PhyloMath Lecture 3

Norm Wickett (10 February 2004)

Before we got started, we took a scenic diversion from the "main line" of the class and took a look at the idea of **expected values**.

To do this, consider first a simple average of 10 numbers: 3 5 2 2 2 2 1 4 2 2
The average of this is simply the sum of these numbers divided by 10, which can be rewritten as:

$$\frac{3(1) + 5(1) + 2(6) + 1(1) + 4(1)}{10}$$

or: 3(1/10) + 5(1/10) + 2(6/10) + 1(1/10) + 4(1/10) where each term is a product of the value and its probability (observed proportion). For a simple average then, the expected value of a random variable (y) can be expressed as a sum:

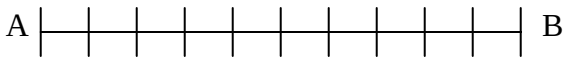
$$E(y) = \Sigma_y (y)pr(y)$$

However, we usually expect that the values we observe are drawn from some probability distribution (eg. Binomial or Poisson), which is the term pr(y) in the above sum. We usually don't have a simple average like this example, so the expected value is essentially what we would expect the mean to be, given a theory.

Note added by Paul: When pr(y) comes from a probability distribution (e.g. the binomial distribution), the expected value represents the value you expect for the simple average. Imagine drawing 20 samples each comprising 10 numbers drawn from a binomial distribution. Each sample would yield different observed proportions, but these observed proportions would hover around the pr(y) values predicted by the binomial model, and thus the 20 simple averages computed would hover around the E[y] value computed using the model.

Back to, and forward from, the "Dumb Substitution Model"

If you take a look back at Jon "Richie" Richmond's notes from the previous class, you will notice the last section treats a simple model of calculating an expected number of substitutions over a given amount of time. Remember that we are interested in two sequences, A and B, separated by 10 trillion years.



Using the Bernoulli Probability Distribution and ten divisions (thus there can either be one substitution or no substitutions in an interval), we see that the maximum number of substitutions in 10 trillion years is 10. This is a little silly. A more realistic model would treat the time interval as an infinite number of divisions, where a substitution is possible at any point in time.

The probability distribution for this is the Poisson distribution:

$$Pr(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where λ is the expected number of substitutions, and is equal to pn in the binomial model (p = the probability of a substitution in the interval, n = the number of intervals).

Note added by Paul: Note that the expected amount of evolution is the same for both models, the difference is in the number of intervals and the probability of a substitution in any given interval.

Derivation of the Poisson distribution

The objective here is to go from a simple binomial distribution (which we used last week) to the poisson distribution, which more accurately describes the probability of substitutions over the time period we are interested in. It is tempting here to just drop in a TAMO[1], but since we are interested in the nitty gritty here, we'll see if we can take the binomial distribution and see how it becomes the poisson distribution as the number of intervals (n) approaches infinity, which can be expressed as:

$$\mathrm{Lim}_{n\to\infty} \binom{n}{y} p^y (1-p)^{n-y}$$

Since λ = pn, p = λ/n, and we can do a little algebra to get:

$$\lim_{n\to\infty} \left[\frac{n!}{y!(n-y)!}\right] \left[\frac{\lambda}{n}\right]^y \left[1 - \frac{\lambda}{n}\right]^{n-y}$$

you can take the term $\frac{\lambda^y}{y!}$ out and rewrite the limit as:

$$\frac{\lambda^y}{y!} \lim_{n\to\infty} \left[\frac{n(n-1)(n-2)\ldots(n-y)(n-y-1)\ldots(2)(1)}{n^y (1-\lambda/n)^y \ldots(n-y)(n-y-1)\ldots(2)(1)}\right] (1-\lambda/n)^n$$

you'll notice that the last (n-y) terms in the square brackets above cancel, (and noticing that there are exactly *y* terms left above and y *n*s below) giving you:

$$\frac{\lambda^y}{y!} \lim_{n\to\infty} \left[\frac{(n/n)(n-1/n)(n-2/n)\ldots(n-y+1/n)}{(1-\lambda/n)^y}\right] (1-\lambda/n)^n$$

If n approaches infinity, the term in the square brackets goes to 1 and $(1-\lambda/n)^n$ goes to $e^{-\lambda}$.

After a lot of algebra, we see that the limit we started with goes to $\frac{\lambda^y}{y!}(e^{-\lambda})$, which is the Poisson distribution.

[1] TAMO was coined by our own Kent Holsinger, and stands for Then A Miracle Occurs

Why is this important? Because the Poisson distribution forms the basis of all the substitution models (e.g. Jukes-Cantor, F-81…), so we can better understand the transition probabilities, and how we get E(y), the expected number of substitutions.

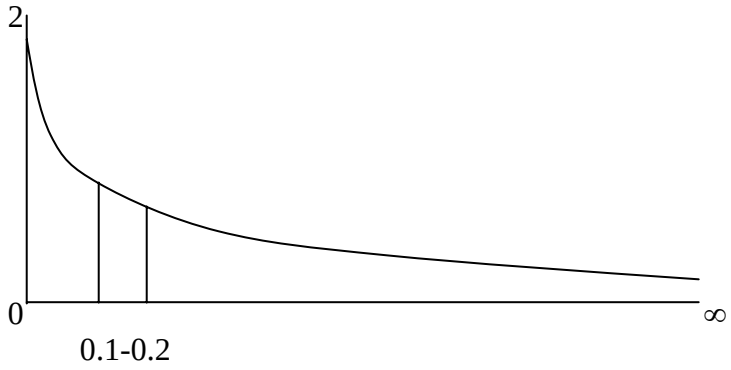
Binomial: $E(y) = pn$ (p = probability of a substitution occurring, n = number of intervals)

Poisson: $E(y) = \mu t$ (μ = instantaneous rate of substitution, t = time)

Probability densities

At this point you may be asking yourself, "Why is μ a rate, whereas p is a probability?"

A probability describes a discrete event, for example, the probability of a substitution occurring in an interval of time. If that interval becomes infinitesimally small, the probability of a substitution at that point becomes zero. In order to describe the probability of a substitution over continuous time, we need to use a probability density, rather than a probability. A probability density is a function (of x, say) that can generate probabilities if integrated over a specified interval on the x-axis. If the curve is integrated over a given interval, the probability of an event (eg. a substitution) happening in that interval can be determined.



So the probability that an event occurs between 0.1 and 0.2 on the x-axis is the integral of the curve (an exponential distribution in this case) from 0.1 to 0.2.

Note added by Paul: In reading Norm's excellent treatment, I realized that I described some aspects of μ poorly, so I have eliminated part of Norm's discussion here and will treat the topic again in the next lecture.

In the derivation of the Poisson distribution above, λ is equivalent to μt.

Now that we know how a Poisson distribution better describes the probability of substitutions, we can understand the probabilities of specific substitution types (or transition probabilities). Remember we defined two classes of transition probabilities under the Jukes-Cantor model: a)

the probability of a nucleotide staying the same[2] (eg. A→A) and b) the probability of a nucleotide changing (eg. A→C). This change happens between two sequences separated by time (t) with a certain substitution rate (α). You may remember:

$$P_{AA}(t) = {}^1/_4 + {}^3/_4 e^{-4\alpha t}$$
$$P_{AC}(t) = {}^1/_4 - {}^1/_4 e^{-4\alpha t} \quad (1)$$

These transition probabilities change depending on the model of evolution and they are what we actually use when we run a maximum likelihood analysis.

Derivation of the transition probabilities

The objective of the next part of this lecture is to understand how these probabilities are derived using the Poisson distribution. For this, it is important to distinguish between a *substitution* and a *disruption*.

A **substitution** is the change of a nucleotide (eg. A→C) at a particular site in a nucleotide sequence that becomes fixed in a population or a species.
A **disruption** can be thought of as a mistake that may temporarily disrupt a nucleotide site such that it may or may not change and become fixed. Disruption have four possible outcomes, all equally probable under the JC model. A disruption yields A→A "changes" just as often as A→C, A→G or A→T changes. Another way to say this is, after a disruption event, the probability of having any of the four bases is ¼.

For our purposes here we will let α = substitution rate, and μ = disruption rate.

If we consider the change A→C, the rate at which this happens is $\alpha = {}^1/_4\mu$.

Note added by Paul: why is α equal to ¼ μ? Because one expects only one fourth of all disruptions to end in C. Note that the results are the same if you consider any substitution. For example, the rate at which G→T substitutions is also α, and again, disruptions yield G→A, G→C, G→G and G→T with equal frequency, and only ¼ of these are G→T substitutions.

How does this help us figure out the substitution probabilities (A→A and A→C,G,T)

To calculate the probability of A→C over time t, it is simply Pr(at least one disruption occurred AND the base inserted after the final disruption was C). The probability of the first event (i.e. "at least one disruption occurred") is just 1 – Pr(no disruptions occurred), and to get this we can substitute 0 for (y) in the Poisson distribution:

[2] Note added by Paul: the phrase "staying the same" is a little misleading, because this transition probability also includes cases in which the nucleotide changed to something else, perhaps several times, but then ended up in the same state that was present at time 0. Perhaps a more accurate phrasing is "the probability of ending up with the same base you started with"

$$1 - \frac{(\lambda)^0 e^{-\lambda}}{0!} \quad = \quad 1 - \frac{(\mu t)^0 e^{-\mu t}}{0!} \quad = \quad 1 - e^{-\mu t}$$

The probability of the second event ("the base inserted after the final disruption was C") is simply ¼.

Here we need a brief review of joint and conditional probabilities. To calculate the joint probability of, for example, both event B **AND** event D occurring simultaneously, we can calculate either Pr(B) Pr(D| B) or Pr(D) Pr(B| D) *(check out figure 1 at the end of the notes)*. The vertical bar in conditional probabilities is read "given that". Taking event B to be "at least one disruption occurred" and event D to be "base inserted after the final disruption was C", we can write

Pr(B AND D) = Pr(B) Pr(D|B)

which in words is

Pr("at least one disruption occurred" **AND** "base inserted after the final disruption was C")
= Pr("at least one disruption occurred") Pr("base inserted after the final disruption was C" **GIVEN THAT** " at least one disruption occurred")

This gives:

$$(1 - e^{-\mu t})\, {}^1\!/_4 \quad = \quad {}^1\!/_4 - {}^1\!/_4\, e^{-\mu t}$$

$P_{AA}(t)$ = Pr(no disruptions) + Pr(ended up with an A **AND** at least one disruption occurred), so we need to sum two probabilities (corresponding to the two ways to get no difference in state) as follows:

$$P_{AA}(t) = e^{-\mu t} + {}^1\!/_4(1 - e^{-\mu t}) \quad = \quad {}^1\!/_4 + {}^3\!/_4\, e^{-\mu t}$$

Because $\alpha = {}^1\!/_4\mu$, we can substitute $e^{-\mu t}$ in the above equations and we get the transitions probabilities above (1).

Note added by Paul: the first part of the transition probability $P_{AA}(t)$ can be thought of as

Pr("no disruptions occurred over time t") Pr("ended with an A" GIVEN THAT "no disruptions occurred over time t")
= Pr("no disruptions occurred over time t")(1)
= $e^{-\mu t}$ (1)
= $e^{-\mu t}$

Note added by Paul: I told Norm he could stop here, as we will revisit the concept of sojourn times again in the next lecture.

Figure 1: Conditional and unconditional probabilites