CS 294-8 Computational Biology for Computer Scientists **Spring 2003**

Lecture 4: January 30

Lecturer: Gene Myers *Scribe: Evan Chang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

4.1 Polymerase Chain Reaction (PCR)

The Polymerase Chain Reaction (PCR) is technique for selective amplification of DNA *in vitro*. In essence, PCR amplification utilizes the natural machinery of polymerase, the ability to synthesize short oligonucleotides, and some sequence knowledge to obtain a pure sample of a specific fragment of DNA. In principle, one molecule of DNA would be enough to carry out the reaction (though we expect at least a few copies). Recall, DNA polymerase requires a primer (a short existing sequence from which to extend) and proceeds from $5' \rightarrow 3'$ (*i.e.* $3' \rightarrow 5'$ along the template strand). PCR combines the original duplex, an excess of synthetic oligonucleotide primers (usually 18-22 nucleotides in length, see figure 4.1), dNTPs (the four nucleoside triphosphates), and a DNA polymerase capable of withstanding high temperature. It proceeds in the following cycle:

1. Heat the solution to denature the double-stranded DNA.
2. Cool to allow hybridization of the primers followed by elongation. It is much more likely the primers will hybridize than for the original strands to re-hybridize with each other because the primers are in great excess.
3. Repeat.

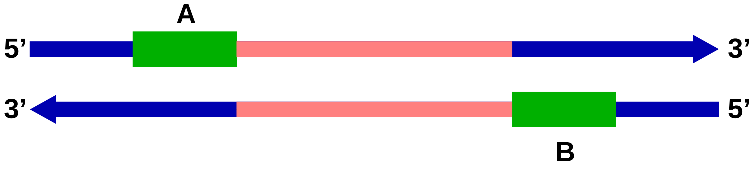


Figure 4.1: PCR Primers. The green segments indicate the primer sequence (usually 18-22 nucleotides) that could be synthesized to replicate the pink region.

It is clear the number of copies of the DNA grows exponentially in the number of cycles. Figure 4.2 illustrates the PCR process.

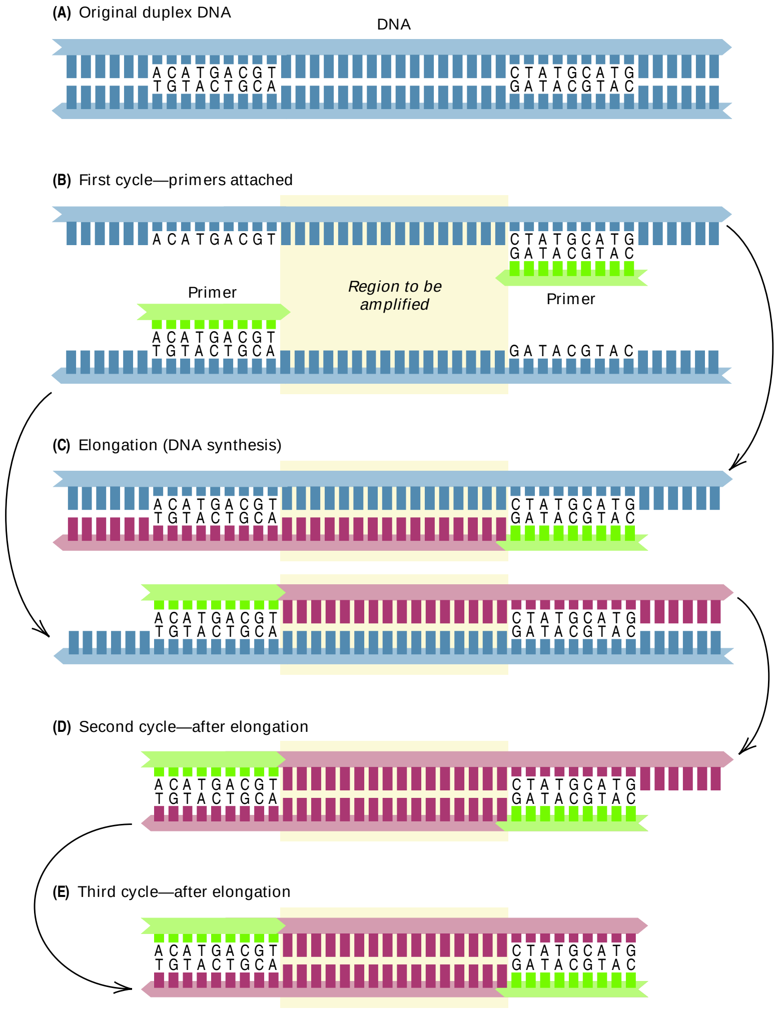


Figure 2.20 Role of primer sequences in PCR amplification. (A) Target DNA duplex (blue), showing sequences chosen as the primer-binding sites flanking the region to be amplified. (B) Primer (green) bound to denatured strands of target DNA. (C) First round of amplification. Newly synthesized DNA is shown in pink. Note that each primer is extended *beyond* the other primer site. (D) Second round of amplification (only one strand shown); in this round, the newly synthesized strand terminates at the opposite primer site. (E) Third round of amplification (only one strand shown); in this round, both strands are truncated at the primer sites. Primer sequences are normally about twice as long as shown here.

Figure 4.2: PCR. (*Source*: Hartl and Jones, page 61 [HL01])

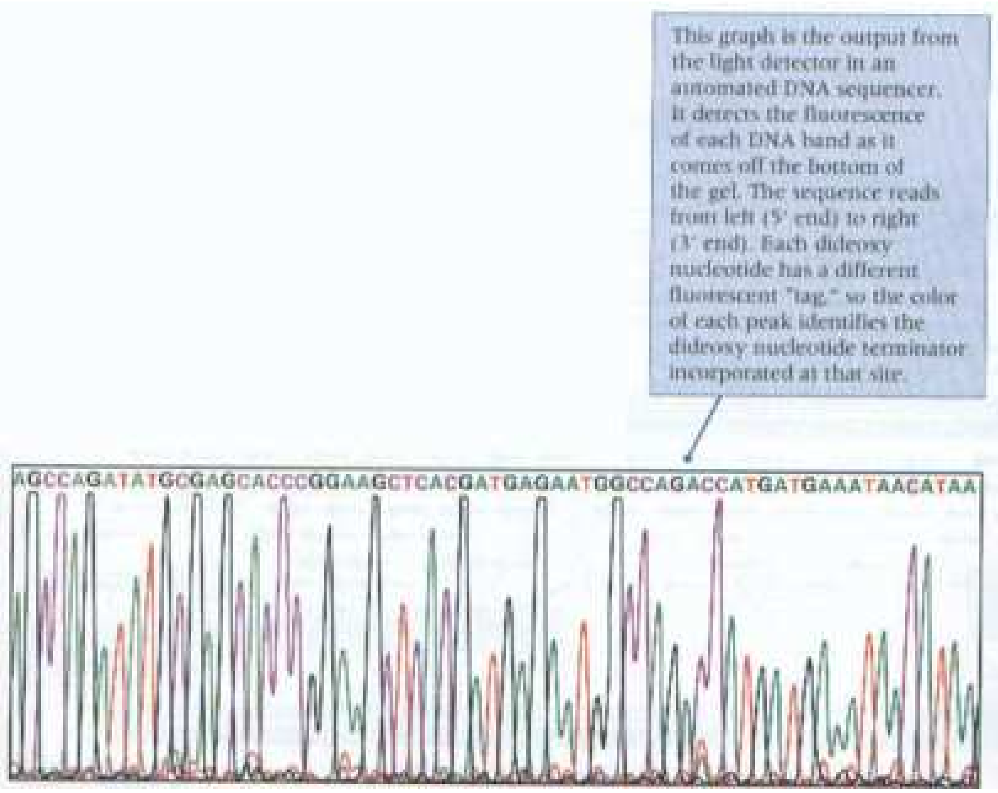
4.2 Application: FBI CODIS Fingerprints

DNA fingerprinting is based an specific type of DNA polymorphism called *simple tandem repeat polymorphism* (STRP). With STRPs, genetic differences result in a different number of copies of some short sequence at particular loci in the genome. Each STRP may differ in the sequence, the length of the repeating unit, and the minimum and maximum number of repeats that occur in the population. By knowing the sequence around the STRP site, PCR primers can be designed to amplify a STRP site (as well as control the range of the expected fragment size). Size differences between resulting fragments can be used to distinguish individuals. The FBI CODIS test uses 13 such STRP sites (along with information about the distributions of the STRPs among different ethnicities). Slides about CODIS from lecture are available at `http://inst.eecs.berkeley.edu/~cs294-8/Materials/CODIS.ppt`.

4.3 DNA Sequencing

The basic idea of the Sanger ladder sequencing method is to produce DNA copies of varying length that stop at a particular base. For example, consider a synthesis reaction that was always forced to end at an adenine (A) residue. Then, if the length of a particular daughter fragment is n, position n in the complement sequence must be A (*i.e.* T in the template sequence). Stopping the reaction at a particular base is accomplished by using *dideoxyribonucleoside triphosphates* (ddNTPs). Dideoxyribose lacks the 3'-hydroxyl group that prevents attachment to the next nucleotide.

Thus, sequencing is accomplished by inserting the desired fragment into a vector (*e.g.* plasmid), combining this with polymerase, a sequencing primer (created from the known the sequence of the vector), dNTPs, and a small amount (≈ 1%) of ddNTPs with a particular fluorescent dye for each base, and then carrying out electrophoresis in a sequencing machine. These machines detect the fluorescent dye by laser light as the fragments run off the gel. This process generates *traces* that can read to determine the sequence (see figure 4.3). It should be clear that the signal would get weaker as the length of the sequence increases. Currently, we are able to sequence only up to 600-800bp.



This graph is the output from the light detector in an automated DNA sequencer. It detects the fluorescence of each DNA band as it comes off the bottom of the gel. The sequence reads from left (5' end) to right (3' end). Each dideoxy nucleotide has a different fluorescent "tag," so the color of each peak identifies the dideoxy nucleotide terminator incorporated at that site.

Figure 4.3: Sequencing Traces. (*Source*: Hartl and Jones, page 245 [HL01])

4.3.1 Paired-End Sequencing

A slight variation of the sequencing protocol described above called *paired-end sequencing* can be used to obtain some information about the space between fragments. Using a double-stranded insert and a sequencing primer at each end in two separate reactions, we sequence 600-800bp from of each end. With the knowledge of the length of the insert (typically, 10,000bp), we also determine the space between these two fragments.

4.3.2 Shotgun Sequencing

Since we are limited to sequencing DNA fragments of 600-800bp in length, to sequence longer segments (or whole genomes), a technique called *shotgun sequencing* has been developed. The basic idea is to take a random sampling and assemble based on overlaps. We define the cover c:

$$c = \frac{R \cdot \bar{L}}{G}$$

where $\bar{L}$ is the number of base-pairs per read (typically, 500bp), R is the number of sequencing reactions (typically, 2 million), and G is the length of the genome. To reduce the likelihood of gaps to an acceptable level, the parameters are chosen such that $c = 10$. A summary of the shotgun sequencing protocol is as follows:

1. Nebulate the original DNA through an aerosol mister to break the DNA into pieces (say approximately 10kbp or 2kbp).
2. Purify based on size with gel electrophoresis.
3. Insert fragments into vectors.
4. Isolate and clone.
5. Run sequencing reactions.
6. Assemble sequences using overlaps.

References

[HJ01] D.L. Hartl and E.W. Jones. *Genetics: An Analysis of Genes and Genomes.* Jones and Bartlett, fifth edition, 2001.