Phylomath Lecture 5 – February 24, 2004
Jessica Budke

Program used during last week's lab – Sojourn time simulator

This is a PHP program associated with the Phylomath course website. There are no direct links to this program. To find the program, type in <sojsim.html> after the address for the course website.
(Try not to run to big a simulation or share with the world. Running from the EEB web server.)

When we looked at this sojourn time simulator last week we noted that it calculated the expected values for several statistics. Two of these values were the **expected proportion of similarities** and **expected proportion of differences.**

How are these two values calculated?

Before we determined how to calculate these values we reexamined the function of the Sojourn time simulator (STS). First the STS starts with a nucleotide sequence, and then a second sequence is simulated. A set length of time over which the sequence is allowed to evolve is determined. Sojourn times are calculated and at the end of each sojourn a disruption occurs. Based on the disruptions that have occurred, sequence 2 is determined. By comparing sequence one to two we can observe if there is a similarity between the starting and ending bases or if there is a difference between the two of them.

Example of products from the Sojourn time simulator

Sites	Seq 1	Timelines with sojourn times	Seq 2	
1	A		A	Similarity
2	C		C	Similarity
3	G		T	Difference

Using results such as these the **expected proportion of similarities** and **expected proportion of differences** were calculated.

We started with the **longer** way to calculate the **expected proportion of differences.**

$$\Pr(\text{difference}) = \sum_j \sum_{j \neq i} \Pr(\text{starting base}) \Pr(\text{ending base} \mid \text{starting base}, t, \mu)$$

$$= 12 \left(\tfrac{1}{4}\right)\left(\tfrac{1}{4} - \tfrac{1}{4} e^{-\mu t}\right)$$

$$= 12/4 \left(\tfrac{1}{4}\right)(1-e^{-\mu t})$$

$$= \tfrac{3}{4}(1-e^{-\mu t})$$

$\sum_j$ = the sum of the number of all possible starting bases = 4

$\sum_{j \neq i}$ = the sum of the number of all possible ending bases = 3

We then calculated the same **expected proportion of differences** by a **shorter** method.

$$\Pr(\text{diff}) = \Pr\left(\begin{array}{c}\text{first sojourn time less}\\ \text{than the length of the branch}\end{array} \,\middle|\, \mu, t\right) \Pr\left(\begin{array}{c}\text{the last disruption yielded}\\ \text{a different base than start}\end{array}\right)$$

$$= [\,1 - \Pr(\text{no sojourn time} < t)]\ (¾)$$

$$= (1-e^{-\mu t})\ (¾)$$ [Reminder: $e^{-\mu t}$ = the probability of zero disruptions before time(t)]

Using either the shorter method or the longer method resulted in the same final equation. The **expected proportion of similarities** can be calculated by subtracting the **expected proportion of differences** from one.

It was mentioned that regardless of how many disruptions occur over time t, it is only the last one that matters.

Recap of how to derive $e^{-\mu t}$

There is a more detailed explanation and set of formulas to derive $\mathbf{e^{-\mu t}}$ in the notes from February 10th (second page).

Binomial Formula: $\Pr(y) = \binom{n}{y} p^y (1-p)^{n-y}$

$$\lambda = pn \qquad n \to \infty, \quad p \to 0$$

Poisson Probability $\Pr(y) = \dfrac{\lambda^y e^{-y}}{y!}$

$$\Pr(0) = \frac{\lambda^0 e^{-y}}{0!} = e^{-\lambda} \qquad \lambda = \mu t$$

Simulation: the parameters μ, t are known values and they are used to simulate data
Estimation Method: the data is known and used to estimate the parameters μ, t

Estimation Methods

1. Method of Moments
2. Least Squares
3. Maximum Likelihood
4. Minimax Method (methods based on minimizing costs)
5. Bayesian Posterior Summaries

1. Method of Moments

For this type of estimation method the goal is to estimate the following:

$d = 3/4\ \mu t = 3\alpha t$ = expected number of substitutions per site across a branch of length t and rate μ

The value of d is used because there may be multiple rates that could be used in place of μ. For example, a model allowing transition- and transversion-type substitutions to occur at different rates. Using the expected number of substitutions per site (d) is an effective standard way to specify a branch length that has the same meaning across many different models.

$E(x)$ = 1st moment = mean = expected value

$E(x^2)$ = 2nd moment

In order to determine the expected number of substitutions across a branch of length t, we again start with two sequences (either simulated or real life data) and determine the actual proportion of similarities and differences between the sequences (call this statistic p).

$p = pr(diff)$ [Observed proportion of differences]

$E(p) = ¾\ (1-e^{-\mu t})$ [the 1st moment → Expected proportion of differences]

Once again everything in this class leads to algebra.

$p = ¾\ (1-e^{-\mu t})$ [We substitute the observed proportion of differences for the expected proportion of differences and then solve for the unknown value μt]

$(4/3)p = 1-e^{-\mu t}$

$\ln[1- (4/3)p] = \ln(e^{-\mu t})$

$\ln [1- (4/3)p] = -\mu t$

$-\ln [1- (4/3)p] = \mu t$ [Multiply both sides of the equation by ¾]

$-(¾) \ln [1-(4/3)p] = (¾)\mu t = d$ [Then the right side of the equation is equal to d]

This equation allows us to determine the expected number of substitutions across a branch of length t and rate μ when we know the observed proportion of differences (p) between our sequences.

Here is an example that we performed later in the class for calculating the **expected number of substitutions per site.** (This may be located in a different place in your own notes.)

This example is based on a handout (see next page) that Paul gave us during class, regarding 30 nucleotides of the ψη-globin gene of gorilla and orangutan. From these sequences we can determine the number of differences and similarities between the two.

$$\left.\begin{aligned} n_{diff} &= 2 \\ n_{sim} &= 28 \end{aligned}\right\} \text{out of } 30$$

d = -3/4 ln [1 – (4/3) (2/30)] = -3/4 ln (0.91111) = 0.069817

d = 3αt = 3(0.02327) = 0.06981 [αt =0.02327; from the handout]

2. Method of Maximum Likelihood

- makes use of all the data

$L = Pr(\text{ data} \mid \mu, t)$ [We are determining the probability of the data given μ and t. Remember that μ and t cannot be estimated separately. They must be estimated as a product]

$$= [Pr(\text{diff} \mid \mu t)]^{n_{diff}} [Pr(\text{sim} \mid \mu t)]^{n_{sim}}$$

$$= \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} - \frac{1}{4}e^{-\mu t}\right)\right]^{n_{diff}} \left[\left(\frac{1}{4}\right)\left(\frac{1}{4} + \frac{3}{4}e^{-\mu t}\right)\right]^{n_{sim}}$$

n_{diff}=number of sites that are different from seq 1 to seq 2

n_{sim}=number of sites that are similar from seq 1 to seq 2

[The first ¼ inside the brackets is the probability of the given starting base]

$$= \left[\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(1 - e^{-\mu t}\right)\right]^{n_{diff}} \left[\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(1 + 3e^{-\mu t}\right)\right]^{n_{sim}}$$

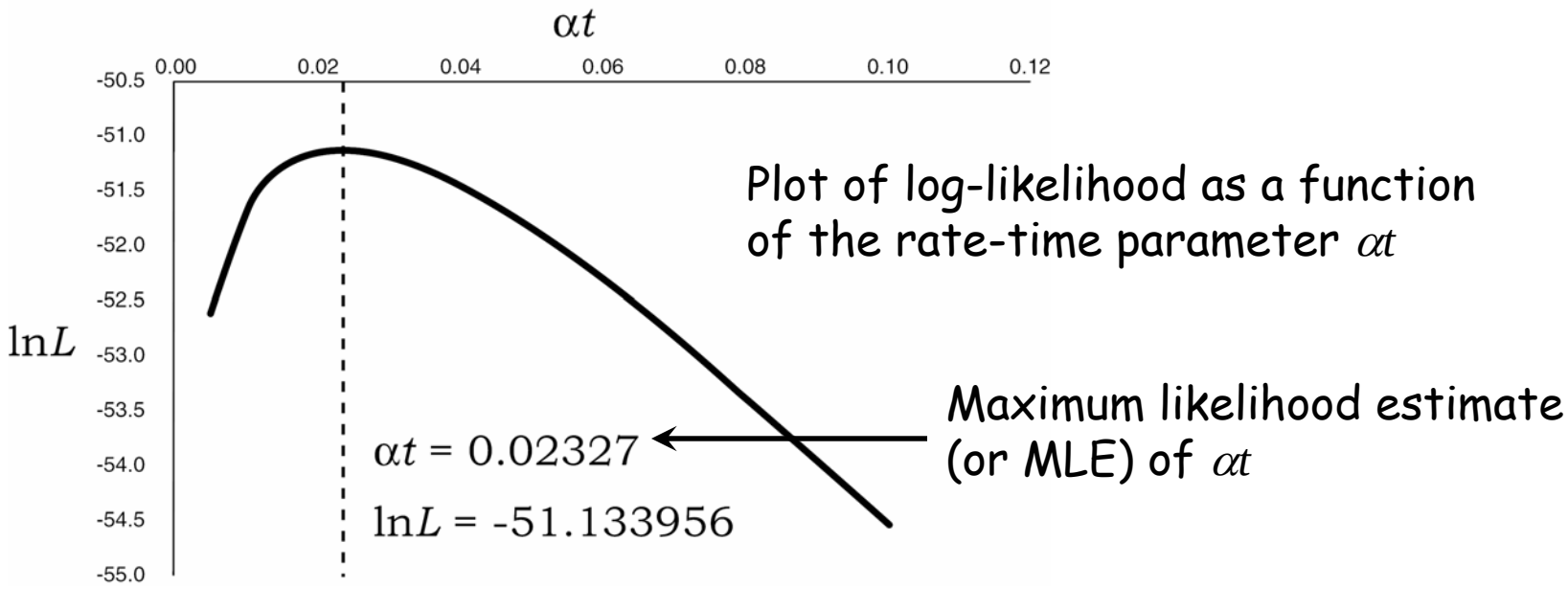
$$= \left(\frac{1}{16}\right)^{n} (1 - e^{-\mu t})^{n_{diff}} (1 + 3e^{-\mu t})^{n_{sim}}$$

An example

First 30 nucleotides of the $\psi\eta$-globin gene of gorilla and orangutan:

```
gorilla   GAAGTCCTTGAGAAATAAACTGCACACACTGG
orangutan GGACTCCTTGAGAAATAAACTGCACACACTGG
```

$$L = \left[\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4} + \tfrac{3}{4}e^{-4\alpha t}\right)\right]^{28}\left[\left(\tfrac{1}{4}\right)\left(\tfrac{1}{4} - \tfrac{1}{4}e^{-4\alpha t}\right)\right]^{2}$$



$$\ell nL = n\ \ell n\left(\frac{1}{16}\right) + n_{diff}\ \ell n(1 - e^{-\mu t}) + n_{sim}\ \ell n\ (1 + 3\,e^{-\mu t})$$

We didn't finish working through the rest of this equation during class. The end result should be the same equation that we resulted in using the Method of Moments. If you would like to complete the equation yourself, take the derivative of lnL with respect to μt and set it to 0, then solve for μt. (Paul decided to spare us the rest of the math.) If you would like to be walked through the equation see Paul. Otherwise you can just take it at face value that the equation results in:

$$d = \left(\frac{3}{4}\right)\mu t\ = -\left(\frac{3}{4}\right)\ell n\left[1 - \left(\frac{4}{3}\right)p\right]$$

Note added by Paul: I hate for you guys to take anything at face value, so let's start next Tuesday by filling in this gap!

Refer again to the handout with the sequences from the gorilla and orangutan.

We discussed why the graph of the log likelihood curve has the shape that it does.

The maximum likelihood is where the slope of the curve is 0

If αt goes to 0 then the rate of subst = 0 or time = 0

Thus the probability of seeing the 2 differences that we did observe is zero; the $\ell nL = -\infty$

If αt goes to ∞ then the probability of observing *only* 2 differences is difficult to explain.

Rate Heterogeneity

Some parts of most genes are under strong stabilizing selection while other parts are under relaxed selection. This causes different rates of change for different sections of a sequence.

Simulation Phase

pinv or pinvar (in PAUP) = Proportion of invariable (invariant) sites

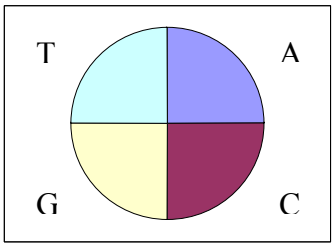
Let pinvar = 0.5 for purposes of this simulation
u ~ uniform (0,1); if u is less than 0.5, site will be **invariable**; if u is any other value, site is **variable** (which means it *can* vary, but is *not obligated* to vary)

Site	Used to determine →	Invariable?	Used to determine →	Start	End	Used to ← determine
1	0.172*	Yes	0.285	C	C	---
2	0.935*	No	0.964	T	T	0.831*
3	0.060	Yes	0.576	G	G	---
4	0.441	Yes	0.758	T	T	---
5	0.491	Yes	0.502	G	G	---
6	0.753	No	0.051	A	A	0.734*

* not the random numbers generated in class.

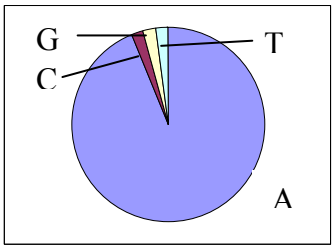
Starting Base: To choose the starting base (using site 6 as an example), we referred back to our choice tables, where each base has a ¼ chance of being the starting base.

For Site #6



Pr (A)

Ending Base: If the site is invariable then the ending base is automatically the same as the starting base. If the site is variable then the probability of the ending base depends on the starting base and μt.



Pr (A | A, μt)

μt = 0.1 (chosen arbitrarily)

Pr (same) = ¼ + ¾ $e^{-\mu t}$ = 0.9286 Pr (particular difference) = ¼ – ¼ $e^{-\mu t}$ = 0.0238

One problem is that the proportion of differences can be obtained by subtraction given the proportion of similarities (and vice versa). Thus we only have one degree of freedom.

p = probability of a difference
1- p = probability of a similarity

We can thus estimate only one parameter in the model when only two sequences are analyzed. Using more sequences in the analysis gives us more than two data patterns, which makes it possible to estimate more parameters in our model. Thus this inability to estimate both d and pinvar only applies to analyses of 2 sequences.

We then discussed the relationship between pinvar and μt. For two sequence analyses, these two values are highly correlated, to such an extent that one cannot be found without the other.

pinvar is directly related to μt
    when pinvar is high, μt also high
    when pinvar is low, μt also low

Note added by Paul: Why is this? When pinvar is high, it accounts for many of the constant sites (sites showing a similarity between the two sequences), leaving mt to explain the remaining sites, most of which show a difference. Thus, μt would need to be high to account for the fact that the sites that are variable nearly all show a difference between the two sequences. Now consider the case when pinvar is low, say 0.0. If pinvar is zero, then all the sites are at least potentially able to vary, and μt will be lower because it needs to explain the fact that many of these potentially variable sites did not in fact show a difference across the branch.

We ended the class with a bit about relative rates. However this topic will be covered in greater detail in upcoming lectures so this information will not be included here.