Session 1

Tree basics

Email = oscar.lao@cnag.crg.eu
Skype = oscar.lao

Before we start

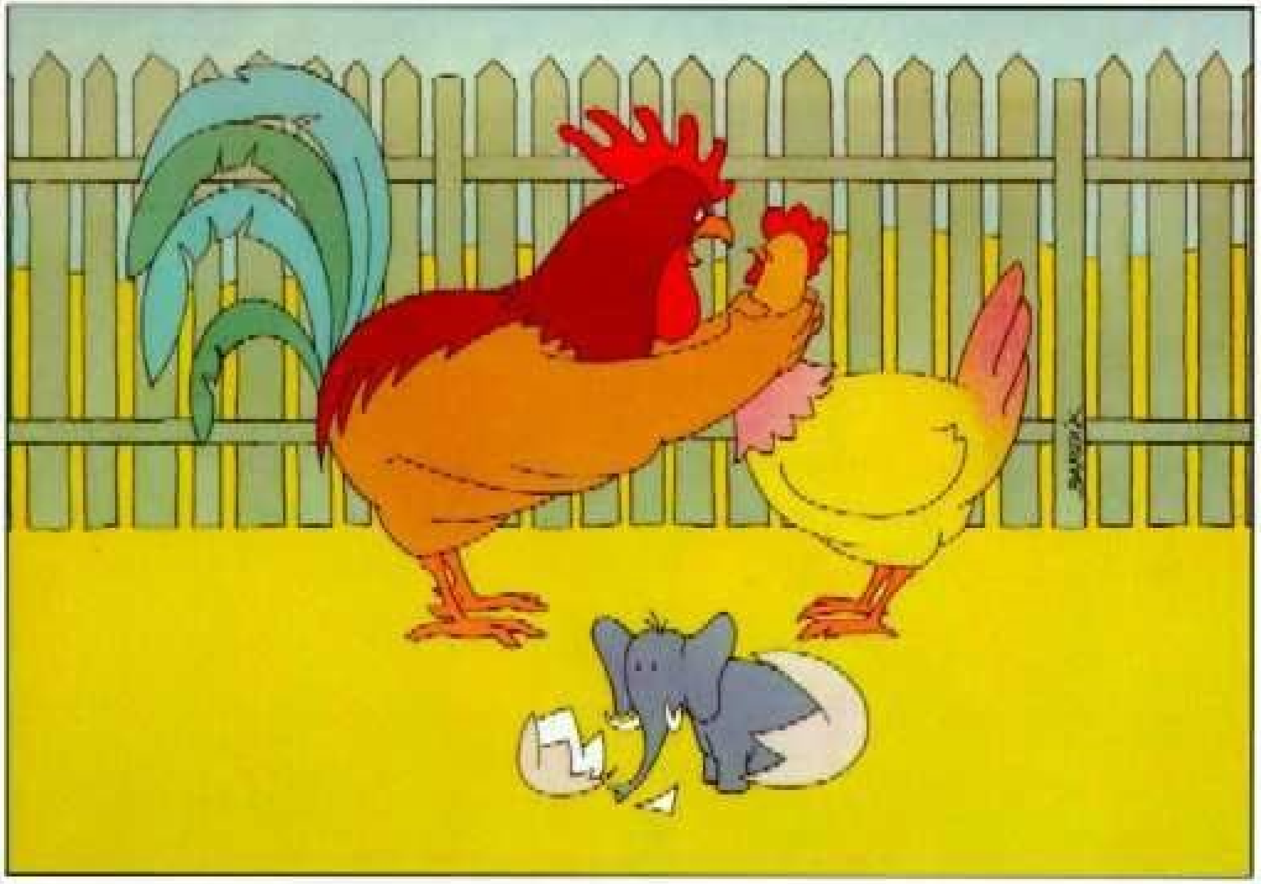
- Some basic rules
  - Please, if you want to contact me, use the oscar.lao@ibe.upf-clisc.es email (I also check the ESCI one, but less).
- Classroom dynamic
  - Each session is divided in two parts of ~55 minutes.
  - Two hours of theory. Two hours of practical (bring your own laptop!)
- Assessments

- Project shared with ASAB
  - To be done in groups of ~four people.
- Exams
  - Midterm and final exam

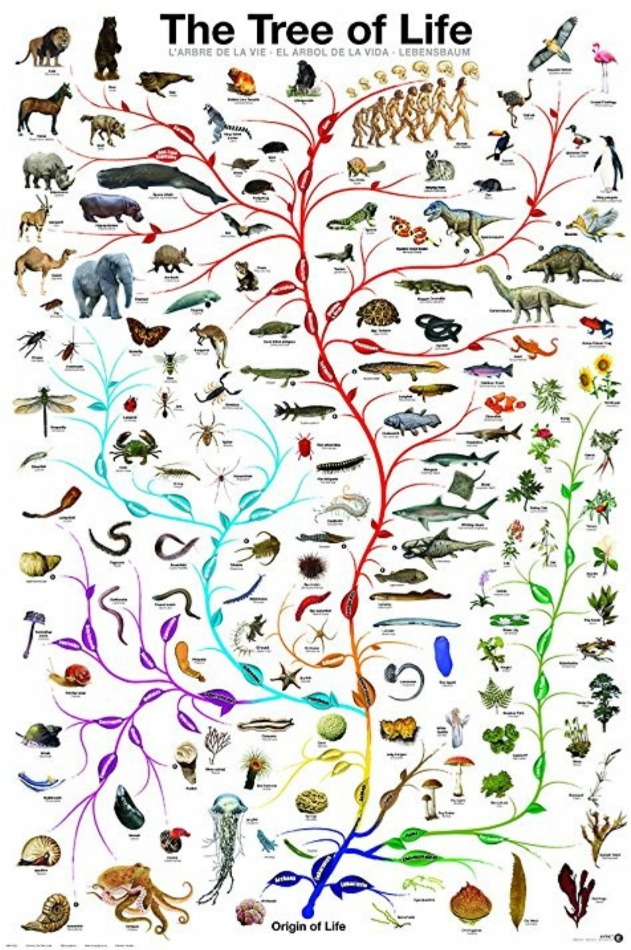
Before we start

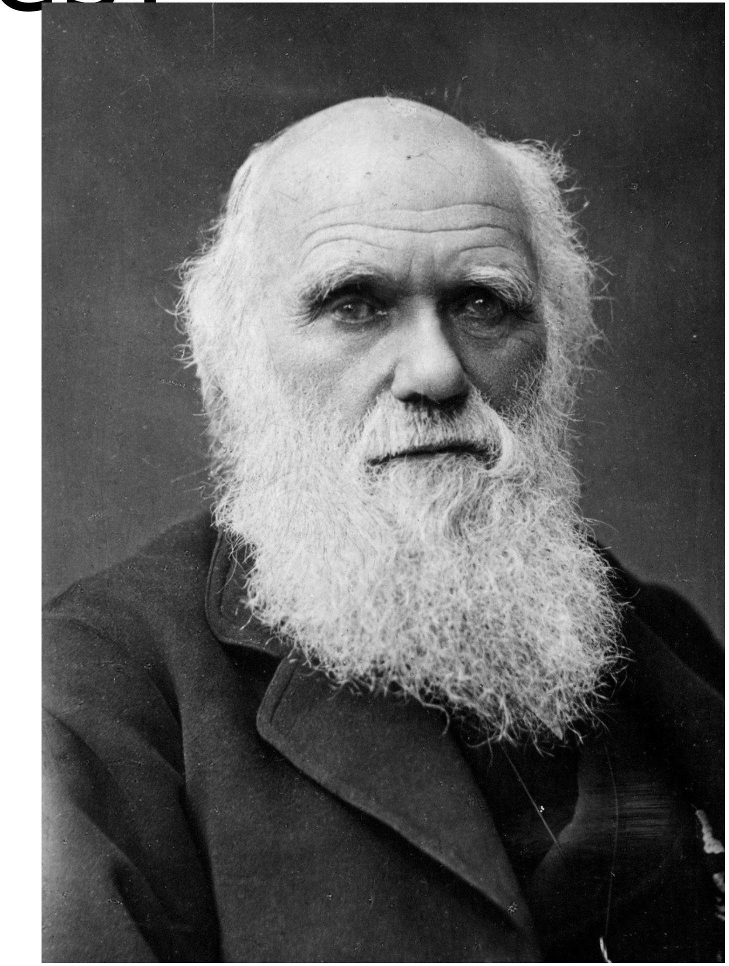
- Weekly Assignments are individual, NOT BY PAIRS.
- Plagiarism
- Python
- Comment your code

WHY IS THIS A JOKE?

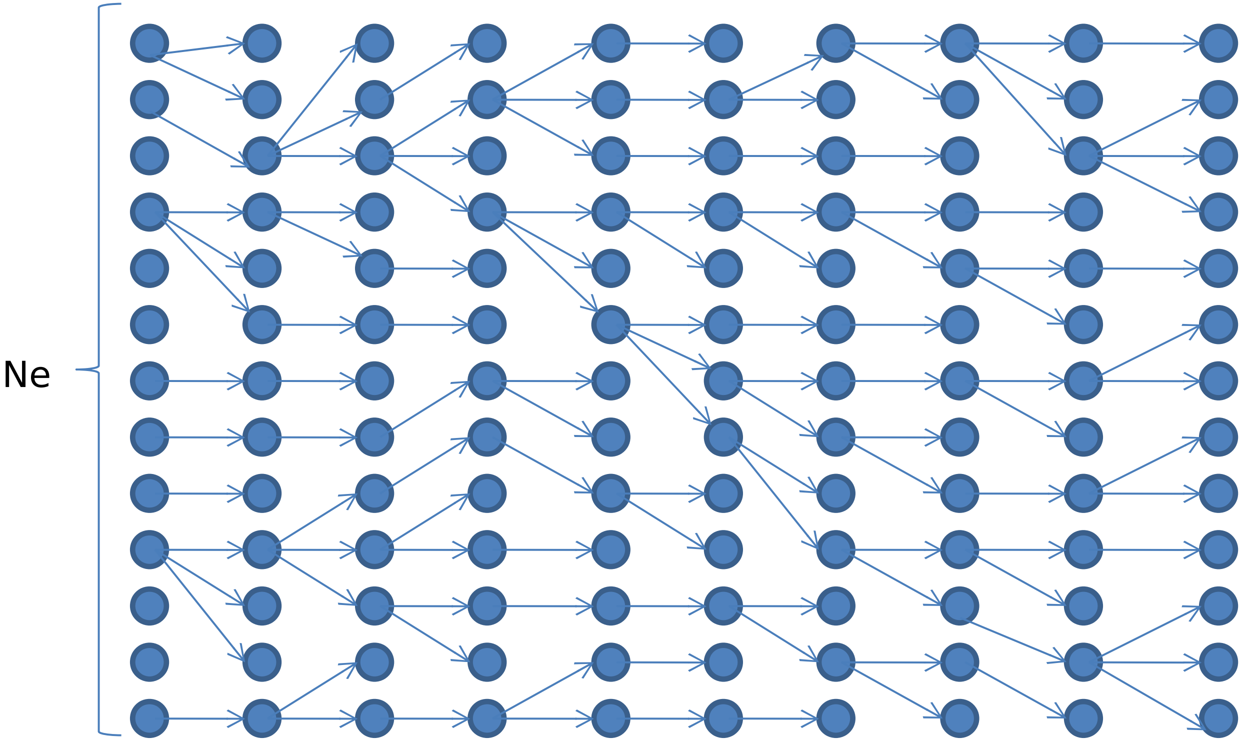


Why do we have to think in trees?

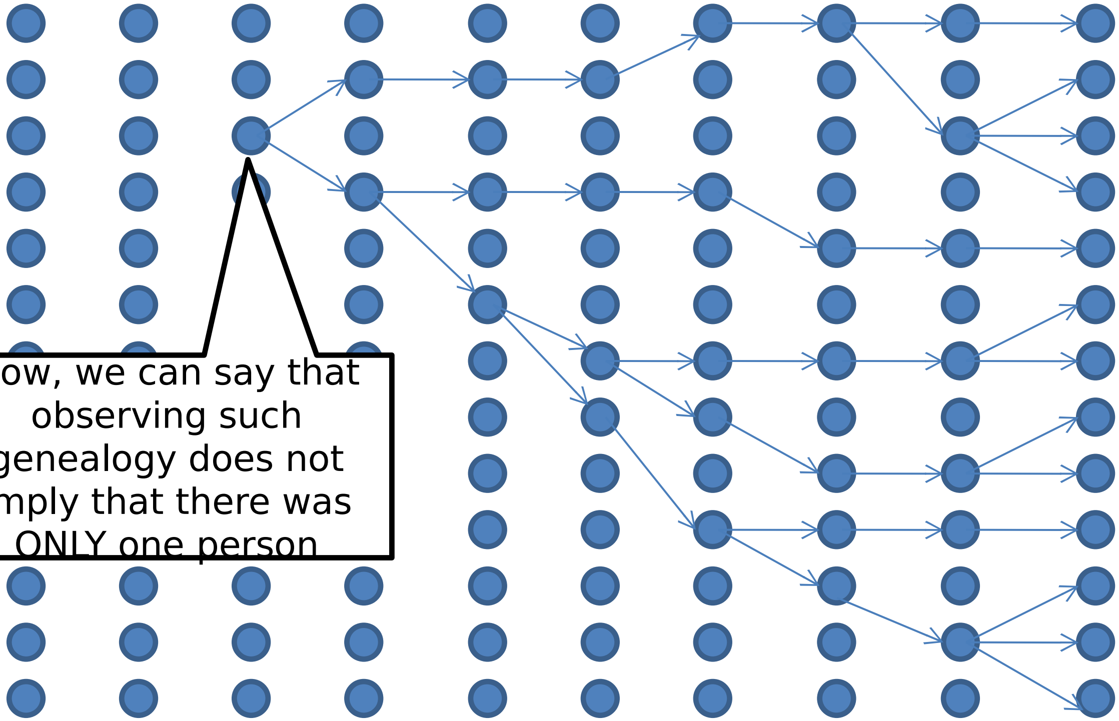




Why do we have to think in trees?

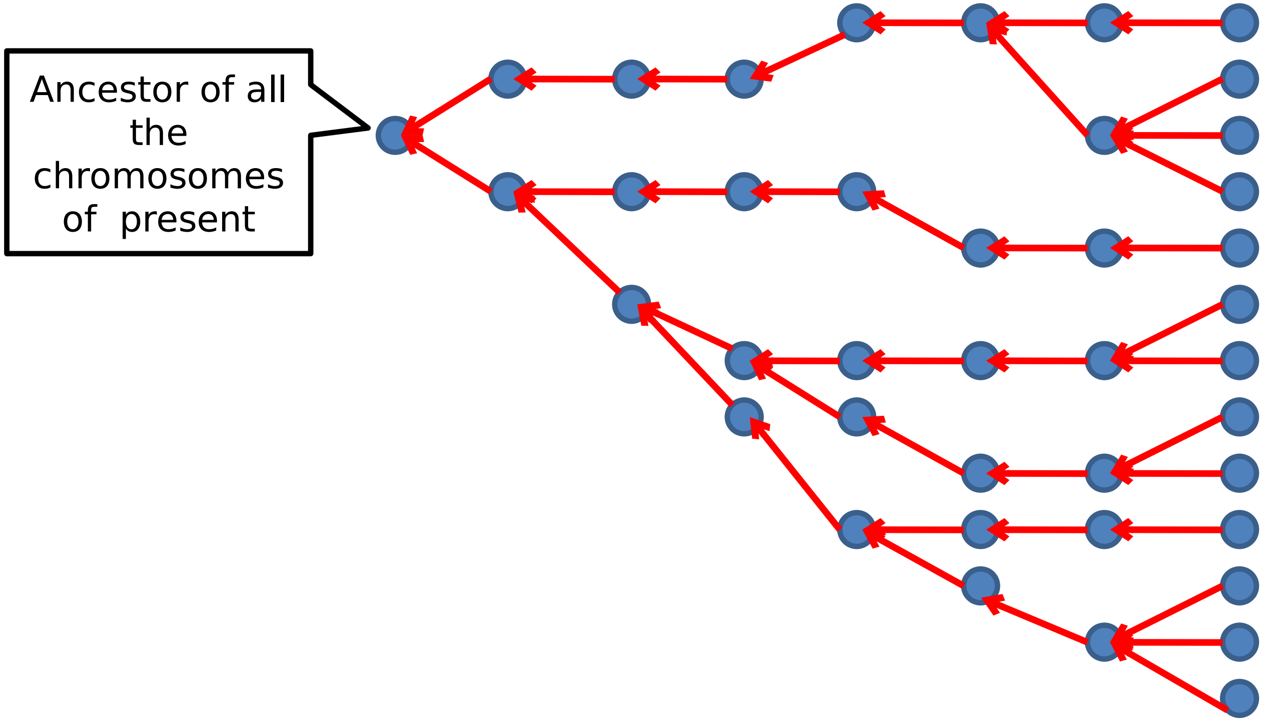


Why do we have to think in trees?



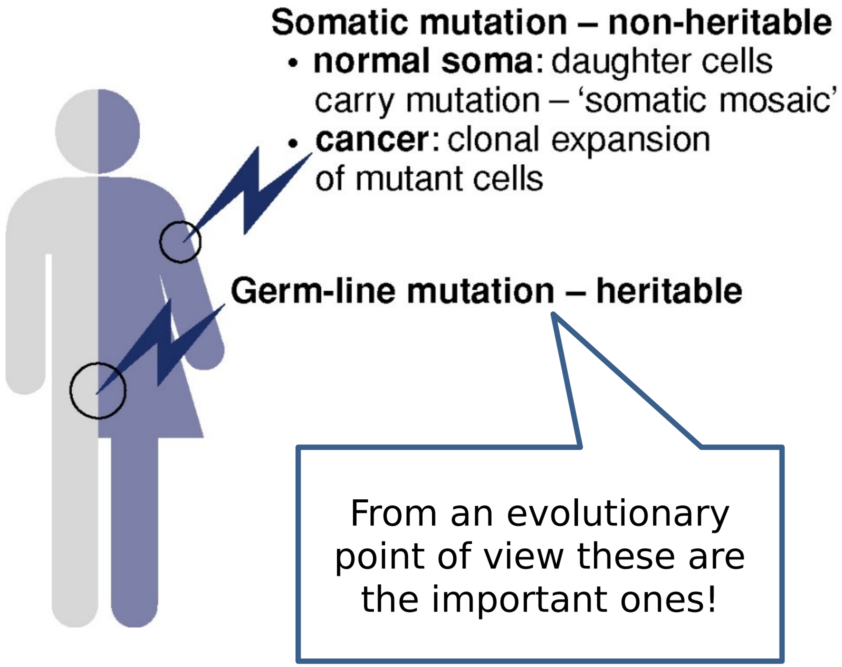
Now, we can say that observing such genealogy does not imply that there was ONLY one person

Why do we have to think in trees?



Ancestor of all the chromosomes of present

Why do we have to think in trees?



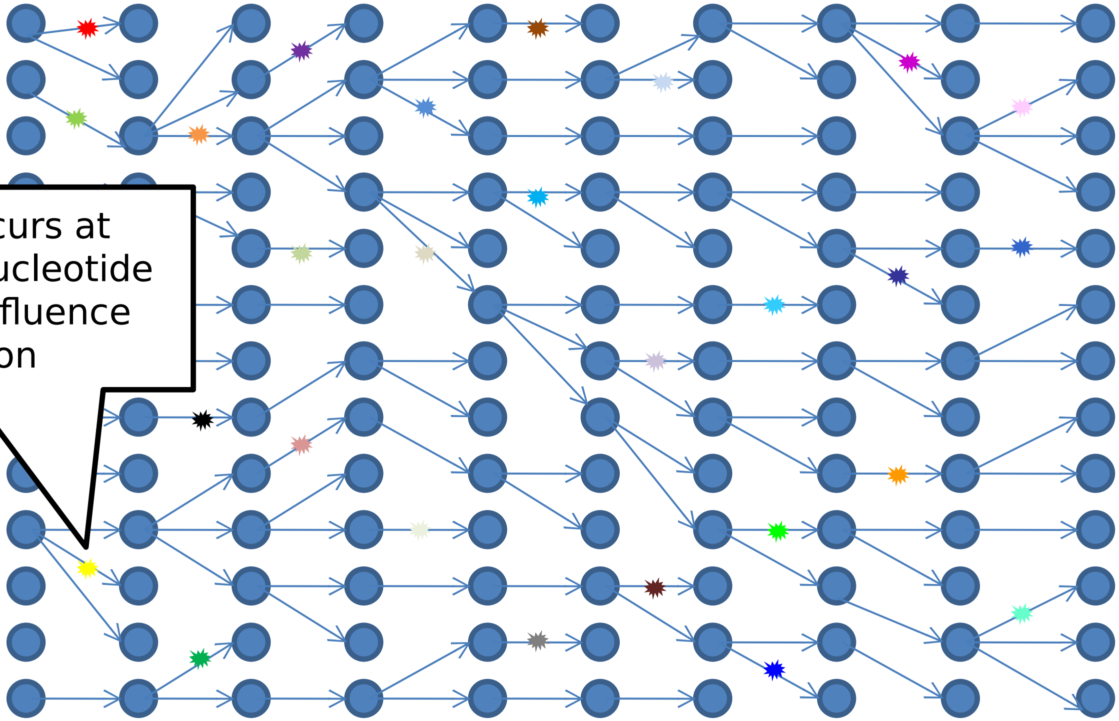
Somatic mutation – non-heritable

- **normal soma**: daughter cells carry mutation – 'somatic mosaic'
- **cancer**: clonal expansion of mutant cells

Germ-line mutation – heritable

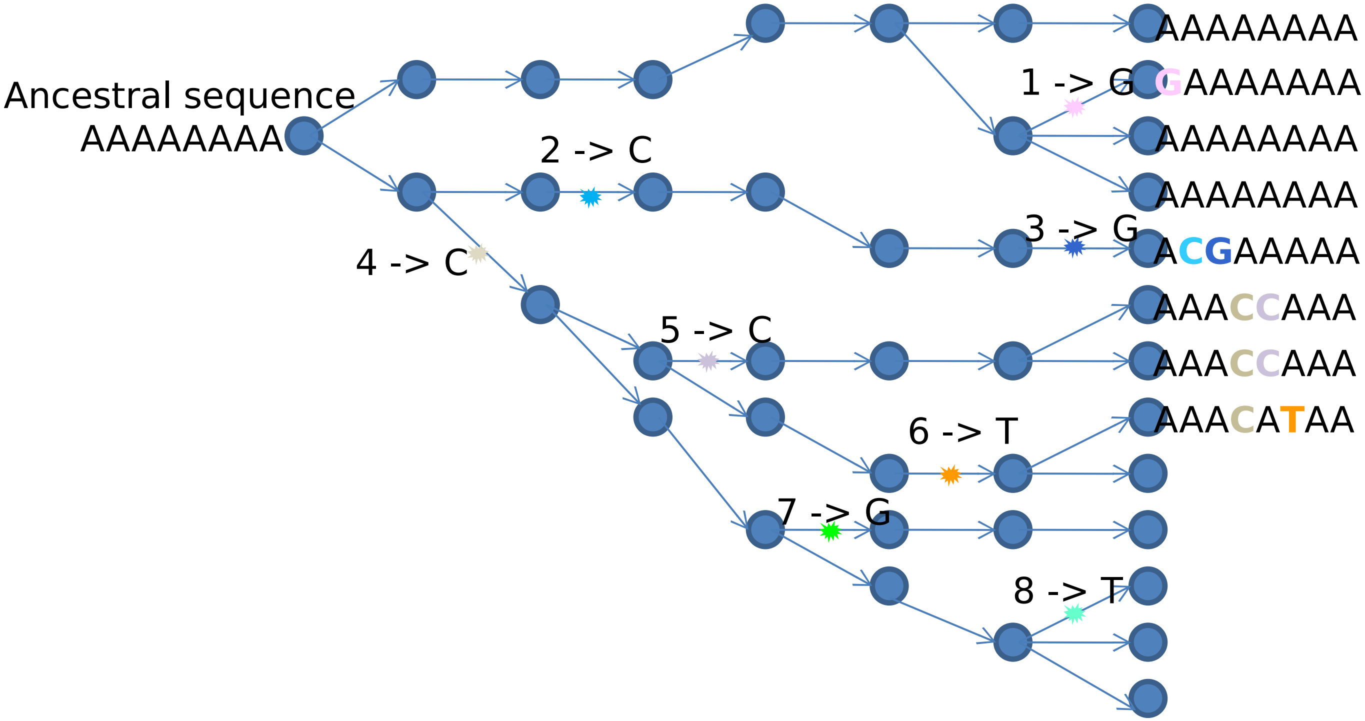
From an evolutionary point of view these are the important ones!

Why do we have to think in trees?

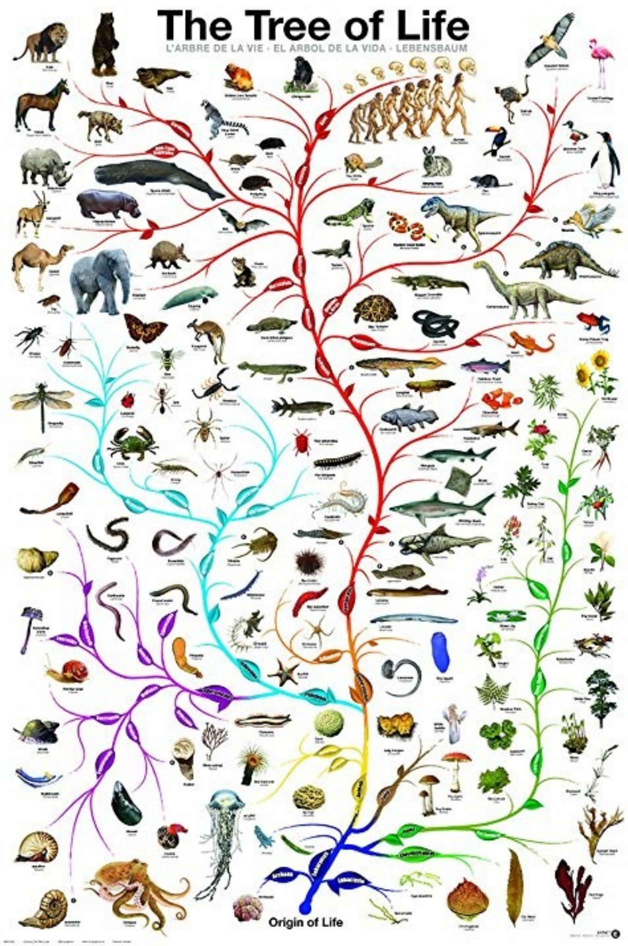


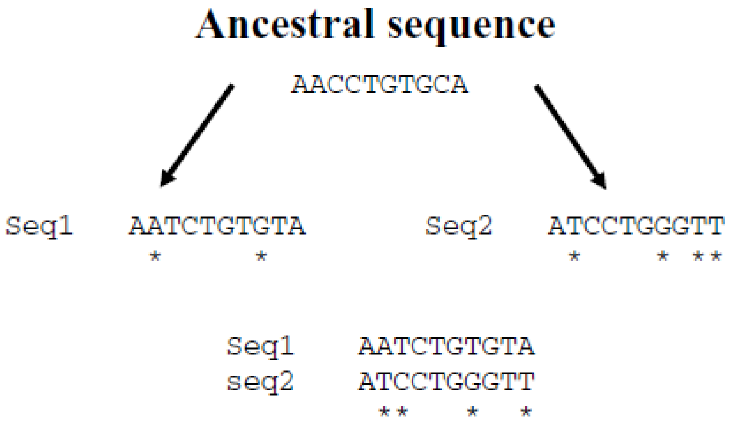
A mutation occurs at random in one nucleotide that does not influence reproduction

Why do we have to think in trees?



Why do we have to think in trees?



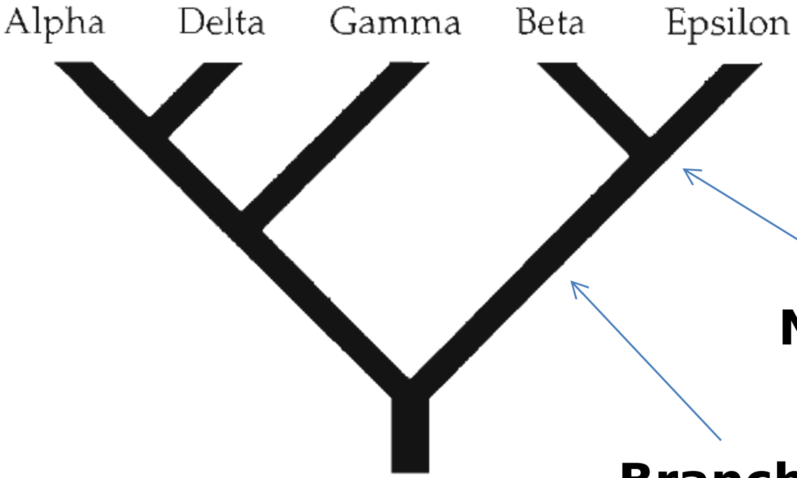

Ancestral sequence

```
                 AACCTGTGCA

Seq1   AATCTGTGTA            Seq2   ATCCTGGGTT
        *     *                      *    *  **

             Seq1   AATCTGTGTA
             seq2   ATCCTGGGTT
                     **   *  *
```

P-distance = 4

What is a tree?

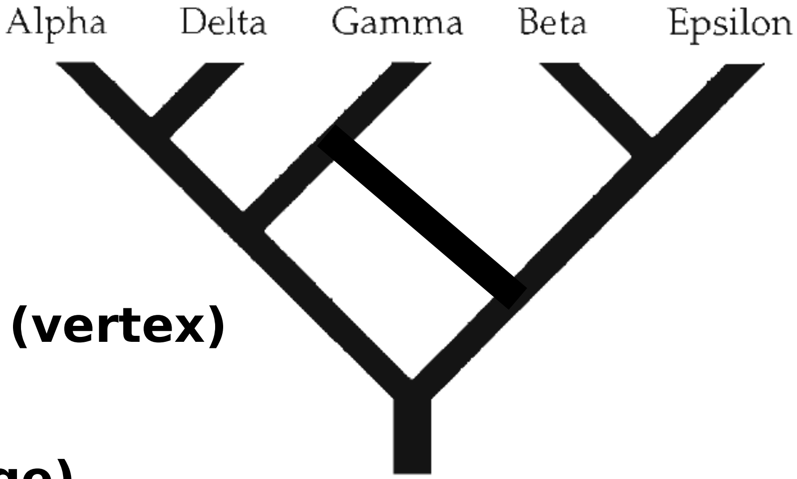
Leaf (vertex)



Node (vertex)

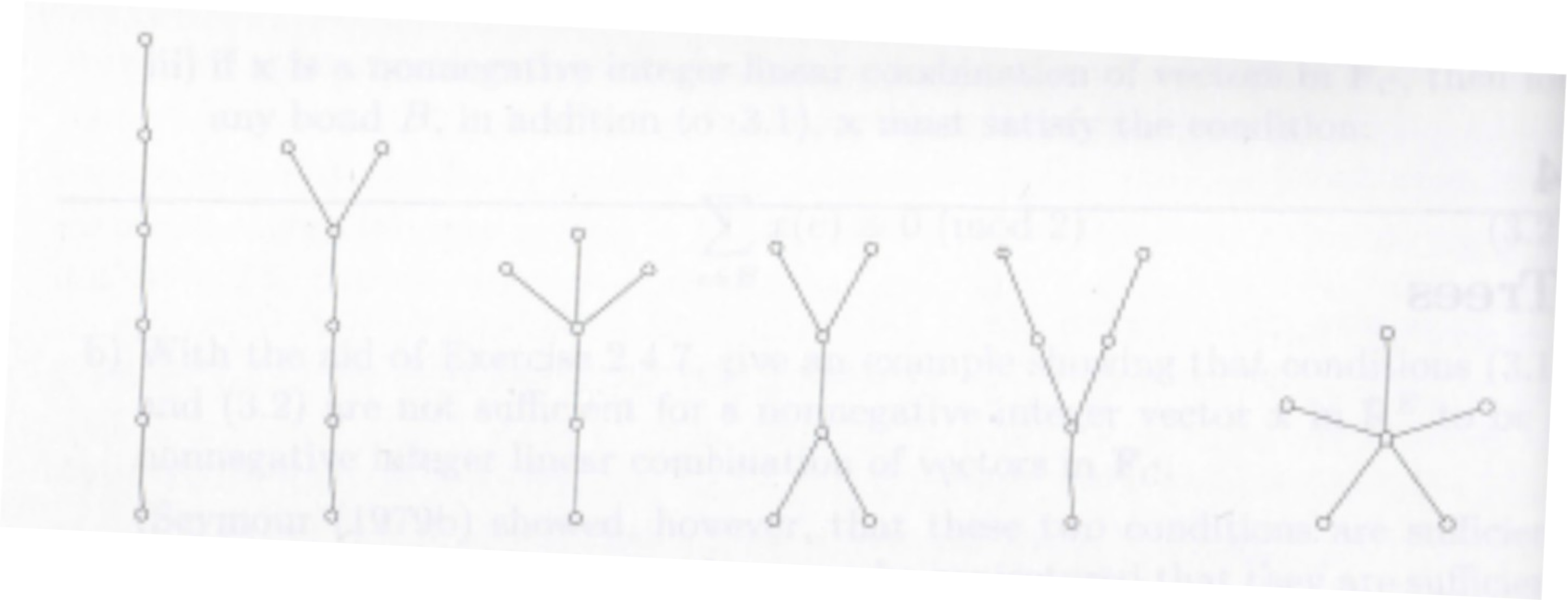
Branch (edge)

Root



Is this a tree? Why?

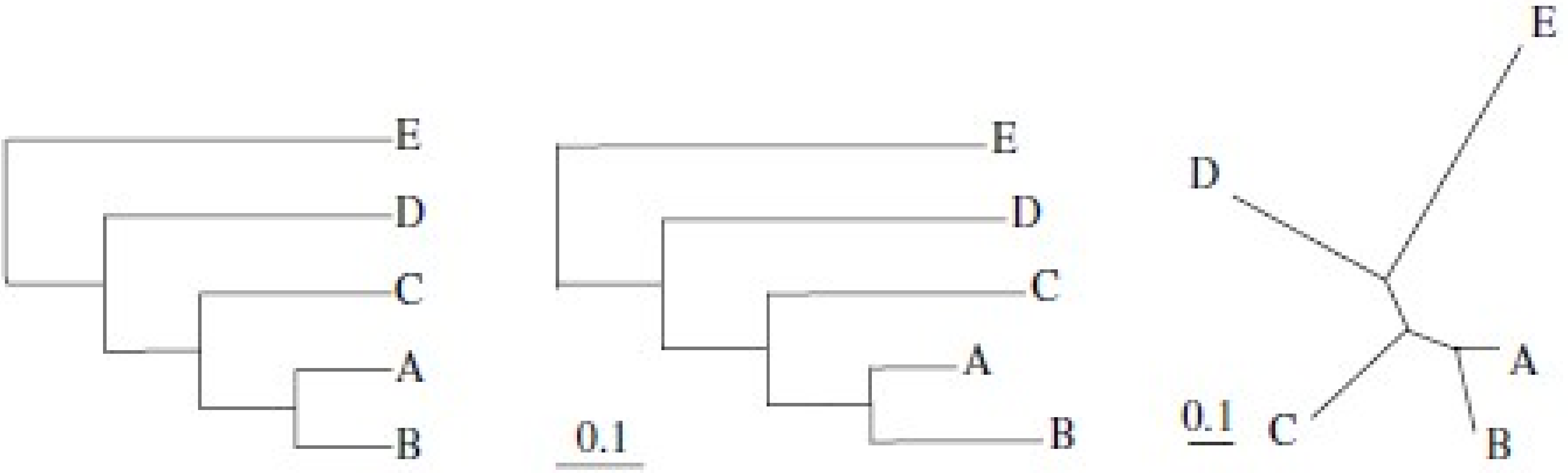
What is a tree?



What is a tree

- Type of graph
- Connected acyclical graph
  - Leaf: vertex of degree one
- Depicts the relationship between Operational Taxonomic Unit (OTUs)

Ways of describing a tree



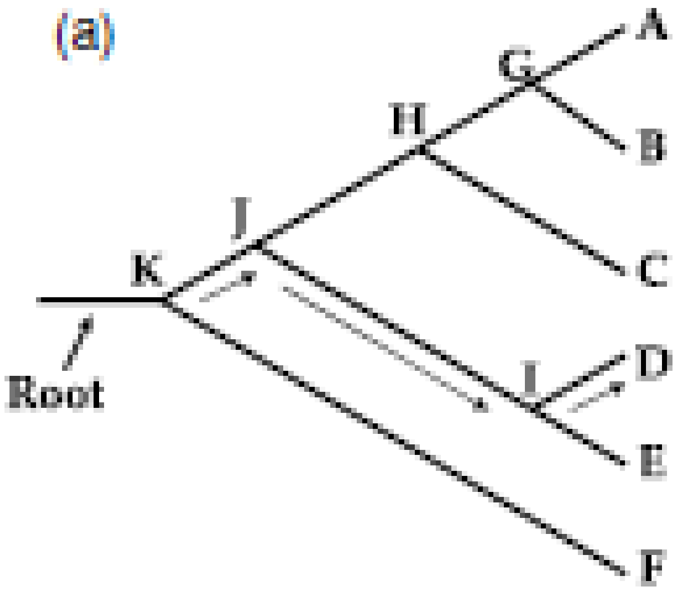
(a) Cladogram

(b) Phylogram

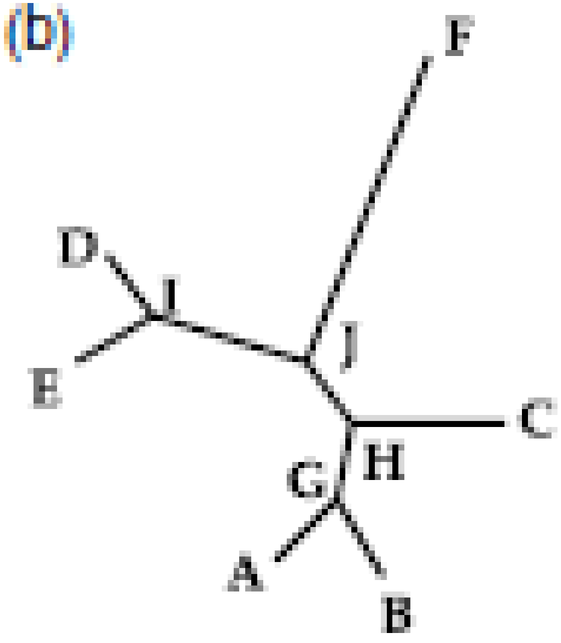
(c) Unrooted tree

Ways of describing a tree

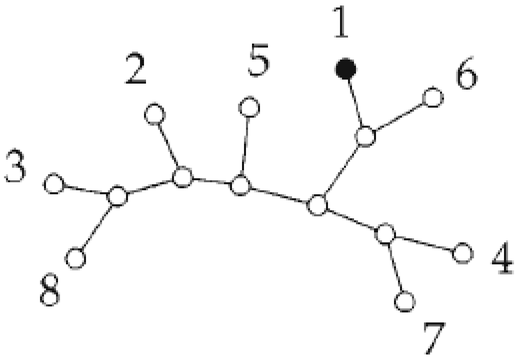
Rooted



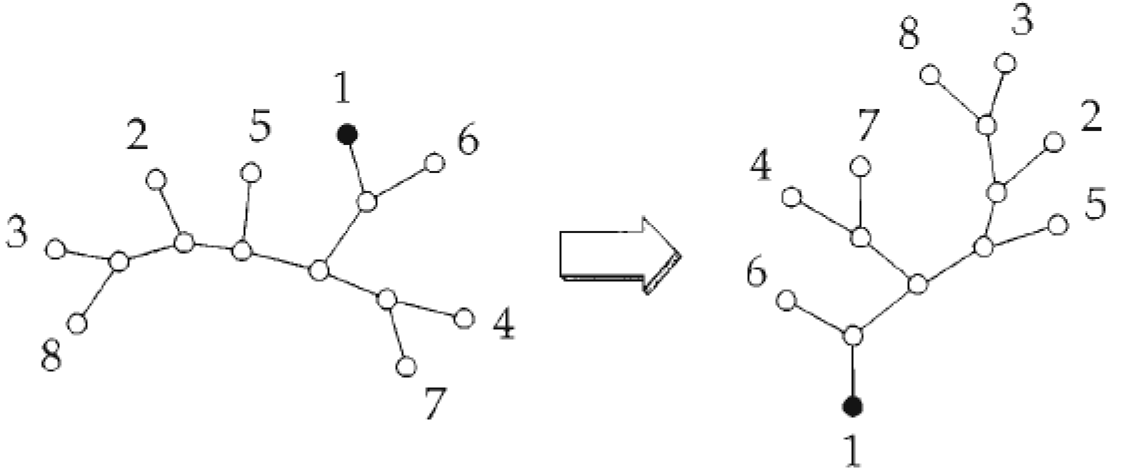
Unrooted



How do we root a tree?



How do we root a tree?

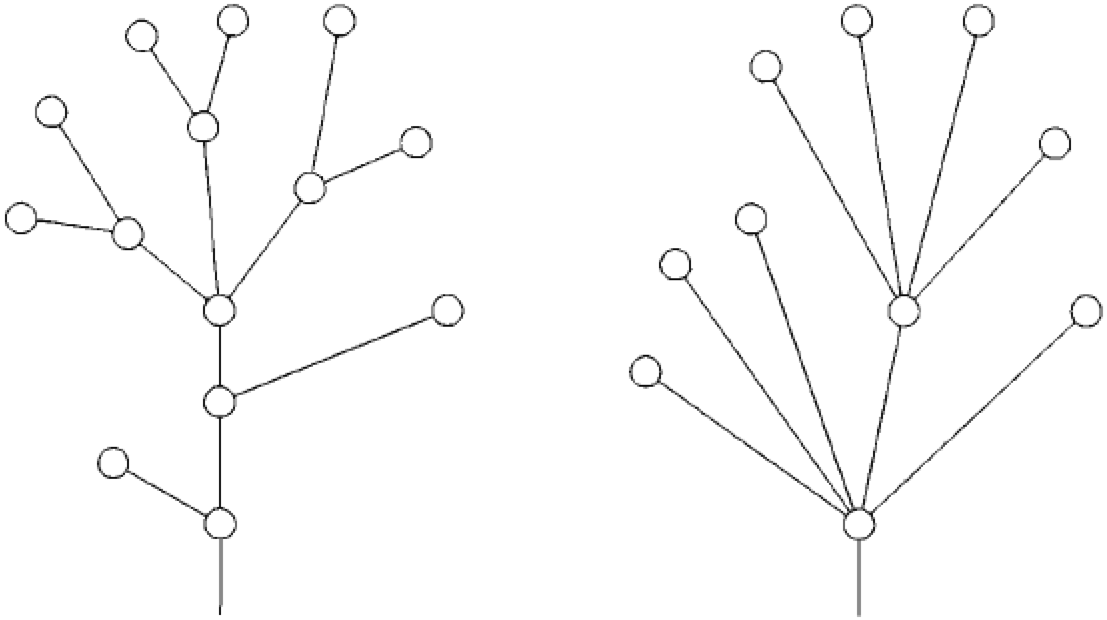


In the context of graph theory, what is a root?

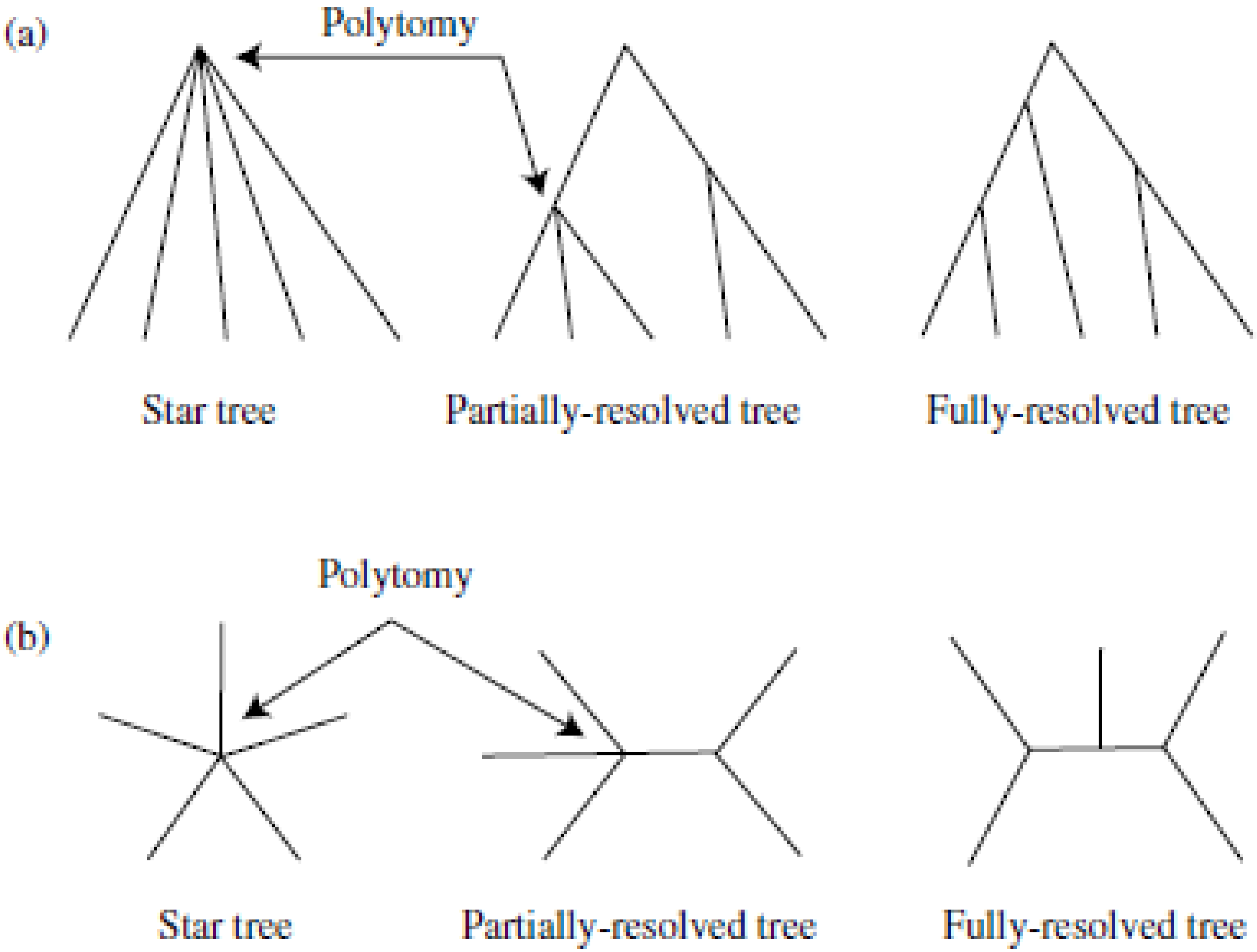
-

In the context of a tree, what does a root mean?

Bifurcated vs multifurcated

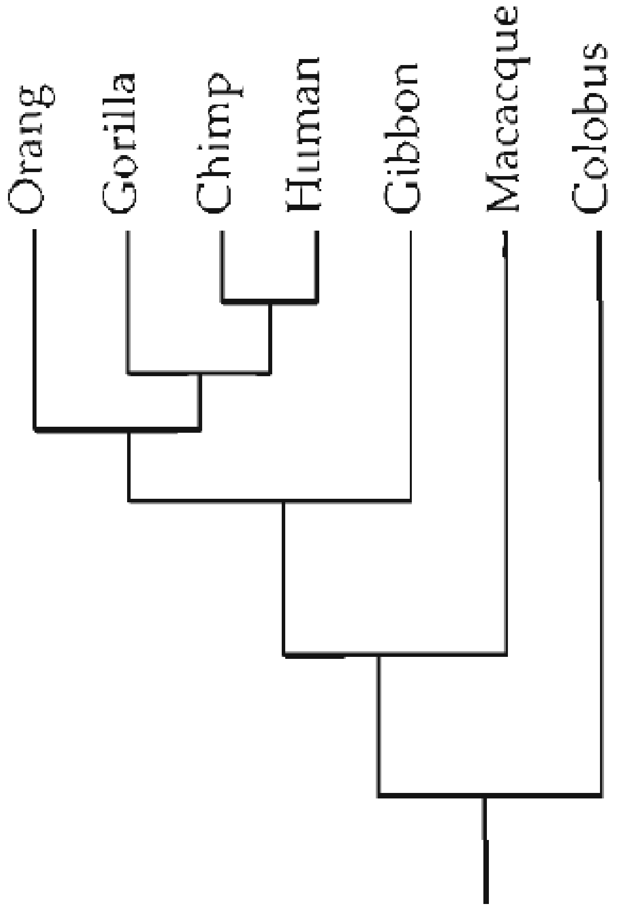


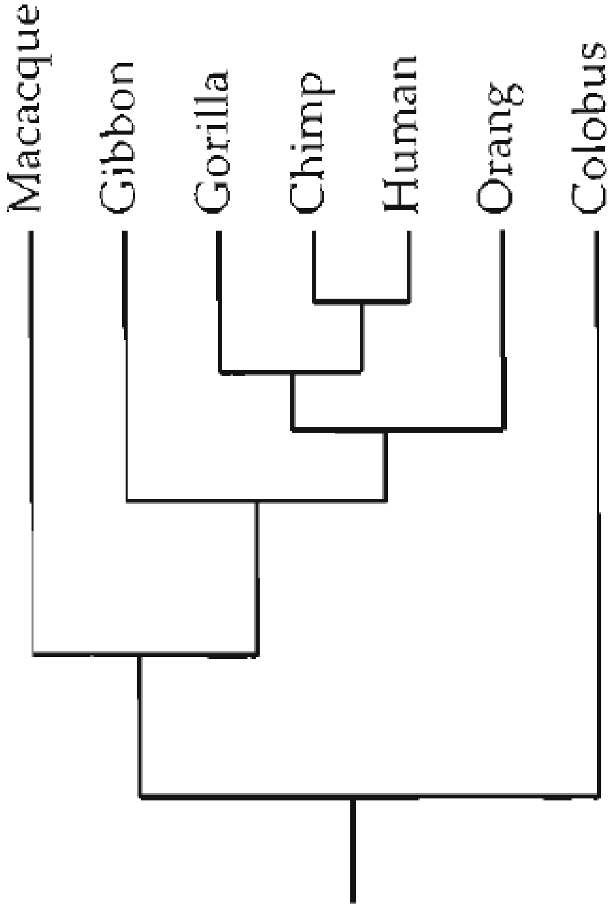
Resolved vs unresolved



How do we interpret a tree?

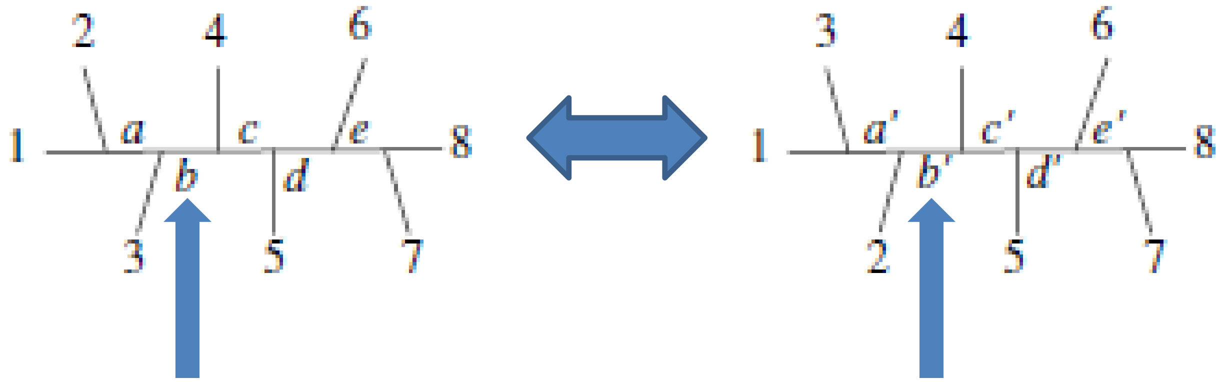
- Tree topology





How do we interpret a tree?

- Distance between trees



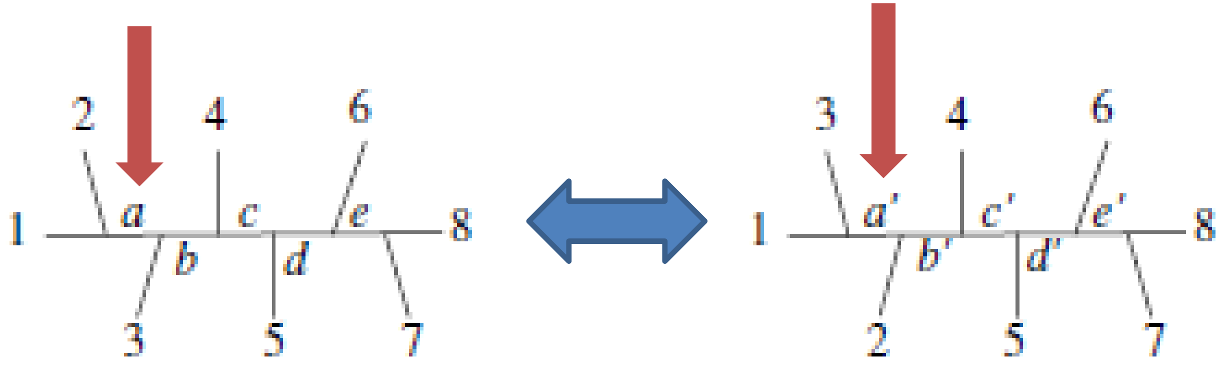
(1,2,3),(4,5,6,7,8)

(1,2,3),(4,5,6,7,8)

total number of bipartitions that are in one tree but not in the other

How do we interpret a tree?

- Distance between trees

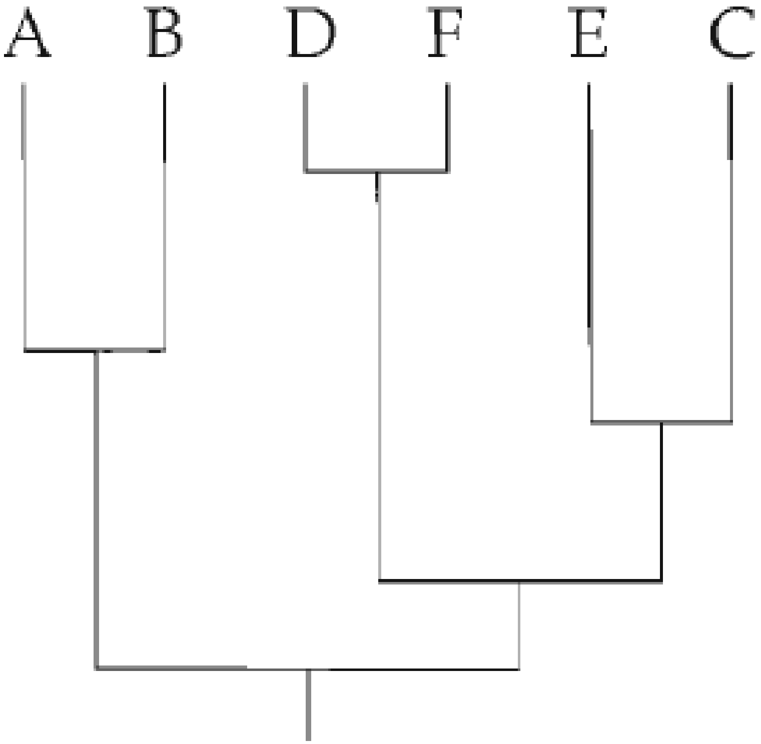


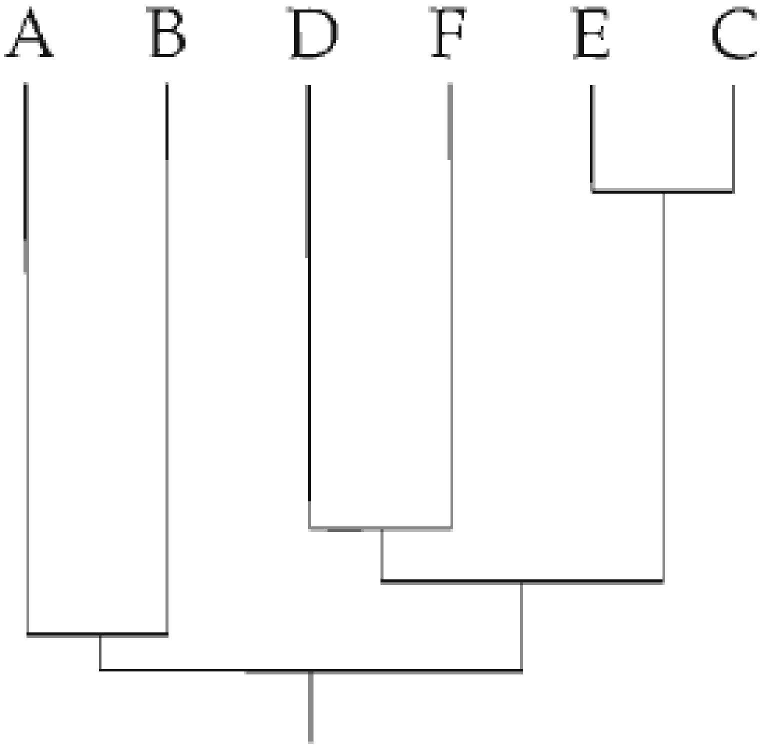
(1,2),(3,4,5,6,7,8)

(1,3),(2,4,5,6,7,8)

How do we interpret a tree?

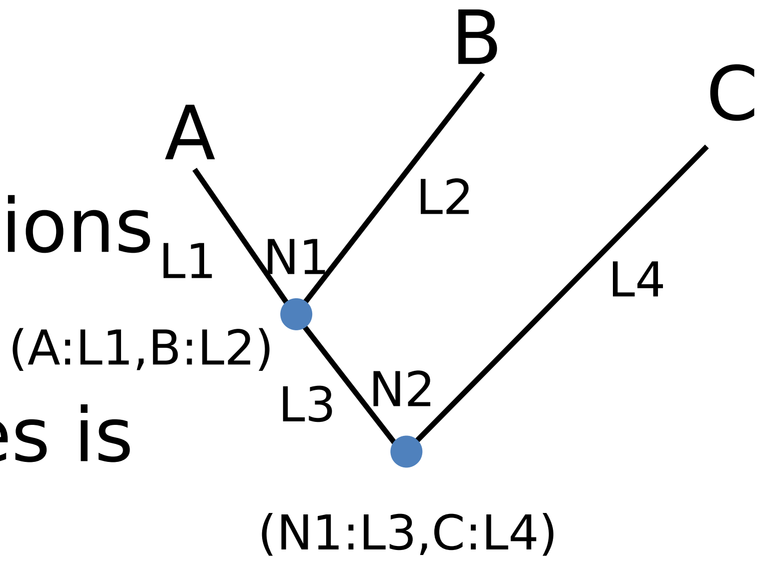
- Branch length





Ways of describing a tree

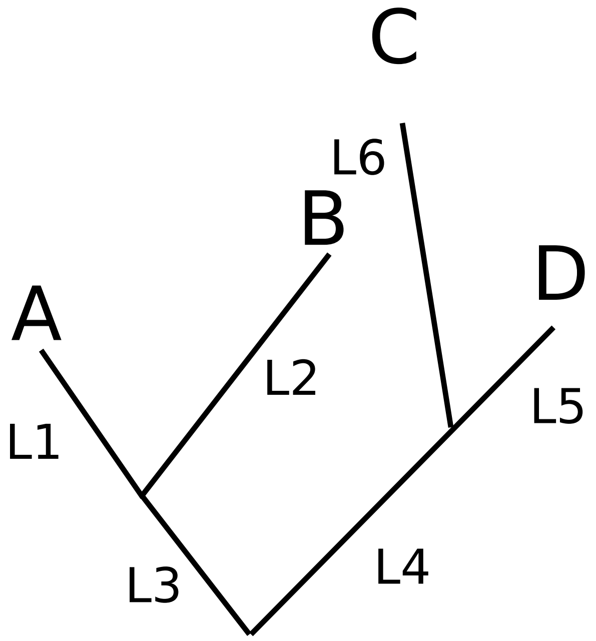
- Newick format
  - Each internal node is defined by the connections (,)
  - Length between nodes is defined by :



((A:L1,B:L2):L3,C:L4)

(C:L4,(A:L1,B:L2):L3)

Ways of describing a tree



Ways of describing a tree

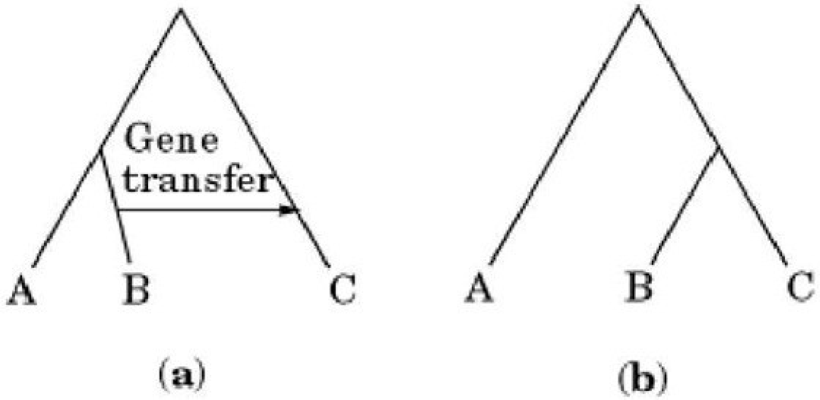
(A:L1,((C:L6,B:L3):L2,D:L5):L4)

Trees vs reality

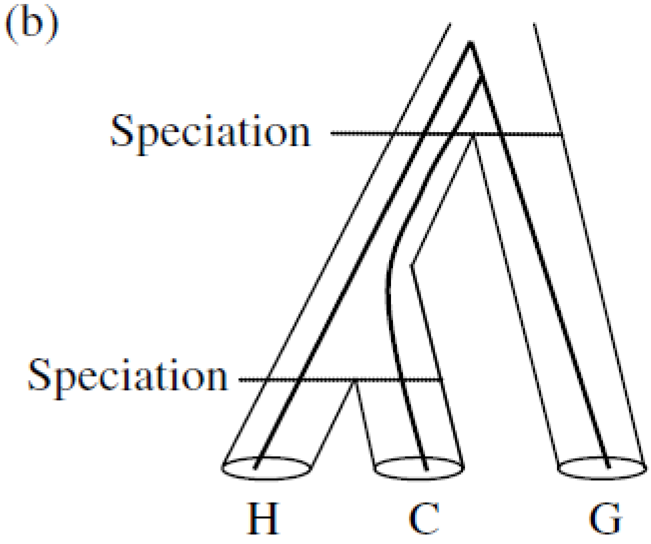
When a tree is not reflecting reality?

Trees vs reality

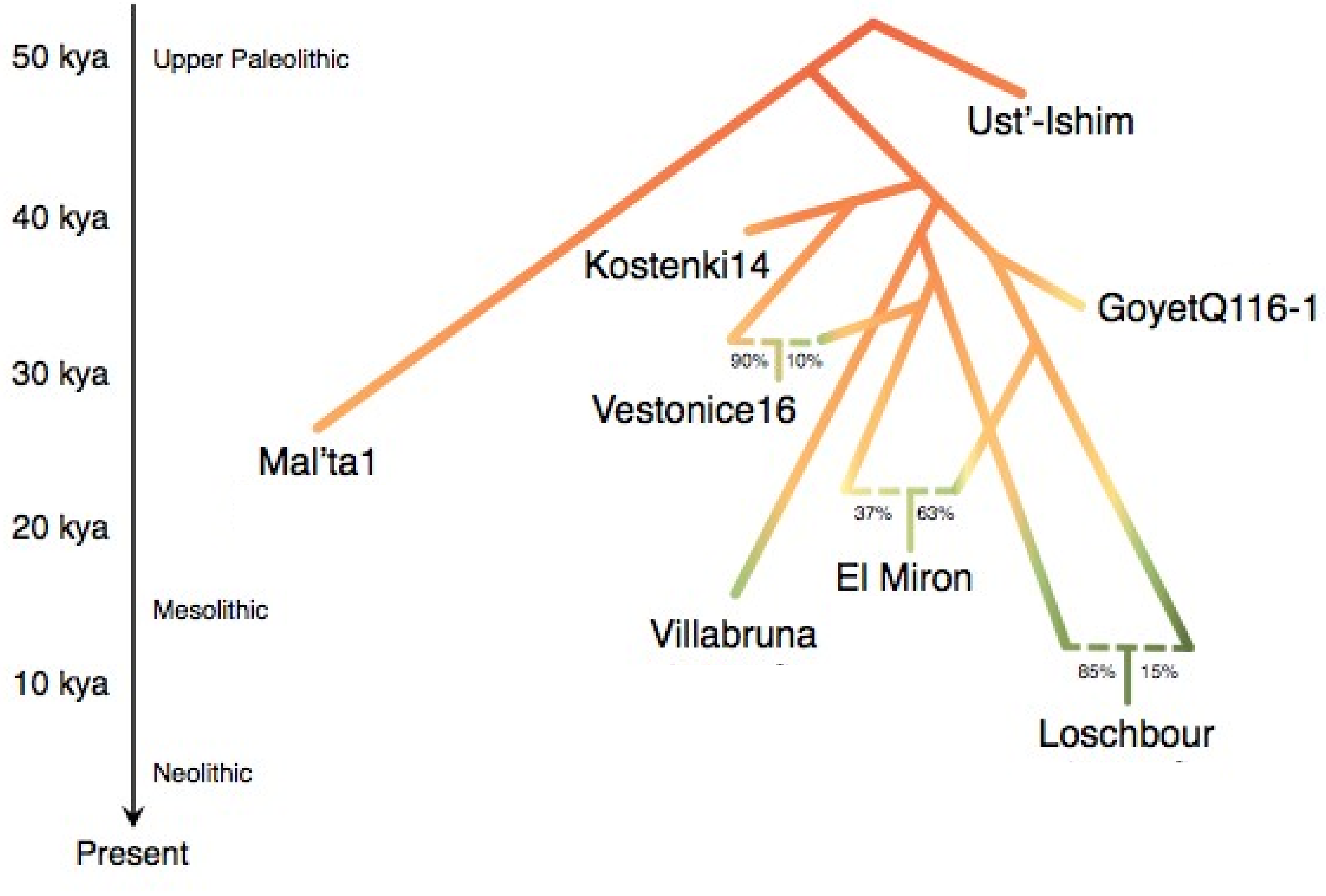
Horizontal gene transfer



Recent speciation

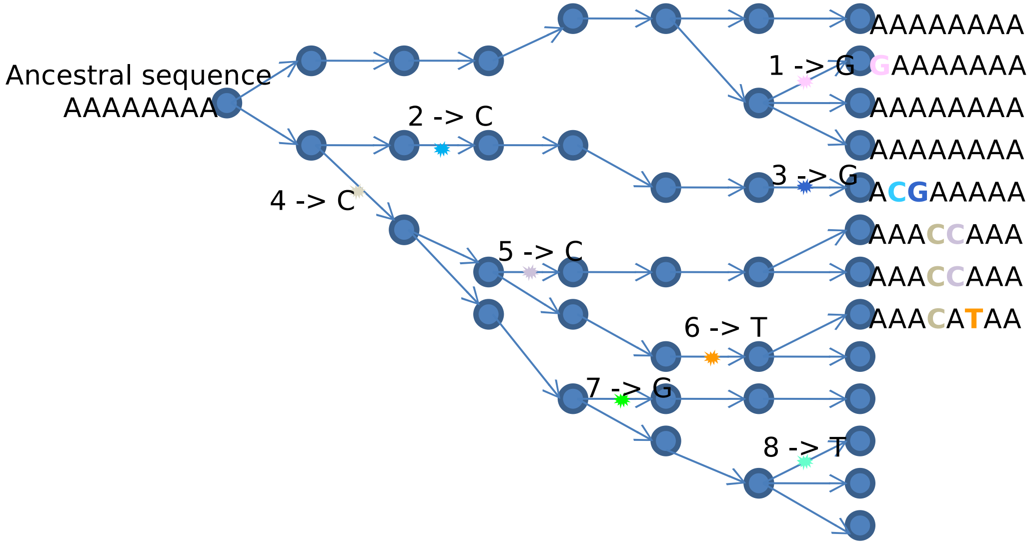


Trees vs reality

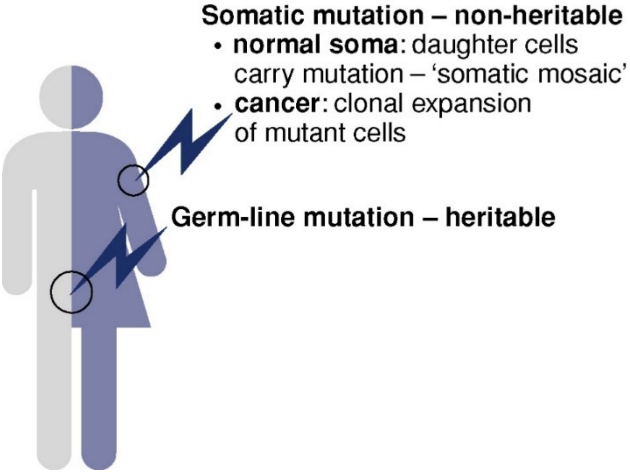


TIME FOR A BREAK

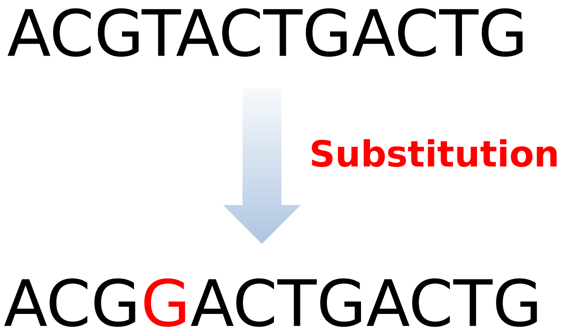
Remember: each locus observed in all current sequences comes from a single ancestor



More than just substitutions


Somatic mutation – non-heritable
• normal soma: daughter cells carry mutation – 'somatic mosaic'
• cancer: clonal expansion of mutant cells
Germ-line mutation – heritable

Human evolutionary genetics

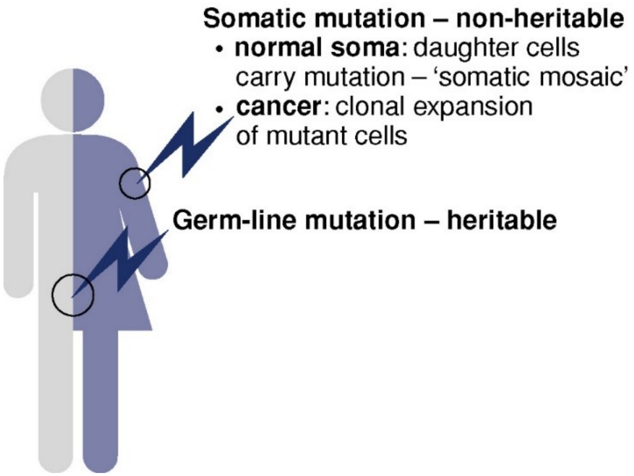


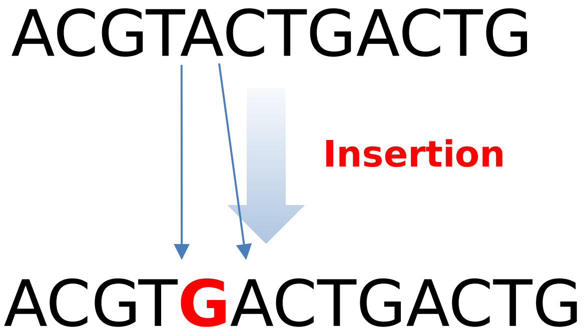
ACGTACTGACTG

Substitution

ACGGACTGACTG

More than just substitutions


Somatic mutation – non-heritable
• normal soma: daughter cells carry mutation – 'somatic mosaic'
• cancer: clonal expansion of mutant cells
Germ-line mutation – heritable

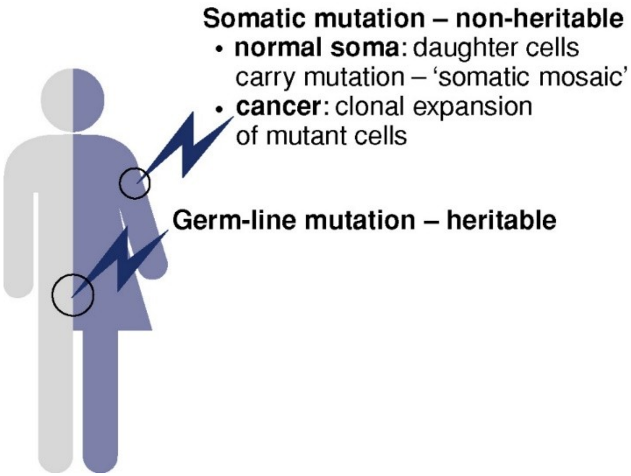


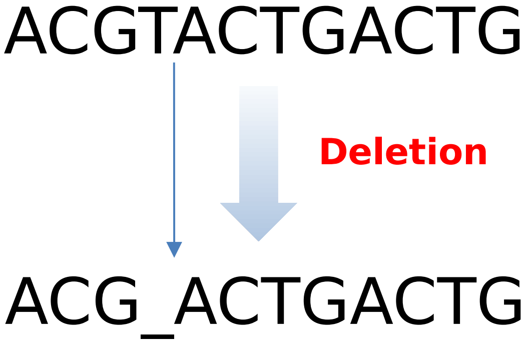
ACGTACTGACTG

Insertion

ACGT**G**ACTGACTG

More than just substitutions


Somatic mutation – non-heritable
• normal soma: daughter cells carry mutation – 'somatic mosaic'
• cancer: clonal expansion of mutant cells
Germ-line mutation – heritable

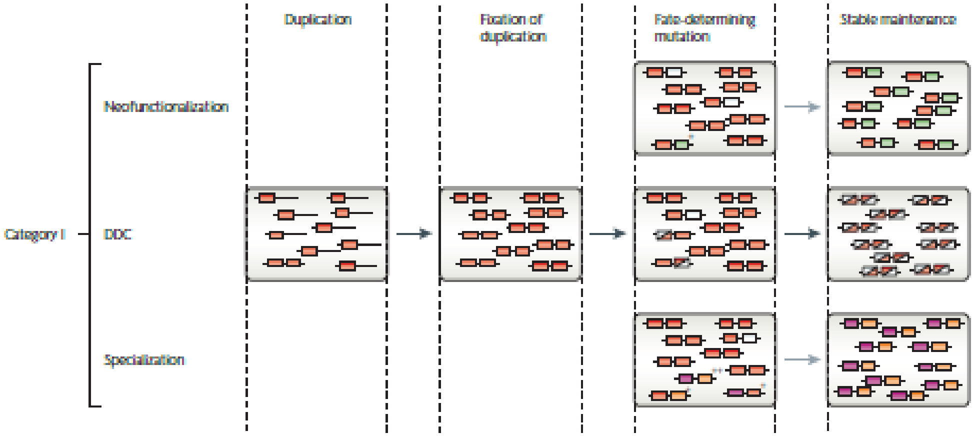


ACGTACTGACTG

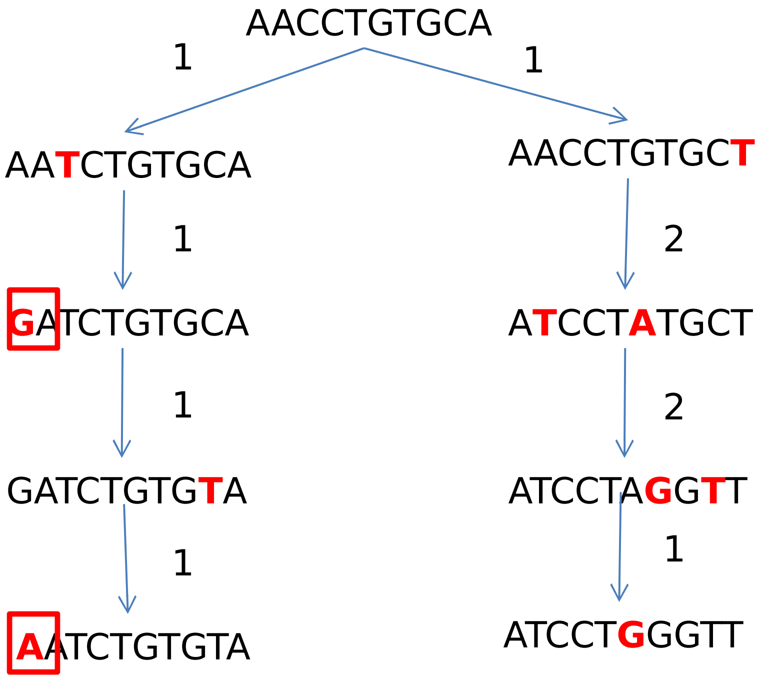
Deletion

ACG_ACTGACTG

More than just substitutions: gene duplications and deletions



Genetic divergence is a measure of **TIME** divergence



AACCTGTGCA

1

AA**T**CTGTGCA

1

GATCTGTGCA

1

GATCTGTG**T**A

1

AATCTGTGTA

1

AACCTGTGC**T**

2

A**T**CCT**A**TGCT

2

ATCCTA**G**G**T**T

1

ATCCT**G**GGTT

How to model the relationship between mutation and time

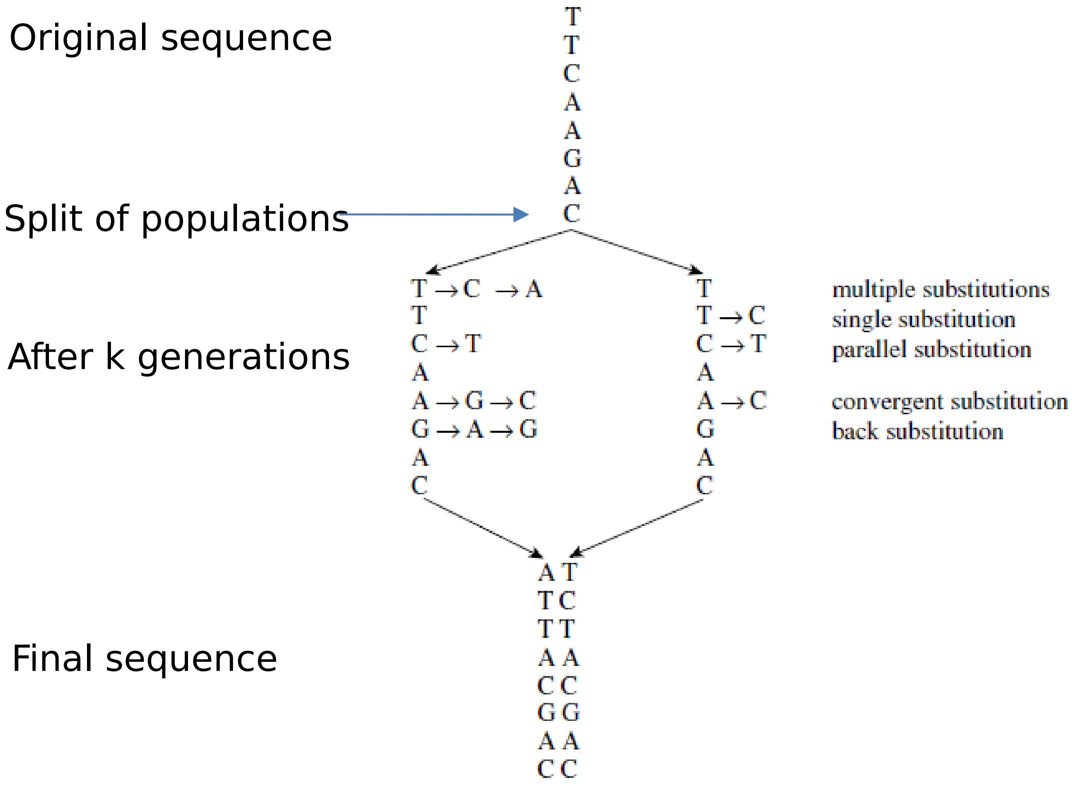
- Assume t, an amount of time
- Assume μ, a mutation rate
- Which is the probability that *n* mutations occur in a branch of *t* length with μ mutation rate?

p

Why is this formula right?

Why is this formula wrong?

Recurrent mutations blur everything!



Original sequence

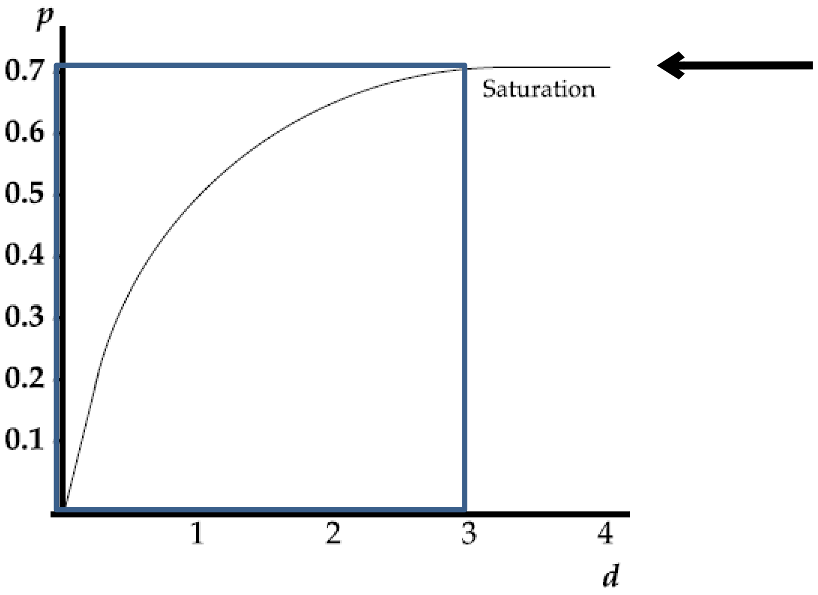
Split of populations

After k generations

Final sequence

From Phylogenetic Handbook: *"The proportion of different homologous sites is called observed distance, sometimes also called p-distance, and it is expressed as the number of nucleotide differences per site. The p-distance is a very intuitive measure. Unfortunately, it suffers from a severe shortcoming: if the degree of divergence is high, p-distances are generally not very informative with regard to the number of substitutions that actually occurred"*

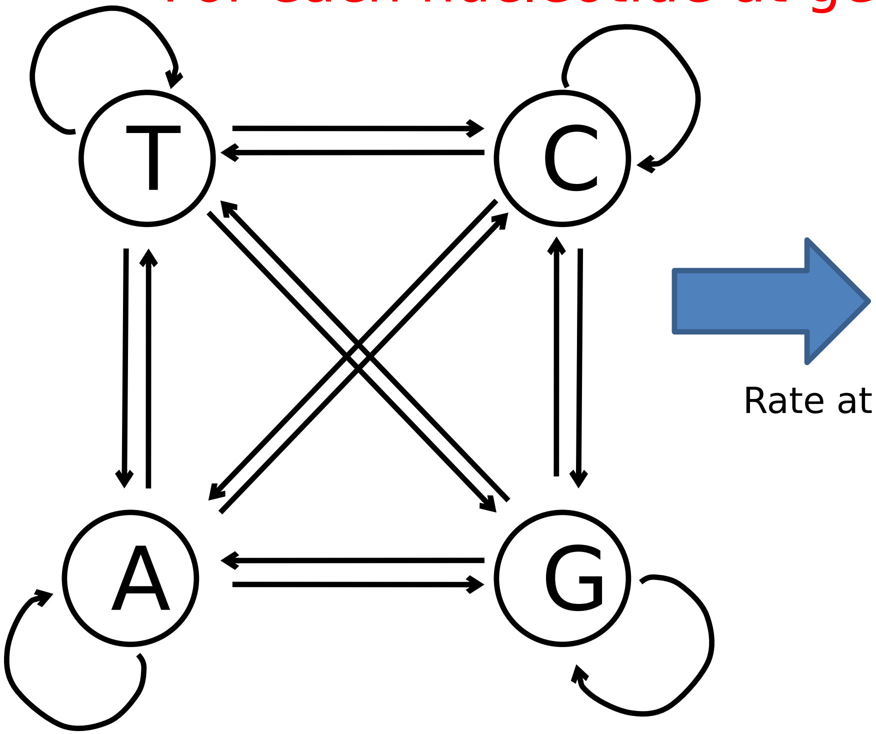
Recurrent mutations blur everything!



Once we reach this point, the genetic distance is no more a proxy for the time of separation

How to model this process?

For each nucleotide at generation t



Substitution matrix

$Q=$

Comes from

	A	C	T	G
A				
C				
T				
G				

Rate at which one nucleotide change to another

Transition matrix

$P=$

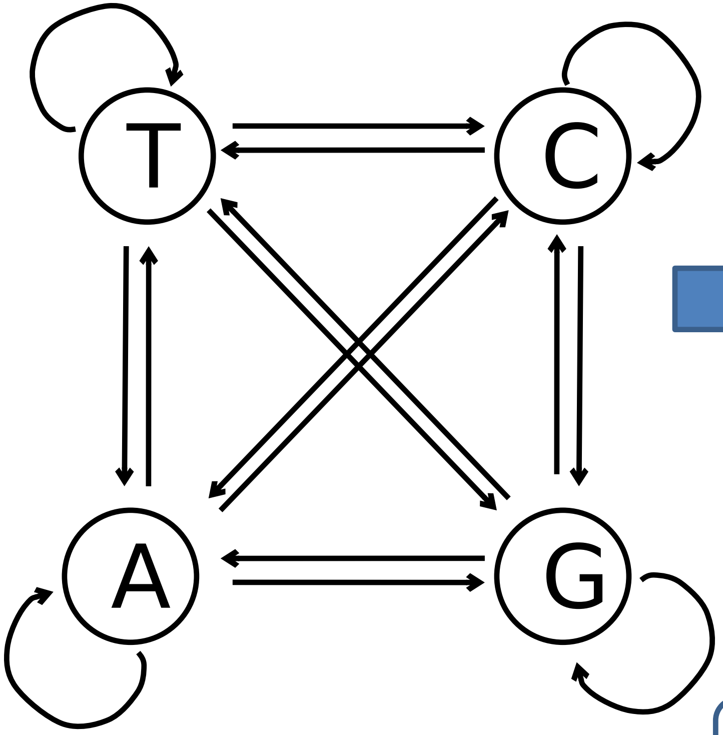
Comes from

	A	C	T	G
A				
C				
T				
G				

Probability to change of one nucleotide from one generation to anot

How to model this process?

For each nucleotide at generation t



Substitution matrix

$Q=$

Comes from

	A	**C**	**T**	**G**
A				
C				
T				
G				

Instantaneous rate of substitution from nucleotide G to nucleotide C

How to model this process?

For each nucleotide at generation t

Transition matrix

Comes from

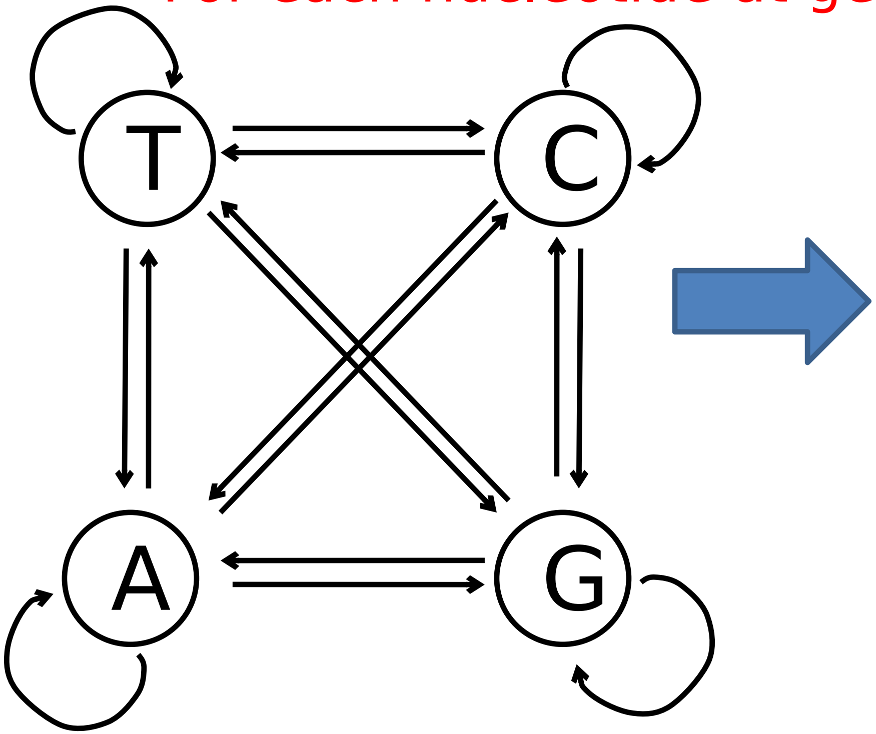
$$P = \begin{array}{c|c|c|c|c} & \mathbf{A} & \mathbf{C} & \mathbf{T} & \mathbf{G} \\ \hline \mathbf{A} & & & & \\ \hline \mathbf{C} & & & & \\ \hline \mathbf{T} & & & & \\ \hline \mathbf{G} & & & & \end{array}$$

Probability that if I am at time t in G, I will move to C at t+1

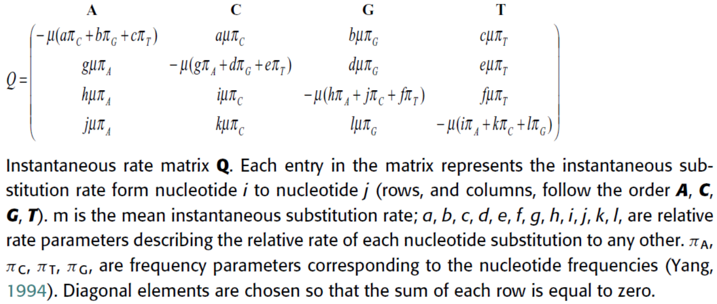
$$P(t)=e^{Qt}=\sum_{n=0}^{\infty} Q^n \frac{t^n}{n!}$$

How to model this process?

For each nucleotide at generation t



Substitution matrix



$$Q=\left(\begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \\ -\mu(a\pi_C+b\pi_G+c\pi_T) & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ g\mu\pi_A & -\mu(g\pi_A+d\pi_G+e\pi_T) & d\mu\pi_G & e\mu\pi_T \\ h\mu\pi_A & i\mu\pi_C & -\mu(h\pi_A+j\pi_C+f\pi_T) & f\mu\pi_T \\ j\mu\pi_A & k\mu\pi_C & l\mu\pi_G & -\mu(i\pi_A+k\pi_C+l\pi_G) \end{array}\right)$$

Instantaneous rate matrix **Q**. Each entry in the matrix represents the instantaneous substitution rate form nucleotide *i* to nucleotide *j* (rows, and columns, follow the order ***A***, ***C***, ***G***, ***T***). m is the mean instantaneous substitution rate; *a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *k*, *l*, are relative rate parameters describing the relative rate of each nucleotide substitution to any other. π_A, π_C, π_T, π_G, are frequency parameters corresponding to the nucleotide frequencies (Yang, 1994). Diagonal elements are chosen so that the sum of each row is equal to zero.

$$\Lambda=\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_4 \end{pmatrix},$$

$$P(t)=e^{Qt}=e^{U^{-1}(\Lambda t)U}=U^{-1}e^{\Lambda t}U$$

How to model the change over time?

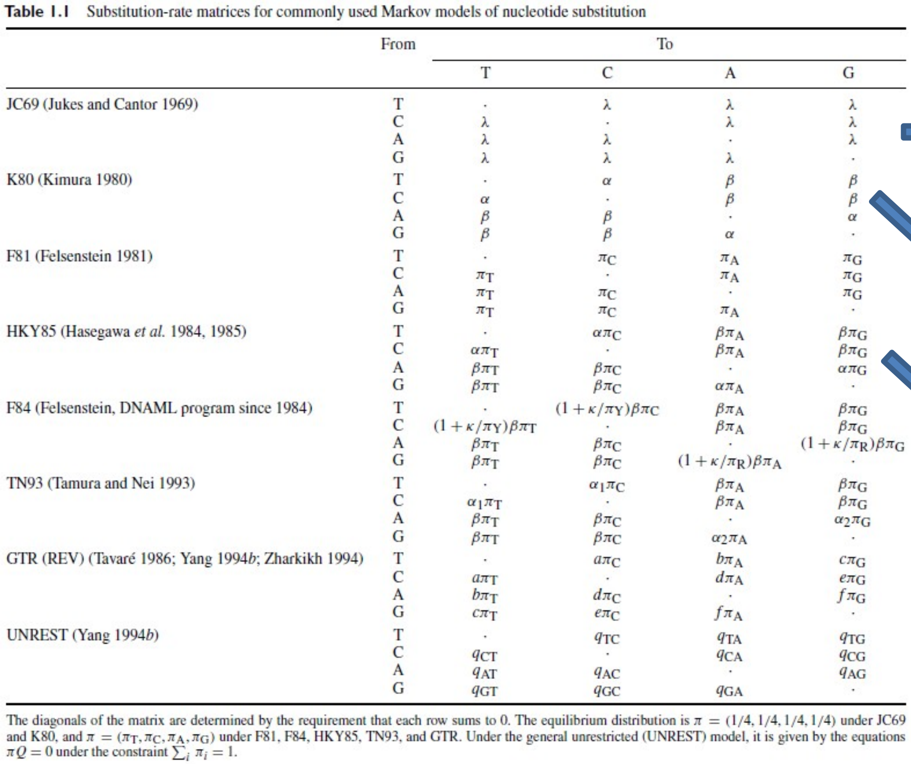
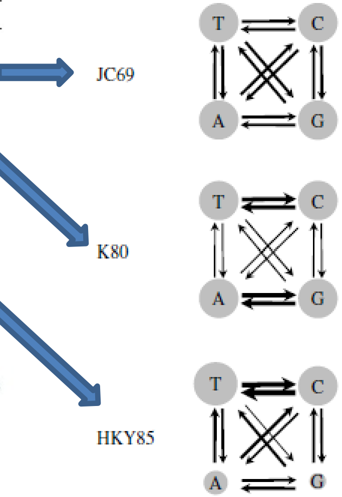


Table 1.1 Substitution-rate matrices for commonly used Markov models of nucleotide substitution

	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	·	λ	λ	λ
	C	λ	·	λ	λ
	A	λ	λ	·	λ
	G	λ	λ	λ	·
K80 (Kimura 1980)	T	·	α	β	β
	C	α	·	β	β
	A	β	β	·	α
	G	β	β	α	·
F81 (Felsenstein 1981)	T	·	π_C	π_A	π_G
	C	π_T	·	π_A	π_G
	A	π_T	π_C	·	π_G
	G	π_T	π_C	π_A	·
HKY85 (Hasegawa *et al.* 1984, 1985)	T	·	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	·
F84 (Felsenstein, DNAML program since 1984)	T	·	$(1+\kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$(1+\kappa/\pi_Y)\beta\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$(1+\kappa/\pi_R)\beta\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$(1+\kappa/\pi_R)\beta\pi_A$	·
TN93 (Tamura and Nei 1993)	T	·	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
	C	$\alpha_1\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
	A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha_2\pi_G$
	G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	·
GTR (REV) (Tavaré 1986; Yang 1994*b*; Zharkikh 1994)	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
	C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
	A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
	G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·
UNREST (Yang 1994*b*)	T	·	q_{TC}	q_{TA}	q_{TG}
	C	q_{CT}	·	q_{CA}	q_{CG}
	A	q_{AT}	q_{AC}	·	q_{AG}
	G	q_{GT}	q_{GC}	q_{GA}	·

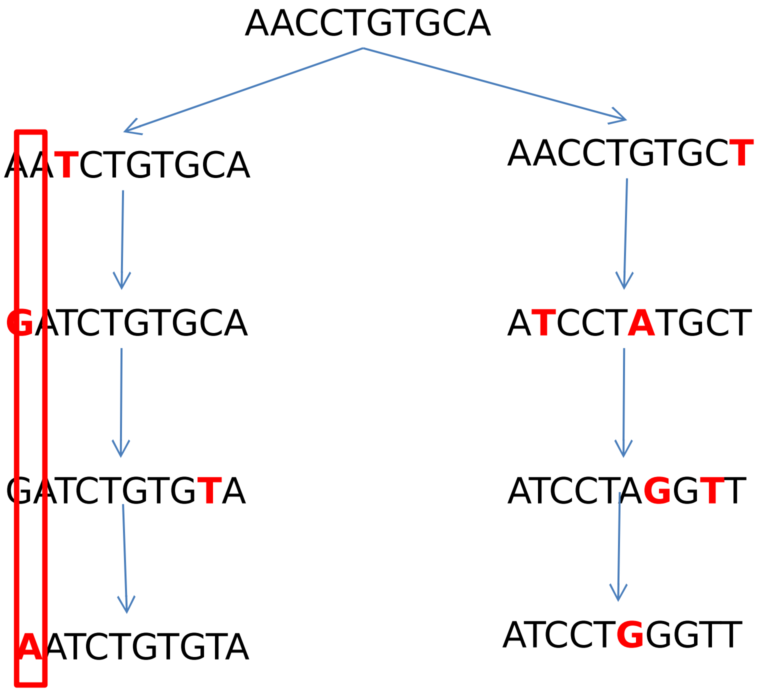
The diagonals of the matrix are determined by the requirement that each row sums to 0. The equilibrium distribution is $\pi = (1/4, 1/4, 1/4, 1/4)$ under JC69 and K80, and $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ under F81, F84, HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by the equations $\pi Q = 0$ under the constraint $\sum_i \pi_i = 1$.



Remember: divergence is a measure of **TIME** of divergence

The trajectory of a particular nucleotide follows a Markov Chain

A -> G -> G -> A



AACCTGTGCA

AA**T**CTGTGCA

AACCTGTGC**T**

GATCTGTGCA

A**T**CCT**A**TGCT

GATCTGTG**T**A

ATCCTA**G**G**T**T

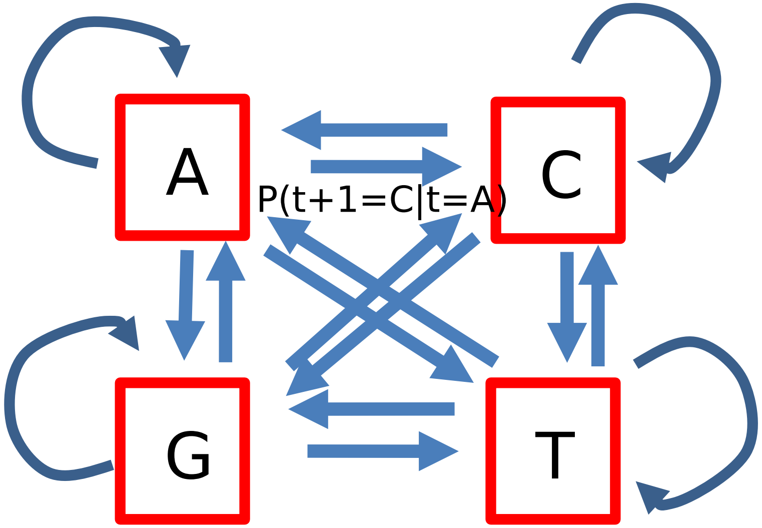
AATCTGTGTA

ATCCT**G**GGTT

Markov Chains

	t0		t1		t2		t3		t4		t5		t6
Markov chain	A	→	A	→	C	→	A	→	G	→	G	→	T

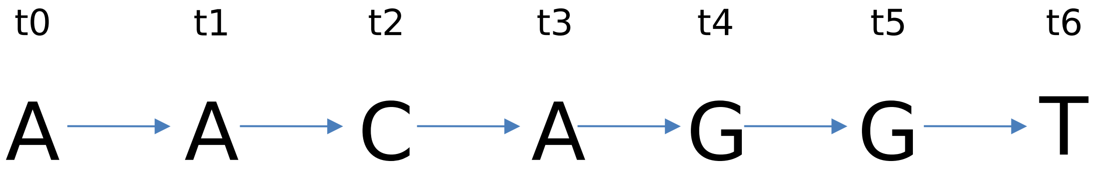
Four possible states



Probability transition matrix

Markov Chains

Markov chain



$P(AACAGGT) = P(AACAGG)\,P(T \vee AACAGG)$

$P(AACAGGT) = P(AACAG)\,P(G \vee AACAG)\,P(T \vee AACAGG)$

From P(X;Y) = P(Y)P(X|Y)

$P(AACAGGT) = P(A)\,P$ ¿

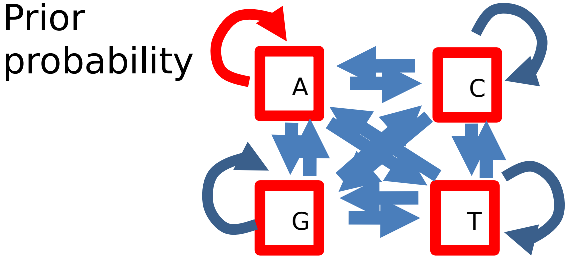
However, in a MC, the probability at a position depends ONLY on the previous state

Markov Chains

t0 t1 t2 t3 t4 t5 t6

Markov chain A → A → C → A → G → G → T

$$P(AACAGGT) = P(A)P¿$$



Markov chains

time-homogeneous time continuous stationary

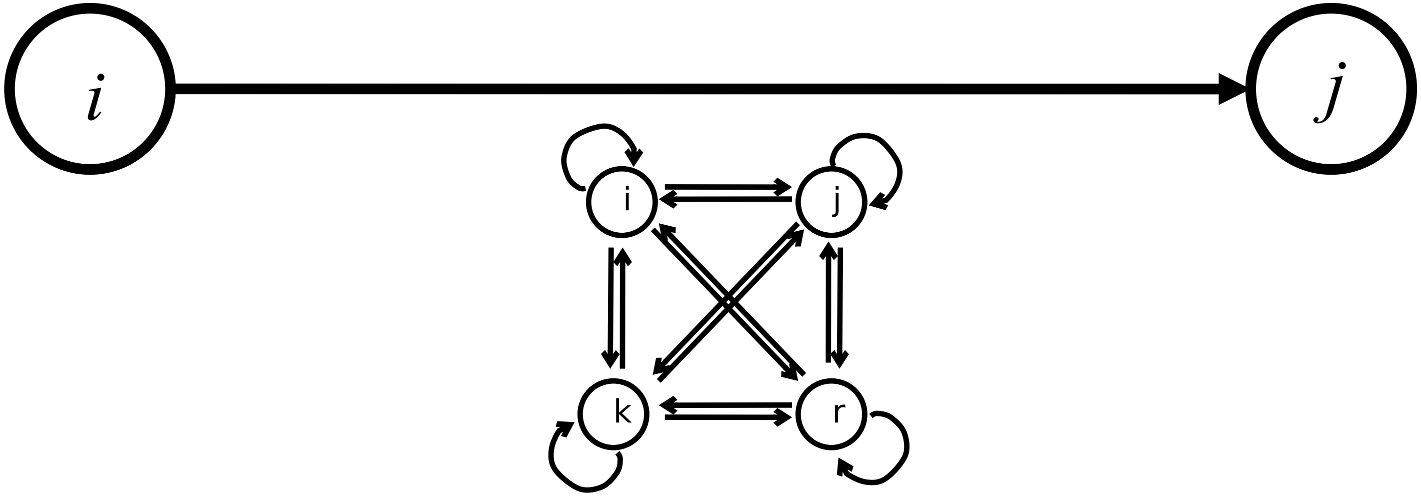
Underlying assumptions:

(1) At any given site in a sequence, the rate of change from base *i* to base *j* is independent
from the base that occupied that site prior *i* (*Markov property*).
(2) Substitution rates do not change over time (***homogeneity***).
(3) The relative frequencies of A, C, G, and T (π_A, π_C, π_G, π_T) are at equilibrium (***stationarity***).

How to compute the probability of changing from nucleotide *i* to *j* after t_1+t_2 times?

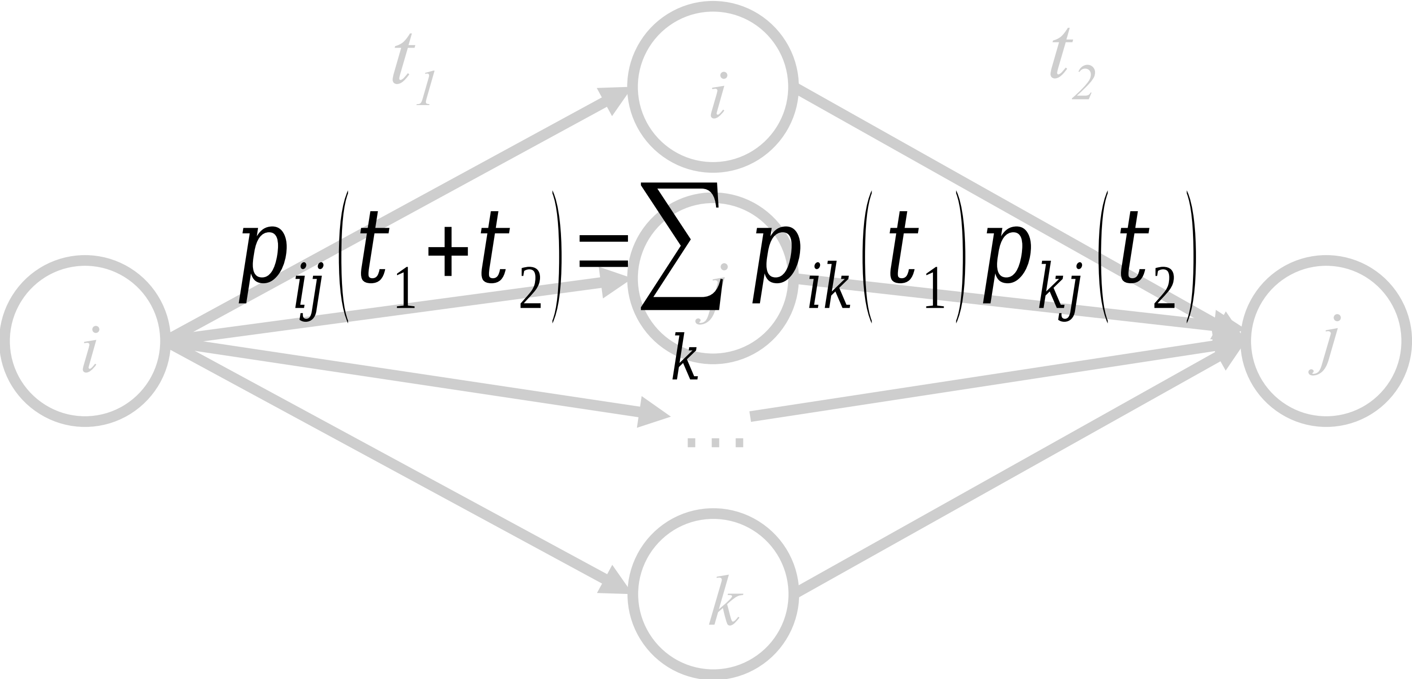
The Chapman-Kolmogorov theorem

$$t_1+t_2$$



How to compute the probability of changing from nucleotide *i* to *j* after t_1+t_2 times?

The Chapman-Kolmogorov theorem

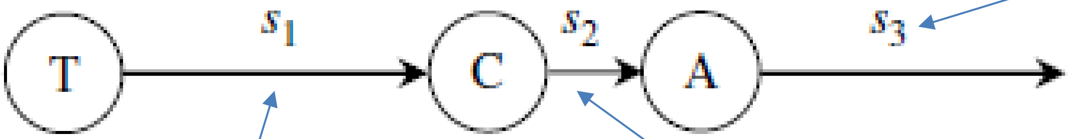


$$p_{ij}\left(t_1+t_2\right)=\sum_k p_{ik}\left(t_1\right)p_{kj}\left(t_2\right)$$

Can you propose an algorithm forward in time for simulating the evolution of a sequence over t generations?

Can you propose another algorithm forward in time for simulating the evolution of a sequence over t generations?

Imagine a single nucleotide and different s times



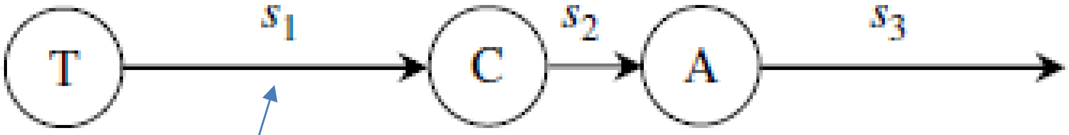
Which is the probability I stay A for s3 generations?

Which is the probability I stay T for s1 generations?

Which is the probability I stay C for s2 generations?

Can you propose another algorithm forward in time for simulating the evolution of a sequence over t generations?

Imagine a single nucleotide and different s times



When I am going to change in probability?

The **exponential distribution**

RANDOM VARIABLE GENERATOR

- Transformation method
  - A function of a random variable is itself a random variable

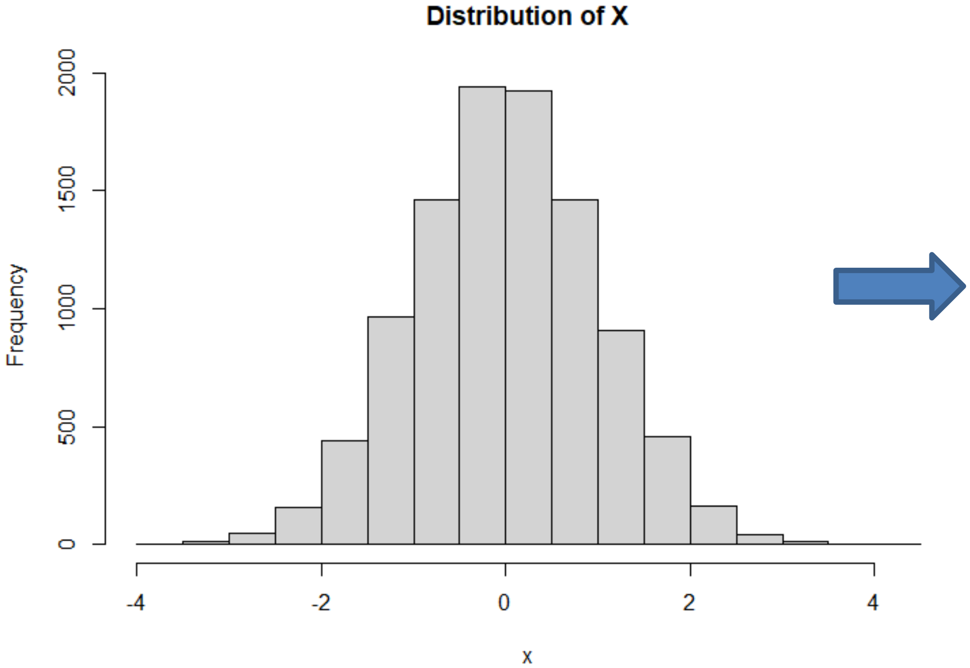
$u \quad U(0,1)$ Random variable uniformly distributed in the range [0,1]

$CDF_{\theta}(x)$ Cumulative density function of variable X

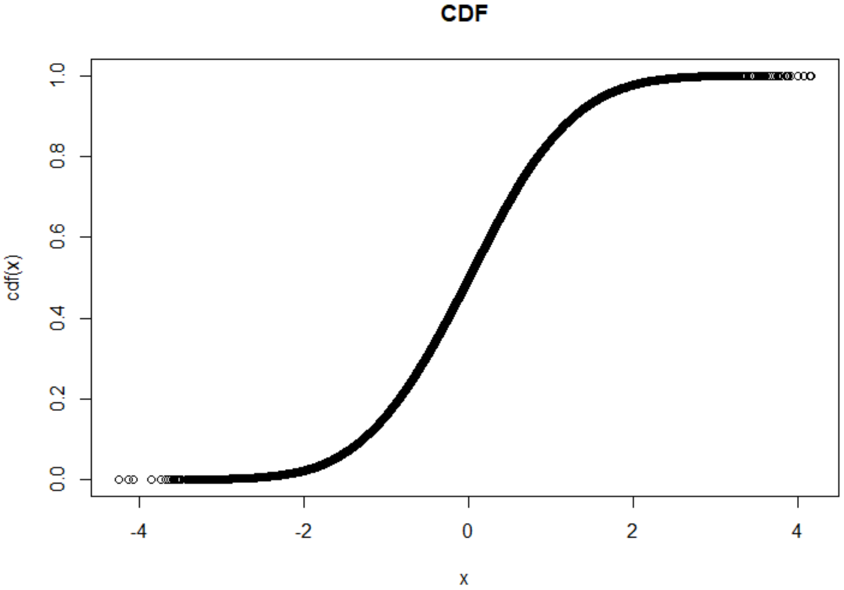
How to get a random sample of x?

RANDOM VARIABLE GENERATOR

Probability density function f(x) from which I want to get random samples



Cumulative density function from which I want to get random samples



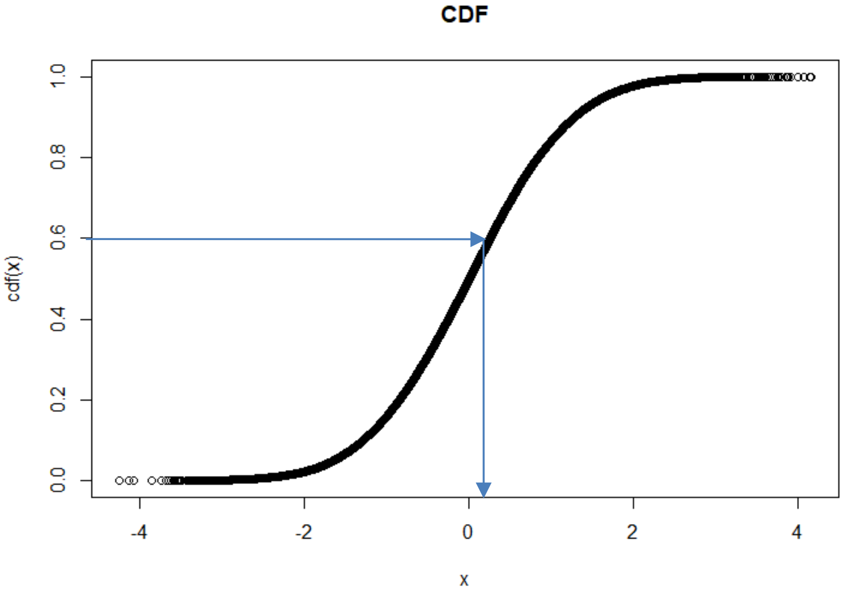
$N(\mu=0,\sigma=1)$ ← $\theta \in \{\mu,\sigma\}$ Parameters that define a normal distribution

RANDOM VARIABLE GENERATOR

Transformation method

$$u \quad U(0,1)$$

Cumulative density function from which I want to get random samples



$$x \quad F_\theta$$

Inverse distribution of the cumulative density function using parameters θ from which we want to get random samples

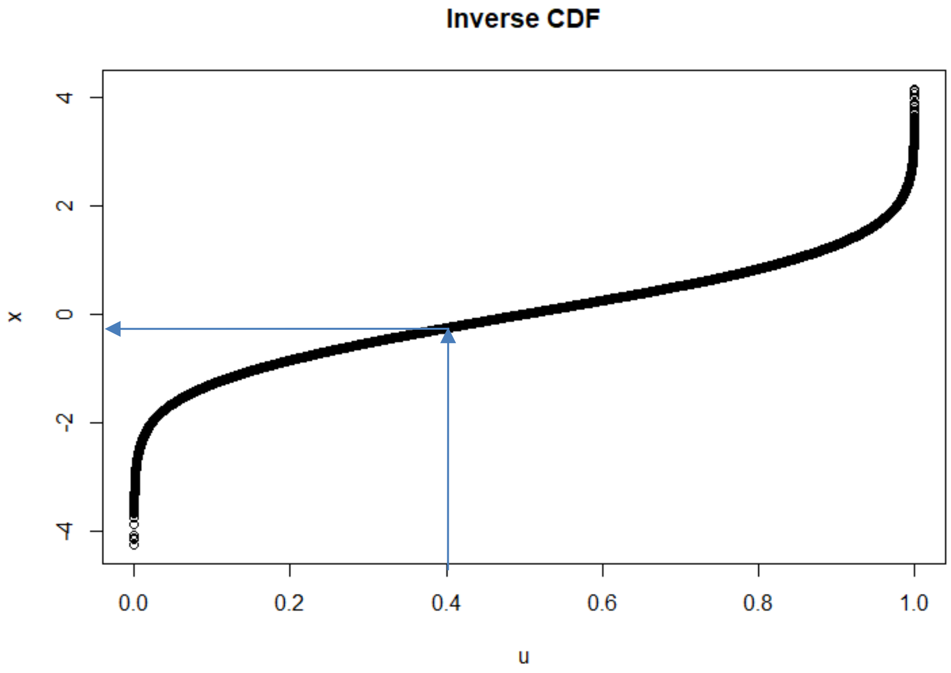
$$x = CDF_\theta^{-1}(u)$$

$$u \quad U(0,1)$$

RANDOM VARIABLE GENERATOR

Transformation method

Inverse Cumulative density function from which I want to get random samples



RANDOM VARIABLE GENERATOR

Bernoulli distribution with parameter p

$$u \quad U\left(0,1\right)$$

$$x = \begin{cases} 0\, if\; u < p \\ 1\, if\; u \geq p \end{cases}$$

RANDOM VARIABLE GENERATOR

Example: The Exponential distribution

$$f(x) = \theta^{-1} e^{\frac{-x}{\theta}}$$

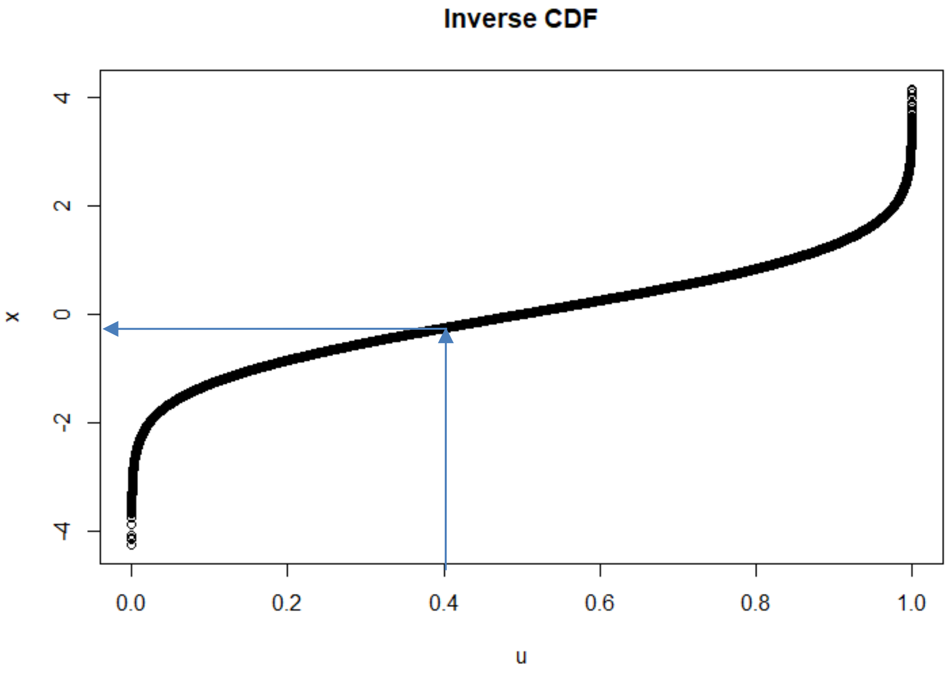
$$CDF_{\theta}(x) = 1 - e^{\frac{-x}{\theta}}$$

$$u = 1 - e^{\frac{-x}{\theta}}$$

$$x = ?$$

RANDOM VARIABLE GENERATOR

Inverse Cumulative density function from which I want to get random samples



For this particular case where the function is Norr

$$N(\theta)$$

$$\theta \in \{\mu, \sigma\}$$

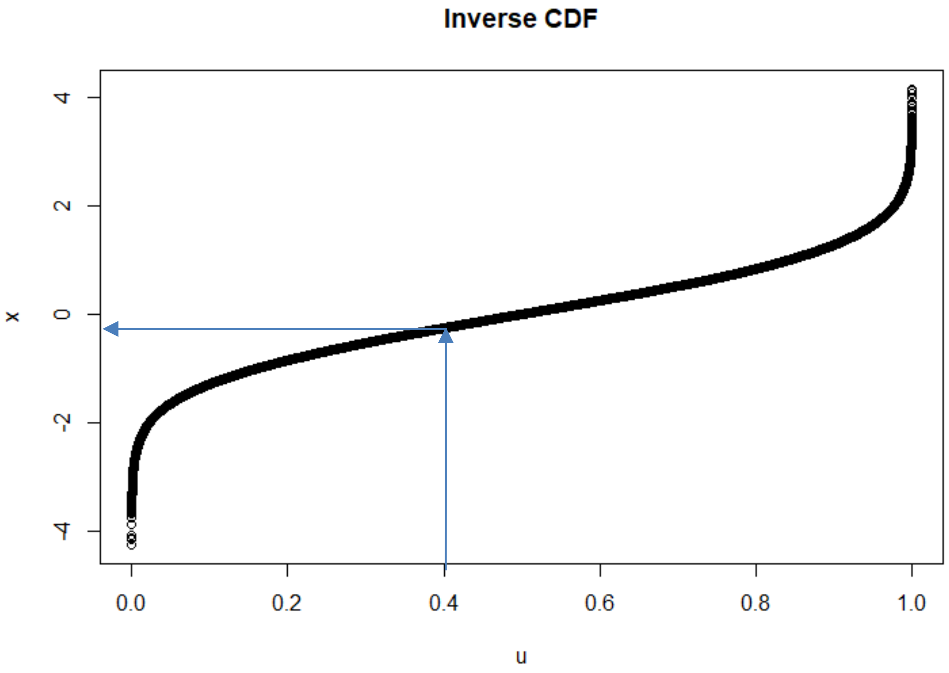
$$\Phi^{-1}(u) = \sqrt{2}\, erf^{-1}(2u-1)$$

$$CDF_{\theta}^{-1}(u) = \mu + \sigma\, \Phi^{-1}(u)$$

Not very easy to compute

RANDOM VARIABLE GENERATOR

Approximations (for Normal distribution)



Box and Muller (1958)

$$u_1 \quad U(0,1)$$

$$u_2 \quad U(0,1)$$

$$x_1 = \sqrt{-2\log(u_1)}\,\sin(2\pi u_2),$$

$$x_2 = \sqrt{-2\log(u_1)}\,\cos(2\pi u_2)$$

RANDOM VARIABLE GENERATOR

Simulating random variables from an arbitrary discrete distribution with a finite number of cells (categories)

$$p_{A,C,T,G} = \{0.35, 0.15, 0.4, 0.1\}$$

1st Approach

- Generate the CDF
  - Sort from the smallest to the largest value
  - Estimate CDF
- Generate
- Pick category x such that

RANDOM VARIABLE GENERATOR

Approach

Problem: For each trial, it requires looking until we suffix the conditio

$$\underbrace{CDF =}_{\text{Unlikely events, but I have to check them}} \text{¿}$$

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

“Any discrete distribution with n cells can be expressed as an equiprobable mixture of n two-point distributions”

$$p_i = \frac{1}{n} \sum_{m=1}^{n} q_i^{(m)}$$

Probability of cell i

Point distribution

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4

Each row must add to 1

Maximum value of the column

To which row the other cell refers to (“alias”)

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
					1
					1
					1
					1

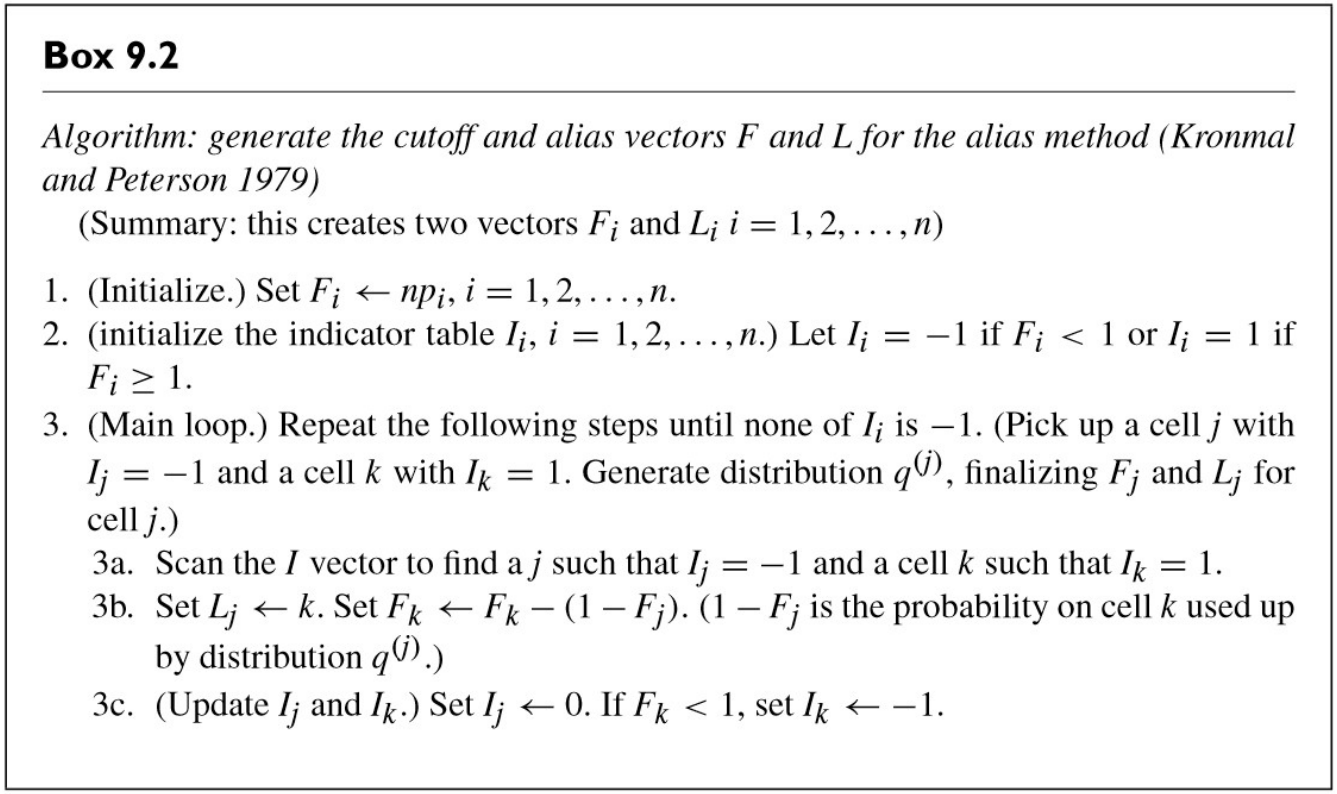
From each row, only two cells can be occupied.
One of the cells must be a column with the same id as row

Maximum value of the column

To which row the other cell refers to

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197



Box 9.2

Algorithm: generate the cutoff and alias vectors F and L for the alias method (Kronmal and Peterson 1979)

(Summary: this creates two vectors F_i and L_i $i = 1, 2, \ldots, n$)

1. (Initialize.) Set $F_i \leftarrow np_i$, $i = 1, 2, \ldots, n$.
2. (initialize the indicator table I_i, $i = 1, 2, \ldots, n$.) Let $I_i = -1$ if $F_i < 1$ or $I_i = 1$ if $F_i \geq 1$.
3. (Main loop.) Repeat the following steps until none of I_i is -1. (Pick up a cell j with $I_j = -1$ and a cell k with $I_k = 1$. Generate distribution $q^{(j)}$, finalizing F_j and L_j for cell j.)
   - 3a. Scan the I vector to find a j such that $I_j = -1$ and a cell k such that $I_k = 1$.
   - 3b. Set $L_j \leftarrow k$. Set $F_k \leftarrow F_k - (1 - F_j)$. ($1 - F_j$ is the probability on cell k used up by distribution $q^{(j)}$.)
   - 3c. (Update I_j and I_k.) Set $I_j \leftarrow 0$. If $F_k < 1$, set $I_k \leftarrow -1$.

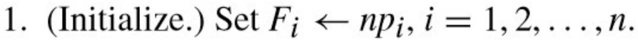
RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

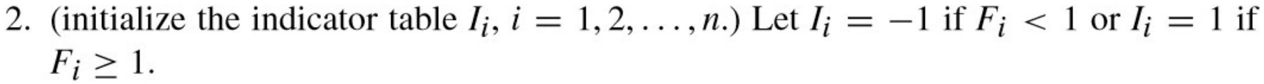


1. (Initialize.) Set $F_i \leftarrow np_i$, $i = 1, 2, \ldots, n$.

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4	1.2	0.8	1.6	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

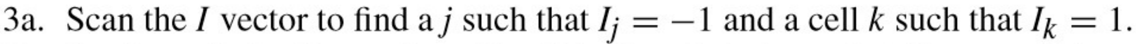


2. (initialize the indicator table I_i, $i = 1, 2, \ldots, n$.) Let $I_i = -1$ if $F_i < 1$ or $I_i = 1$ if $F_i \geq 1$.

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4	1.2	0.8	1.6	
	-1	1	-1	1	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197



3a. Scan the I vector to find a j such that $I_j = -1$ and a cell k such that $I_k = 1$.

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4	1.2	0.8	1.6	
	-1	1	-1	1	

RANDOM VARIABLE GENERATOR

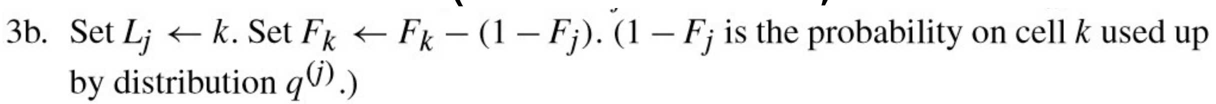
pproach: The alias method (Walker 1974; Kronmal and Peterson 197

3b. Set $L_j \leftarrow k$.

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4	1.2	0.8	1.6	
	4				
	-1	1	-1	1	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

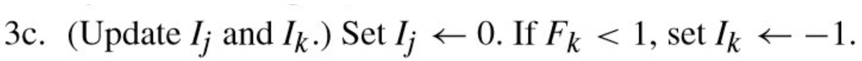


3b. Set $L_j \leftarrow k$. Set $F_k \leftarrow F_k - (1 - F_j)$. $(1 - F_j$ is the probability on cell k used up by distribution $q^{(j)}$.)

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			1-0.4	1
	0.4	1.2	0.8	1.6-(1-0.4)	
	4 -1	1	-1	1	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

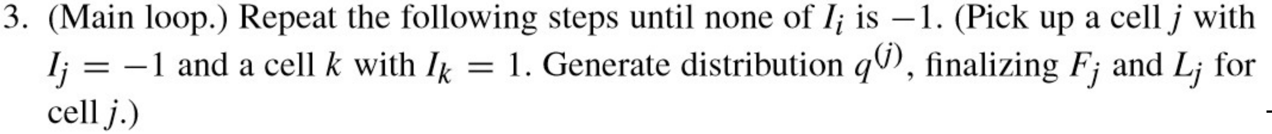


3c. (Update I_j and I_k.) Set $I_j \leftarrow 0$. If $F_k < 1$, set $I_k \leftarrow -1$.

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	1
				1	1
	0.4	1.2	0.8	1	
	4				
	0	1	-1	0	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197



3. (Main loop.) Repeat the following steps until none of I_i is -1. (Pick up a cell j with $I_j = -1$ and a cell k with $I_k = 1$. Generate distribution $q^{(j)}$, finalizing F_j and L_j for cell j.)

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	
				1	
	0.4	1.2	0.8	1	
	4				
	0	1	-1	0	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	1
				1	1
	0.4	1.2	0.8	1	
	4				
	0	1	-1	0	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	**A**	**C**	**T**	**G**	**Sum**
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	1
		1			1
		0.2	0.8		1
				1	
	0.4	1.0	0.8	1	
	4		2		
	0	0	0	0	

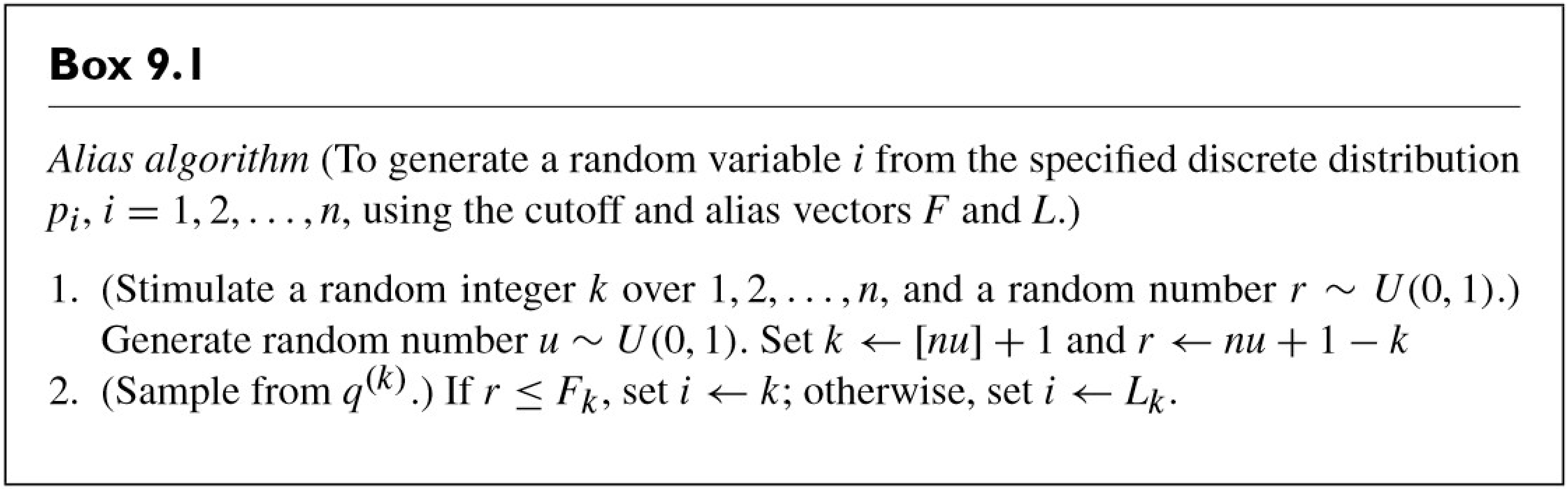
RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	1
		1			1
		0.2	0.8		1
				1	
	0.4	1.0	0.8	1	
	4		2		
	0	0	0	0	

RANDOM VARIABLE GENERATOR

pproach: The alias method (Walker 1974; Kronmal and Peterson 197



Box 9.1

Alias algorithm (To generate a random variable i from the specified discrete distribution p_i, $i = 1, 2, \ldots, n$, using the cutoff and alias vectors F and L.)

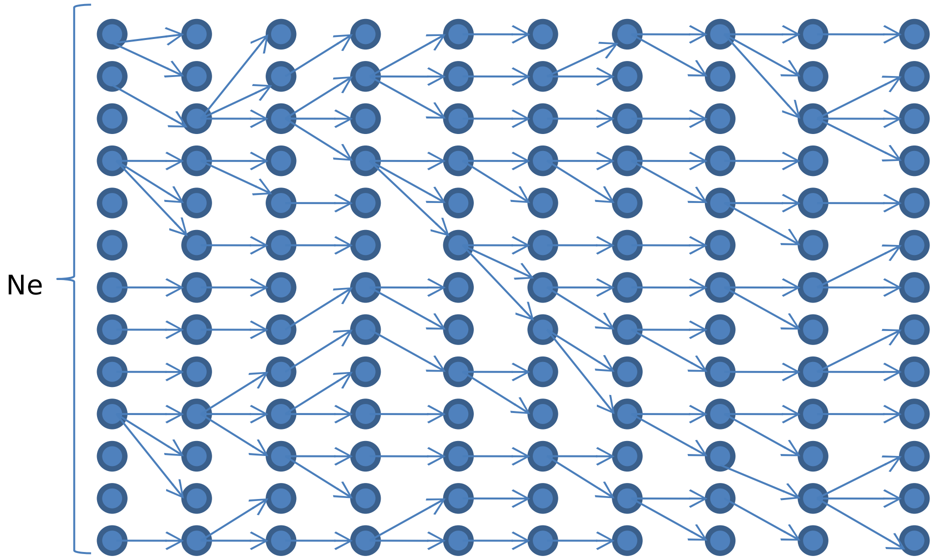
1. (Stimulate a random integer k over $1, 2, \ldots, n$, and a random number $r \sim U(0, 1)$.) Generate random number $u \sim U(0, 1)$. Set $k \leftarrow [nu] + 1$ and $r \leftarrow nu + 1 - k$
2. (Sample from $q^{(k)}$.) If $r \leq F_k$, set $i \leftarrow k$; otherwise, set $i \leftarrow L_k$.

RANDOM VARIABLE GENERATOR

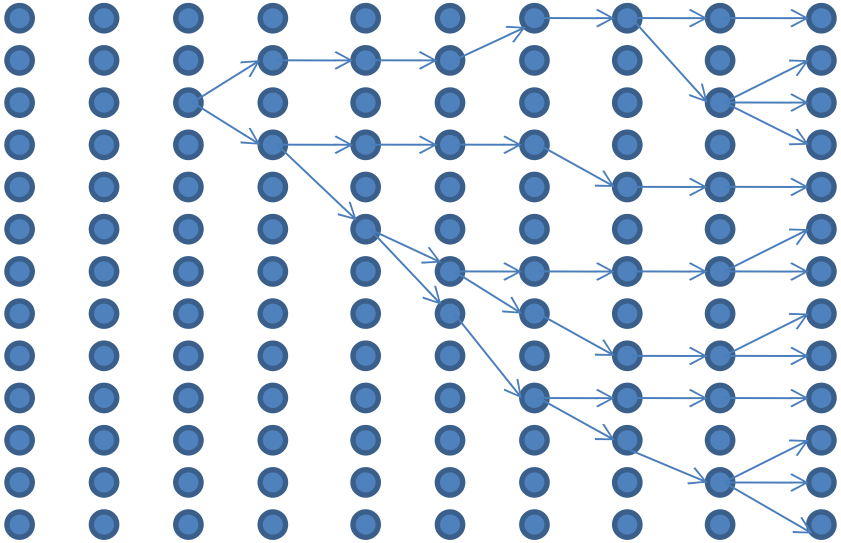
pproach Bis: Extension of the Alias by Vose (A Linear Algorithm For Generating Ra
rs With a Given Distribution)

i	A	C	T	G	Sum
	0.1	0.3	0.2	0.4	1
	0.4	1.2	0.8	1.6	4
	0.4			0.6	1
		1			1
		0.2	0.8		1
				1	
	0.4	1.0	0.8	1	
	4		2		
	0	0	0	0	

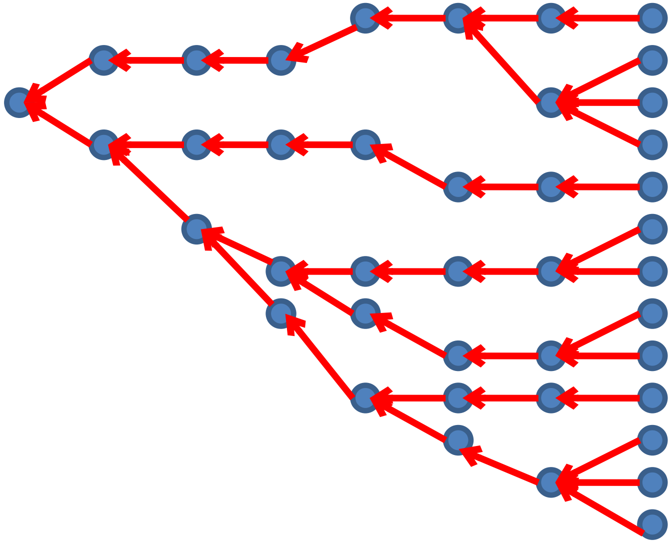
A backward approach



A backward approach



A backward approach



Can you propose an algorithm backward in time for simulating the evolution of K sequences until complete coalescence?