CODEML WORKSHOP

Practical activity #1

TOPICS:

- Introduction to the Linux command line
- How do we measure evolution on a genetic sequence?
- What is codeml?
- Branch models
- Using codeml to test a hypothesis

LINUX COMMAND LINE

INTRODUCTION

- navigating: cd
- listing: ls
- making folders: mkdir
- creating: nano [filename]
- moving: mv
- visualizing: less, cat, more
- manipulating, editing files: sed, sort

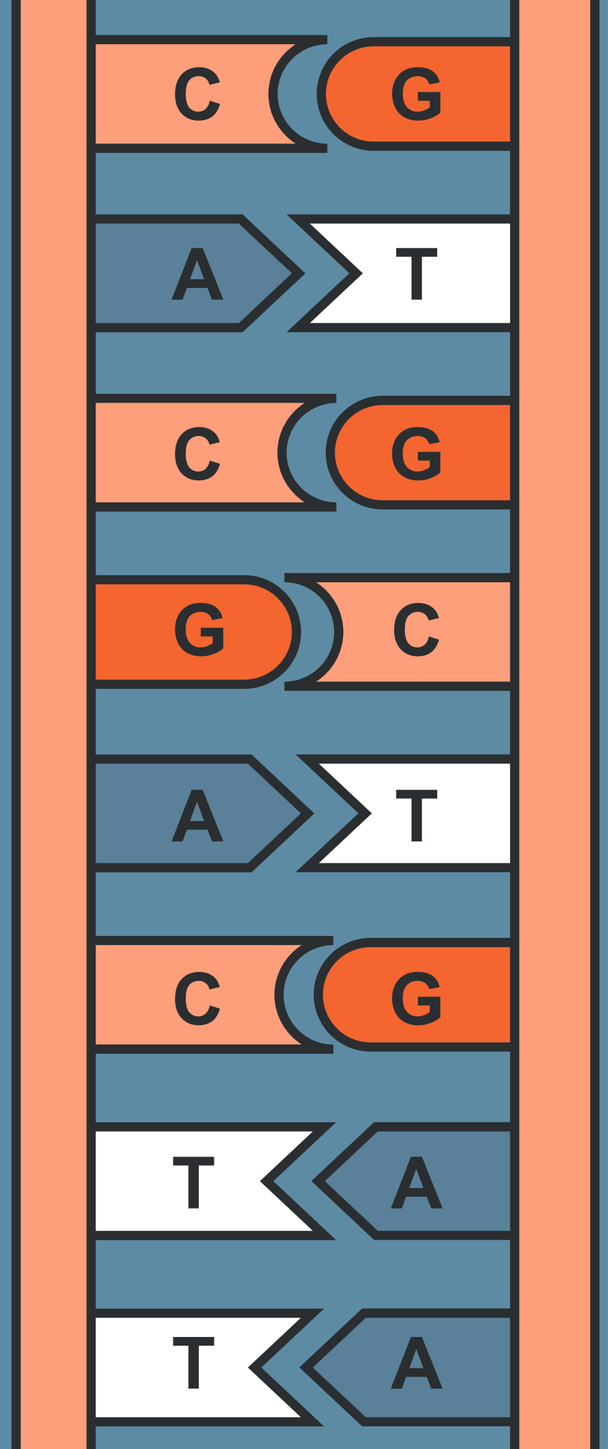
HOW DO WE MEASURE EFOLUTION?

ATGAGGTGCCACGTCGCTTCCAGCTGCCTCGTGGTCTGA

start codon

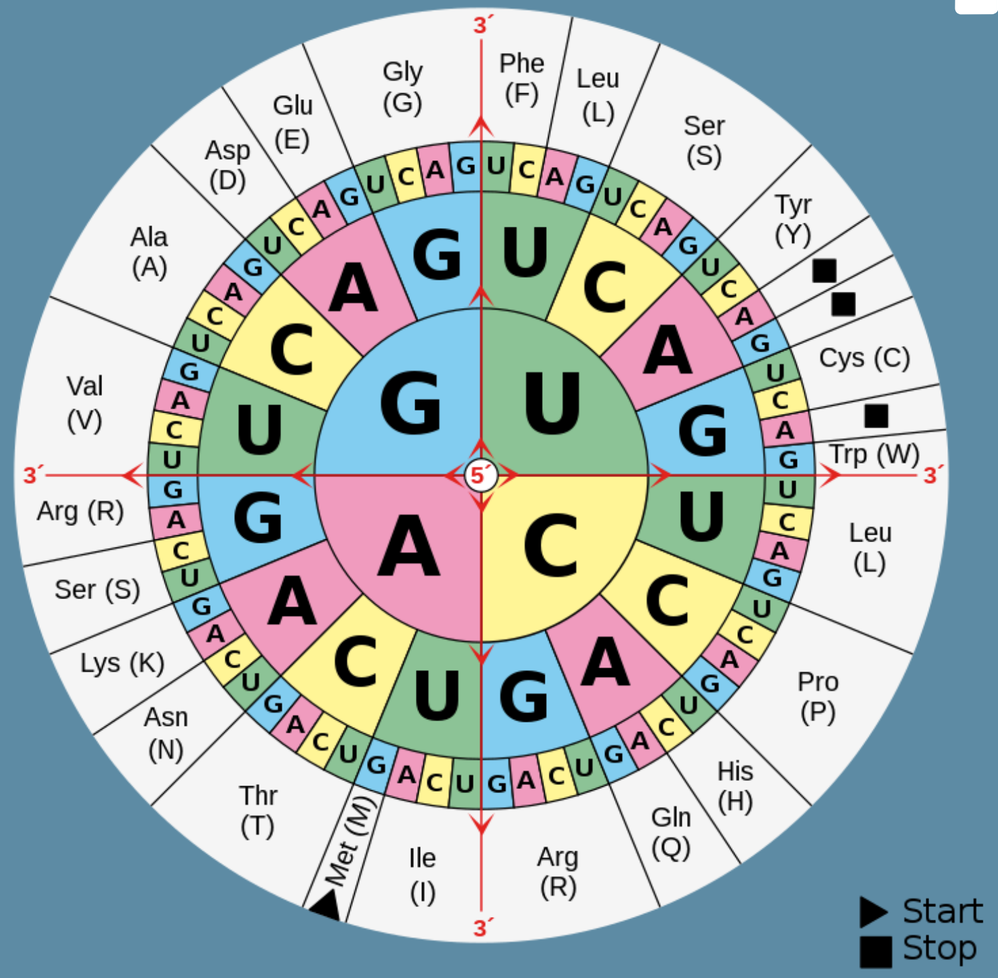
stop codon

Each 3 nucleotides code for one amino acid



THE GENETIC CODE

IS REPETITIVE



This means that different triplets can code for the same amino acid

GGA
GGG } Gly

GCG } Ala

SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?

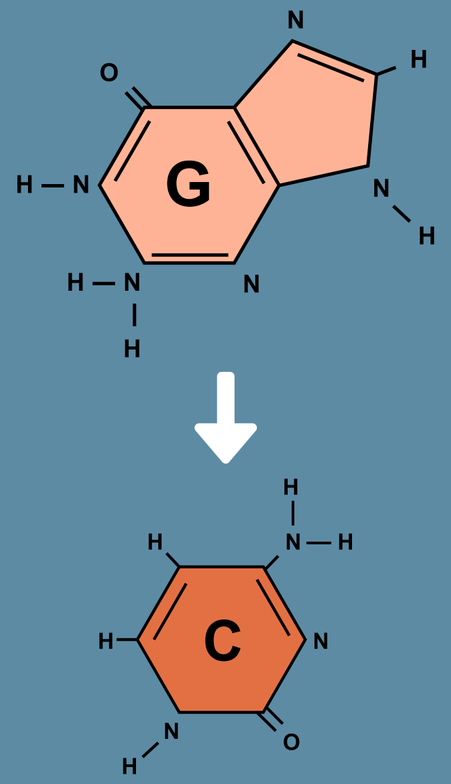
We can have two types of substitution:
synonymous or non-synonymous

GGA → GGG

Thr Thr

GGG → GCG

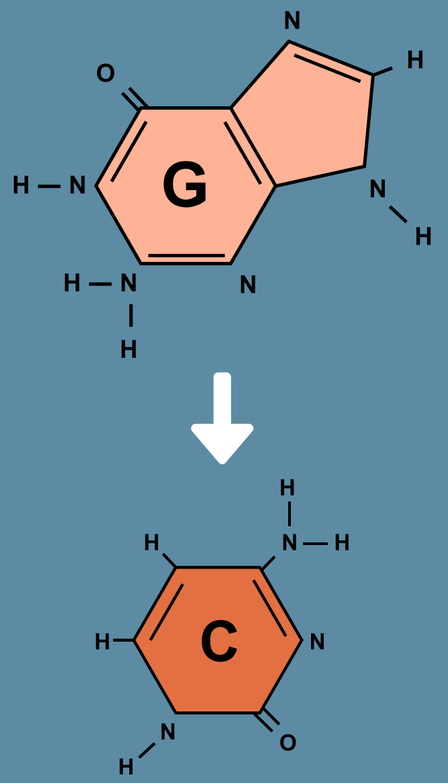
Thr Asn

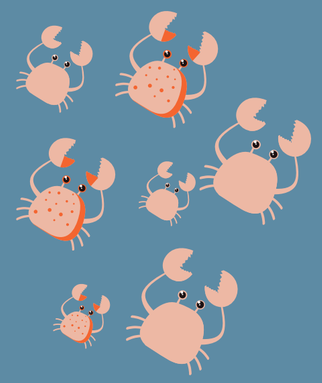


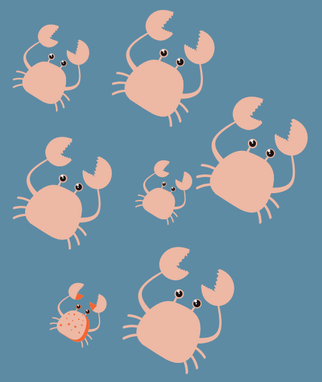
For example, a G (guanine) might be swapped for a C (cytosine).

SO WHAT HAPPENS

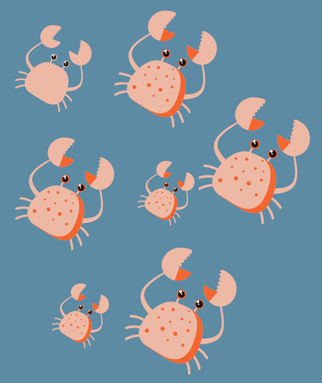
WHEN A NUCLEOTIDE CHANGES?







Increased predation by exposure

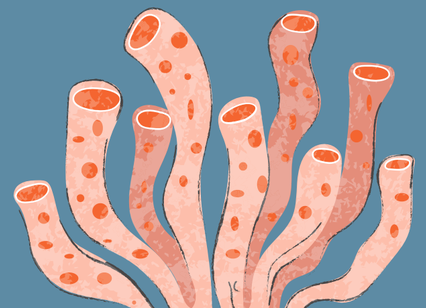


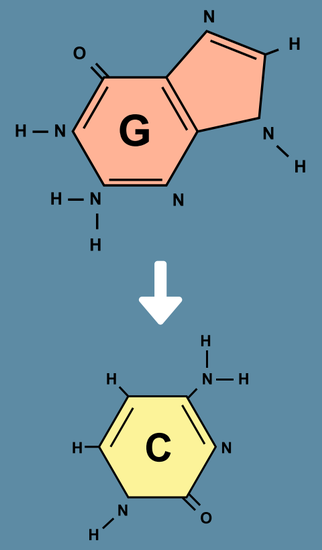
Decreased predation by camouflage or signaling

GGG → GCG

Allele 1 plain

Allele 2 spots

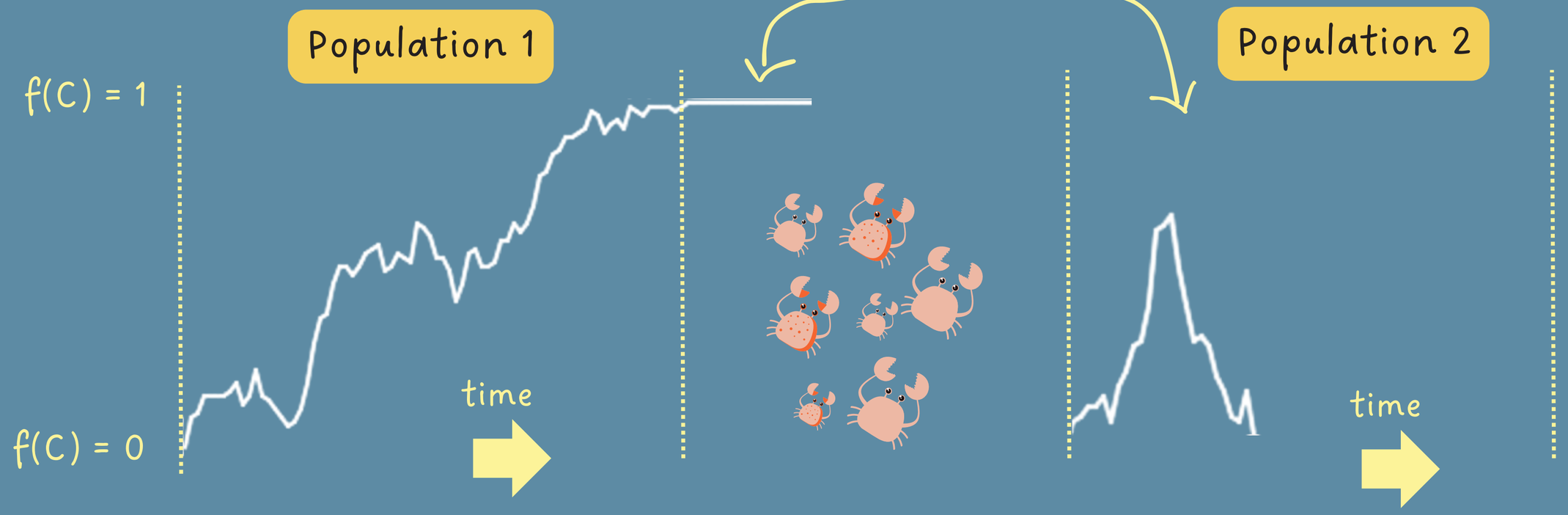


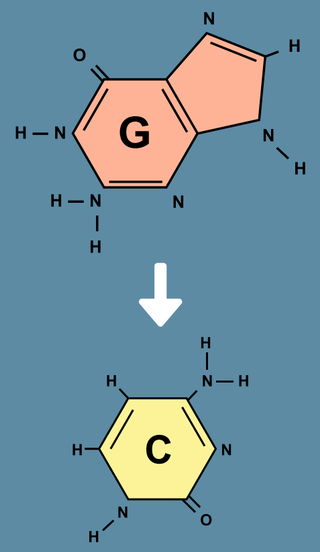


SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?

A mutation can be fixed or lost in a population

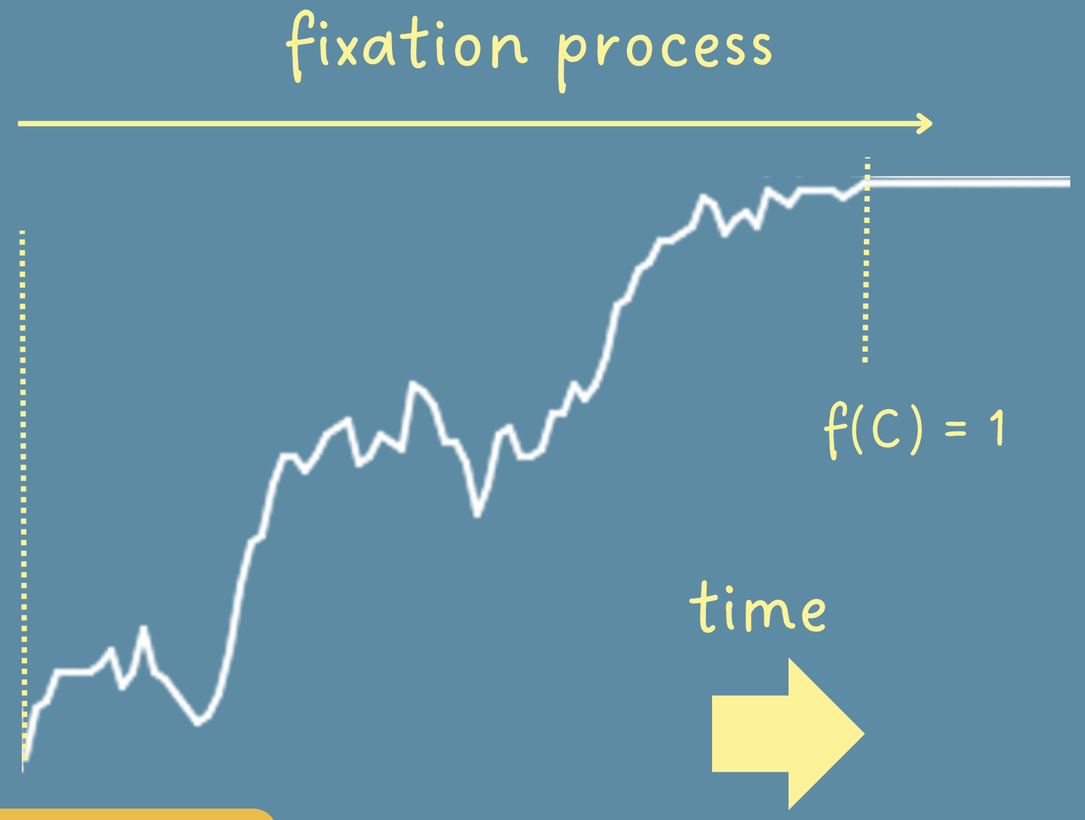




SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?

This graph shows the frequency of the allele C in a population trhough time



Mutation event: G -> C

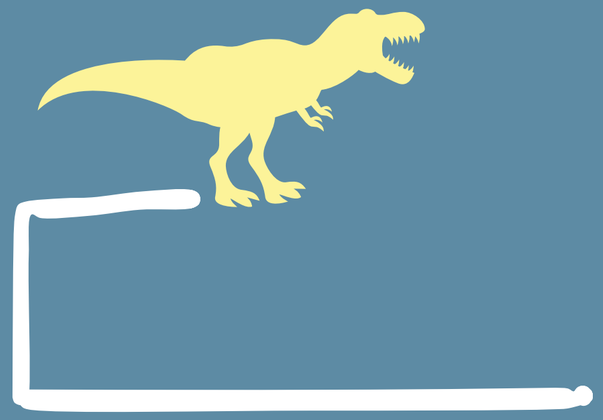
Substitution

the outcome of fixation process

change in "state" of the population (G -> C)

HOW DO WE "SEE" THIS

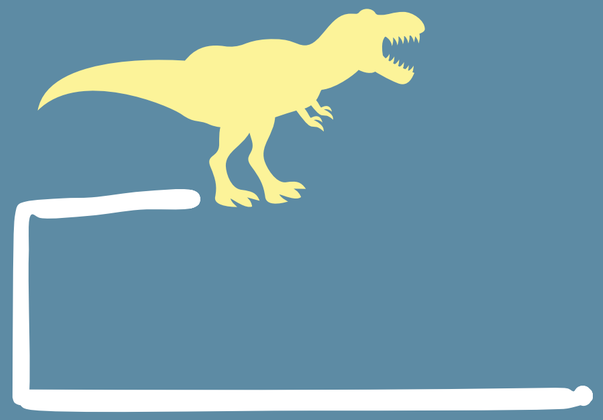
OVER LONG EVOLUTIONARY TIME?



A

HOW DO WE "SEE" THIS

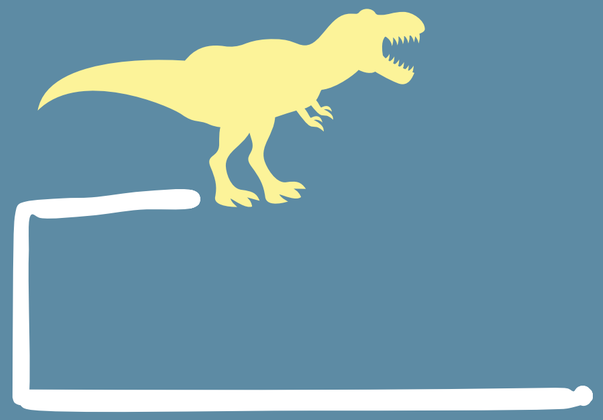
OVER LONG EVOLUTIONARY TIME?



A ⟶ C

HOW DO WE "SEE" THIS

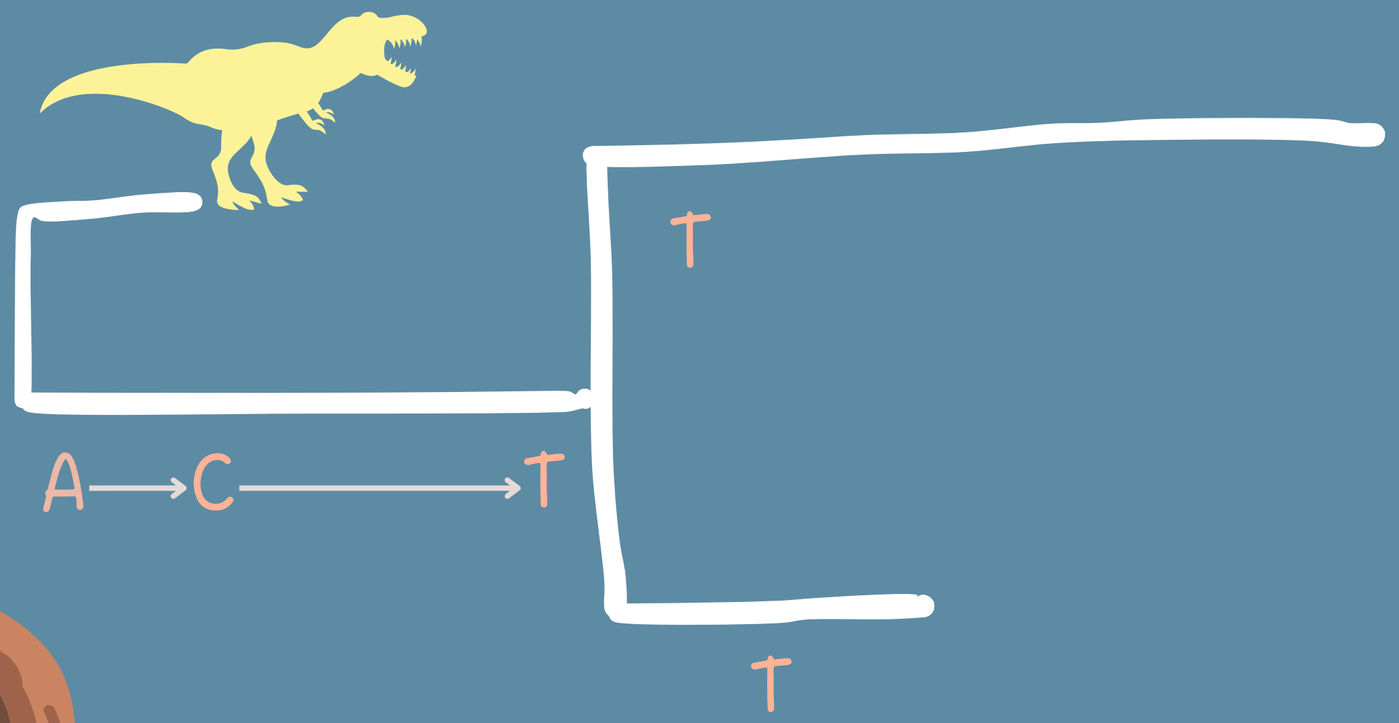
OVER LONG EVOLUTIONARY TIME?



A → C → T

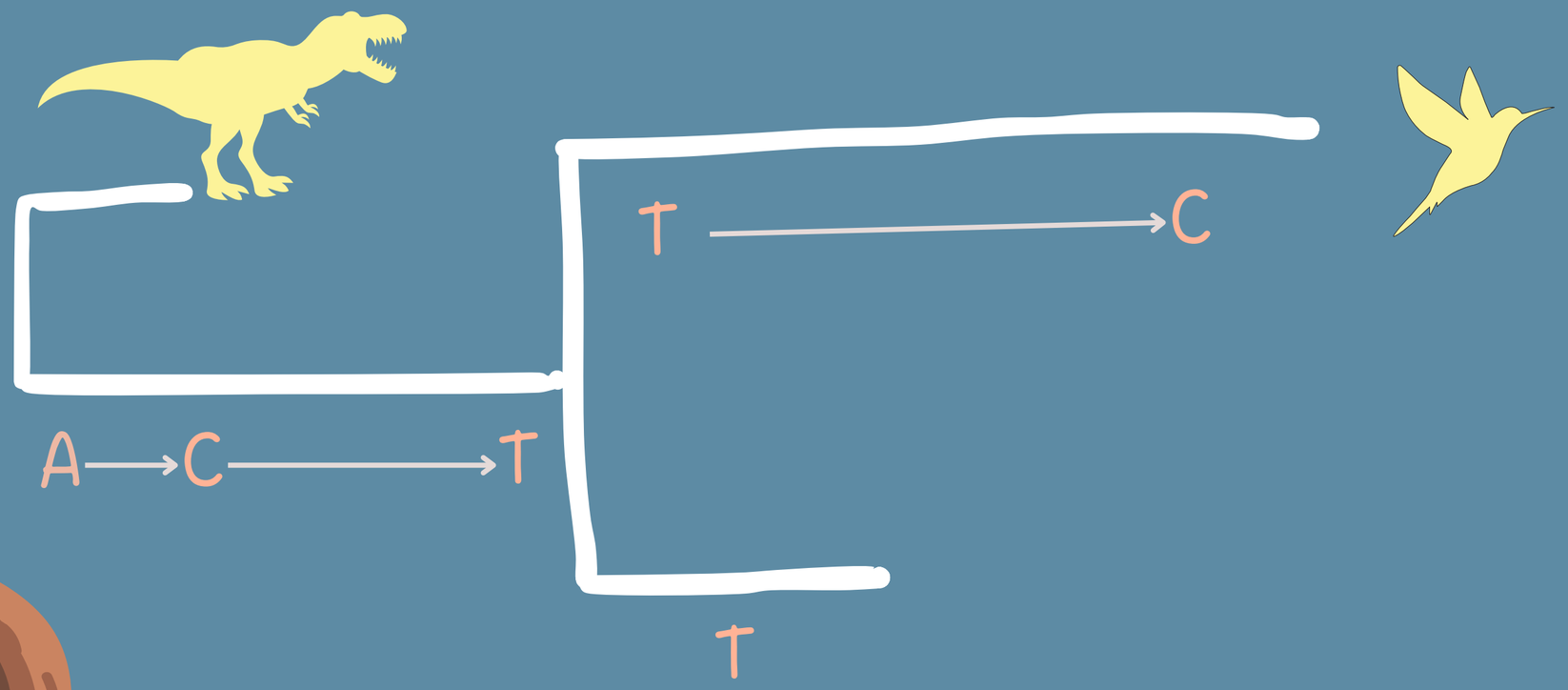
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



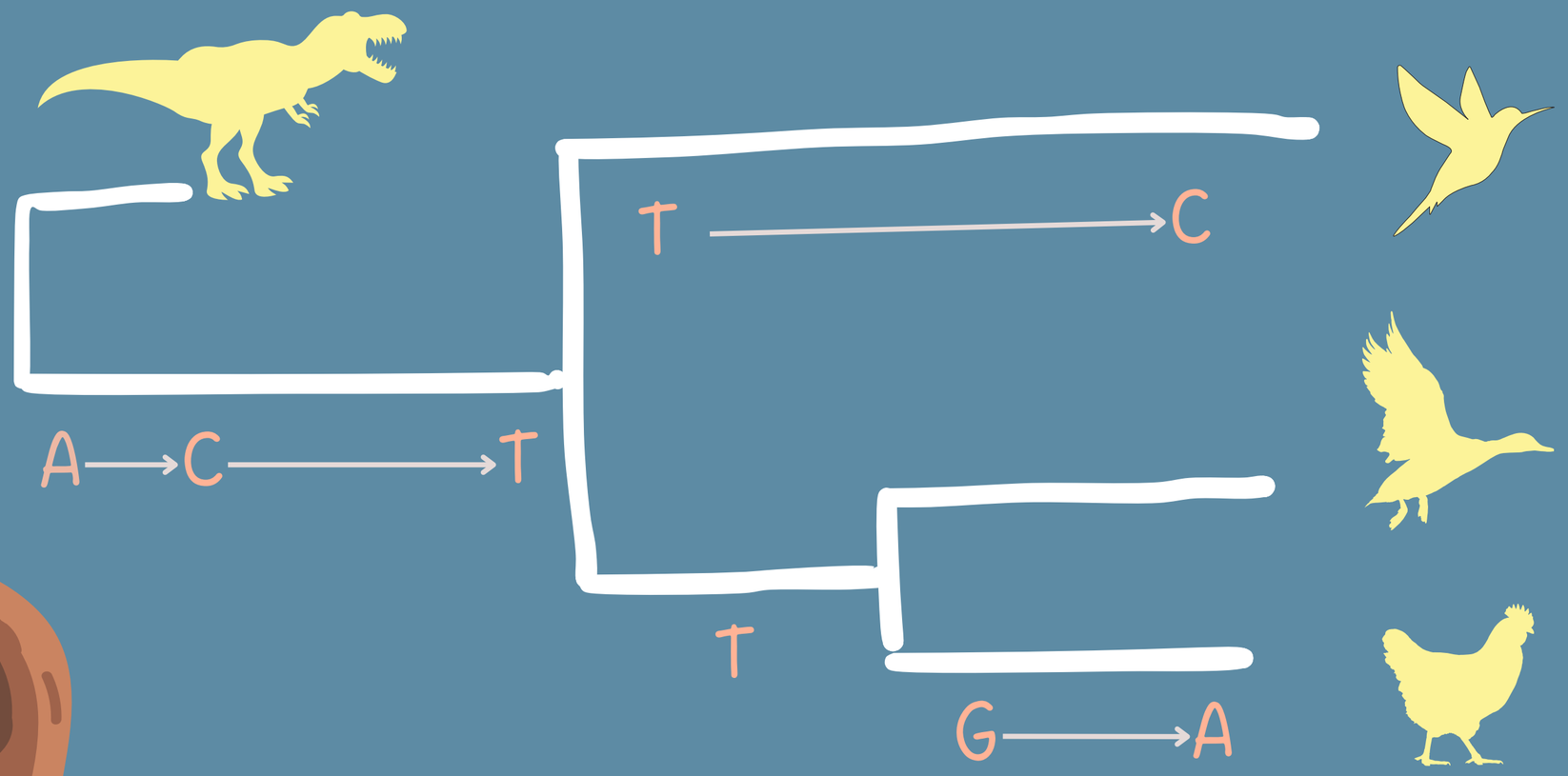
HOW DO WE "SEE" THIS

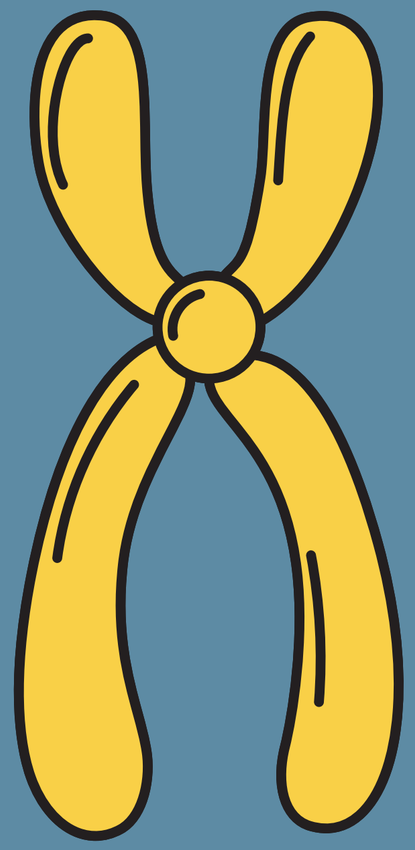
OVER LONG EVOLUTIONARY TIME?



HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?





HOW DO WE MEASURE EVOLUTION ON A SEQUENCE?

ω: a ratio for measuring natural selection

How is it calculated?

dN/dS

dn = rate of non synonymous substitutions
ds = rate of synonymous substitutions

What does it mean?

Selective regimes

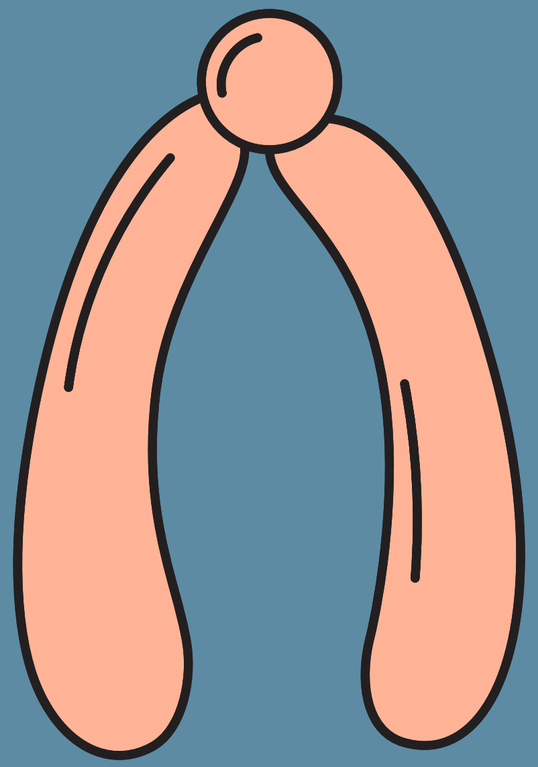
$\omega > 1$ positive (diversifying) selection
$\omega \cong 1$ neutral evolution
$\omega < 1$ negative (purifying) selection

WHAT IS CODEML?

A program that implements codon and amino acid substitution models

↓

Model = potential (mathematical) explanation of how the sequence is evolving



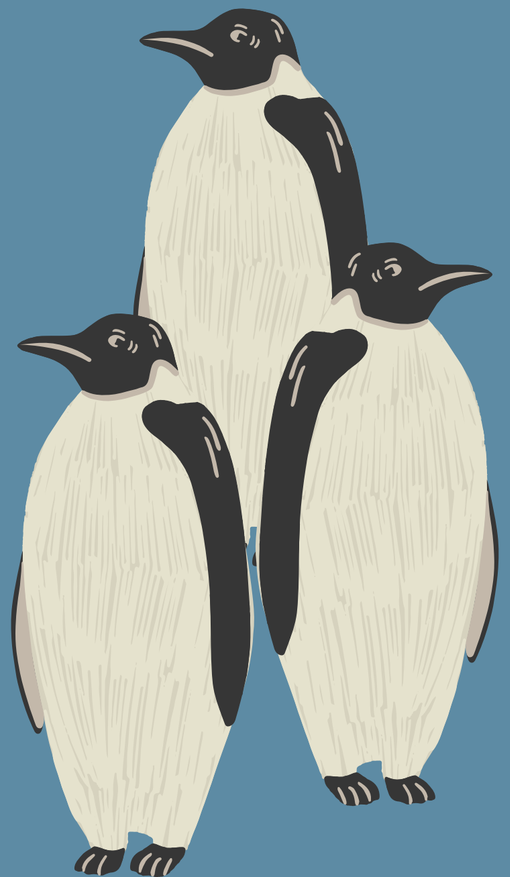
WHAT CAN WE DO
WITH CODEML?

Fit model to data: what does the data tell us about the evolutionary process?

(ex: is high omega or low omega better?)

WHAT CAN WE DO
WITH CODEML?

Hypothesis-testing: does the data allow me to reject the null? (neutral evolution)



WHAT CAN WE DO
WITH CODEML?

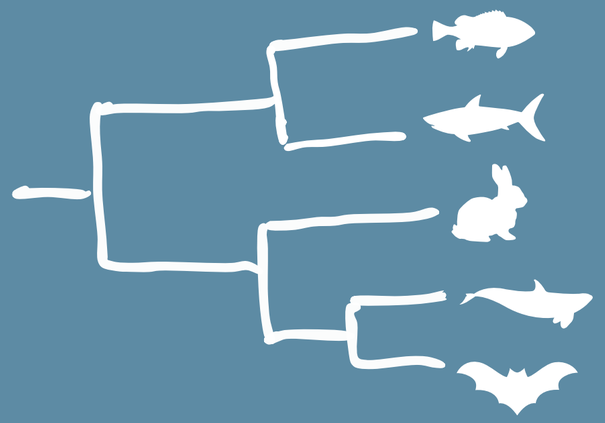
Investigating the signal: which particular sites are under positive selection?

If the null is rejected - given evidence that some sites are evolving adaptively, which ones?

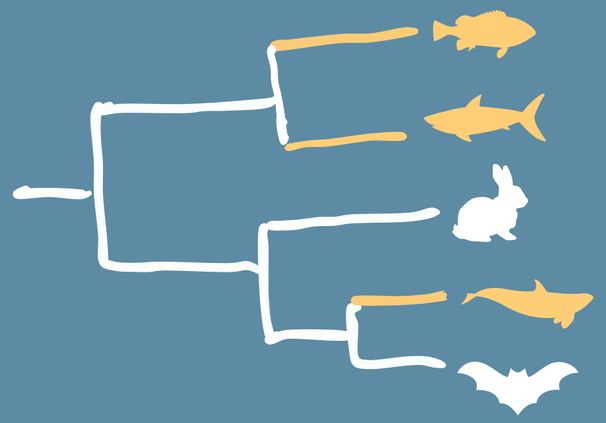
BRANCH-MODELS

Estimate omega across branches only

Model 0



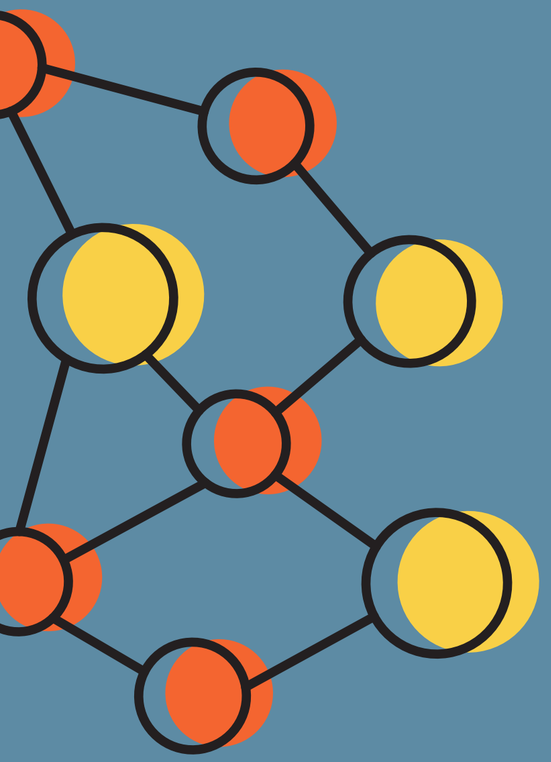
2-model



por exemplo...

ω_1	ω_1	ω_2
0.324	0.324	0.563

SITE-MODELS



Estimate omega across sites only

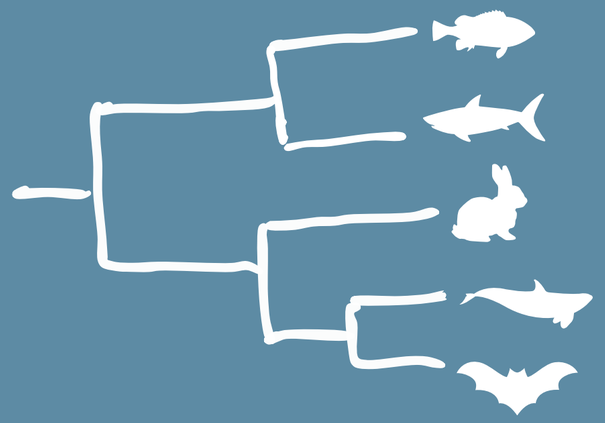
Model M1a

Model M2a

SITE-MODELS

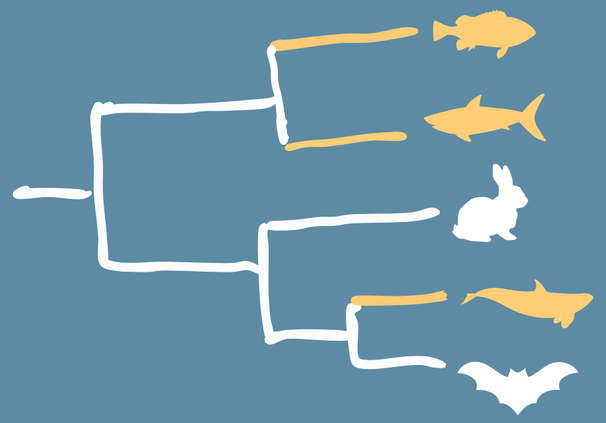
Estimate omega across branches and sites

A-model null



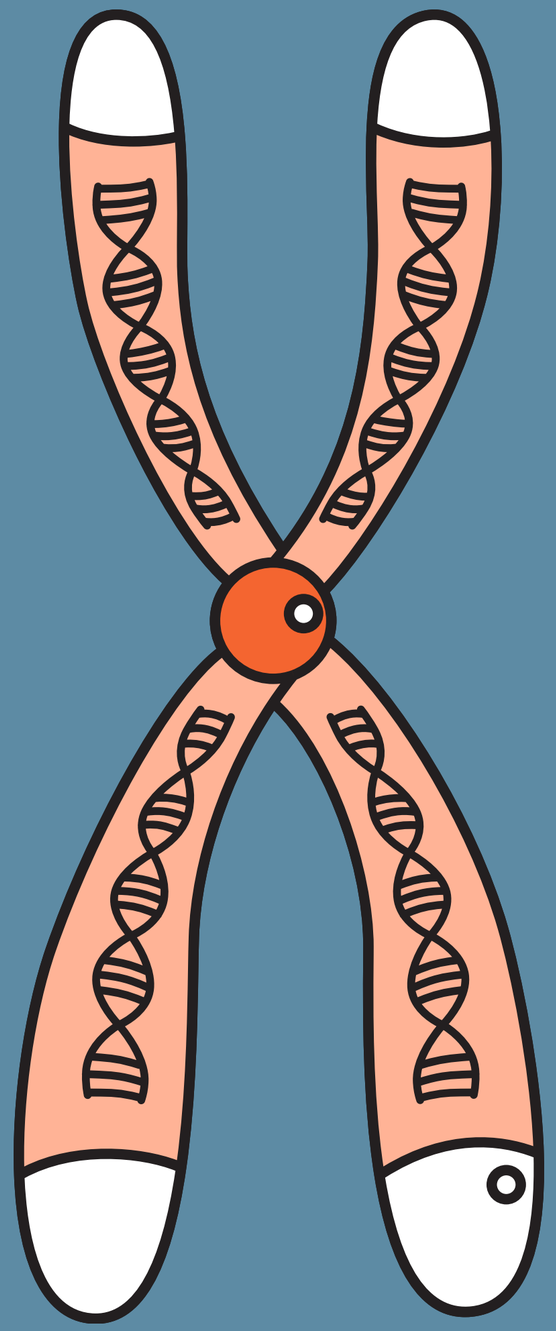
ω_1

A model



ω_1 0.324

ω_2 0.563



HOW TO RUN CODEML?

1. Make a folder for each run containing: an alignment, a gene tree and the control file

2. Make sure you don't have any gaps, stop codons & that names match between alignment and tree

3. Run the control file from inside your folder with the command codeml control_file.ctl

HOW TO GET AN ALIGNMENT

MAFFT

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

RIMD

Download version
- Mac OS X
- Windows
- Linux
- Source

Online version
- **Alignment**
- mafft --add
- Merge
- Phylogeny
- Rough tree

Merits / limitations
Algorithms
Tips
Benchmarks
Feedback
Follow

This service was unstable due to maintenance, 18:00 – 21:00, May 23, JST.

To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

If you need an MSA of only a specific region, then try extracting the region first (2022/Oct). ***New!***

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:
Paste protein or DNA sequences in fasta format. Example

or upload a **plain text** file: Choose File No file chosen

☐ Use DASH to add homologous structures (protein only)

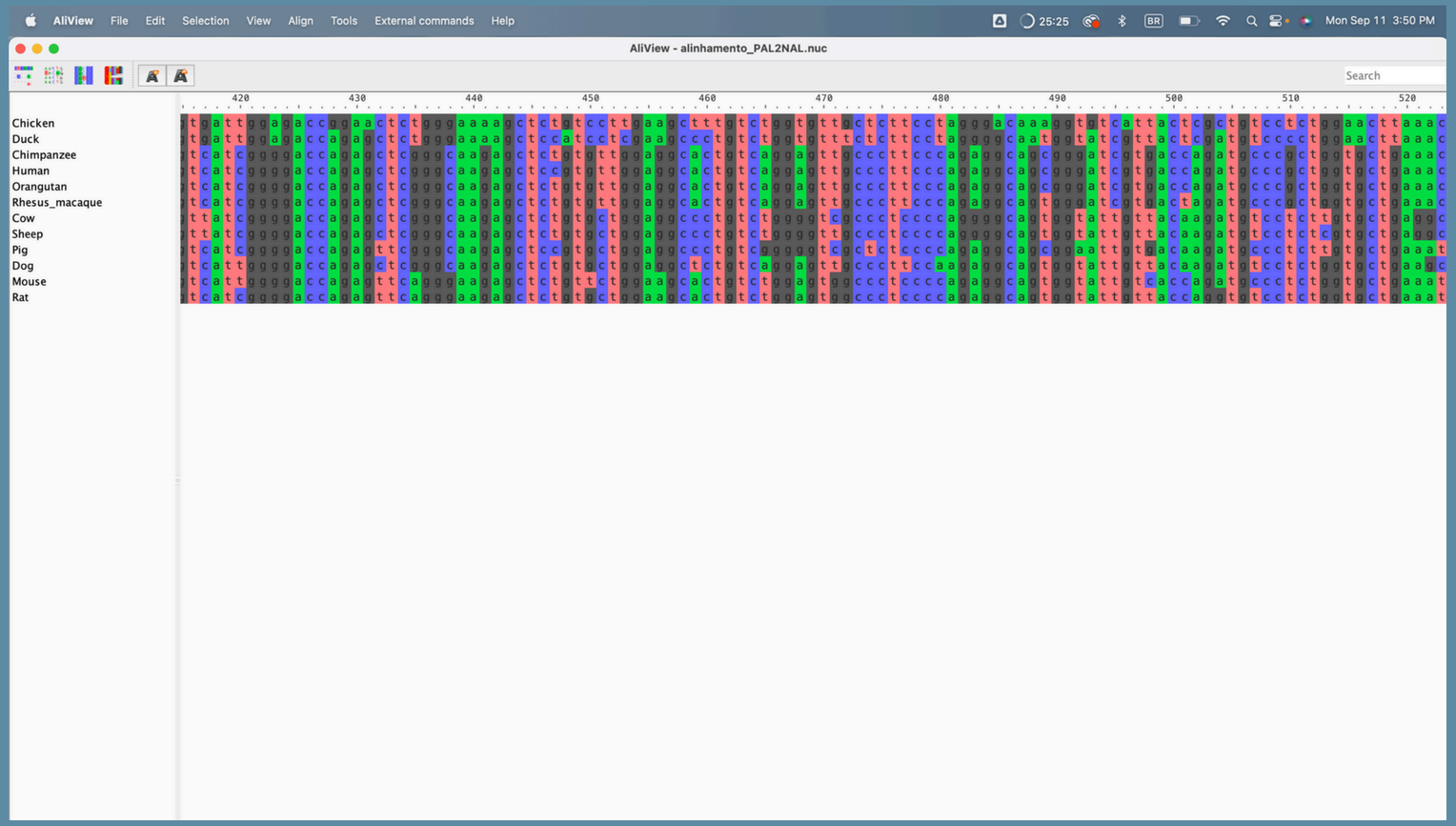
◉ Ouput original plus DASH sequences ○ Output original sequences only

☐ Give structural alignment(s) externally prepared

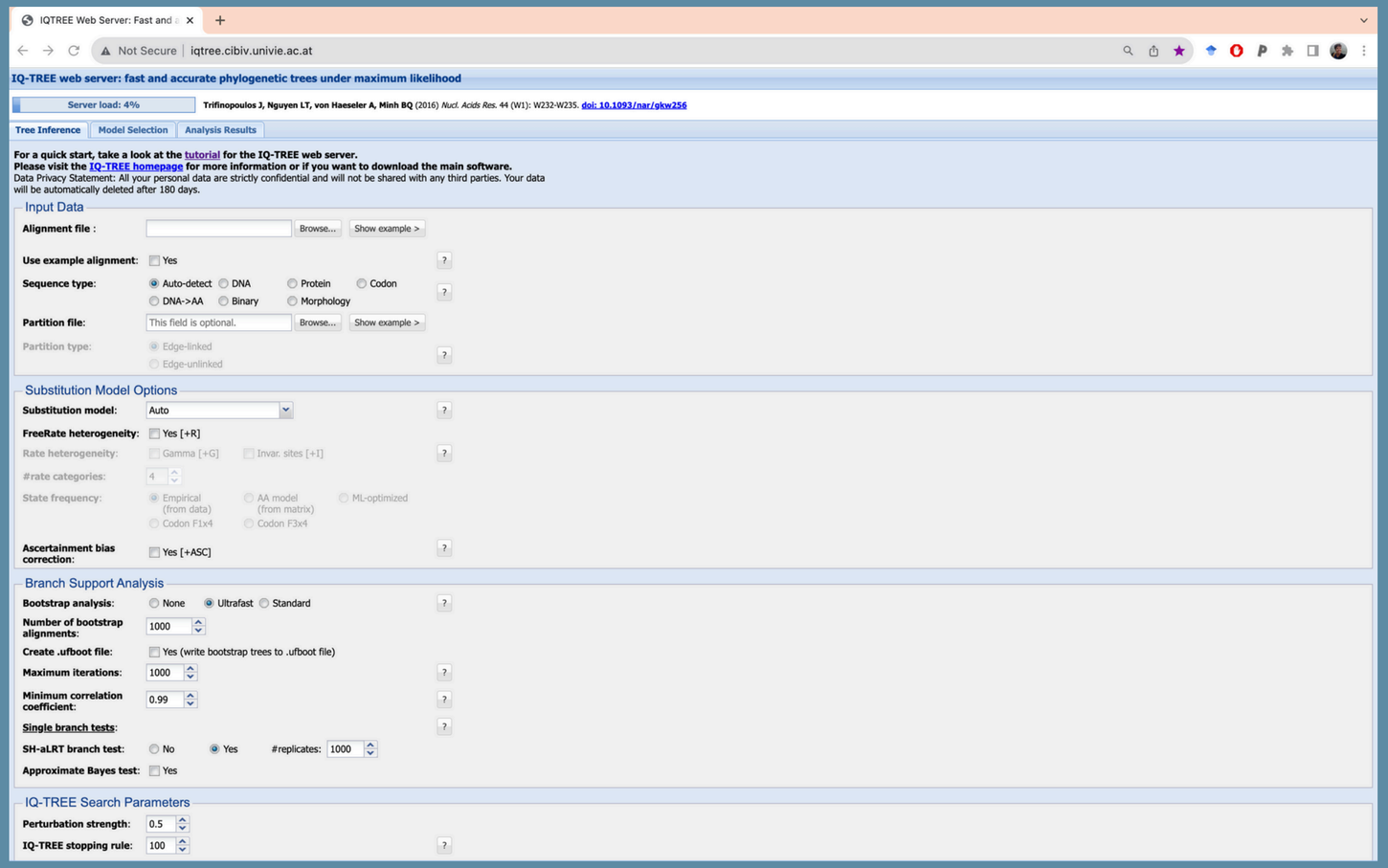
☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, *etc.*) Help

HOW TO GET AN ALIGNMENT

ALIVIEW



HOW TO GET A TREE

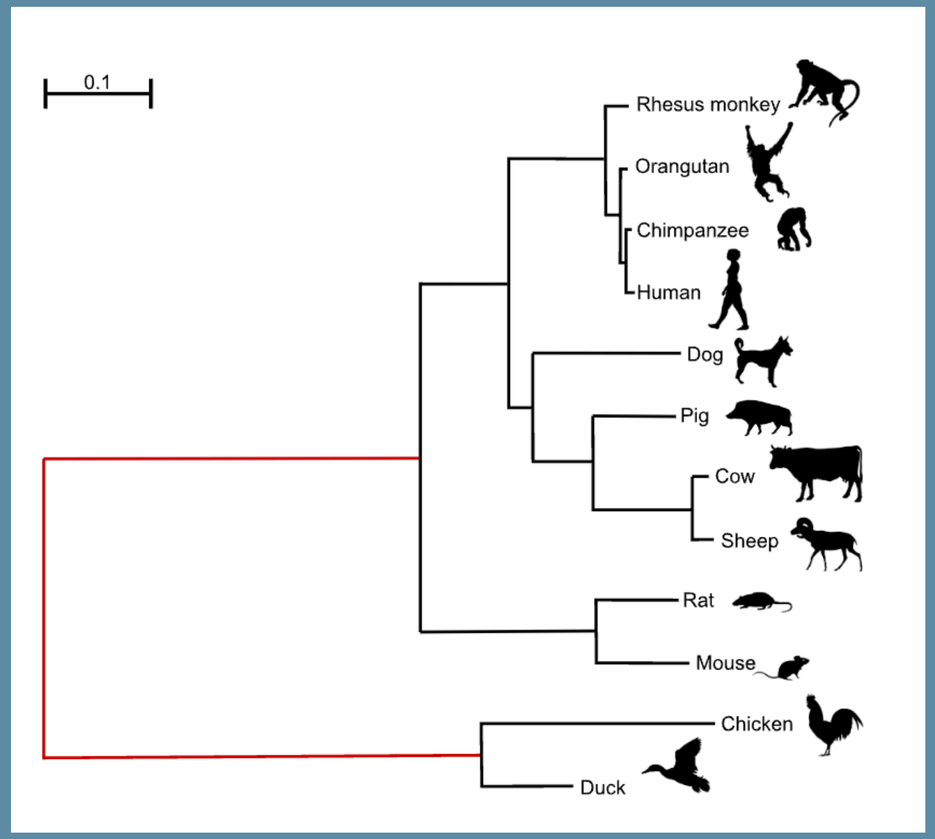


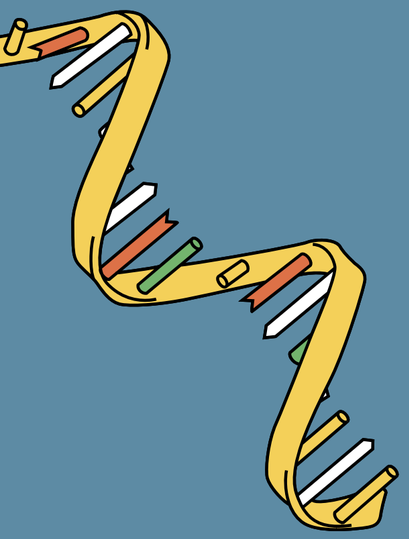
HOW DO WE USE CODEML

TO TEST A HYPOTHESIS?

Questions:

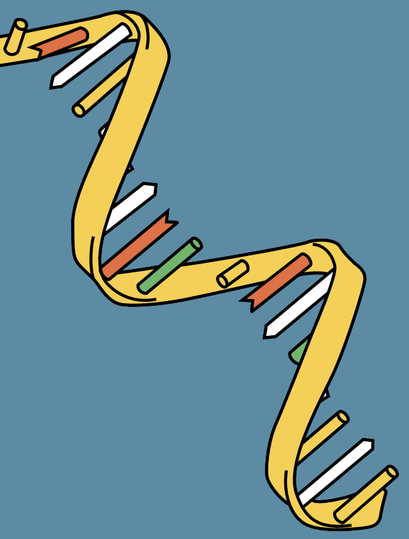
- Did Mx evolved in these species to inhibit a set of species-specific pathogens?
- What factors drive Mx evolution in different animal lineages?





THE CODEML CONTROL FILE

```
seqfile   = Mx_aln.phy          * Path to the alignment file
treefile = Mx_unroot.tree       * Path to the tree file
outfile = out_M0.txt            * Path to the output file
noisy  = 3                      * Display moderate information on the screen
verbose = 1                     * Detailed output file
seqtype = 1                     * Codon data
Ndata  = 1                      * One gene alignment
icode  = 0                      * Universal genetic code
cleandata = 0                   * Do not remove sites with ambiguity data
model  = 0                      * One ω for all branches (M0 and site models)
NSsites = 0                     * One ω for all sites (M0 and branch model)
CodonFreq = 2                   * Use F3x4 model
clock  = 0                      * Assume no clock
fix_omega = 0                   * Enables option to estimate omega
omega  = 0.5                    * Initial omega value
```



THE CODEML CONTROL FILE

seqfile = Mx_aln.phy

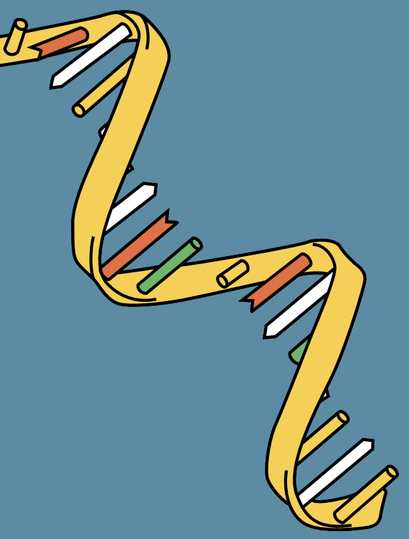
change this with the name or path to your sequence

treefile = Mx_unroot.tree

change this with the name or path to your tree

outfile = out_M0.txt

change this with the name or path to your output file



THE CODEML CONTROL FILE

model = 0

change this number to specify the branch model you want to run

NSsites = 0

change this number to specify the site model you want to run

When you change both at the same time, you can specify a third type of model: the branch-site!