Computational Biology

Lecturers:
Tim Vaughan & Carsten Magnus

Teaching Assistants:
Etthel Windels, Antoine Zwaans,
Adrian Lison & Chaoran Chen

Computational Evolution
Department of Biosystems Science and Engineering

HS 2022

How to study evolution?

The easiest way to study something is by observation.

How to study evolution?

The easiest way to study something is by observation.

- Wetlab
  - Very realistic;
  - Time-consuming and expensive;
  - Impossible (sometimes).
- Simulation
  - A virtual experiment in which we mimic a (biological) process on a computer to study its properties
  - Not necessarily realistic
  - Allows us to:
    - generate data with given assumptions;
    - test predictive properties of models.

How to study evolution?

The easiest way to study something is by observation.

- Wetlab
  - Very realistic;
  - Time-consuming and expensive;
  - Impossible (sometimes).
- Simulation
  - A virtual experiment in which we mimic a (biological) process on a computer to study its properties
  - Not necessarily realistic
  - Allows us to:
    - generate data with given assumptions;
    - test predictive properties of models.

Today we will simulate evolution!

The tree of great apes

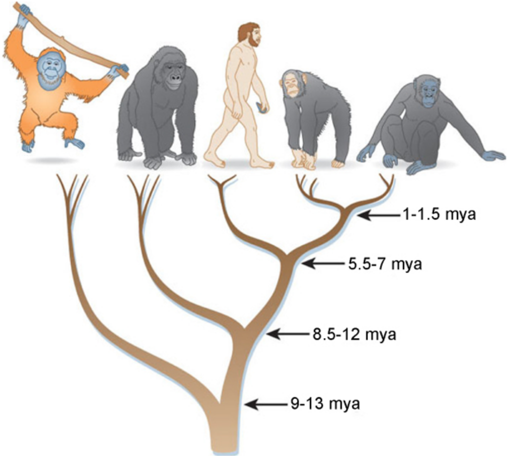
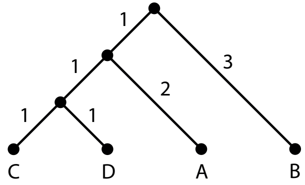


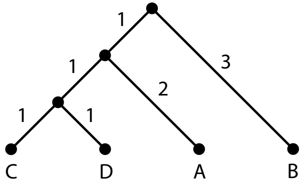
Figure adapted from [Paabo, 2003]

Storing trees: Newick format



- Format for tree representation
- To record a tree in Newick format:
  - Assign a label to each tip
  - Choose two tips that are a cherry (e.g. C and D)
  - Replace selected tips with a new tip of the form (tip1:branch1,tip2:branch2) (e.g. $(\mathrm{C}:1,\mathrm{D}:1)$)
    - Branch length to the new tip is the branch length to the cherry
  - Repeat until the full tree is rewritten
- What is the Newick format for the rooted tree above?

Storing trees: Newick format



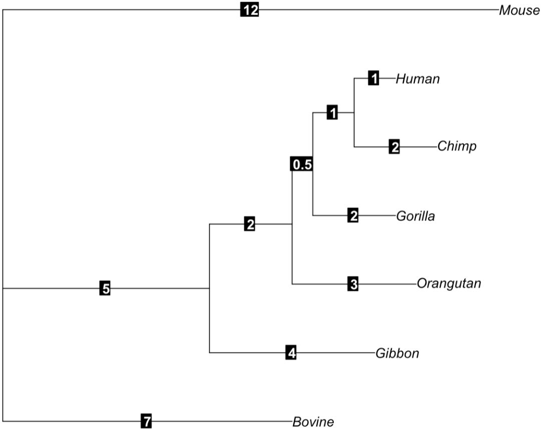
- Format for tree representation
- To record a tree in Newick format:
  - Assign a label to each tip
  - Choose two tips that are a cherry (e.g. C and D)
  - Replace selected tips with a new tip of the form (tip1:branch1,tip2:branch2) (e.g. $(\mathrm{C}:1, \mathrm{D}:1)$)
    - Branch length to the new tip is the branch length to the cherry
  - Repeat until the full tree is rewritten
- What is the Newick format for the rooted tree above? $(((\mathrm{C}:1, \mathrm{D}:1):1, \mathrm{A}:2):1, \mathrm{B}:3)$;

Storing trees: Newick format

- Draw the tree given by the newick string: $(Bovine : 7, (Gibbon : 4, (Orangutan : 3, (Gorilla : 2, (Chimp : 2, Human : 1) : 1) : 0.5) : 2) : 5, Mouse : 12);$

Storing trees: Newick format

- Draw the tree given by the newick string: $(\mathrm{Bovine}: 7, (\mathrm{Gibbon}: 4, (\mathrm{Orangutan}: 3, (\mathrm{Gorilla}: 2, (\mathrm{Chimp}: 2, \mathrm{Human}: 1): 1): 0.5): 2): 5, \mathrm{Mouse}: 12);$



Evolution Simulation Algorithm

Steps:

1. **Initialization of the starting sequence**:
   - Sample a starting nucleotide for each position in the sequence
2. **Iterative simulation** of sequence evolution, along all branches of the tree
   - Compute the transition probability matrix $P(t_b)$.
   - Sample a new nucleotide for each position in the sequence.

Step 1: Initialization of the starting sequence

1a. Sample a starting nucleotide n

Step 1: Initialization of the starting sequence

1a. Sample a starting nucleotide n

From the vector of equilibrium frequencies of nucleotides

	T	C	A	G
Π	0.22	0.26	0.33	0.19

Step 1: Initialization of the starting sequence

1a. Sample a starting nucleotide $\mathfrak{n}$

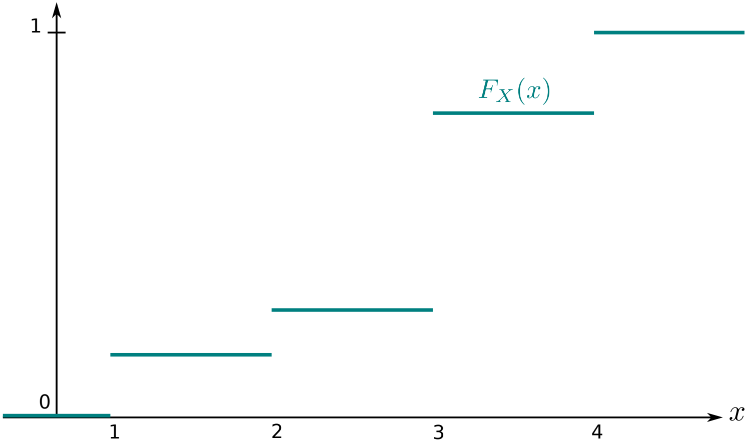
From the vector of equilibrium frequencies of nucleotides

	T	C	A	G
Π	0.22	0.26	0.33	0.19

Knowing Π, how do we sample a nucleotide?

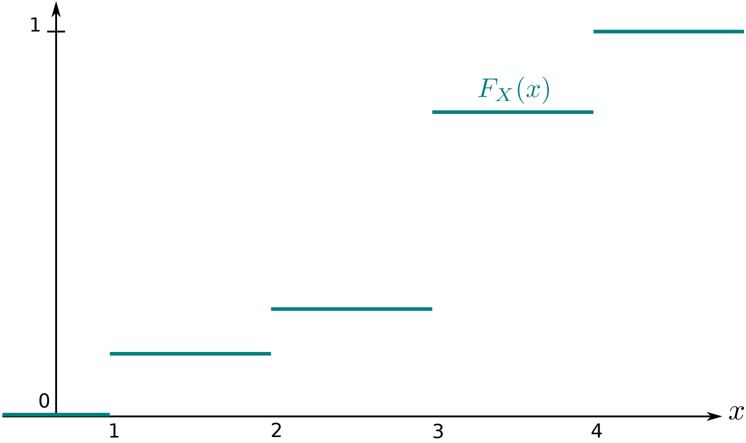
Inverse transform method

- Sample u from $U(0, 1)$;



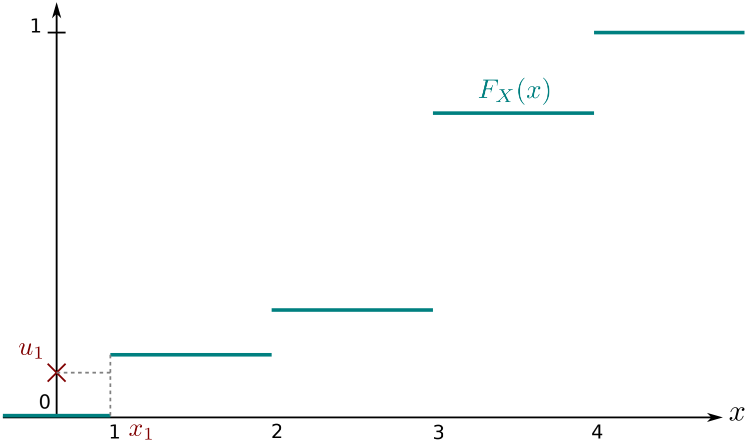
Inverse transform method

- Sample u from $U(0,1)$;
- Transform u into a sample from the desired distribution using the **CDF** $==$ **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



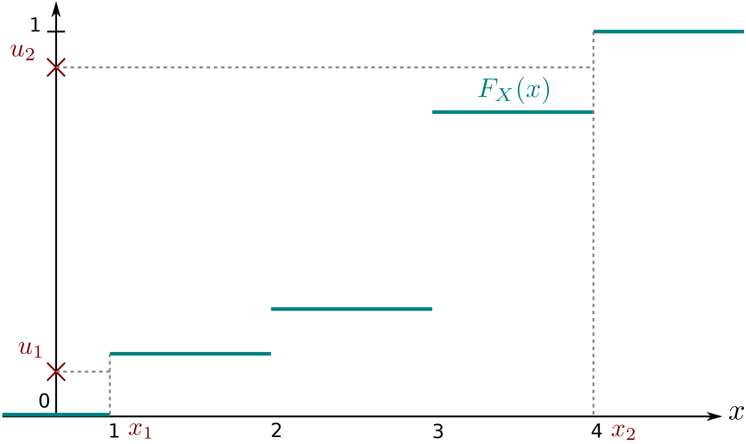
Inverse transform method

- Sample u from $U(0, 1)$;
- Transform u into a sample from the desired distribution using the **CDF** $==$ **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



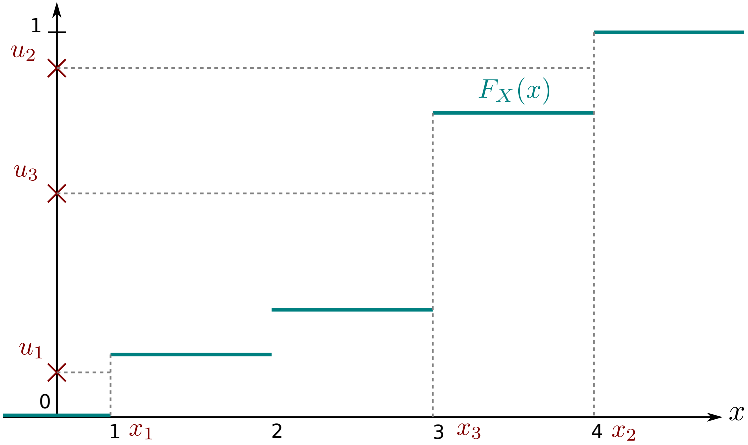
Inverse transform method

- Sample u from $U(0, 1)$;
- Transform u into a sample from the desired distribution using the **CDF** == **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



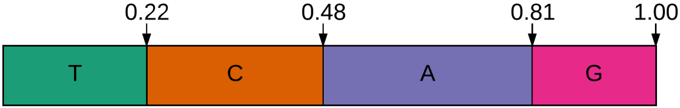
Inverse transform method

- Sample u from $U(0,1)$;
- Transform u into a sample from the desired distribution using the **CDF** == **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



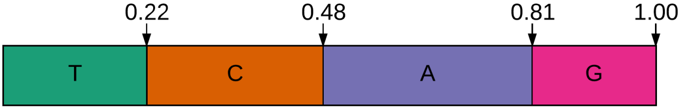
Sampling discrete random variables

	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



Sampling discrete random variables

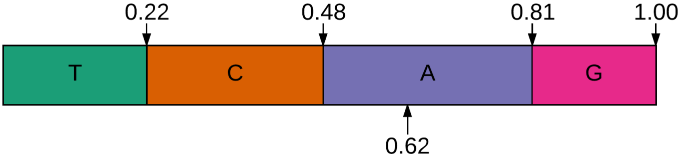
	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



Sample u from $\mathcal{U}(0, 1)$.

Sampling discrete random variables

	T	C	A	G
π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00

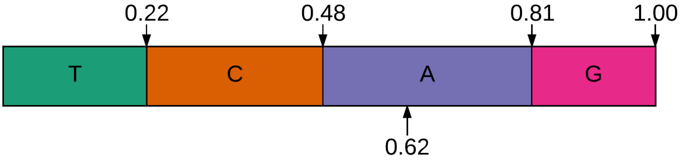


Sample u from $\mathcal{U}(0, 1)$.

E.g. $u = 0.62$.

Sampling discrete random variables

	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



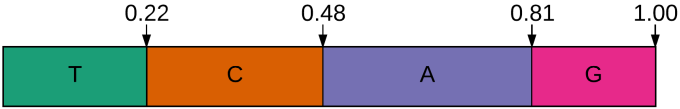
Sample u from $\mathcal{U}(0,1)$.

E.g. $u = 0.62$.

Select nucleotide **A**.

Sampling discrete random variables

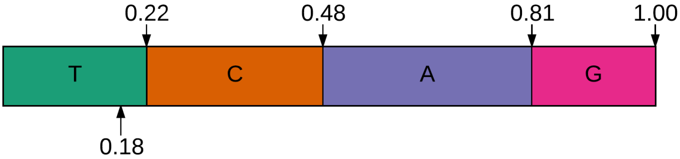
	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



Sample u from $U(0, 1)$.

Sampling discrete random variables

	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00

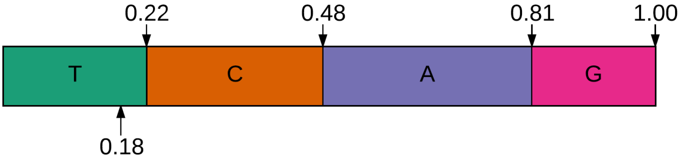


Sample u from $U(0, 1)$.

E.g. $u = 0.18$.

Sampling discrete random variables

	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



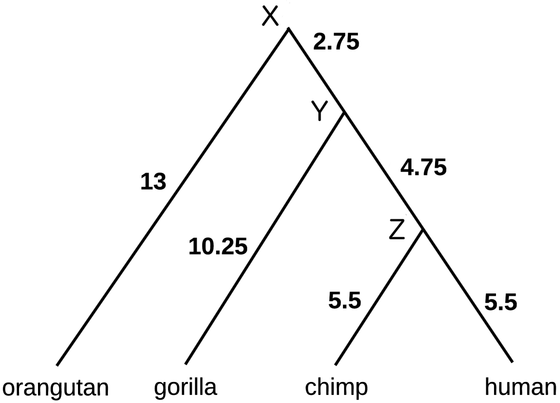
Sample u from $\mathcal{U}(0, 1)$.

E.g. $u = 0.18$.

Select nucleotide **T**.

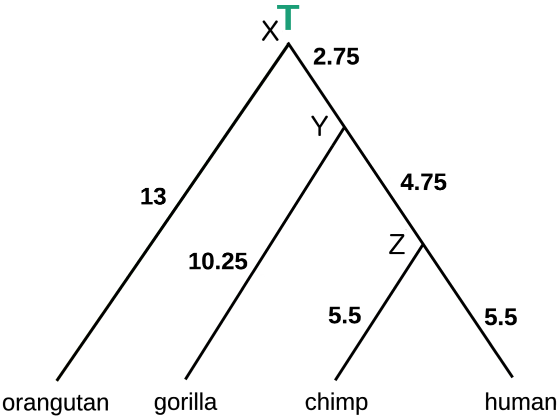
Step 1: Initializing the starting sequence

1b. Place n on the root node;



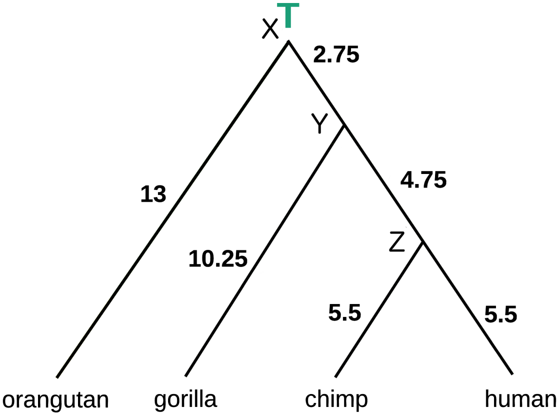
Step 1: Initializing the starting sequence

1b. Place n on the root node;



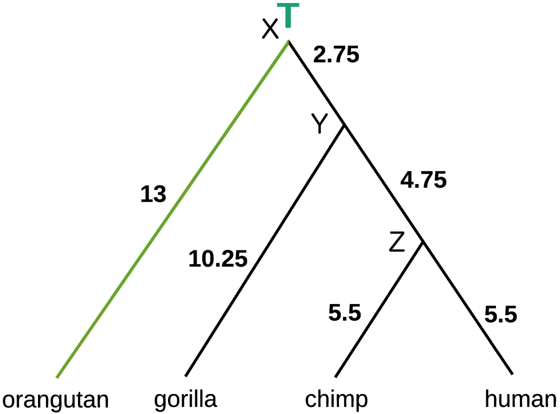
Step 2a: Choose the next branch for simulation

```
Get a branch $b$ with a nucleotide at the start;
$t_b$ = length($b$);
$n$ = nucleotide at start of branch $b$;
```



Step 2a: Choose the next branch for simulation

```
Get a branch $b$ with a nucleotide at the start;
$t_b$ = length($b$);
$n$ = nucleotide at start of branch $b$;
```



Step 2b-d: Sample the new nucleotide

$P(t_b) = e^{Qt_b}$;

Sample new nucleotide n_{new} from row n in $P(t_b)$;

Place n_{new} at the end of branch b;

Step 2b-d: Sample the new nucleotide

```
$P(t_b) = e^{Qt_b}$;
Sample new nucleotide $n_{new}$ from row $n$ in $P(t_b)$;
Place $n_{new}$ at the end of branch $b$;
```

To sample new nucleotide n_{new} we will need the substitution rate matrix Q, and transition probability matrix P.

Substitution rate matrix – TN93

$\Pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ - equilibrium frequencies.

α_1, α_2 - transition rates.

β - transversion rate.

$$Q_{TN93} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} \cdot & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \cdot \end{pmatrix} \end{array}$$

The diagonals are set such that each row sums up to zero, e.g. $q_{TT} = -(\alpha_1\pi_C + \beta\pi_A + \beta\pi_G)$.

Substitution rate matrix – TN93

$$\Pi = (0.22, 0.26, 0.33, 0.19)$$
$$\alpha_1 = 44.229,\ \alpha_2 = 21.781$$
$$\beta = 1$$

$$Q_{TN93} = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & A & \mathrm{G} \end{array} \\ \begin{pmatrix} -0.01957 & 0.01873 & 0.00054 & 0.00031 \\ 0.01584 & -0.01669 & 0.00054 & 0.00031 \\ 0.00036 & 0.00042 & -0.00752 & 0.00674 \\ 0.00036 & 0.00042 & 0.01170 & -0.01249 \end{pmatrix} \end{array}$$

Note: the matrix is scaled to 0.0135 substitutions per mya so that we get reasonable sequences.

Transition probability matrix – TN93

$\Pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ - equilibrium frequencies.

α_1, α_2 - transition rates.

β - transversion rate.

t_b - branch length.

$$\mathrm{P}(t_b) = e^{t_b \mathrm{Q}_{TN93}(\alpha_1, \alpha_2, \beta, \Pi)}$$

Substitution rate matrix – TN93

$$\Pi = (0.22, 0.26, 0.33, 0.19)$$

$$\alpha_1 = 44.229,\ \alpha_2 = 21.781$$

$$\beta = 1$$

$$t_b = 13\,\text{mya}$$

$$P_{TN93}(13\,\text{mya}) = \begin{array}{c} \\ T \\ C \\ A \\ G \end{array} \begin{array}{c} \begin{array}{cccc} T & C & A & G \end{array} \\ \begin{pmatrix} 0.795 & 0.194 & 0.007 & 0.004 \\ 0.164 & 0.824 & 0.007 & 0.004 \\ 0.005 & 0.005 & 0.913 & 0.077 \\ 0.005 & 0.005 & 0.134 & 0.856 \end{pmatrix} \end{array}$$

Sampling substitution times

We start with nucleotide **T**, so we are interested in row T:

$$P_{\text{TN93}}(13\,\text{mya}) = \begin{array}{c} \\ \text{T} \\ \text{C} \\ A \\ \text{G} \end{array} \begin{array}{c} \begin{array}{cccc} \text{T} & \text{C} & A & \text{G} \end{array} \\ \begin{pmatrix} 0.795 & 0.194 & 0.007 & 0.004 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \end{pmatrix} \end{array}$$

Sampling substitution times

We start with nucleotide **T**, so we are interested in row T:

$$P_{\mathrm{TN93}}(13\,\mathrm{mya}) = \begin{array}{c c} & \begin{array}{cccc} \mathrm{T} & \mathrm{C} & A & \mathrm{G} \end{array} \\ \begin{array}{c} \mathrm{T} \\ \mathrm{C} \\ A \\ \mathrm{G} \end{array} & \begin{pmatrix} 0.795 & 0.194 & 0.007 & 0.004 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \end{array}$$

Sample new nucleotide n_{new} with the weights $[p_{\mathsf{T}\to\mathsf{T}}, p_{\mathsf{T}\to\mathsf{C}}, p_{\mathsf{T}\to\mathsf{A}}, p_{\mathsf{T}\to\mathsf{G}}]$

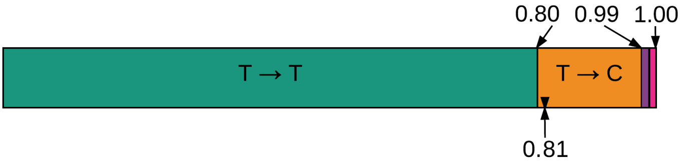
Getting the substitution

Sample u from $\mathcal{U}(0, 1)$.
E.g. $u = 0.81$.

Getting the substitution

Sample u from $\mathcal{U}(0, 1)$.

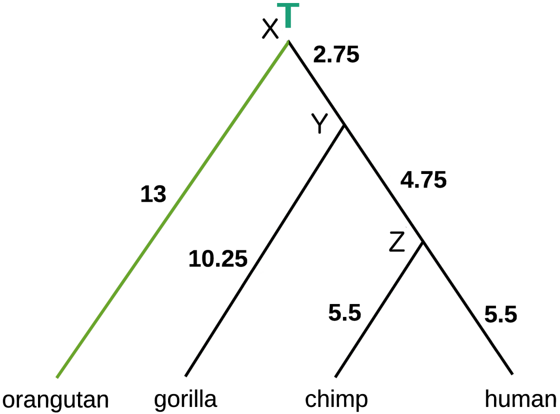
E.g. $u = 0.81$.



Selected substitution is T → C.

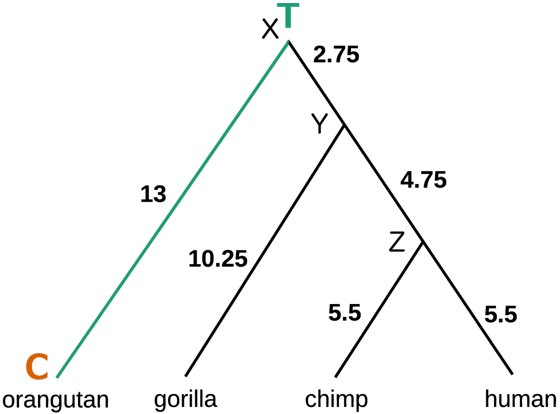
Step 2b-d: Sample the new nucleotide

$P(t_b) = e^{Qt_b}$;
Sample new nucleotide n_{new} from row n in $P(t_b)$;
Place n_{new} at the end of branch b;



Step 2b-d: Sample the new nucleotide

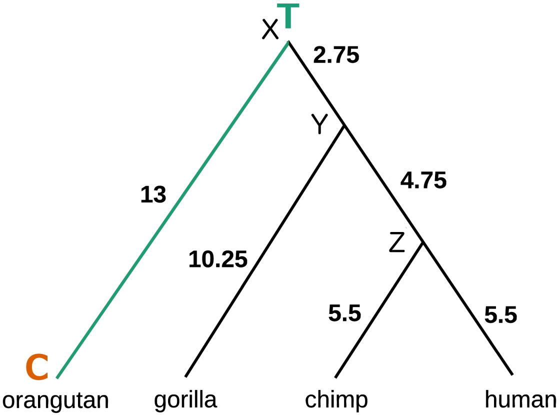
```
$P(t_b) = e^{Qt_b}$;
Sample new nucleotide $n_{new}$ from row $n$ in $P(t_b)$;
Place $n_{new}$ at the end of branch $b$;
```



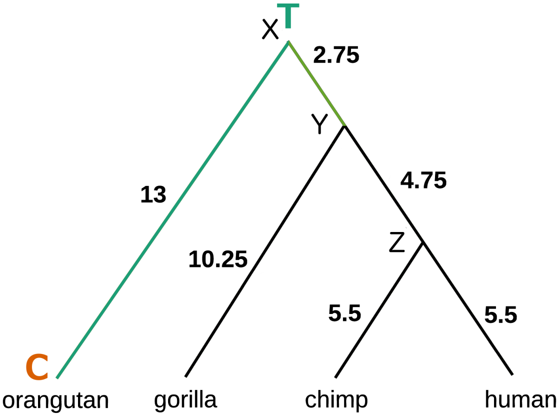
Repeat step 2

```
while not all branches are used do
    Get a branch $b$ with a nucleotide at the start;
    $t_b$ = length($b$);
    $n$ = nucleotide at start of branch $b$;
    $P(t_b) = e^{Qt_b}$;
    Sample new nucleotide $n_{new}$ from row $n$ in $P(t_b)$;
    Place $n_{new}$ at the start of the daughter branches of $b$;
end
```

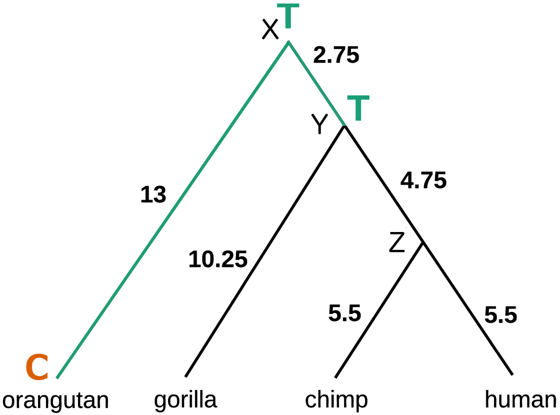
Evolution



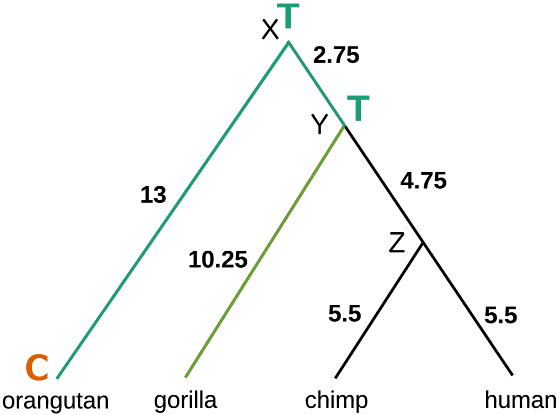
Evolution



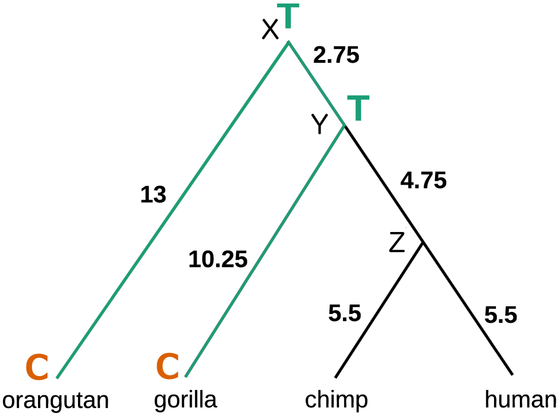
Evolution



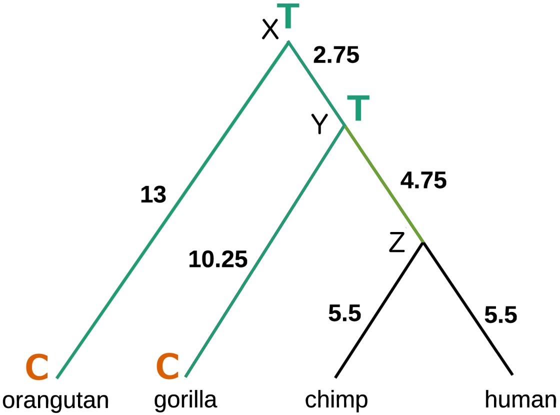
Evolution



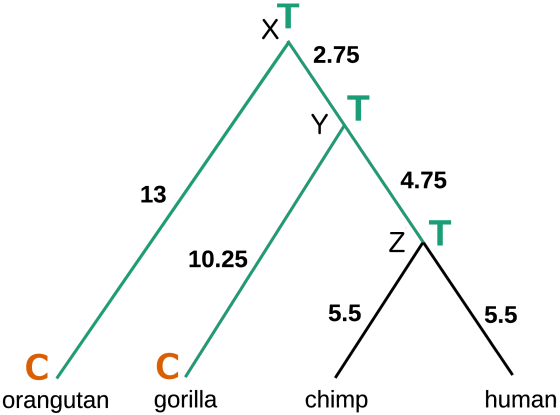
Evolution



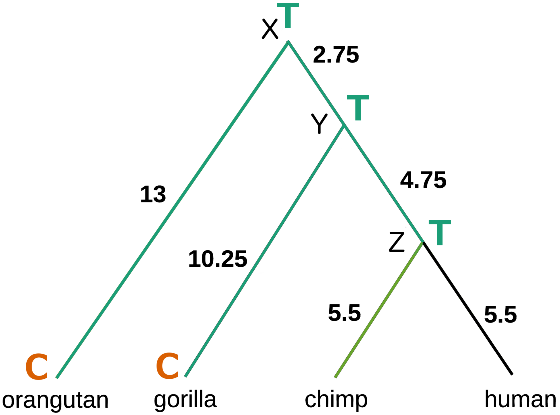
Evolution



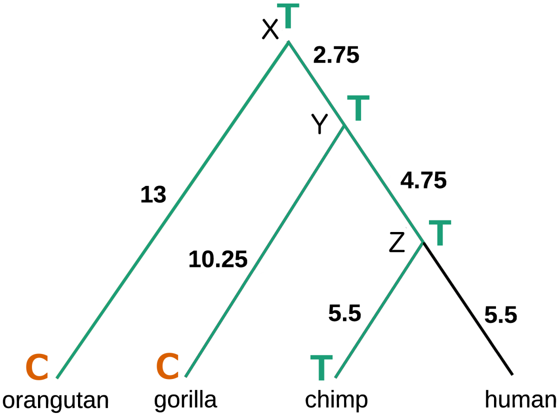
Evolution


X
T
2.75
Y
T
13
4.75
Z
T
10.25
5.5
5.5
C
C
orangutan
gorilla
chimp
human

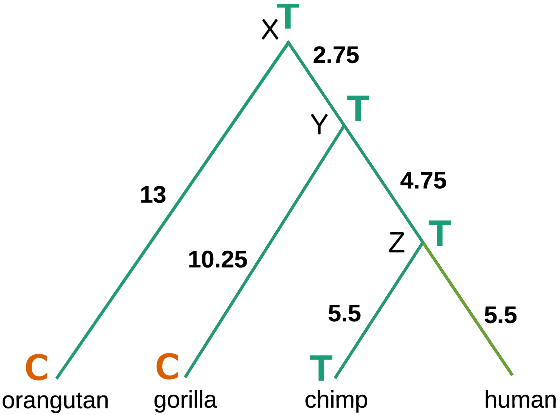
Evolution


X
T
2.75
Y
T
4.75
13
10.25
Z
T
5.5
5.5
C
C
orangutan
gorilla
chimp
human

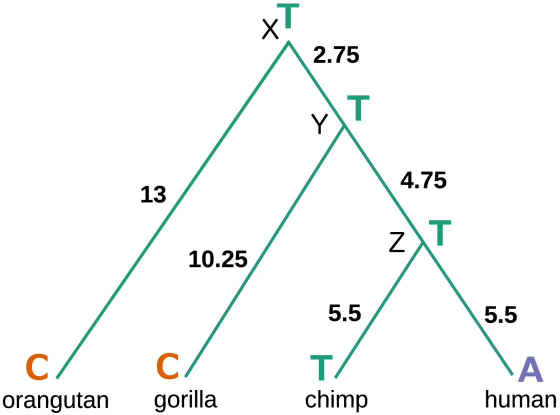
Evolution


X
T
2.75
Y
T
13
4.75
Z
T
10.25
5.5
5.5
C
C
T
orangutan
gorilla
chimp
human

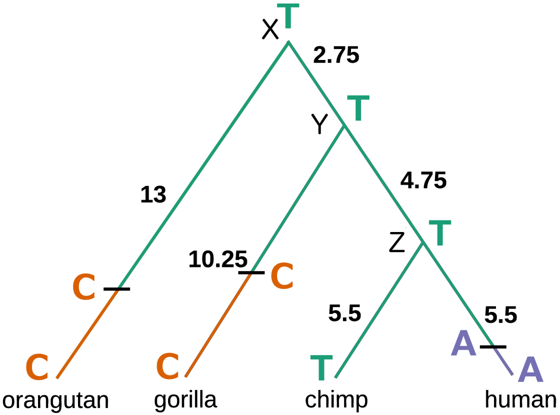
Evolution


X
T
2.75
Y
T
13
4.75
Z
T
10.25
5.5
5.5
C
C
T
orangutan
gorilla
chimp
human

Evolution


T
X
2.75
T
Y
4.75
13
T
Z
10.25
5.5
5.5
C
C
T
A
orangutan
gorilla
chimp
human

Evolution


X
T
2.75
Y
T
13
4.75
Z
T
10.25
C
C
5.5
5.5
A
C
C
T
A
orangutan
gorilla
chimp
human

Exercise for today

1. Use a random number generator to “roll dice”
2. Evolve a character along the tree;

All of the characters together will produce an alignment.

Algorithm

```
$N$ = number of sites in the alignment;
$Q$ = substitution rate matrix;
for $i = 1$ to $N$ do
    Sample a nucleotide $n$ from the initial distribution;
    Add $n$ to the sequence of the root node;
end
while not all branches are visited do
    Get a branch $b$ with a sequence at the start;
    $t_b$ = length($b$);
    $P(t_b) = e^{Qt_b}$;
    for $i = 1$ to $N$ do
        $n$ = nucleotide at position $i$ at the start of branch $b$;
        Sample new nucleotide $n_{new}$ from row $n$ in $P(t_b)$;
        Place $n_{new}$ at the end of sequences in the daughter
          branches of $b$;
    end
end
```

References I

- Paabo, S. (2003). The mosaic that is our genome. *Nature*, 421(6921):409–12.