When a fluorophore emits light as it relaxes from an excited energy state to a ground energy state, such relaxation occurs after the molecule has spent a certain amount of time in the excited state (see Sect. 3.1). The average time spent in the excited state is known as the fluorescence lifetime of the molecule (LAKOWICZ, 1999) and it is statistically the same for all molecules having the same structure and exposed to the same environmental conditions. A common characteristic (although not necessarily assumable) is that the statistical relaxation of a fluorophore follows an exponential decay profile when examined over several excitation/relaxation cycles. For this case, the fluorescence lifetime is then specified as the exponential time constant where 63% of the relaxations occur more quickly than this lifetime average and 37% occur more slowly.

The lifetime of common visible and near infrared fluorophores ranges from 0.5–4 nanoseconds and is dependent on their chemical structure. The ability to discriminate among fluorophores is impacted by the ratio of their lifetimes as well as the number of photons available to produce the composite lifetime profile histogram (KÖLLNER and WOLFRUM, 1992; KÖLLNER, 1993).

The use of energy transfer to allow common excitation for multiple dyes has been successfully commercialized (see Sect. 3.3). As researchers examine alternative approaches for facilitating common excitation as well as increasing the number of available dye choices, the exploitation of fluorescence lifetime discrimination for DNA sequencing shows promising potential because a fluorophore's lifetime is independent of concentration and mul-

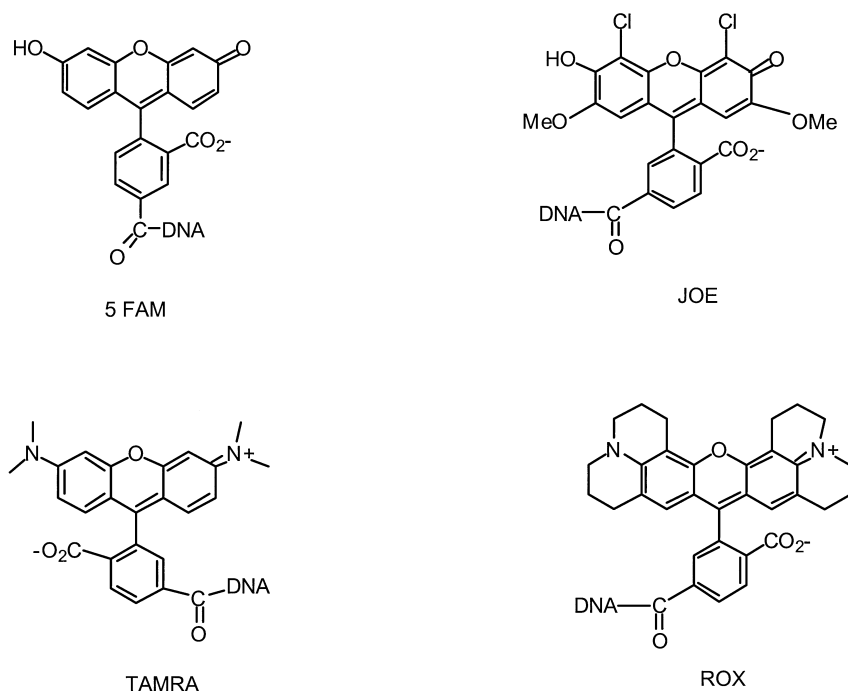


Fig. 6. Structures of dyes FAM, JOE, TAMRA, and ROX. FAM and JOE are members of the fluorescein family while TAMRA and ROX are members of the rhodamine family. All four dyes must be purified from isomers that contain alternate sites for the reactive functionality which ultimately couples the dye to DNA. Shown are the 5-isomer for FAM and the 6-isomer for JOE, TAMRA, and ROX.

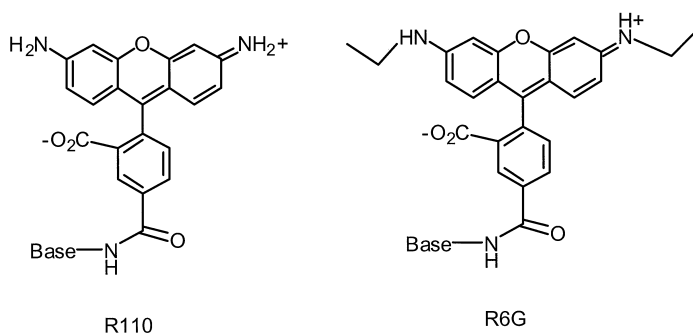


Fig. 7. Structures for R110 and R6G dyes. Both dyes are members of the rhodamine family. Shown are the 5-isomers.

tiple dyes having overlapping spectral emission can be distinguished (CHANG et al., 1993; HAN et al., 1993; SAUER et al., 1994; LEGENDRE et al., 1996; SOPER et al., 1996; MÜLLER et al.,

1997; NUNNALLY et al., 1997; FLANAGAN et al., 1998). Methods for “on-the-fly” lifetime measurements of labeled DNA fragments have been described for capillary electrophoresis

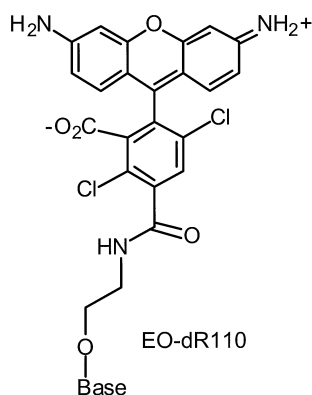


Fig. 8. Structure for dichloro-R110 dye linked to a nucleotide base. See text for the effects of adding the two chlorides to R110 (shown in Fig. 7). Similar dichloro modifications have been made to dyes TAMRA, ROX, and R6G (shown in Figs. 6 and 7).

embodiments (LI and MCGOWN, 1996; LI et al., 1997) and slab gel embodiments (LASSITER et al., 2000). Besides enabling common excitation, fluorescence lifetime discrimination also permits the use of common spectral detection optics. Both spectral and lifetime discrimina-

tion can be combined in a single design embodiment to take advantage of the strengths of each approach (PRUMMER et al., 2000; LASSITER et al., 2000).

4 Biochemistry of DNA Sequencing

The efficient completion of large DNA sequencing projects is now a reality due in great part to the development of fluorescence-based dideoxynucleotide sequencing chemistries coupled with instrumentation for real time detection of dye-labeled DNA fragments during gel electrophoresis (see Sect. 5). The commercially available automated sequencers (Sect. 5, Tabs. 2 and 3) can be divided into two groups based on the number of fluorescent dyes used in a sequencing reaction.

The first type uses the one-dye/four-lane approach in which the identity of the chain-terminating nucleotide at each position is determined by running four separate reactions each of which contains the same fluorescent

Tab. 1. Dye Absorption and Emission Properties (Aqueous Environment) for Several Commercial Dyes Available for DNA Sequencing. Absorption and Emission Maxima are Approximate and may be Dependent on Solvent, Solvent Properties (e.g., pH), and the Biomolecule to which they are Attached

Dye	Absorption Max	Emission Max	Dye Family
FAM	490–495 nm	515–520 nm	fluorescein
R110	500–505 nm	525–530 nm	rhodamine 110
dR110	NA	530–535 nm	rhodamine 110
JOE	520–525 nm	550–555 nm	dichlorodimethylfluorescein
R6G	525–530 nm	555–560 nm	rhodamine 6G
dR6G	NA	560–565 nm	rhodamine 6G
TAMRA	550–555 nm	580–585 nm	tetramethylrhodamine
dTAMRA	NA	590–595 nm	tetramethylrhodamine
ROX	580–585 nm	605–610 nm	X-rhodamine
dROX	NA	615–620 nm	X-rhodamine
Cy5	650–655 nm	665–670 nm	pentamethine carbocyanine
Cy5.5	670–675 nm	690–695 nm	pentamethine carbocyanine
IRDye 700	685–690 nm	710–715 nm	pentamethine carbocyanine
IRDye40	765–770 nm	785–790 nm	heptamethine carbocyanine
IRDye41	795–800 nm	820–825 nm	heptamethine carbocyanine
IRDye800	795–800 nm	820–825 nm	heptamethine carbocyanine

NA, information not available

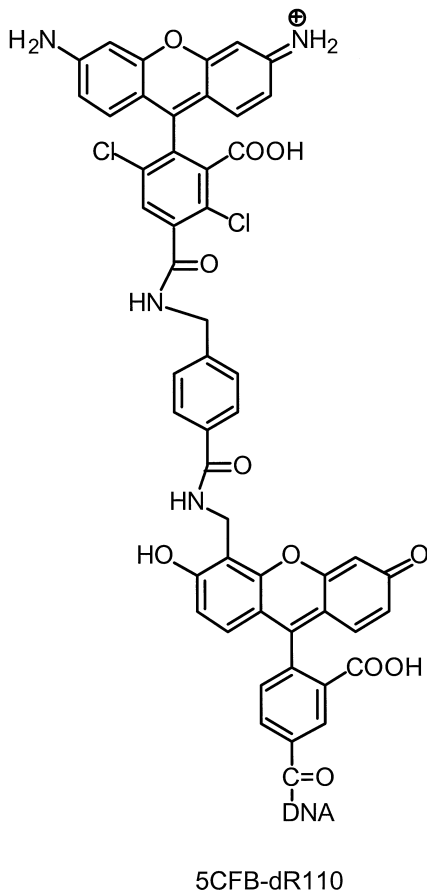


Fig. 9. Structure of dichloro-R110 linked to 4'-aminomethylfluorescein (LEE et al., 1997; ROSENBLUM et al., 1997). This dual dye configuration permits fluorescence resonant energy transfer from fluorescein (donor) to R110 (acceptor) and is a member of a commercially available family of dyes trademarked as BigDyes™ (PE Biosystems). Other BigDyes™ are synthesized with dTAMRA, dROX, and dR6G as acceptors, all of which contain the dichloro modification.

dye but a different dideoxynucleotide (ddATP, ddTTP, ddGTP, ddCTP). The four completed sequencing reactions are loaded in separate lanes of a slab gel (Sect. 2, Fig. 1), and the automated sequencer must then be able to align the raw data from all four lanes precisely enough to determine the correct base sequence (Sect. 6.2).

The second type employs the four-dye/one-lane approach in which a single combined reaction is performed using a fluorescent label specific for each of the four dideoxynucleotides. The combined sequencing reaction can be analyzed in a single gel lane or capillary (Sect. 5.4), and the automated sequencer must first correct for the different mobilities of the four dye-labeled DNA fragment sets before calling bases (HAWKINS et al., 1992).

4.1 Sequencing Applications and Strategies

DNA sequencing is a fundamental technique in genome analysis and it has major applications which fall into two general classes:

- (1) *de novo* sequencing of unknown DNA and
- (2) resequencing segments of DNA for which the sequence is already known.

In both cases, the DNA to be sequenced is first cloned into a viral or plasmid vector, or is part of an amplified PCR fragment (Sect. 4.2).

The approach used to sequence unknown DNA is termed the sequencing strategy and it should provide the correct consensus sequence on both strands of the target DNA using a minimal number of sequencing reactions with minimal overlap (Sect. 4.1.1). Large-scale sequencing projects make use of one or more sequencing strategies to completely characterize the entire genome of an organism (SULSTON et al., 1992; FLEISCHMANN et al., 1995; VENTER et al., 1998; KUKANSKIS et al., 2000). On the other hand, many laboratories employ methods of resequencing to characterize the variability of smaller, known DNA segments in order to find mutations or verify recombinant DNA constructs (Sect. 4.1.2).

4.1.1 New Sequence Determination

The selection of a sequencing strategy usually depends on the size of the target DNA. For example, random shotgun sequencing is currently the method used in most large-scale DNA sequencing projects (MARTIN-GALLAR-

DO et al., 1994; VENTER et al., 1998). In shotgun sequencing, a large segment of target DNA (e.g., a medium-sized BAC clone of 100–120 kilobases) is randomly fragmented by physical shearing or enzymatic digestion to fragment sizes in the range of 1–5 kilobases. These smaller fragments are then subcloned into bacteriophage M13 or plasmid vectors (Sects. 4.2.1 and 4.2.2). The cloned inserts are sequenced from “universal” primer binding sites in the flanking vector DNA, and the resulting sequence information compiled by computer into contiguous sequence (i.e., “contigs”) in order to reassemble the original large target DNA.

This method generates rapidly 95% of the desired sequence, but becomes less efficient as each subsequent random subclone is more likely to yield sequence information already obtained. Typically, each base in the target DNA sequence is read an average of four to six times during this “working draft” phase of the shotgun sequencing project. However, gaps or unresolved regions will still remain which can be filled in by directed approaches during the “finishing” phase of the sequencing project (HUNKAPILLER et al., 1991; SULSTON et al., 1992; ROACH et al., 1999; KUKANSKIS et al., 2000).

Advantages of shotgun sequencing include no requirement for prior knowledge of the insert sequence and no limitation on the size of the starting target DNA. Additionally, a high degree of parallel processing and automation can be implemented during the initial random phase, with only one or two oligonucleotide sequencing primers required.

Primer walking is a fully directed sequencing strategy. It provides an efficient way to obtain new sequence information, and is a good choice for the primary sequencing of small regions (1–3 kilobases) of genomic or cDNA clones, or as a secondary approach to achieve closure and resolve local ambiguities after an initial shotgun sequencing phase. Other approaches such as the enzymatic nested deletion method (LIU and FLEISCHMANN, 1994) or transposon insertion (MARTIN et al., 1994) have also been used for small-scale *de novo* sequencing.

The primer-directed method is initiated by sequencing the target DNA from one end

using a vector-specific standard primer (VOSS et al., 1993b). A new walking primer is designed using the most distant, reliable sequence data obtained from the first sequencing reaction with the standard primer. This walking primer is then used to sequence the next unknown section of the DNA template. In theory, this primer walking process can be repeated many times to sequence extensive tracts of DNA. However, its use is generally limited to smaller projects because the successive rounds of sequence analysis, primer design, and primer synthesis are too expensive and time-consuming (ANSORGE et al., 1997).

The major benefits of primer walking are that no subcloning is required, the location and direction of each sequencing run is known, and the degree of redundancy needed to obtain final sequence is minimized. Moreover, read lengths greater than 1,000 bases have been reported (NISHIKAWA and KAMBARA, 1992; GROTHUES et al., 1993; ZIMMERMANN et al., 1994; MIDDENDORF et al., 1995; CARRILHO et al., 1996; KLEPÁRNIK et al., 1996; ROEMER et al., 1997, 1998; SALAS-SOLANO et al., 1998b; ZHOU et al., 2000) thus reducing the number of walking primers needed to finish a sequencing project.

4.1.2 Confirmatory Sequencing

The major purpose of DNA sequencing in many laboratories is to resequence small regions of interest (<1 kilobase) using cloned DNA or a PCR product as the template. Resequencing is useful for applications such as confirming plasmid constructs, screening the products of site-directed mutagenesis experiments, or comparing sequences of wild-type and mutant variants associated with genetic disease (LARDER et al., 1993; Perkin-Elmer/ABI, 1995; PLASCHKE et al., 1998). Since the target region has often been characterized, it is possible to design a primer so that the sequence of interest is within 100–150 bases of the sequencing primer. This will provide optimum resolution in the raw sequence data generated by the automated DNA sequencer, and thus the highest base calling accuracy that can be obtained (Sect. 6).

4.2 DNA Template Preparation

In the first step of a Sanger dideoxy sequencing reaction, the primer is annealed to a single-stranded DNA template (Sect. 2). DNA in this form can be purified directly from viruses such as bacteriophage M13 which have single-stranded genomes. On the other hand, double-stranded DNA such as a plasmid vector containing the target insert must first be converted to the single-stranded form either by alkali or heat denaturation prior to sequencing (CHEN and SEEBURG, 1985; ANSORGE et al., 1997).

The material presented in this section is intended to serve only as a general guide for preparing DNA templates. Specific protocols and applications can be found in several molecular biology manuals (SAMBROOK et al., 1989; AUSUBEL et al., 1992; FANNING and GIBBS, 1997; ANSORGE et al., 1997; WILSON and MARDIS, 1997a, b).

4.2.1 Single-Stranded DNA Template

Several variants of the bacteriophage M13 were constructed for the purpose of generating a DNA template for dideoxy sequencing (MESSING, 1983). The DNA to be sequenced is cloned into the double-stranded replicative form of the phage, transformed into *E. coli*, and harvested in large quantity from the culture medium in the form of phage particles containing single-stranded DNA (MESSING and BANKIER, 1989). The purified DNA is ideal for sequencing as it is single-stranded so that no complementary strand exists to compete with the sequencing primer during the annealing step. Moreover, a universal sequencing primer hybridizes to a complementary portion of the phage DNA immediately adjacent to the multiple cloning site. M13 is still used extensively for high-throughput sequencing applications (MARTIN-GALLARDO et al., 1994).

4.2.2 Double-Stranded DNA Template

Many methods have been developed to isolate and purify plasmid DNA from bacteria (SAMBROOK et al., 1989). Generally, the process involves five steps:

- (1) insert foreign (target) DNA into the plasmid vector,
- (2) transform a suitable bacterial strain with the recombinant plasmid,
- (3) grow the bacterial culture,
- (4) harvest and lyse bacteria, and
- (5) purify the plasmid DNA.

For sequencing applications, double-stranded plasmid DNA containing the target sequence must be of high purity. Contaminating salt, RNA, protein, DNases, and polysaccharides from the host bacteria can inhibit dideoxy sequencing reactions and produce low signal, high background, or spurious bands. Plasmid DNA purified through a cesium chloride gradient is suitable for sequencing provided that residual salt is removed from the DNA by ethanol precipitation. Commercial plasmid purification kits using anion-exchange resins or silica gel membrane technologies are available from Qiagen Inc. (Valencia, CA) or Promega Corp. (Madison, WI). These kits are easy to use and provide high quality DNA.

4.2.3 Vectors for Large-Insert DNA

Cloning vectors capable of replicating large DNA inserts, such as cosmids (DNA inserts 35 to 45 kilobases), P1-derived artificial chromosomes (PACs; DNA inserts from 100 to 150 kilobases), and bacterial artificial chromosomes (BACs; DNA inserts up to 300 kilobases), have been developed for use in genome mapping and large-scale DNA sequencing projects (CRAXTON, 1993; IOANNOU et al., 1994; SHIZUYA et al., 1992). These large-insert clones can be used to construct subclone libraries and then sequenced by the shotgun approach (WILSON and MARDIS, 1997b) (Sect. 4.1.1).

It is also important to sequence directly on these large DNA clones (BOYSEN et al., 1997; FAJAS et al., 1997). Sequence information from the ends of large-insert clones is used in the initial mapping phase of a sequencing project by detecting clones with overlapping sequence. Additionally, closing gaps and low quality regions in the “draft” sequence of a large-insert clone can be accomplished more efficiently by sequencing directly off the cosmid or BAC clone. This process eliminates the need to find the specific subclone sequence or to generate a new subclone library covering the gap.

4.2.4 PCR Products

The polymerase chain reaction (PCR) permits a region of DNA located between two distinct priming sites to be amplified (MULLIS and FALOONA, 1987). The product of this *in vitro* nucleic acid amplification is termed the PCR product. If equal amounts of the two primers are used, the PCR product will be a linear double-stranded DNA molecule typically less than 3 kilobases in size which can serve as template for DNA sequencing (INNIS et al., 1990; FANNING and GIBBS, 1997).

The PCR reaction mix contains significant amounts of reagents such as primers, nucleotides, enzyme, and even unwanted amplified products which must be completely removed from the PCR product before it can be successfully sequenced. Thus, the PCR product should be checked on an agarose gel to verify the presence of a single band of the expected size. Then, the PCR product is purified using a commercial PCR purification kit (e.g., Promega Corp. Wizard® DNA Clean-Up System) or by PEG precipitation (WILSON and MARDIS, 1997a). Alternatively, PCR products can be purified by agarose gel (AUSUBEL et al., 1992).

4.3 Enzymatic Reactions

4.3.1 DNA Polymerases

In the original Sanger dideoxy sequencing protocol, the Klenow fragment of *E. coli* DNA polymerase I was used for primer extension/

termination reactions. The quality of the DNA sequence obtained with the Sanger method was significantly improved by the development of a modified T7 DNA polymerase (Sequenase® v2.0, United States Biochemical, Cleveland OH and Amersham Pharmacia Biotech, Piscataway NJ) which has enhanced processivity and a striking uniformity of termination patterns particularly when manganese ions are used as a cofactor (TABOR and RICHARDSON, 1987, 1989; VOSS et al., 1989). Both the Klenow fragment and modified T7 DNA polymerase catalyze the synthesis of DNA sequencing fragments in a single pass as the enzyme moves along the template DNA. However, these enzymes are also thermolabile, and thus cannot be used in cycle sequencing protocols which produce an amplification of signal by reusing repeatedly small amounts of the template DNA (CRAXTON, 1991). Modified T7 DNA polymerase is effective for sequencing difficult regions with repeats that cause premature “stops” in cycle sequencing reactions (WILSON and MARDIS, 1997a).

Cycle sequencing methods that utilize the thermostable *Thermus aquaticus* (Taq) DNA polymerase have been developed (INNIS et al., 1988; CRAXTON, 1991). The use of a thermostable DNA polymerase allows repeated rounds of high temperature DNA synthesis involving thermal denaturation of the double-stranded template DNA, primer annealing, and extension/termination of the reaction products. For each cycle, the amount of product DNA will be roughly equivalent to the amount of primed template. Thus, a significant benefit of cycle sequencing is that only small amounts of DNA template are required since the number of sequencing reaction products (i.e., “the signal”) are linearly amplified during the 20–40 cycles of synthesis. For example, 20–30 ng of a small PCR product or 2–3 µg of a large BAC clone provide sufficient template DNA to complete a cycle sequencing reaction. Moreover, performing the cycle sequencing reactions at elevated temperatures minimizes sequencing artifacts due to secondary structure in the template DNA.

Until recently, the main disadvantage of cycle sequencing was the poor performance of the native Taq DNA polymerase which tends

to incorporate dideoxynucleotides unevenly as compared to deoxynucleotides. As a result, sequencing patterns generated with these enzymes are not uniform (i.e., variable peak heights or band intensities) which reduces the base calling accuracy in automated DNA sequencers (Sect. 6). However, new genetically modified thermostable polymerases with a high affinity for dideoxynucleotides have been introduced (REEVE and FULLER, 1995; PARKER et al., 1996). These enzymes, Thermo Sequenase™ from Amersham Pharmacia Biotech and AmpliTaq FS™ from PE Biosystems (Foster City, CA), incorporate ddNTPs at rates similar to dNTPs resulting in uniform peak heights and, therefore, longer, more accurate sequence read lengths. Additionally, the reduced discrimination against ddNTPs that has been engineered into ThermoSequenase and AmpliTaq FS may manifest itself in the greater acceptance of fluorescent dye-labeled terminators (Sect. 2) as substrates in the enzymatic sequencing reaction (REEVE and FULLER, 1995).

4.3.2 Labeling Strategy

Automated DNA sequencing uses fluorescent dyes (Sect. 3) for the detection of electrophoretically resolved DNA fragments. There are three methods for labeling DNA sequencing reaction products:

- (1) dye-labeled primer sequencing (SMITH et al., 1986; ANSORGE et al., 1986) in which the fluorescent dye is attached to the 5' end of the oligonucleotide primer,
- (2) dye-labeled terminator sequencing (PROBER et al., 1987; LEE et al., 1992) in which the fluorophores are attached to the dideoxynucleotides or a non-nucleotide terminator (ROEMER et al., 2000), and
- (3) internal labeling (VOSS et al., 1993b, 1997; STEFFENS et al., 1995) in which a dye-labeled deoxynucleotide is incorporated during the synthesis of a new DNA strand.

Each labeling method has advantages and disadvantages.

Dye-labeled primer sequencing has benefited from the new DNA polymerases which do not discriminate between deoxynucleotides and dideoxynucleotides (Sect. 4.3.1). The sequencing electropherograms generated using these enzymes with dye-primers have very even peak heights which makes the base calling easy and reliable. Furthermore, signal uniformity allows heterozygote detection to be based on peak heights as well as the presence of two bases at the same position (Perkin Elmer/ABI, 1995). One disadvantage of the dye-primer method is a greater likelihood of increased background level (e.g., spurious bands) because nucleotide chains which terminate prematurely will add to the level of false terminations. Also, the four-dye/one-lane approach for automated sequencing (Sect. 4) requires four separate extension reactions and four dye-labeled primers per template.

The main advantages of dye-terminator sequencing are convenience, since only a single extension reaction is required per template, and the synthesis of a dye-labeled primer is not necessary. In fact, custom unlabeled primers with preferred hybridization sites can be used with dye terminators. Moreover, false terminations (i.e., DNA fragments terminated with a dNTP rather than a ddNTP) are not observed since these products are unlabeled. Finally, sequencing with dye terminators provides a way to read through most compressions. Presumably the large fluorophore at the 3' end of the DNA fragment modifies or eliminates the in-gel secondary structure that causes compressions (WILSON and MARDIS, 1997a). The major disadvantage of dye terminators is that the pattern of termination varies between DNA polymerases and is less uniform than for dye-labeled primers.

4.3.3 The Template-Primer-Polymerase Complex

An important factor in the relative success of a sequencing reaction is the number of template-primer-polymerase complexes formed during the course of a sequencing reaction. The formation of this complex is necessary in order to produce dye-labeled extension products. A significant number of problems asso-

ciated with DNA sequencing reactions can be traced to one or more of these key elements.

For example, the ability of an oligonucleotide primer to bind to the template and interact with the DNA polymerase is a major factor in the overall signal strength of the reaction. Primers should be designed without inverted repeats or homopolymeric regions, a base composition of about 50% GC, no primer dimer formation, and one or more G or C residues at the 3' end of the primer. These factors affect the stability of the primer–template interaction, and thus determine the number of primer–template complexes available to the DNA polymerase under a given set of conditions. For cycle sequencing with thermostable polymerases, it is important to design the primer with an annealing temperature of at least 50°C. Lower annealing temperatures tend to produce higher background and stops in cycle sequencing.

The amount of DNA template used in the dideoxy sequence reaction needs to be within an appropriate range. If the amount of template is too low, then few complexes will form and the overall signal level will be too low for automatic base calling. Additionally, higher amounts of a lower quality template (e.g., salt contaminant carried over from the DNA preparation) may be inhibitory to the DNA polymerase resulting in lower signal levels.

The most common factors which limit sequence read length and base calling accuracy in automated DNA sequencers are impure DNA template, incorrect primer or template concentrations, suboptimal primer selection and annealing, and poor removal of unincorporated dye-labeled dideoxynucleotides.

4.3.4 Simultaneous Bidirectional Sequencing

Simultaneous bidirectional sequencing (SBS), also termed “doublex” sequencing (WIEMANN et al., 1995; VOSS et al., 1997), is a sequencing method in which both strands of duplex DNA (plasmid or PCR product) are sequenced simultaneously by combining a forward and reverse primer (each labeled with a different fluorescent dye) in the same sequencing reaction. An automated DNA sequencing

system with dual lasers, such as the LI-COR Model 4200 (Lincoln, NE) or the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) two-dye DNA sequencer, can be used to detect and analyze both the forward and reverse sequences of a bidirectional reaction in parallel (ROEMER et al., 1997; ANSORGE et al., 1997).

The benefits of the SBS method are threefold. First, SBS doubles the amount of sequence information from a single sequencing reaction. Second, since confirming sequence can be generated in the same reaction, it is easier to resolve ambiguities in one strand using the sequence of the complementary strand. Third, time and reagent consumption are halved by combining the forward and reverse sequencing reactions.

5 Fluorescence DNA Sequencing Instrumentation

5.1 Introduction

In principle, there are only three components of a fluorescence detection system:

- (1) the excitation energy source,
- (2) the fluorescent sample, and
- (3) the fluorescence emission energy detector.

In practice, all of these components are sophisticated subsystems whose designs are coordinated to deliver maximum information throughput with optimized signal vs. noise discrimination (to achieve high accuracy and data quality). A brief discussion of these components is provided here to allow an overview of the parameters involved in proper instrumentation design for DNA sequencing. For a detailed description of general fluorescence-based instrumentation, a comprehensive textbook such as that authored by LAKOWICZ (1999) should be consulted. For a recent review of near-infrared fluorescence instrumentation refer to MIDDENDORF et al. (1998).