CHEMENG 355 Spring 2018

Problem Set 2 Due April 26, 2018

Problem 1: Error-Prone PCR

When performing PCR, it is important to ensure fidelity in DNA replication to obtain accurate products. Let's investigate how mutation rates over the course of replication impact PCR accuracy. Using a high-fidelity DNA polymerase, we plan to amplify a 1000-bp gene from plasmid DNA, starting with 10 pg (ca. 10 million copies) of double-stranded template DNA. For simplicity, you may assume that the polymerase always terminates replication at the end of gene of interest.

a) Assume the polymerase makes no mistakes. Derive a formula for the total number of PCR amplicons that are produced from *n* copies of template dsDNA after *k* PCR cycles (i.e., k = 1 after 1 PCR cycle, where one cycle encompasses denaturation, annealing, and extension).

Suppose a distracted researcher uses manganese chloride ($MnCl_2$) instead of magnesium chloride ($MgCl_2$). The Mn^{2+} causes the polymerase to make mistakes much more frequently. Unintentionally, this researcher thus performs error-prone PCR, which is commonly used to construct mutant libraries.

b) For a single cycle of PCR, what is the probability that a completely correct copy of the gene is generated?

c) Assume the polymerase makes a mistake once every 10000 nucleotide additions ($\mu = 10^{-4}$) in the presence of Mn^{2+}. Derive a formula for the total number of correct PCR amplicons that are produced from *n* copies of template dsDNA after *k* PCR cycles. *Hint:* assume that each nucleotide addition is a statistically independent event.

d) Using your result from part b, write the formula for the total number of mutated PCR amplicons produced from *n* copies of template dsDNA after *k* PCR cycles, where a mutated strand is defined as having one or more errors in the DNA sequence.

e) What is the percentage of mutated strands of DNA after *k* = 1 cycle? 5 cycles? 35 cycles?

f) How does this percentage of mutated strands after 35 cycles compare to a typical PCR reaction involving $MgCl_2$ and a high-fidelity polymerase ($\mu = 4.4 \times 10^{-7}$)?

g) (Bonus) Using $\mu = 10^{-4}$, after 20 cycles of PCR, on average, how many strands will contain exactly 3 errors? Approximate your answer to 3 significant figures.

Problem 2: Searching Genomic Libraries

Tardigrades, also known as “water bears” are some of the toughest creatures you may have never heard of. These small aquatic species have an extraordinary ability to withstand extreme stresses, which include extreme temperatures (~-273 °C to ~100 °C), pressures, complete immersion in organic solvents, irradiation, and even being expelled into open space by the impact of a comet. (Hashimoto *et al.*, Nat. Comm., 2016, 7, 1-14)

You are a grad student who has had the good fortune to have discovered a new, highly stress-resistant tardigrade species from a sample from a pond outside of your lab. It is known that much of the tardigrade’s stress resistance derives from a protective protein coating on the surface of the tardigrade body.

Obviously, being able to determine the gene encoding for this rare protective protein would be of interest to many human scientists. You decide to create a gene library using the tardigrade genomes that you isolate from your pond sample.

a) The size of the tardigrade genome is 100 Mbp. Let’s say you have materials to generate 10,000 clones. How big should your genome fragments be to have a 99% probability that the protein coat-coding gene is contained in this library?

b) One tacit assumption here is that the fragment length is significantly longer than the length of our target gene. If the fragments are shorter, the chance of a fragment containing only a portion of the gene is higher. If we do not make this assumption, and let G be the length of the target gene (and *g* be the ratio of the target gene length to length of the genome), how does the library equation change?

c) Coverage (C) is defined as the average number of times a given nucleotide from the genomic sequence is represented in a genomic library and is therefore simply $C = N*F/L$, where N is the number of clones, L = the length of the genome, and F is the average length of each fragment. For 10x coverage of the tardigrade genome, what is the probability that a nucleotide is represented in the fragment pool?

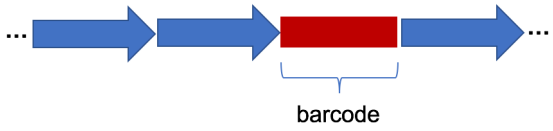
   Hint 1: Answer is independent of genome length
   Hint 2: $\ln(1 + x) \approx x$ for $|x| \ll 1$

d) Bonus: These rare proteins are certainly useful, but it might not be immediately obvious how one would construct a functional assay to know that they had been produced in a clone. Think of two possible ways that one could do this based on information given in the problem statement.

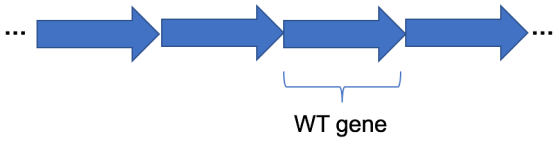
Problem 3: Preparing DNA Libraries by PCR

In class we discussed how a genomic library could help you find a specific gene of interest (e.g. the gene that encodes for GFP). Another way to find genes that have a certain function is to use a mutant library. One example of a mutant library is the yeast deletion collection (produced by the Stanford Genome Technology Center) which consists of 5000 unique yeast strains that are identical except that each contains a different single null mutation. In this collection, there is at least one strain for almost every non-essential gene in the yeast genome where the wild-type gene has been individually deleted and replaced with a unique known barcode (a synthetic DNA sequence). Barcodes can be read by PCR and sequencing to determine which gene has been deleted in each particular strain.

DNA from an example mutant strain, where a single gene has been replaced with a synthetic DNA barcode sequence (in red):
(we'll discuss how genes can be edited e.g. to make a mutation and/or to add a barcode sequence in future lectures)



DNA from Wild Type (WT) yeast strain (not mutated):



Researchers have used this collection for over a decade to study the effects of perturbing individual genes on yeast physiology. Suppose you are interested in determining which genes are important for the growth of yeast at higher temperatures as an example of a stress condition.

In the yeast genome, there are genes that are crucial for survival at high temperature. Cells in which these genes are deleted will not grow well under this stress condition compared to normal growth conditions. If we use PCR and sequencing to determine the barcodes present in the strains that grow in the normal conditions compared to those that grow in the stress condition, it should allow us to find genes that are important for growth at high temperatures.

a) You begin by taking an aliquot of the yeast deletion collection. You need to ensure that you gather enough cells such that all library members are well-represented. What is the minimum number of cells you need if 99% of all null mutant strains is each to appear at least once in the final library pool?
(Hint: to do this, write a few lines of computer code or use a software tool to simulate randomly picking from a list of 5000 library members N times, where N is an initial guess at the number of cells you need, and adjust your guess for N to get close to 99%.)

b) We are interested in quantifying the abundance of each library member (i.e. each barcode) in the cell population before and after exposure to the stress condition. To amplify the barcodes in each pool, we would like to begin with, on average, 1000 copies of each barcode to ensure that inefficiency in PCR and sequencing does not alter each member's representation. What concentration of total library DNA is needed to attain those 1000 copies? Assume these copies are contained in a volume of 50 microliters.

c) Over-amplifying a DNA library can cause depletion of primers, which leads to the formation of chimeric products and other undesirable effects. Practically, this means limiting the final DNA concentration achieved by PCR to less than 1 nM, which is enough for subsequent sequencing. How many cycles of PCR are necessary to amplify the initial DNA to this concentration?

d) You now sequence the pool of barcodes obtained from the cell population before and after exposure to high temperatures. How can you tell which genes are important for cell survival under these conditions? Are you able to say something about the relative importance of these genes?