# **Computer exercises – Tree of life 2017**

In this week we will

1. Install the MEGA7 program

2. Solve the first MEGA tutorial 2

3. Create your own influenza data set in Fasta format

4. Align sequences using MEGA’s implementation of Clustal

5. Solve MEGA tutorial 3 (http://www.megasoftware.net/webhelp/walk\_through\_mega/estimating\_evolutionary\_distances\_from\_nucleotide\_sequences.htm)

6. Estimate genetic distances for nucleotide sequences

## **Install MEGA**

MEGA 7.0 is the most popular free phylogenetic program available. Go to [www.megasoftware.net](http://www.megasoftware.net) and request a download. You can choose versions for Macintosh, Windows and Linux

Now solve MEGA tutorial 2

## **GENBANK**

GenBank is a collection of useful databases that can be accessed through from Align/Show Web Browser in MEGA. In order to use GenBank efficiently, it is a good idea to use specific search fields. These are specified as e.g. [Author], [pdat] (publication time) and [Title] and use logical operators to combine search terms, e.g. AND, OR, NOT. For example, the search “Tataru [Author] AND Markov [Title]”in the Pubmed database should return two papers.

Solve the following questions:

## 1. **Find and download the paper by Craig Venter from 2001 describing sequencing of the first human genome. How many pages is the paper and where was it published?** **How can the sequence be obtained -**

Accession number (of the data, republished in 2015 and potentially updated): CH471055 AADB02000000

The Sequence of the Human Genome

J. Craig Venter et al.

Science 16 Feb 2001:

Vol. 291, Issue 5507, pp. 1304-1351

DOI: 10.1126/science.1058040

## 2. **Find the paper on the high coverage archaic Denisovan sequence published by Svante Paabos group and answer the same questions. Who were the Denisovans?**

## It has 8 pages and it was published in Nature. The sequence can be obtained on GeneBank (accession codes: ERP000318, ERP000121 and FR696050).

Archaic hominine group from Denisova cave in Siberia. The paper concluded that that Denisovans were sister group to Neandertals.

## 3. **How many DNA sequences and protein sequences have been determined in the dolphin (you might want to use the latin name)**

## Nucleotide: 1904

## Protein: 632

## **4. Find the Taxonomic position of Dolphins and find the number of DNA sequences deposited in GenBank for each of the dolphin species**

## Delphinus capensis, Nucleotide 491, protein 277

## Delphinus delphis nucleotide 1310, protein 342

## Delphinus delphis ponticus nucleotide 5

## Delphinus tropicalis nucleotide 33, protein 13

## **5. Find sequences from FOXP2, which is a protein involved in human speech. How many different species has the nucleotide sequence been determined for? Try downloading the sequences (in FASTA format). Save them (or add sequences to the MEGA alignment editor). You should download equally sized sequences of the same kind (i.e. either genomic sequences, mRNA sequences or protein sequences, not a mixture). You are free to download either protein or nucleotide sequences. Create a file called foxp2.fas**

In the protein database: Animals 1217, Bacteria 1.

In the nucleotide database: Animals 1756, Bacteria 1.

## 6. **Create your own influenza sequence data set from Genbanks flu database** [**http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi**](http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi)**. Choose protein coding region, Type A, canine and cat, any region, HA protein, any H subtype, any N subtype and full length only. While this analysis can be carried out in MEGA own browser it may be more easy to do it from your own browser and then later import the results to MEGA**

## a. **How many sequences do you get ?** Choose 10 of these from different species and from different subtypes and download them both as protein coding and protein in Fasta format as flucoding.fas and fluprotein.fas. We will use these for the alignment exercise and for phylogenetic exercises in the coming weeks

96 protein coding region sequences were found.

**BLAST**

Blast stands for Basic Local Alignment Search Tool and is a way to search for homologous sequences in a database using a query sequence (for example if you have a gene and will look for orthologous genes in different species or members of the same gene family in the same species). It is probably the most widely used computational tool in (molecular) biology.

BLAST is also integrated in MEGA which now has a built in web browser interface. The advantage of browsing from MEGA being that you can also import directly the sequences you found in MEGA (using the big “ADD to alignment red button”)

## **1. Use Genbank to retrieve the mitochondrial genome from the Neanderthal and then do a BLAST search to find the most closely related sequences. What are the most closely related sequences?**

Homo sapiens microRNA 99%, Homo sapiens MT-RNR2-like 2 95%

## 

## **Alignment**

You have downloaded flu data sets and you should now align these within MEGA’s alignment explorer. If you doubleclick a file it may open directly in MEGA’s alignment explorer

First open your flucoding.fas.

0. What the differences between MUSCLE and CLUSTAL alignments?

MUSCLE does multiple alignments while CLUSTAL does pairwise alignments. Thereby, MUSCLE is supposed to give a better alignment.

**1. Align this at the DNA level using Clustal. Look carefully at all the parameters and try to change them and see how that affects the alignment. Tip: Before you realign you should first remove all gaps. Try to judge the quality of your alignment by looking at it both at the DNA level and at the protein level.**

a. Try to see how you can manually edit your alignment afterwards. Export your best alignment in Megaformat as flucoding.meg for later use

**2. Now remove gaps and choose to translate to amino acids before aligning. Aligning is now at the protein level so Clustal will show different options. Try different combinations and evaluate the alignment as before. Does it seem to be an advantage to align at the protein level and then translate back to nucleotide sequences? Export your best alignment in mega format**

3. Repeat some of the previous questions using the Muscle alignment tool.

4. Perform alignment also of your FoxP2 data set and export in Mega format

5. Now open your fluprotein.fas dataset, align it and export it as fluprotein.meg

**Estimating distances**

**First go carefully through MEGA tutorial 3**

**Link to the tutorial 3:** [**http://www.megasoftware.net/webhelp/walk\_through\_mega/estimating\_evolutionary\_distances\_from\_nucleotide\_sequences.htm**](http://www.megasoftware.net/webhelp/walk_through_mega/estimating_evolutionary_distances_from_nucleotide_sequences.htm)

**then**

First open flucoding.meg in MEGA. We now want to estimate pairwise distances between all 10 sequences. Choose this option and you will see that there are a lot of choices you can make. Make sure you understand all of these, e.g

**1. what is complete versus pairwise deletion?**

In *MEGA*, there are two ways to treat gaps. One is to delete all of these sites from the data analysis. This option, called the *Complete-Deletion*, is generally desirable because different regions of DNA or amino acid sequences evolve under different evolutionary forces. The second method is relevant if the number of nucleotides involved in a gap is small and if the gaps are distributed more or less randomly. In that case it may be possible to compute a distance for each pair of sequences, ignoring only those gaps that are involved in the comparison; this option is called *Pairwise-Deletion*. The following table illustrates the effect of these options on distance estimation with the following three sequences:

http://www.megasoftware.net/mega4/WebHelp/part\_iv\_\_\_evolutionary\_analysis/constructing\_phylogenetic\_trees/handling\_missing\_data\_and\_alignment\_gaps/rh\_alignment\_gaps\_and\_sites\_with\_missing\_information.htm

**2. what is the variance estimation method?**

It is used to specify whether to compute Distances only or Distances and Standard Errors using the selected estimation method.

http://www.megasoftware.net/webhelp/centraldialogbox\_hc/hc\_dist\_analysis\_option\_dialog.htm

**3. what does it mean that transitions and transversions can be used?**

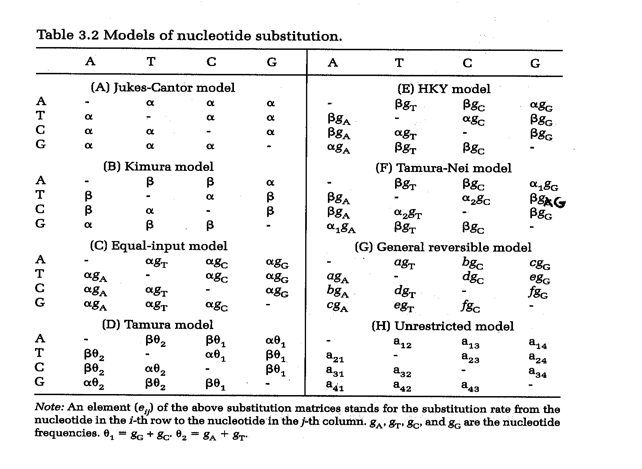
Depending on the distance model or method selected, the evolutionary distance can be teased into two or more components. Thanks to this parameter we are able to choose whether to test our data using just changes on transitions, just changes on transversions or both of them.

You can now try to make distance matrices using p distances, Jukes-Cantor, Kimura 2 parameter and Tamura Nei models, more detailed information on these are found in the Book Chapters 2 and 3 of Nei and Kumar. Make sure you know what each of these models are doing, e.g.

**1. what are the parameters in these models?**

All the models can be found here:

[**https://www.megasoftware.net/mega4/WebHelp/part\_iv\_\_\_evolutionary\_analysis/computing\_evolutionary\_distances/distance\_models/hc\_models\_for\_estimating\_distances.htm**](https://www.megasoftware.net/mega4/WebHelp/part_iv___evolutionary_analysis/computing_evolutionary_distances/distance_models/hc_models_for_estimating_distances.htm)

****

**Jukes-Cantor model:** assumes the same substitution rate (alpha) across sites and assumes that nucleotides have equal frequencies (25% each).

**Kimura model:** takes into account the difference between transitions and transversions. Transitions (exchange of molecules of the same group) are expected to be at higher rates.

**Equal-input model:** takes into account the differences in nucleotide frequencies but assumes the same substitution rate across nucleotides.

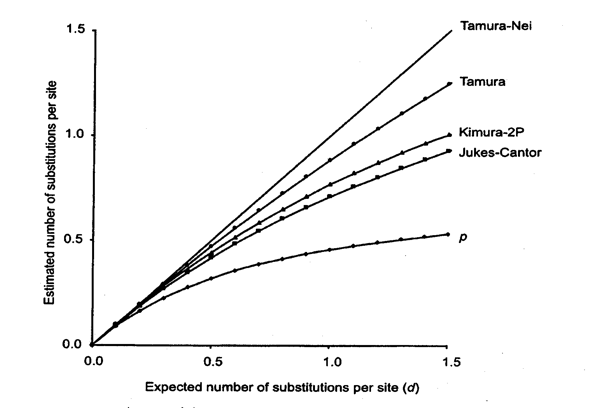
**Tamura Model:** Tamuras 3-parameter model corrects for multiple hits, taking into account differences in transitional and transversional rates and G+C-content bias (1992). It assumes an equality of substitution rates among sites.

**HKY**: Takes into account transition transversion bias, GC content bias (by explicitly stating nucleotide frequencies).

**Tamura-Nei**: The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional substitution rates between purines and transversional substitution rates between pyrimidines. It also assumes equality of substitution rates among sites (see [related gamma model](https://www.megasoftware.net/mega4/WebHelp/part_iv___evolutionary_analysis/computing_evolutionary_distances/distance_models/nucleotide_substitution_models/gamma_distances/hc_tamura_nei_distance_with_gamma_model.htm)).

**2. which model estimates largest distances and why?**

The model that better represents the data will give a higher distance, so models more complicated (with more parameters) tend to calculate higher distance than the simplest model (p-distance).

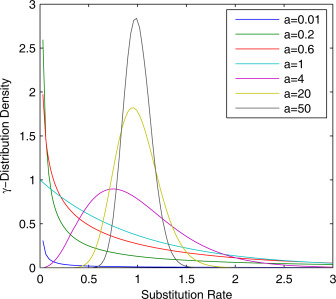


Now try on the same data to estimate the transition/transversion bias as well as the gamma parameter using maximum likelihood under the Models menu.

**1. What does the result tell you regarding which substitution model to use?**

**a. How do you interpret the value of R, the ratio of transitions to transversions?**

**2. Does the gamma rate parameter suggest large variation in underlying substitution rate?**

****

when **gama**<1, the distribution indicates that most of the sites have very low substitution rates despite the existence of a few sites with higher substitution rates; (2) when **gama**>1, the distribution shows that the majority of the sites have intermediate rates around 1, except for the fact that some sites may exhibit extreme rates (very low or high); (3) when **gama** goes to infinity, the distribution becomes a simpler type that all sites have the same rate.

Now open fluprotein.meg instead and estimate distances at the protein level.

**1. What differences do you see between p distance and Poisson corrected distance?**

**a. What are the distances at the protein level compared to the distances at the nucleotide level?**