Feedback — Quiz 2: Models of Evolution (computer)

Help Center

You submitted this quiz on **Sat 20 Jul 2013 7:35 AM PDT**. You got a score of **12.00** out of **12.00**.

Overview

In this exercise we will explore a number of different, but closely related, models of evolution. Using such models it is possible to estimate the number of unseen mutational events and thereby obtain genetic distances that have been corrected for superimposed substitutions. It is, however, important to realize that these corrections are based on the assumption that we observe approximately the expected amount of change - if, for instance, 20 mutational events end up leading to no observable changes then it is impossible to guess the actual amount of change regardless of which correctional scheme we employ. Using more and longer sequences helps ensuring that the observed change is closer to the expected change, so the correction is more likely to be accurate with adequate amounts of data. The same models also play an important role in phylogenetic reconstruction based on maximum likelihood and Bayesian techniques.

Getting started, biological background.

1. **Open a Terminal window in Ubuntu:**

   Make sure to maximize the window: the analyses we will perform give lots of output to the screen, so having a nice and large shell window makes it easier to keep track of what happens.

2. **Construct working directory:**

   ```
   cd /home/student
   ```

   ```
   mkdir models
   ```

Note: The first command is to make sure that you start in your home directory before creating the working directory for this exercise. This is not strictly necessary if you have just opened a Terminal, since you will then already be in the right place. But, better safe than sorry...

3. **Change to working directory**

```
cd models
```

4. **Copy files for exercise:**

```
cp ~/data/primatemitDNA.nexus ./primatemitDNA.nexus
```

```
cp ~/data/titv.data ./titv.data
```

The tilde (~) is a brief way of specifying your home directory (in this case:`/home/student`).

5. **Inspect sequence file:**

```
nedit primatemitDNA.nexus &
```

This file contains an aligned set of mitochondrial DNA sequences from man, chimpanzee, gorilla, orangutan and gibbon. Mitochondria are cellular organelles that are bounded by a lipid membrane and contain their own genome. Mitochondrial DNA is related to certain bacterial genomes, and it is believed that the original mitochondrium was a primitive bacterial cell that was engulfed by an early ancestor of eukaryotic cells and that the pair subsequently went on to form a constant symbiotic relationship.

Mitochondrial DNA has a higher rate of substitution than nuclear DNA. This makes it useful for investigating phylogenetic relationships between closely related species, such as the five primates included in the present data set. Close the `nedit` window when you are done.

6. **Inspect additional data file:**

```
nedit titv.data &
```

This file contains a single header line and one column of numbers giving estimated times of divergence between man and chimpanzee, man and gorilla, man and orangutan, and man and gibbon. (Divergence times are in millions of years). This file will be used later in the

exercise when we investigate how various distance measures increase over time. **Note:** If the `nedit` window is too narrow, then the column headings will wrap over two lines. Make sure to make the window as wide as possible in order to understand the structure of this file. Close the `nedit` window when you are done.

Question 1

The Jukes and Cantor model.

The Jukes and Cantor model of evolution has the following rate matrix:

[Math Processing Error]

We will now use the gnuplot program to explore some features of evolution occurring according to this model.

1. **Start gnuplot program:**

   ```
   gnuplot
   ```

   You have previously used the gnuplot program to plot data sets from a file. The program can also be used to plot functions that are given in symbolic form.

2. **Examine expected amount of change as a function of branch length:**

   For the Jukes and Cantor model the following equation gives the probability, *[Math Processing Error]*, that a given site will display observable change, expressed as a function of branch length, *[Math Processing Error]*:

   [Math Processing Error]

   Here, *[Math Processing Error]* is measured in substitutions per site. *[Math Processing Error]* is also the *expected* fraction of sites showing observable change along a branch of length *[Math Processing Error]*: if any single site has probability *[Math Processing Error]* of changing, then on average *[Math Processing Error]* sites will have changed in a sequence of length *[Math Processing Error]*. We will now

explore how the *expected* amount of observable change depends on the branch length:

```
set xlab "Branch length"
```

```
set ylab "Observed difference"
```

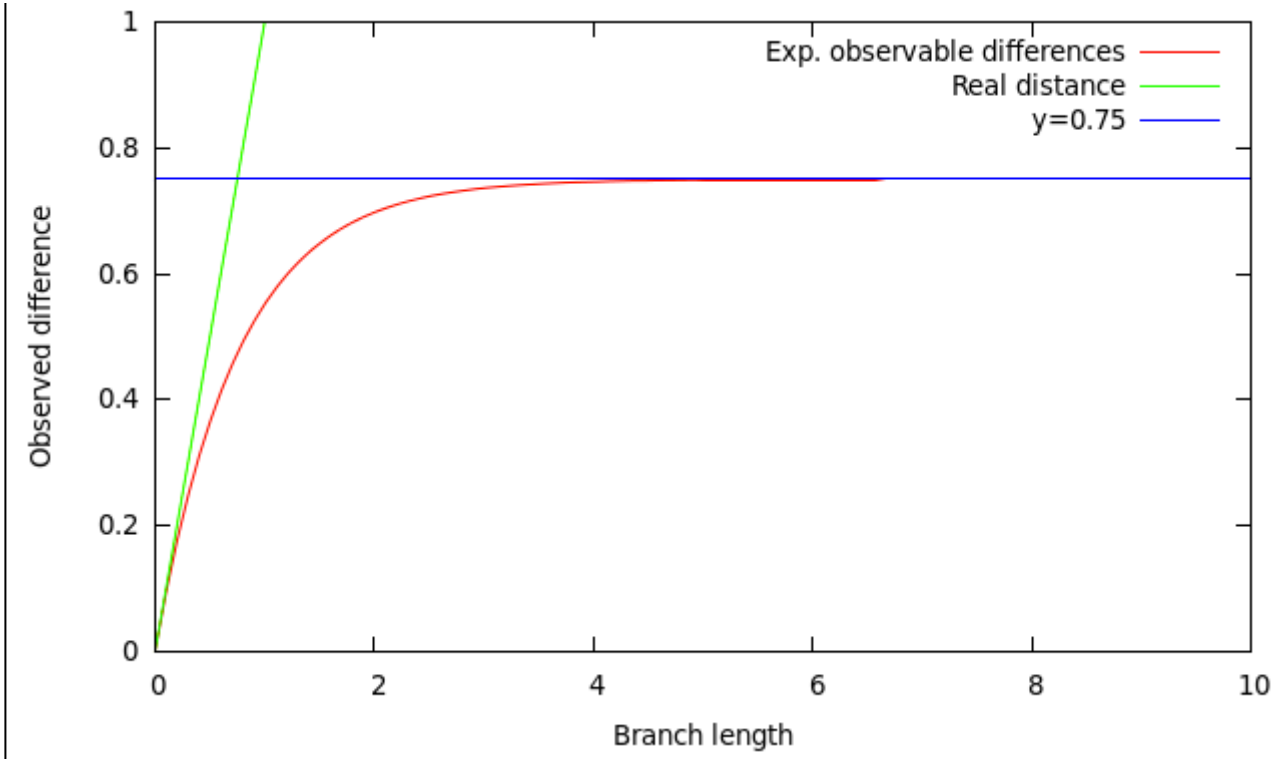
```
plot [0:10] [0:1] 0.75*(1-exp(-1.33*x)) tit "Exp. observable differences" , x tit "Real distance", 0.75 tit "y=0.75"
```

The plot command above has to be entered on a single line! In this expression, x represents *[Math Processing Error]*, i.e., the branch length (the actual amount of change that has occurred). The curve we have plotted thus gives the expected observed difference as a function of the actual amount of change.

- **Question:** Which of the following statements are true?

Your Answer		Score	Explanation
☐ The Jukes and Cantor correction will have a limited effect when the branch length is large.	✔	0.20	
☑ The graph of the observed differences plateaus at 3/4.	✔	0.20	
☑ For small branch lengths, the expected observable difference rises almost linearly and is very close to the real distance.	✔	0.20	
☑ Sequences can become no more than 75 % different according to the Jukes and Cantor model.	✔	0.20	
☐ The graph of the observed differences plateaus at 1.	✔	0.20	
Total		1.00 / 1.00	

Question Explanation



You will notice that the observed change increases from zero to an equilibrium value of 0.75. According to the Jukes and Cantor model, sequences can therefore become no more than 75% different, which is the same as being 25% identical. Twenty-five percent identity is of course the similarity we would get from constructing two random sequences containing 25% of each base and simply placing them side by side. You will also note that for low values of x, the observable difference rises almost linearly. Try to inspect this in greater detail:

```
plot [0:1] [0:1] 0.75*(1-exp(-1.33*x)) tit "Exp. observable differences" , x tit "Real distance", 0.75 tit
"y=0.75"
```

Note how the expected observed difference is very close to the real distance up to about x=0.1 changes per site. This means there is only a limited effect of JC corrections when a pair of sequences differ at less than 10% of their sites.

Question 2

- **Jukes and Cantor model: Examine estimated branch length as a function of observed difference:**

Above we examined how the (expected) observed distance depended on the real distance. We will now examine how the real distance can be estimated from the observed distance. This is done by solving the above equation for *[Math Processing Error]*, giving us an expression that allows us to estimate the real amount of change as a function of the observed change:

[Math Processing Error]

Note that this correction will only work if the observed difference is approximately as expected. Consider this: In the dice-rolling simulation we found that if there has been 0.67 changes per site then the expected observed difference is 0.44. However, as you saw in the simulation, the actual observed difference can be different from the expected 0.44 (say, 0.33 or 0.58). If the observed difference is not the same as the expected observed difference, then we will obviously also get the wrong number even after correction.

We will now plot (estimated) real change as a function of observed difference (this is the inverse of what you did before):

```
set xlab "Observed difference"
```

```
set ylab "Real distance"
```

```
plot [0:0.8]  -0.75*log(1-1.33*x) tit "Estimated real distance", x tit "Observed distance"
```

(The function "log" means the natural logarithm in gnuplot). Note how the correction becomes increasingly more important as the observed distance increases. Also note that this correction does not allow the observed distance to rise above 0.75, although that situation may arise in real data. Above 75% difference the corrected distance is not defined. When using JC corrected distances for phylogenetic reconstruction, you should therefore beware of this situation.

- **Question:** Use the equation above to estimate the actual distance if the observed distance is 0.1, 0.4, and 0.6 respectively. Enter replies on a single line, separated by spaces. The answers have to be entered in the correct order.

You entered:

0.107 0.572 1.207

Your Answer		Score	Explanation
0.107	✔	0.33	
0.572	✔	0.33	
1.207	✔	0.33	
Total		1.00 / 1.00	

Question 3

The Kimura 2 parameter model.

The Kimura 2 parameter model of evolution has the following rate matrix:

[Math Processing Error]

Note how transitions (A/G and C/T) have a different rate than transversions (A/C, A/T, C/G, and G/T). Based on this matrix, the expected ratio of transitions to transversions is:

[Math Processing Error]

Meaning that if transitions and transversions had the same rate (Jukes and Cantor), then we would expect:

[Math Processing Error]

Empirically, this is typically not the case. In fact one often sees *[Math Processing Error]* and, for mitochondrial DNA, a typical value is *[Math Processing Error]* (meaning that *[Math Processing Error]* is 20 times higher than *[Math Processing Error]*)! We will now use the gnuplot program to explore some features of evolution occurring according to this model.

It can be shown that under the K2P model, the chance of observing a transition and a transversion respectively depends on *[Math Processing Error]* and *[Math Processing Error]* in the following way:

[Math Processing Error]

[Math Processing Error]

where *[Math Processing Error]* and *[Math Processing Error]*.

1. **Examine expected amount of change as a function of branch length:**

```
set xlab "Real distance"
```

```
set ylab "Observed difference"
```

We will now examine how the expected amount of transitions and transversions change with time when R=10. Fortunately, gnuplot can be used to compute and keep track of the values of R, A, and B. In the gnuplot window enter the following:

```
R = 10
```

```
A = (-2*R-1.0)/(R+1.0)
```

```
B = (-2)/(R+1.0)
```

You can check the computed values of A and B by using the `print` command:

```
print A
```

```
print B
```

You should have obtained values of approximately A = -1.909 and B = -0.1818. You can now plot the curves showing how the expected amount of transitions and transversions change as a function of the branch length (the actual amount of change):

```
plot [0:40] [0:1]  0.25-0.5*exp(A*x)+0.25*exp(B*x) tit "Transitions", 0.5 - 0.5 * exp(B*x) tit "Transversions", 0.25 t
it "y=0.25", 0.5 tit "y=0.50", 0.75 tit "y=0.75", 0.25-0.5*exp(A*x)+0.25*exp(B*x) + 0.5 - 0.5 * exp(B*x) tit "Total, o
bserved difference"
```

Several interesting things are going on in this plot. First of all, note that I have added a third curve showing the total observed difference. This is simply the sum of the observed transitions and transversions.

Second, as was the case for the Jukes and Cantor model, the total observed difference increases to a maximum value of 0.75 (corresponding to 25% similarity).

Third, note that the expected amount of transitional differences first rise rapidly and then decline slowly to an equilibrium value of 0.25. Transversional differences rise slowly to an equilibrium value of 0.5. The equilibrium values are determined by the fact that when sufficient time has passed sequence similarities will essentially be random; since there are twice as many possible transversions as transitions, these will in the end make up two thirds of all observed changes. Early on, before this stage is reached, the much higher rate of transitions will cause them to make up the vast majority of all observed changes, and only after considerable time has elapsed will the transversions catch up.

Question: From the plot, estimate the real distance (x-axis) at which the transition and transversion lines cross. (Hint: If you move the cursor over the plot, the x and y coordinates of the cursor are shown in the pane at the bottom of the gnuplot window). Enter reply as nearest integer.

You entered:

6

Your Answer		Score	Explanation
6	✔	1.00	
Total		1.00 / 1.00	

Question 4

- **Kimura 2 parameter model: Experiment with other transition/transversion rate ratios:**

The exact behaviour of the relationship between the two types of change depends on the relative rates of transition and transversion. You should now repeat the above analysis with (1) R=2, and (2) R=0.5. Remember to recompute A and B after entering the new value of R (You will probably find it useful that previously entered gnuplot commands can be recalled by pressing the up-arrow). Recall that R=0.5 means that transitions and transversion occur with the same rate, *[Math Processing Error]*. For each of these two cases rerun the plot command and consider the changes.

Question: Based on the two plots which of the following statements are true?

Your Answer		Score	Explanation
☑ For R = 2, the transition and transversion lines cross at around 1.5 substitutions per site.	✔	0.25	
☑ For R = 0.5, the transition and transversion lines never cross each other.	✔	0.25	
☐ For both R=2 and R = 0.5, the transition and transversion lines never cross each other.	✔	0.25	
☐ For R = 2, the transition and transversion lines cross at around 5 substitutions per site.	✔	0.25	
Total		1.00 / 1.00	

Question 5

Question: When R=0.5, the Kimura 2 parameter model is in fact equivalent to another model - which one?

You entered:

Jukes and Cantor

Your Answer		Score	Explanation
Jukes and Cantor	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

When *[Math Processing Error]* we have:

[Math Processing Error]

meaning that transitions and transversions have the same rate. What model has that assumption

Question 6

- **Kimura 2 parameter model: Examine how apparent transition/transversion ratio changes with branch length:**

The apparent transition transversion ratio is simply the observed number of transitions divided by the observed number of transversions. The following plot command shows this number as a function of branch length for the case *[Math Processing Error]* (I have simply taken the expression for observed transitions and divided it by the expression for observed transversions):

```
set ylab "Observed transition/transversion ratio"
```

```
plot [0:4]  (0.25-0.5*exp(-1.909*x)+0.25*exp(-0.1818*x))/ (0.5 - 0.5 * exp(-0.1818*x)) notit
```

Note how the *apparent* ratio is close to the real ratio, *[Math Processing Error]*, when not much change has occurred (i.e., for small x).

The model of evolution that we have explored here is not a particularly complicated one - in fact it only has two free parameters. Nevertheless, you will by now appreciate that it is capable of displaying fairly un-intuitive behaviour. Stating our hypothesis about this biological system in explicit mathematical terms is what allowed us to explore this thoroughly.

Question: What value does the apparent transition/transversion ratio approach asymptotically? (You will need to plot with a wider x-range to see this).

You entered:

0.5

Your Answer		Score	Explanation
0.5	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

Recall that you can change the x-range in gnuplot using the following notation (here the x-range is 0 to 15):

```
plot [0:15] x**2
```

Question 7

Analysis of mitochondrial data set

In this part of the exercise we will explore a real mitochondrial data set containing sequences from man, chimpanzee, gorilla, orangutan, and gibbon. We will investigate how the use of different models of evolution affects the estimated distance matrix. Since mitochondrial DNA is known to have very different transition and transversion rates, we will pay special attention to this aspect.

Note: Make sure to quit gnuplot (by entering "q" at the gnuplot prompt) before proceeding with this part of the exercise.

1. **Prepare editor window:**

```
nedit titv.data &
```

(Make sure to make the nedit window as wide as possible - otherwise the header line will be wrapped over two lines). This file contains a header line and a column listing the estimated divergence time between man and each of the other four primates (in millions of years). These estimates are associated with a fair amount of uncertainty, but the implied branching order is almost certainly correct. You will be using this file for entering various measures that you compute from the data file.

2. **Start PAUP*, load data file:**

```
paup primatemitDNA.nexus
```

3. **Define outgroup:**

```
outgroup gibbon
```

4. **Activate outgroup rooting and select how tree will be printed:**

```
set root=outgroup outroot=monophyl
```

5. **Select distance-based tree-reconstruction:**

```
set criterion=distance
```

6. **Select uncorrected distances under the least squares criterion:**

```
dset distance=p objective=lsfit
```

7. **Short digression on PAUP* online help system:**

   We interrupt this exercise for a brief announcement: By now you should be familiar with many of the commands used in PAUP, but you probably do not have an overview of the long list of possible options that can be specified. Fortunately, PAUP has a command that is useful in this context:

```
dset ?
```

   Here I have used dset as an example, but typing *any* command followed by a question mark ("?") will give you a list of all the possible options for that command, along with a list of the current values. This is very useful if you want to experiment with different settings in an analysis. When you want to learn more about the individual settings, you can download the command reference in PDF from this site:

   PAUP* command reference document

   One final thing that may be good to know: PAUP* accepts abbreviated commands as long as the abbreviation is unambiguous. That means you can for instance write `set crit=dist` instead of the full `set criterion=distance`, and `desc` instead of `describetrees`.

8. **Construct least squares tree:**

```
alltrees
```

   This is a small data set so we can use exhaustive searching.

9. **Inspect tree:**

```
describetrees all/plot=phylogram
```

   This tree reflects our current belief about how these organisms are related.

10. **Print distance matrix, note distances from human:**

```
showdist
```

The `showdist` command lists the distance matrix computed according to the currently active distance-setting (as specified in the `dset` command above).

Question: What are the p-distances for the following pairs of sequences (enter numbers on a single line, separated by spaces. Use at least two significant digits): human/chimpanzee, human/gorilla, human/orangutan, human/gibbon.

Note: Also copy the entries giving the p-distance between human and each of the other four primates into the proper place in the `titv.data` file. (The numbers should all be in a single column under the "`p_dist`" header).

You entered:

0.08705 0.09933 0.15748 0.17969

Your Answer		Score	Explanation
0.08705	✔	0.25	
0.09933	✔	0.25	
0.15748	✔	0.25	
0.17969	✔	0.25	
Total		1.00 / 1.00	

Question 8

- **Select uncorrected distances, counting only transitions:**

```
dset subst=ti
```

The option `subst=ti` specifies that only transitional substitutions should be counted. The previously issued "`distance=p`" is still the active setting. You can verify this by typing "`dset ?`" and checking the value listed for `distance`.

- **Print distance matrix, note transitional distances from human:**

```
showdist
```

In this distance matrix only the transitions have been counted for each pair of taxa.

Question: What are the transition-distances for the following pairs of sequences (enter numbers on a single line, separated by spaces. Use at least two significant digits): human/chimpanzee, human/gorilla, human/orangutan, human/gibbon.

Note: Also enter the numbers in the column labeled "Transitions(P)" in the file.

You entered:

0.08147 0.09040 0.11725 0.12946

Your Answer		Score	Explanation
0.08147	✔	0.25	
0.09040	✔	0.25	
0.11725	✔	0.25	
0.12946	✔	0.25	
Total		1.00 / 1.00	

Question 9

- **Select uncorrected distances, counting only transversions:**

```
dset subst=tv
```

- **Print distance matrix, note transversional distances from human:**

```
showdist
```

Question: Again enter the distances from everything to human below (separated by spaces, and using at least two significant digits) and in the column labeled Transversions(Q) in the file.

You entered:

0.00558 0.00893 0.04023 0.05022

Your Answer		Score	Explanation
0.00558	✔	0.25	
0.00893	✔	0.25	
0.04023	✔	0.25	
0.05022	✔	0.25	
Total		1.00 / 1.00	

Question 10

- **Compute JC-corrected distances:**

As we saw above, it is possible to come up with model-based corrections for the effect of multiple substitutions that allow us to estimate the real amount of change from the observed amount of change. For the JC-model, the equation for the corrected distance is:

[Math Processing Error]

Question: For each of the four lines in the `titv.data` file, and based on the numbers in the column labeled p_dist, compute the JC-corrected distance. Enter the results in the column labeled "JC" in the titv.data file. Also enter the values below in the usual order (human:chimp, human:gorilla, human:orangutan, human:gibbon), separated by spaces, and using at least two significant digits.

You entered:

0.093 0.107 0.177 0.205

Your Answer		Score	Explanation
0.093	✔	0.25	
0.107	✔	0.25	
0.177	✔	0.25	
0.205	✔	0.25	
Total		1.00 / 1.00	

Question 11

- **Compute K2P corrected distance:**

As was the case for the JC model, we can also compute estimated real distances under the K2P model. This can be done using the following equation:

[Math Processing Error]

Question: Using the numbers in columns P and Q, you should now use this equation to compute the K2P-corrected distance estimates. Enter the results in the column labeled K2P in the file. Also enter the values below in the usual order (human:chimp, human:gorilla, human:orangutan, human:gibbon), separated by spaces, and using at least two significant digits. Make sure to save the file after all results have been entered.

You entered:

0.095 0.1097 0.182 0.211

Your Answer		Score	Explanation
0.095	✔	0.25	
0.1097	✔	0.25	
0.182	✔	0.25	
0.211	✔	0.25	
Total		1.00 / 1.00	

Question 12

- **Plot distances:**

First, open a second Terminal and cd to the models directory.

```
cd models
```

```
gnuplot
```

```
set xlab "Time since divergence (MY)"
```

```
set ylab "Genetic distance (substitutions/site)"
```

```
plot 'titv.data' u 2:3 tit "p-distance" w linespoi, 'titv.data' u 2:4 tit "Transitions (P)" w linespoi, 'titv.data' u 2:5
 tit "Transversions (Q)" w linespoi, 'titv.data' u 2:6 tit "JC" w linespoi, 'titv.data' u 2:7 tit "K2P" w linespoi
```

We have here plotted the total difference, the observed transitional and transversional difference, as well as the JC- and K2P-corrected distances as a function of estimated divergence times.

Question: Do the two different correction schemes result in the same estimates of the real distance?

Your Answer		Score	Explanation
Yes	✔	1.00	
No			
Total		1.00 / 1.00	