Notes – 10/15/2007

By Bart Trzynadlowski (trzy@u)

See lecture slides for background material. Below are some comments regarding things that were mentioned in class which may not be detailed in the slides.

Structure of Random Sequences:

Earlier, we assumed that random sequences could be generated by taking permutations of an input sequence. All nucleotides were fair game and could be rearranged in any order relative to each other. In reality, pairs or triplets of nucleotides appear with different frequency than individual nucleotide statistics would suggest. For example, assuming G and C each have a 1/4 chance of occurring, GC together would not necessarily be 1/16.

$$P_{GC} \neq P_G \cdot P_C$$

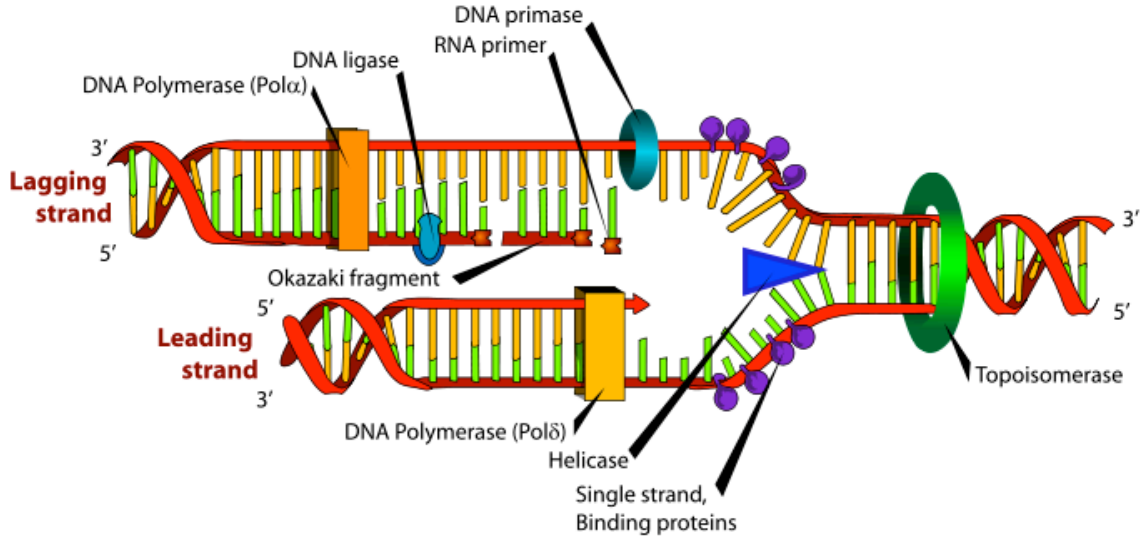
Therefore, an algorithm for generating random sequences may have to take such structural information into account.

Weekly Bio Interlude:

This week's bio interlude is in the slides for lectures 2-3.

DNA replication is handled by DNA polymerase.

Synthesis is 5' to 3' end (the new portion) and the polymerase crawls along from 3' to 5'. On leading stand, synthesis proceeds in a continuous manner but on lagging strand, it happens in short bursts: Okazaki fragments are formed in between sequences of RNA primer.



The above image is from Wikipedia (public domain image.) The article on DNA replication is here (I thought it served as a good complement to the lecture): http://en.wikipedia.org/wiki/DNA_replication

Probability, Parameterized Distributions, and MLE:

Taken from last year's notes by Larry Jean; a good overview of what we covered in class. Almost all of it is in the slides, however.

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 samples are 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