



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`



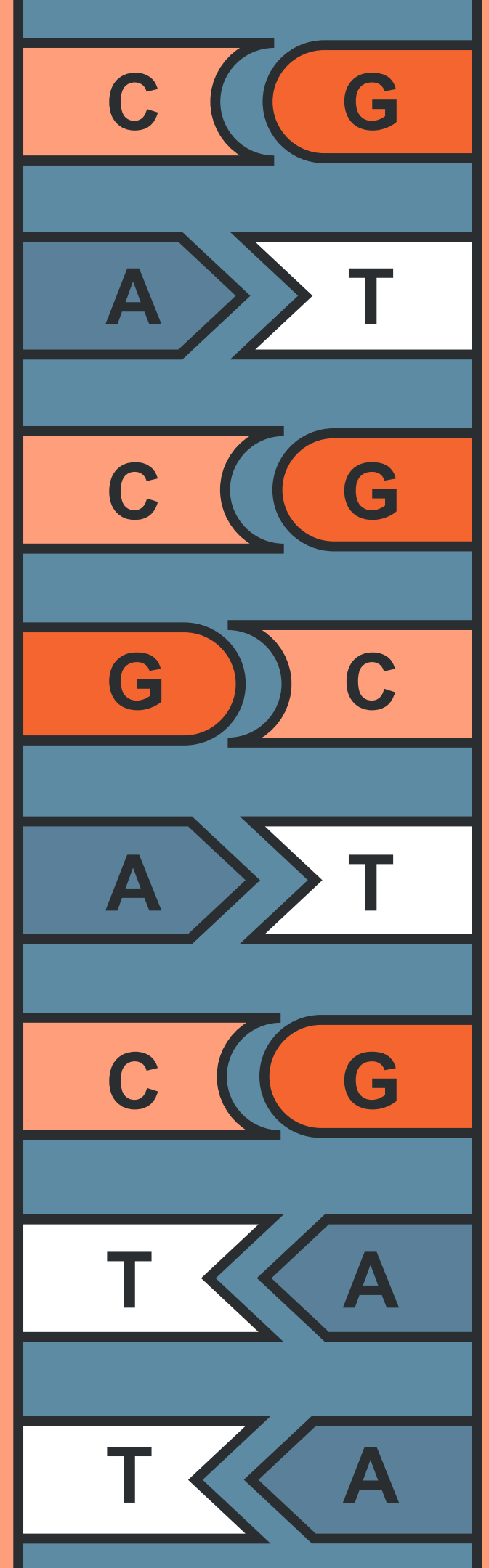
HOW DO WE MEASURE EVOLUTION?

ATGAGGTGCCACGTCGCTTCCAGCTGCCTCGTGGTCTGA

start codon

stop codon

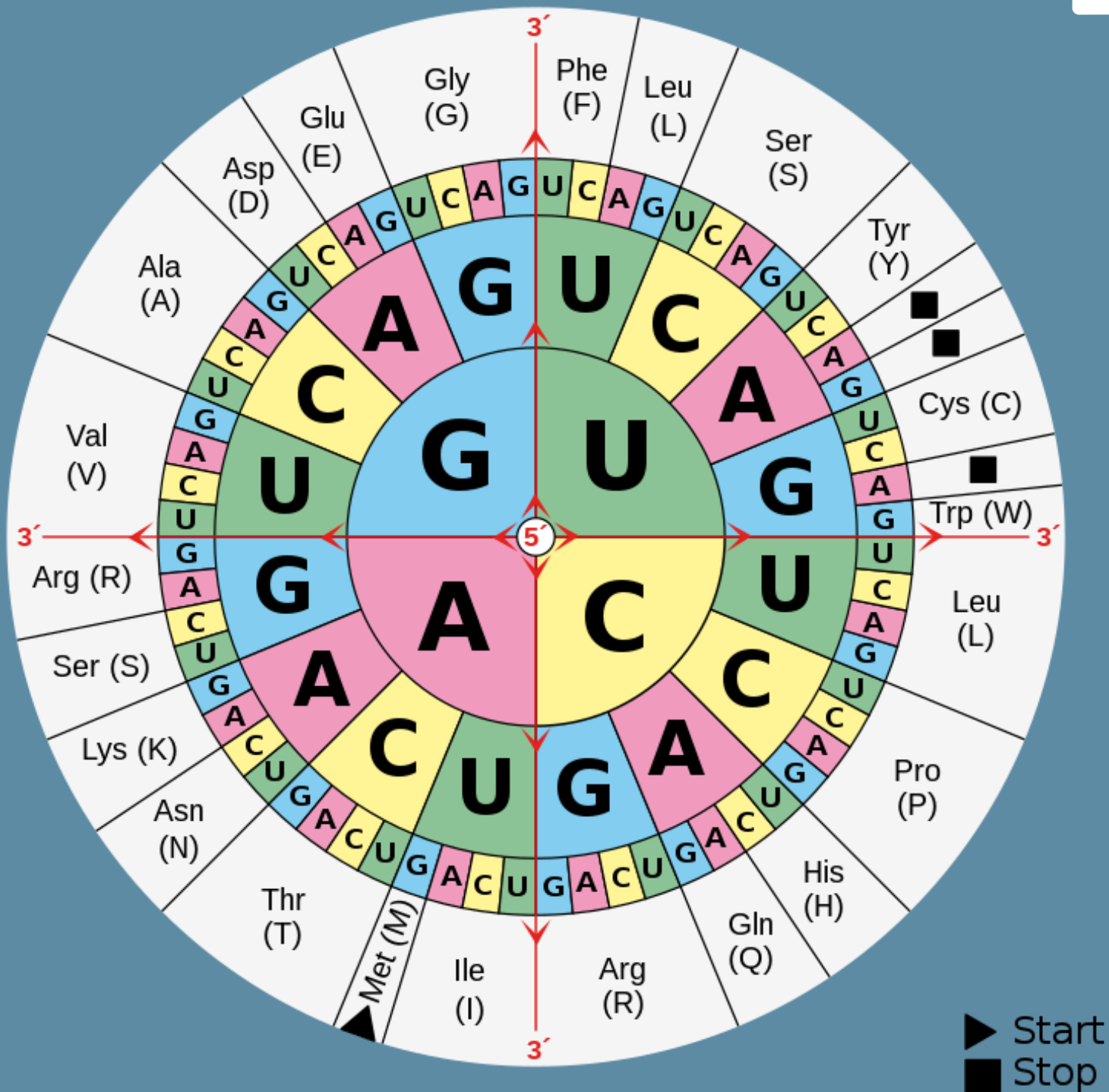
Each 3 nucleotides code for one amino acid



THE GENETIC CODE

IS REPETITIVE

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



GGA } Gly
GGG }
GCG } Ala

SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?

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

GGA → GGG

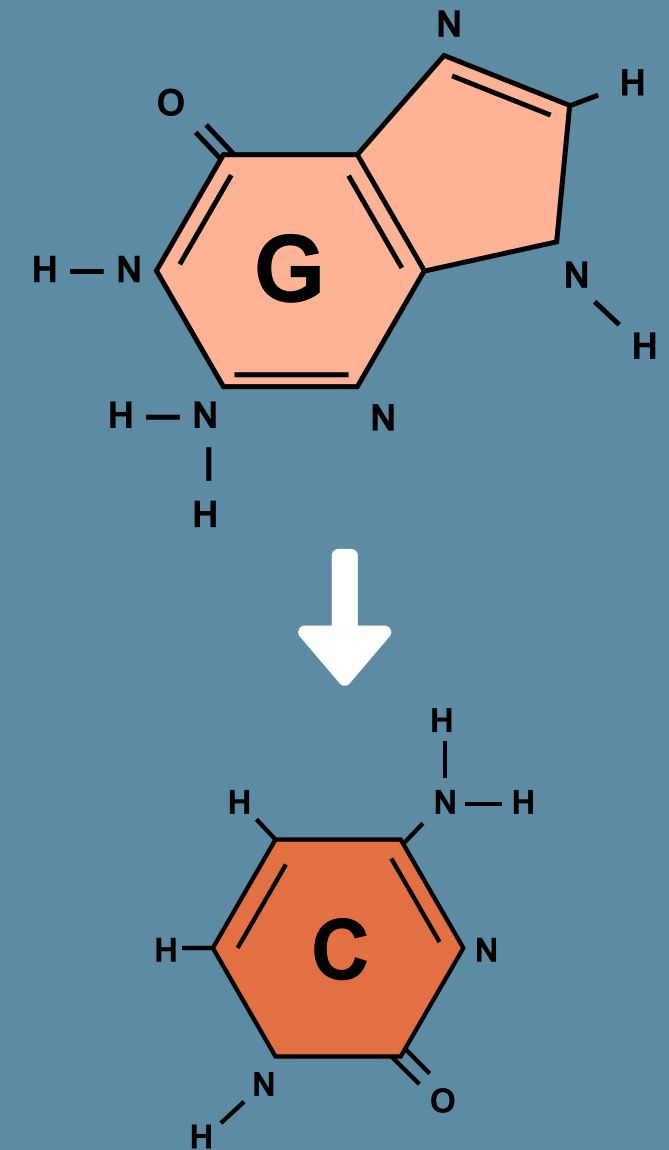
Thr

Thr

GGG → GCG

Thr

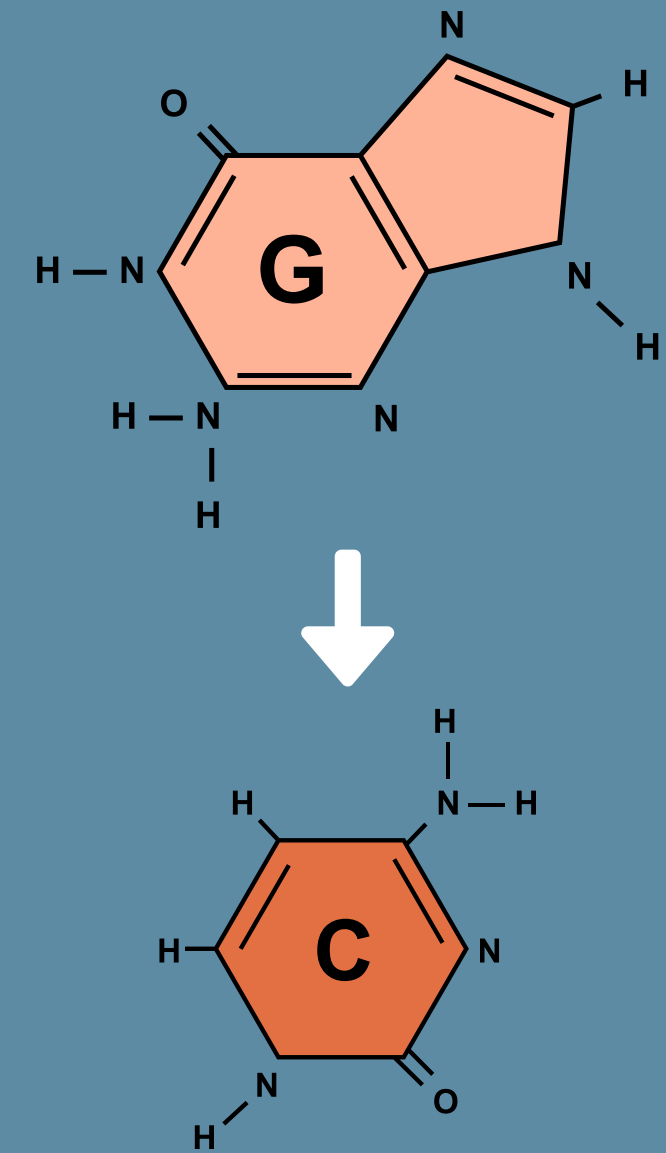
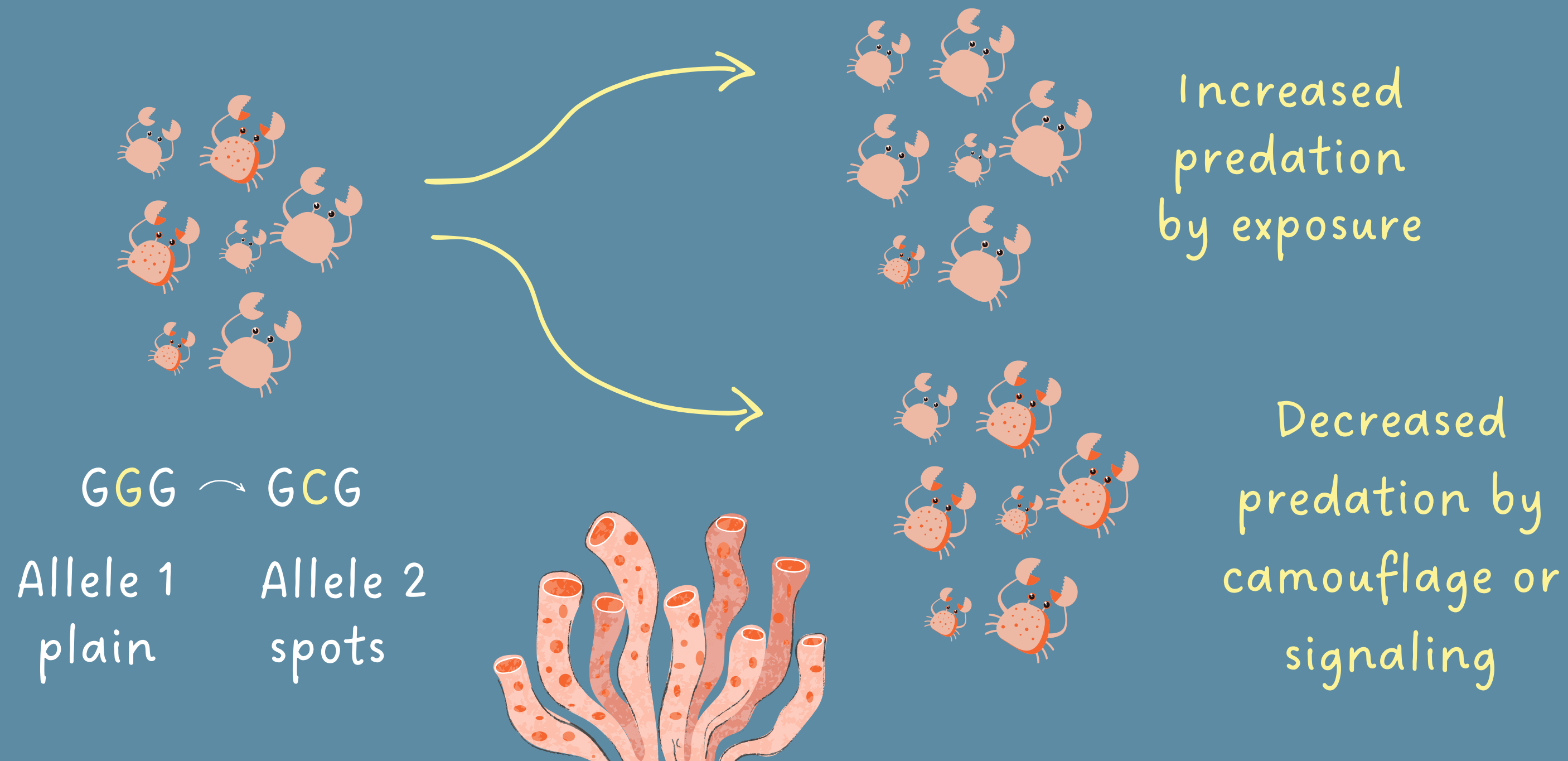
Asn

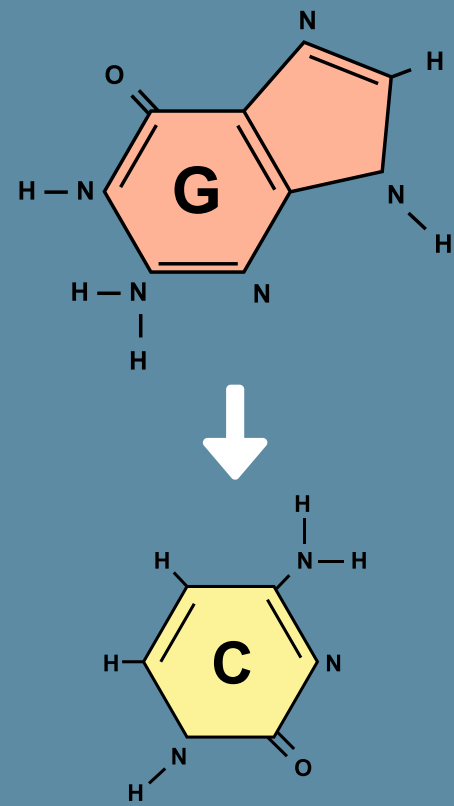


For example, a G (guanine) might be swapped for a C (cytosine).

SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?

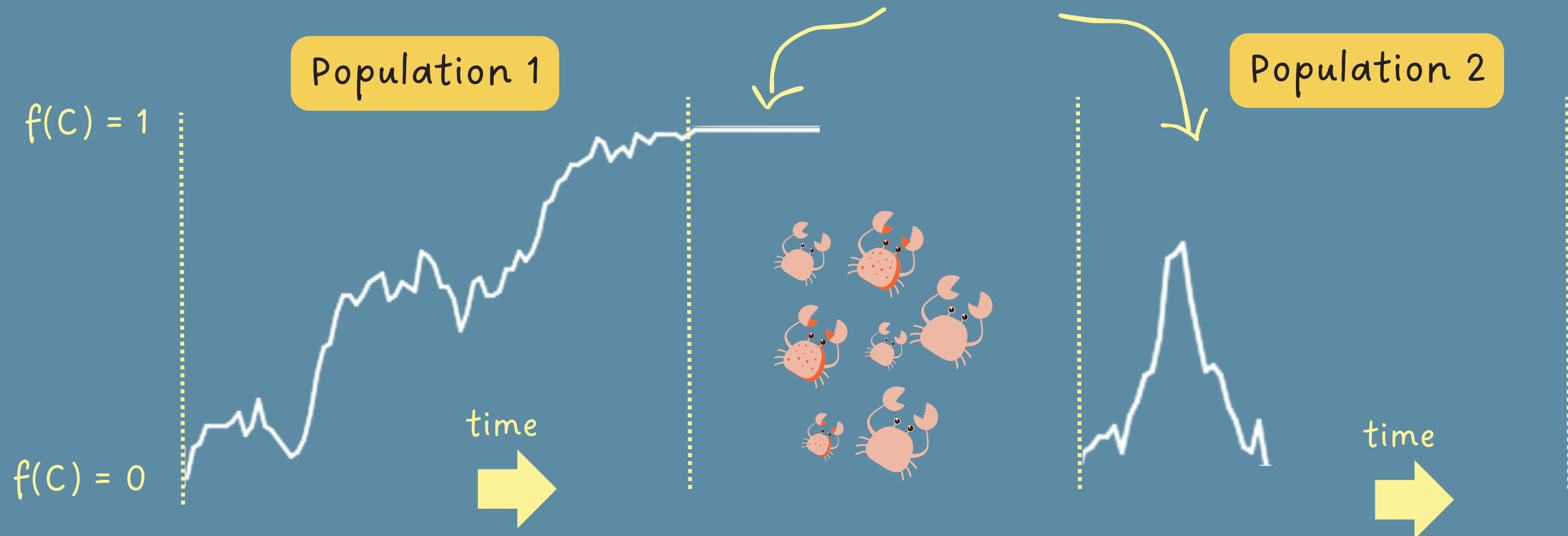




SO WHAT HAPPENS

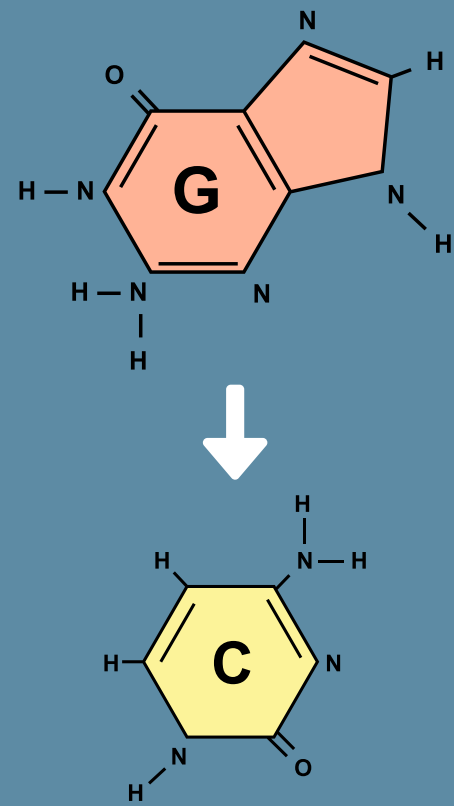
WHEN A NUCLEOTIDE CHANGES?

A mutation can be fixed or lost in a population



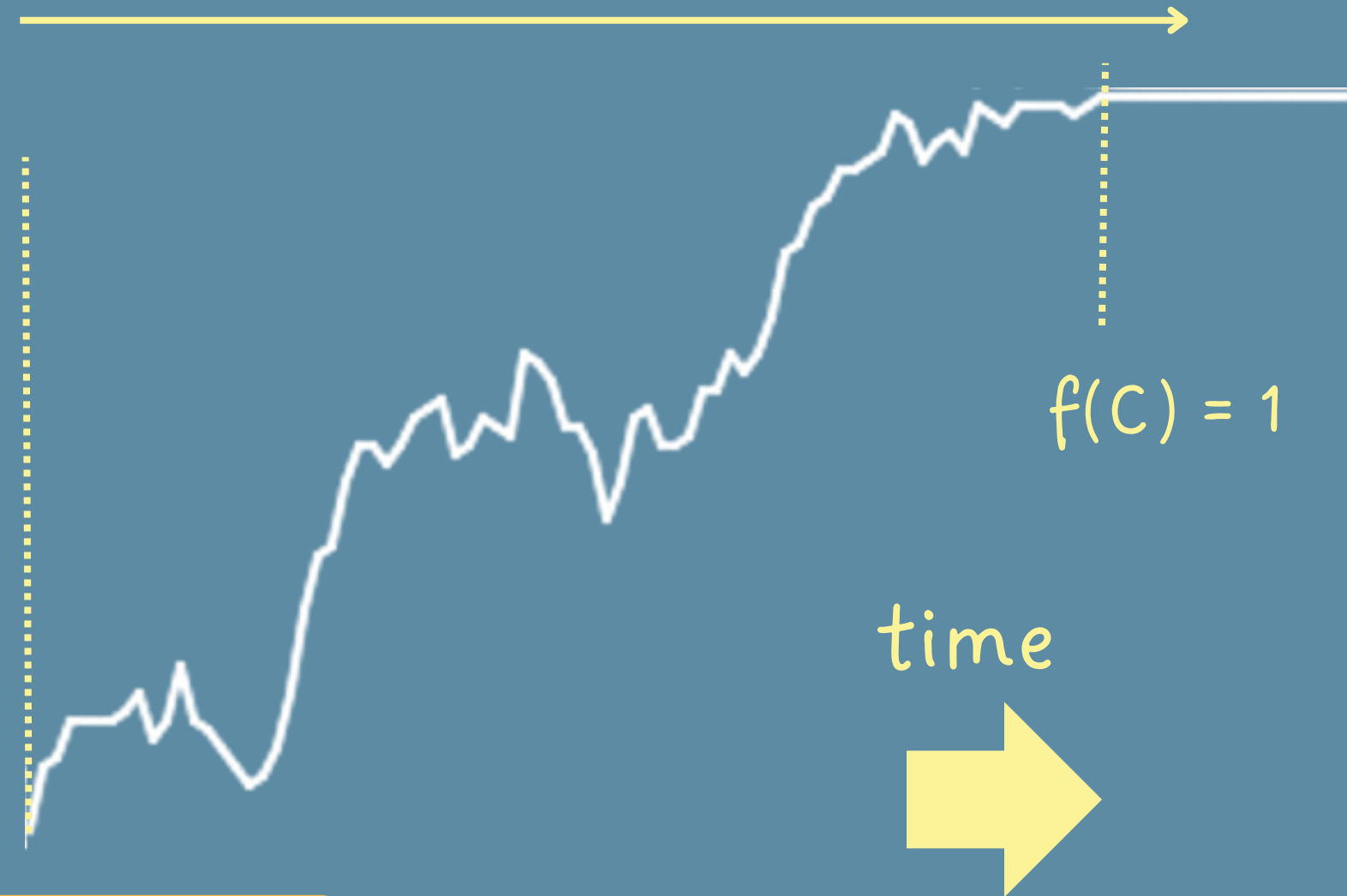
SO WHAT HAPPENS

WHEN A NUCLEOTIDE CHANGES?



fixation process

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



Substitution

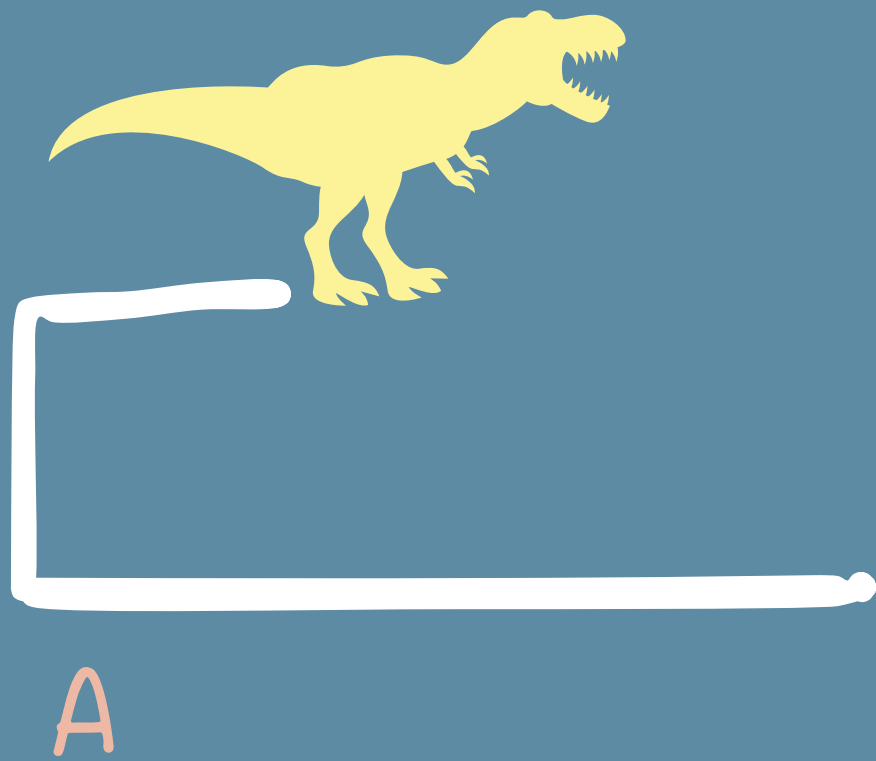
the outcome of
fixation process

change in "state"
of the population
(G → C)

Mutation event: G → C

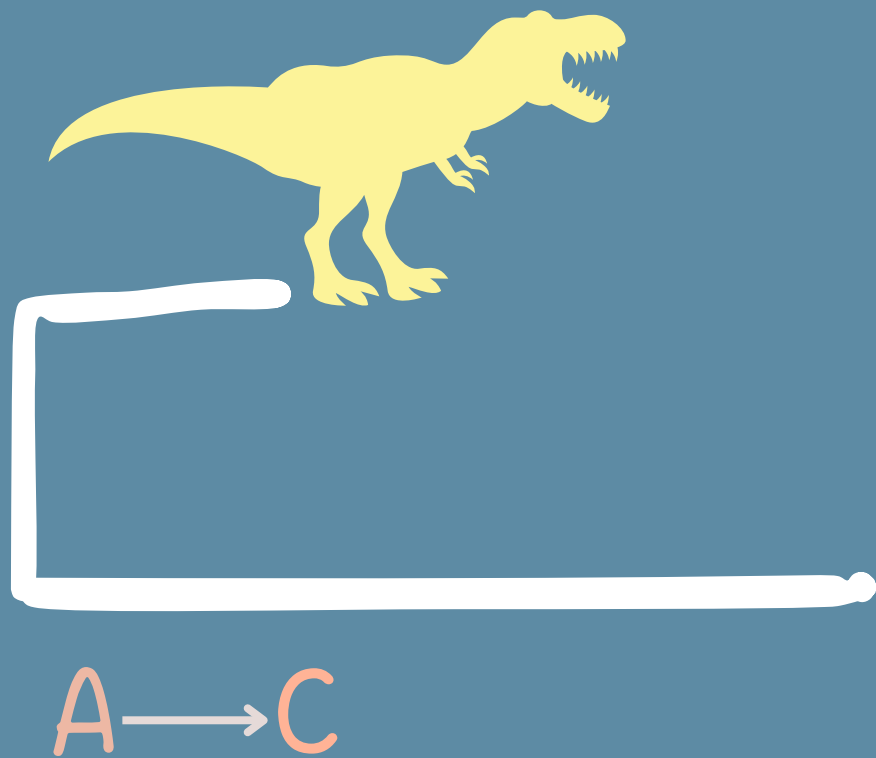
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



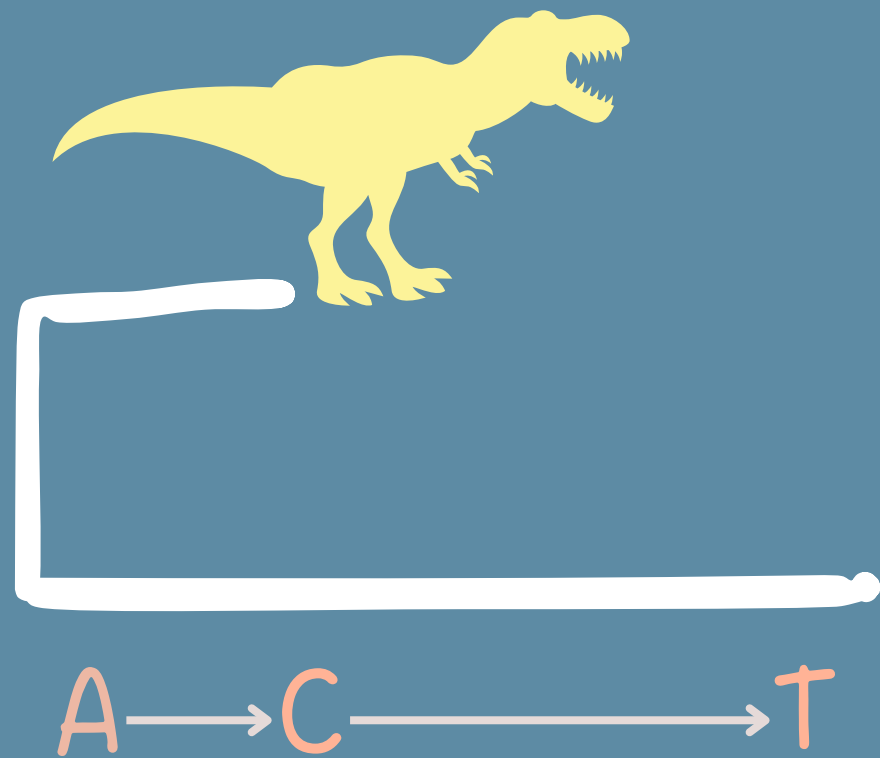
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



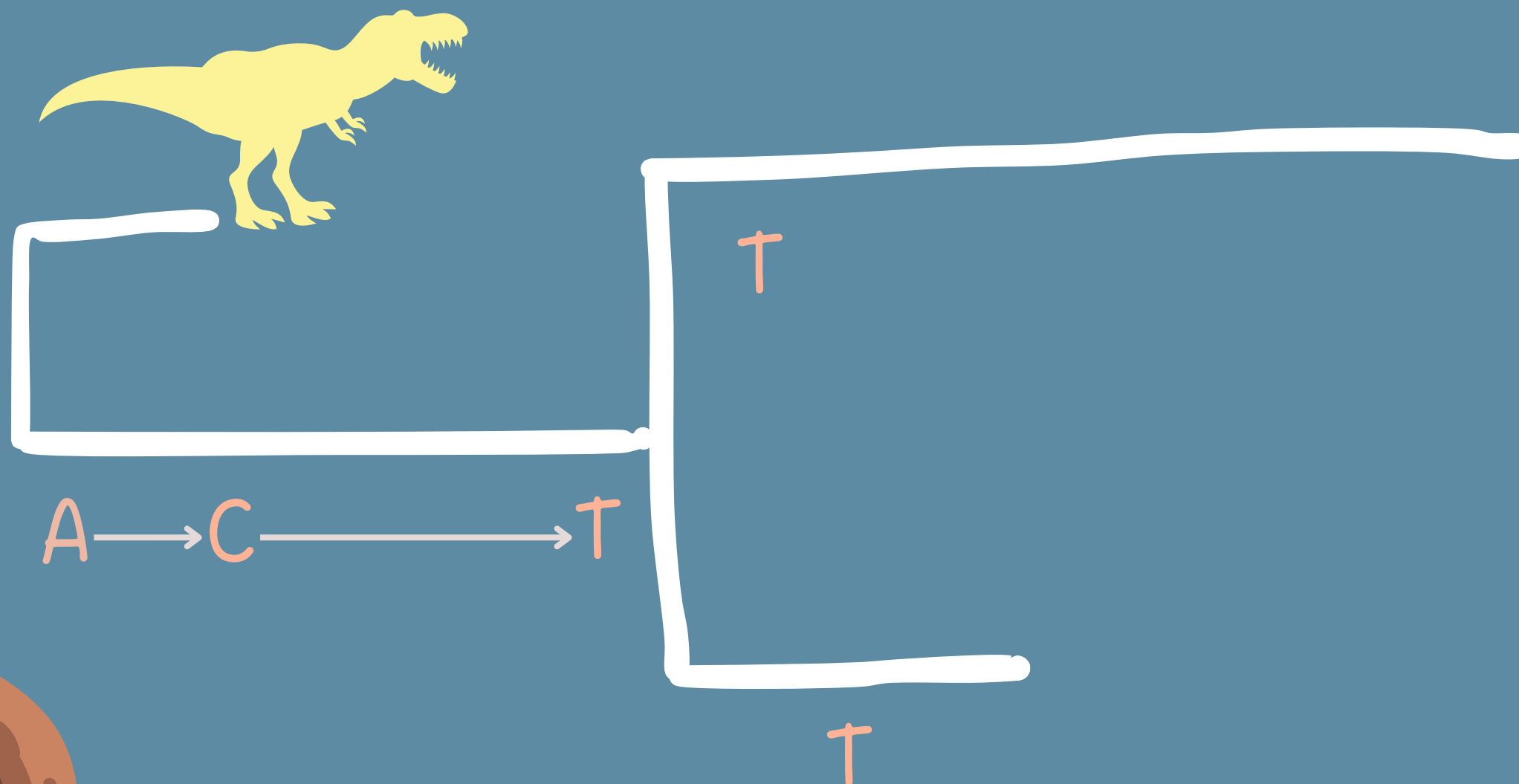
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



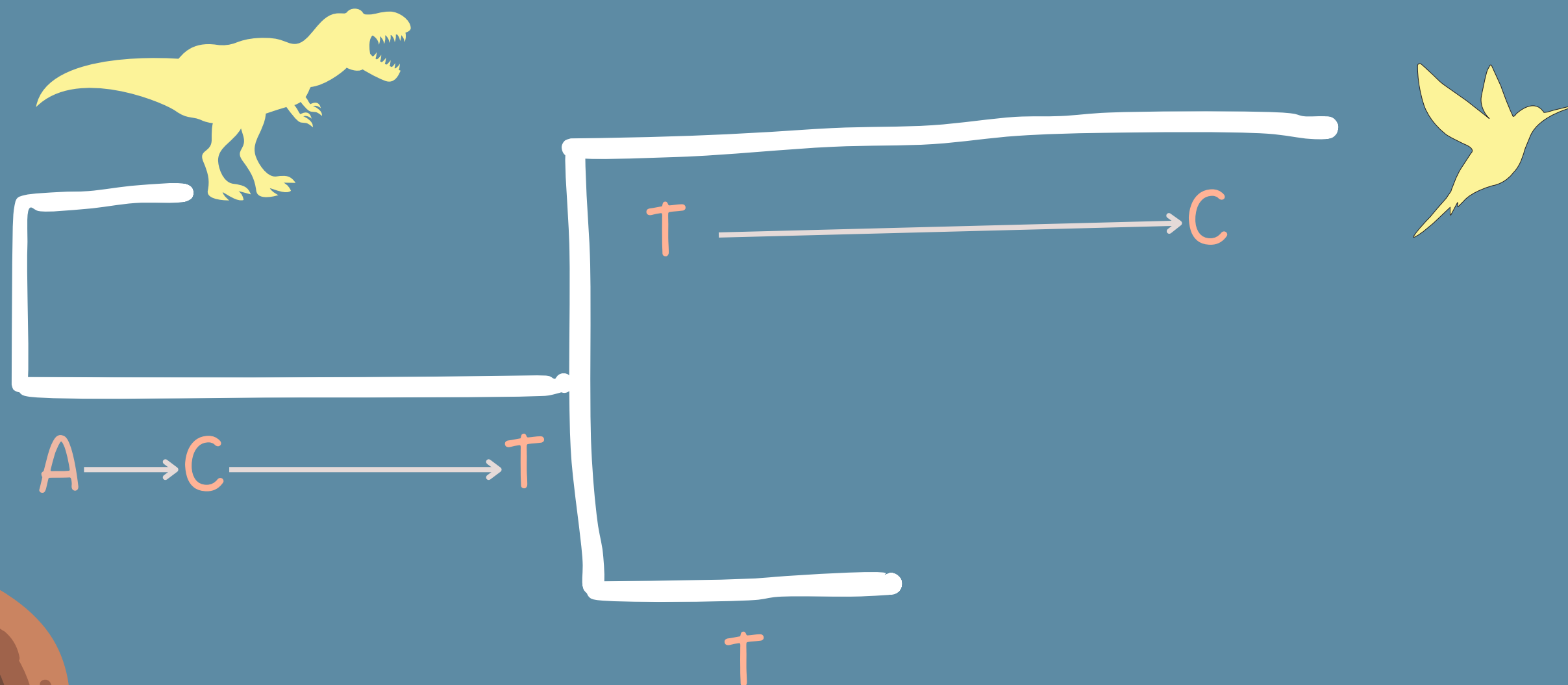
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



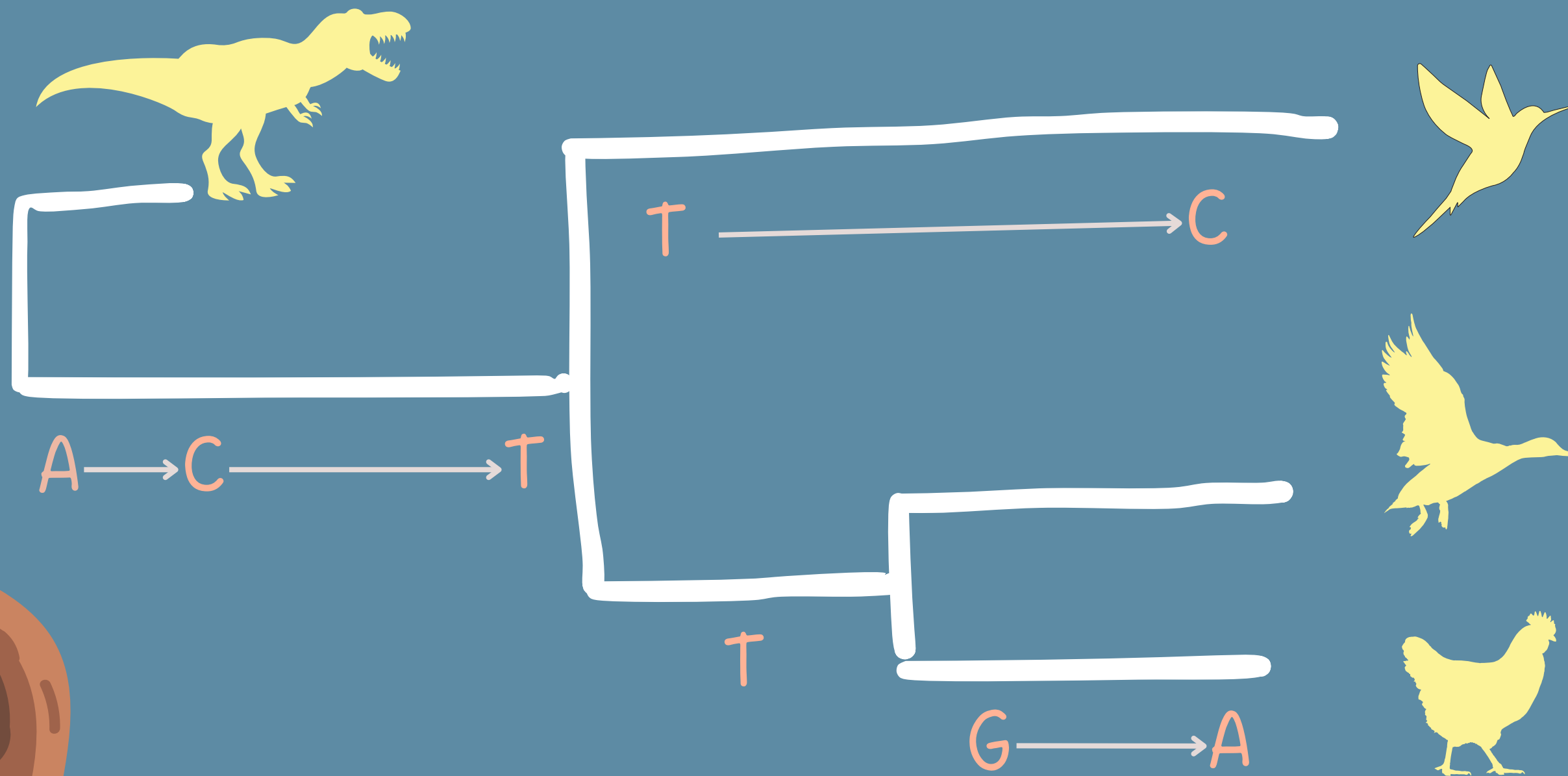
HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?

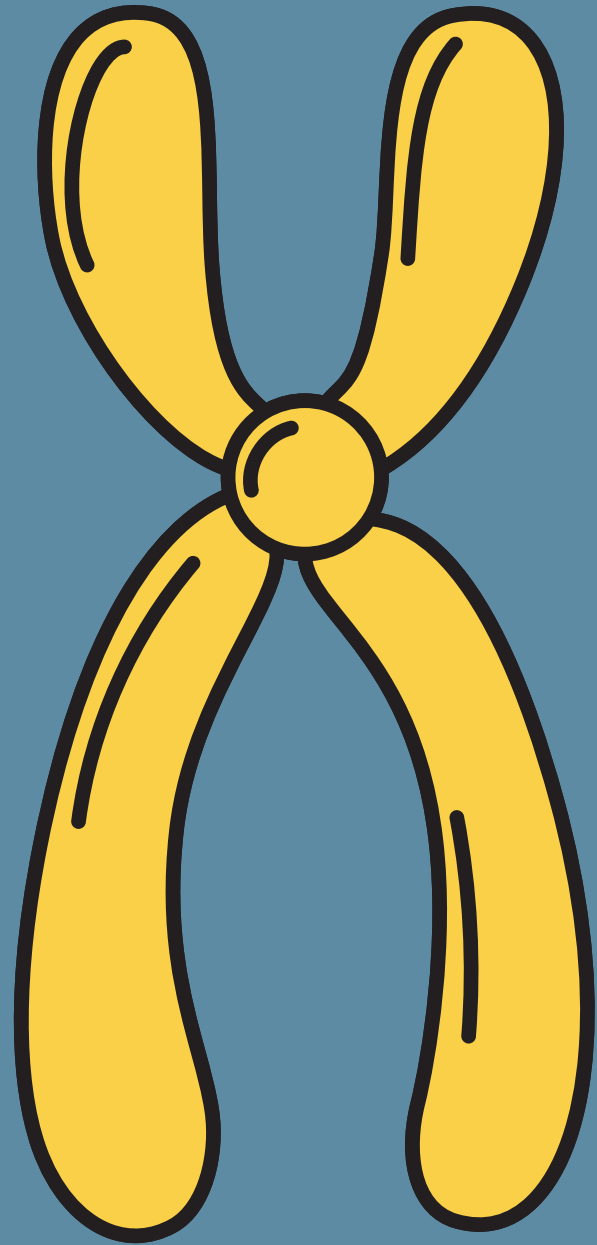


HOW DO WE "SEE" THIS

OVER LONG EVOLUTIONARY TIME?



HOW DO WE MEASURE EVOLUTION ON A SEQUENCE?



How is it
calculated?

ω : a ratio for
measuring natural
selection

What does it mean?

dN/dS

Selective regimes

dN = rate of non synonymous substitutions
 dS = rate of synonymous substitutions

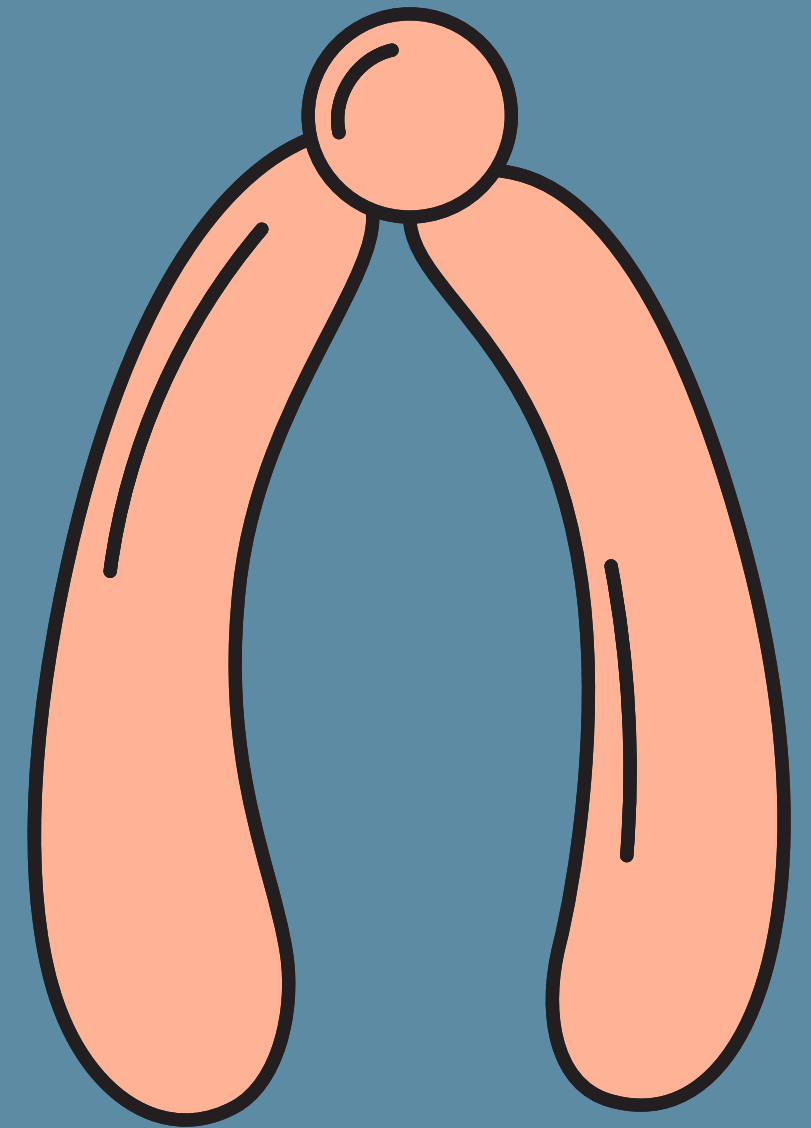
$\omega > 1$ positive (diversifying) selection
 $\omega \cong 1$ neutral evolution
 $\omega < 1$ negative (purifying) selection

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?

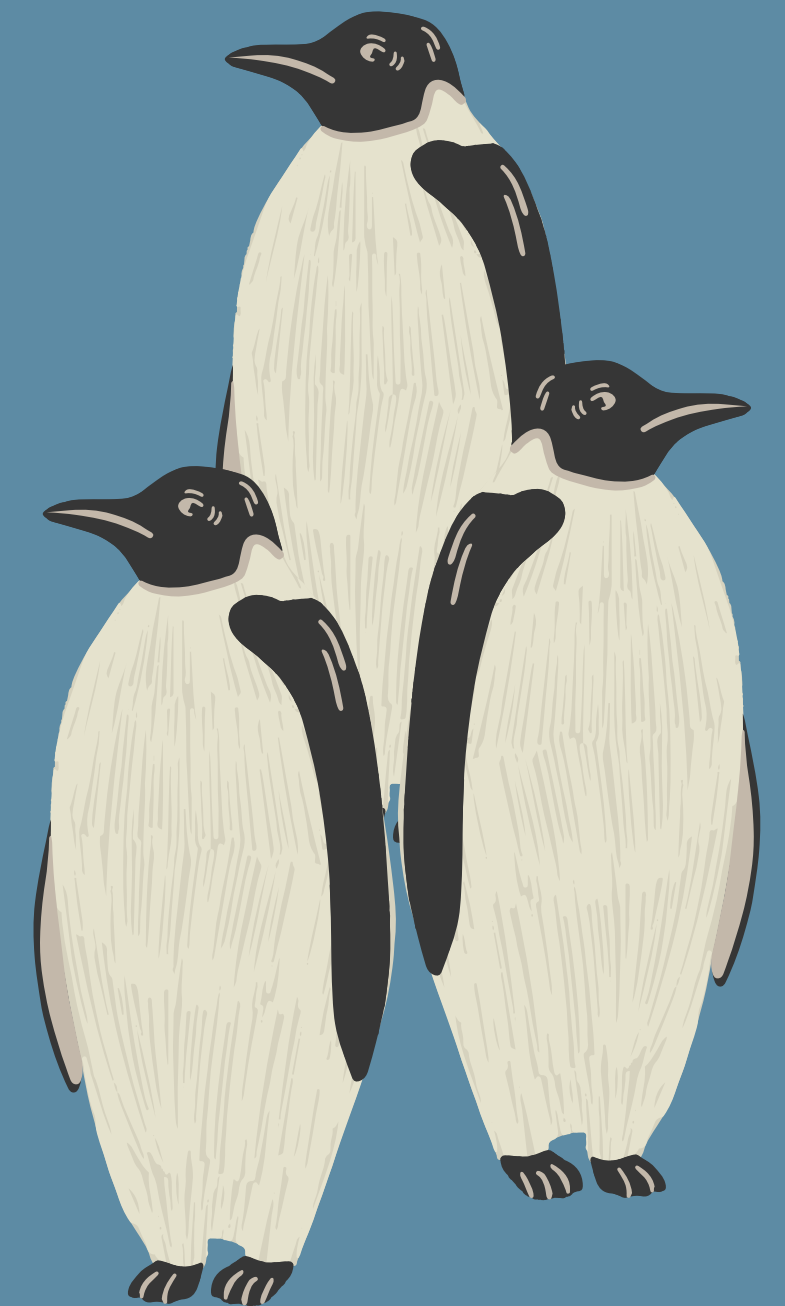
Fit model to data: what does the data tell us about the evolutionary process?

(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?

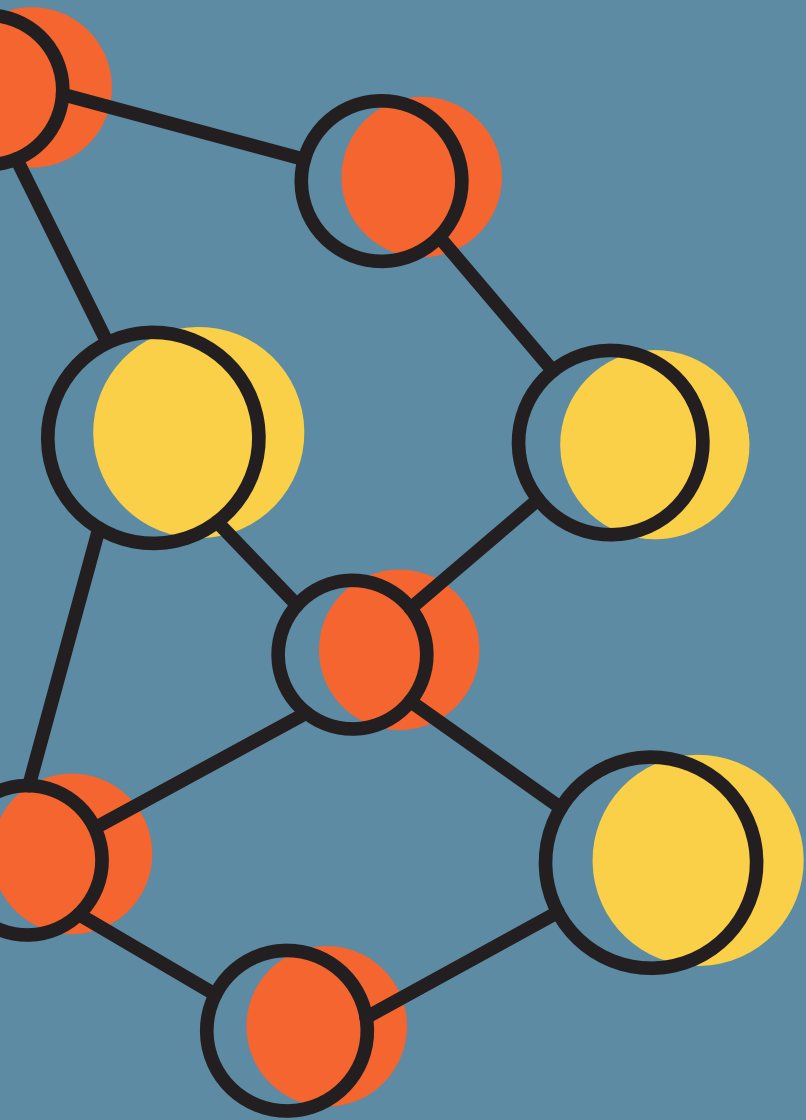
Investigating the signal: which particular sites are under positive selection?

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



BRANCH-MODELS

Estimate omega across **branches** only



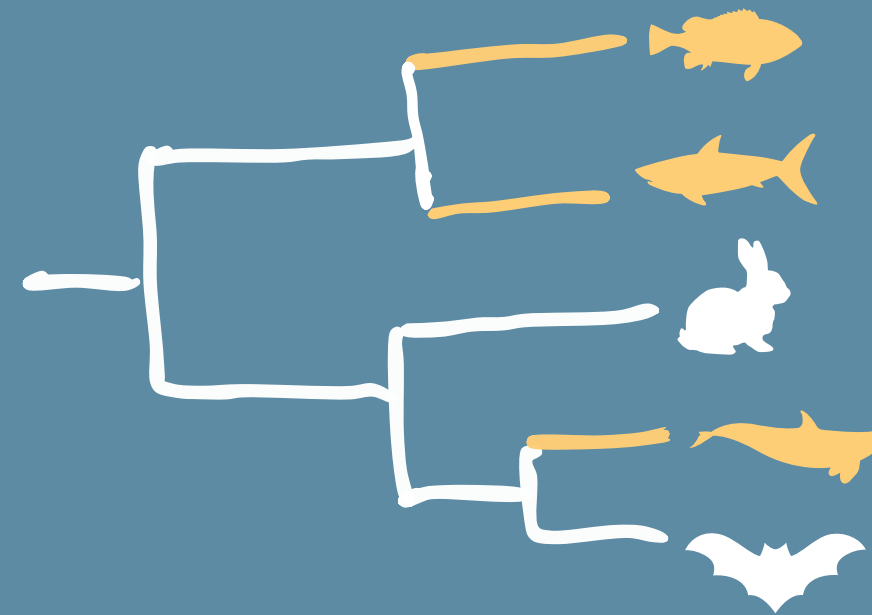
Model 0



ω_1

0.324

2-model



ω_1

0.324

ω_2

0.563

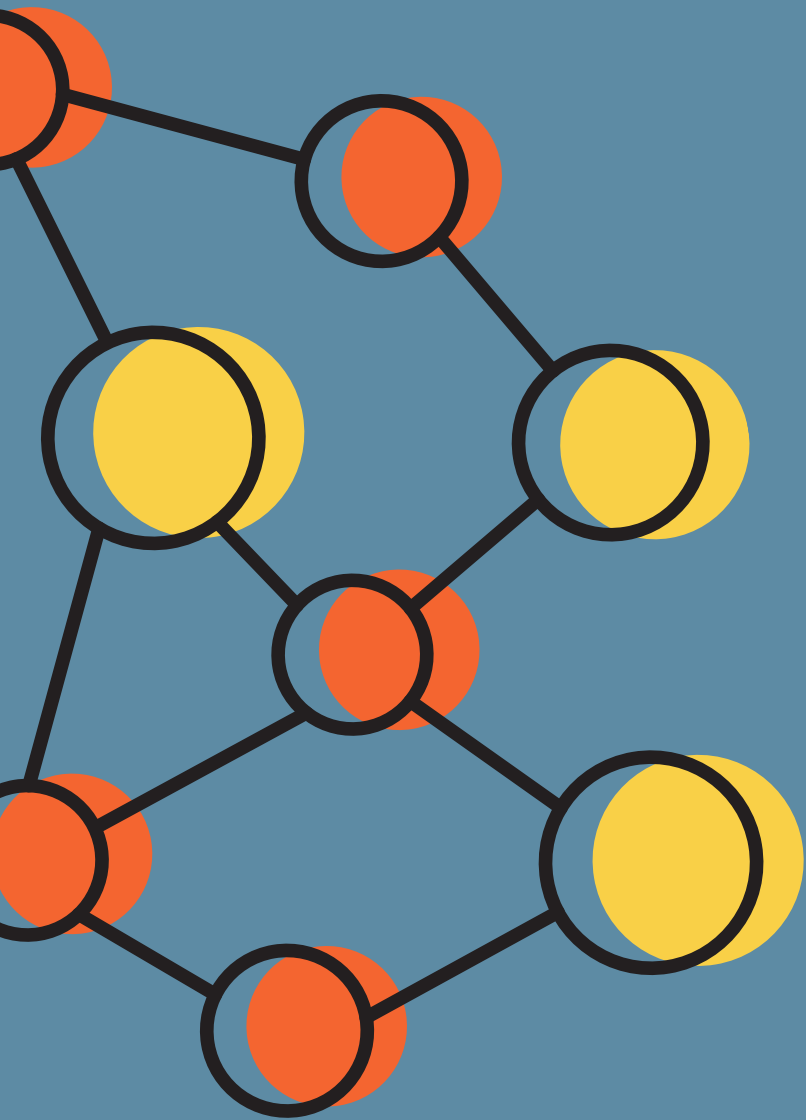
por exemplo...

SITE-MODELS

Estimate omega across **sites** only

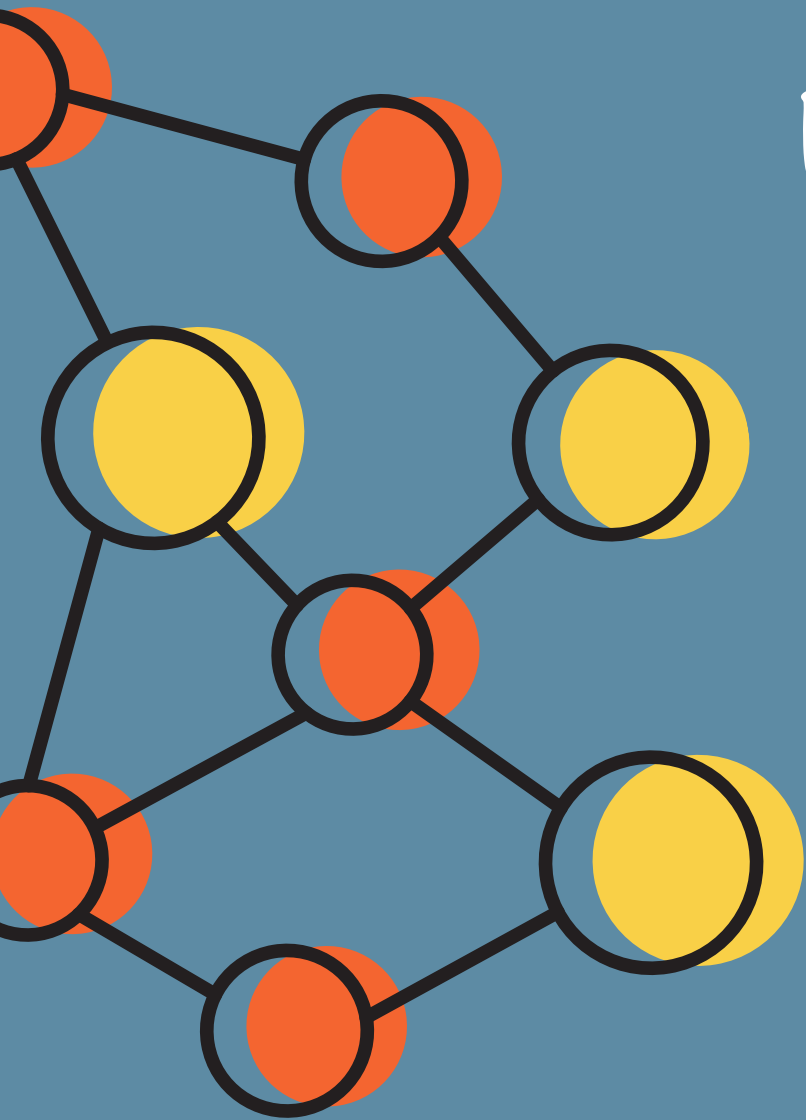
Model M1a

Model M2a

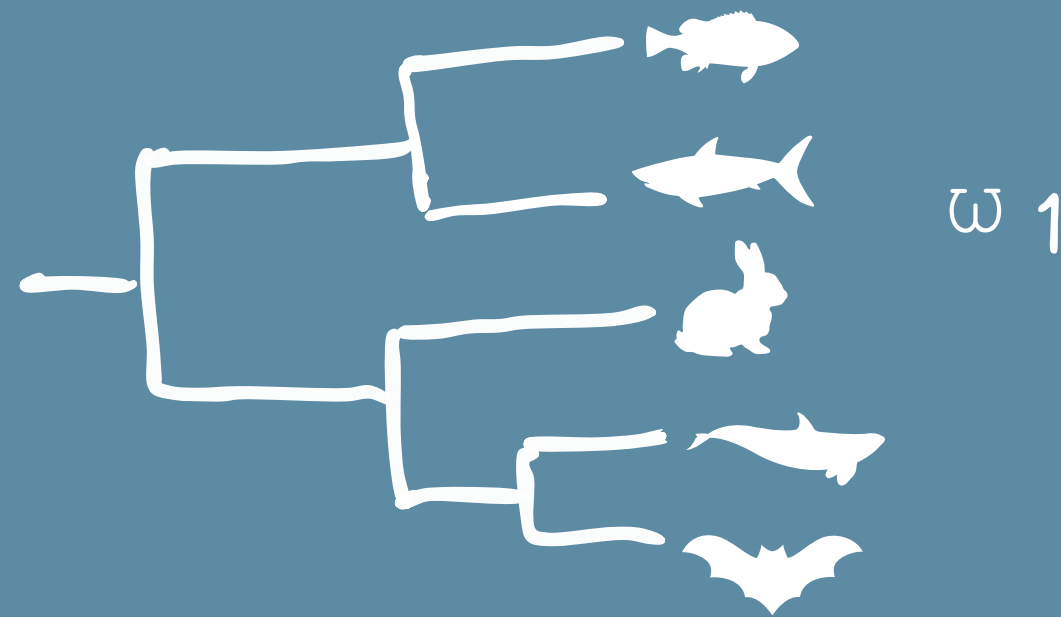


SITE-MODELS

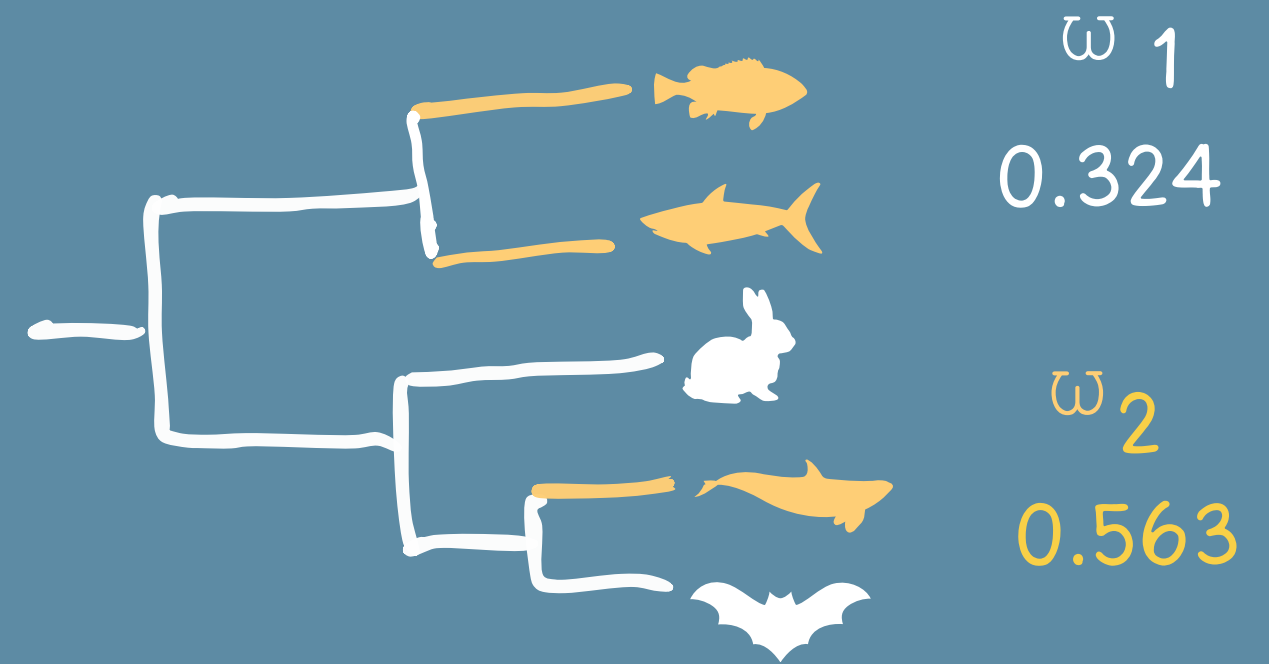
Estimate omega across **branches** and **sites**

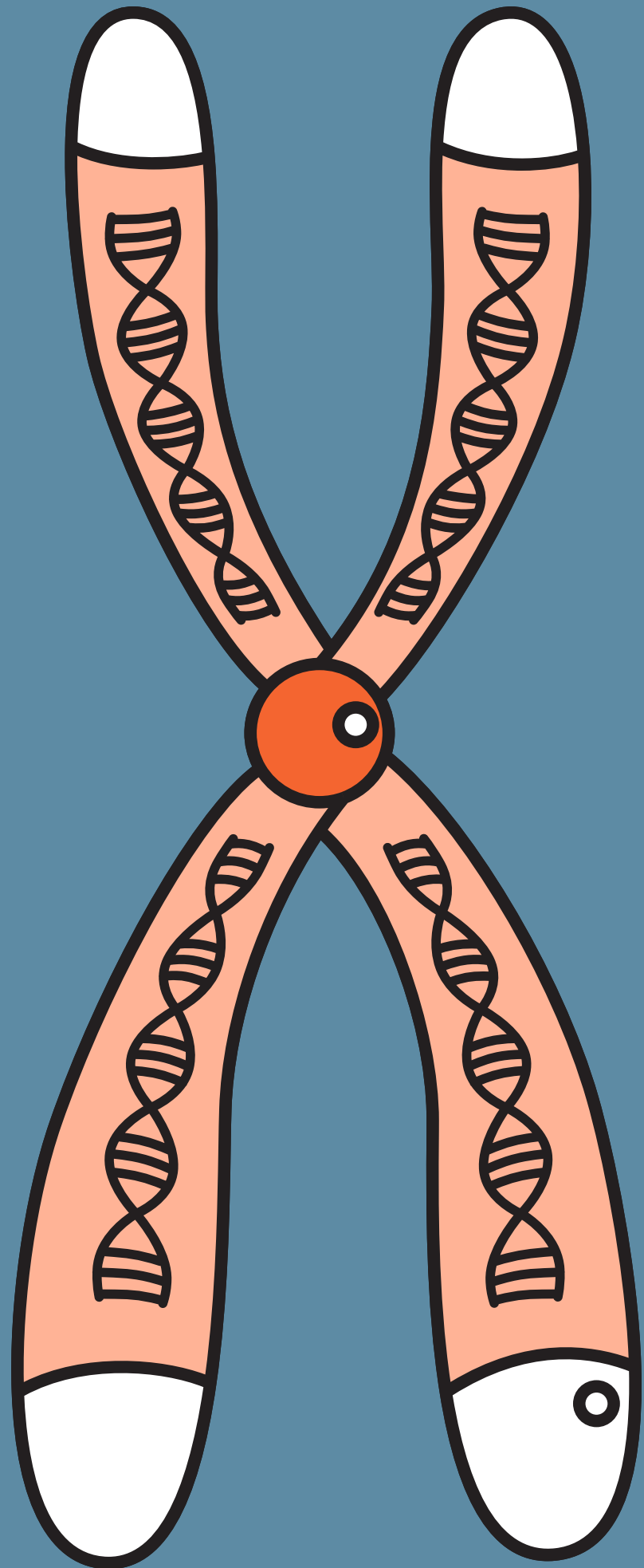


A-model null



A model





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

[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

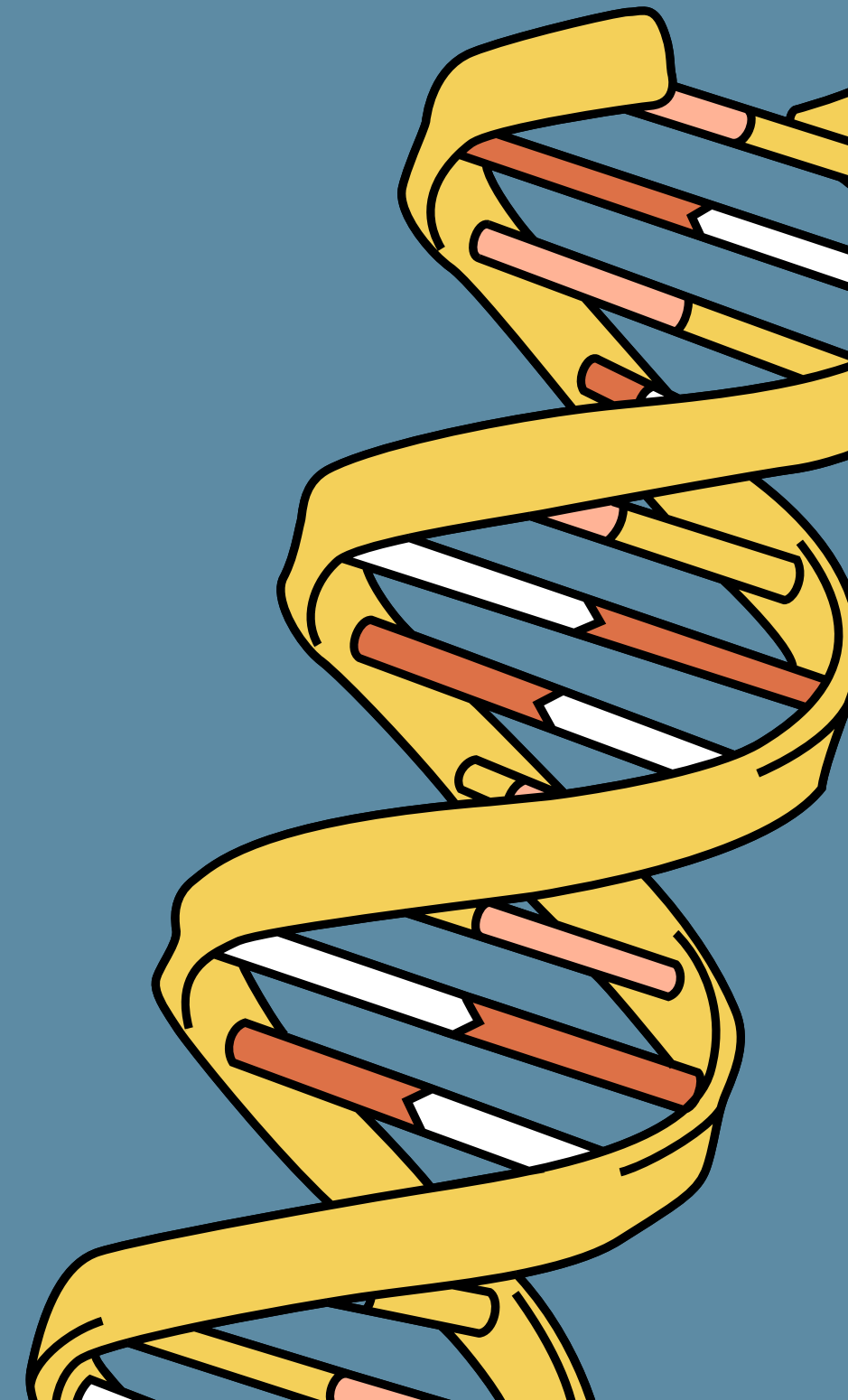
☐ Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

A stylized illustration of a DNA double helix. The sugar-phosphate backbones are represented by thick yellow ribbons that spiral around each other. The nitrogenous base pairs are shown as horizontal bars between the ribbons, with some colored orange and others white. The entire structure is set against a solid blue background.

HOW TO GET AN ALIGNMENT

ALIVIEW



HOW TO GET A TREE



IQ-TREE Web Server: Fast and accurate phylogenetic trees under maximum likelihood

Server load: 4% Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. [doi:10.1093/nar/gkw256](https://doi.org/10.1093/nar/gkw256)

Tree Inference | Model Selection | Analysis Results

For a quick start, take a look at the [tutorial](#) for the IQ-TREE web server.
Please visit the [IQ-TREE homepage](#) for more information or if you want to download the main software.
Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file :

Use example alignment: ☐ Yes

Sequence type: ☒ Auto-detect ☐ DNA ☐ Protein ☐ Codon ☐ DNA->AA ☐ Binary ☐ Morphology

Partition file: This field is optional.

Partition type: ☒ Edge-linked ☐ Edge-unlinked

Substitution Model Options

Substitution model: Auto

FreeRate heterogeneity: ☐ Yes [+R]

Rate heterogeneity: ☐ Gamma [+G] ☐ Invar. sites [+I]

#rate categories: 4

State frequency: ☒ Empirical (from data) ☐ AA model (from matrix) ☐ ML-optimized
☐ Codon F1x4 ☐ Codon F3x4

Ascertainment bias correction: ☐ Yes [+ASC]

Branch Support Analysis

Bootstrap analysis: ☐ None ☒ Ultrafast ☐ Standard

Number of bootstrap alignments: 1000

Create .ufboot file: ☐ Yes (write bootstrap trees to .ufboot file)

Maximum iterations: 1000

Minimum correlation coefficient: 0.99

Single branch tests:

SH-aLRT branch test: ☐ No ☒ Yes #replicates: 1000

Approximate Bayes test: ☐ Yes

IQ-TREE Search Parameters

Perturbation strength: 0.5

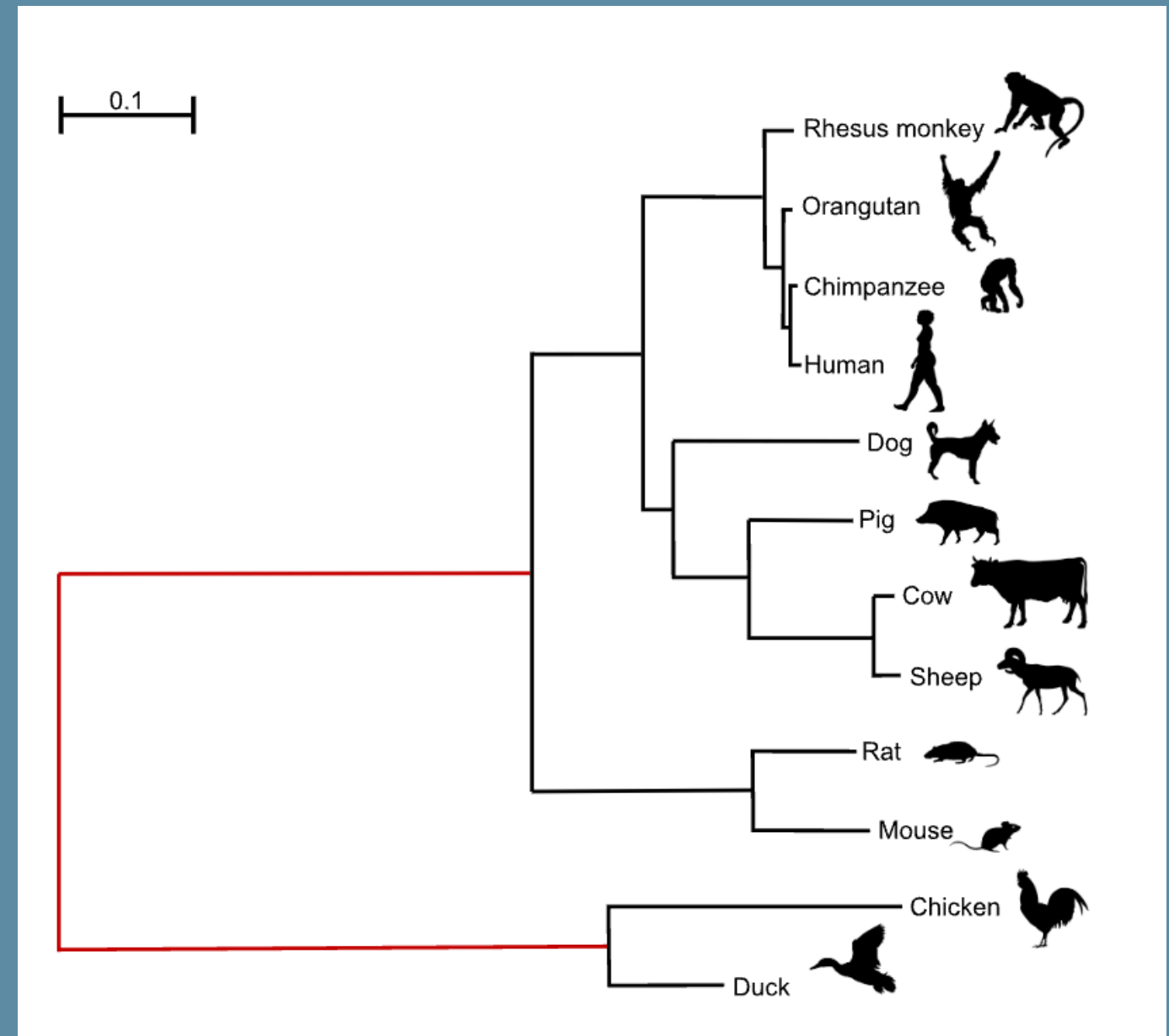
IQ-TREE stopping rule: 100

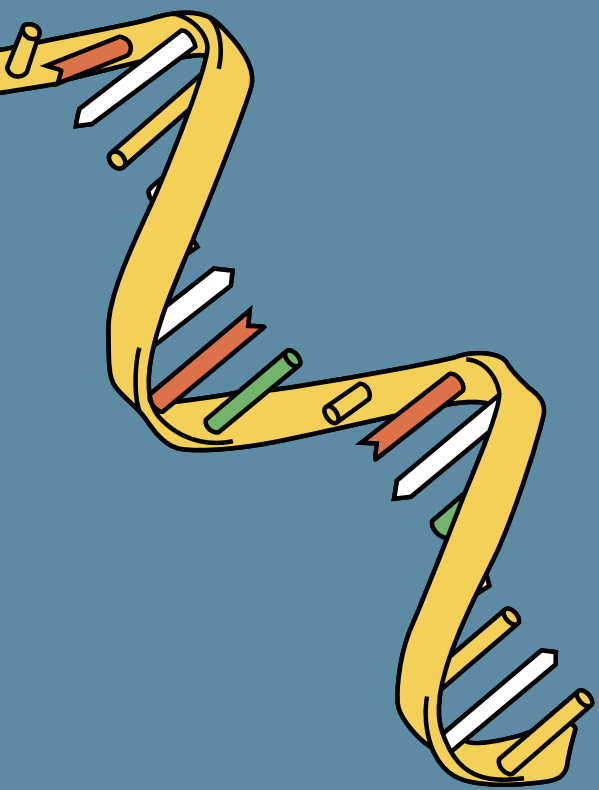


HOW DO WE USE CODEML TO TEST A HYPOTHESIS?

Questions:

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



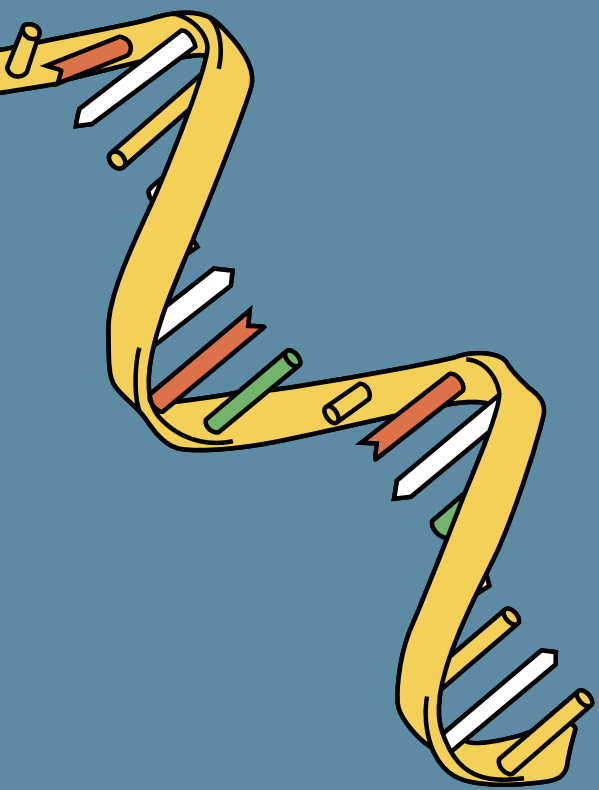


THE CODEML CONTROL FILE



```
seqfile = Mx_aln.phy
treefile = Mx_unroot.tree
outfile = out_M0.txt
noisy = 3
verbose = 1
seqtype = 1
Ndata = 1
icode = 0
cleandata = 0
model = 0
Nssites = 0
CodonFreq = 2
clock = 0
fix_omega = 0
omega = 0.5
```

- * Path to the alignment file
- * Path to the tree file
- * Path to the output file
- * Display moderate information on the screen
- * Detailed output file
- * Codon data
- * One gene alignment
- * Universal genetic code
- * Do not remove sites with ambiguity data
- * One ω for all branches (M0 and site models)
- * One ω for all sites (M0 and branch model)
- * Use F3x4 model
- * Assume no clock
- * Enables option to estimate omega
- * 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



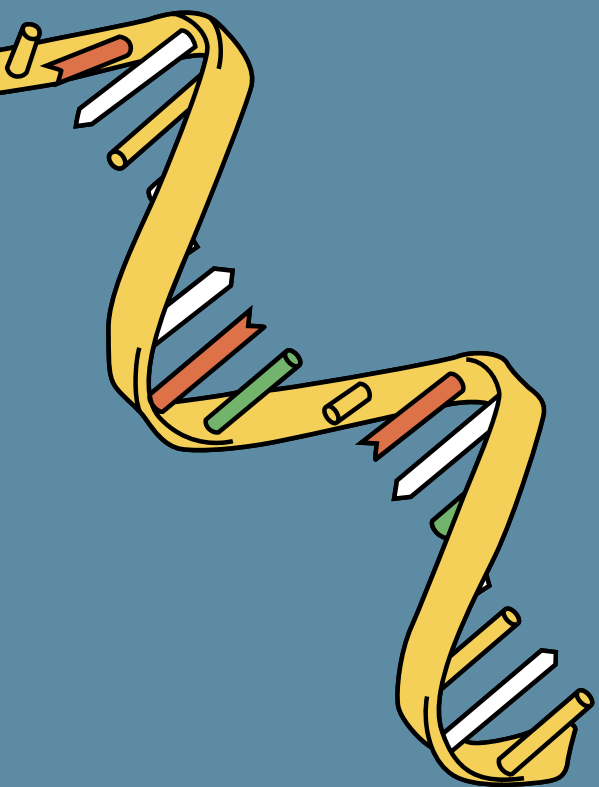
change this with the name or path to your tree

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!