CSE 527 Larry Jean
10/16/06 leijean@amath.washington.edu

Lecture 6

PCR:

- Double stranded DNA, heat, comes apart, cool down, primer specific to parts of DNA attach, polymerase start replicating
- One strand gives two copies of same DNA between primers
- Need to have unique primers

Gel Electrophoresis:

- DNA backbone negatively charged, put in electric field, migrate toward positive pole
- Get separation of nucleotides
- Basis for DNA sequencing (Sanger)
- OH group on nucleotide triphosphate of DNA
    - Bridge to next phosphate group to create sequence

Fluorescent Sequencing:

- Suppose primer at one end, polymerase copying. If a small fraction of the pool of DNA bases are "defective" (ie, the necessary OH group removed, a so-called di-deoxy nucleotide), when one of these bases happens to be inserted, chain growth will stop.
- If furthermore, the di-deoxy nucleotides are fluorescently tagged, say all ddAtp's are one color, all ddGtp's are another color, etc, then, can determine where the A's & G's are in the sequence are by examining distance between colors via gel electrophoresis.
- Trace of colors for nucleotides at different positions give nucleotide sequence
- 4 color sequence fluorescent + gel electrophoresis separate by size
- Advantage: fast and cheap!
    - Question: gel electrophoresis can't tell how many nucleotide in sequence
    - Answer: can make it longer to get greater resolution, use different gel materials
- *Problem*: can't do entire chromosome by fluorescent because resolution gets bad at greater lengths
- *Solution*: "Whole Genome Shotgun"

Whole Genome Shotgun:

- Cut genome into pieces (a, b, c, d, etc.), clone them into organisms (ie, bacteria stable in lab), fragments can be pieced back together according to similarity of sequences
- can read 600 nucleotides individually in one run
- *Problem*: not 100% accurate
    - Eg. 5 a's in a row can get a stretch of peaks, hard to identify exactly how many a's
- *Confusion*: do they fold up? Sequencing artifacts different if fold vs. unfold
- *Complication*: genome full of repetitive fragments (ie, short tandem repeats) artificial juncture of fragments in chimeric DNA fragments
- "top down" strategy: clone of large chunks → subclone → subclone → … → map back
    - Danger: repeats become large scale problem in recovery
    - Can get fairly good estimates of length of fragment to get around this problem
- Centromere too complicated to map, probably will never be fully sequenced
- Cost driven sequencing of genome very hot now
- Other methods for sequencing genome:
    - Student: Look into "pyro" method
    - Pull molecule into nanopore and read off the sequence
- *Goal*: make it fast and cheap to sequence nucleotides

Probability Basics:

- Sample space discrete or continuous
- Distribution describes probability of event occurring
- Probability density(mass) function continuous(discrete)
    - function value does not give probability
    - area under curve up to x gives probability of sample <= x
- Normal Distribution
    - symmetric about mean
    - variance (sigma^2) defines how spread out curve is, or how close tend to be from mean
    - "square" in "(x-mu)" term gives equal spread from left and right of mean
    - bell shaped
- Expected Value
    - expected outcome in the long run
    - eg. If roll 6 = win(+1) and otherwise lose(-1), the expected value is -2/3, ie, will lose 2/3 on average in the long run
- continuous → integral; discrete → summation
- *Important*: expected value of random variable itself
- Population vs. Sample
    - big difference between mean and variance
    - can't always perform operation on population, but ok on sample
    - as sample size gets large, sample mean tends to population mean

Parameter Estimation:

- Distributions defined by parameters
- eg. Normal distribution: theta=(mean, variance)
    - estimate population mean/variance from sample
- How to estimate parameters? Many ways; we'll look at two: MLE and EM

Maximum Likelihood Estimation (MLE):

- Likelihood of sequence of operations is product of probability density of each operation under assumed parameters (assume sample independent)
- *Goal*: find value of theta that maximizes likelihood of observed data
    - eg.1: Height of sample in room is 5ft, if assumed theta=10ft, then likelihood low
    - eg.2: Coin flip:
        - flip 1000 times, head turns up 642 times. If assumed theta=prob(head)=0.5, then NOT likely to maximize likelihood of data. If theta=0.642, then more likely
- Need to find theta that maximizes L, the likelihood function (smooth function of theta)
- How to maximize L?
    - Take derivative of L with respect to theta and set equal zero is point that maximizes L (or minimizes it…)
    - Problem: differentiating L: big product is a mess
    - Solution: take differentiation of log L (log likelihood), which is derivative of summation (much easier to work with)
- *Warning*: max/min of logL can be on boundary, so need to verify that it is indeed max
- eg3: x_i normally distributed, know variance, mean is parameter
    - Convention: use theta_hat to represent the theta that maximizes likelihood
- eg4: x_i normally distributed, both variance and mean unknown parameters
    - Take partial derivatives wrt theta1 and theta2 separately
    - Result: theta1=sample mean, theta2=sample variance
    - Problem: theta2 is consistent but biased!
    - If pick mean in the middle, highly unlikely that two samples will have the same distance away from mean. The sample mean is an unbiased estimate of the population mean (e.g., it's equally likely to be too large as to be too small), but its placement in the middle of the sample will tend to underestimate the variance.
    - In the extreme where sample size is 1, then theta1 still makes sense as an estimate of population mean, but variance estimate is zero!
- *Important*: MLE is great, but not perfect (of course); need to choose method appropriated for individual need
- *Goal*: have measured data → have model → likelihood of data → choose parameter (theta) that maximizes likelihood
- A more complex example:
    - Mixture of two normal distributions
    - Not good: assume one distribution among all data
    - Better: have multiple distributions, identify which point came from which distribution.
        - might be hidden variables and need to separate them.
        - Sample separated but don't know which came from which. Start by guessing, then re-estimate
        - More next time…