Computational Molecular Evolution

Tutorial

by Matyas Pajkos

Dosztanyi Lab

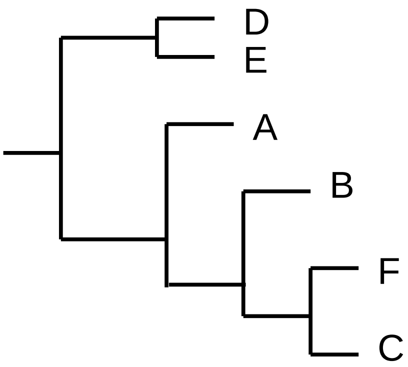
2016 January

The tutorial is in pdf format that is why you have to write your results of this tutorial in a computer file (Office Word, WordPad).

Task 1.

Mini Quiz:

In phylogenetics it is important to use the Newick format correctly. Have a look at the following tree and think about the correct Newick string of it.



Question: Select the correct character string(s). Which of them code the tree?

1. ((((C , F) , B) , A) , (D , E));
2. (((A , B) , (F , C)) , (D , E));
3. ((A , (B , (F , (C)))) , (D , E));
4. ((A , (B , (F , C))) , (E , D));
5. ((B , (A , (F , C))) , (D , E));

Task 2.

Prepare and view trees in FigTree viewer:

FigTree is designed as a graphical viewer of phylogenetic trees. Download the FigTree viewer using the link below.

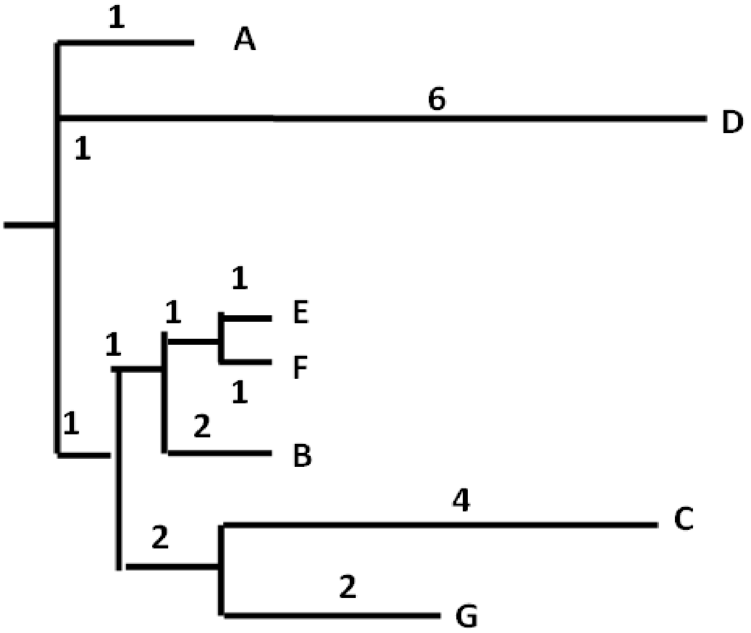
http://tree.bio.ed.ac.uk/download.html?name=figtree&id=90&num=2

If you have Linux, open a new terminal and just type the following commands (These commands install the treeviewer FigTree, and a Java-plugin for Firefox):

```
sudo apt-get install figtree
```

```
sudo apt-get install icedtea-plugin
```

If the viewer is downloaded open a text file (WordPad) and type the rooted tree below in Newick format and thereafter save it with "newick" extension (file.newick).



If it is done, open the saved tree file applying the downloaded viewer:
"File" → "open" → select the tree file
Question: Did you get the same tree topology?

If you managed to build the same topology display the branch lengths:
Click to "branch label"
Question: Did you observe the same branch lengths?

The program has several options for altering the display of the tree, including viewing the tree as unrooted, and altering the rooting interactively. After you've played around with the possibilities for a while you should close the window.

Task 3.

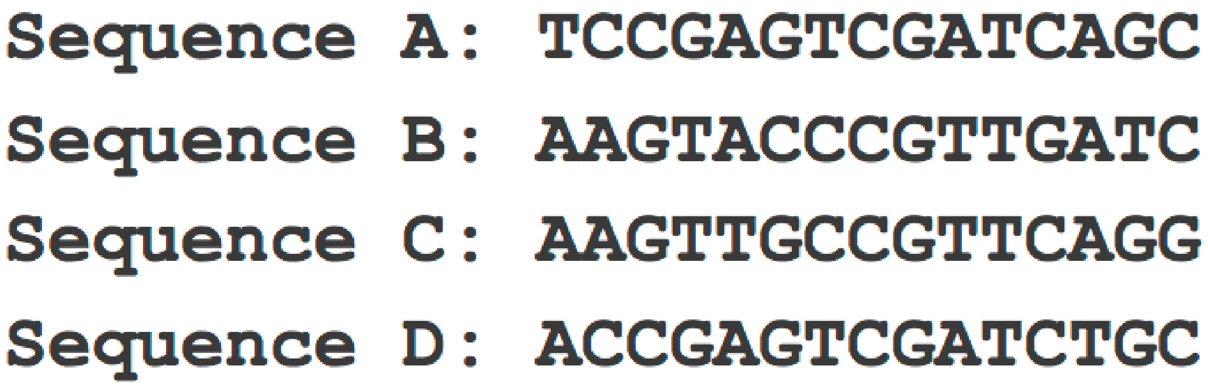
Prepare distance matrix:

Prepare a distance matrix similar to the one shown below. You can construct your own very simply on a piece of paper, or in your computer result file:

	A	**B**	**C**	**D**
A	-			
B		-		
C			-	
D				-

Count pairwise distances:

Below you will find a DNA alignment. Count the distance between each pair of sequences and enter the numbers in your distance matrix. (Remember that the distance between a pair of sequences is defined as the number of nucleotide positions where the two sequences differ).



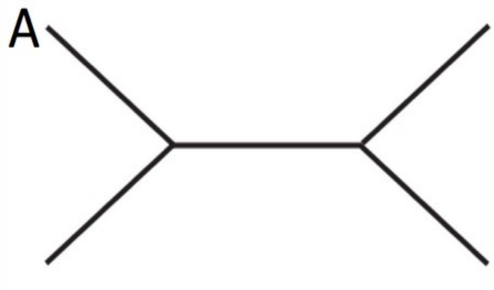
```
Sequence A:  TCCGAGTCGATCAGC
Sequence B:  AAGTACCCGTTGATC
Sequence C:  AAGTTGCCGTTCAGG
Sequence D:  ACCGAGTCGATCTGC
```

Question: What distances did you find? Enter the distances separated by spaces. The distances should be entered in the same order as they appear in the distance matrix (AB, AC, AD, BC, BD, CD).

Task 4.

Reconstruct phylogenetic tree using distance matrix:

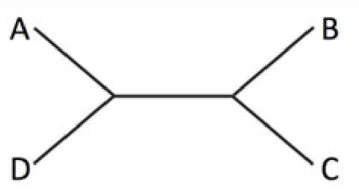
Prepare a tree sketch, similar to the one shown below or just have a look at this sketch:



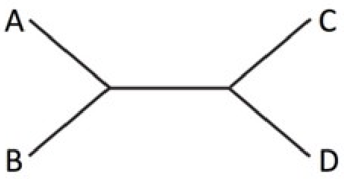
Use the distance matrix from task 3 as the basis for reconstructing an unrooted phylogenetic tree for the sequences A, B, C, and D. Specifically, add sequence labels (B, C, D) to your tree sketch, in the proper positions (I have already indicated the position of A). **Note:** branch lengths are not drawn to scale on the tree sketch!

Question: Which of the following trees is the correct tree?

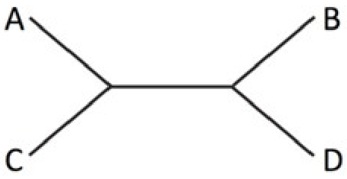
1



2



3



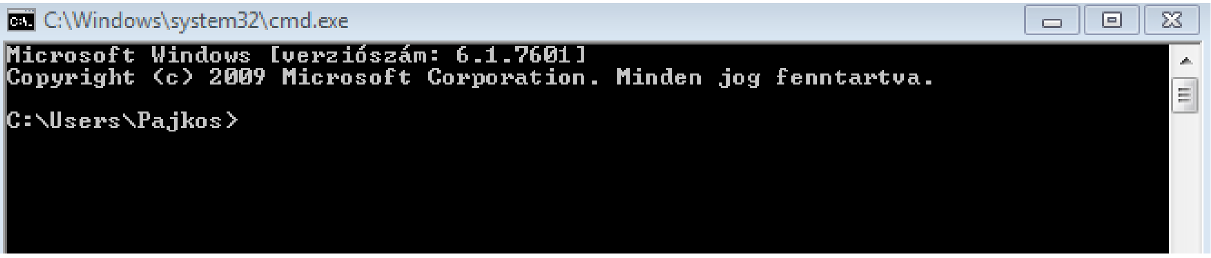
Task 5.

Introduction:

PAML is a package of programs for phylogenetic analyses of DNA or protein sequences using maximum likelihood. **Codeml** is a part of the PAML package, which is a suite of programs for detection of positive selection using DNA or protein data.

To use Codeml and compute positive selection a terminal interface is needed. First of all open a windows console:

Click the Start button, click All Programs, click Accessories, and then click Run, type cmd, click OK

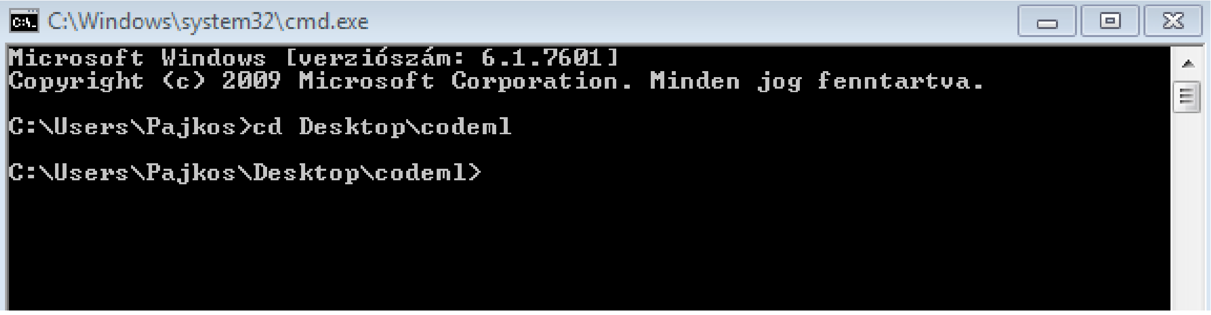


```
Microsoft Windows [verziószám: 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Minden jog fenntartva.

C:\Users\Pajkos>
```

In the Open box, type the commands below to get the location of Codeml. In this case I downloaded the program to my desktop:

cd Desktop\codeml



```
Microsoft Windows [verziószám: 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Minden jog fenntartva.

C:\Users\Pajkos>cd Desktop\codeml

C:\Users\Pajkos\Desktop\codeml>
```

Estimation of different dN/dS in Lysozyme of primates:

Genomic blotting and enzymatic amplification show that the genome of the langur monkey (like that of other primates) contains only a single gene for lysozyme c, in contrast to another group of foregut fermenters, the ruminants, which have a multigene family encoding this protein. Therefore, the langur stomach lysozyme gene has probably evolved recently (within the period of monkey evolution) from a conventional primate lysozyme.

In this exercise you predict molecular selection of the lysozyme gene applying the null- and branch-model of the Codeml algorithm. Firstly, you use the null-model (M0) that is the most basic model of evolution. It computes dN and dS values among branches assuming an identical, gloabl dN/dS ratio. Chang directory to lysozyme\codeml-M0 typing the following command:

```
cd lysozyme\codeml-M0
```

Then, just type the command as follow to compute the global omega. This will start the codeml program using the settings in the ctl-file:

```
codeml lysozyme_M0.ctl
```

If you can notice the time used the computation is done. Open the result file (the file with '.result' extension, the other generated files are not important now) and have a look at the variables computed.

Question: What is the omega value? Based on the omega value decide whether the lysozyme gene is under positive or negative selection.

Adaptive (positive selection) evolution of lysozyme has involved remodeling of amino acid sequences. Lysozyme c has been recruited as a digestive enzyme in the stomachs of creatures needing to retrieve nutrients from microorganisms in fermented food. In this chapter you test the evolution of the langur monkey lysozyme gene using the branch-model.

Chang directory to lysozyme\codeml-branch typing the command below:

```
cd ..\codeml-Branch
```

Open the tree file using the FigTree and have a look at the tree topology. Find the clade that includes Can_colobus and Pne_langur. Open the tree file in a text editor and find the parenthesis pair that defines this clade. If you managed to get it, add two necessary characters after the right parenthesis labeling the branch that has to be tested.

```
)#1
```

If it is done save the file and close it, finally launch the Codeml:

```
codeml lysozyme_MB.ctl
```

Open the result file and answer the question below.

Question: What is the omega value? Based on the omega value decide whether the lysozyme gene is under positive or negative selection during evolution of the analyzed branch.

Detection of positively selected sites in gp120:

In this exercise you are going to investigate features of HIV-1 evolution. You will do this by analyzing a large set of env-genes from HIV-1, subtype B. Specifically, the DNA sequences analyzed here correspond to a region surrounding the hypervariable V3 region of the gp120 protein. Like other retroviruses, particles of HIV are made up of 2 copies of a single-stranded RNA genome packaged inside a protein core, or capsid. The core particle also contains viral proteins that are essential for the early steps of the virus life cycle, such as reverse transcription and integration. A lipid envelope, derived from the infected cell, surrounds the core particle. Embedded in this envelope are the surface glycoproteins of HIV: gp120 and gp41. The gp120 protein is crucial for binding of the virus particle to target cells, while gp41 is important for the subsequent fusion event. It is the specific affinity of gp120 for the CD4 protein that targets HIV to those cells of the immune system that express CD4 on their surface (*e.g.*, T-helper lymphocytes, monocytes, and macrophages).

The role gp120 plays in infection and the fact that it is situated on the surface of the HIV particle, means it is an obvious target for the immune response. That means that there may be a considerable selective pressure on gp120 for creating immune-escape mutants, where amino acids in the gp120 epitopes have been substituted. In this exercise you will construct a maximum likelihood tree that we will subsequently use to investigate whether you can detect such a selective pressure on parts of gp120, again using maximum likelihood methods.

In outline, you will now use the following steps to investigate whether there is any evidence for positively selected positions in your data set:

1. Fit model-1 (null-hypothesis), which assumes there is not any positive selection.
2. Fit model-2 (alternative-hypothesis), which assumes some positive selection.
3. Assess the strength of evidence for the two models using LRT-based model probabilities
4. If model-2 is better: identify the positively selected codons

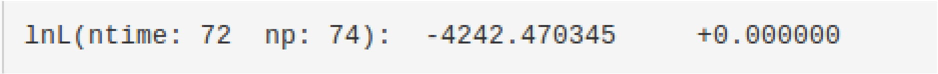
Chang directory:

cd ..\..\gp120\codeml-Site

Then launch Codeml for the gp120 data. Depending on your computer, this will take some minutes to finish. This launch includes both the alternative-hypothesis and null-hypothesis.

codeml codeml_SM.ctl

Wait for the run to finish. Depending on your computer, this take some minutes to finish. Then look at the result file. This file contains a wealth of information concerning your analysis. The top part of the file gives an overview of your sequences, codon usage and nucleotide frequencies. You can ignore this information for now, and move on to the interesting part, namely the model-1 likelihoods and parameter values. If you locate a line that looks a bit like the one shown below,



```
lnL(ntime: 72  np: 74):  -4242.470345     +0.000000
```

identify the number of "free parameters", *K*, used in model model-1: This is indicated by "np", and is 74 in the example shown above (most of these parameters are branch lengths in the tree; specifically, the number of branch length parameters is indicated by "ntime", and is 72 in this example). Also note the log-likelihood of the fitted model. This is the number right after the parenthesis, and is -4242.470345 in the example here.

Question: What are the values of K and lnL for model-1?

Find likelihood and K for model-2. Scroll past the model-1 output until you get to the results for model-2.

Question: What are the values of K and lnL for model-2?

Assess strength of evidence for model-1 and model-2. Model-2 will always have a better (higher) log-likelihood than model model-1 because model-2 has more free parameters. You should now use the recipe given below to compute LRT.

Chi-square value: $\mathbf{2\times(lnL_1 - lnL_0)}$

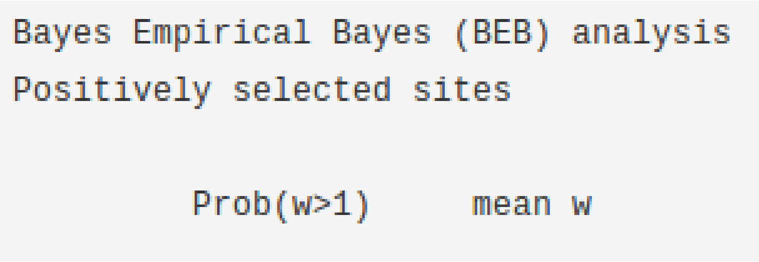
Degree of freedom: $np_1 - np_0$

Use these parameters and do some statistics applying this online interface:

http://www.danielsoper.com/statcalc3/calc.aspx?id=11

Question: What is p-value? Is this value smaller than 0.5? According to this result, is model-2 better than model-1?

If your model-2 is clearly better than model-1 (I firmly believe it should be if you did things according to my instructions...), then you have evidence for the existence of positively selected sites in the gp120 gene. Now scroll down to the end of the result file and locate a list similar to this one (note: this is the "Bayes Empirical Bayes" table, not the "Naive Empirical Bayes" table:



```
Bayes Empirical Bayes (BEB) analysis
Positively selected sites

         Prob(w>1)     mean w
```

This gives you a list of which residues (if any) that were found to belong to the positively selected *dN*/*dS*-class. Also listed is the probability that the site really is in the codon class where *dN*/*dS*>1, and a weighted average of the *w* at the site. Using only DNA sequences you have now identified likely epitopes on the gp120 protein.

Question: List all sites having more than 95% probability of belonging to the positively selected class.