

Lets define x_1 , x_2 and x_3 as the frequencies of elements in states A, B and C respectively

After one step (iteration) $x_1 \cdot p_{12}$ elements will be converted from A to B, $x_1 \cdot p_{13}$ from A to C and so on.

The new frequency of A elements (x_1') will be given by the proportion of those elements that were in A and remain in A after one step ($x_1 \cdot p_{11}$) plus those that were B and changed to A ($x_2 \cdot p_{21}$) plus those that were C and changed to A ($x_3 \cdot p_{31}$)

In other words the new x_1 is:

$$x_1' = x_1 \cdot p_{11} + x_2 \cdot p_{21} + x_3 \cdot p_{31}$$

Similarly

$$x_2' = x_1 \cdot p_{12} + x_2 \cdot p_{22} + x_3 \cdot p_{32}$$

$$x_3' = x_1 \cdot p_{13} + x_2 \cdot p_{23} + x_3 \cdot p_{33}$$

Probabilities in two steps:

What is the probability of going from one state to another in two steps?

One has to consider the different pathways

Suppose you want to calculate the probability of going from A to B in two steps.

There are 3 different pathway:

1: A->B->B

2: A->A->B

3: A->C->B

the respective probabilities for each pathway are

1: $p_{12} * p_{22}$

2: $p_{11} * p_{12}$

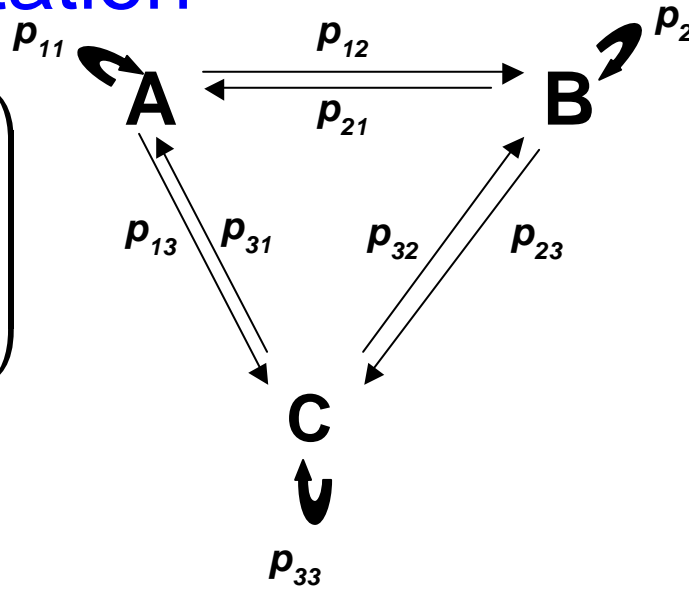
3: $p_{13} * p_{32}$

The global probability of going from A to B in two steps is the sum of the probabilities of the three pathways, namely:

$$p_{12} * p_{22} + p_{11} * p_{12} + p_{13} * p_{32}$$

Matrix notation

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$



$$\mathbf{x} = (x_1 \quad x_2 \quad x_3)$$

Multiplying a vector and a matrix $\mathbf{x} * P$ gives a vector that will be called \mathbf{x}'

By the rules of vector-matrix multiplication this is done in the following way:

$$\mathbf{x} * P = (\underbrace{x_1 * p_{11} + x_2 * p_{21} + x_3 * p_{31}}_{x_1'}, \underbrace{x_1 * p_{12} + x_2 * p_{22} + x_3 * p_{32}}_{x_2'}, \underbrace{x_1 * p_{13} + x_2 * p_{23} + x_3 * p_{33}}_{x_3'})$$

Therefore $\mathbf{x} * P$ gives the new frequencies for the different states (A,B,C) after one step.

So we can write that:

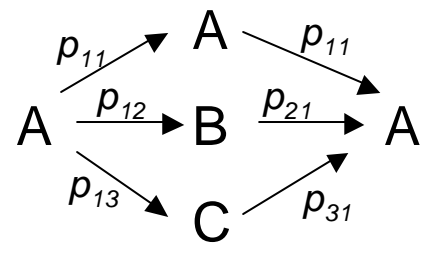
$$\mathbf{x}_{t+1} = \mathbf{x}_t * P$$

Matrix Matrix Multiplication

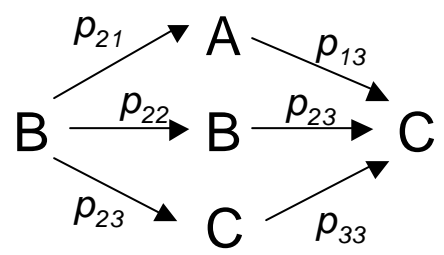
$$P * P = P^2 = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} * \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$

$$P^2 = \begin{pmatrix} p_{11} & \cdot & \cdot \\ \cdot & \cdot & p_{23} \\ \cdot & \cdot & \cdot \end{pmatrix}$$

$$p_{11}^{(2)} = p_{11} * p_{11} + p_{12} * p_{21} + p_{13} * p_{31}$$



$$p_{22}^{(2)} = p_{21} * p_{13} + p_{22} * p_{23} + p_{23} * p_{33}$$



Therefore P^2 gives the probabilities of going from one state to the other in two steps

Likewise $x * P^2$ gives the frequencies for the different states after two steps

Summarizing we can say that:

$$\mathbf{x}_t^* \mathbf{P} = \mathbf{x}_{t+1}$$

$$\mathbf{x}_t^* \mathbf{P}^2 = \mathbf{x}_{t+2}$$

$$\mathbf{x}_t^* \mathbf{P}^n = \mathbf{x}_{t+n}$$

In general we are interested in the steady state (equilibrium). This steady state is accomplished after a very large number of steps.

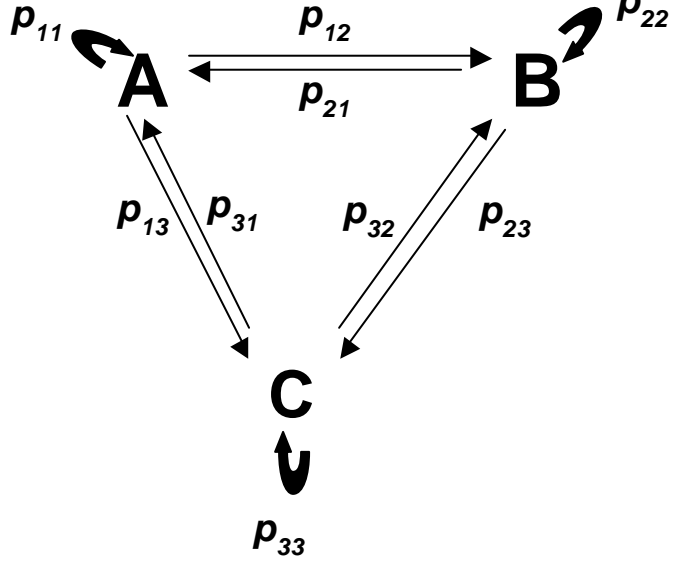
$$\mathbf{x} \mathbf{P}^n = \hat{\mathbf{x}}$$

$n \rightarrow \infty$

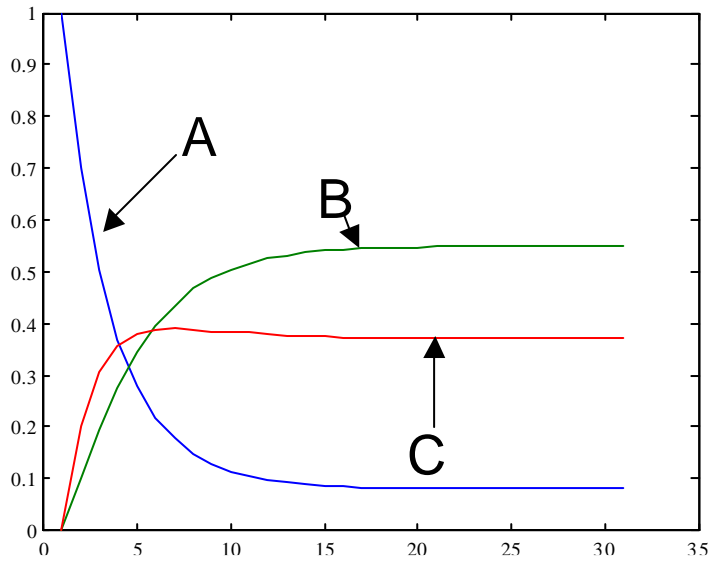
When $\mathbf{x} = \hat{\mathbf{x}}$, we can also express this situation in the following way $\mathbf{x}_t^* \mathbf{P} = \mathbf{x}_{t+1}$ being $\mathbf{x}_{t+1} = \mathbf{x}_t$

Examples

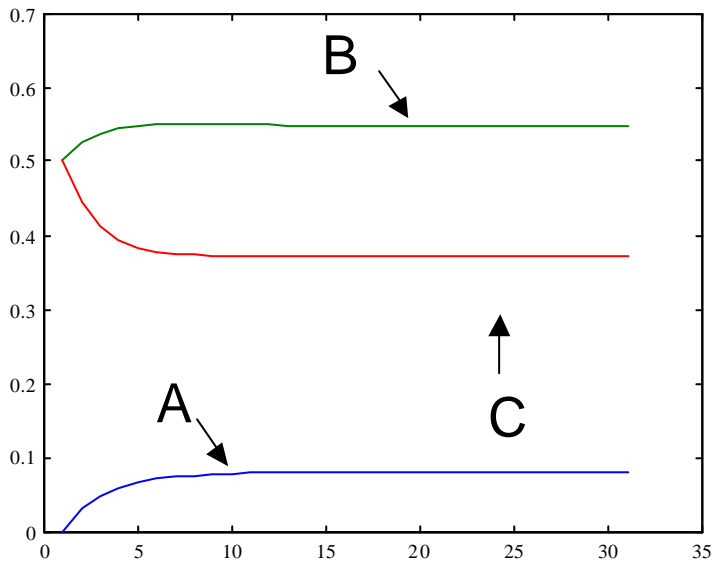
$$P = \begin{pmatrix} 0.7 & 0.1 & 0.2 \\ 0.01 & 0.85 & 0.14 \\ 0.05 & 0.20 & 0.75 \end{pmatrix}$$



$$X = (1 \quad 0 \quad 0)$$

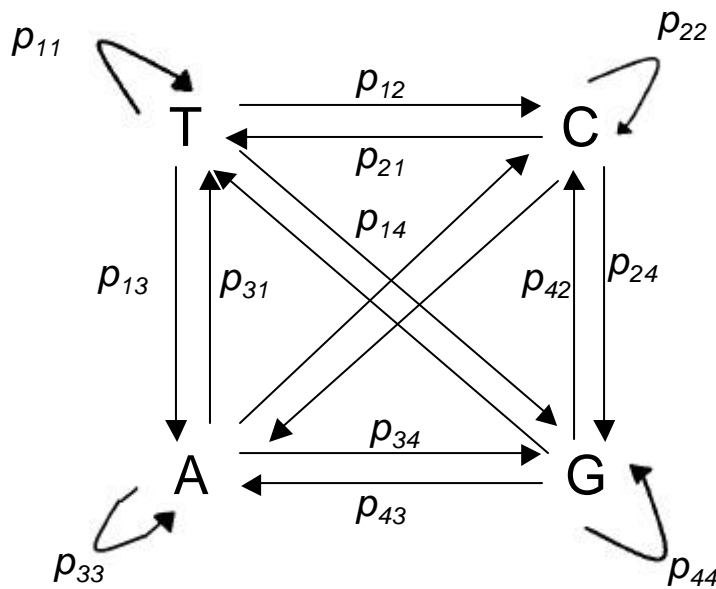


$$X = (0 \quad 0.5 \quad 0.5)$$



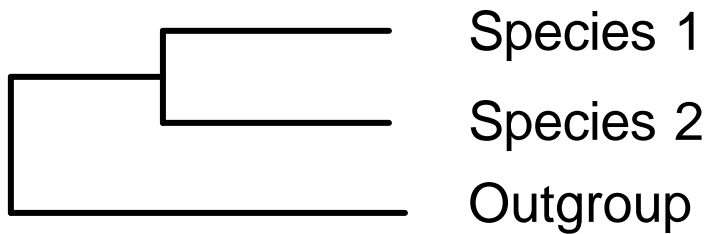
Estimating the equilibrium state from real DNA data

The goal is to obtain the matrix having probabilities of changing from each nucleotide to the others



/to From	T	C	A	G
T	p_{11}	p_{12}	p_{13}	p_{14}
C	p_{21}	p_{32}	p_{23}	p_{24}
A	p_{31}	p_{32}	p_{33}	p_{34}
G	p_{41}	p_{42}	p_{43}	p_{44}

For this purpose we first build a matrix of observed substitutions



TACATCCGGATGCAC TACTCGAT Sp1
 TACGTTCCGATACAGTTCCCGCT Sp2
 TACGTTCCGATGCAGTACGGGAT Outgroup
 * * * + * + * + * * * + * * + * + * + * + *

The number of substitutions among the different nucleotides can be estimated by maximum parsimony or by maximum likelihood

| /to
From | T | C | A | G |
|-------------|---|---|---|---|
| T | 5 | 1 | 0 | 0 |
| C | 0 | 4 | 0 | 1 |
| A | 1 | 1 | 3 | 0 |
| G | 0 | 1 | 2 | 2 |

Expected number of substitutions

| /to
From | T | C | A | G |
|-------------|----------|----------|----------|----------|
| T | p_{11} | p_{12} | p_{13} | p_{14} |
| C | p_{21} | p_{32} | p_{23} | p_{24} |
| A | p_{31} | p_{32} | p_{33} | p_{34} |
| G | p_{41} | p_{42} | p_{43} | p_{44} |

$$T=2300 \quad C=3000 \quad A=2500 \quad G=3100$$

We are going to have $p_{11} * 2300$ that remain in T,
 $p_{12} * 2300$ that change from T to C, $p_{13} * 2300$ that change
from T to A and $p_{14} * 2300$ that change from T to G

| /to
From | T | C | A | G |
|-------------|-----------------|-----------------|-----------------|-----------------|
| T | $p_{11} * 2300$ | $p_{12} * 2300$ | $p_{13} * 2300$ | $p_{14} * 2300$ |
| C | $p_{21} * 3000$ | $p_{32} * 3000$ | $p_{23} * 3000$ | $p_{24} * 3000$ |
| A | $p_{31} * 2500$ | $p_{32} * 2500$ | $p_{33} * 2500$ | $p_{34} * 2500$ |
| G | $p_{41} * 3100$ | $p_{42} * 3100$ | $p_{43} * 3100$ | $p_{44} * 3100$ |

| From/to | T | C | A | G | Total |
|---------|---|---|---|---|-------|
| T | 5 | 1 | 0 | 0 | 6 |
| C | 0 | 4 | 0 | 1 | 5 |
| A | 1 | 1 | 3 | 0 | 5 |
| G | 0 | 1 | 2 | 2 | 5 |

| From/to | T | C | A | G |
|---------|------------|------------|------------|------------|
| T | 5/6 | 1/6 | 0/6 | 0/6 |
| C | 0/5 | 4/5 | 0/5 | 1/5 |
| A | 1/5 | 1/5 | 3/5 | 0/5 |
| G | 0/5 | 1/5 | 1/5 | 3/5 |

Practical exercise number 1

Goals:

- 1-To obtain a nucleotide substitution matrix from an alignment using Maximum Likelihood method
- 2- To normalize this matrix in order to obtain a probabilistic matrix so that the equilibrium state in base composition can be calculated.
- 3- To compare the observed base composition with the composition expected at equilibrium

Tools:

- 1-the program baseml belonging the PAML package
- 2- The command grep and MS-Excel to deal with the output of baseml
- 3- The computer program Matlab to work with matrices and vectors

Working with baseml

This computer program works with control file called baseml.ctl, that is a simple text file.

baseml.ctl

seqfile = primate.nuc

treefile = primate.tre

outfile = out2 * main result file

noisy = 9 * 0,1,2,3: how much rubbish on the screen

verbose = 1 * 1: detailed output, 0: concise output

runmode = 0 * 0: user tree; 1: semi-automatic; 2: automatic

* 3: StepwiseAddition; (4,5):PerturbationNNI

model = 4 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85

* 5:T92, 6:TN93, 7:REV, 8:UNREST, 9:REVu; 10:UNRESTu

Mgene = 0 * 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff

fix_kappa = 0 * 0: estimate kappa; 1: fix kappa at value below

kappa = 5 * initial or fixed kappa

fix_alpha = 1 * 0: estimate alpha; 1: fix alpha at value below

alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)

Malpha = 0 * 1: different alpha's for genes, 0: one alpha

ncatG = 8 * # of categories in the dG, AdG, or nparK models of rates

nparK = 0 * rate-class models. 1:rK, 2:rK&fK, 3:rK&MK(1/K), 4:rK&MK

clock = 0 * 0:no clock, 1:clock; 2:local clock; 3:TipDate

nhomo = 0 * 0 & 1: homogeneous, 2: kappa for branches, 3: N1, 4: N2

getSE = 0 * 0: don't want them, 1: want S.E.s of estimates

RateAncestor = 2 * (0,1,2): rates (alpha>0) or ancestral states

Small_Diff = 7e-6

cleandata = 1 * remove sites with ambiguity data (1:yes, 0:no)?

* ndata = 5

* icode = 0 * (with RateAncestor=1. try "GC" in data,model=4,Mgene=4)

* fix_blength = -1 * 0: ignore, -1: random, 1: initial, 2: fixed

method = 0 * 0: simultaneous; 1: one branch at a time

PhyML uses alignments in Phylip format like the following one

```
3 120
Species1
AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATT
CTGCCTAGCAAACCTCAAACCTACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGG
species2
AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATT
CTGCCTAGCAAACCTCAAATTATGAACGCACCCACAGTCGCATCATAATTCTCTCCAAGG
Species3
AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCACGGACTTACATCATCATTATTATT
CTGCCTAGCAAACCTCAAACCTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG
```

And a tree file in the following format

```
((Species1: 0.0547, Species: 0.05128), Species3: 0.07977);
```

Note: if the option **runmode** in baseml is different from 0, the program will build a phylogenetic tree and will not infer the ancestral states and substitutions

The results of baseml will be stored in two files, the one you species as outputfile and a second text file called **rst**

The file rst contains the counting of each type of substitution along each branch of the phylogenetic tree, as well as the nucleotides that remain unchanged.

Regretfully this information is not in matrix format.

So, you have to use grep and Excel to count the numbers of each type of substitution and the number of nucleotides that remain unchanged to obtain a Matrix.

1-Open the file rst with the notepad editor and find out which combination of characters you should use to grep the information you need.

2-use the command grep and direct the output to a text file

3-open this text files with MS-Excel and organize the information to get a matrix with the counting of each type of substitution.

Working with Matlab

4-Copy and paste this matrix to Matlab.

Matlab is an interactive program for numerical computation and data visualization. There are many different toolboxes available which extend the basic functions of Matlab into different application areas.

5- In Matlab normalize the matrix (dividing each element in a row by the sum of the row)

6- Iterate the vector-matrix multiplication $[x_{t+1}=x_t*P]$ until equilibrium is reached. Try with different initial conditions (i.e. different base compositions)

7-Compare the equilibrium base composition with the real