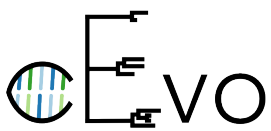
Computational Biology

Lecturers:
Tanja Stadler, Tim Vaughan & Carsten Magnus

Teaching Assistants:
Antoine Zwaans, Adrian Lison,
James Munday & Marcus Overwater

Computational Evolution
Department of Biosystems Science and Engineering

HS 2023



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

How to study evolution?

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

The easiest way to study something is by observation.

How to study evolution?

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

The easiest way to study something is by observation.

- Wetlab
  - Very realistic;
  - Time-consuming and expensive;
  - Impossible (sometimes).

How to study evolution?

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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 Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

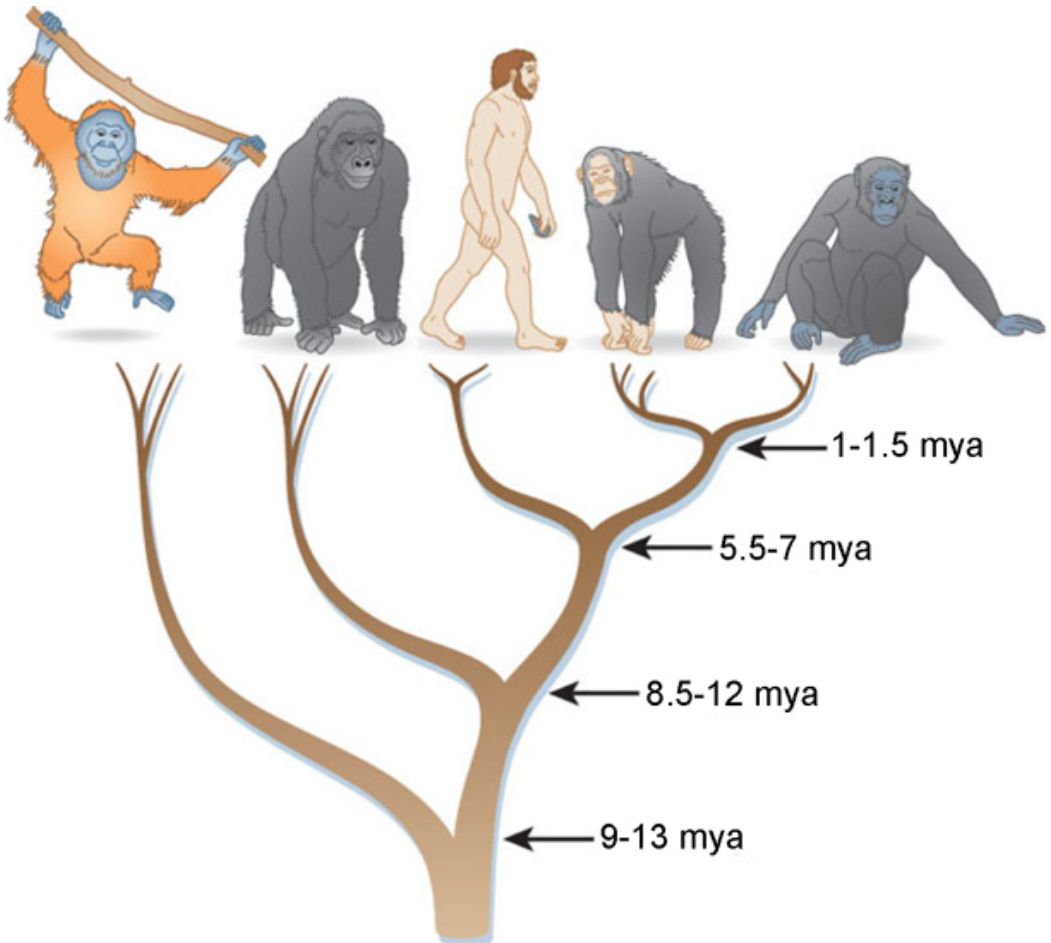
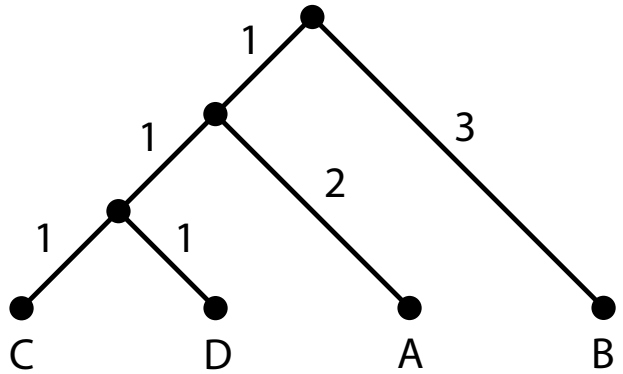


Figure adapted from [?]

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

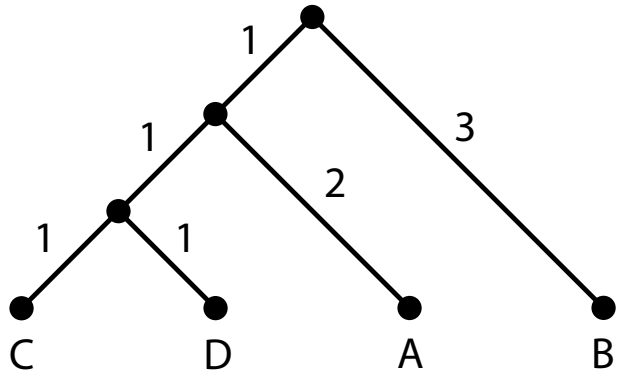
Storing trees: Newick format



- 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?

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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. $(C:1, D:1)$)
    - 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? $(((C:1, D:1):1, A:2):1, B:3);$

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

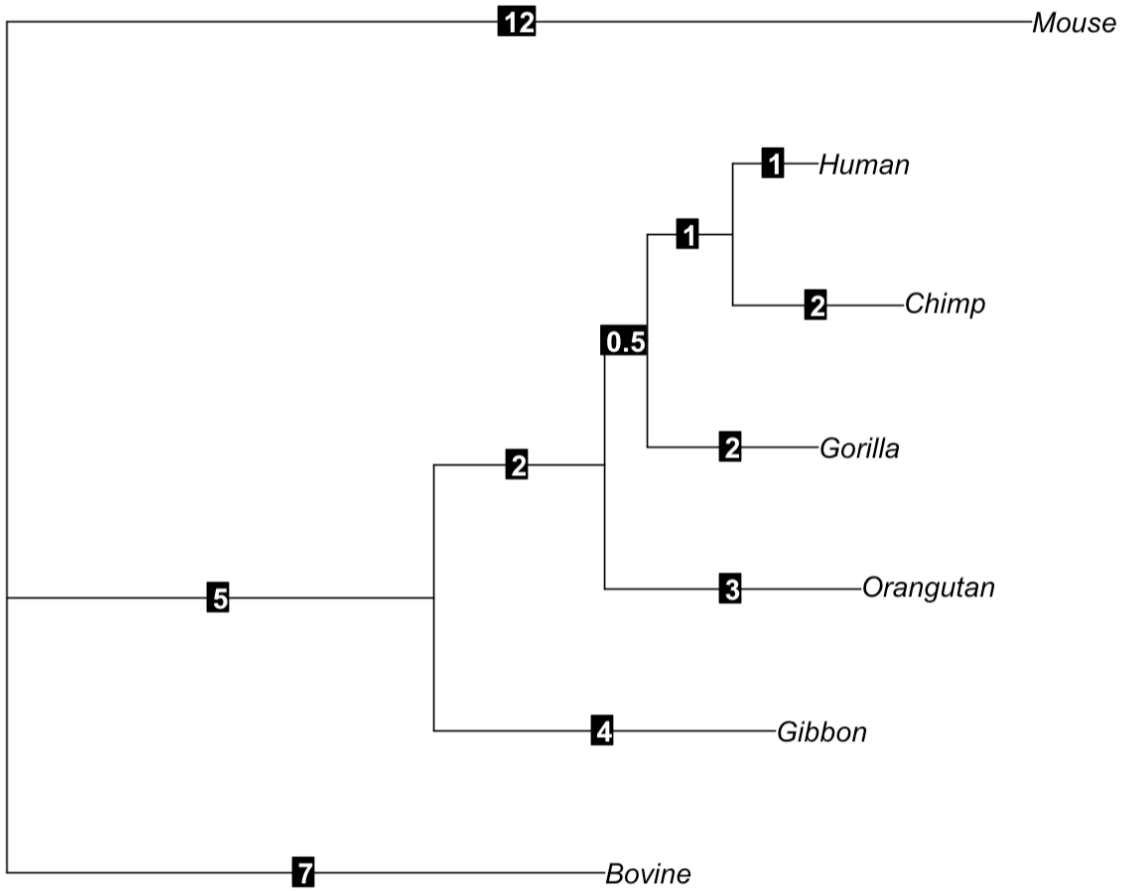
Storing trees: Newick format

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

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Storing trees: Newick format

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



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution Simulation Algorithm

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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 $\mathbf{P}(\mathbf{t_b})$.
   - Sample a new nucleotide for each position in the sequence.

Step 1: Initialization of the starting sequence

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

Step 1: Initialization of the starting sequence

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

Step 1: Initialization of the starting sequence

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

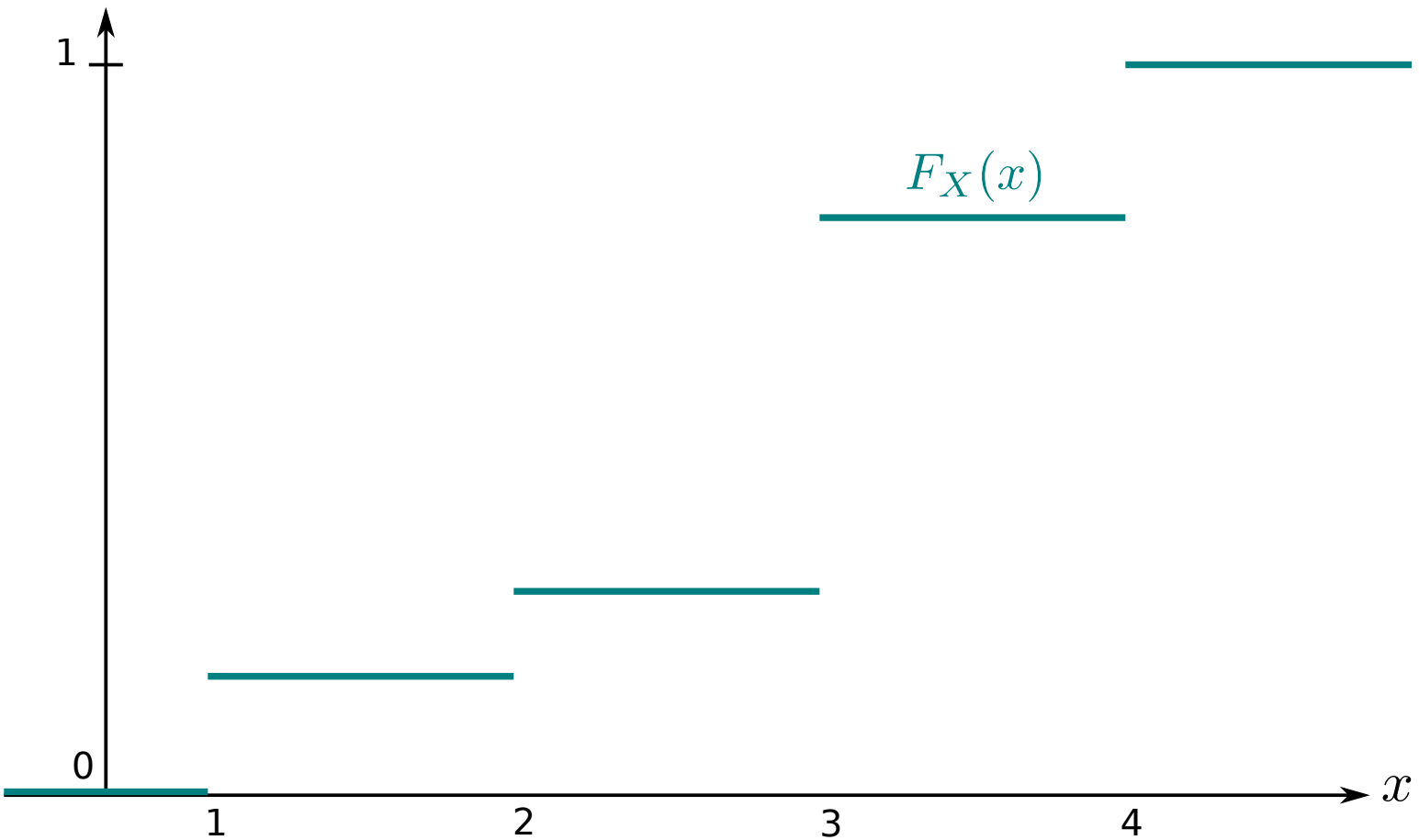
From the vector of equilibrium frequencies of nucleotides

	T	C	A	G
Π	0.22	0.26	0.33	0.19

Knowing Π, how do we sample a nucleotide?

Inverse transform method

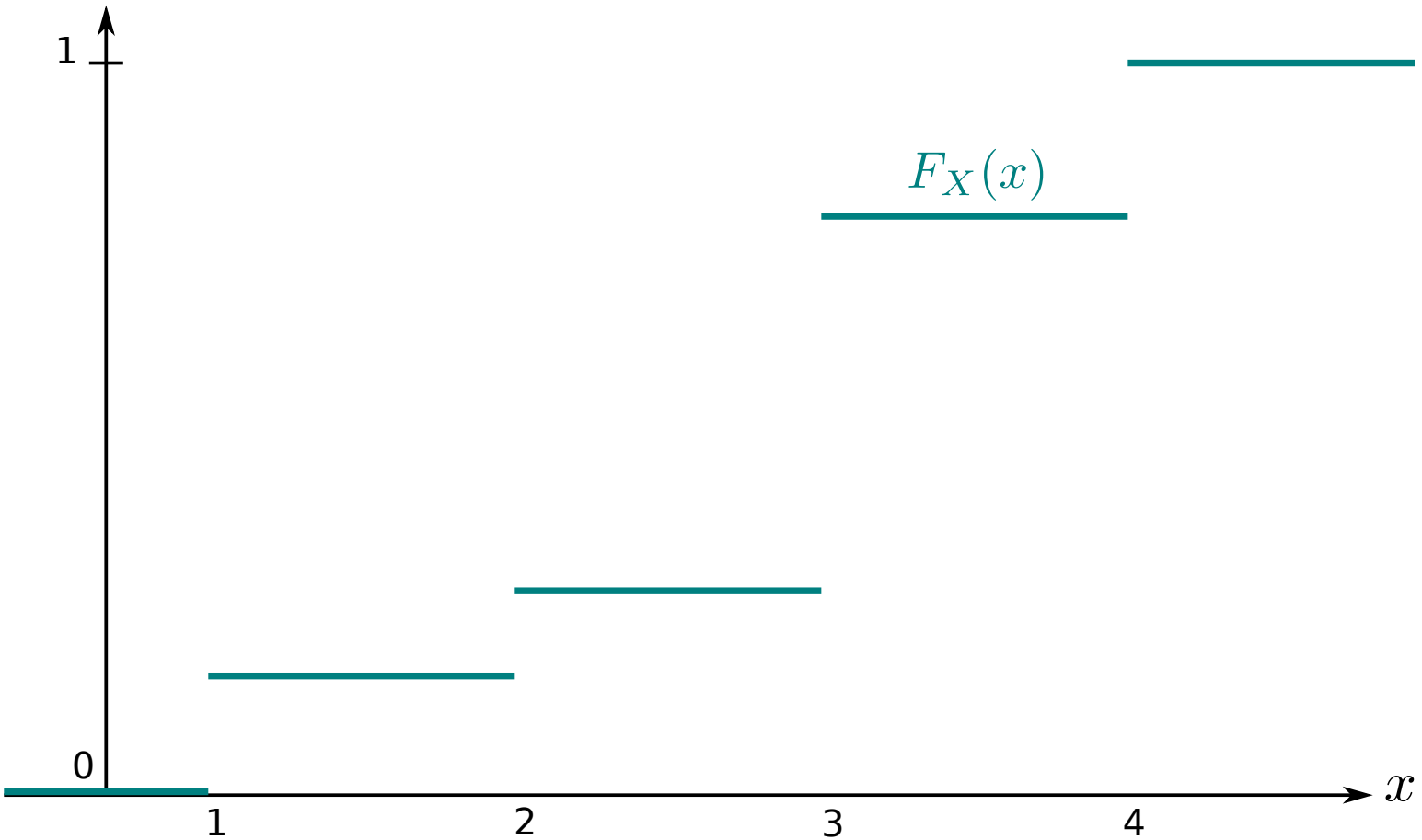
- Sample $\mathfrak{u}$ from $\mathfrak{U}(0,1)$;



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Inverse transform method

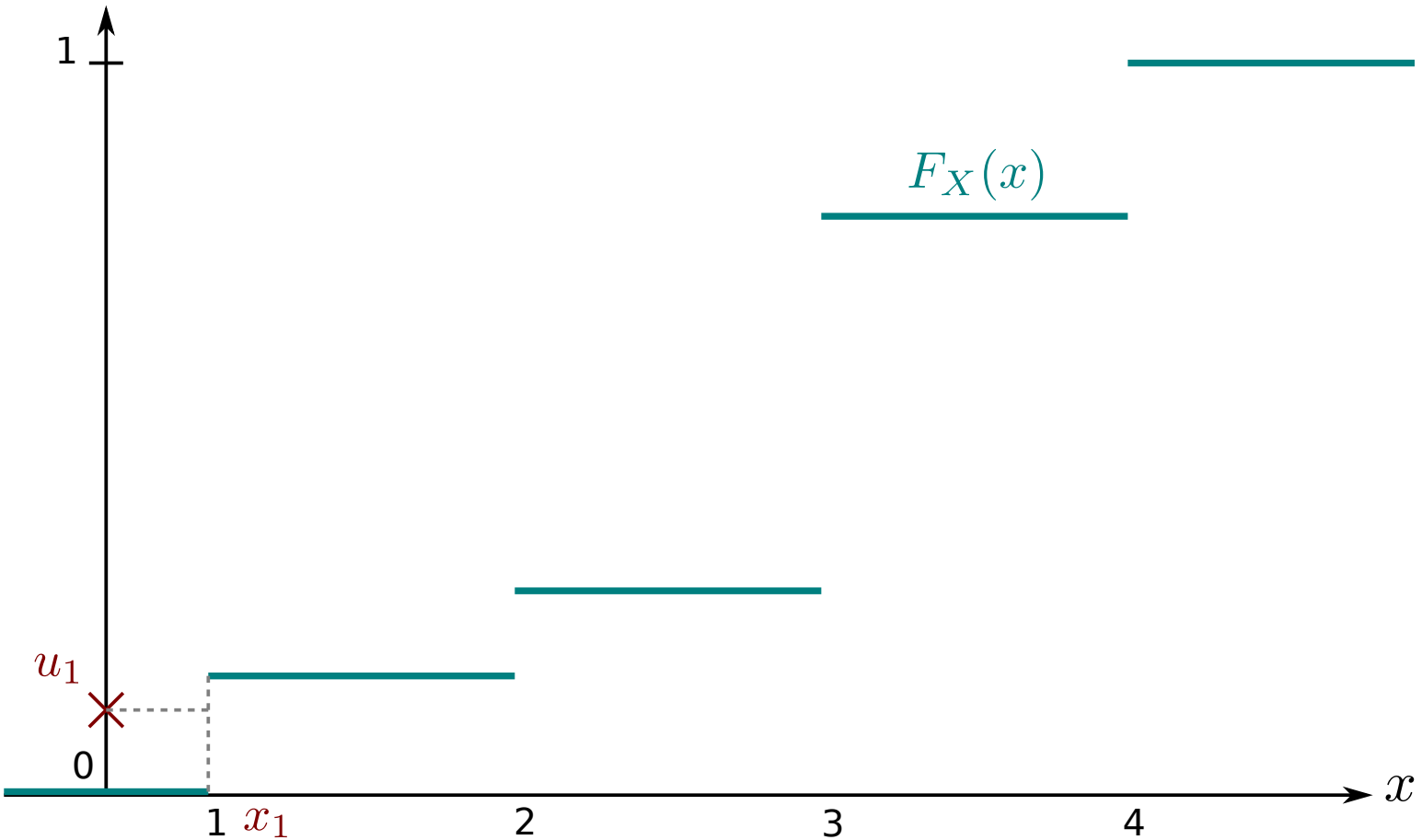
- Sample $\mathfrak{u}$ from $\mathcal{U}(0, 1)$;
- Transform $\mathfrak{u}$ into a sample from the desired distribution using the **CDF** == **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Inverse transform method

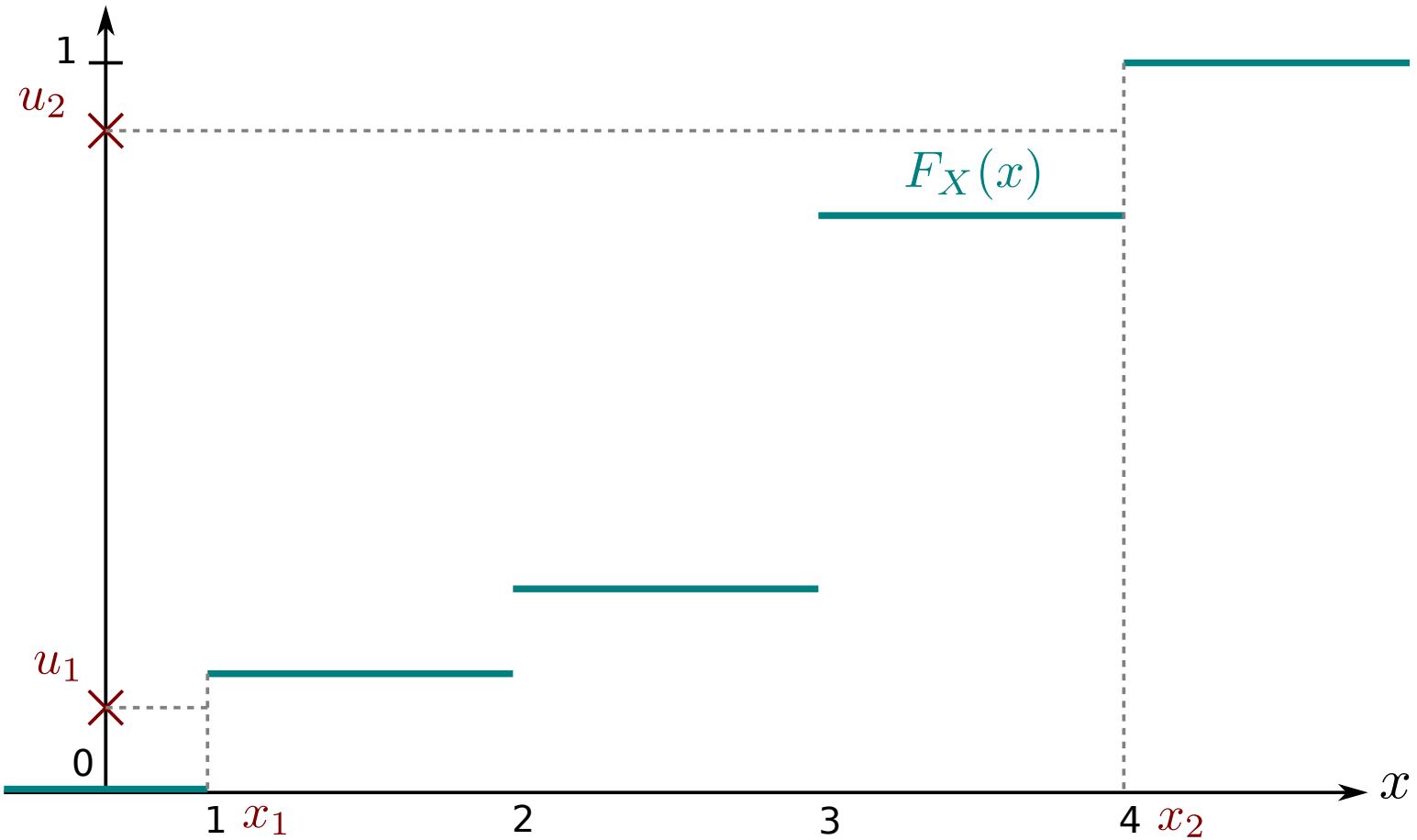
- Sample $\mathfrak{u}$ from $\mathcal{U}(0, 1)$;
- Transform $\mathfrak{u}$ into a sample from the desired distribution using the **CDF** == **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Inverse transform method

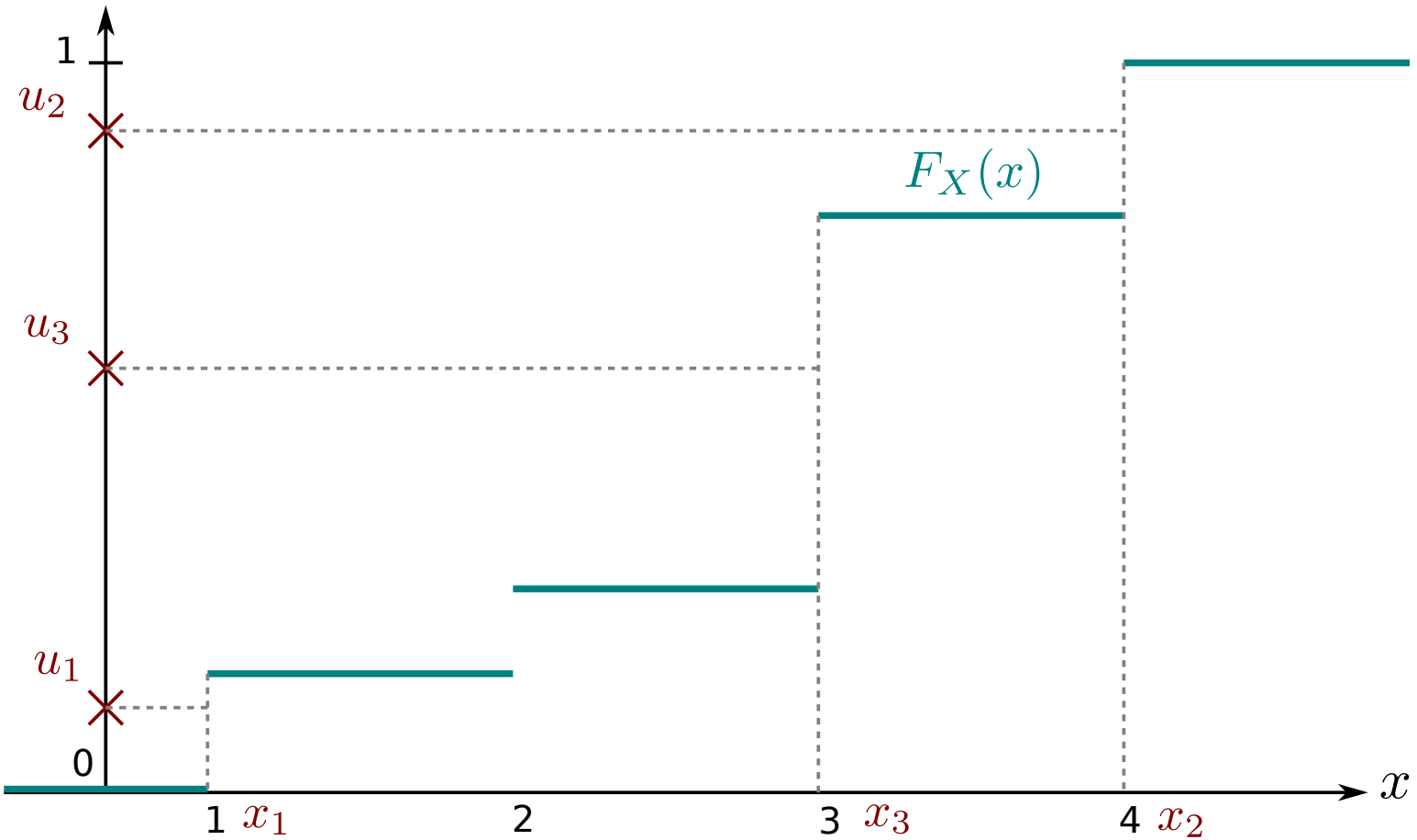
- Sample $\mathfrak{u}$ from $\mathcal{U}(0, 1)$;
- Transform $\mathfrak{u}$ into a sample from the desired distribution using the **CDF** == **C**umulative **D**istribution **F**unction $F_X(x) = P(X \leqslant x)$.



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Inverse transform method

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

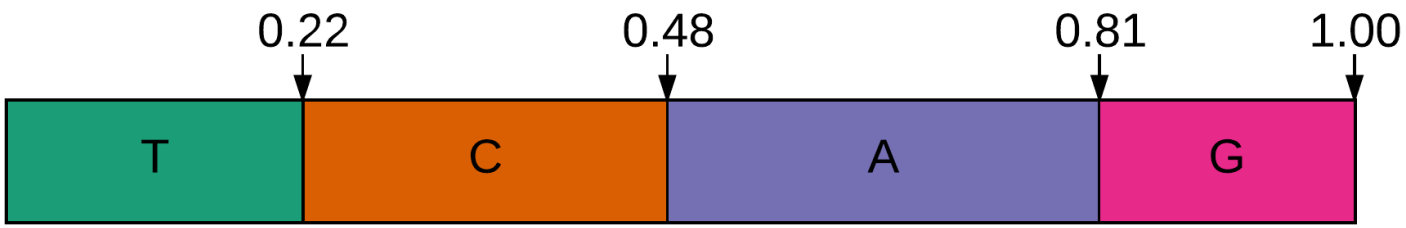


The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

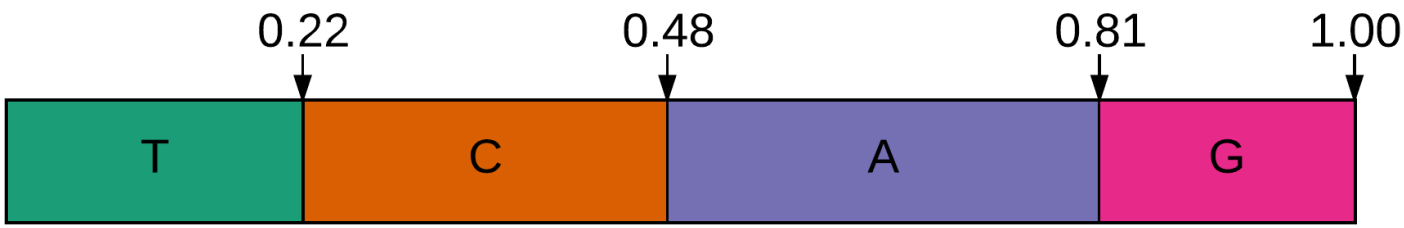
	T	C	A	G
Π	0.22	0.26	0.33	0.19
CDF	0.22	0.48	0.81	1.00



Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

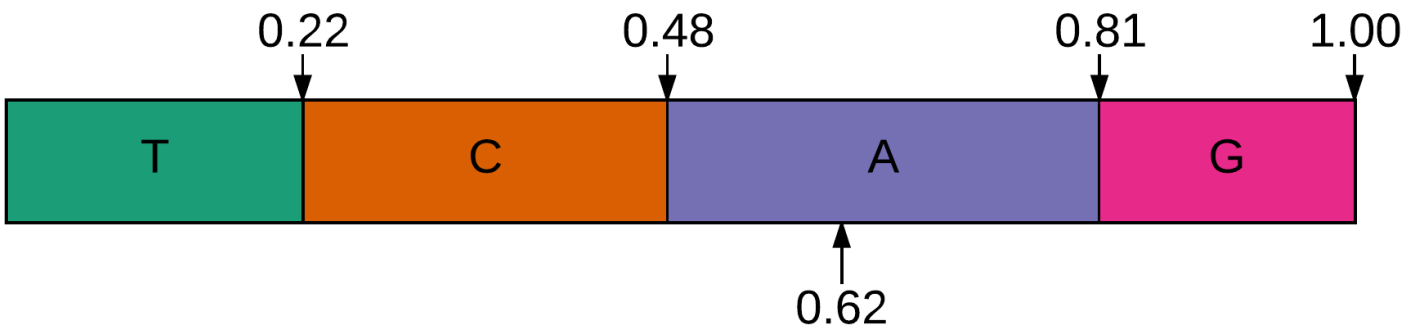


Sample u from $\mathcal{U}(0, 1)$.

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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



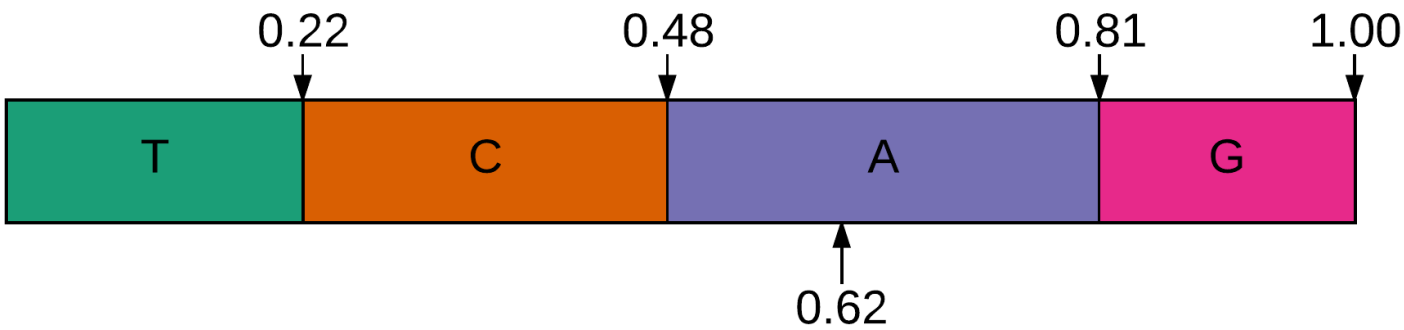
Sample u from $U(0, 1)$.

E.g. $u = 0.62$.

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

	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 $U(0, 1)$.

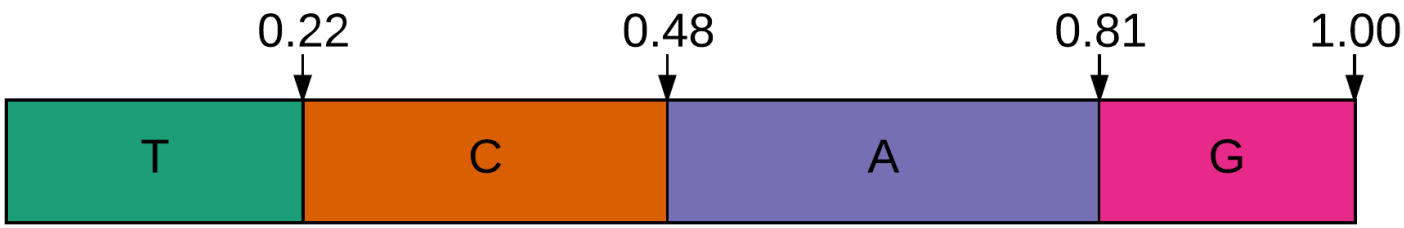
E.g. $u = 0.62$.

Select nucleotide **A**.

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

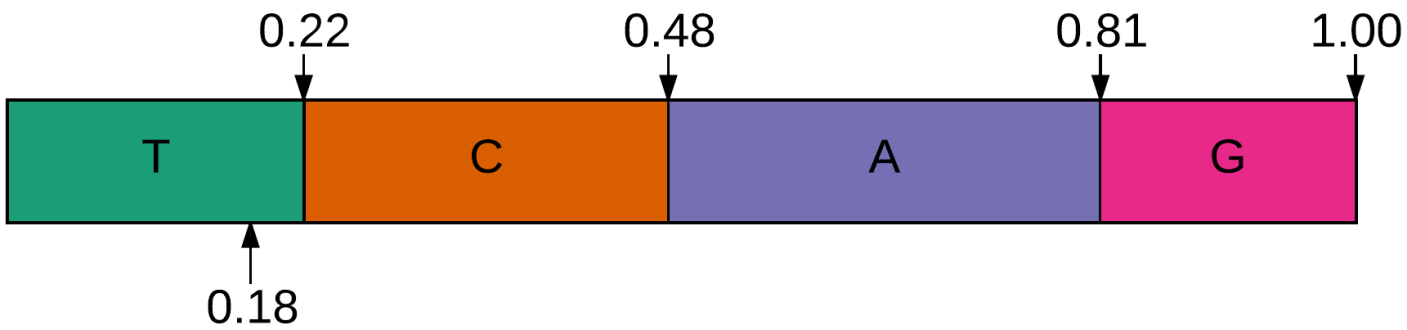


Sample u from $\mathcal{U}(0, 1)$.

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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



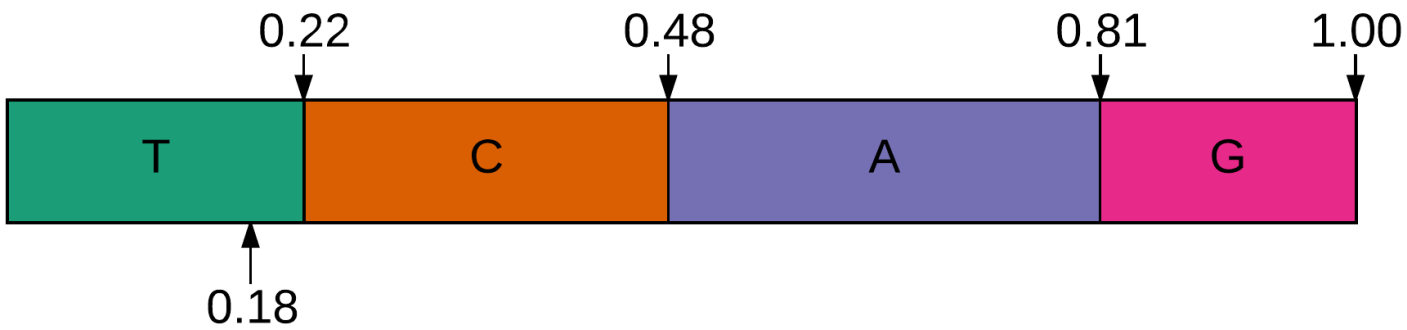
Sample u from $U(0, 1)$.

E.g. $u = 0.18$.

Sampling discrete random variables

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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



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

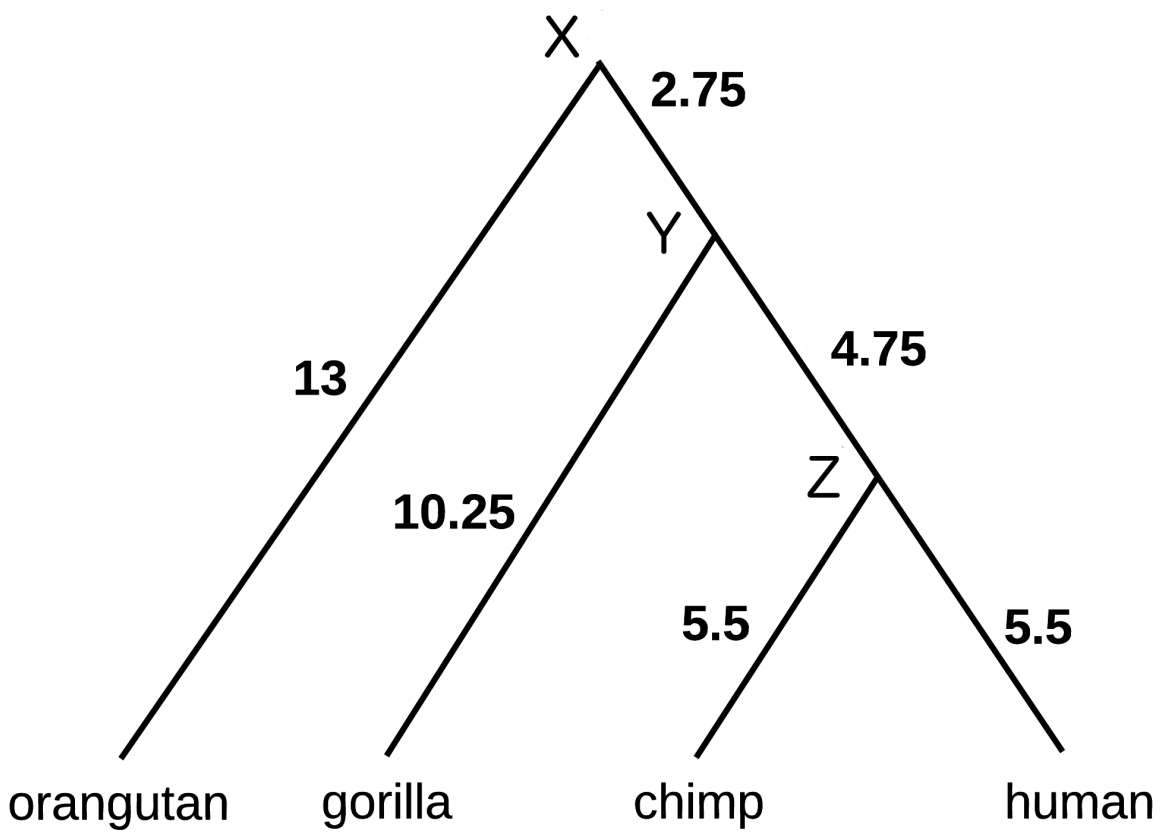
E.g. $\mathfrak{u} = 0.18$.

Select nucleotide **T**.

Step 1: Initializing the starting sequence

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

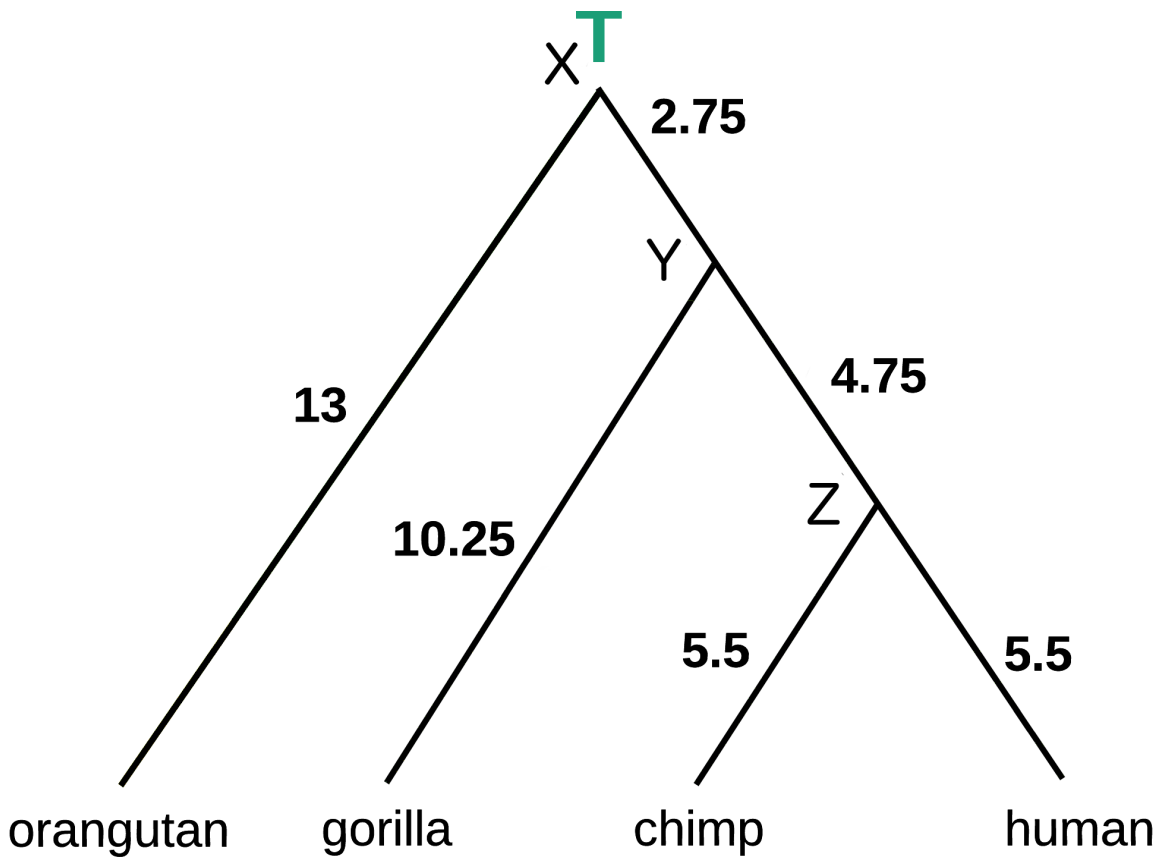
1b. Place $\mathfrak{n}$ on the root node;



Step 1: Initializing the starting sequence

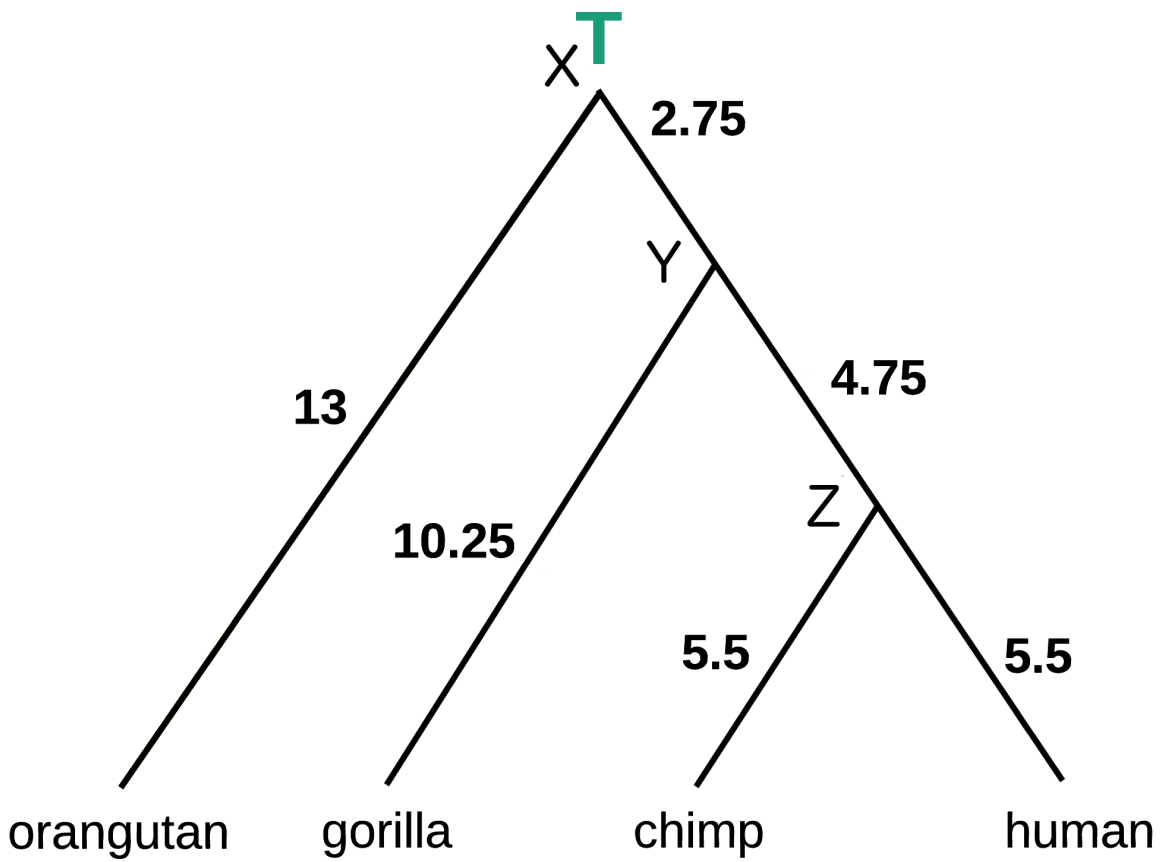
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

1b. Place $\mathfrak{n}$ on the root node;



Step 2a: Choose the next branch for simulation

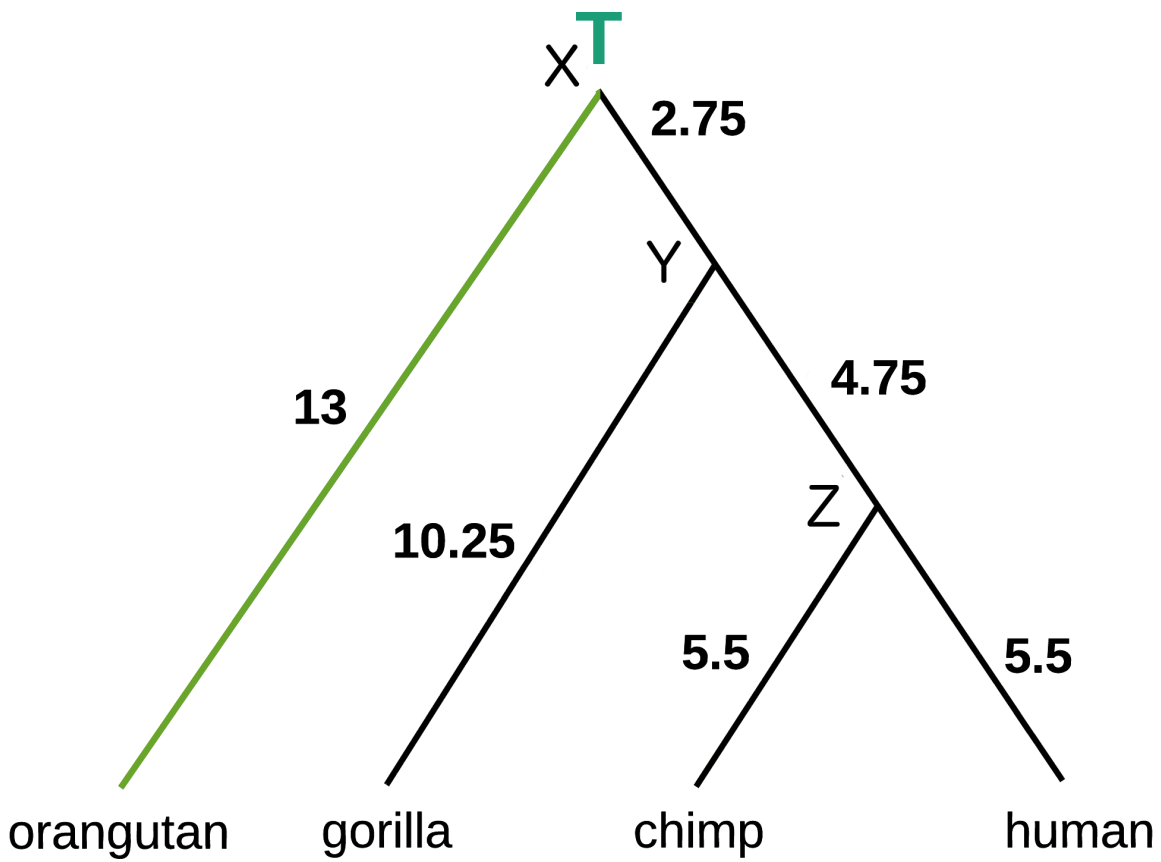
Get a branch $\mathfrak{b}$ with a nucleotide at the start;
$\mathfrak{t}_{\mathfrak{b}} = \text{length}(\mathfrak{b})$;
$\mathfrak{n}$ = nucleotide at start of branch $\mathfrak{b}$;



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Step 2a: Choose the next branch for simulation

Get a branch $\mathfrak{b}$ with a nucleotide at the start;
$\mathfrak{t}_{\mathfrak{b}} = \text{length}(\mathfrak{b})$;
$\mathfrak{n}$ = nucleotide at start of branch $\mathfrak{b}$;



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Step 2b-d: Sample the new nucleotide

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

$\mathrm{P}(\mathbf{t_b}) = \mathbf{e}^{\mathrm{Q}\mathbf{t_b}}$;
Sample new nucleotide $\mathbf{n}_{new}$ from row $\mathbf{n}$ in $\mathrm{P}(\mathbf{t_b})$;
Place $\mathbf{n}_{new}$ at the end of branch $\mathbf{b}$;

To sample new nucleotide $\mathbf{n}_{new}$ we will need the substitution rate matrix Q, and transition probability matrix P.

Substitution rate matrix – TN93

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

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

$\beta = 1$

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

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

Transition probability matrix – TN93

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

α_1, α_2 - transition rates.

β - transversion rate.

t_b - branch length.

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

Substitution rate matrix – TN93

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

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

$\beta = 1$

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

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

Sampling substitution times

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

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

Sampling substitution times

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

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

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

Getting the substitution

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

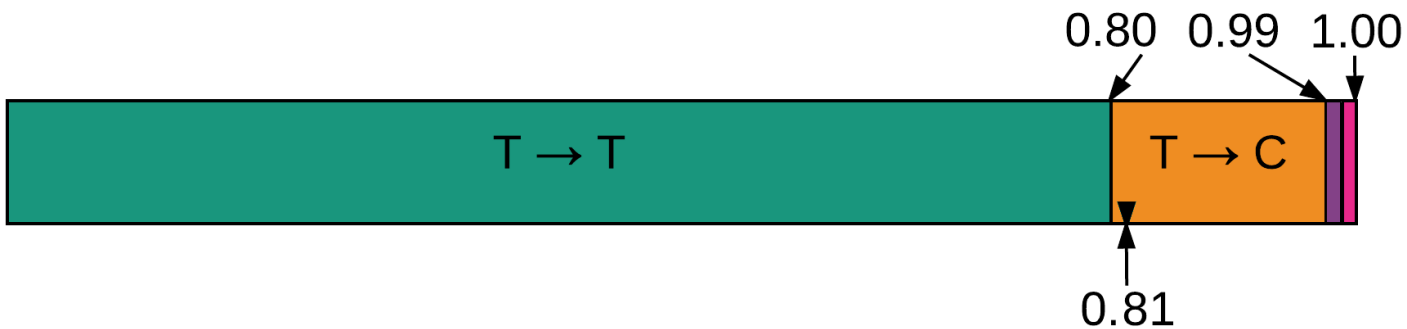
E.g. $\mathfrak{u} = 0.81$.

Getting the substitution

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

E.g. $\mathfrak{u} = 0.81$.

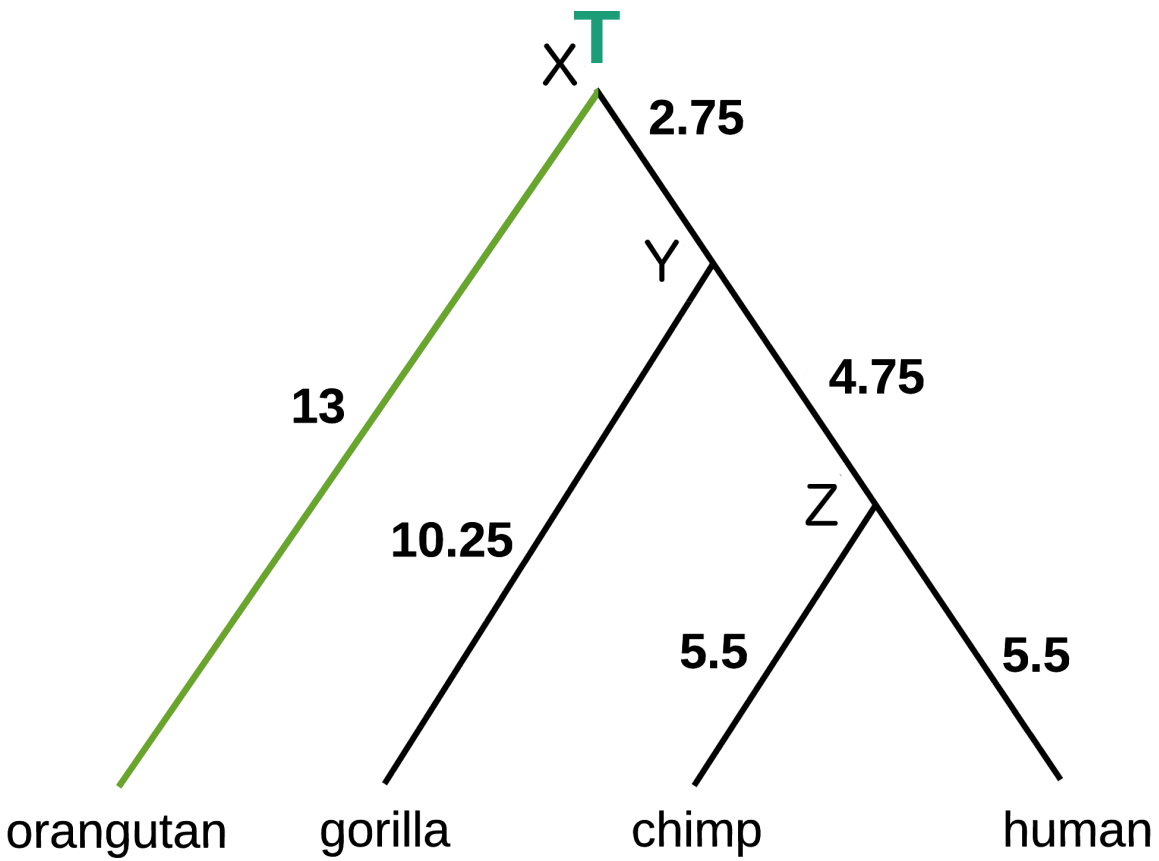


Selected substitution is T $\rightarrow$ C.

Step 2b-d: Sample the new nucleotide

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

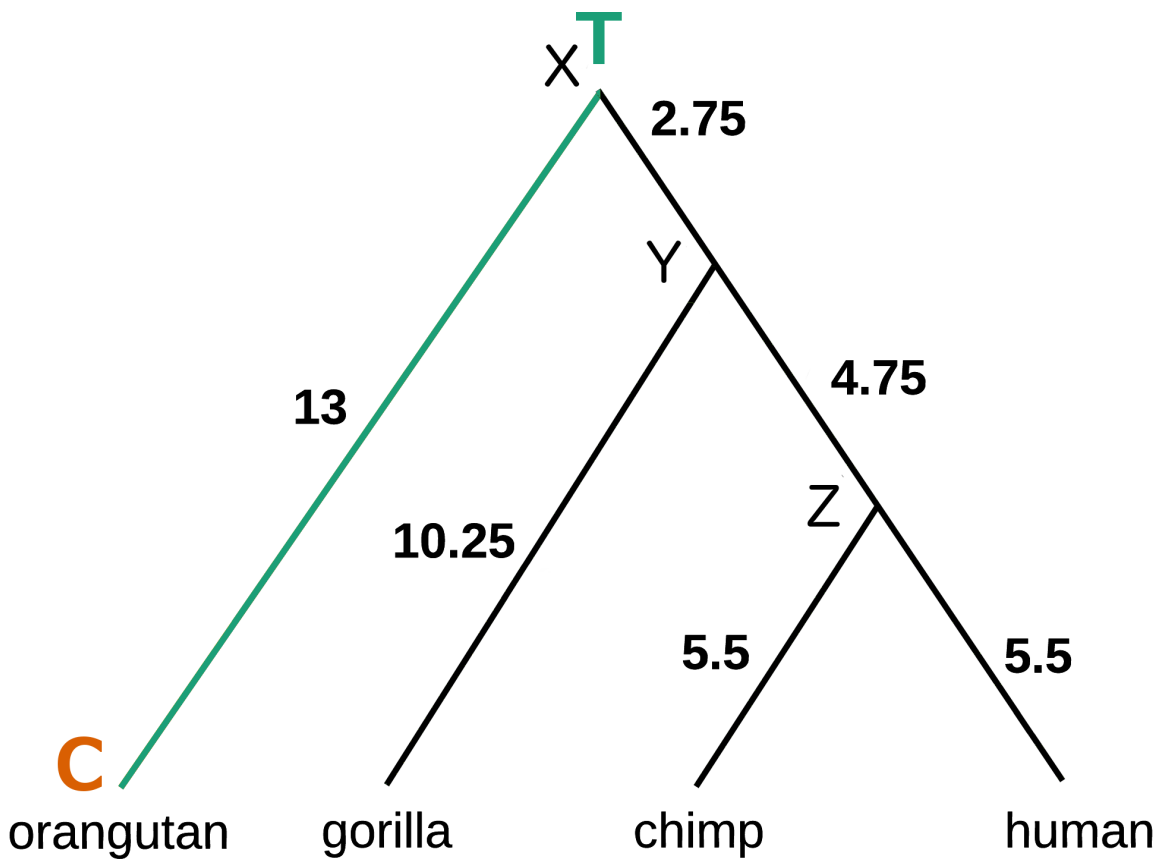
$P(t_b) = e^{Qt_b}$;
Sample new nucleotide n_{new} from row n in $P(t_b)$;
Place n_{new} at the end of branch b;



Step 2b-d: Sample the new nucleotide

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

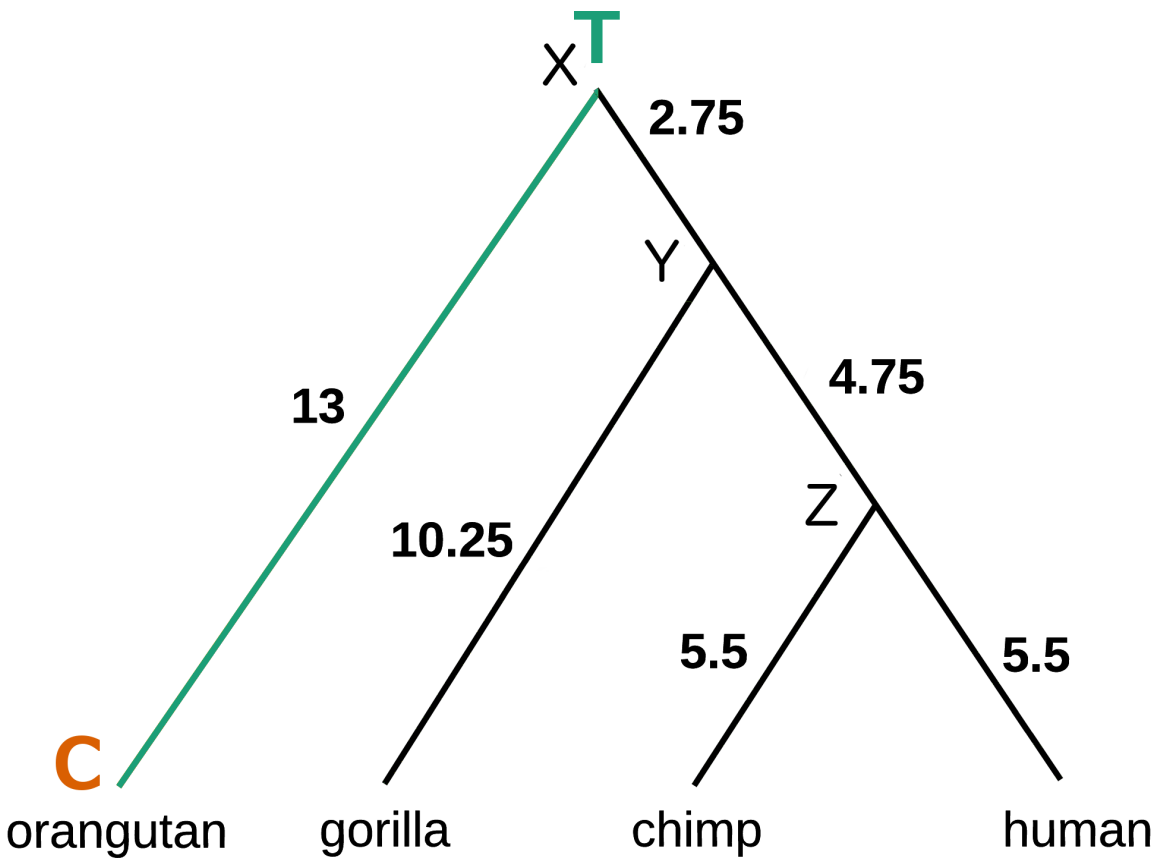


Repeat step 2

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

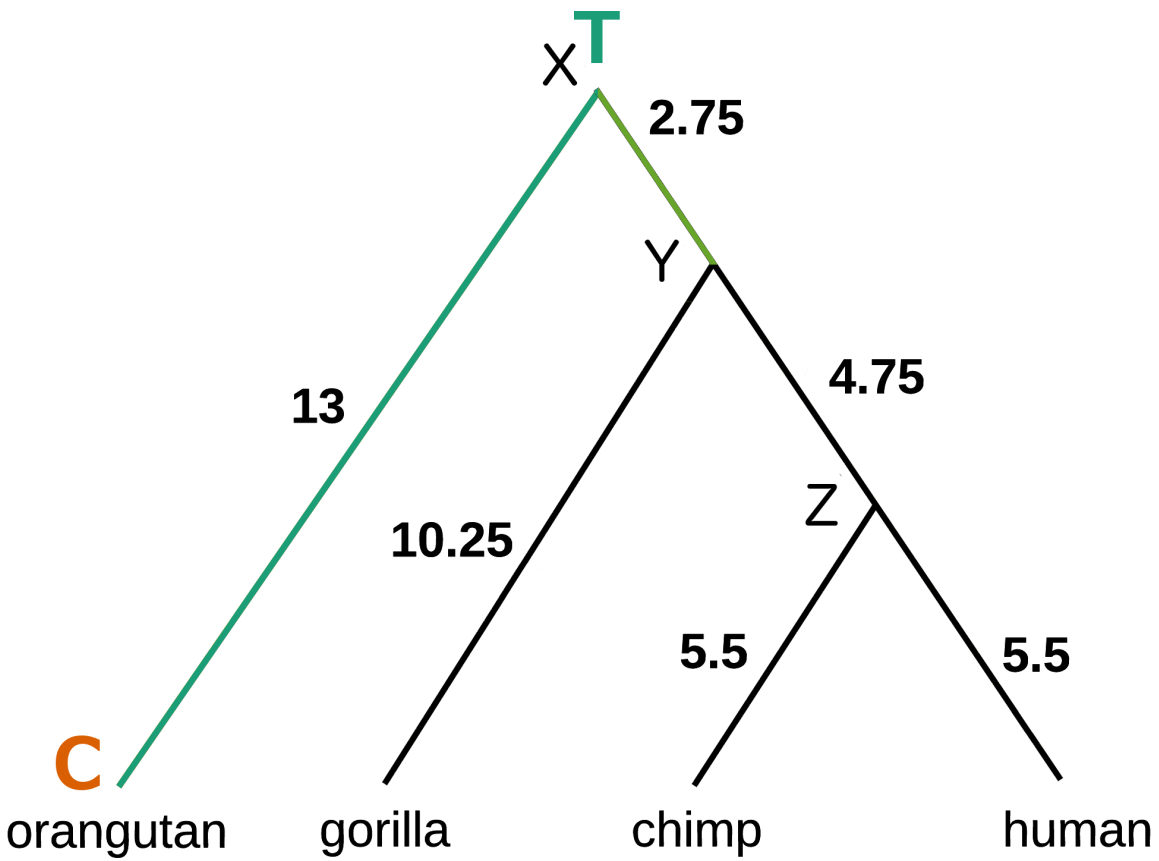
while *not all branches are used* **do**
    Get a branch $\mathbf{b}$ with a nucleotide at the start;
    $\mathbf{t_b} = \text{length}(\mathbf{b})$;
    $\mathbf{n}$ = nucleotide at start of branch $\mathbf{b}$;
    $\mathrm{P}(\mathbf{t_b}) = \mathbf{e}^{\mathrm{Q}\mathbf{t_b}}$;
    Sample new nucleotide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

Evolution



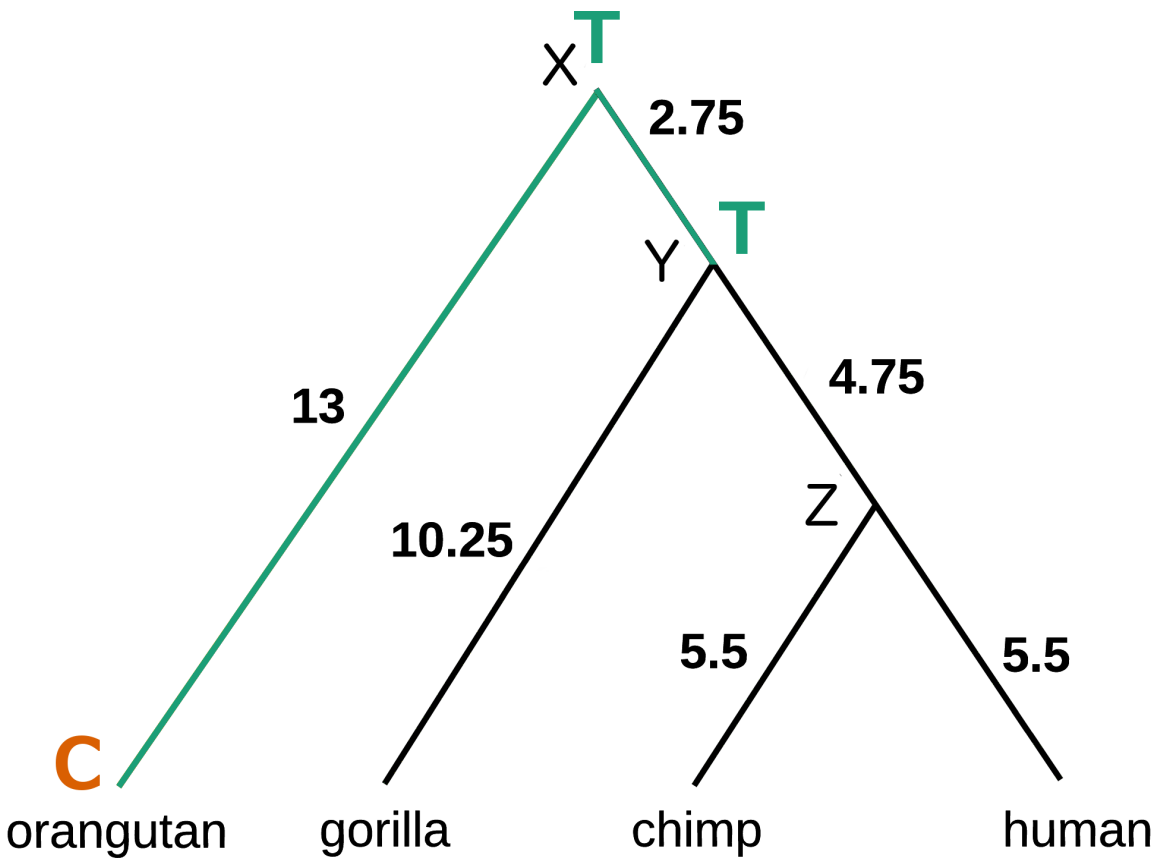
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



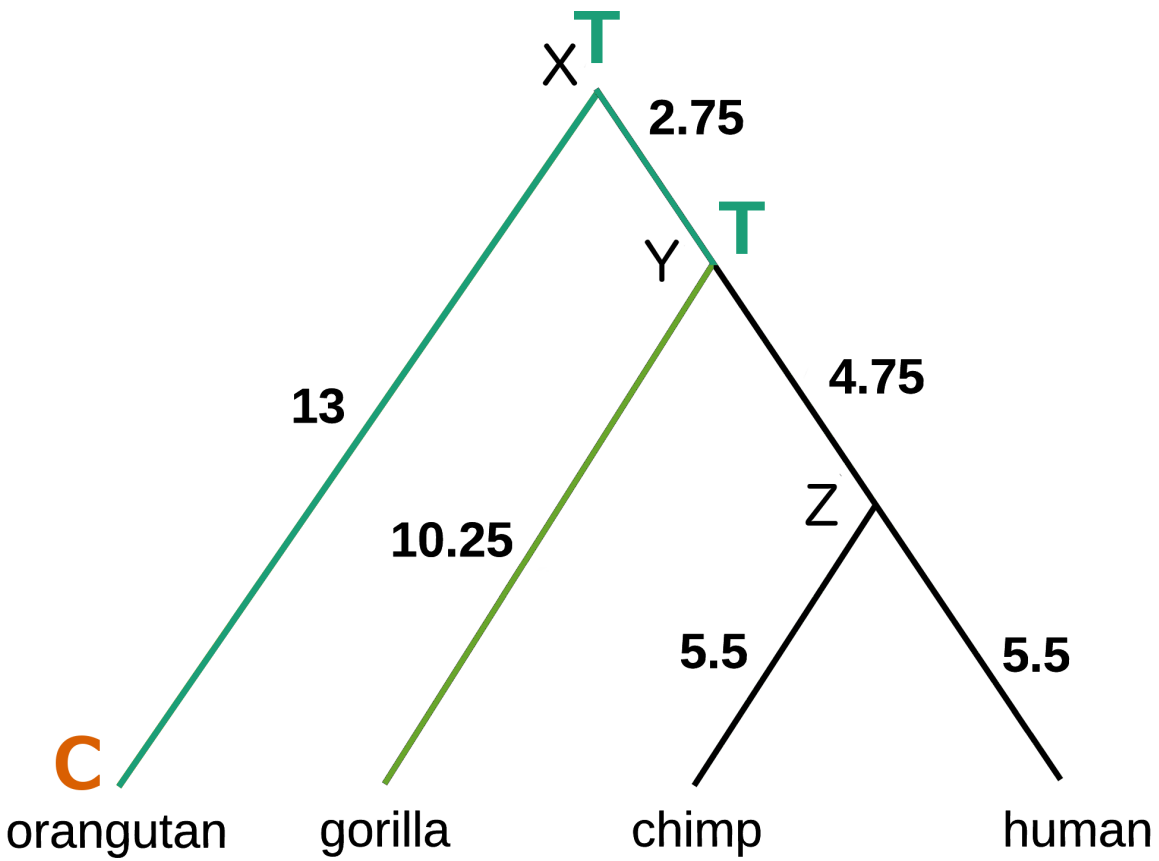
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



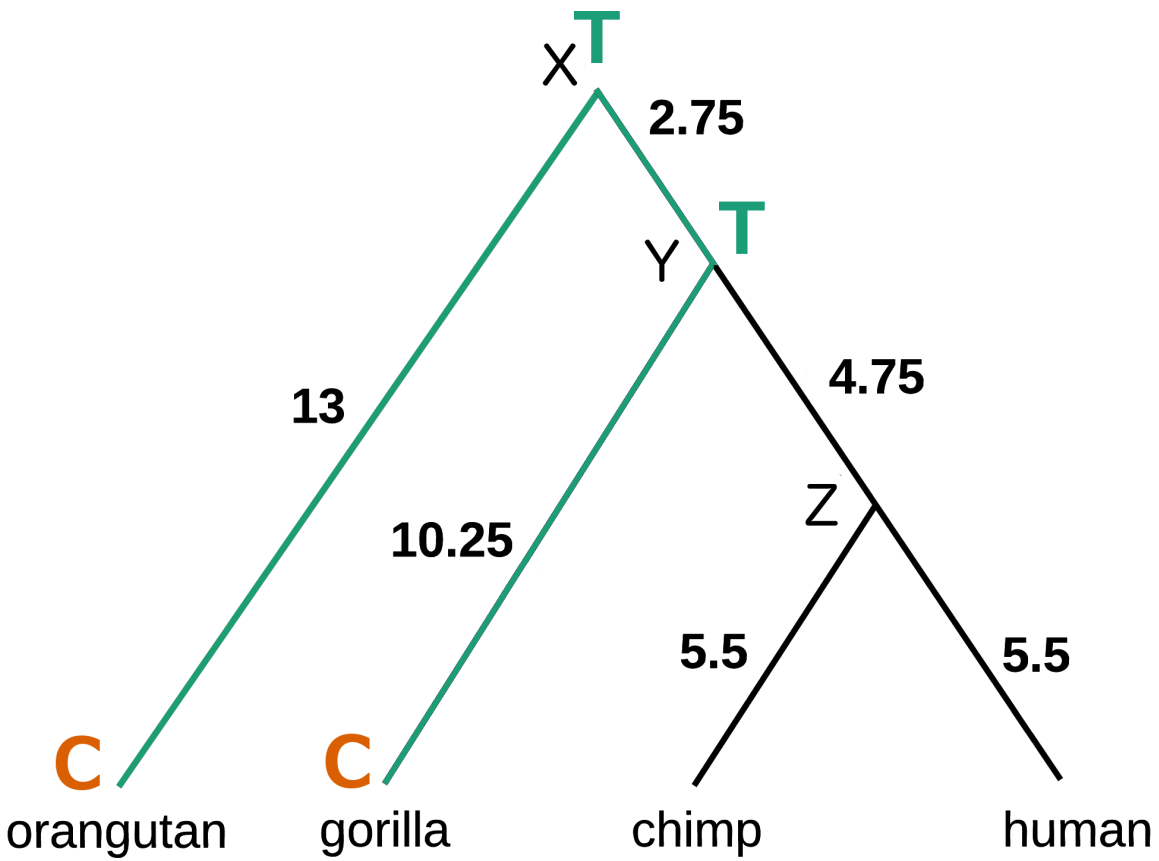
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



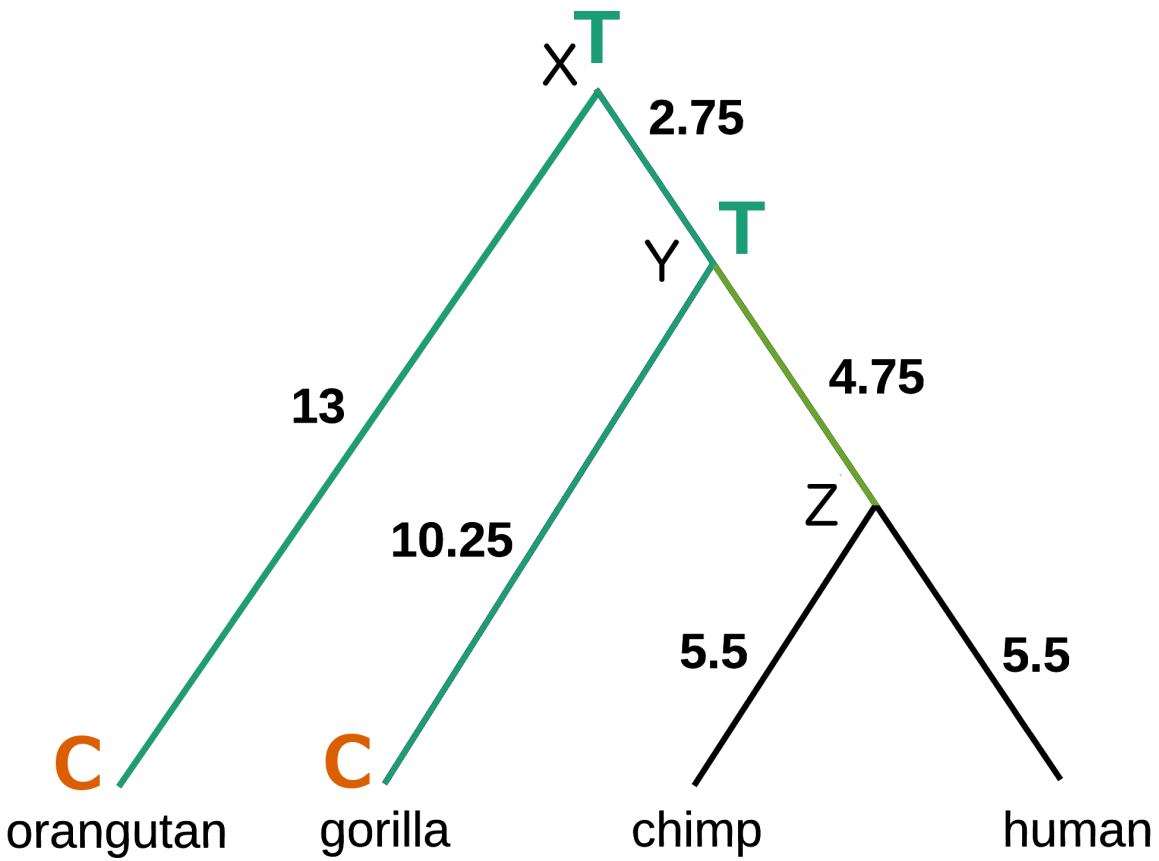
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



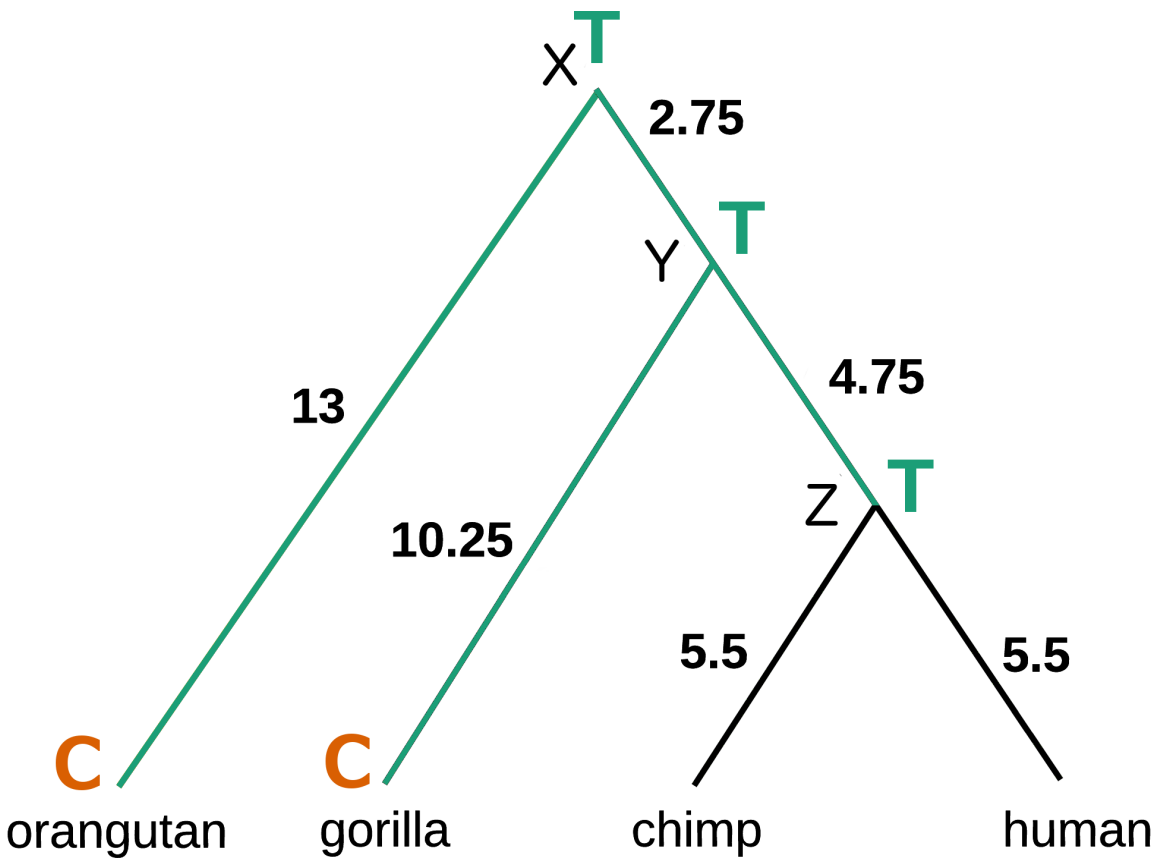
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



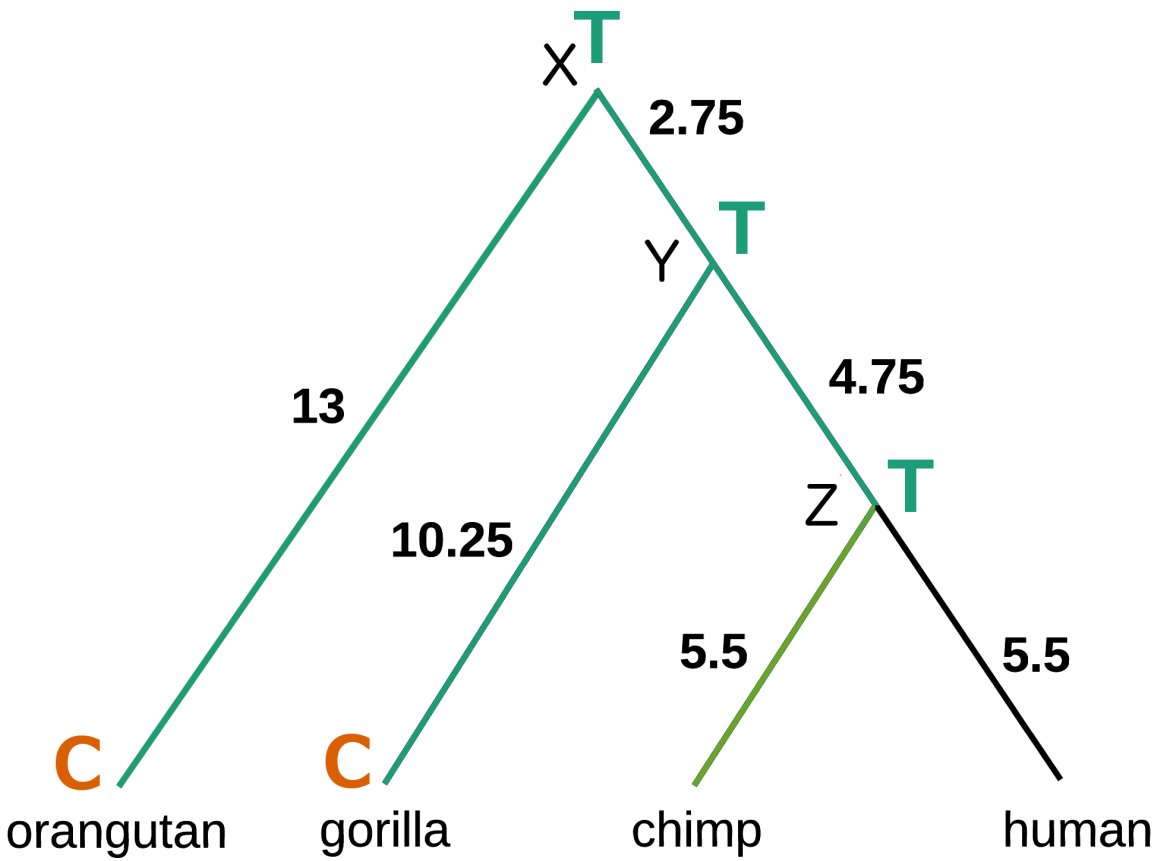
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



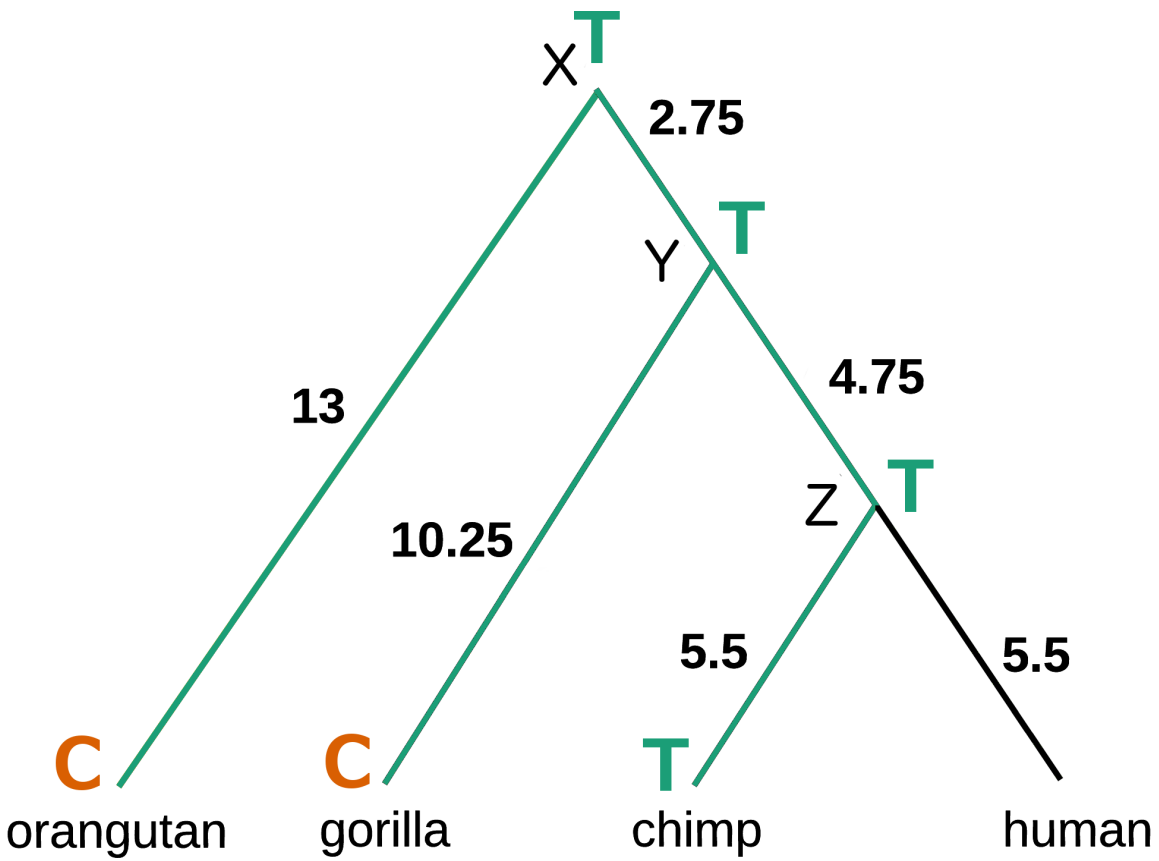
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



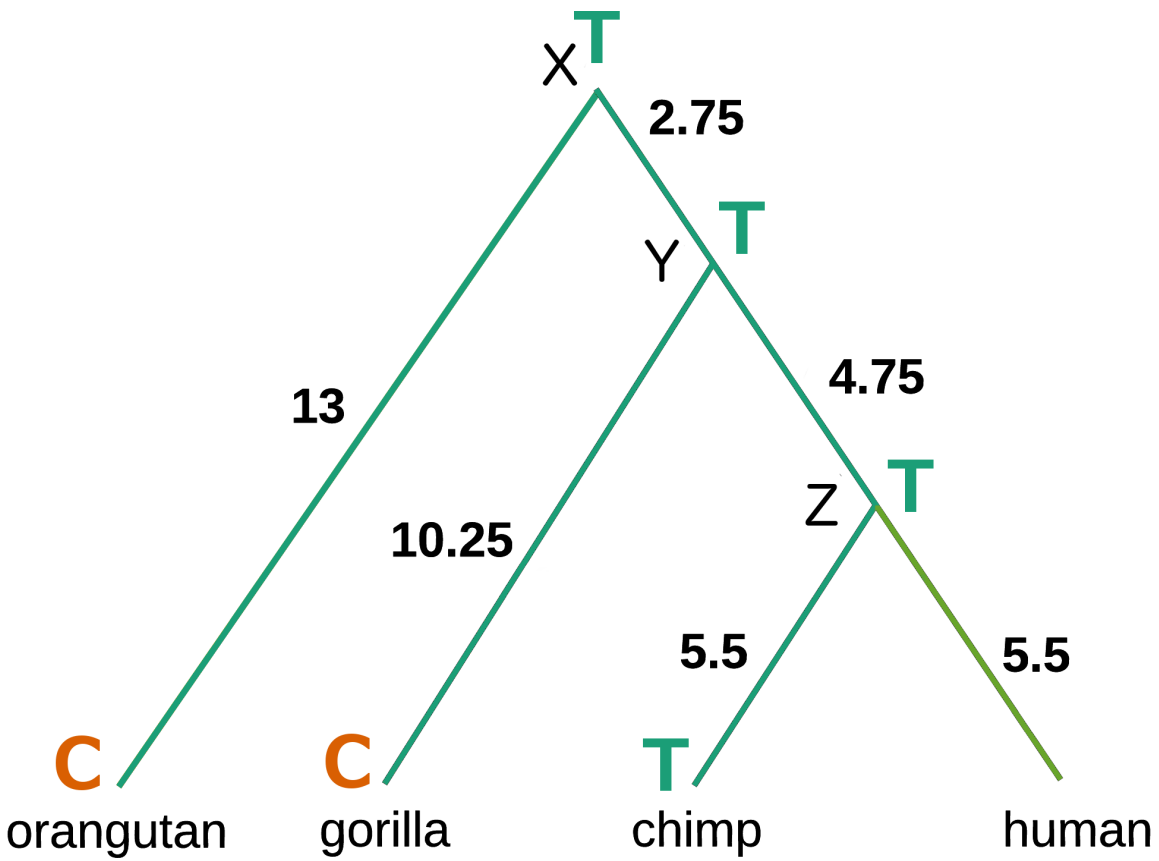
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



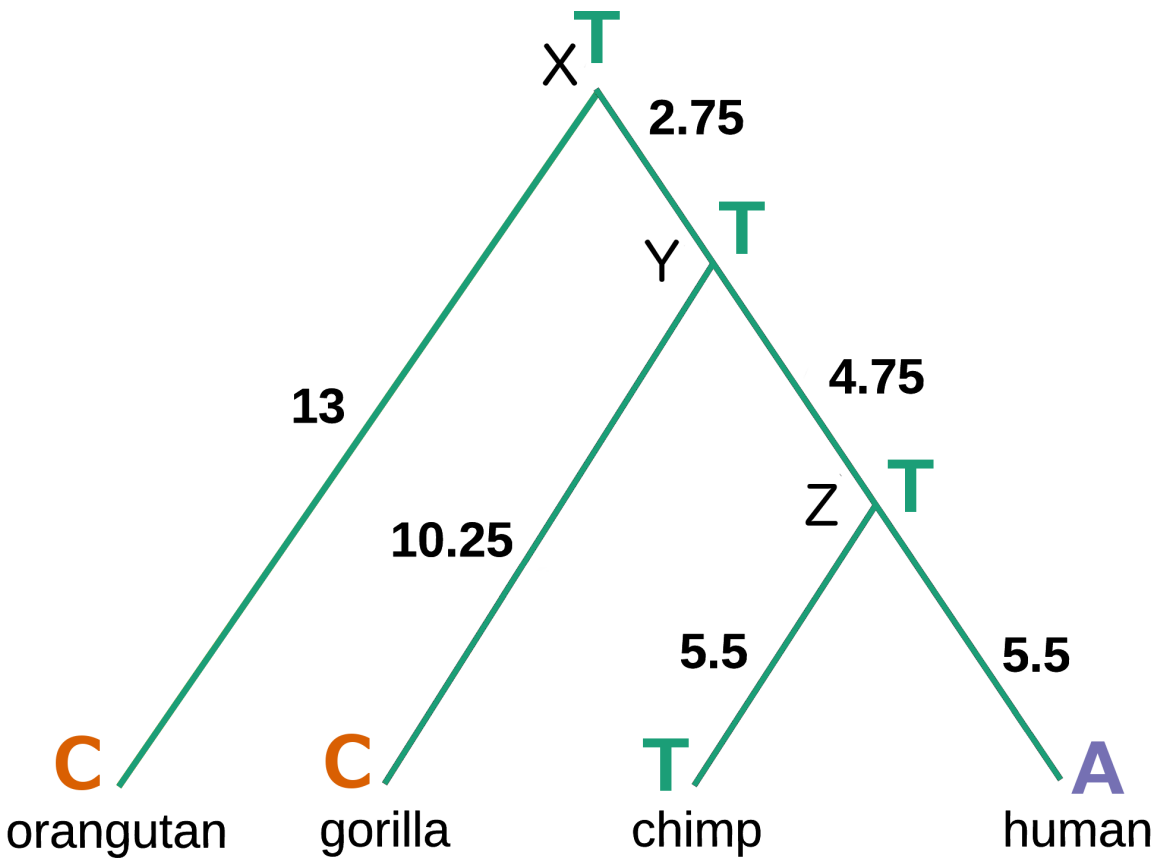
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



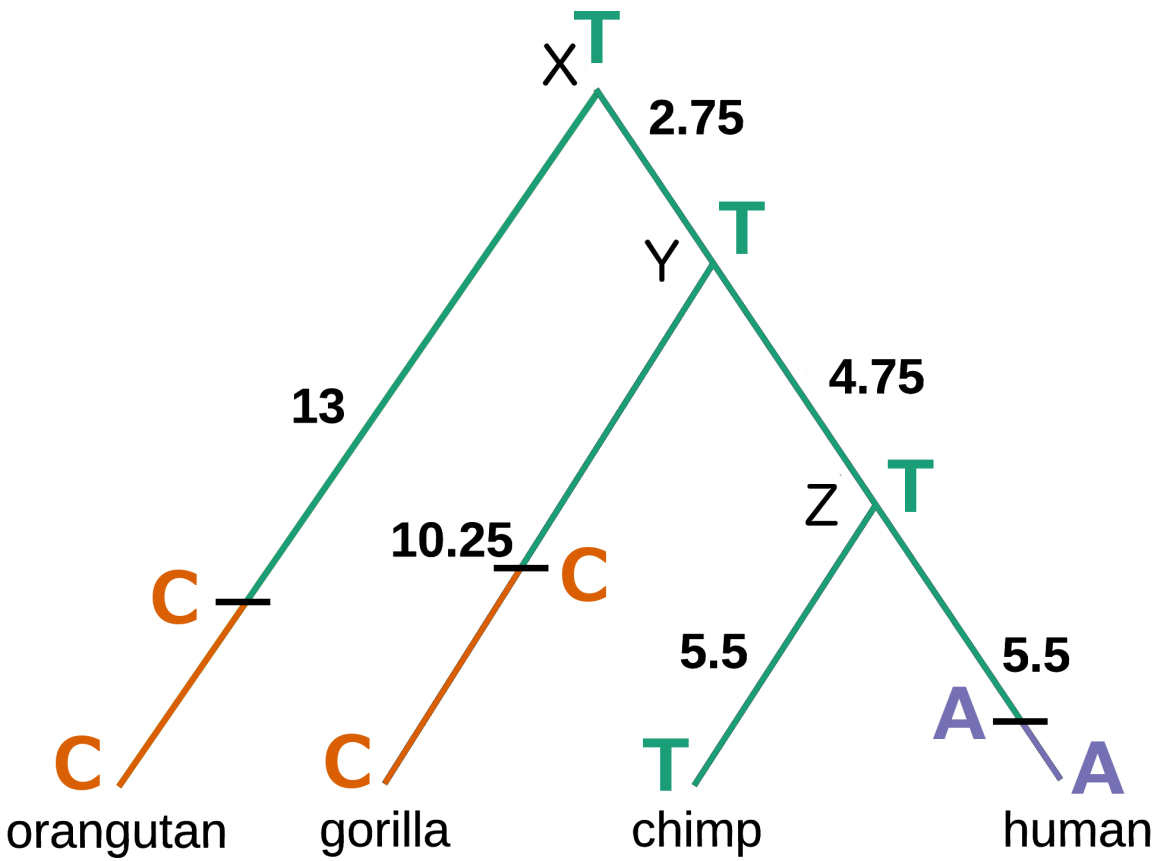
The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Evolution



The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

Exercise for today

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

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

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References

N = number of sites in the alignment;
Q = substitution rate matrix;
for $\mathrm{i} = 1$ **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;
    $\mathrm{t_b} = \mathrm{length}(\mathrm{b})$;
    $\mathrm{P}(\mathrm{t_b}) = e^{\mathrm{Q}\mathrm{t_b}}$;
    for $\mathrm{i} = 1$ **to** *N* **do**
        n = nucleotide at position i at the start of branch b;
        Sample new nucleotide n_{new} from row n in $\mathrm{P}(\mathrm{t_b})$;
        Place n_{new} at the end of sequences in the daughter branches of b;
    end
end

References I

The Simulation Game
Studying evolution
Simulating evolution
Initializing the starting sequence
Simulating the substitutions
Pen and paper exercise
Algorithm
References