Bi 1X, Spring 2011

Assignment 3: PCR and Restriction Enzymes

Polymerase Chain Reaction (PCR)

One of the key steps in our experiment is amplifying the environmental genomic DNA that we have extracted from the Caltech ponds. We will achieve this amplification using the ubiquitous process of Polymerase Chain Reaction (PCR). PCR is a powerful technique that, along with restriction enzymes and a couple of other "tools," has transformed modern molecular biology through its ability to amplify specific pieces of DNA, even if there is only a small amount of "template" DNA to start with. As shown in Fig. 1 below, through a repetitive series of temperature cycles, DNA polymerase acts as a molecular Xerox machine copying molecules again and again. The questions of this problem set will familiarize you with the basic breakthrough concept of PCR as well as the restriction enzymes.

References on PCR:

- http://www.idtdna.com/support/technical/TechnicalBulletinPDF/Polymerase_Chain_Reaction.pdf
- http://www.idtdna.com/support/technical/TechnicalBulletinPDF/A_Basic_PCR_Protocol.pdf

Problem 1: Understanding PCR

A. Templates, amplicons and anchored products in a PCR

While it is true that the total number of molecules in a PCR reaction increases exponentially as a function of the cycle number, closer inspection reveals that this population is actually the sum of three distinct populations at any given moment in the PCR tube: there is the "original template", the "amplicons" and the "anchored products", all of which are targets for further amplification. The original template is simply the original molecule(s) from which we are trying to amplify a certain sequence. The amplicon is the double stranded DNA that stretches from the sequence of the forward primer to the sequence of the reverse primer; this is the desired end product. The anchored product is different from the amplicon in that only one side is bounded by the forward or reverse primers—there is no cue to tell the DNA polymerase when to stop replicating on the other side. Therefore, anchored products have an undetermined length that will be longer than the amplicon length. Note that anchored products are amplified off of the original template, whereas amplicons are actually formed from replication of either anchored products or other amplicons and not of the original template. Figure 1 below depicts the PCR amplification for the first three cycles and shows the original template, the anchored products and the amplicons.

i. **How many original templates, anchored products, amplicons and total molecules exist in the PCR tube at the start of the reaction and at the end of the first, second and third cycles (i.e. n=0,1,2 and 3) in Figure 1?** For example, for n=1 (at the end of the first

cycle) there is one original (double stranded) template molecule, one anchored product (comprised of two single stranded DNAs) and zero amplicons, for a total of two molecules. You may assume that the amplification efficiency is 100%, i.e., at the end of every cycle the number of molecules exactly doubles (as depicted in Figure 1) and that we start out with one original template molecule and primers in excess.

Summary of components at the end of the first four cycles:

Cycle number	0	1	2	3	4
Original template	1	1	1	1	1
Anchored product	0	1	2	3	4
Amplicon	0	0	1	4	11
Total number of molecules	1	2	4	8	16

ii. How many molecules of the various classes exist at the end of the fourth (n=4) cycle? Can you write down a general formula for the number of various molecules (template, amplicon, anchored product and total number of molecules) as a function of the cycle number n? (You can make the same assumptions as in part i)

In an ideal reaction which is 100% efficient, the total number of molecules doubles every cycle:

$$N_{tot}(n) = N_{tot}(0)2^n = 2^n$$

Where for simplicity we assumed that $N_{tot}(0)=1$, i.e. we start out with a single molecule of the original template. The number of original template molecules is of course constant and always equals 1:

$$N_{originaltemplate}(n) \equiv 1$$

Since anchored products are amplified each cycle only off the original template, their number increases linearly with the cycle number n. Thus

$$N_{anchoredproduct}(n) = n$$

This is therefore the linear component of the PCR amplification. The number of amplicons as a function of the cycle number n will be given by considering the total number of molecules in a given cycle:

$$N_{tot}(n) = N_{originaltemplate}(n) + N_{anchoredproduct}(n) + N_{amplicon}(n)$$

Thus

$$N_{amplicon}(n) = N_{tot}(n) - \left[N_{originaltemplate}(n) + N_{anchoredproduct}(n)\right] = 2^n - n - 1$$

iii. What is the percent of anchored products out of the total number of molecules after n=35 cycles? Given your result, can anchored products be neglected after 35 cycles?

After 35 cycles the anchored product constitutes only of $n/2^n \cong 10^{-7}\%$ of the final product and is thus completely negligible.

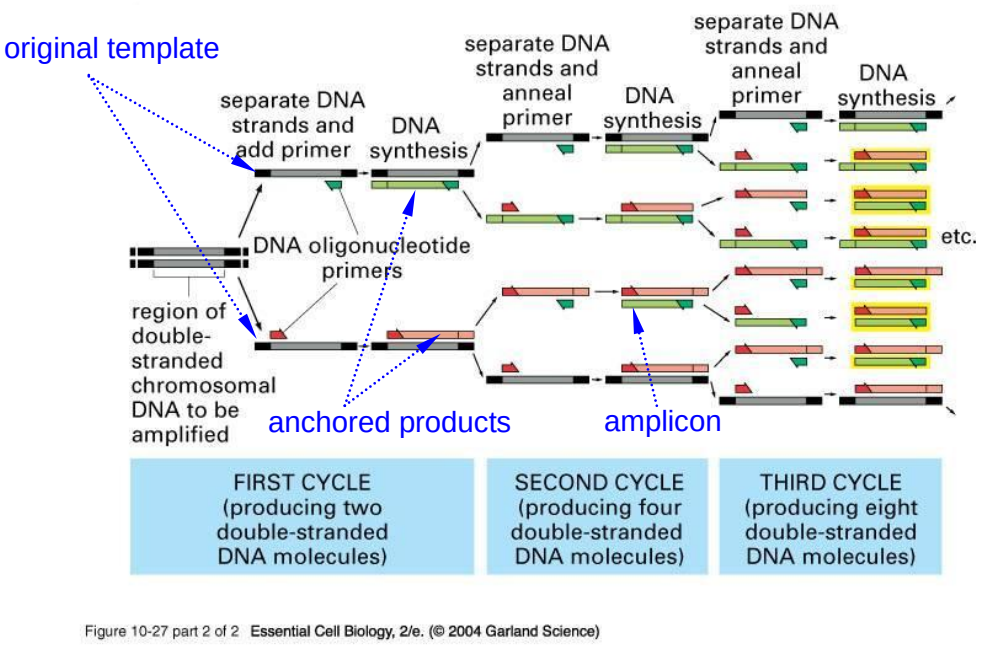


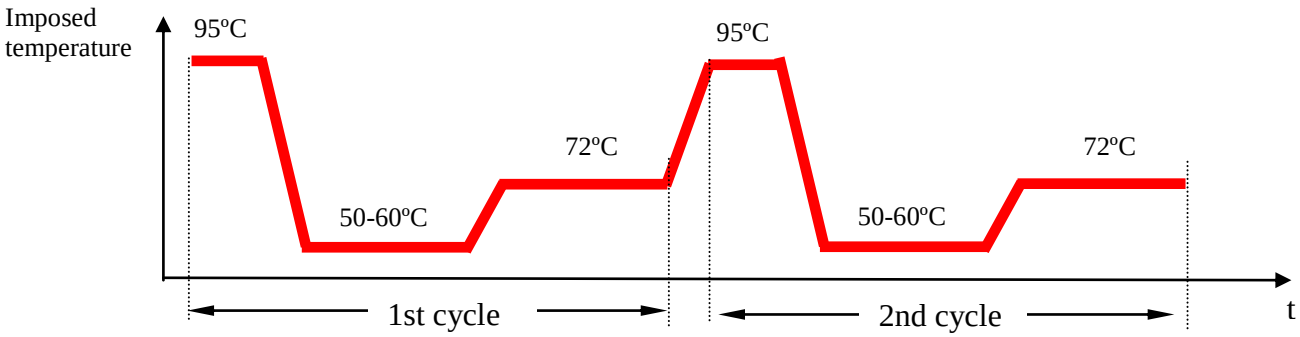
Figure 10-27 part 2 of 2 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Figure 1. The first three cycles in a PCR reaction.

B. The PCR program

Each cycle in the PCR consists of three precisely timed phases that are executed at certain carefully chosen temperatures. Figure 2 is a graph of the imposed reaction temperature as a function of time for any given cycle.

i. **Based on Figure 2, draw a schematic diagram of the temperature profile imposed on the PCR reaction as a function of time for the first two cycles.**



ii. **<u>Briefly</u> explain the purpose of each of the three phases and why the specific temperature was chosen for each phase (you may consult the references listed at the beginning of this problem set).**

- **Denaturation (or melting) phase:** the reaction mix is heated to ~95ºC and double stranded DNA is separated into single strands. No replication takes place at this stage since all DNA is single stranded and the DNA polymerase has no targets to bind to.

- **Annealing phase** – the temperature is reduced to the "annealing temperature" (e.g. 50-60ºC) where primers anneal to their specific targets on the single stranded DNAs. The annealing temperature is chosen to be a few degrees below the melting temperature of the primers (i.e. the temperature needed to break the hydrogen bonds of the primers with a complementary targets). Thus as the temperature is reduced, the primers look for good targets to anneal to, and stay annealed to their corresponding targets at the actual annealing temperature. Note that the forward and reverse primers are designed to have similar melting temperatures.

- **Elongation (or extension) phase** - DNA polymerase present in the reaction mix is always looking for free 3' ends present along the DNA to which it can bind to and replicate the single stranded portion of the DNA (DNA polymerase always elongates in the 3' to 5' direction). In the case of PCR, these are the 3' ends of the primers annealed to the single stranded DNA. This process is set at the optimal operational temperature for DNA polymerase, 72ºC, although one can also use lower temperatures. Although the elongation temperature is usually above primer melting temperature, the primers do not melt off the DNA since the elongation stabilizes the association of primers to their targets.

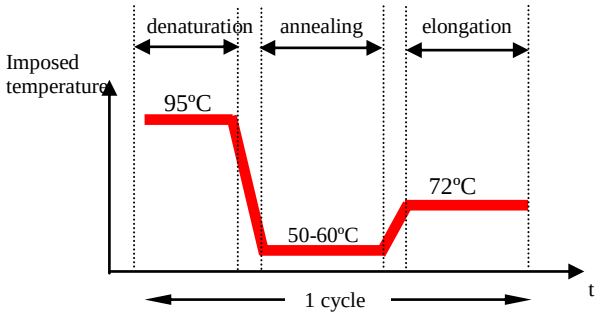


Figure 2. A typical imposed temperature profile during one cycle of a PCR reaction.

Problem 2: Exponential amplification

PCR is an exponential process: for an ideal reaction, the total number of molecules doubles every cycle. Exponential amplification adds up very quickly and allows us to amplify and detect a minute quantity of template, even a single molecule.

A. Assuming you start out with a single copy of the template, compute the minimum number of cycles required to generate at least 100 ng of amplicon DNA, which is the minimum amount of DNA needed in order to detect the product on an agarose gel. Assume that the template is 1000 base pairs long (1 base pair has a mass of 650 daltons; 1 dalton = $1.66 \cdot 10^{-24}$ g). You may also assume that the amplification is ideal, i.e., the number of molecules doubles every cycle.

We first need to compute the mass of a single template. Since each base pair has a mass of 660 Da, a 1000 bp long template will have a mass of 1000 bp X 650 Da = 650 kDa. To covert this into units of ng, we recall that 1 Da = 1/NA grams, where NA is Avogadro's number and is equal to $6.022 \cdot 10^{-23}$ mol^{-1}. Therefore 1 Da ~ $1.66 \cdot 10^{-24}$ g. Thus, one template has a mass of:

Mass of 1 template = $(650 \cdot 10^{3}$ Da) X $(1.66 \cdot 10^{-24}$ g/Da) = $1.08 \cdot 10^{-18}$ g.

Therefore we need to generate more than $(100 \cdot 10^{-9}$ g)/($1.08 \cdot 10^{-18}$ g) ~ $9.3 \cdot 10^{10}$ copies of template to obtain at least 100ng of product.

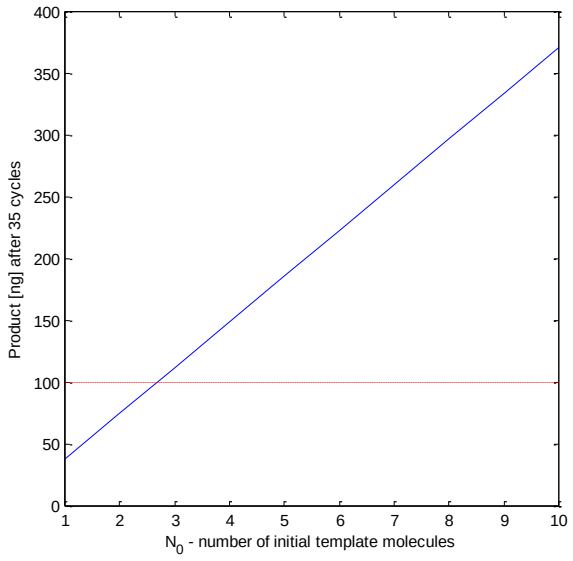
Assuming that amplification is 100% efficient, that is at the end of every cycle the total number of templates exactly doubles, then the minimum number of cycles needed to generate 9.1·1010 copies of the template is

$2^n > 9.3 \cdot 10^{10} \rightarrow n > \log_2(9.1 \cdot 10^{10}) \rightarrow n = 37$.

Therefore a minimum of 37 cycles is required to generate at least 100ng of product DNA. The number of PCR cycles is usually between 30 and 40 cycles. Note that this estimation is approximate since in reality the efficiency of amplification is reduced as the reaction progresses until amplification saturates at some point. Also, the more cycles we add the more chance that nonspecific products will appear. In practice, when dealing with cases where the number of initial template molecules is very low, more sensitive and more quantitative fluorescence based detection methods exist as an alternative to running the product on a gel that can allow detection at much earlier cycles.

B. Plot a graph of the total mass of amplified molecules (in units of ng) at the end of cycle 35 as a function of the number of initial template molecules for 1,2,3 … 10 initial template molecules. How many template molecules are required to pass the threshold for detection on an agarose gel? (You can make the same assumptions as in part A)

Below we plot the number of initial template molecules N0 as a function of the product mass after 35 cycles $M(n=35)=N0 \cdot 2^{35} \cdot (1.08 \cdot 10^{-18}g) \cdot 10^{9}$ ng/g = 37.1·N0. From the graph below it is apparent that at least 3 molecules are required to pass the detection threshold when using an agarose gel.



Problem 3: Restriction Enzymes

A. Restriction enzymes are proteins that recognize specific sequences at which they cut the DNA. Two commonly-used restriction enzymes are HindIII and EcoRI.

i. **Look up the recognition sequences that these enzymes cut and make a sketch of the pattern of cutting they carry out.**

To answer this question go to the Enzyme Finder in the New England Biolabs (NEB) website (http://www.neb.com/nebecomm/EnzymeFinderSearchByName.asp) and enter the name of the restriction enzyme you are interested in (make sure you search by "name"). Find the double-stranded (two rows) sequence; the forward slashes indicate where the backbone of the DNA is cleaved.

HindIII recognition site:

5′...A▼AGCTT...3′
3′...TTCGA▲A...5′

EcoRI recognition site:

5′...G▼AATTC...3′
3′...CTTAA▲G...5′

ii. **Consider the roughly 48,000 bp genome of λ (lambda) phage (a bacterial virus) and make an estimate, based on simple probability arguments, of the lengths of the fragments that you would get if the DNA is cut by either HindIII or EcoRI (remember that lambda DNA is linear). There is a precise mathematical way to do this that you**

should derive, and it depends upon the length of the recognition sequence. Based on your derivation, explain why a 5 cutter (i.e. a restriction enzyme that recognizes a 5 base pair sequence) will have shorter fragments than an 8 cutter.

Recognition sites for HindIII and EcoRI are 5'AAGCTT3' and 5'GAATTC3', respectively. To estimate the length of fragments resulting from cutting by either HindIII or EcoRI, we calculate the probability of the occurrence of these six consecutive bases in the lambda genome. Since there are four "natural" bases, the probability for recognition N base sequences is $P_N = (1/4)^N$. which implies that mean length of a fragment is 4^N bp. For a piece of DNA length L, the number of fragments on average will be $L/(4^N)$. However, because lambda DNA is linear, the number of fragments will be $L/(4^N) + 1$. For each HindIII and EcoRI, the number of fragments is calculated to be $L/(4^6) + 1 = 13$ for linear, lambda DNA (for circular DNA, the number of fragments would be 12) with fragment sizes of $4^N = 4^6 \sim 4000$ bp. For 5- and 8-cutters, the number of DNA fragments are predicted to be 48 and 2, respectively (or 47 and 1 for circular DNA).

B. We did this digestion experiment with lambda DNA in class. In preparation for viewing the digested results next week, find the actual fragment lengths obtained after the lambda genome is treated ("digested") with these restriction enzymes by going to the New England Biolabs website:

http://www.neb.com/nebecomm/tech_reference/restriction_enzymes/sites/Lambda_sites.pdf

Look up the tables that they have for identifying the sites on the lambda genome that are cut by these different enzymes. **Calculate the fragment sizes for**:

a. **a single digestion with HindIII**
   HindIII: 23130, 25157, 27479, 36895, 37459, 44141 (6 cutting sites)
   The fragment sizes are: 23130, 2027, 2322, 9416, 564, 6682, 4361.

b. **a single digestion with EcoRI**
   EcoRI: 21226, 26104, 31747, 39168, 44972 (5 cutting sites)
   The fragment sizes are: 21226, 4878, 5642, 7421, 5804,3530

c. **a double digestion with both HindIII and EcoRI simultaneously**
   From the restriction map, we can easily get the precise length of the fragments: **21226bp, 1904bp, 2027bp, 947bp, 1375bp, 4268bp, 5148bp, 564bp, 1709, 4973bp, 831bp, 3530bp**

C. Based on the fragment sizes you found in part B (a. and b.) is the prediction that you made in part A, ii. a good one?

The actual number of cutting sites for HindIII and EcoRI are 6 and 5, respectively, which are within factor of two with our prediction. This shows that our assumption that the genome as a random sequence of A's,T's,C's, and G's is a poor assumption. From our calculation, the average fragment length is predicted to be $4^N = 4^6 \sim 4000$ bp.

D. Below is an image of a gel electrophoresis sample showing the following (from the left):

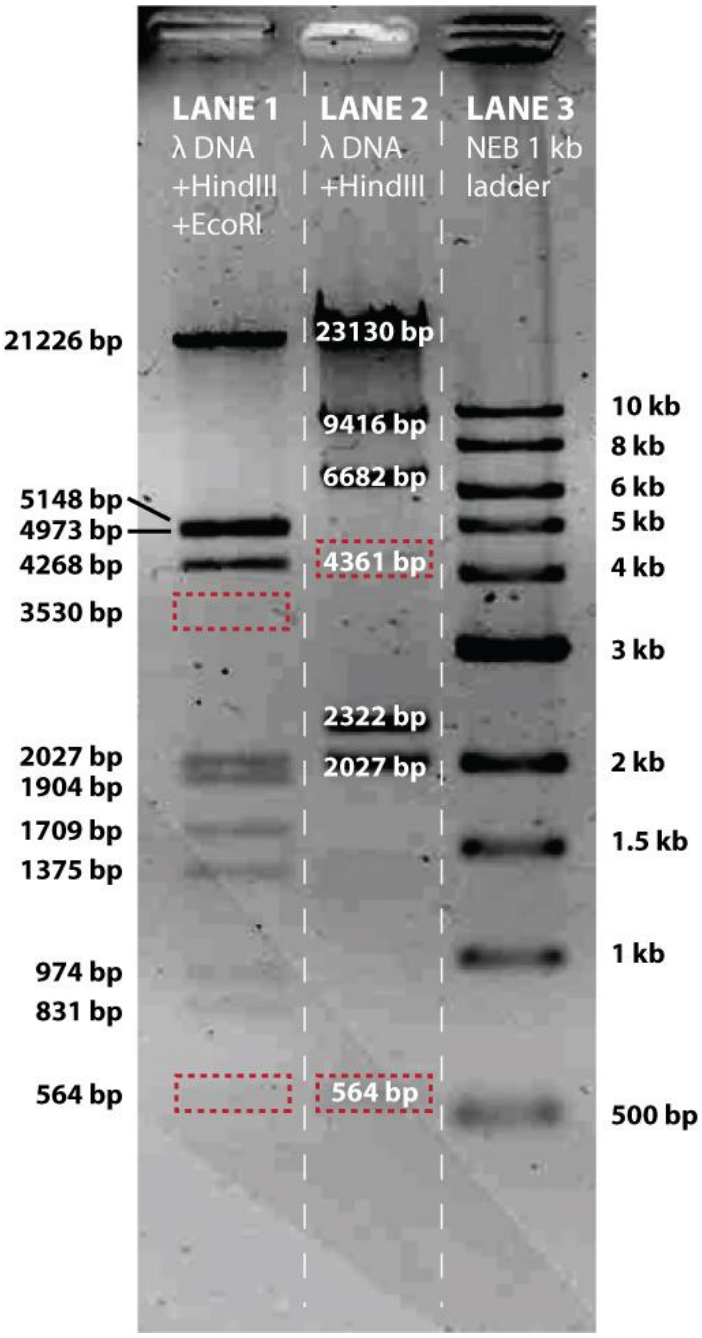
Lane 1: a double digestion of λ DNA with HindIII and EcoRI simultaneously.
Lane 2: a single digestion of λ DNA with HindIII
Lane 3: 1kb NEB ladder.

Each horizontal band in the gel corresponds to a fragment of a different size. You can estimate the size of the fragments by comparing them to the fragments in the right most lane (lane 3), which is a sequence of fragments of known size, called a ladder. To obtain the fragment sizes for this ladder (called a 1kb NEB) look up the following NEB site:

http://www.neb.com/nebecomm/products/productN3232.asp

i. **Match the fragment lengths for the ladder based on the website above with the actual pattern of fragments in the right most lane**.

ii. **Now use the ladder as a reference and try to match the fragment lengths you calculated in part B with the actual bands that appear in the remaining two lanes. Note that for experimental reasons, some of the bands may be too faint to make out. These have been boxed in red on the diagram below. When referring to the gel, make sure you explicitly identify the fragments that you think correspond to your expected results.**


LANE 1
λ DNA
+HindIII
+EcoRI
LANE 2
λ DNA
+HindIII
LANE 3
NEB 1 kb
ladder
21226 bp
5148 bp
4973 bp
4268 bp
3530 bp
2027 bp
1904 bp
1709 bp
1375 bp
974 bp
831 bp
564 bp
23130 bp
9416 bp
6682 bp
4361 bp
2322 bp
2027 bp
564 bp
10 kb
8 kb
6 kb
5 kb
4 kb
3 kb
2 kb
1.5 kb
1 kb
500 bp