Feedback — Quiz 4: Maximum Likelihood (computer)

Help Center

You submitted this quiz on **Sat 20 Jul 2013 7:12 AM PDT**. You got a score of **2.00** out of **2.00**.

Biological background

This data set consists of an alignment of full length mitochondrial DNA from human (53 sequences), chimpanzee (1 sequence), bonobo (1 sequence), and Neanderthal (1 sequence). The Neanderthal DNA was extracted from archaeological material, specifically 38,000 year old bones found at Vindija in Croatia (all sequence data was taken from this paper: Green et al., Cell, 2008).

The view emerging from most anatomical, archaeological, and DNA-based studies places Neanderthals as a different species from Homo sapiens. This is in agreement with the "Out-of-Africa hypothesis", which states that Neanderthals coexisted without mating with modern humans who originated in Africa somewhere between 100,000 to 200,000 years ago. There is, however, also some anatomical and paleontological research which supports the so-called "multi-regional hypothesis", which propounds that some populations of archaic Homo evolved into modern human populations in many geographical regions. Consequently, Neanderthals could have contributed to the genetic pool of present-day Europeans. We will use the present data set to consider this issue.

Getting started.

- **Open a Terminal window on 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.

- **Construct working directory, copy files:**

```
cd ~student
```

```
mkdir likelihood
```

```
cd likelihood
```

```
cp ~/data/neanderthal.nexus ./neanderthal.nexus
```

Question 1

Neanderthal DNA: Maximum Likelihood Tree

- **Start PAUP and load data set:**

```
paup neanderthal.nexus
```

- **Remove subset of sequences to reduce computational burden:**

```
delete 5-40
```

This command removes 36 human sequences (sequence number 5 to sequence number 40) from the data set. We do this in order to reduce the time needed to finish the analysis. In the remaining data set we now have 17 human sequences, one chimpanzee, one bonobo, and one Neanderthal.

- **Specify substitution model:**

In the analysis performed here, we have reason to believe that the Kimura 2 parameter model is a fair description of how the sequences evolve (i.e., transitions and transversions have separate rates). We furthermore have evidence that different sites evolve at quite different rates, and we want to model this using a gamma distribution. Moreover, we will request that the transition/transversion rate ratio and the gamma shape parameter are estimated from data. (Although we will not discuss the issue further at this point, it is important to realize that there are techniques for stringent selection of the best model, and that one should never just randomly select one. We will return to such techniques later in the course when we discuss model selection. For now, however, you should just accept that K2P + gamma is an adequate model for the present data set). To specify the substitution model, enter the following at the PAUP prompt:

```
set criterion=likelihood
```

```
lset nst=2 tratio=estimate basefreq=equal rates=gamma shape=estimate
```

In order to search for a maximum likelihood tree, we must first give a detailed description of the assumed substitution model. Since this is the first time we do this, I will give a rather thorough description of each part of the command.

First `lset` ("likelihood settings") is the command used in PAUP to specify likelihood models, just as `dset` was used to specify settings for the distance criterion.

Secondly, we specify that we want a model with two different types of substitution rates (`nst=2`) and where the frequency of each base is 25% (`basefreq=equal`). You will recognize this as the K2P model. Note that, by default, PAUP assumes that `nst=2` means that we want to make a distinction between transitions and transversions. It is also possible to specify models with two types of substitutions that are NOT transitions and transversions respectively. One example would be: `lset nst=6 rmatrix=estimate rclass=(a a a b b b)`. I will not explain this example in detail at this point.

Third, we request that the transition/transversion ratio should be estimated from the data (`tratio=estimate`).

Finally, we specify that we want to use a model where substitution rates at different sites follow a gamma distribution (`rates=gamma`), and that we want the shape of this distribution to also be estimated from the data (`shape=estimate`).

- **Specify outgroup and rooting options:**

```
outgroup Pan_troglodytes Pan_paniscus
```

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

The chimpanzee and the bonobo form the outgroup.

- **Start heuristic search of tree space using nearest neighbor interchange (NNI):**

  ```
  hsearch swap=nni rseed=765
  ```

  This step will take a couple of minutes (depending on your computer). Grab a cup of coffee! For large datasets you sometimes have to wait hours or even days for a maximum likelihood analysis to finish. (Note that you can follow the progress of the hill-climbing algorithm by inspecting the number reported in the last column of the output from PAUP, labeled "Best trees". This lists the negative log of the likelihood, *[Math Processing Error]*. (Recall that since likelihoods are numbers between 0 and 1, log-likelihoods will be negative numbers, and therefore negative log-likelihoods will be positive numbers. Yeah, well...). As the likelihood increases, this number will decrease. Note that PAUP uses more decimal places to decide when to stop the search, than are shown in this output, and you may therefore get the same value printed for several steps.

- **Question:** What is the negative log-likelihood, *[Math Processing Error]*, for the best tree found using NNI?

You entered:

35036.94

Your Answer		Score	Explanation
35036.94	✓	1.00	
Total		1.00 / 1.00	

Question Explanation

The negative of the log likelihood is a positive number

Question 2

Is the Neanderthal sequence placed inside or outside the clade of human sequences?

Your Answer		Score	Explanation
○ The Neanderthal sequence is inside the clade of human sequences.			
● The Neanderthal sequence is placed outside the clade of human sequences.	✓	1.00	
Total		1.00 / 1.00	