Exercises

1) How is the word length (k) expected to affect the sensitivity and speed of BLAST? longer = less sensitive, faster

2) How many hash keys are needed for blastn with k=11 and for blastp with k=3? 4^11 and 20^3

3) How would you change BLAST parameters if you were trying to find very distantly related homologs?

   Expect threshold (E-value) increase

   Comparison matrix (BLOSSUM62/BLOSSUM45) 45

   Word length (k) smaller

   Filtering low complexity (on/off) on

4) Whats the probability of observing a BLAST score greater than one observed with an E-value of 5? 1-exp(-5)=.9933

$$P(S \geq x) = 1 - \exp\left(-Kmne^{-\lambda x}\right) \qquad E = Kmne^{-\lambda S}$$

Exercises

5) Given the BLOSUM62 scoring matrix, extend the seed to find an ungapped alignment until a drop in score. Write the alignment and score.

```
DKSQVDVIVLVGGSTKVQKLVTDY
seed         GGS
NNLWRNGWRLAGGSSIVQWSRHYA
          .....664.....16
       .....-340664.....20
       .....-3406641-3..21
```

Alignment: LVGGST
LAGGSS

BLOSUM 62 scoring matrix

(positive values are shaded)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Where we are

- String search
- Alignment
- Substitution rates (time models)
- Phylogenetics
- Sequence annotation/motifs (space models)

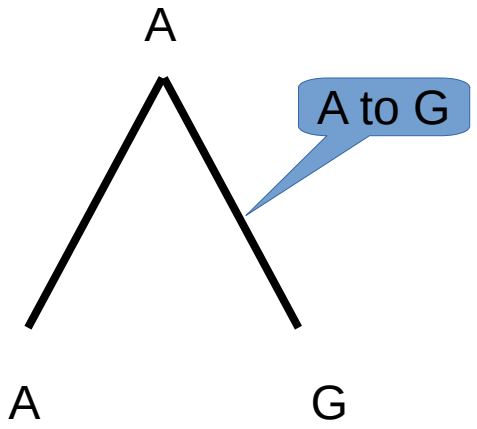
Today's objectives

- Introduction to comparative genomics
- Substitution rates vs differences
- Jukes-Cantor model
- Markov Models
- Markov substitution models
- Maximum likelihood

Introduction

What is a nucleotide substitution model

- model of how a sequence changes over time
- infer prior events, such as ancestral state
- infer model parameters (mutation, selection)



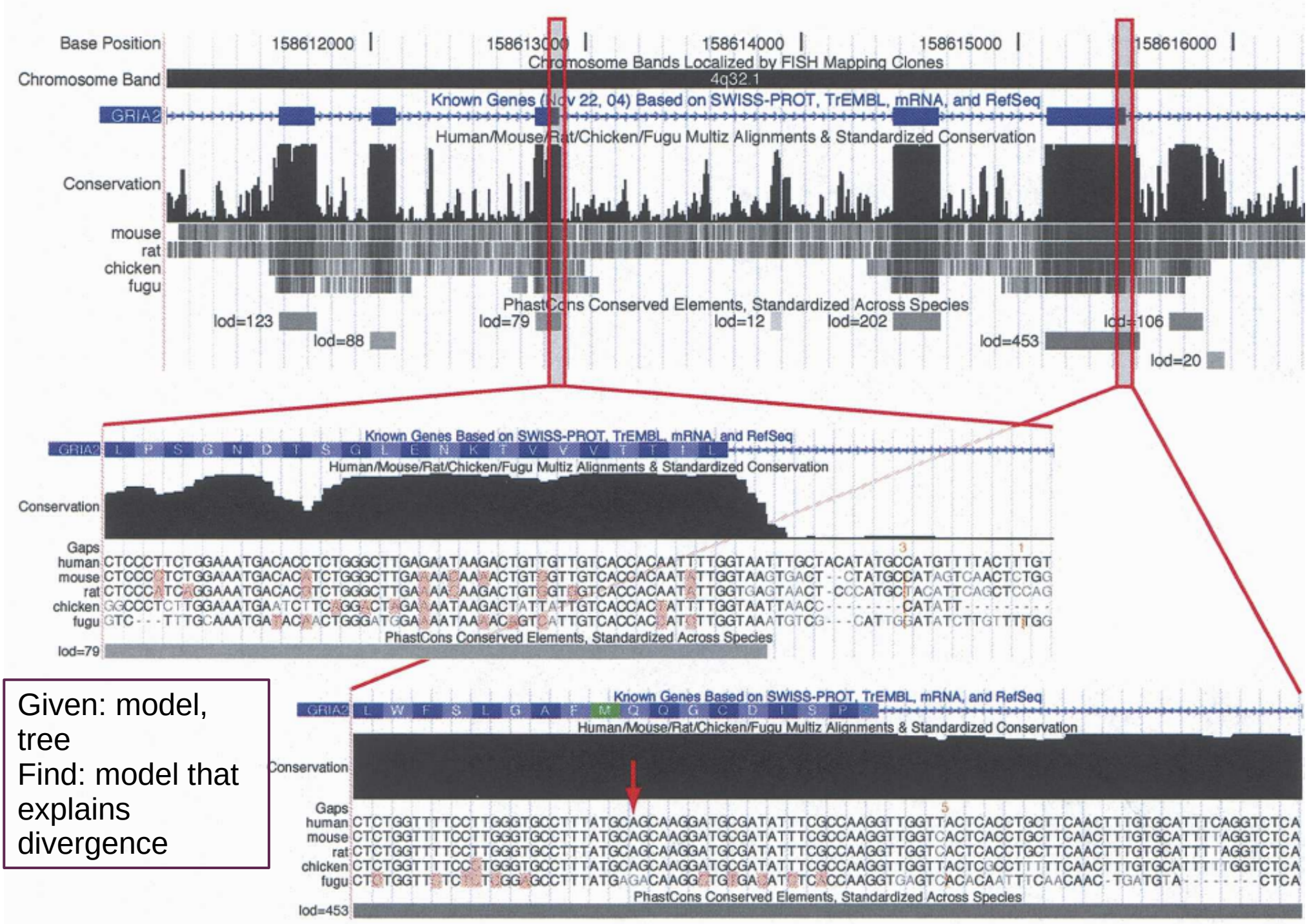
Why do we need substitution models

- construct evolutionary relationships: phylogenetic trees
- date historical events: HIV into humans, neanderthals, tree of life
- identify functional sequences and changes within them, e.g. disease mutations, cancer drivers

Nucleotide Substitution Models

```
CATAGCTAGCAT
CATACCT--CAT
```

- Probability of nucleotide substitutions given alignment.
- Estimate time of divergence
- Estimate selection (constraint) on DNA/protein sequences
- Gaps (indels) are ignored/removed



Given: model, tree
Find: model that explains divergence

Functional sequences are conserved: they evolve slower than non-functional sequences

Percent identity

Problem: model of nucleotide change

- Percent identify (%ID) is the number of differences / number of sites

```
AGTCGTCGACGACC
|||   |       |.|
AGT--T-----AGC
```

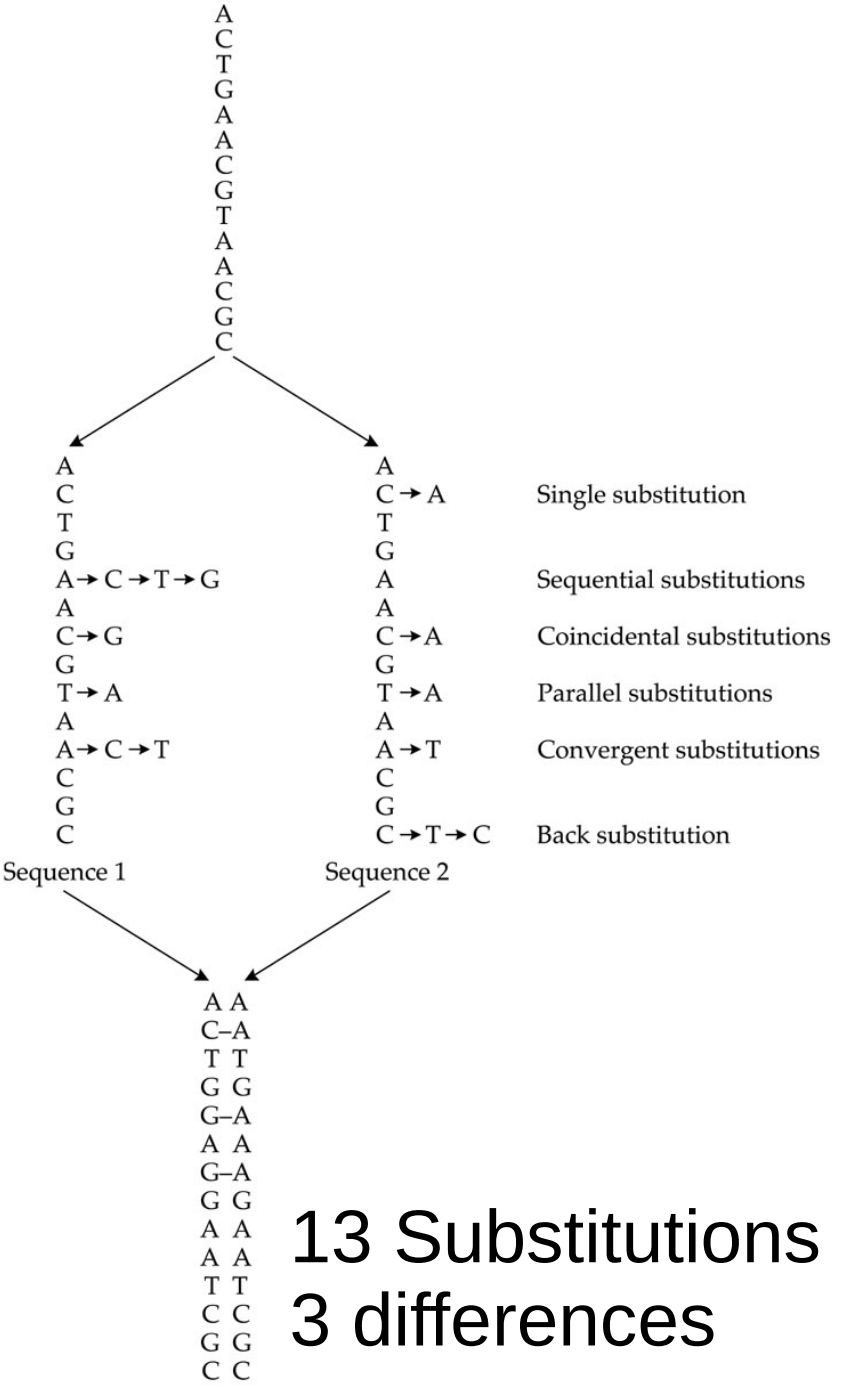
6/7 = 86%
Gaps removed

- %ID decreases with time
- Mutations occur at a constant rate (assumption) and so accumulate linearly with time

Problem: %ID is not linear with time and changes can be missed

Multiple Hits

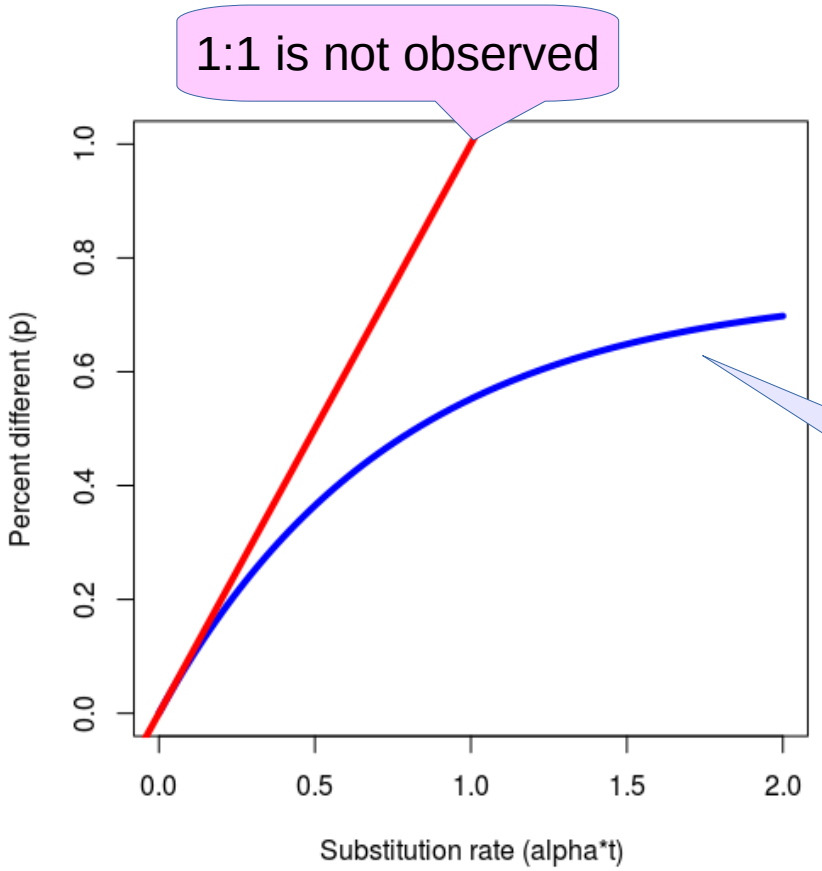
Ancestor (Unknown)	Current (Observed)	Observed Differences	Actual Changes
AAAAA	AAAAT AAAAA	1	1
AAAAA	AAAAC AAAAG	1	2
AAAAA	AAAAT → AAAAC AAAAA	1	2
AAAAA	AAAAC AAAAC	0	2
AAAAA	AAAAA AAAAT → AAAAA	0	2



13 Substitutions
3 differences

Nucleotide Substitution Models

Percent identity = 1 – percent difference



Estimate (model) substitutions based on alignment

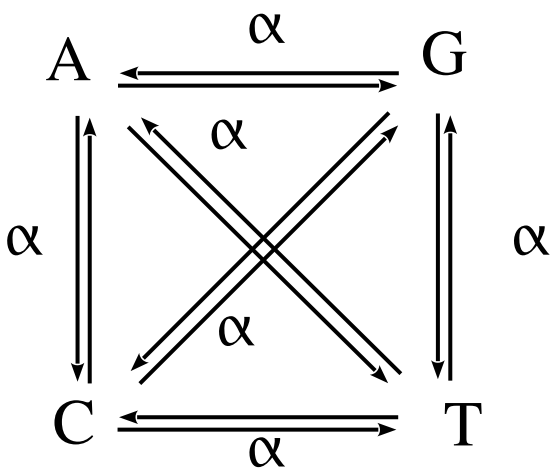
Substitution rate =
substitutions per site

Percent differences
< substitution rate

Solution: find a model such that
Substitution rate = function(%diff)

Jukes and Cantor

```
CATAGCTAGCAT
CATACCT--CAT
```



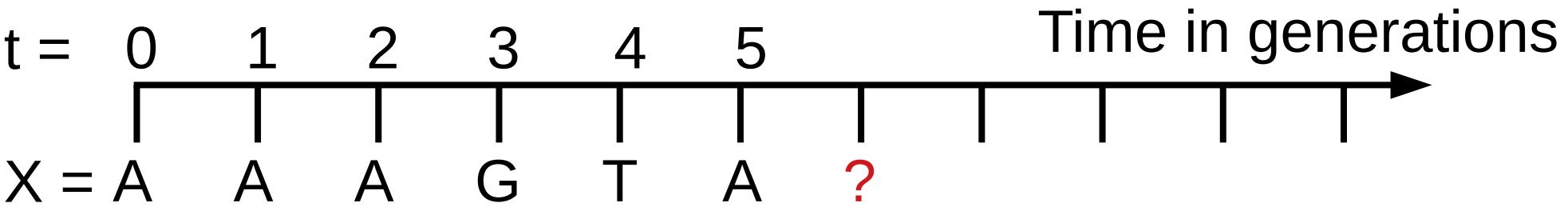
Purines

Pyrimidines

Jukes and Cantor (JC69) Model (1969)
Assumptions of JC model.

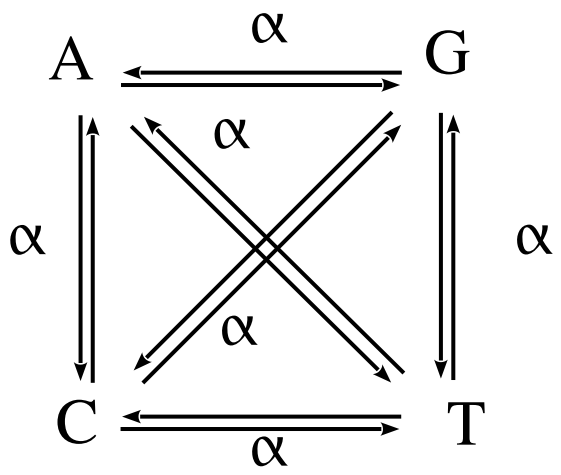
1) Equal base frequencies
2) Equal mutation rates between the bases
3) Constant mutation rate (time & sequence)
4) No selection
5) Sites evolve independent of one another

Discrete time model



$P_{ii}(t) = P(X = i, t \mid X = i, t = 0)$

Probability of going from state i to state i (P_{ii}) in t generations

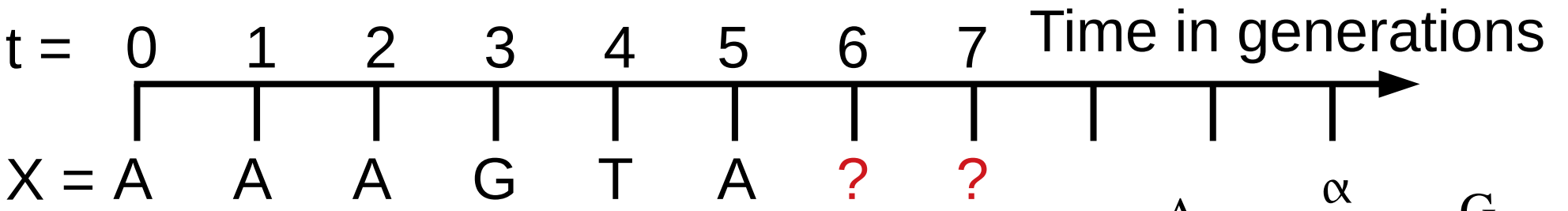


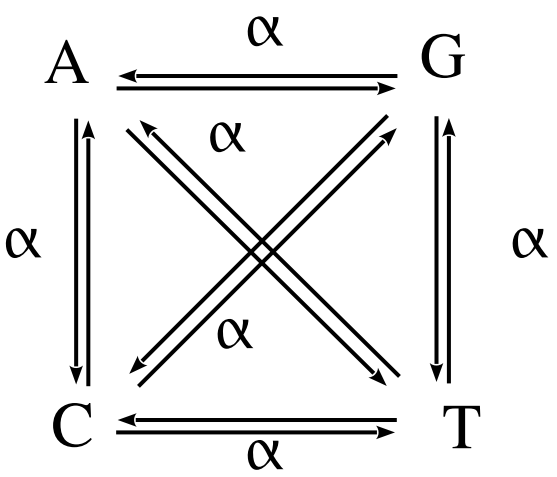
P (no change in 1 generation) = 1-3α

P_{AA} = 1-3α (probability of going from A to A in one generation)

P_{AT} = α

Discrete time model





$P_{AA} = 1-3\alpha$

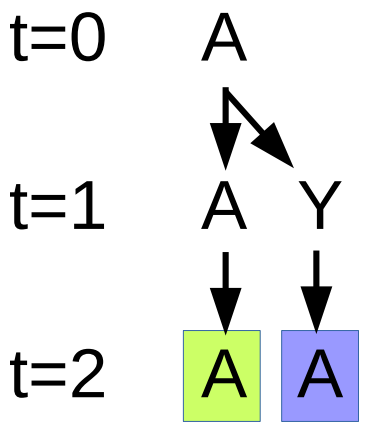
$P_{AT} = \alpha$

$P_{A(6)} = 1-3\alpha$

$P_{A(7)} = P_{A(6)} * P_{AA} + P_{Y(6)} * P_{YA}$

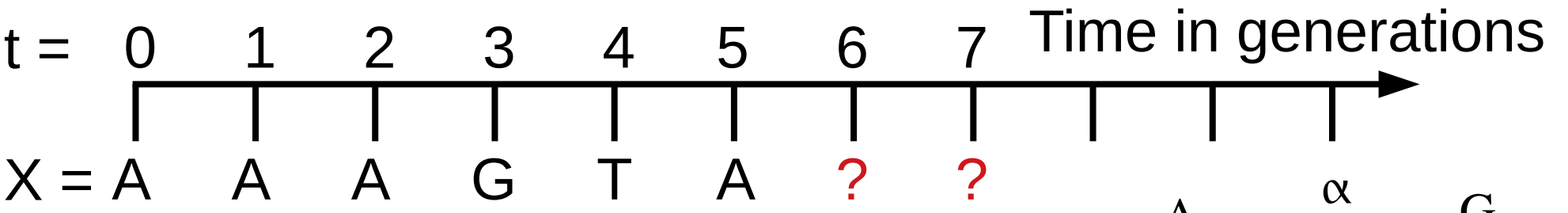
$P_{A(6)}$ = probability of A at t=6

Y = G or C or T



When two events, A and B, are mutually exclusive, the probability that A or B will occur is the sum of the probability of each event:
P(A or B) = P(A) + P(B)

Discrete time model



$P_{AA} = 1-3\alpha$

$P_{A(7)} = P_{A(6)}*P_{AA} + P_{Y(6)}*P_{YA}$

$P_{A(7)} = P_{A(6)}(1-3\alpha)+(1-P_{A(6)})\alpha$

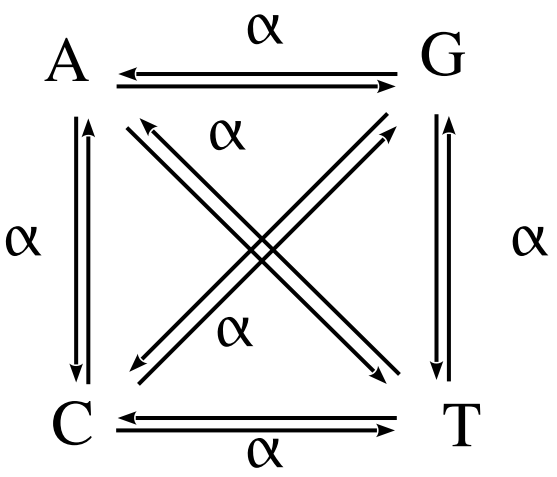
$P_{A(t+1)} = P_{A(t)}(1-3\alpha) + (1-P_{A(t)})\alpha$

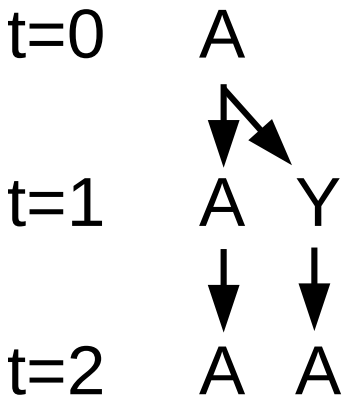
$P_{A(t+1)} - P_{A(t)} = -4\alpha P_{A(t)}+\alpha$

$dP_{A(t)}/dt = \alpha-4\alpha P_{A(t)}$ (continuous time)

$P_{A(t)} = 1/4 + (P_{A(0)} – 1/4)e^{-4\alpha t}$ (solution)

$P_{AA}(t) = 1/4 + (1 – 1/4)e^{-4\alpha t} = 1/4 + 3/4e^{-4\alpha t}$





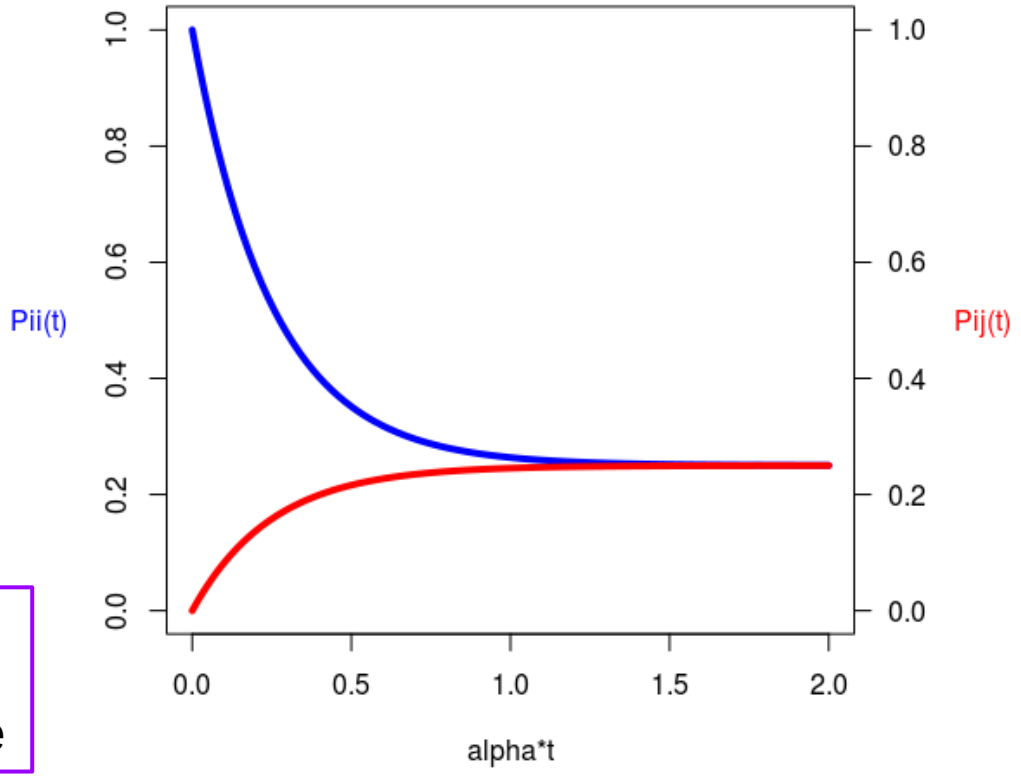
Jukes and Cantor

$$P_{ii}(t) = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

$$P_{ij}(t) = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$$

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3} p\right)$$

p = observed percent differences
K = substitution rate = alpha*t
Substitution rate = mutation rate * time



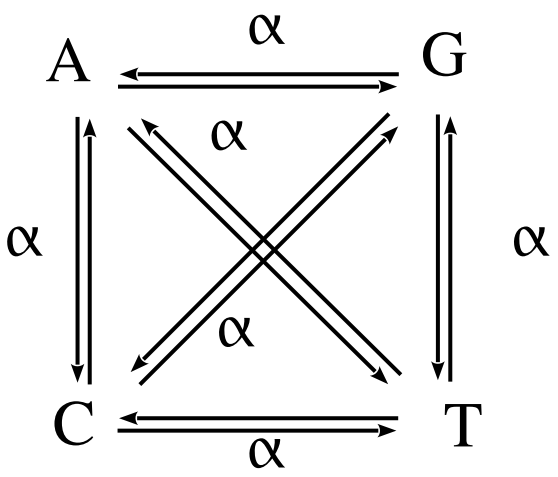
$P_{ii}(t) = P(X = i, t \mid X = i, t = 0)$

If the nucleotide residing at a certain site in a DNA sequence is i at time 0, what is the probability, $P_{ii}(t)$, that this site will be occupied by i at time t?
Substitution rate = number of substitutions per site

Problem: assumptions

```
CATAGCTAGCAT
CATACCT--CAT
```

Transitions: AG, TC
Transversions: AC, AT, GC, GT
GC content varies

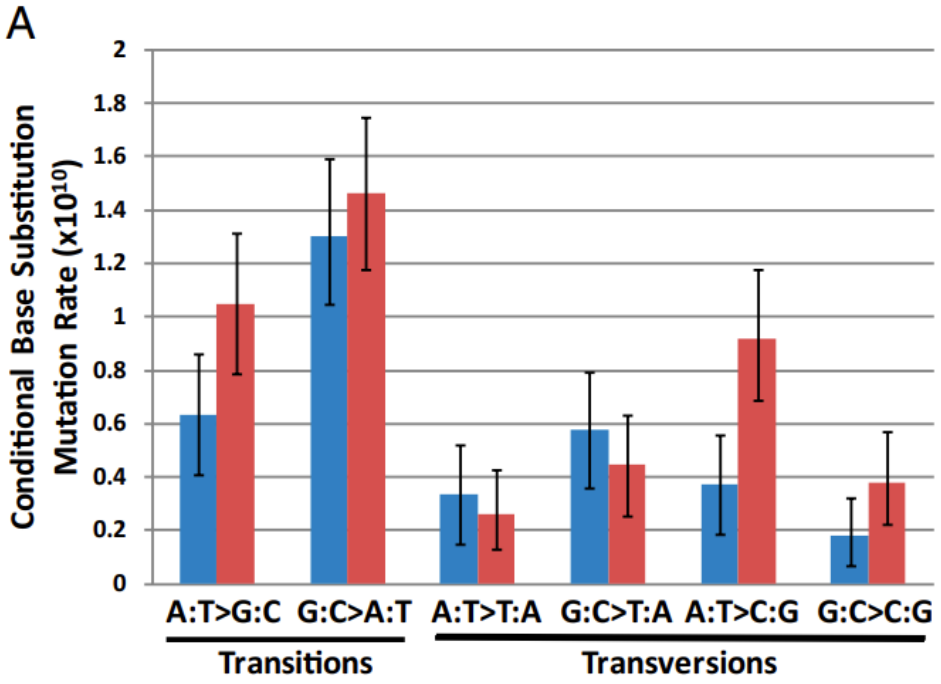


Jukes and Cantor (JC69) Model (1969)
Assumptions of JC model.

1) Equal base frequencies
2) Equal mutation rates between the bases
3) Constant mutation rate (time & sequence)
4) No selection
5) Sites evolve independent of one another

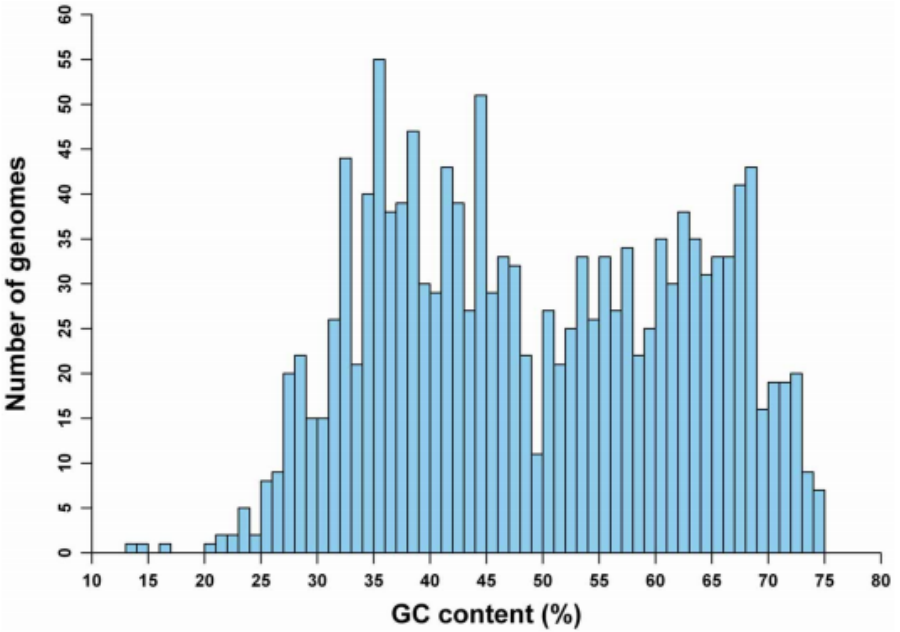
Mutation rates and nucleotide frequencies

Transitions > Transversions



https://www.pnas.org/content/pnas/109/41/E2774.full.pdf

Variation in GC content



Distribution of 1442 archaeal and bacterial genomes in terms of GC content

https://www.frontiersin.org/articles/10.3389/fmicb.2013.00269/full

Transitions/Transversions ratio
Expected = 2/4 = 0.5
Observed = 2

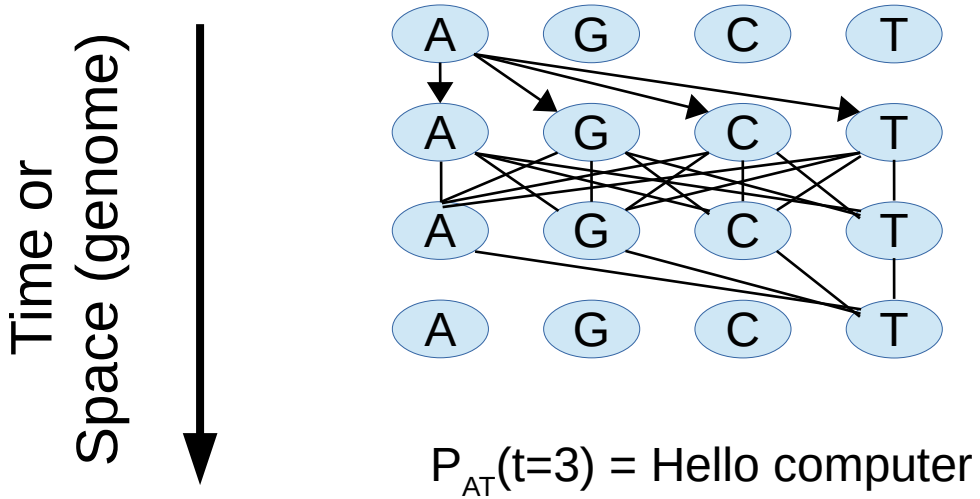
Equal base frequencies
Expected GC = 50%
Observed = 25-75%

A is paired with T, G is paired with C.
An A to G transition on the + strand = T to C on - strand

5'AGCTT 5'GGCTT
3'TCGAA 3'CCGAA

Solution: Markov models

- Random variables, probability theory
- Types of random variables
- Sequences of random variables, stochastic processes
- Markov chains (sequential in time or space), matrix algebra



$P_{AT}(t=3)$ = Hello computer

With Markov Model we can:

- calculate transitions for any length of time
- accounting for all possible paths
- accommodate complex models

Random Variables

A random variable is a variable whose values depend on outcomes of a random phenomenon

Discrete random variables:
e.g. roll of dice (outcome): 5, 3, 1, 4, 2
e.g. toss a coin: H, H, T, T, H
e.g. winning a lottery: L, L, L, L, W

For each event (E) in the sample space (S):

$0 < P(E) \leq 1$

$P(S) = 1$

sum of all event probabilities = 1

Continuously distributed random variables
e.g. speed of car going by
e.g. waiting time for the mailman
e.g. height of humans

A random variable is independent and identically distributed (IID) if they share the same probability distribution and are independent.

Probability of random variables

Probability of an event A can be written as P(A)
If all events have equal chance P(A) = # ways to get A / total # outcomes

1) What's the probability of rolling a 3?
Dice has 6 sides, each with Probability of 1/6

2) What's the probability of rolling two 3s?
1 way to get {3,3}, 36 total outcomes = 1/36

If two events are independent then P(A and B) = P(A)*P(B)
P (rolling two 3s) = P (roll = 3) * P (roll = 3) = 1/6*1/6 = 1/36

3) What's the probability of rolling a 3 and 5?
2 ways to get a 3 and 5: {3,5}, {5,3} = 2/36,
4) What about sequential 3,5? sequential = 1/6*1/6 = 1/36

If two events are mutually exclusive, P(A or B) = P(A) + P(B)

4) What's the probability of getting a 3 or 5, with one roll: 1/6 + 1/6 = 2/6

Types of random variables

Bernoulli: Suppose that a trial, or an experiment, whose outcome can be classified as either a “success” or as a “failure” is performed with probability p and 1 − p. The outcome is a Bernoulli random variable.

2 outcomes, with probability p and 1 − p, e.g. coin toss

Binomial: Suppose that n independent trials, each of which results in a “success” with probability p and in a “failure” with probability 1 − p, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, p).

Many outcomes, e.g. toss a coin 10 times, # heads is a binomial random variable: Binomial(10, 0.5)

$$P(X=k)=\binom{n}{k}p^{k}(1-p)^{(n-k)} \qquad \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

n choose k:
Number of combinations of k (success) given n (trials)
Order doesn't matter

Binomial distribution

Binomial: Binomial(n = 3, p = 0.3), X is the number of heads with probability = 0.3.

P(X = 2)

$$P(X=k)=\binom{n}{k}p^k(1-p)^{(n-k)}$$

P(X = 2) = 3 * 0.3^2 * (1-0.3)^1 = 0.189

With three tosses there are 8 possible outcomes:
HHH = 0.027
HHT = 0.063
HTH = 0.063
HTT = 0.147
THH = 0.063
THT = 0.147
TTH = 0.147
TTT = 0.343

X = 2 for 3 outcomes (of 8)

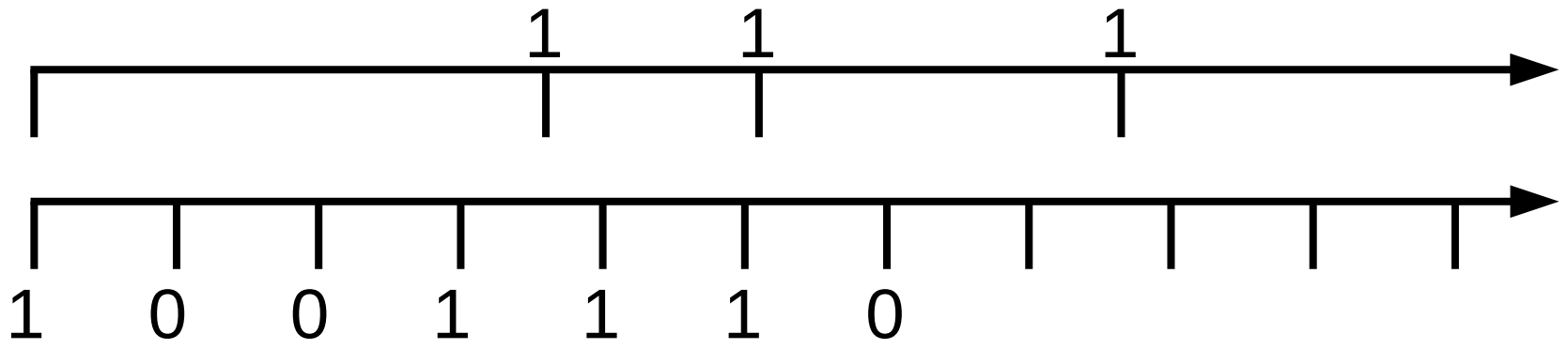
P(X = 2) = 0.063*3 = 0.189

P(X = 0) = 0.343
P(X = 1) = 0.441
P(X = 2) = 0.180
P(X = 3) = 0.027

Bernoulli and Poisson Process (sequential random variables)

Poisson process: a counting process of the number of events over some time interval (continuous time)

Waiting times between events are exponentially distributed with mean 1/rate (Markov process)



Bernoulli process: a sequence of Bernoulli random variables (discrete time)

- Both are memoryless, they do not depend on prior events
- $P(X_{t+1} = 1 \mid X_t = 0) = P(X_{t+1} = 1 \mid X_t = 1) = P(X_{t+1} = 1)$
- $P(X > m + n \mid X > m) = P(X > n)$ or $P(X>40|X>30)=P(X>10)$
- Poisson is a continuous time analog of Bernoulli

Memorylessness

Prior events do not change the probability of a random variable (Markov property-only the current state):

Gambler's fallacy: the incorrect belief that, if a particular event occurs more frequently than normal during the past, it is less likely to happen in the future

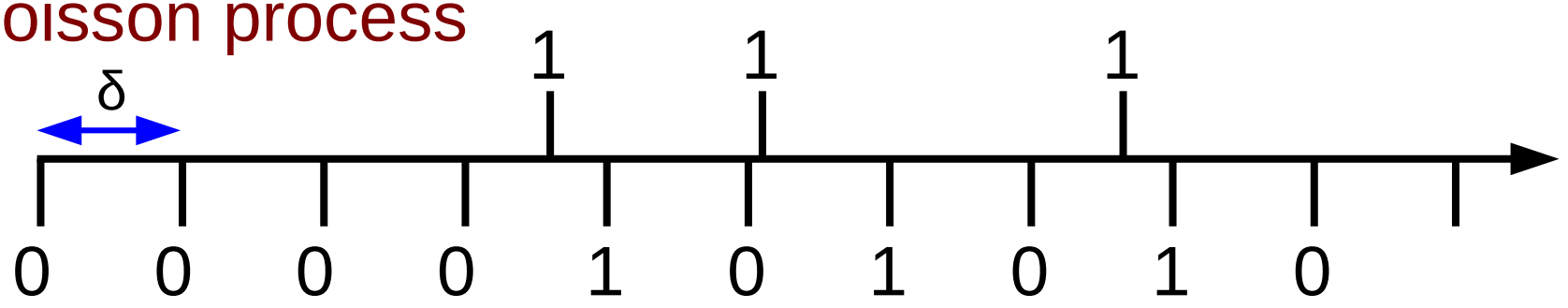
P(heads | H) = 0.5
P(heads | HHHHH) = 0.5

P(mutation in 1,000,010 years | no mutation in million years) = P(mutation in 10 years)

Conversion of discrete to continuous time

Converting to continuous time will help us find solutions to the nucleotide substitution models. Lets figure out how to convert between discrete and continuous time.

Poisson process



Lets look at a Poisson process, but instead of looking at events which happen in continuous time, lets look at events in discrete time intervals δ (delta)

Poisson Process: the number of events (n) in time t is

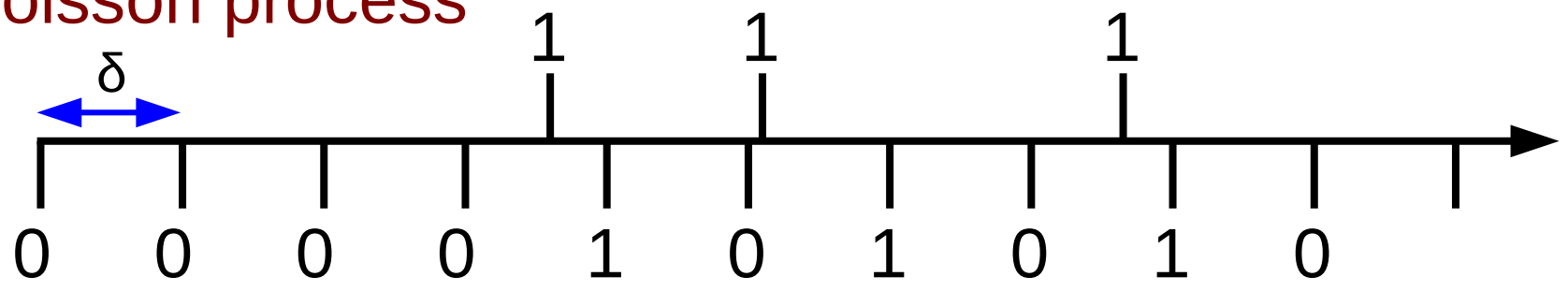
$$P\left(N\left(t\right)=n\right)=\frac{\left(\lambda t\right)^{n}}{n!}e^{-\lambda t}$$

Where λ is the rate

Conversion of discrete to continuous time

Converting to continuous time will help us find solutions to the nucleotide substitution models. Lets figure out how to convert between discrete and continuous time.

Poisson process



Lets look at a Poisson process, but instead of looking at events which happen in continuous time, lets look at events in discrete time intervals δ (delta)

This is why Poisson process is so common

As δ becomes very small, Poisson process converges to Bernoulli process with $p = \lambda\delta$ and $n = t/\delta$ such that two events occurring in δ becomes negligible.

Number of events in a Bernoulli process is Binomial(n, p)

Conditional Probability & Memory

If P(A | B) = P(A), then events A and B are said to be independent

If A and B are dependent then: $$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

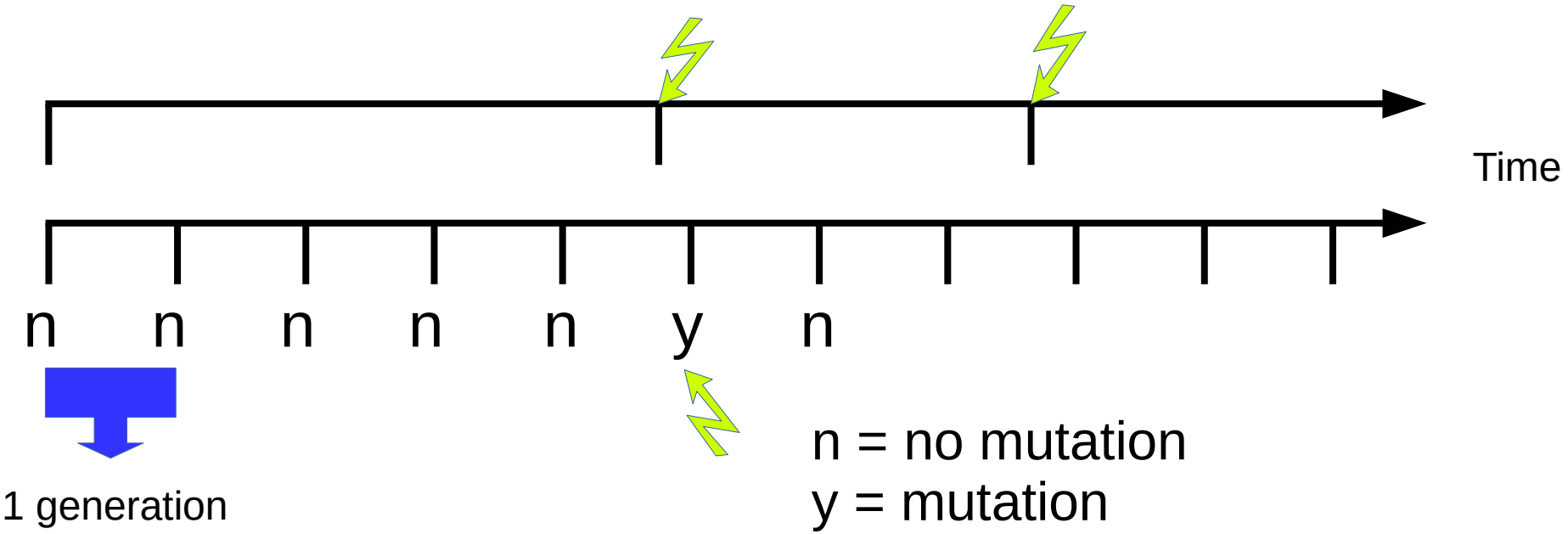
$$P(A \text{ and } B) = P(A|B) \cdot P(B)$$

For variables in a sequence (time or space):

- Memoryless means sequential numbers are not dependent
- 1st order Markov chains: variables depend only on the preceding state. 2nd order Markov chains depend on preceding 2 states, etc.

Mutational process (independent)

Number of mutations are a Poisson process with rate λ over time, expectation or mean value of t*λ



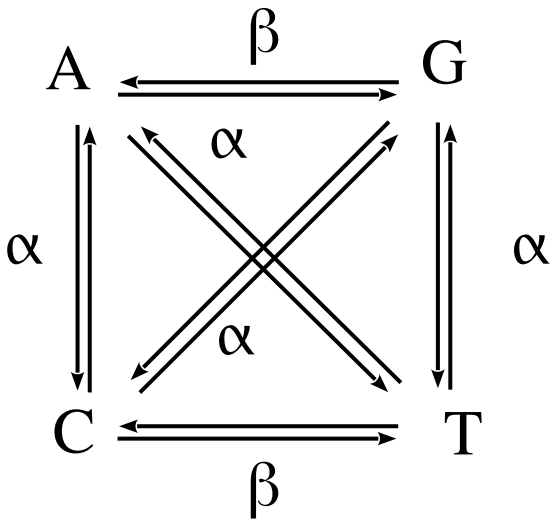
Mutations are a Bernoulli process with p(success) = λ, lets say λ = 1e-8 per site, per generation

Problem: DNA substitution depends on current state

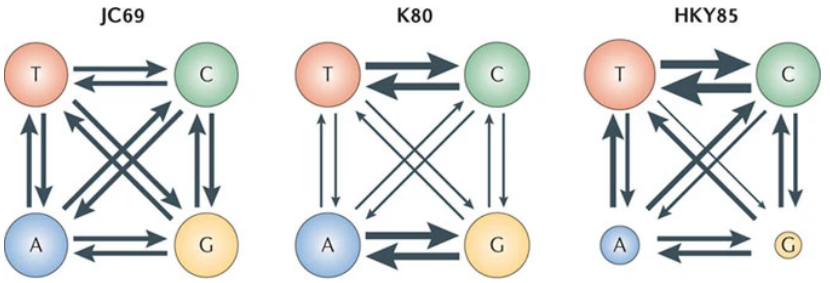
Transitions != Transversions
P(A | T) != P(A | G)
e.g. $\beta > \alpha$
AGAGAGCTCTCTCAGAGAG

GC != AT
P(C | A) != P(A | C)
P(C | T) != P(T | C)
Same for G
We need more parameters!
e.g. $P_{AC} > P_{CA}$
ACGCGACTGTGCACGTGCGAC
ACGCGACTGTGCACGTGCGAC

Thus, A is more rare, C is more common



β = transitions
α = transversions



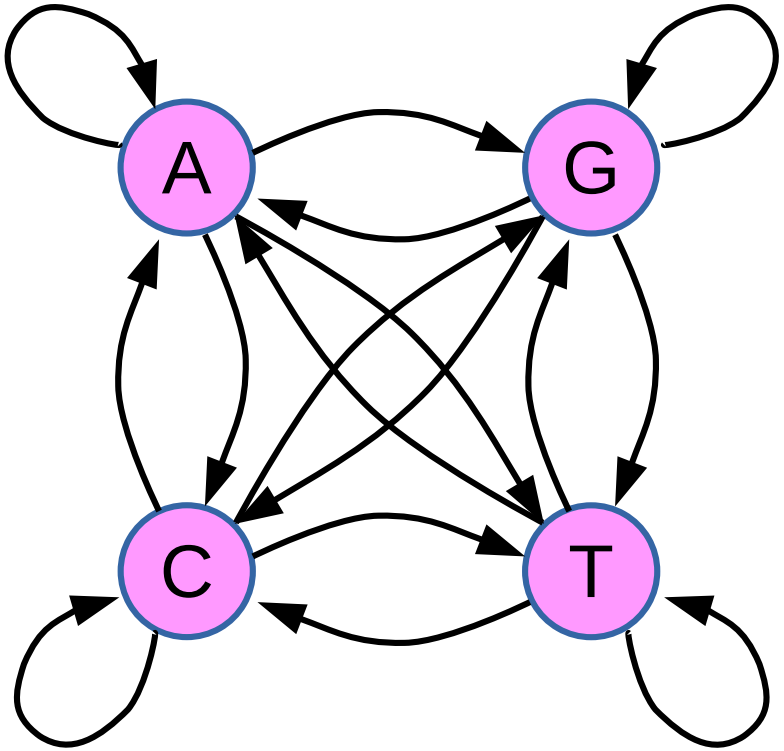
Markov Models

A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event (i.e. the current state).
A Markov chain is the sequences of transitions through time of a Markov process.

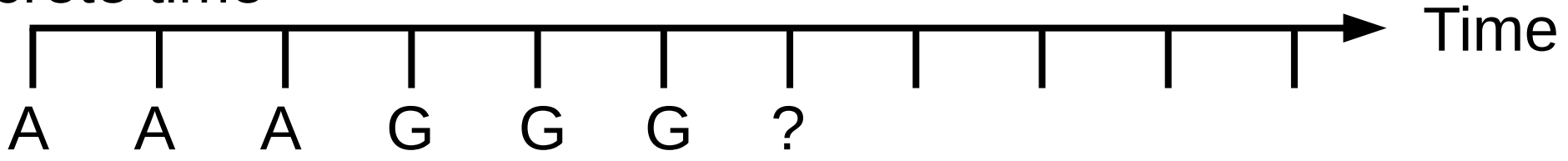
Markov process:
(a) The number of possible outcomes or states is finite.
(b) The outcome at any stage depends only on the outcome of the previous stage.
(c) The probabilities are constant over time.
Transitions determined by transition rate matrix

Markov Models

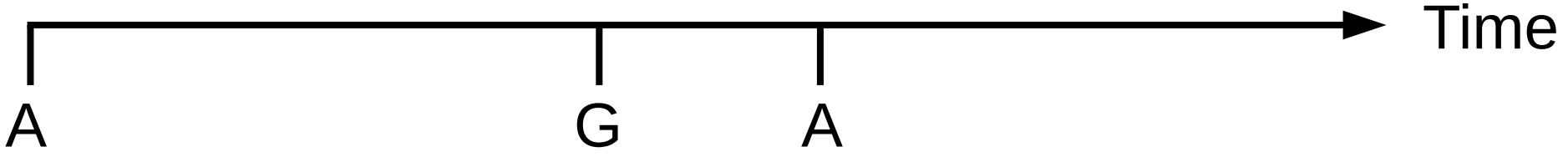
- States {A,G,C,T}
- Transitions between states are governed by transition matrix
- Time between transitions are exponential random variables (continuous)



Discrete time



Continuous time



Markov Chains in Biology

Time

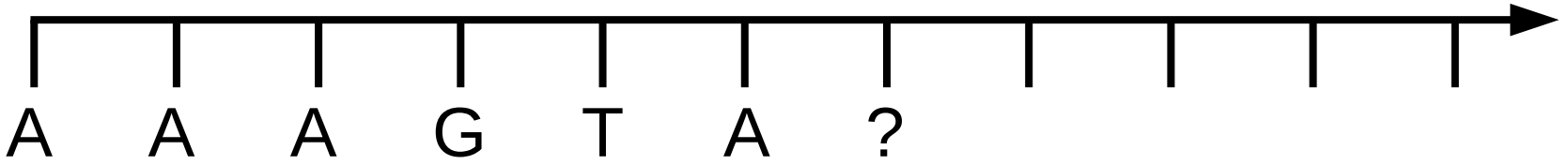
- Birth - death process (ecology)
- Substitution rate (molecular evolution)

Space

- DNA sequence features (genes, hidden)

Discrete Time Markov Chain

A Markov Chain:



$$P_{i \to j} = P\left(X_{n+1} = j \,|\, X_n = i\right)$$

Depends on current state

$$P_{i \to j} = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AT} \\ p_{GA} & p_{GG} & p_{GC} & p_{GT} \\ p_{CA} & p_{CG} & p_{CC} & p_{CT} \\ p_{TA} & p_{TG} & p_{TC} & p_{TT} \end{pmatrix} \quad P_{ij} = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

Jukes-Cantor Model
No Memory

http://setosa.io/blog/2014/07/26/markov-chains/

Exercises

1) Calculate the nucleotide substitution rate using JC69 model and this alignment:

```
CGATCGATCGA
CAAGCCA-CTA
```

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

2) Why is p < K in the JC69 Model?

3) Under the JC69 model, how often would you expect to see an A in one sequence? How often would you expect to see an A aligned to A between two sequences as divergence time goes to infinity?

4) What is the probability of ATG going to ATC given a substitution rate (αt) of 0.5?

5) What are the assumptions of JC model?