# Assignment 4 / Biometry and bioinformatics II / 2014

_____

The goal of this set of exercises is to familiarize with some general molecular evolutionary concepts, nucleotide substitution modelling, composition, codon usage, synonymous, non-synonymous substitutions

You can use either MEGA5-software, which you have been using for assignment set 1, (http://www.megasoftware.net/) or DnaSP5 (http://www.ub.edu/dnasp/, installed in C128)

## Assignment set 4.1

Modelling nucleotide substitutions is an elementary part of (for example) phylogeny reconstructions (with the exception of parsimony methods). Jukes-Cantor model assumes that all substitutions occur with equal probabilities. Derivation is given in lecture slides and, for convenience, here anew.

Derive the two-parameter model including separate parameters for transitions and transversions. In case you consider this as too demanding, it is enough that you construct the strating scheme and the first equations, i.e. the part corresponding equations (1) – (3) in Jukes-Cantor one-parameter model.

Assumption: all nucleotide substitutions occur with equal probabilites, *a*, Jukes-Cantor model (1969)

- The rate of substitution for each nucleotide is $3a$ per unit time

|   | A | T | C | G |
|---|---|---|---|---|
| A |   | *a* | *a* | *a* |
| T | *a* |   | *a* | *a* |
| C | *a* | *a* |   | *a* |
| G | *a* | *a* | *a* |   |

- At time 0: Assumption that at a certain nucleotide site there is A, $P_{A(0)} = 1$

- Question: probability that this site is occupied by A at time $t$, $P_{A(t)}$ ?

- At time 1, probability of still having A at this site is

$$P_{A(1)} = 1 - 3a \tag{1}$$

- $3a$ is the probability of A changing to T, C, or G

- The probability of the site having A at time 2 is

$$P_{A(2)} = (1 - 3a)P_{A(1)} + a\,[1 - P_{A(1)}] \tag{2}$$

This includes two possible courses of events from time points $t=0$ --> $t = 1$ --> $t = 2$

| $t = 0$ | | $t = 1$ | | $t = 2$ |
|---------|---------------|-----------|---------------|-----|
| A | no substitution | A | no substitution | A |
| A | substitution | T or C or G | substitution | A |

- The following recurrence equation holds for any $t$

$$P_{A(t+1)} = (1 - 3a)P_{A(t)} + a[1 - P_{A(t)}] \tag{3}$$

Note that this holds also for $t = 0$, because $P_{A(0)} = 1$ and thus

$$P_{A(0+1)} = (1 - 3a)\,P_{A(0)} + a\,[1 - P_{A(0)}] = 1 - 3a$$

which is identical with equation (1).

- The amount of change in $P_{A(t)}$ per unit time, rewriting equation (3):

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3aP_{A(t)} + a[1 - P_{A(t)}] = -4aP_{A(t)} + a \tag{4}$$

- Approximating the previous discrete-time model by a continuous-time model, by regarding $\Delta P_{A(t)}$ as the rate of change at time $t$. With this approximation equation (4) is rewritten as

$$dP_{A(t)} / dt = -4aP_{A(t)} + a \tag{5}$$

- The solution of this first-order linear differential equation is

$$P_{A(t)} = \tfrac{1}{4} + (P_{A(0)} - \tfrac{1}{4})e^{-4at} \tag{6}$$

- The starting condition was A at the given site, $P_{A(0)} = 1$, consequently

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{3}{4}e^{-4at} \tag{7}$$

- Equation (6) holds regardless of the initial conditions, for example if the initial nucleotide is not A, then $P_{A(0)} = 0$, and the probability of having A at time t

$$P_{A(t)} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4at} \tag{8}$$

- Equations (7) and (8) describe the substitution process. If the initial nucleotide is A, then $P_{A(t)}$ decreases exponentially from 1 to ¼ . If the initial nucleotide is not A, then $P_{A(t)}$ will increase monotonically from 0 to ¼ .

- Under this simple model, after reaching equilibrium, $P_{A(t)}=P_{T(t)}=P_{C(t)}=P_{G(t)}$ for all subsequent times.

- Equation (7) can be rewritten in a more explicit form to take into account that the initial nucleotide is A and the nucleotide at time t is also A

$$P_{AA(t)} = \tfrac{1}{4} + \tfrac{3}{4}e^{-4at} \tag{9}$$

- If the initial nucleotide is G instead of A, from equation (8)

$$P_{GA(t)} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4at} \tag{10}$$

Since all the nucleotides are equivalent under the Jukes-Cantor model, the general probability, $P_{ij(t)}$, that a nucleotide will become $j$ at time $t$, given that it was $i$ at time 0, equations (9) and (10) give the general probabilities $P_{ii(t)}$ and $P_{ij(t)}$, where $i \neq j$.

$$P_{ii(t)} = \tfrac{1}{4} + \tfrac{3}{4}\, e^{-4\alpha t} \qquad \text{and} \qquad P_{ij(t)} = \tfrac{1}{4} + \tfrac{1}{4}\, e^{-4\alpha t} \qquad (11)$$

## Number of substitutions, *nucleotide divergence*, between two sequences

● We assume that all sites in sequence evolve at the same rate and follow the same substitution scheme. The number of sites compared between two sequences is denoted by $L$.

● Consider the probability that a nucleotide at a given site at time $t$ is the same in both sequences. Suppose that the nucleotide at a given site was A at time point 0. At time $t$, the probability that a descendant sequence will have A at this site is $P_{AA(t)}$, and consequently the probability that two descendant sequences have A at this site is $P^2_{AA(t)}$. Similarly, the probabilities that both sequences have T, C or G at this site are $P^2_{AT(t)}$, $P^2_{AC(t)}$, and $P^2_{AG(t)}$

● The probability that the nucleotide at a given site at time $t$ is the same in both sequences is

$$I_{(t)} = P^2_{AA(t)} + P^2_{AT(t)} + P^2_{AC(t)} + P^2_{AG(t)} \qquad (12)$$

● From equations (11) we obtain

$$I_{(t)} = \tfrac{1}{4} + \tfrac{3}{4}\, e^{-8\alpha t} \qquad (13)$$

● Equation (13) also holds for T, C or G. Therefore, regardless of the initial nucleotide at a given site, $I_{(t)}$ represents the proportion of *identical* nucleotides between two sequences that diverged $t$ time units ago. The probability that the two sequences are *different* at a site at time $t$ is $p = 1 - I_{(t)}$. Thus

$$p = \tfrac{3}{4}\,(1 - e^{-8\alpha t}) \qquad \text{or} \qquad 8\alpha t = \ln(1 - (4/3)\,p) \qquad (14)$$

● The time of divergence between two sequences is usually not known, and thus estimation of $\alpha$ is not possible. Instead, it possible to calculate **K, which is the number of substitutions per site since the time of divergence between the two sequences**. In the case of the one-parameter model, $K = 2(3\,\alpha t)$, where $3\,\alpha t$ is the number of substitutions per site in a single lineage.

$$\mathbf{K = 6\,\alpha}t = -\tfrac{3}{4} \ln(1 - (4/3)\,p) \qquad (15)$$

where $p$ is the observed proportion of different nucleotides between the two sequences.

## Assignment set 4.2

### Background

**DNA**      **SECOND BASE**      **RNA**

|  | T | C | A | G |  |
|---|---|---|---|---|---|
| **T** | TTT Phe<br>TTC Phe<br>TTA Leu<br>TTG Leu | TCT Ser<br>TCC Ser<br>TCA Ser<br>TCG Ser | TAT Tyr<br>TAC Tyr<br>TAA Stop<br>TAG Stop | TGT Cys<br>TGC Cys<br>TGA Stop<br>TGG Trp | T<br>C<br>A<br>G |
| **C** | CTT Leu<br>CTC Leu<br>CTA Leu<br>CTG Leu | CCT Pro<br>CCC Pro<br>CCA Pro<br>CCG Pro | CAT His<br>CAC His<br>CAA Gln<br>CAG Gln | CGT Arg<br>CGC Arg<br>CGA Arg<br>CGG Arg | T<br>C<br>A<br>G |
| **A** | ATT Ile<br>ATC Ile<br>ATA Ile<br>ATG Met | ACT Thr<br>ACC Thr<br>ACA Thr<br>ACG Thr | AAT Asn<br>AAC Asn<br>AAA Lys<br>AAG Lys | AGT Ser<br>AGC Ser<br>AGA Arg<br>AGG Arg | T<br>C<br>A<br>G |
| **G** | GTT Val<br>GTC Val<br>GTA Val<br>GTG Val | GCT Ala<br>GCC Ala<br>GCA Ala<br>GCG Ala | GAT Asp<br>GAC Asp<br>GAA Glu<br>GAG Glu | GGT Gly<br>GGC Gly<br>GGA Gly<br>GGG Gly | T<br>C<br>A<br>G |

FIRST BASE (left) / THIRD BASE (right)

|  | U | C | A | G |  |
|---|---|---|---|---|---|
| **U** | UUU Phe<br>UUC Phe<br>UUA Leu<br>UUG Leu | UCU Ser<br>UCC Ser<br>UCA Ser<br>UCG Ser | UAU Tyr<br>UAC Tyr<br>UAA Stop<br>UAG Stop | UGU Cys<br>UGC Cys<br>UGA Stop<br>UGG Trp | U<br>C<br>A<br>G |
| **C** | CUU Leu<br>CUC Leu<br>CUA Leu<br>CUG Leu | CCU Pro<br>CCC Pro<br>CCA Pro<br>CCG Pro | CAU His<br>CAC His<br>CAA Gln<br>CAG Gln | CGU Arg<br>CGC Arg<br>CGA Arg<br>CGG Arg | U<br>C<br>A<br>G |
| **A** | AUU Ile<br>AUC Ile<br>AUA Ile<br>AUG Met | ACU Thr<br>ACC Thr<br>ACA Thr<br>ACG Thr | AAU Asn<br>AAC Asn<br>AAA Lys<br>AAG Lys | AGU Ser<br>AGC Ser<br>AGA Arg<br>AGG Arg | U<br>C<br>A<br>G |
| **G** | GUU Val<br>GUC Val<br>GUA Val<br>GUG Val | GCU Ala<br>GCC Ala<br>GCA Ala<br>GCG Ala | GAU Asp<br>GAC Asp<br>GAA Glu<br>GAG Glu | GGU Gly<br>GGC Gly<br>GGA Gly<br>GGG Gly | U<br>C<br>A<br>G |

● Each amino acid is coded by a "triplet of nucleotides", a codon, having three "sites", first, second and third site or position.

● Nucleotide sites (positions) are classified into nondegenerate, twofold degenerate, and fourfold degenerate sites:

● A site is nondegenerate if all possible changes at this site are non-synonymous: nucleotide change => amino acid change, twofold degenerate if one of the three possible changes is synonymous (nucleotide change => no amino acid change), and fourfold degenerate if all possible changes at the site are synonymous

● For example, the first two positions at the codon TTT (Phe) are nondegenerate, while the third position is twofold degenerate. The third position at the codon GTT (Val) is fourfold degenerate.

● By using the data 2 (see the data description in assignment set 1) calculate

    ● Codon usage of the bacteria in the data.

    ● The so called  GC-content of the bacteria in the data.

● The concepts and, in many instances, practical "tools", *codon usage* and *GC-content* will be introduced in more detail during the lecture 8.10. Now we get familiar with these by working out the facts: codons are not used at random, GC-content is not evenly distributed. During the session 3.10 we had a look at these (by using another data) by using DnaSP5.

You get a table of GC-content of all sequences (i.e. one table).  Write about differences you can observe.

But codon usage: each sequence item has it´s own table. You are <u>not</u> supposed to start inspecting tens of such tables!

> Just pick up some tables (not many !) and explain some part of the results you notice. You are not supposed to write an extensive report commenting all codon – amino acid issues. Pick up, for example, some amino acids and inspect them, and write about your observations.

Hint: When you have worked with this data for assignment 1, you have got a clustering structure of the bacteria, i.e. different species and different serotypes of one species (*Streptococcus pneumoniae*).  *Use your clustering structure as framework for selecting some items (sequences, which are different species or different within-species serotypes). This means: pick up sequence items from clearly different clusters.*

---------------------------------------------------------------------------------------------

## Assignment set 4.3

By using data_4 answer the question posed above in 4.2: nucleotide composition and codon usage bias. And, also for some other gene from data_3 (the mitochondrial genome, see assignment set 1): nucleotide composition and codon usage bias.

Inspect codon  usage biases only from some animals, not from the whole data. Take, for example, dog, horse and cow. And the same animals from both genes.

Data_4 –file is the cytochrome-gene from the mitochondrial genome (cut from the data in data_3, which is now given in a bit modified version: data 3_aligned_IUPACedited (see http://www.dnabaser.com/articles/IUPAC%20ambiguity%20codes.html;
The reason: DnaSP does not recognize other than A,C,T,G´s and N – all Y´s and W´s and R´s, i.e. the ”not clear bases” have been replaced by N.)

## Your task is to cut also another gene.

How to cut: One way is to use Clustal, define the coordinates to be included, and save that file as a new fasta-file. In DnaSP: there is the window which allows definition (by using coordinates) of a region to be analysed. The same is true for MEGA5, in which you can also delete the regions which you don´t want to be included. The data_3 coordinate table is, for convenience, given here anew.

| | Nucleotides in AB499817, the first sequence in datafile | Nucleotides taking into account gaps in aligned file |
|---|---|---|
| tRNA-Phe | 1-69 | 1-80 |
| 12S ribosomal RNA | 70-1023 | 81-1090 |
| tRNA-Val | 1024-1090 | 1091-1161 |
| 16S ribosomal RNA | 1091-2670 | 1162-2840 |
| tRNA-Leu | 2671-2745 | 2841-2917 |
| gene ND1 | 2748-3704 | 2919-3882 |
| tRNA-Ile | 3704-3722 | 3882-3901 |
| tRNA-Gln | 3769-3843 | 3948-4025 |
| tRNA-Met | 3845-3914 | 4028-4097 |
| gene ND2 | 3915-4958 | 4098-5143 |
| tRNA-Trp | 4957-5024 | 5142-5215 |
| tRNA-Ala | 5038-5106 | 5232-5301 |
| tRNA-Asn | 5108-5179 | 5310-5386 |
| tRNA-Cys | 5213-5280 | 5419-5495 |
| tRNA-Tyr | 5281-5348 | 5496-5572 |
| gene COI | 5350-6894 | 5574-7140 |
| tRNA-Ser | 6892-6962 | 7132-7216 |
| tRNA-Asp | 6967-7034 | 7222-7292 |
| gene COII | 7035-7718 | 7293-7977 |
| tRNA-Lys | 7736-7802 | 7995-8066 |
| gene ATPase subunit 8 | 7804-8007 | 8068-8276 |
| gene ATPase subunit 6 | 7965-8645 | 8234-8914 |
| gene COIII | 8645-9428 | 8914-9697 |
| tRNA-Gly | 9429-9496 | 9698-9770 |
| gene ND3 | 9497-9843 | 9771-10117 |
| tRNA-Arg | 9843-9911 | 10117-10187 |
| gene ND4L | 9914-10210 | 10191-10487 |
| gene ND4 | 10204-11581 | 10481-11858 |
| tRNA-His | 11580-11650 | 11857-11930 |
| tRNA-Ser | 11651-11710 | 11931-11995 |
| tRNA-Leu | 11711-11780 | 11996-12067 |
| gene ND5 | 11781-13601 | 12068-13895 |
| gene ND6 | 13585-14112 | 13879-14406 |
| tRNA-Gln | 14111-14181 | 14405-14476 |
| gene cytB | 14186-15325 | 14482-15625 |
| tRNA-Thr | 15326-15395 | 15626-15703 |
| tRNA-Pro | 15395-15460 | 15703-15772 |
| D-loop | 15461-16741 | 15773-18424 |

## Assignment set 4.4

The file  data HLA_gene  contains human alleles at one *human leukocyte* gene.

- Synonymous and non-synonymous ?

- Nondegenerate, two-fold degenerate, four-fold degenarte ?
  (note that MEGA5  has buttons-to-be-clicked to get these).