

# CODEML WORKSHOP

Parte prática #1





# TÓPICOS:

- Introdução à linha de comando linux
- Como medimos evolução numa sequência genética?
- O que é o codeml?
- Modelos de ramos, sítios e ramos-sítios
- Usando o codeml para testar hipóteses

# LINHA DE COMANDO

## INTRODUÇÃO

- navegar: `cd`
- listar: `ls`
- criar pastas: `mkdir`
- criar/editar arquivo: `nano`
- mover: `mv`
- visualizar: `less`, `cat`, `more`
- manipular/editar arquivos: `sed`, `sort`



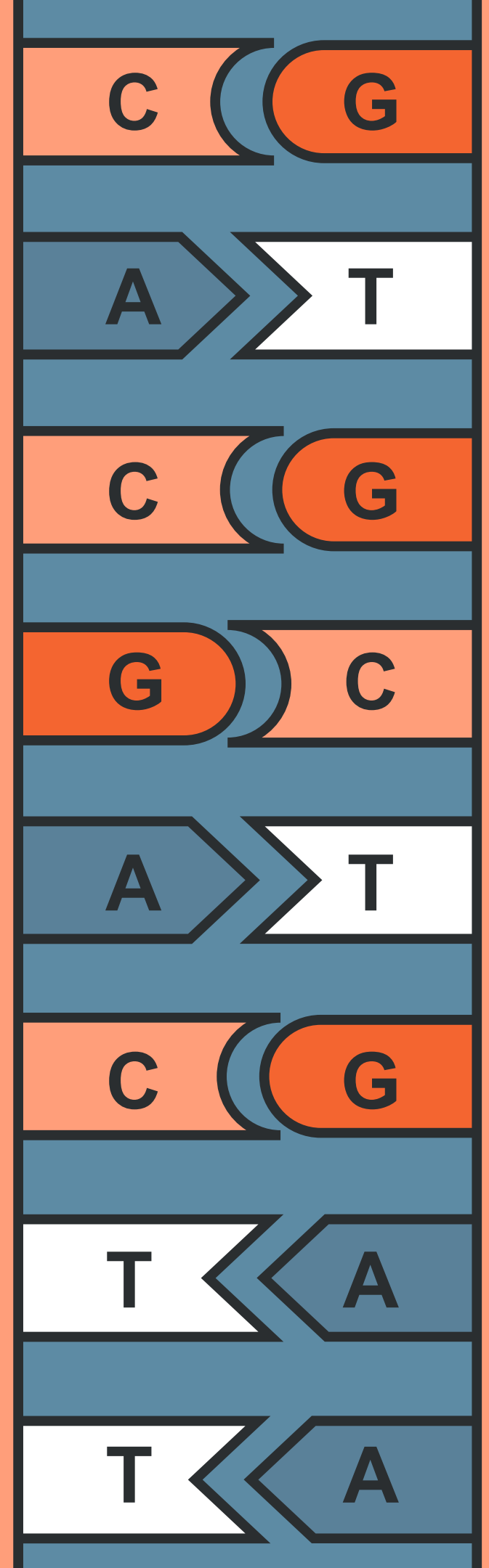
# COMO PODEMOS MEDIR EVOLUÇÃO?

ATGAGGTGCCACGTCGCTTCCAGCTGCCTCGTGGTCTGA

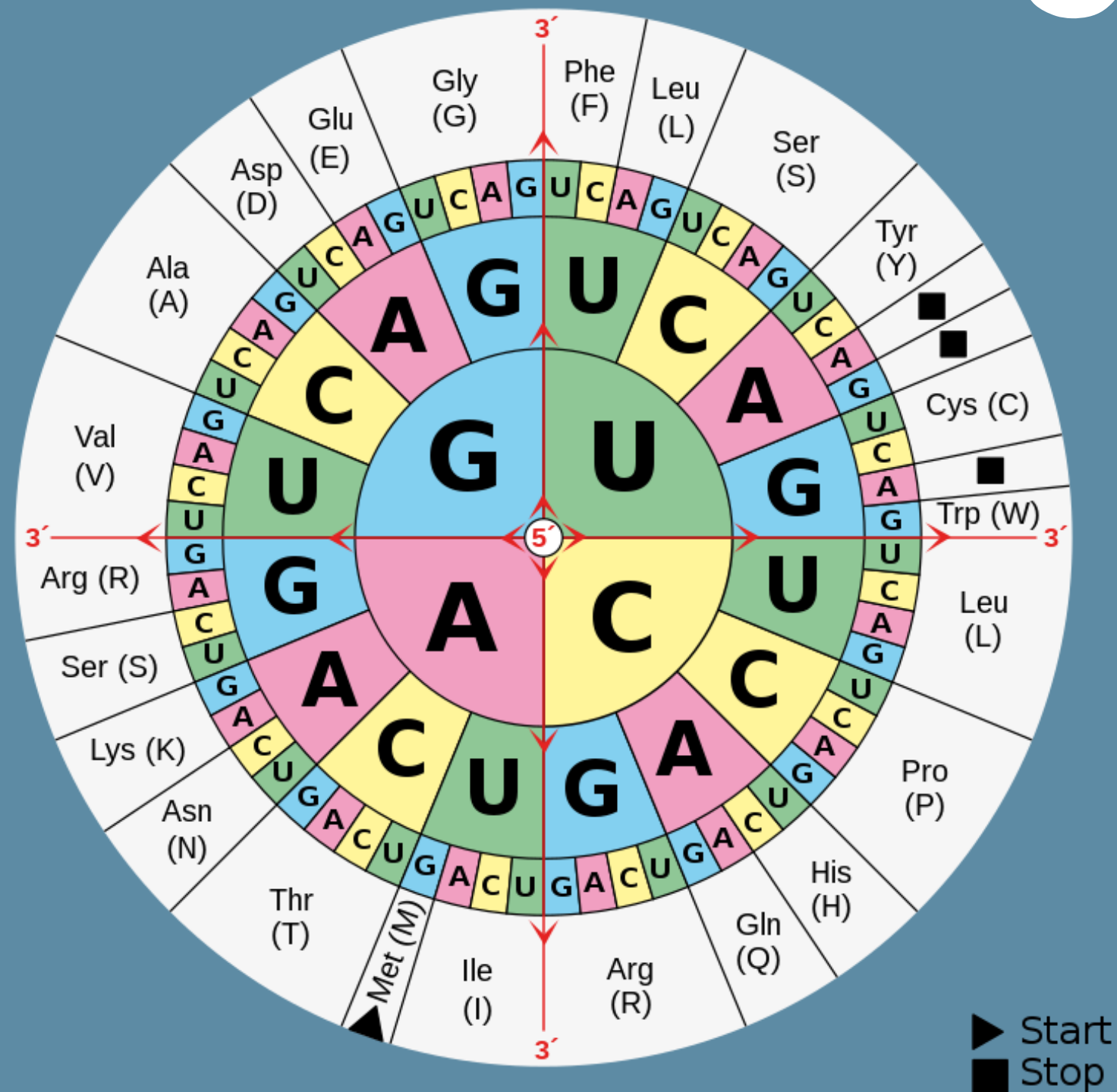
códon de início

códon de parada

Cada 3 nucleotídeos codificam um aminoácido



# O CÓDIGO GENÉTICO É REPETITIVO



Isso significa que diferentes trincas (códon)s podem codificar o mesmo aminoácido

GGA } Gly  
GGG }  
GCG } Ala

# E O QUE ACONTECE QUANDO UM NUCLEOTÍDEO MUDA?

Podemos ter dois tipos de substituições:  
sinônimas ou não-sinônimas

GGA → GGG

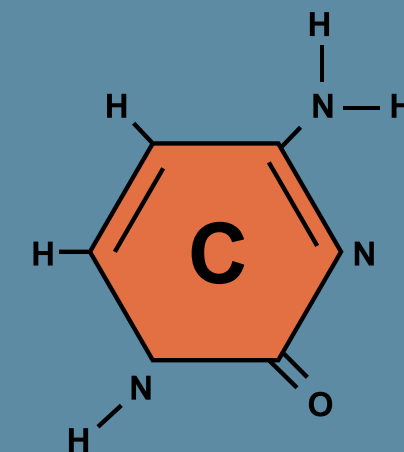
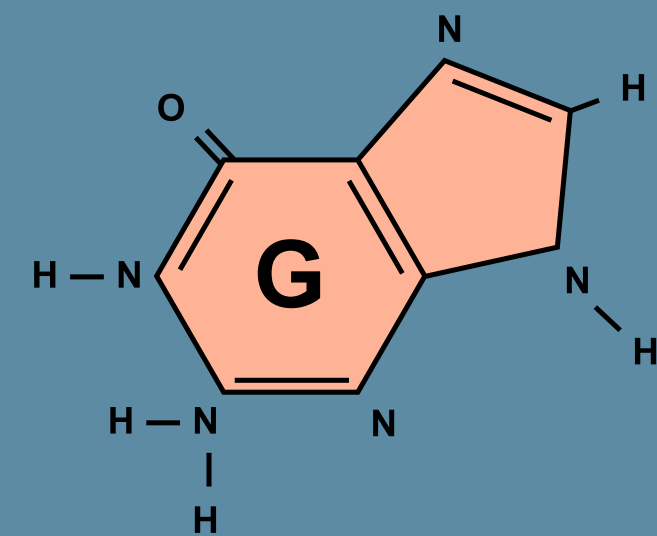
Thr

Thr

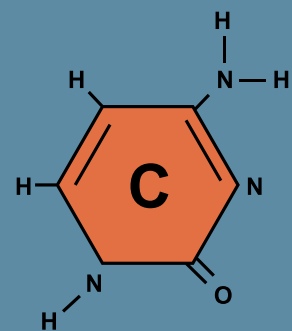
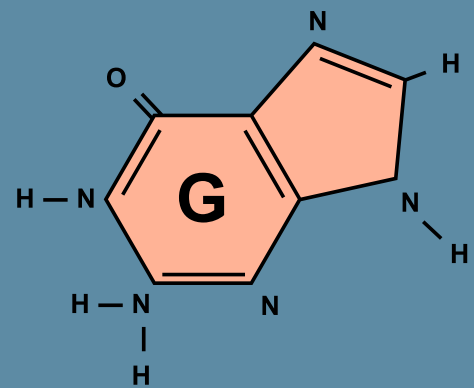
GGG → GCG

Thr

Asn

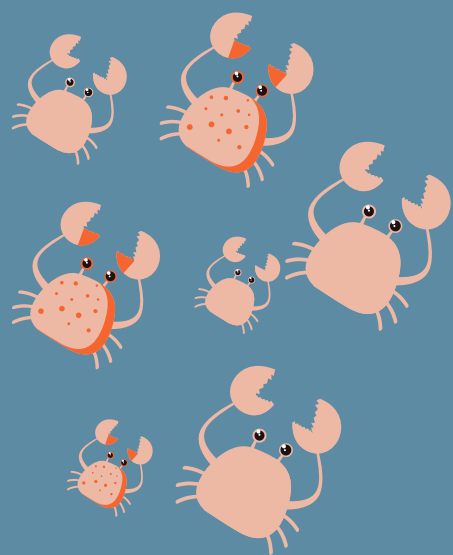


Exemplo de  
substituição: uma  
guanina (G) sendo  
trocada por uma  
citosina (C)



# E O QUE ACONTECE

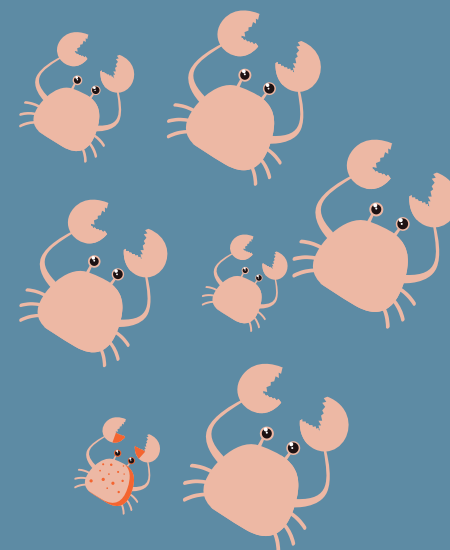
## QUANDO UM NUCLEOTÍDEO MUDA?



GGG → GCG

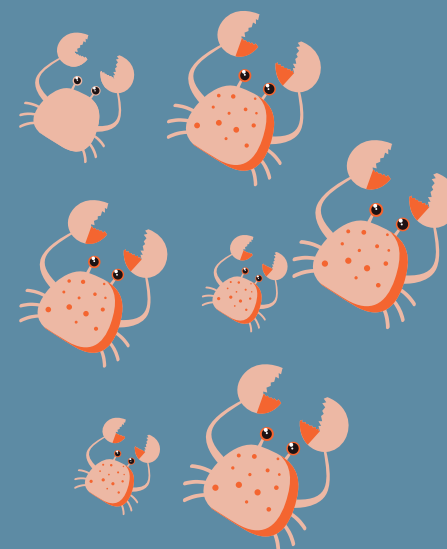
Alelo 1  
rosa

Allelo 2  
pintado



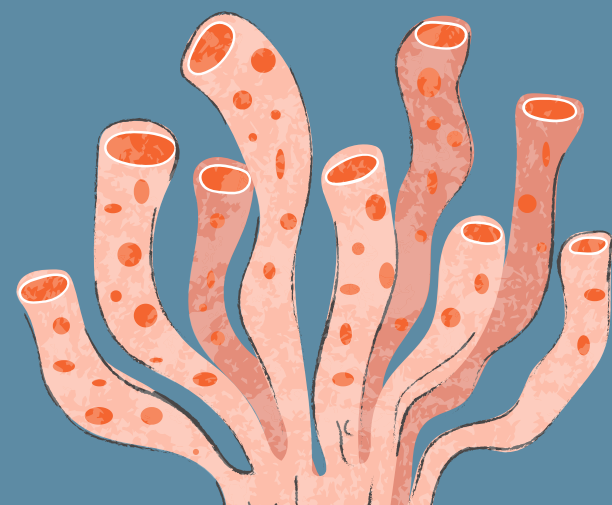
Maior  
predação  
(expostos)

População 2

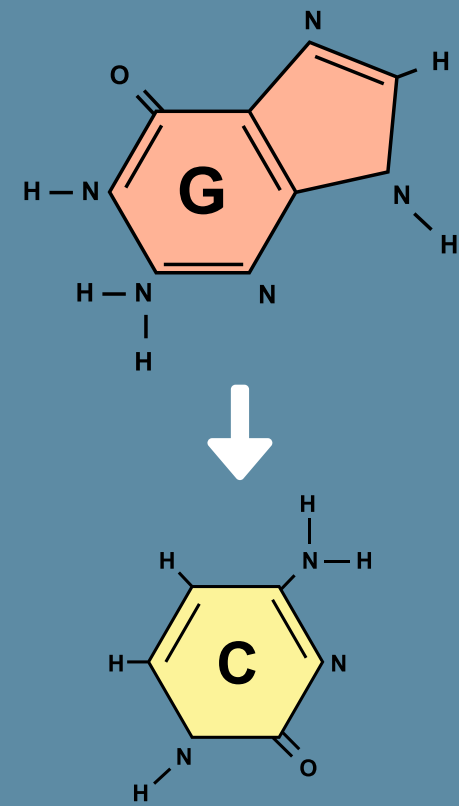


Menor  
predação  
(camuflagem)

População 1



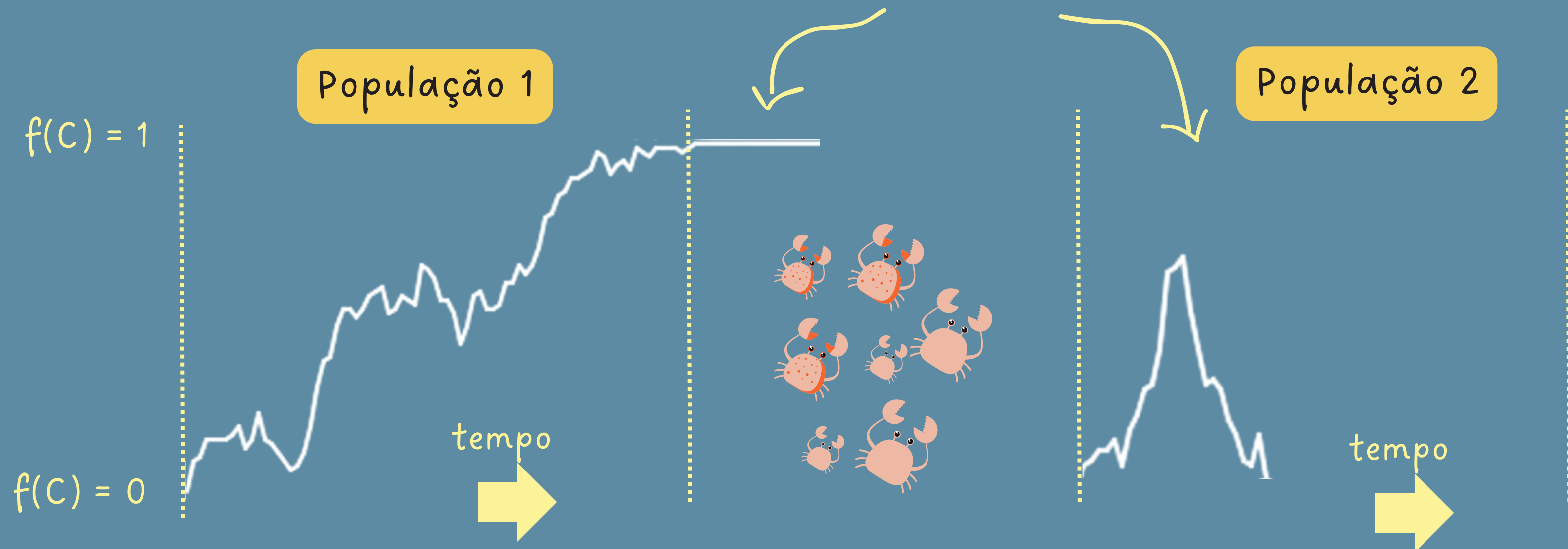




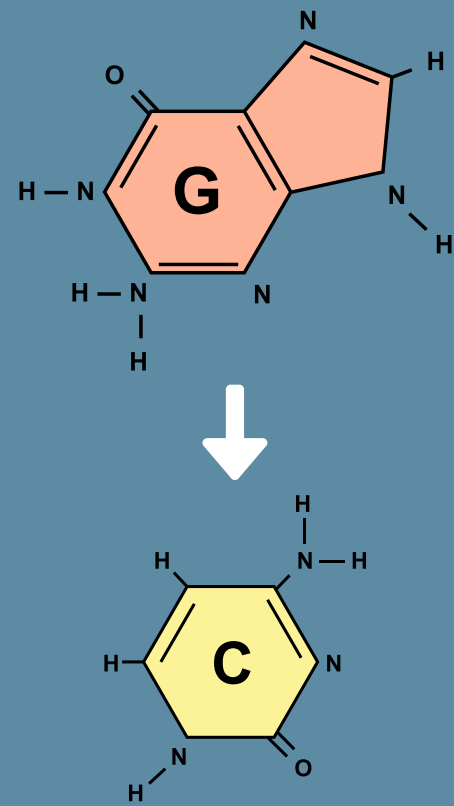
# E O QUE ACONTECE

## QUANDO UM NUCLEOTÍDEO MUDA?

Uma mutação pode ser fixada ou perdida em uma população







# E O QUE ACONTECE

## QUANDO UM NUCLEOTÍDEO MUDA?

processo de fixação

Este gráfico mostra a frequência do alelo C em uma população, ao longo do tempo



**Substituição**

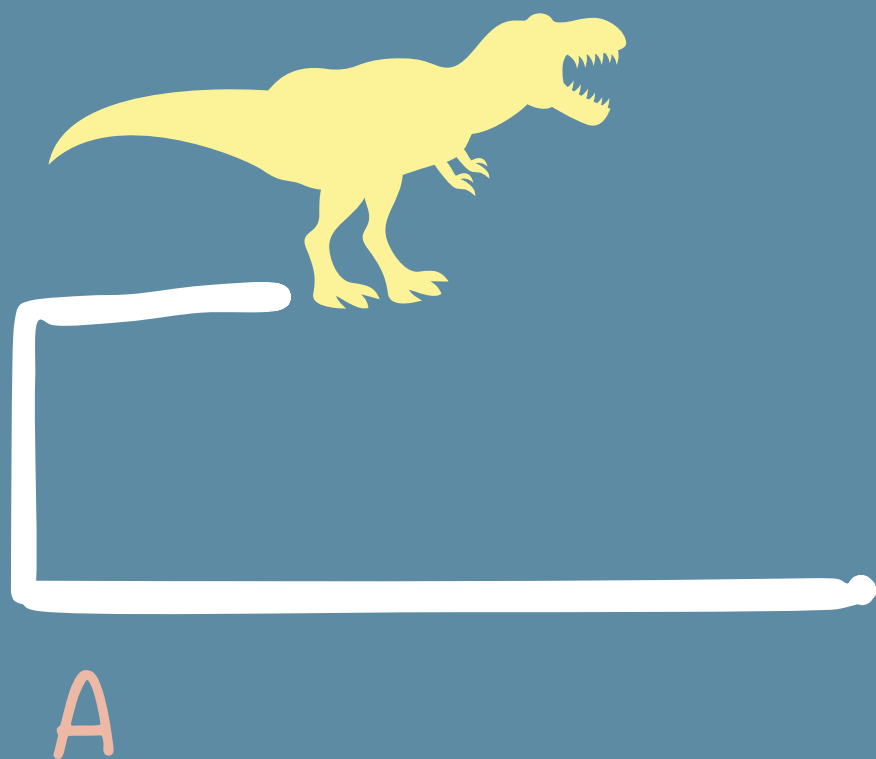
é o resultado do processo de fixação

ou seja, uma mudança no "estado" da população (G → C)

Evento de mutação: G → C

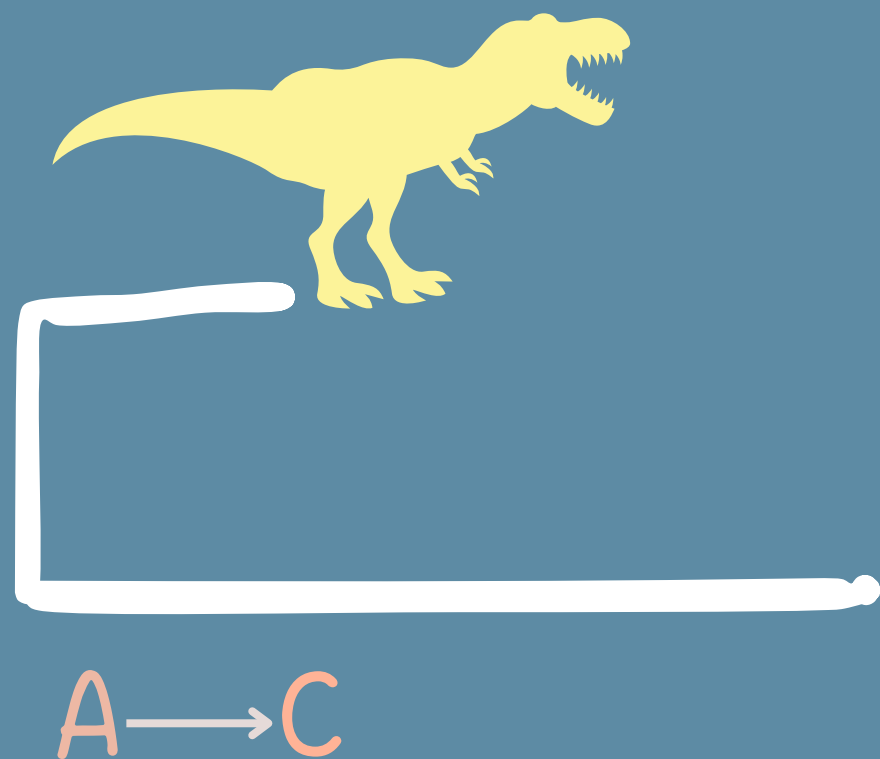
# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



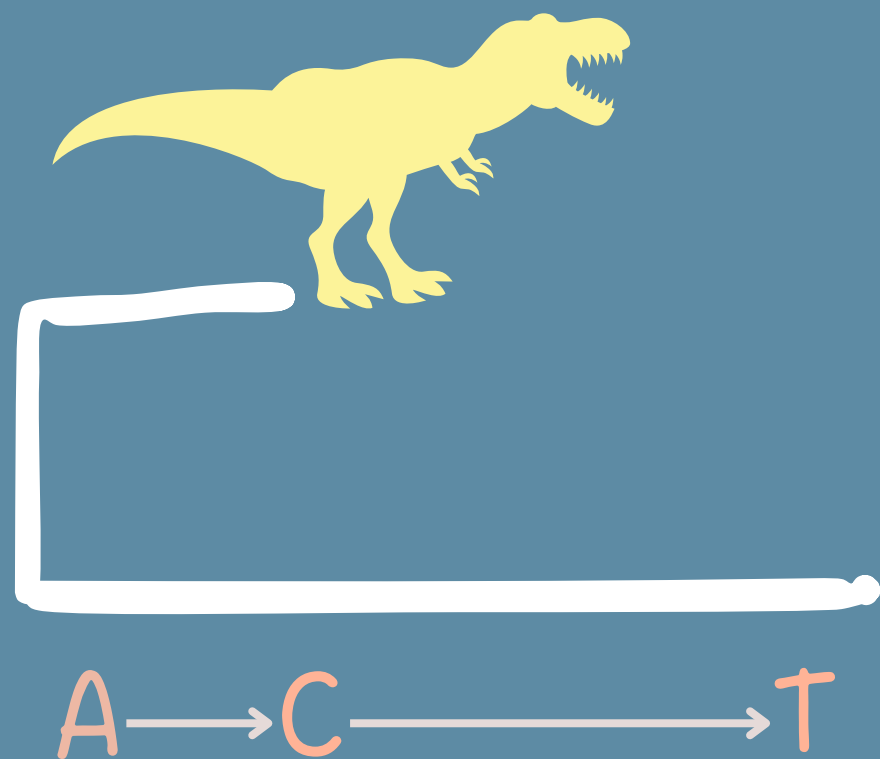
# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



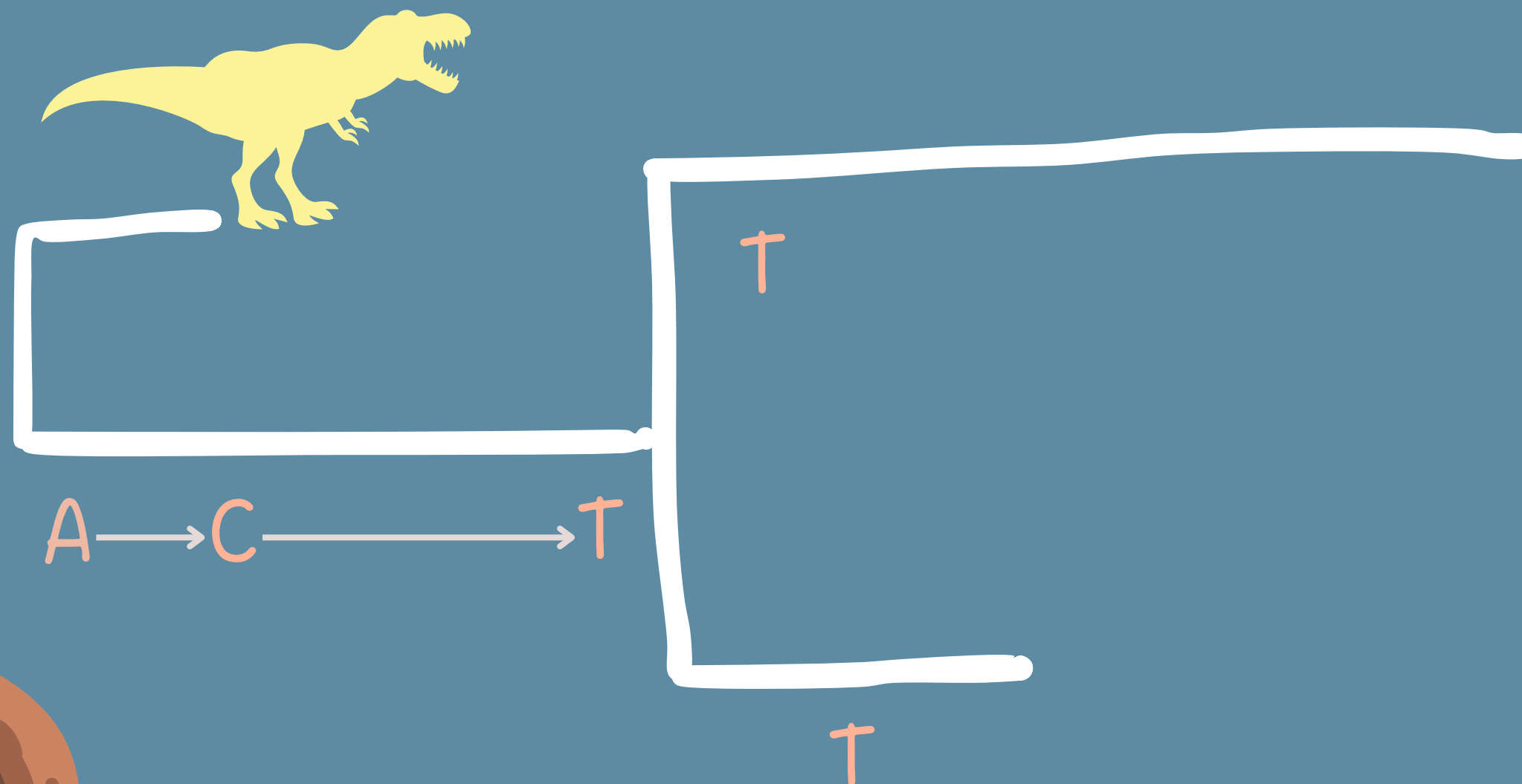
# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



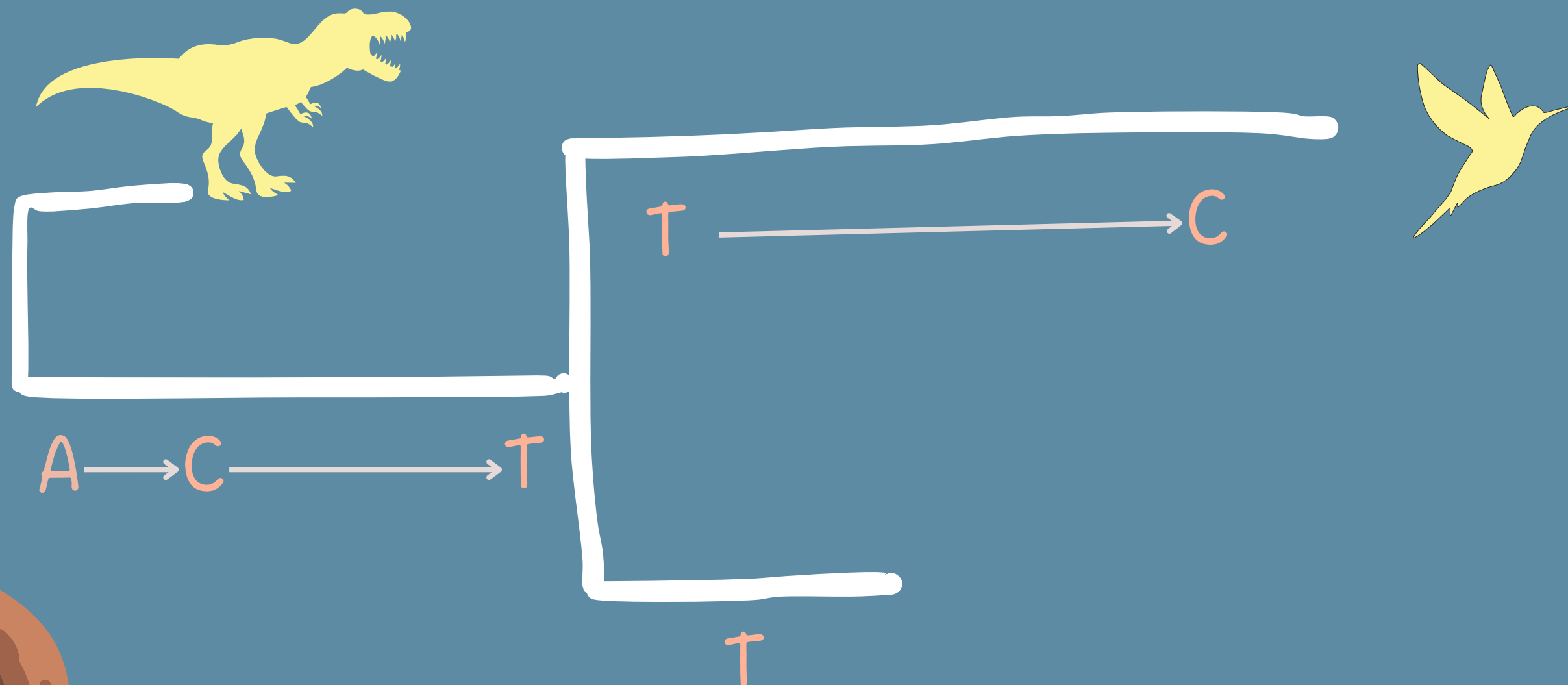
# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



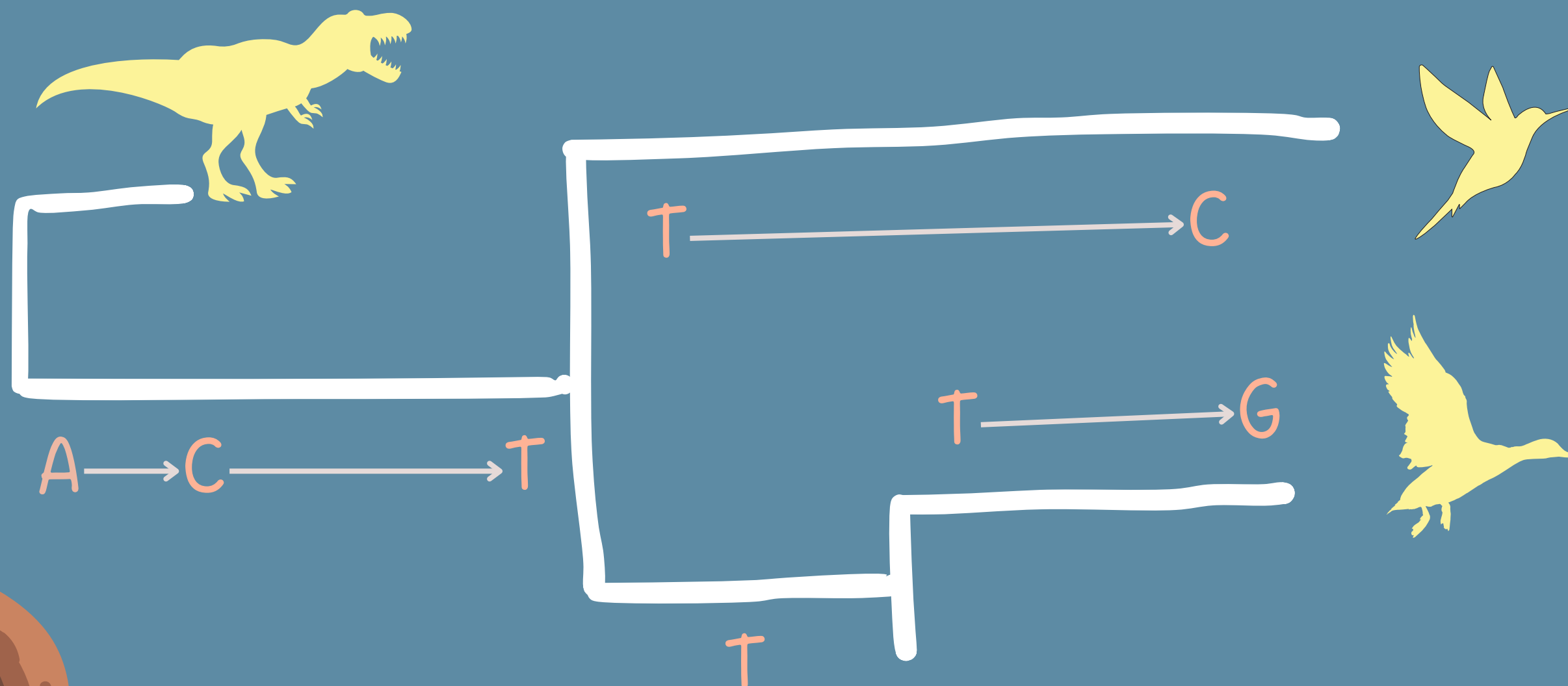
# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



# COMO ISSO ACONTECE

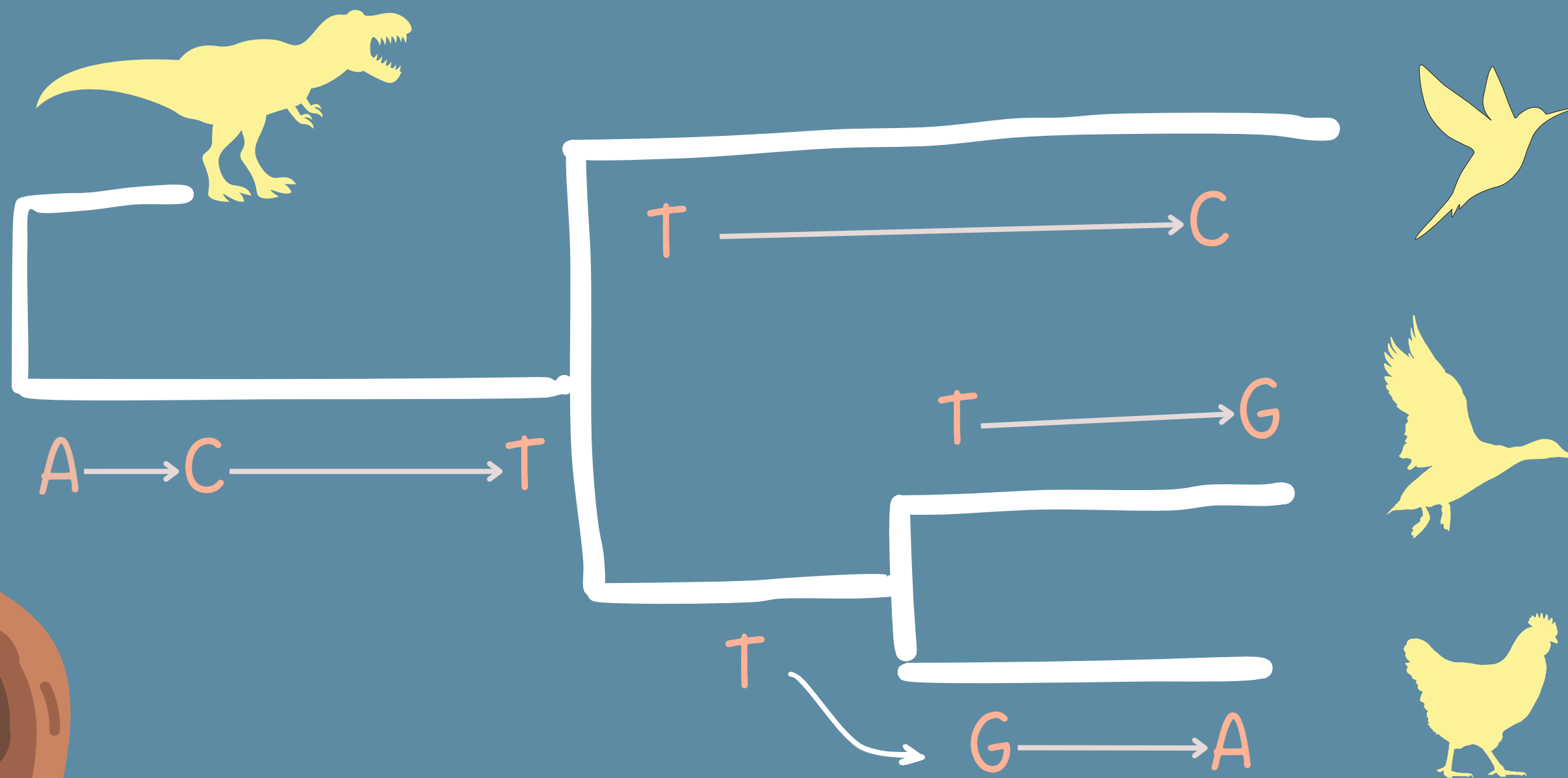
AO LONGO DO TEMPO EVOLUTIVO?



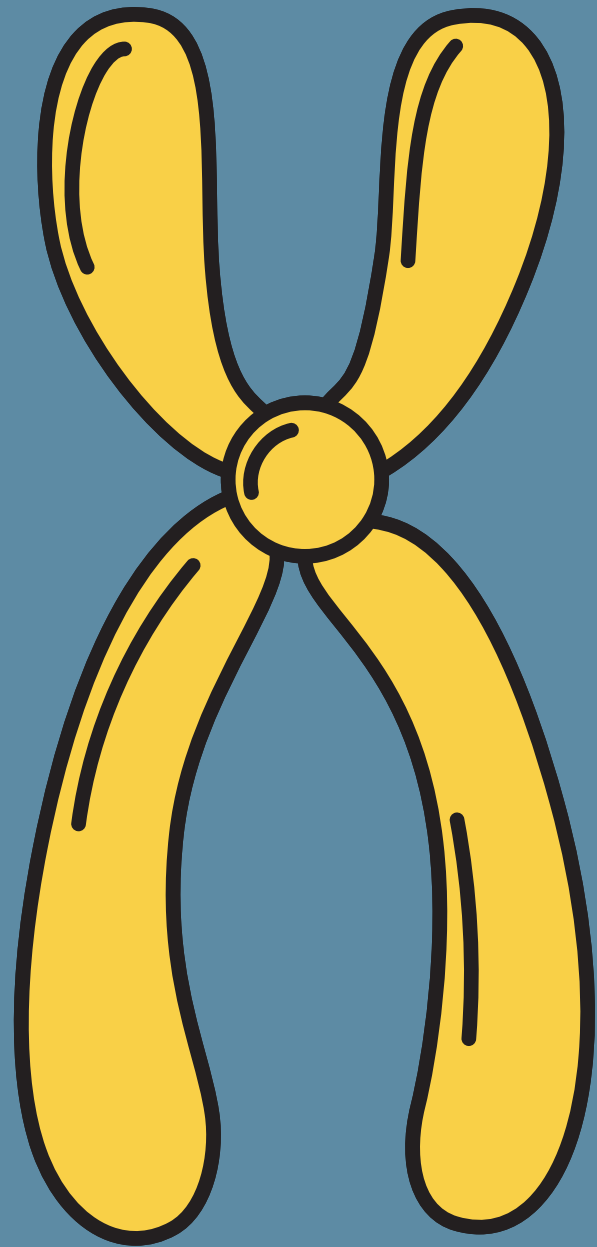


# COMO ISSO ACONTECE

AO LONGO DO TEMPO EVOLUTIVO?



# E COMO MEDIMOS A EVOLUÇÃO EM UMA SEQUÊNCIA?



Como é  
calculado?

$\omega$ : uma medida de  
seleção natural

O que significa?

$dN/dS$

Regimes seletivos

$dN$  = taxa de substituições não-sinônimas  
 $dS$  = taxa de substituições sinônimas

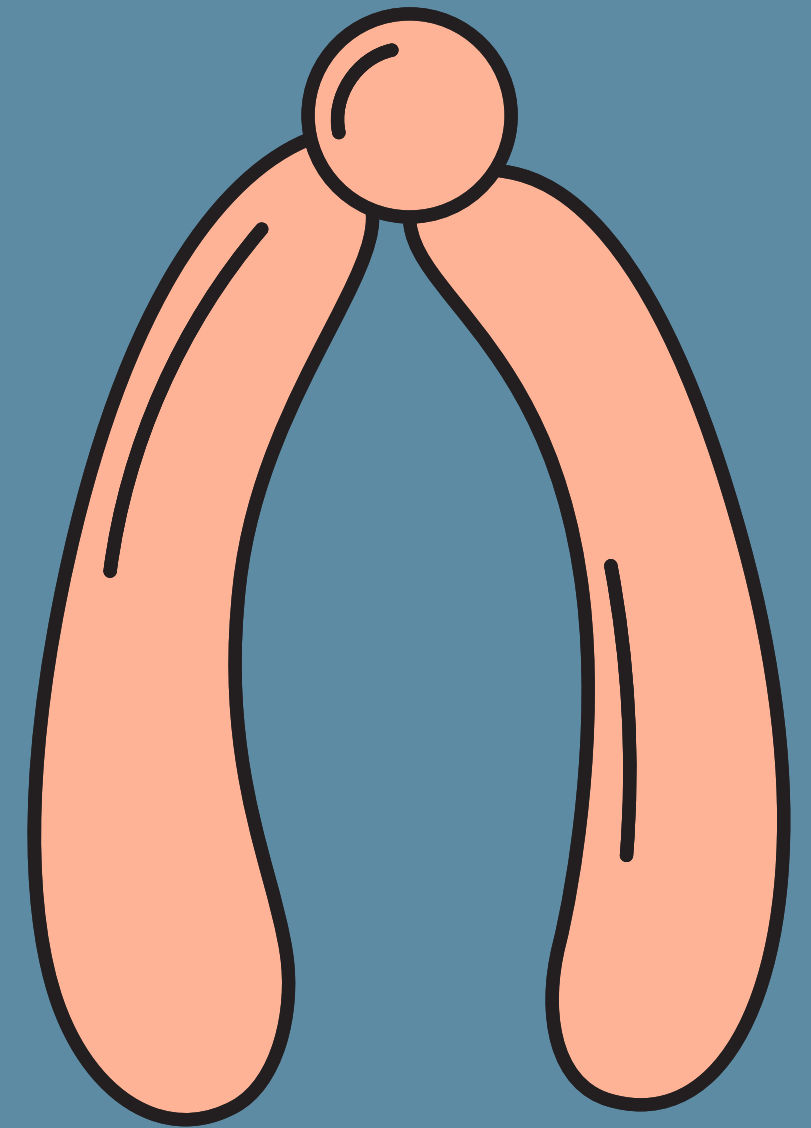
$\omega > 1$  seleção positiva (diversificadora)  
 $\omega \cong 1$  evolução neutra  
 $\omega < 1$  seleção negativa (purificadora)

# O QUE É O CODEML?

Um programa que implementa  
modelos de substituição de  
códon e amino ácidos



Modelo = é uma possível explicação (matemática)  
de como a sequência está evoluindo



# O QUE PODEMOS FAZER COM O CODEML?



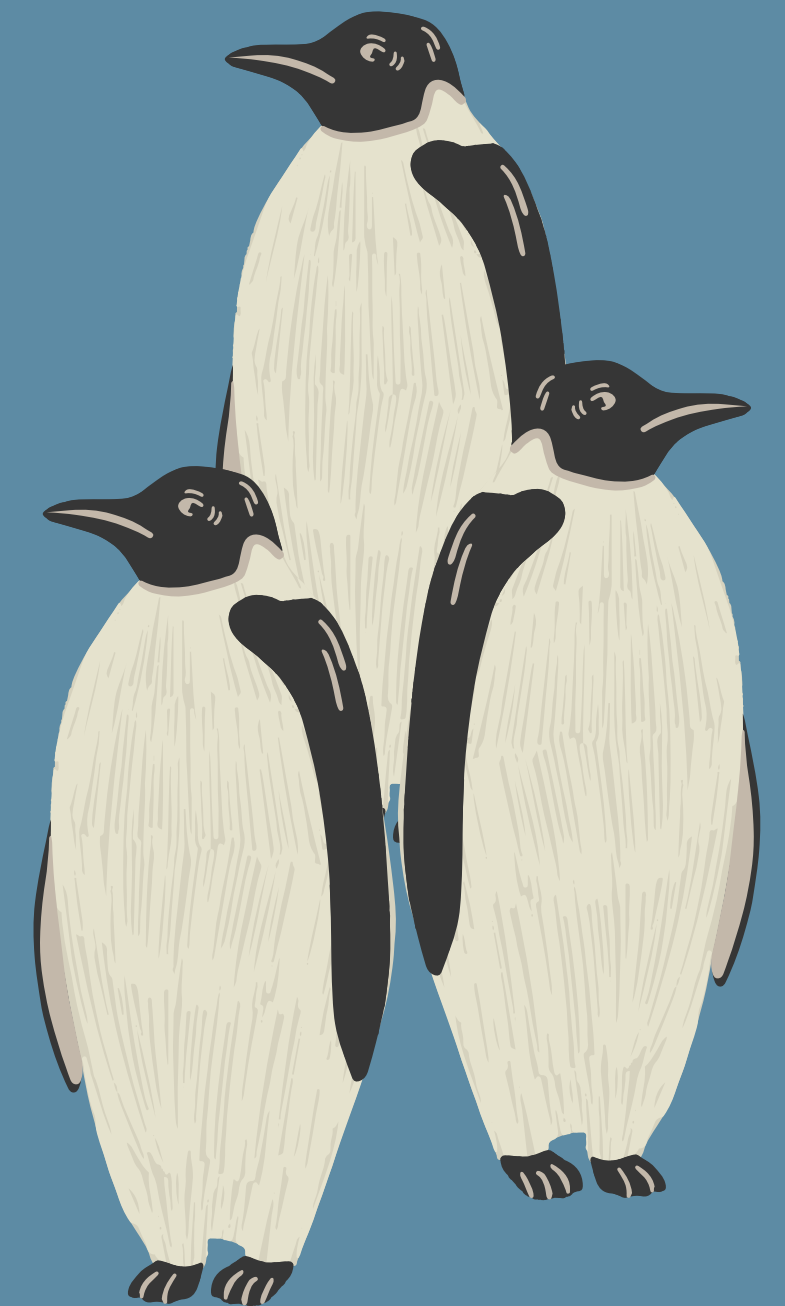
*Ajustar um modelo aos dados:* o que os dados  
(sequências) podem nos dizer sobre o processo evolutivo?

(ex: será que um omega maior ou menor se ajusta  
melhor aos dados?)

# O QUE PODEMOS FAZER COM O CODEML?

Testar hipóteses: os dados me permitem rejeitar a hipótese nula? (evolução neutra)

(ex: será que essa sequência/espécie está evoluindo de forma diferente? Será que está sob seleção positiva?)



# O QUE PODEMOS FAZER COM O CODEML?



Investigar o sinal evolutivo: quais sítios específicos estão sob seleção positiva?

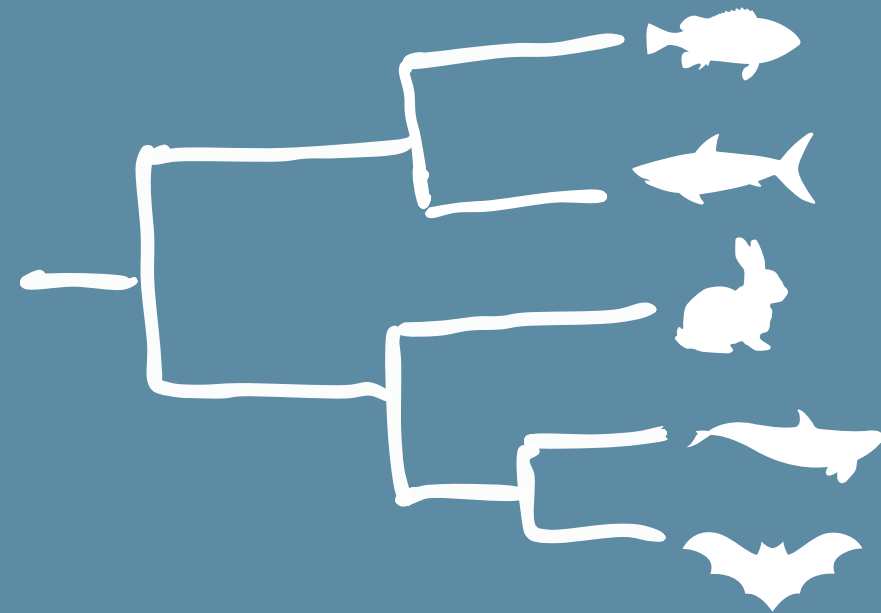
se a hipótese nula (evolução neutra) for rejeitada, significa que alguns sítios estão evoluindo adaptivamente - mas quais???

# MODELOS DE RAMO

Estimam o valor de  $\omega$  apenas para ramos



Model 0

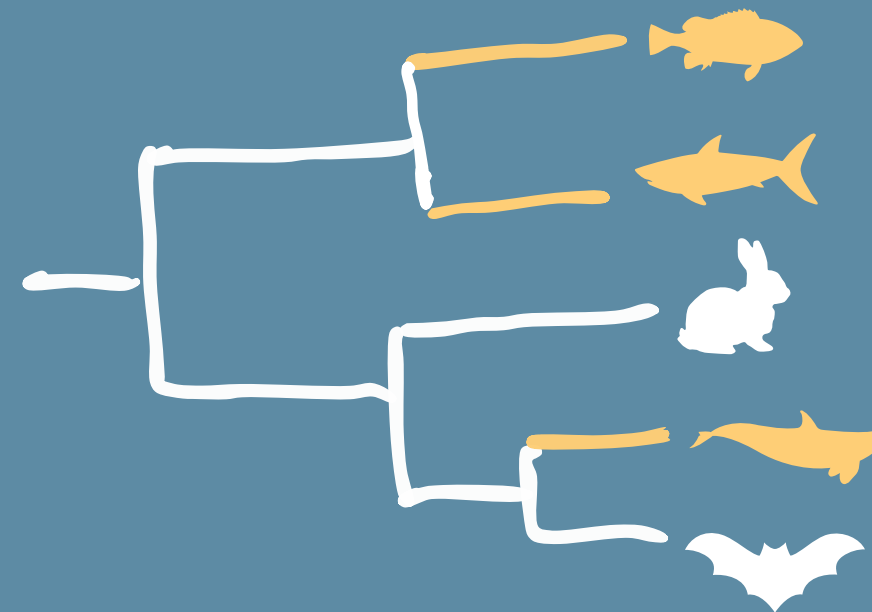


$\omega_1$

por exemplo...

0.324

2-model



$\omega_1$

$\omega_2$

0.324

0.563

Será que  
alguns ramos  
evoluíram  
diferente?



# MODELOS DE SÍTIO

Estimam o valor de  $\omega$  apenas para **sítios**

Model M1a

AAGTCCGAGCTG

classe 1  
 $w < 1$

classe 2  
 $w = 1$

Model M2a

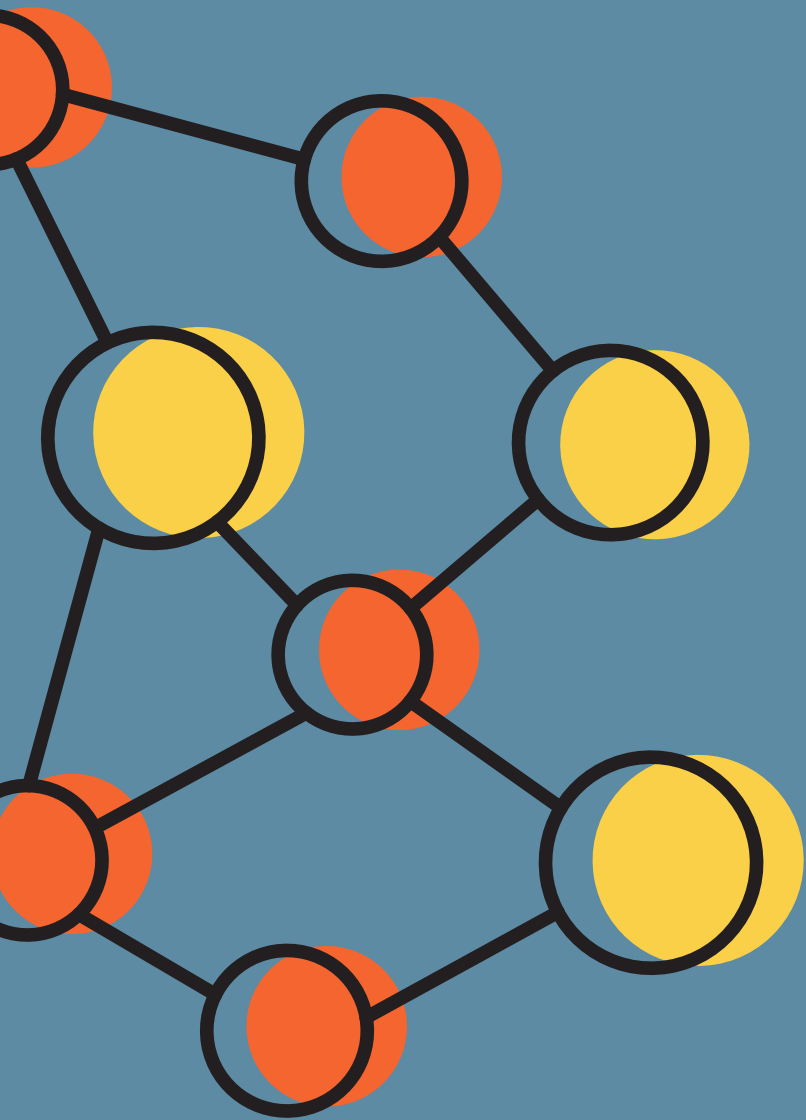
AAGTCCGAGCTG

classe 1  
 $w < 1$

classe 2  
 $w = 1$

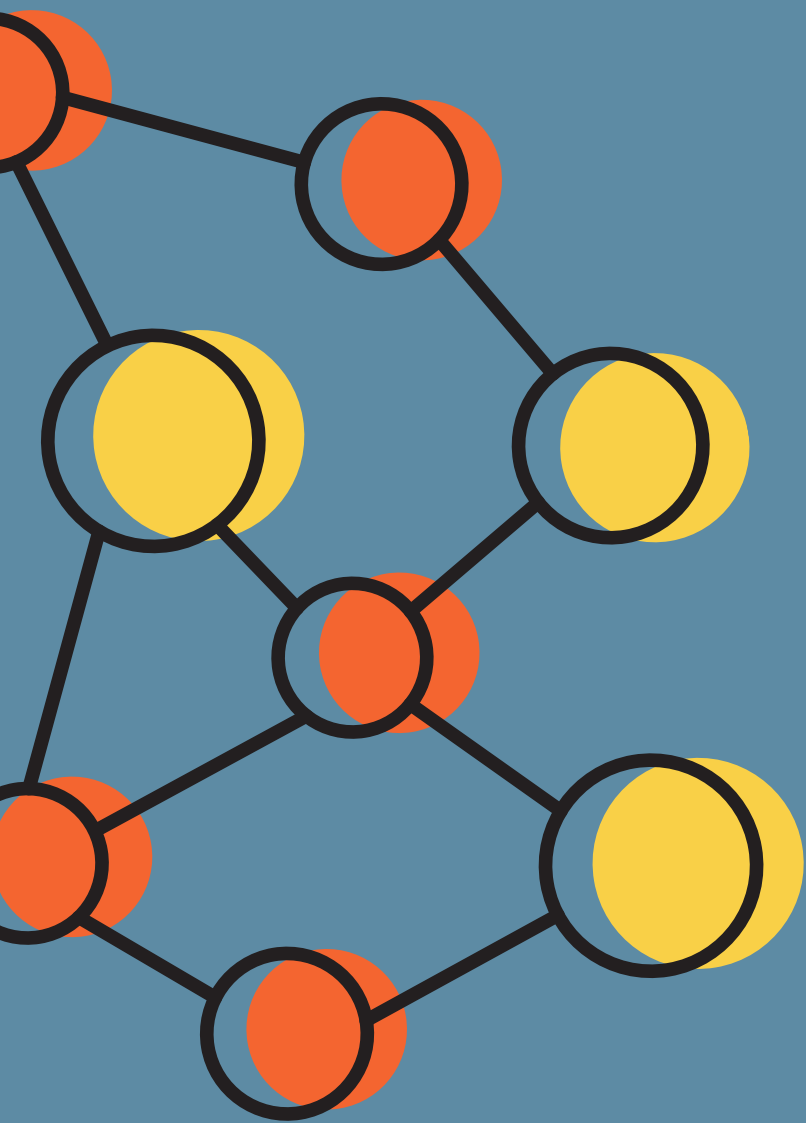
classe 3  
 $w > 1$

Será que alguns sítios na sequência evoluíram diferente, ou sob seleção positiva?

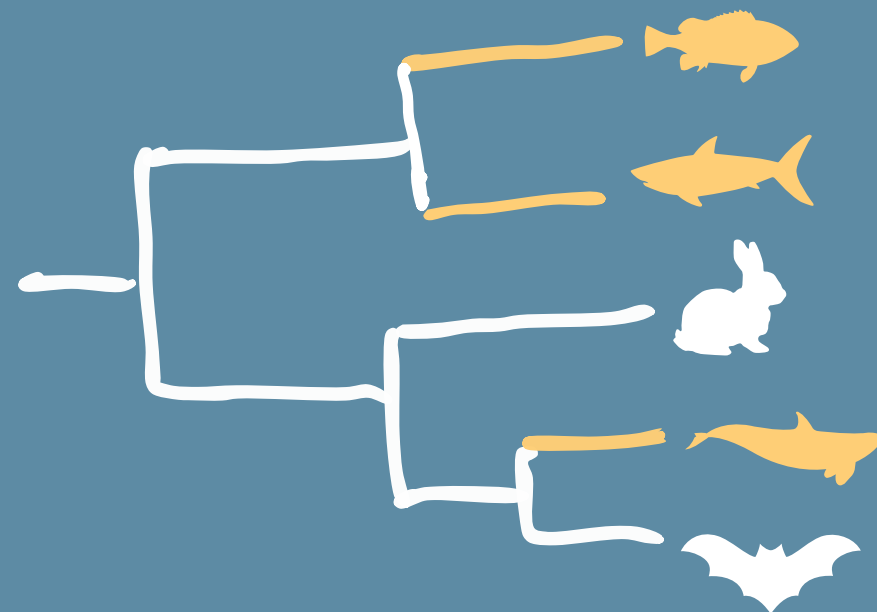


# MODELOS DE RAMO-SÍTIO

Estimam o valor de  $\omega$  em ramos e sítios



A-model null



Classes 2a e 2b:  
 $w = 1$  (foreground)  
 $w < 1$  ou  $w = 1$

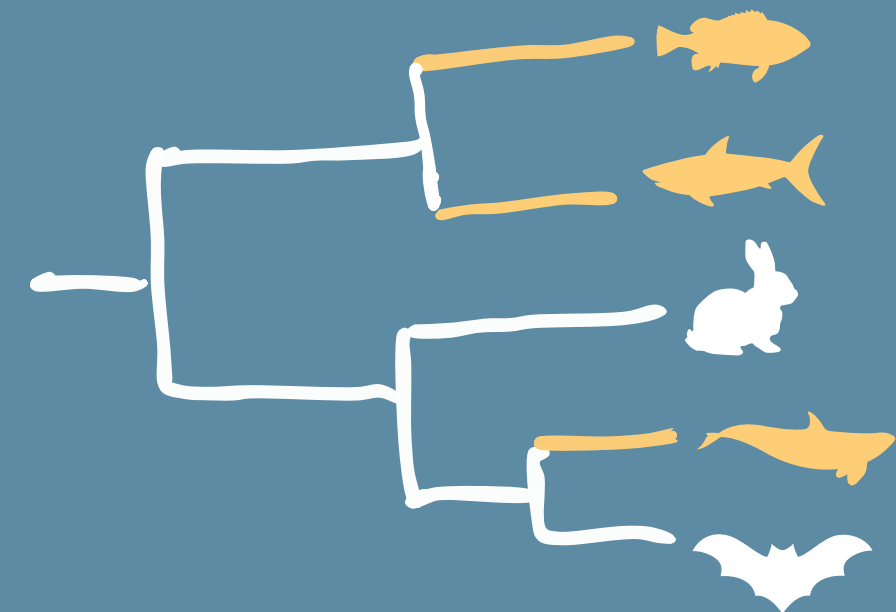
Classe 0

$w < 1$

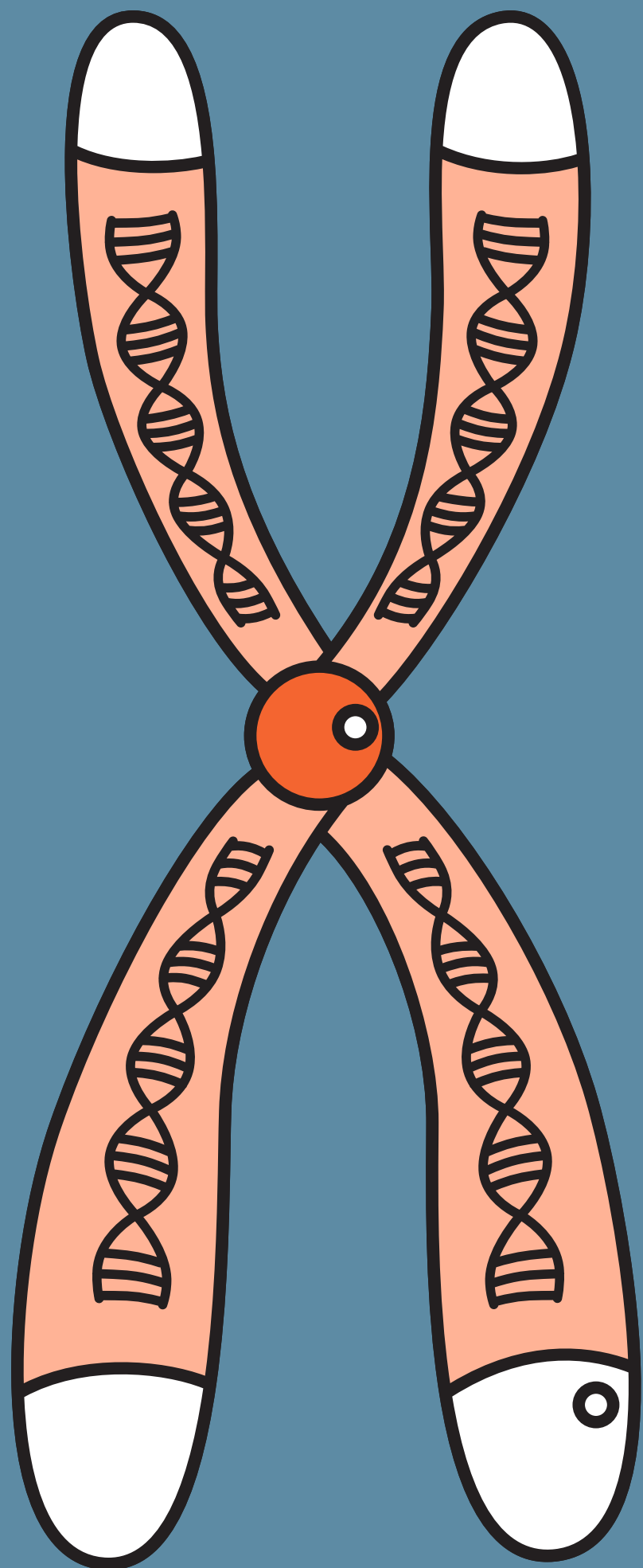
Classe 1

$w = 1$

A model



Classes 2a e 2b  
 $w > 1$  (foreground)  
 $w < 1$  ou  $w = 1$



# COMO RODAMOS O CODEML?

1. Faça uma pasta para cada modelo contendo: um **alinhamento**, uma **árvore de gene** e o **arquivo controle**
2. Verifique se o alinhamento não tem gaps, códons de parada e que os nomes são os mesmos que na árvore
3. Rode o arquivo controle dentro da sua pasta, com o comando: **codeml control\_file.ctl**

# FAZENDO UM ALINHAMENTO

## USANDO O 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 colors including orange, red, and white. The overall style is clean and modern, set against a solid blue background.

# VISUALIZANDO UM ALINHAMENTO

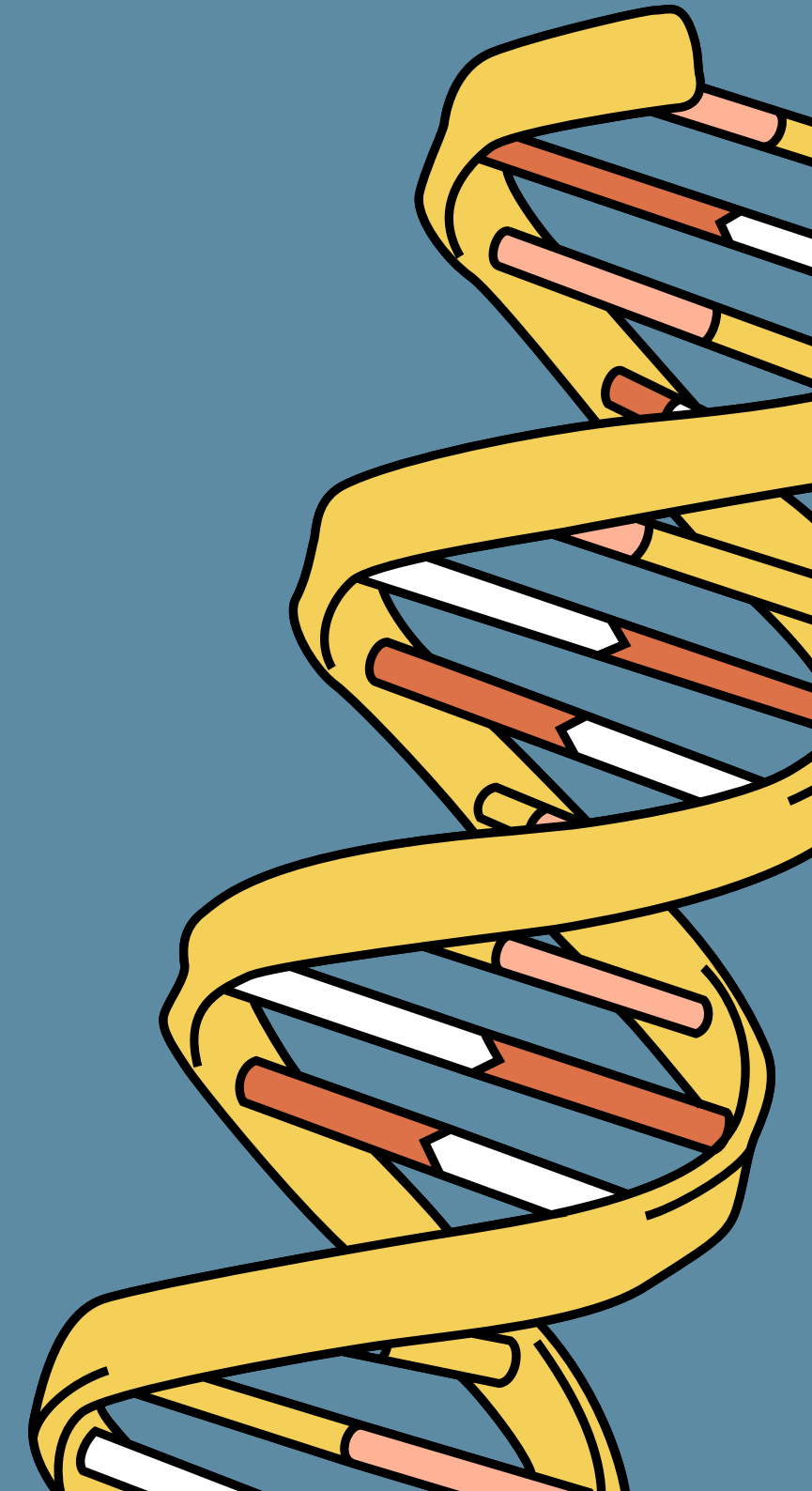
## USANDO O ALIVIEW





# ESTIMANDO UMA ÁRVORE

## USANDO O IQTREE



IQTREE Web Server: Fast and accurate phylogenetic trees under maximum likelihood

Not Secure | iqtree.cibiv.univie.ac.at

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

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:

FreeRate heterogeneity: ☐ Yes [+R]

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

#rate categories:

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:

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

Maximum iterations:

Minimum correlation coefficient:

Single branch tests:

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

Approximate Bayes test: ☐ Yes


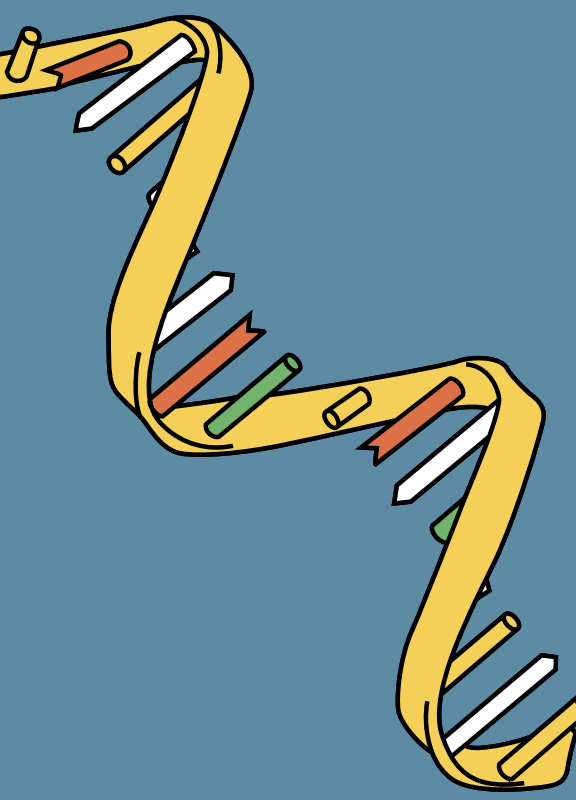
IQ-TREE Search Parameters

Perturbation strength:

IQ-TREE stopping rule:



# 0 ARQUIVO CONTROLE



```
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  $\omega$  for all branches (M0 and site models)
- \* One  $\omega$  for all sites (M0 and branch model)
- \* Use F3x4 model
- \* Assume no clock
- \* Enables option to estimate omega
- \* Initial omega value



# O ARQUIVO CONTROLE

seqfile = Mx\_aln.phy

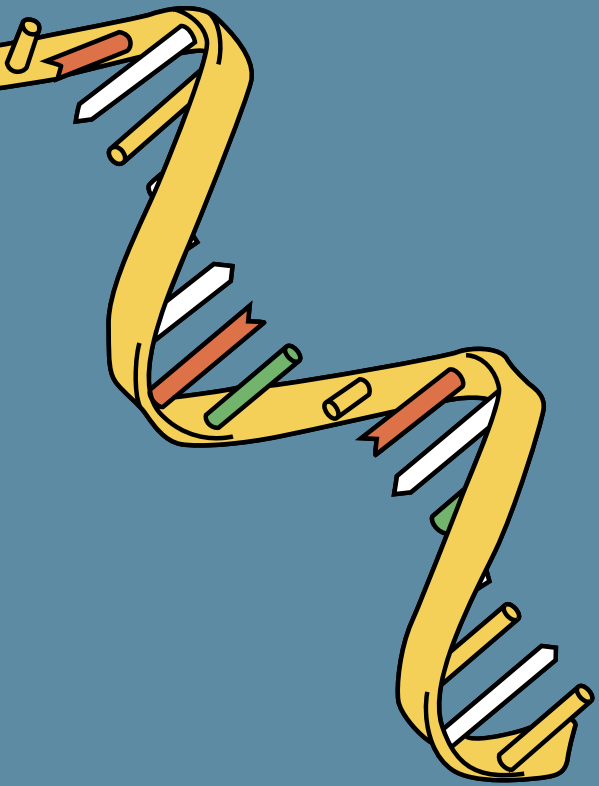
coloque aqui o nome/caminho para a sua sequência

treefile = Mx\_unroot.tree

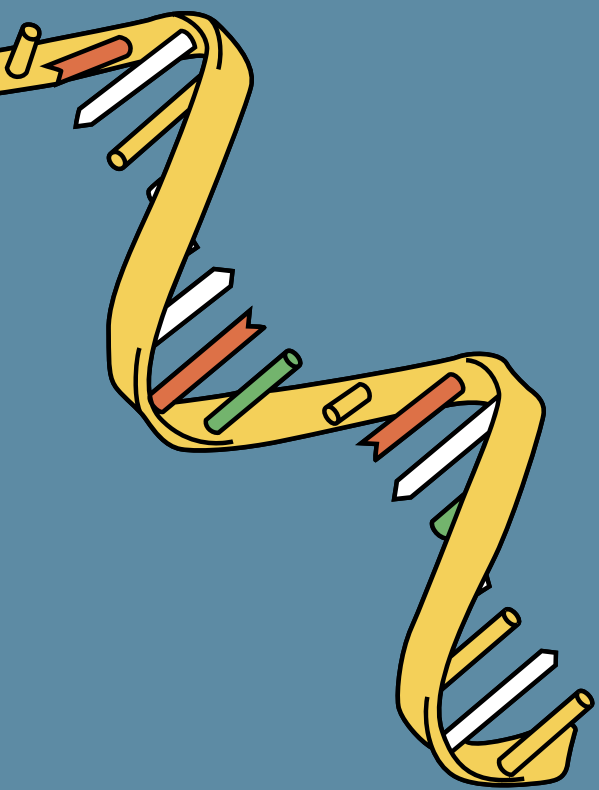
coloque aqui o nome/caminho para a sua árvore

outfile = out\_M0.txt

coloque aqui o nome/caminho para o seu arquivo de resultados



# O ARQUIVO CONTROLE



model = 0

→ coloque aqui o modelo de ramo que  
você quer rodar

NSsites = 0

→ coloque aqui o modelo de sítio que  
você quer rodar



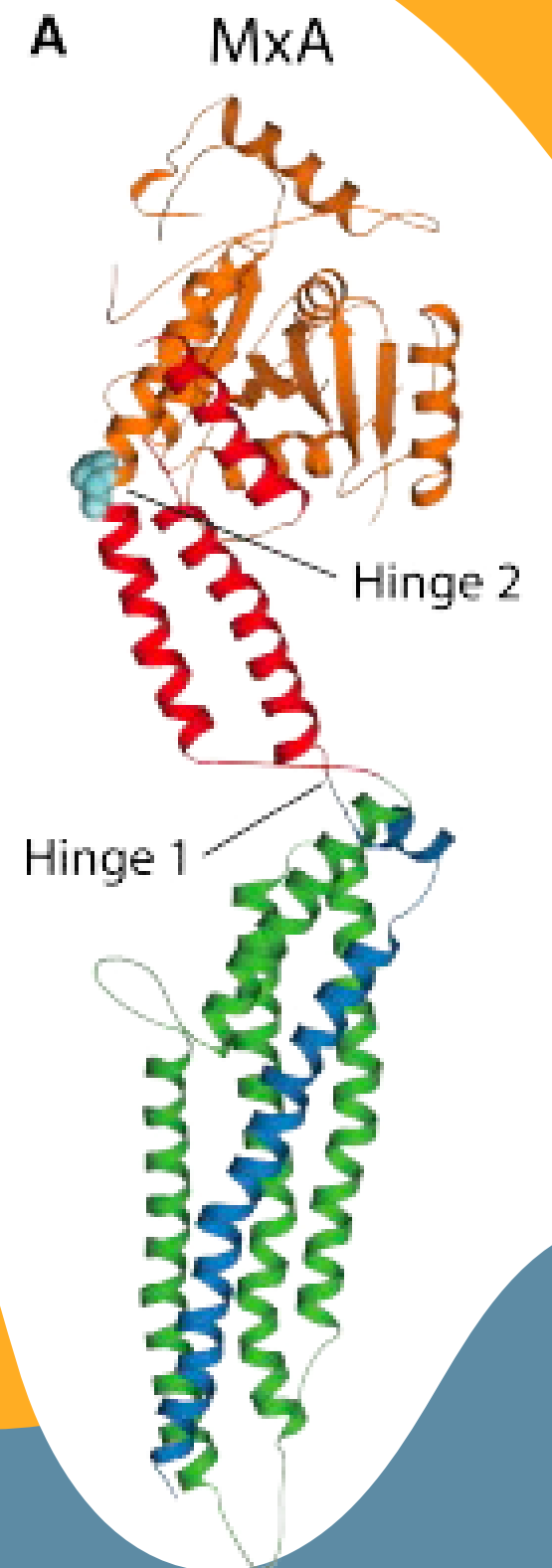
Mudando as duas opções ao mesmo tempo, especificamos  
o modelo de ramo-sítio!

# COMO USAMOS O CODEML PARA TESTAR UMA HIPÓTESE?

O gene myxovirus é responsável pela resposta antiviral em várias espécies

Ele produz uma proteína chamada Mx  
(myxovirus resistance protein)

Estrutura cristalina da  
proteína Mx em humanos

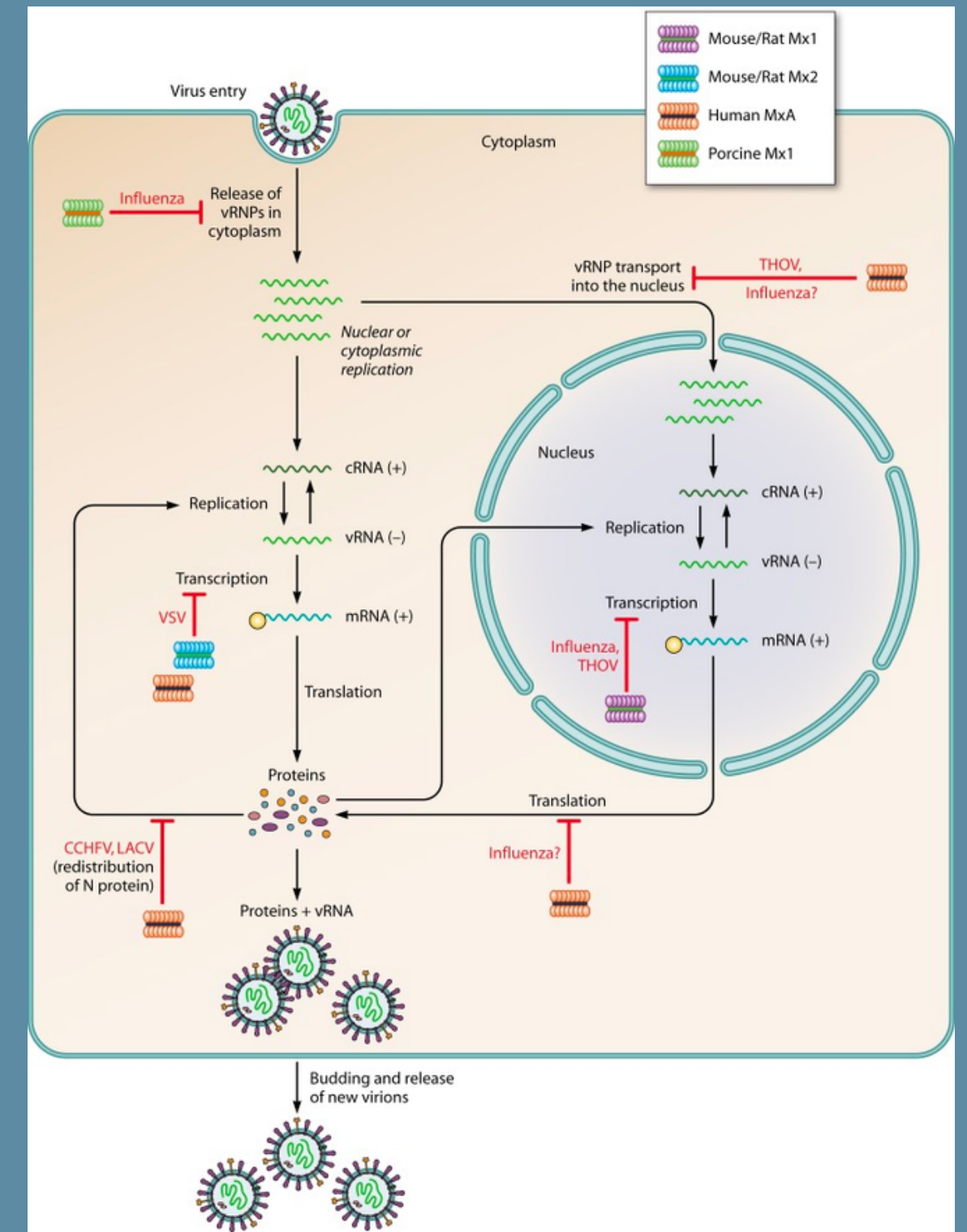


# COMO USAMOS O CODEML PARA TESTAR UMA HIPÓTESE?

O gene myxovirus é responsável pela resposta antiviral em várias espécies

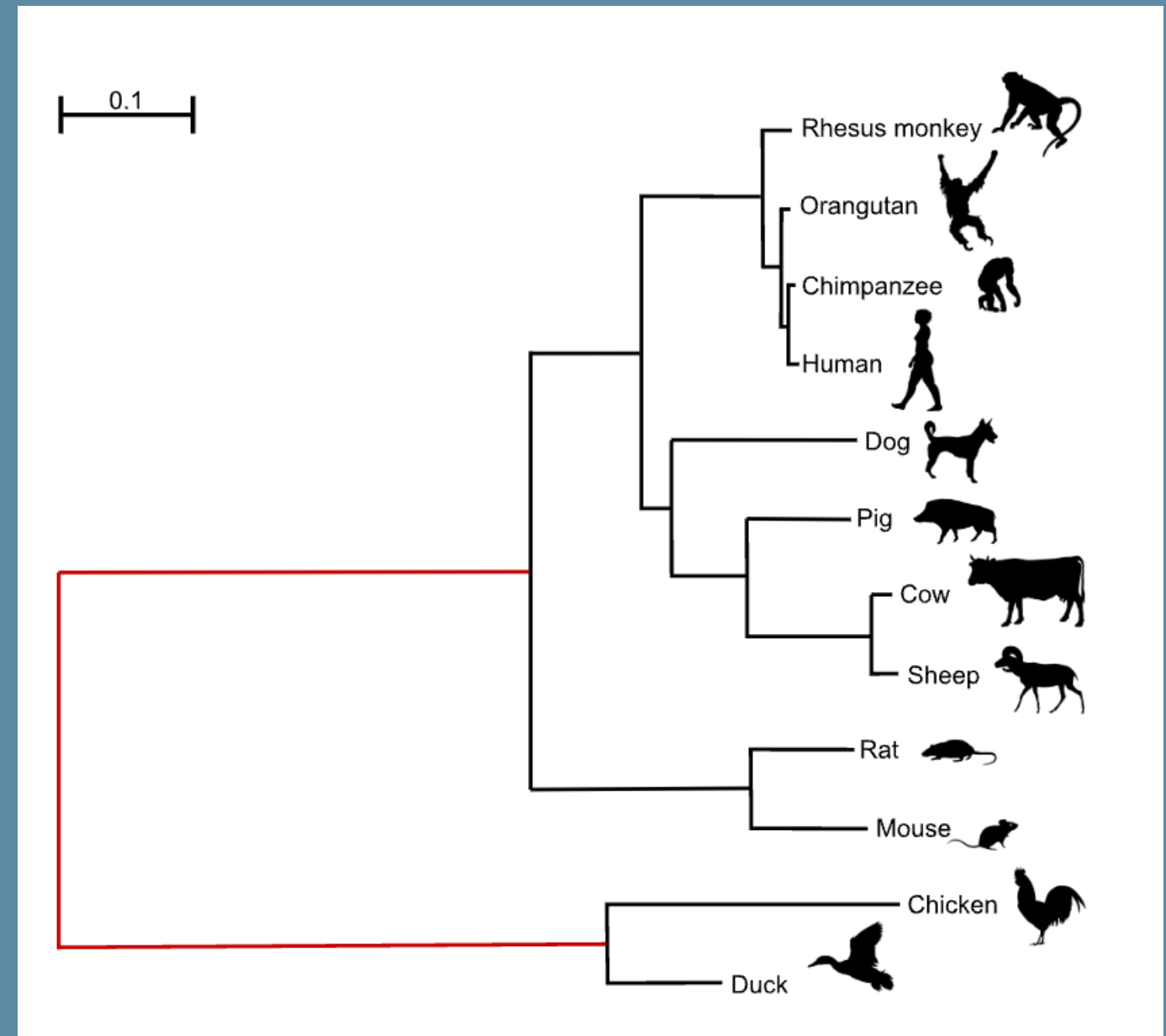
- A proteína Mx tem de 1-3 isoformas, (dependendo da espécie) com diferentes atividades antivirais e localização na célula
- A sua expressão na célula varia entre mamíferos (núcleo) e aves (citoplasma)

Adaptação a patógenos específicos?



# COMO USAMOS O CODEML

## PARA TESTAR UMA HIPÓTESE?



# COMO USAMOS O CODEML PARA TESTAR UMA HIPÓTESE?

Perguntas:

- Será que o Mx evoluiu nessas espécies para combater patógenos específicos de cada uma?
- Que fatores influenciaram a evolução do Mx em diferentes linhagens de animais?

