DNA Sequencing

Susan H Hardin, *University of Houston, Texas, USA*

DNA sequencing is the determination of base order in a DNA molecule. Methods for determining base order involve either chemical degradation or, more commonly, enzymatic synthesis of the region that is being sequenced. Automation of the DNA sequencing process is accelerating the progress of the Human Genome Project.

Introductory article

Article Contents

Introduction

The development of methods that allow one to quickly and reliably determine the order of bases, or the 'sequence', in a fragment of DNA is a key technical advance, the importance of which cannot be overstated. Knowledge of DNA sequence enables a greater understanding of the molecular basis of life. DNA sequence information provides information critical to understanding a wide range of biological processes. The order of bases in DNA specifies the order of bases in RNA, the molecule within the cell that directly encodes the informational content of proteins. Scientists routinely use the DNA sequence information to deduce protein sequence information. Base order dictates DNA structure and its function, and provides a molecular programme that can specify normal development, manifestation of a genetic disease, or cancer.

Knowledge of DNA sequence and the ability to manipulate these sequences has accelerated the development of biotechnology and has led to the development of molecular techniques that provide the tools for asking and answering important scientific questions. The polymerase chain reaction (PCR), an important biotechnique that facilitates sequence-specific detection of nucleic acid, relies on sequence information. DNA sequencing methods allow scientists to determine whether a change has been introduced into the DNA, and to assay the effect of the change on the biology of the organism, regardless of the type of organism that is being studied. Ultimately, DNA sequence information may provide a way to identify individuals uniquely.

DNA sequencing has become so commonplace that the technique itself is often taken for granted. However, this has not always been the case. It was, in fact, almost required that scientists publish or present DNA sequence data before a sequence was considered reliable. Furthermore, the length of the DNA information that it is possible to obtain and the number of sequences that are analysed on a single gel have increased by an order of magnitude. This article provides an overview of DNA sequencing development.

To understand the DNA sequencing process, one must recall several facts about DNA. First, a DNA molecule is composed of four bases, adenine (A), guanine (G), cytosine (C) and thymine (T). These bases interact with each other in very specific ways through hydrogen bonds, such that A interacts with T, and G interacts with C. These specific interactions between the bases are referred to as base pairings. The two strands of a DNA molecule occur in an antiparallel orientation in which one strand is positioned in the 5′ to 3′ direction and the other strand is positioned in the 3′ to 5′ direction. The terms 5′ and 3′ refer to the directionality of the DNA backbone, and are critical to describing the order of the bases. The convention for describing base order in a DNA sequence uses the 5′ to 3′ direction, and is written from left to right. Thus, if one knows the sequence of one DNA strand, the complementary sequence can be deduced.

There are two methods that are typically used to determine DNA sequence; the development of each method resulted in the award of a Nobel Prize. The first uses chemicals to specifically degrade the DNA strand, and is referred to as Maxam–Gilbert DNA sequencing in honour of the inventors, A. Maxam and W. Gilbert. The second method involves specific inhibition of enzymatic DNA synthesis and is referred to as Sanger sequencing in honor is its inventor, F. Sanger. These two sequencing methods are described in more detail below.

Both methods require that the reaction products share a common endpoint. This requirement stems from the separation method used to visualize the reaction products. These reaction products are size-separated by applying an electric current through a gel matrix (electrophoresis), and a common end is necessary to keep the reaction products in register with respect to size mobility, so that the smaller products migrate more rapidly on the gel relative to the larger products. More specifically, either the 5′ or the 3′ end can define the fragment endpoint in a Maxam–Gilbert sequencing reaction, while only the 5′ end defines the fragment endpoint in a Sanger sequencing reaction. The reason for this difference is clarified below.

Maxam–Gilbert DNA Sequencing (Chemical Degradation)

In this method, a singly end-labelled DNA fragment (typically labelled with a radioactive marker) is exposed to base-specific damage. The chemicals used to induce DNA damage are dimethylsulfate (attacks G), sodium hydroxide (attacks A), formic acid (attacks G and A), hydrazine (attacks C and T), and hydrazine in the presence of sodium chloride (attacks C). These treatments are limited so that on average only one of the bases in the strand is damaged. Modification and, ultimately, elimination of the base (but not the sugar) produces a weak point in the DNA molecule that is susceptible to cleavage. Next, the DNA is exposed to piperidine at high temperature to break the strand at the weakened position. Since these reactions are performed on a population of molecules, electrophoresis of the reaction produces a ladder of cleavage products that correspond to the positions of that base along the DNA strand. Products from the different chemical modifications are electrophoresed in adjoining lanes on the gel and read to determine the DNA sequence (**Figure 1**). The smaller fragments indicate the identity of bases closest to the labelled end, and successively larger products indicate the identity of bases farther from this end. Depending on whether the label is located at the 5′ or the 3′ end of the DNA strand, the base order is read from the bottom to the top of the autoradiogram as either 5′ to 3′ or 3′ to 5′, respectively.

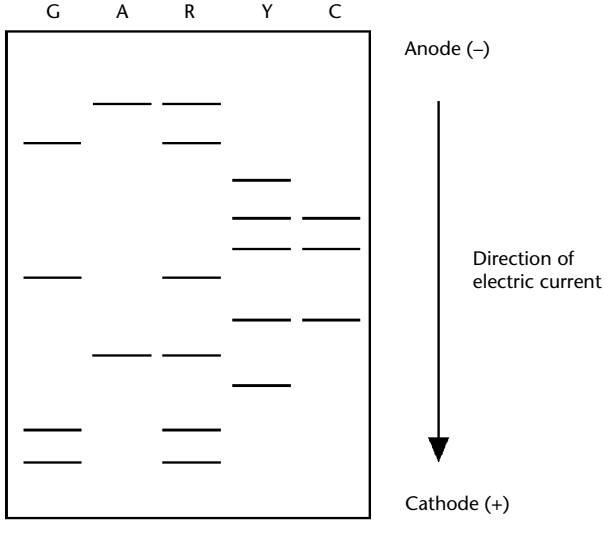

Figure 1 Schematic view of Maxam–Gilbert reaction products. G, A, R, Y and C represent the specific chemical reactions that identify the relative positions of guanine (G), adenine (A), purine ('R'; G and A), pyrimidine ('Y'; C and T) and cytosine (C) bases, respectively. In this example the fragment is labelled at the 5′ end. Reading from the bottom towards the top of the gel, the banding pattern corresponds to the sequence 5′ GGTACGCCTGA 3′.

Maxam–Gilbert sequencing is not routinely used by most investigators for several reasons. First, data produced in chemical sequencing reactions are typically more ambiguous than data produced in enzymatic sequencing reactions. One reason for this is that the chemical reactivity of the bases is influenced by reaction impurities. Therefore, when one reads the sequence from this type of reaction, the relative intensities of the reaction products must be analysed for proper interpretation of base identity. Additionally, this procedure uses hazardous chemicals and high levels of radioactivity. When compared with enzymatic DNA sequencing, Maxam–Gilbert sequencing produces relatively shorter sequence information and the procedures required to generate this information are more labour-intensive.

Sanger DNA Sequencing (Enzymatic Synthesis)

Sanger sequencing is currently the most commonly used method for sequencing DNA. The method exploits several features of a DNA polymerase: its ability to make an exact copy of a DNA molecule; its directionality of enzymatic synthesis (5′ to 3′); its requirement for a DNA strand (a 'primer') from which to begin synthesis; and its requirement for a 3′ OH at the end of the primer. If a 3′ OH is not available, the DNA strand cannot be extended by the polymerase. If a dideoxynucleotide (ddNTP – ddATP, ddTTP, ddGTP, ddCTP,), a base analogue lacking a 3′ OH, is added into an enzymatic sequencing reaction, it is incorporated into the growing strand by the polymerase. However, once the ddNTP is incorporated, the polymerase is unable to add any additional bases to the end of the strand. Importantly, ddNTPs are incorporated into the DNA strand by the polymerase using the same base incorporation rules that dictate incorporation of natural nucleotides, where A specifies incorporation of T, and G specifies incorporation of C (and vice versa).

Once the polymerase incorporates a ddNTP, chain extension stops. To determine DNA sequence, one performs four reactions per template, where each reaction is used to determine the relative position of a specific base along the DNA strand. This is accomplished by adding a different dideoxynucleotide into each reaction containing the DNA polymerase, the DNA primer, and the dNTPs. The ratio between ddNTP and dNTP is critical for determining how many nucleotides (on average) the polymerase is able to incorporate into the DNA molecule before incorporating a ddNTP, thereby terminating chain elongation. The DNA primer, the ddNTPs, or the dNTPs can be either radiolabelled or otherwise tagged to allow detection of the newly synthesized DNA strands. Since these reactions are performed on a population of molecules, electrophoresis of the reaction produces a

ladder of extension products that correspond to the positions of that base along the DNA strand. Products from the different reaction vessels are electrophoresed in adjoining lanes on a sequencing gel, detected, and read to determine the DNA sequence (**Figure 2**). The smaller fragments indicate the identity of bases closest to the 5′ end and, since the DNA polymerase only incorporates bases in the 5′ to 3′ direction, reading the identity of the successively larger products provides the 5′ to 3′ sequence of the extended DNA strand. Typically, approximately 300–400 bases can be determined in a single (manual) reaction.

Automated DNA Sequencing

A major advance in determining DNA sequence information occurred with the introduction of automated DNA sequencing machines. The automated sequencer is used to separate sequencing reaction products, detect and collect (via computer) the data from the reactions, and analyse the order of the bases to automatically deduce the base sequence of a DNA fragment. Automated sequencers detect extension products containing a fluorescent tag, allowing researchers to eliminate radioactivity from the DNA sequencing process. Sequence lengths that can be read using an automated sequencer are dependent upon a variety of parameters, but typically range between 500 and 1000 bases.

As described for Sanger-type sequencing reactions using (primarily) isotopes to detect the extension products, some automated sequencers use four lanes to collect the data from the reactions. However, some machines use differently coloured fluorescent tags to indicate base identity (**Figure 3**). This approach enables a single lane to contain the data for a DNA template and increases fourfold the

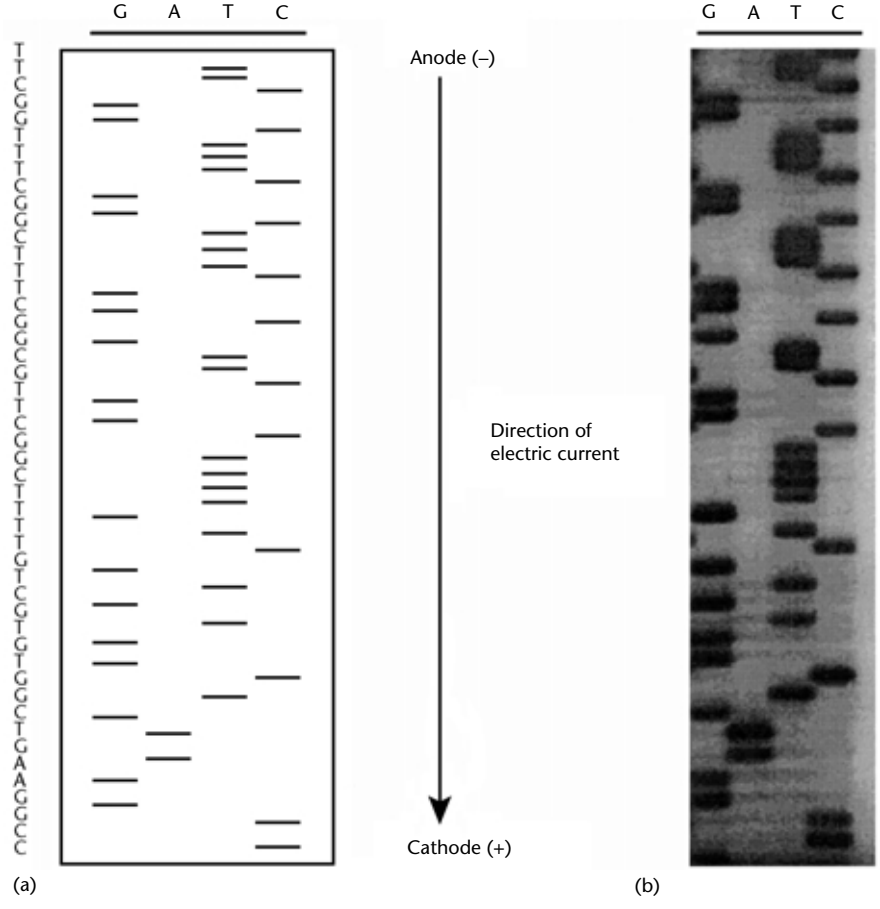

Figure 2 Data produced using Sanger sequencing reaction. G, A, T and C represent the sequencing reaction products resulting from inclusion of ddGTP, ddATP, ddTTP or ddCTP. Since enzymatic synthesis proceeds 5′ to 3′, the smaller fragments identify bases that are closer to the primer (5′ end of the sequence information). (a) Schematic view of Sanger reaction products. The DNA sequence identified by this pattern of bands is indicated. (b) Photograph of corresponding sequence data.

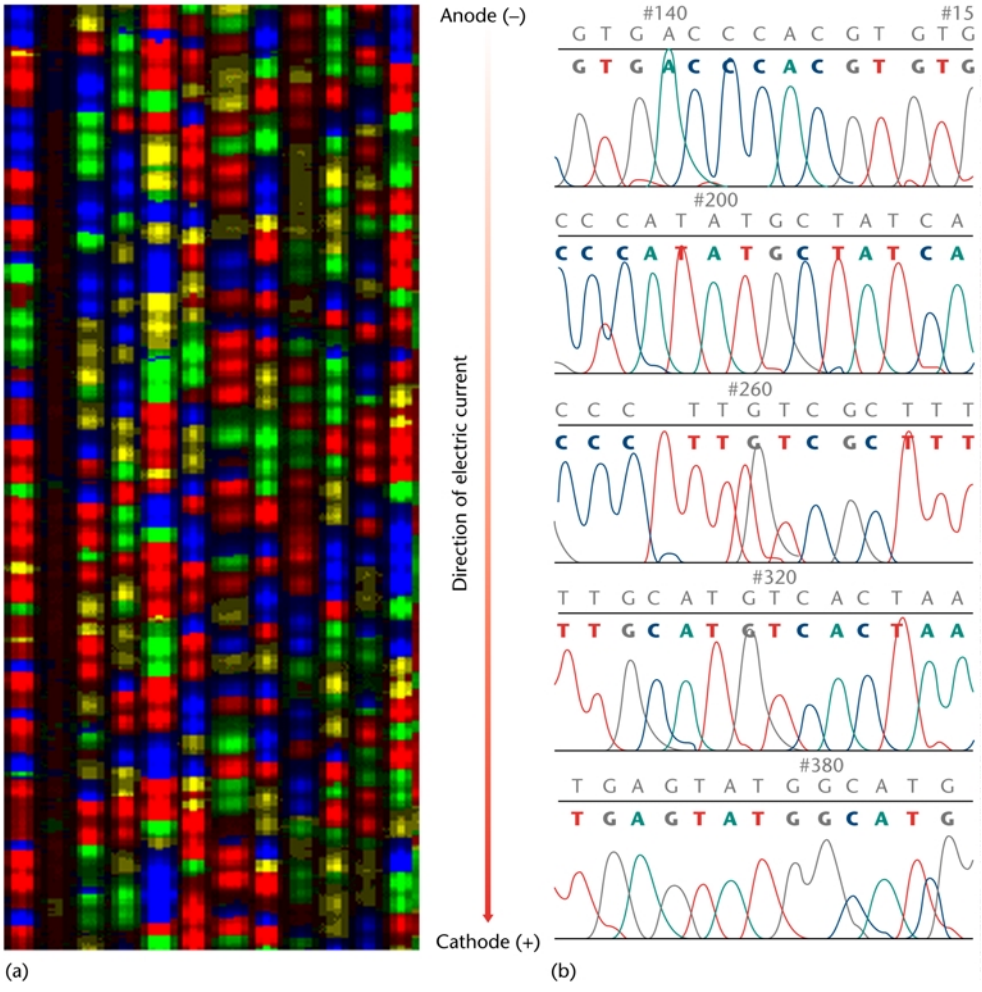

Figure 3 (a) Raw sequence data collected on an automated DNA sequencer (Perkin-Elmer ABI PRISM Model 377). The four colours indicate the relative position of the bases in the DNA fragment. Each four-colour vertical line corresponds to a different sequence reaction. The smaller fragments (nearer the cathode) identify bases that are closer to the primer (5′ end of the sequence information). (b) Portions of a representative, analysed sequence determined by the automated sequencer.

amount of data contained on a gel. This single-lane approach is made possible by the development of fluorescent tags that can be attached either to the DNA primer or to the ddNTP. Since four-colour chemistry is used by more researchers, it is discussed in more detail below.

Dye primer chemistry

When dye primer chemistry is used to detect the sequencing products, fluorescent tags are attached to the sequencing primer. With this chemistry, the primer is synthesized four times and a different tag (corresponding to a different base identity) is attached to the primer during each synthesis. Subsequently, the researcher assembles four separate reactions, each containing the DNA template, a specific ddNTP and colour-coded primer, and the reagents necessary to produce extension products. The primer defines the beginning (5′ end) of the extension product, and the incorporated ddNTP defines base identity at the 3′ end of the molecule. After the reaction is completed, the colour-coded products are pooled and prepared for loading into a single lane on an automated sequencer.

Dye terminator chemistry

When dye terminator chemistry is used to detect the sequencing products, base identity is determined by the fluorescent tag attached to the ddNTP. This type of reaction chemistry is performed in a single tube that

contains the DNA template, the primer, all four fluorescently labelled ddNTPs, and the reagents necessary to produce extension products. The primer defines the beginning (5′ end) of the extension product and the incorporated, colour-coded ddNTP defines base identity at the 3′ end of the molecule. After the reaction is complete, the extension products are prepared for loading into a single lane on an automated sequencer. An advantage of dye terminator chemistry is that extension products are visualized only if they terminate with a dye-labelled ddNTP; prematurely terminated products are not detected. Thus, reduced background signal typically results with this chemistry.

Genome Sequencing

Very often, a researcher needs to determine the sequence of a DNA fragment that is larger than the 500–1000 base average sequencing read length. Not surprisingly, strategies to accomplish this have been developed. These strategies are divided into two major classes, random or directed. Strategy choice is influenced by the size of the fragment to be sequenced.

In random, or shotgun, DNA sequencing, a large DNA fragment (typically one larger than 20 000 base pairs) is broken into smaller fragments that are inserted into a cloning vector. It is assumed that the sum of information contained within these smaller clones is equivalent to that contained within the original DNA fragment. Numerous smaller clones are randomly selected, DNA templates are prepared for sequencing reactions, and fluorescently-labelled primers that will base-pair with the vector DNA sequence bordering the insert are used to begin the sequencing reaction. Subsequently, the sequence of the original DNA fragment is reconstructed by computer assembly of the sequences obtained from the smaller DNA fragments. This strategy is being used extensively to determine the sequence of ordered fragments that represent the entire human genome [http://www.nhgri.nih.gov/HGP/]. However, this random approach is typically not sufficient to complete sequence determination, since gaps in the sequence often remain after computer assembly. A directed strategy (described below) is usually used to complete the sequence project.

A directed, or primer-walking, sequencing strategy can be used to fill gaps remaining after the random phase of large-fragment sequencing, and as an efficient approach for sequencing smaller DNA fragments. This strategy uses DNA primers that anneal to the template at a single site and act as a start site for chain elongation. This approach requires knowledge of some sequence information to design the primer. The sequence obtained from the first reaction is used to design the primer for the next reaction and these steps are repeated until the complete sequence is determined. Thus, a primer-based strategy involves repeated sequencing steps from known into unknown DNA regions; this process minimizes redundancy, and it does not require additional cloning steps. However, this strategy requires the synthesis of a new primer for each round of sequencing.

The necessity of designing and synthesizing new primers, coupled with the expense and the time required for their synthesis, has limited the routine application of primer-walking for sequencing large DNA fragments. Researchers have proposed using a library of short primers to eliminate the requirement for custom primer synthesis. The availability of a primer library would minimize waste of primer, since each primer could be used to prime multiple reactions, and would allow immediate access to the next sequencing primer.

Prospects

One of the original goals of the Human Genome Project was to complete sequence determination of the entire human genome by 2005. However, the project is ahead of schedule and it is expected to produce a 'working draft' of the human genome by 2001. The completed genome sequence is expected by 2003, at least two years ahead of schedule. Technological advances are responsible for the rapid progress of this ambitious project. Progress in all aspects involving DNA manipulation (especially manipulation and propagation of large DNA fragments), evolution of faster and better DNA sequencing methods, development of computer hardware and software capable of manipulating and analysing the data (bioinformatics), and automation of procedures associated with generating and analysing DNA sequences is responsible for this acceleration.

Further Reading

Ball S, Reeve MA, Robinson PS, Hill F, Brown DM and Loakes D (1998) The use of tailed octamer primers for cycle sequencing. *Nucleic Acids Research* **26**: 5225–5227.

Burbelo PD and Iadarola MJ (1994) Rapid plasmid DNA sequencing with multiple octamer primers. *BioTechniques* **16**: 645–650.

Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, the members of the DOE and NIH planning groups (1998) New goals for the US Human Genome Project: 1998–2003. *Science* **282**: 682–689.

Hardin SH, Jones LB, Homayouni R and McCollum JC (1996) Octamer primed cycle sequencing: design of an optimal primer library. *Genome Research* **6**: 545–550.

Jones LB and Hardin SH (1998a) Octamer-primed cycle sequencing using dye-terminator chemistry. *Nucleic Acids Research* **26**: 2824–2826.

Jones LB and Hardin SH (1998b) Octamer sequencing technology: optimization using fluorescent chemistry. *ABRF News* **9**(2): 6–10.

Kieleczawa J, Dunn JJ and Studier FW (1992) DNA sequencing by primer walking with strings of contiguous hexamers. *Science* **258**: 1787–1791.

Kotler LE, Zevin-Sonkin D, Sobolev IA, Beskin AD and Ulanovsky LE (1993) DNA sequencing: modular primers assembled from a library of hexamers or pentamers. *Proceedings of the National Academy of Sciences of the USA* **90**: 4241–4245.

Maxam AM and Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the USA* **74**: 560–564.

Raja MC, Zevin-Sonkin D, Shwartzburd J *et al.* (1997) DNA sequencing using differential extension with nucleotide subsets (DENS). *Nucleic Acids Research* **25**: 800–805.

Sanger F, Nicklen S and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* **74**: 5463–5467.

Siemieniak DR and Slightom JL (1990) A library of 3342 useful nonamer primers for genome sequencing. *Gene* **96**: 121–124.

Smith LM, Sanders JZ, Kaiser RJ *et al.* (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**: 674–679.

Studier FW (1989) A strategy for high-volume sequencing of cosmid DNAs: random and directed priming with a library of oligonucleotides. *Proceedings of the National Academy of Sciences of the USA* **86**: 6917–6921.