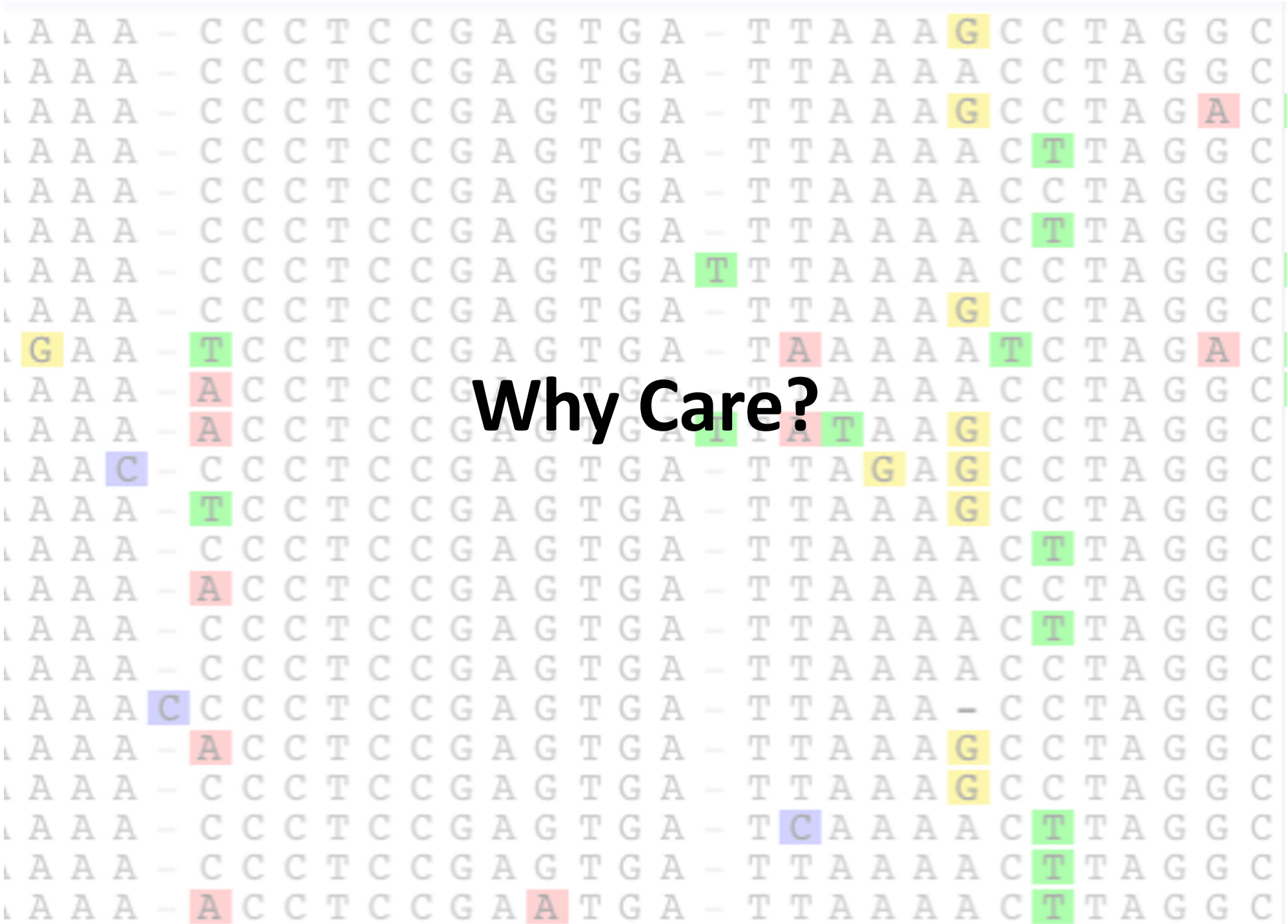
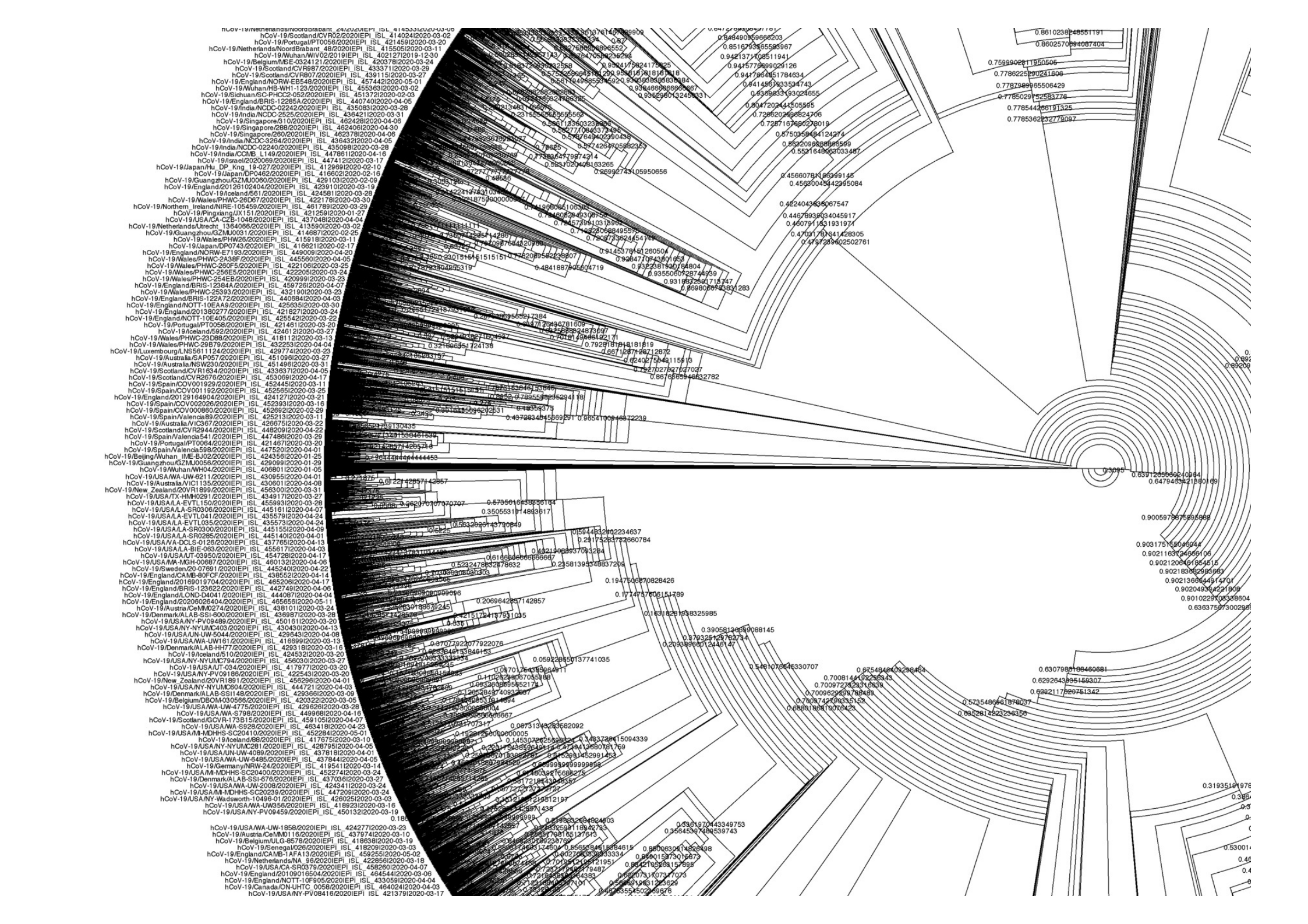
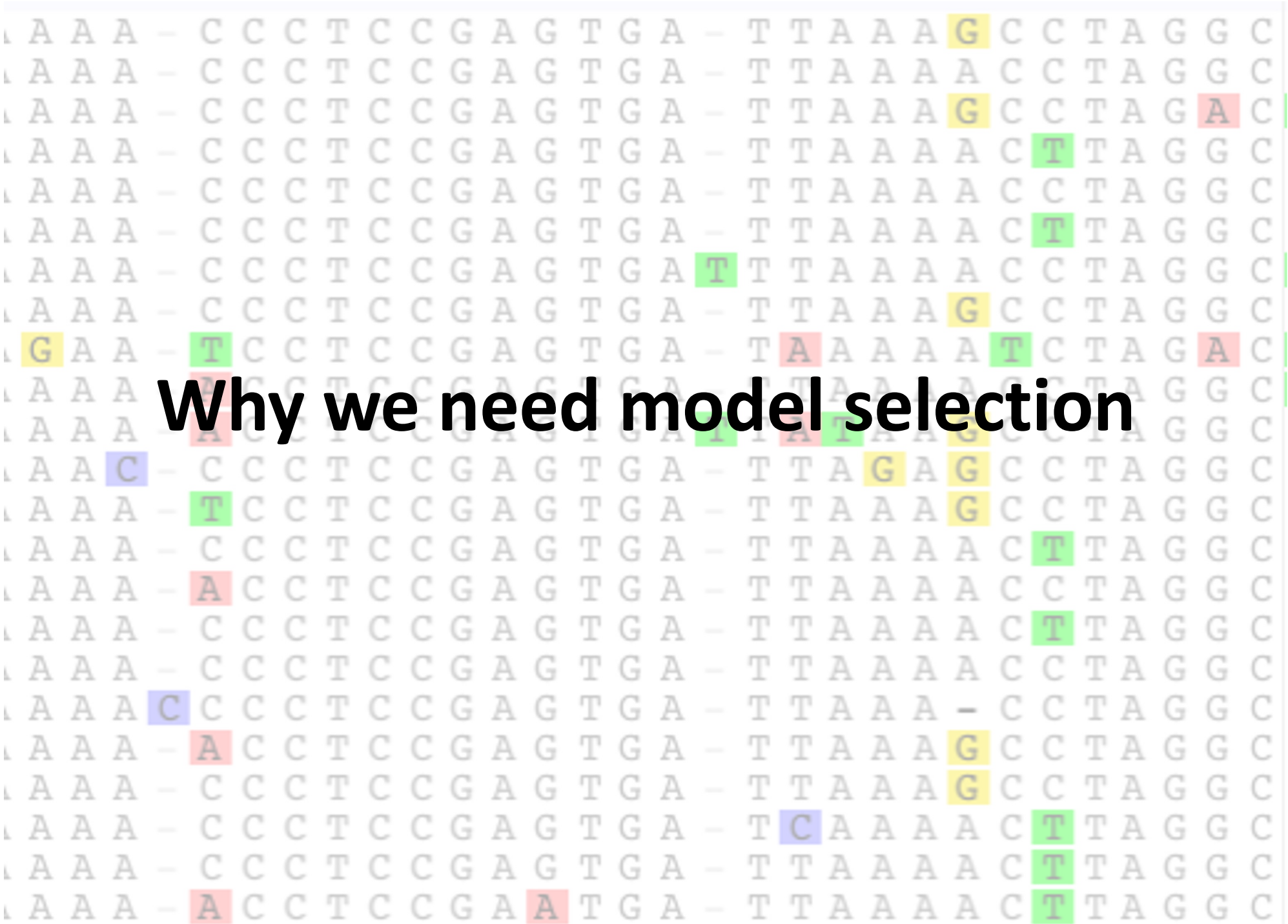
Models of molecular evolution, and how to choose a good one

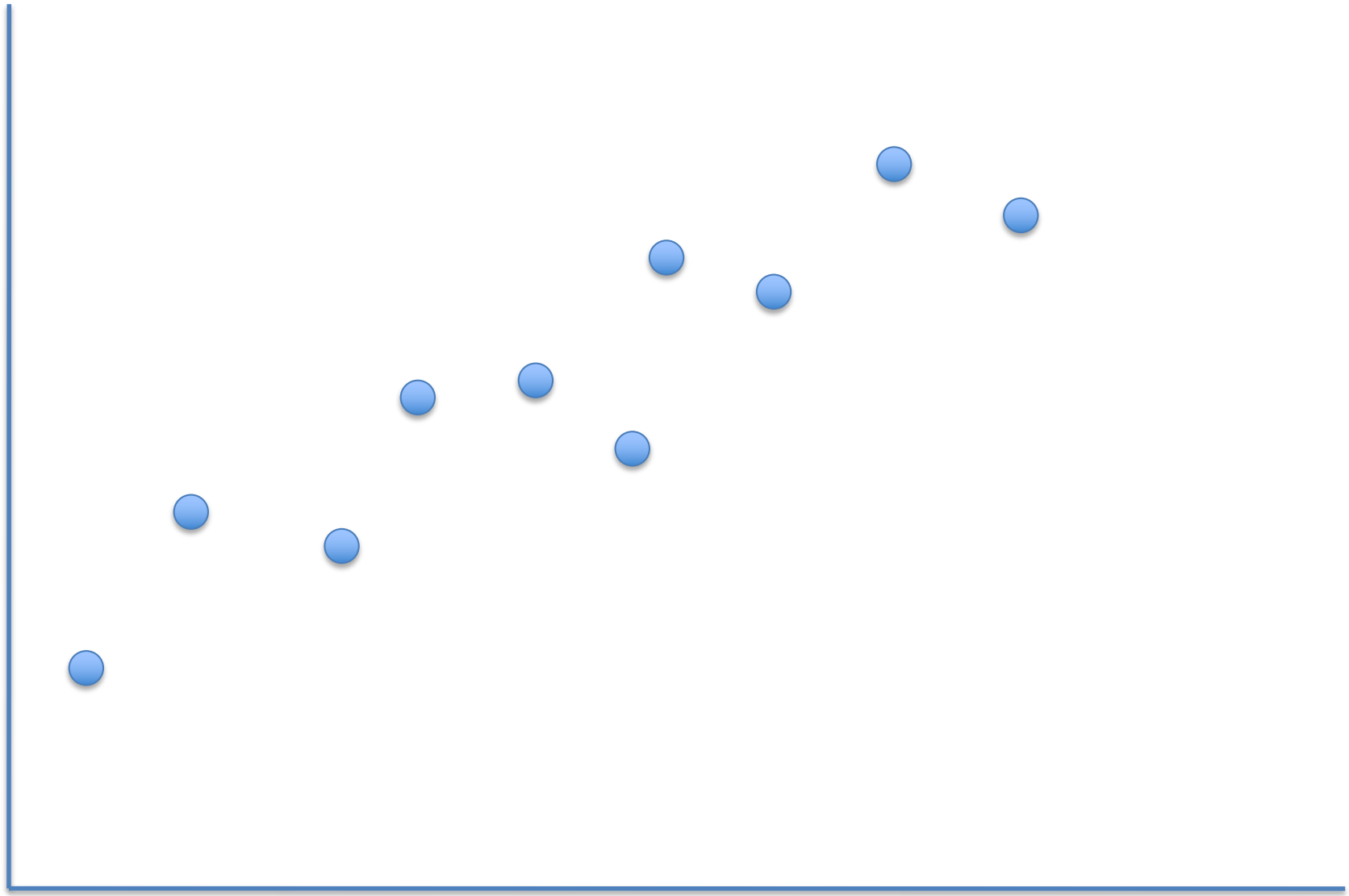
Rob Lanfear
Ecology and Evolution
Australian National University

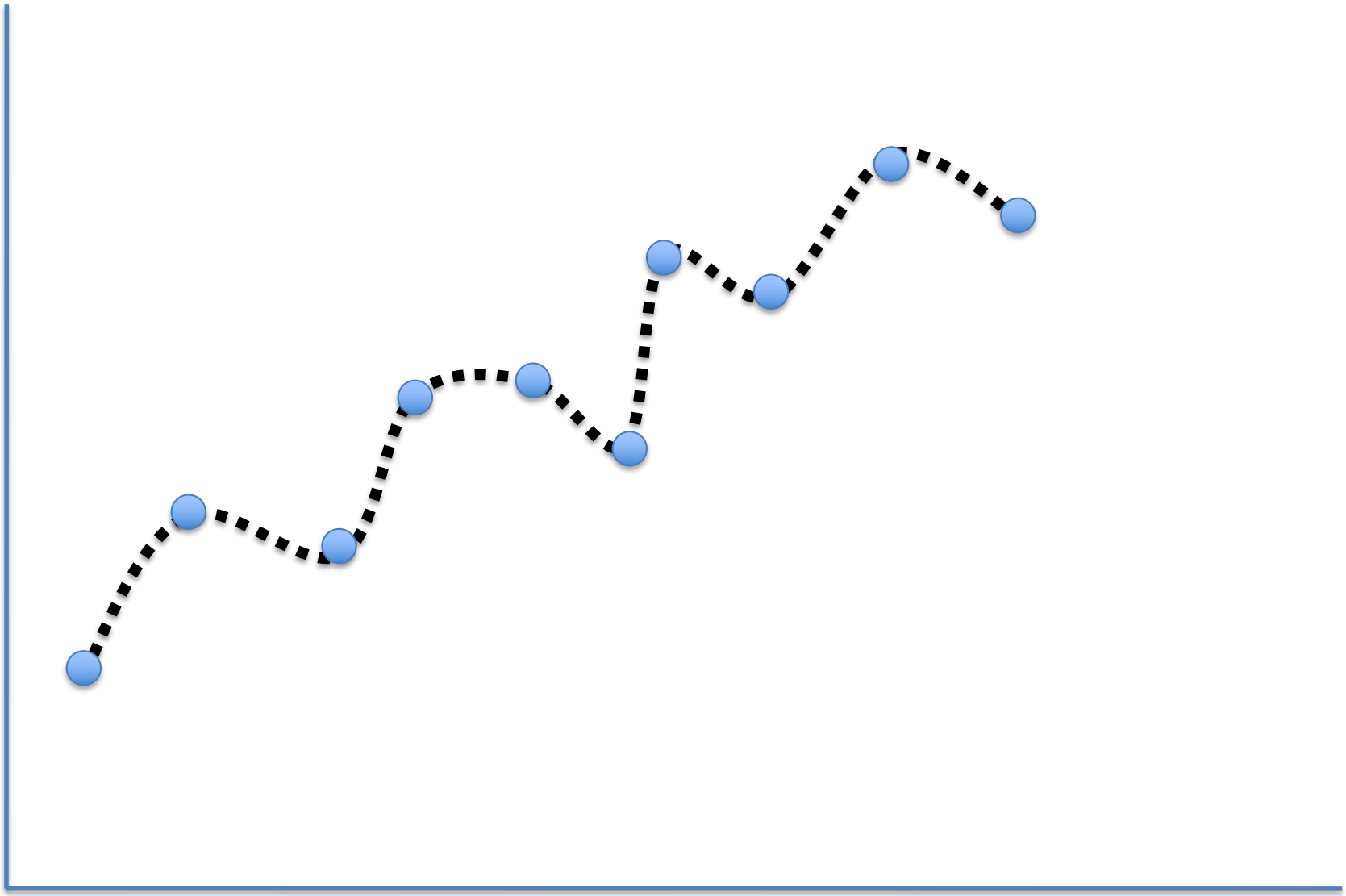

Why Care?

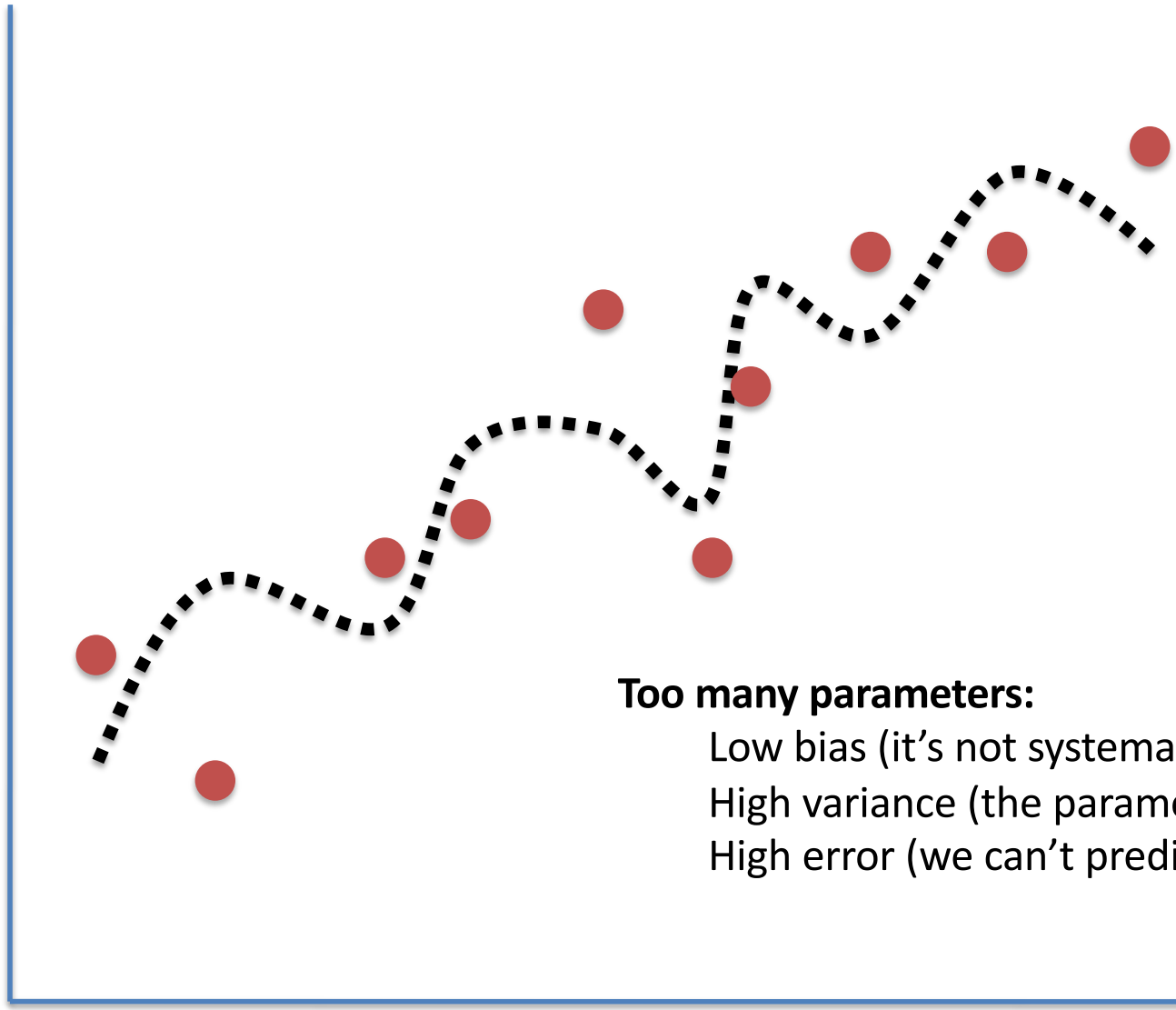




Why we need model selection





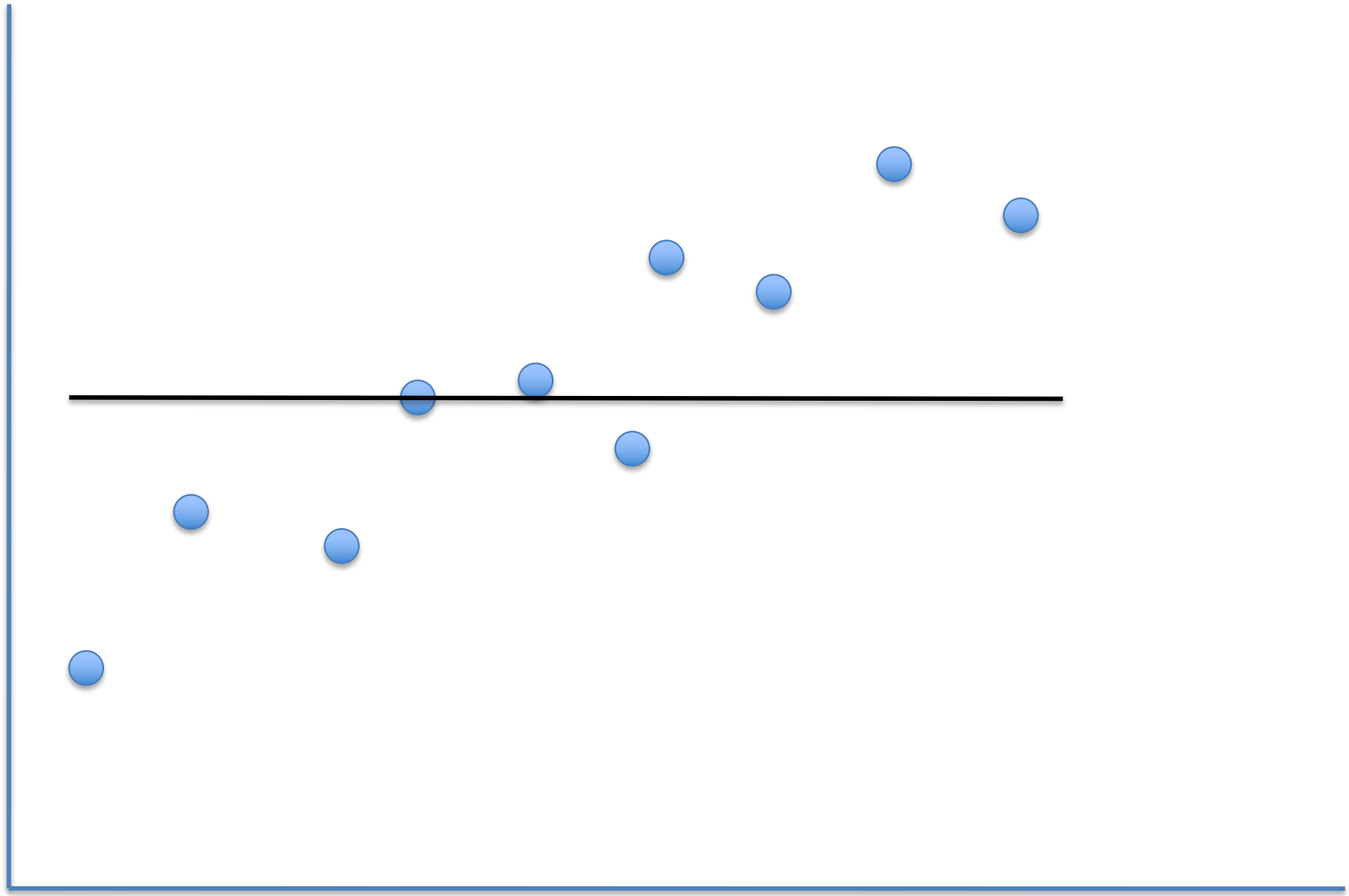


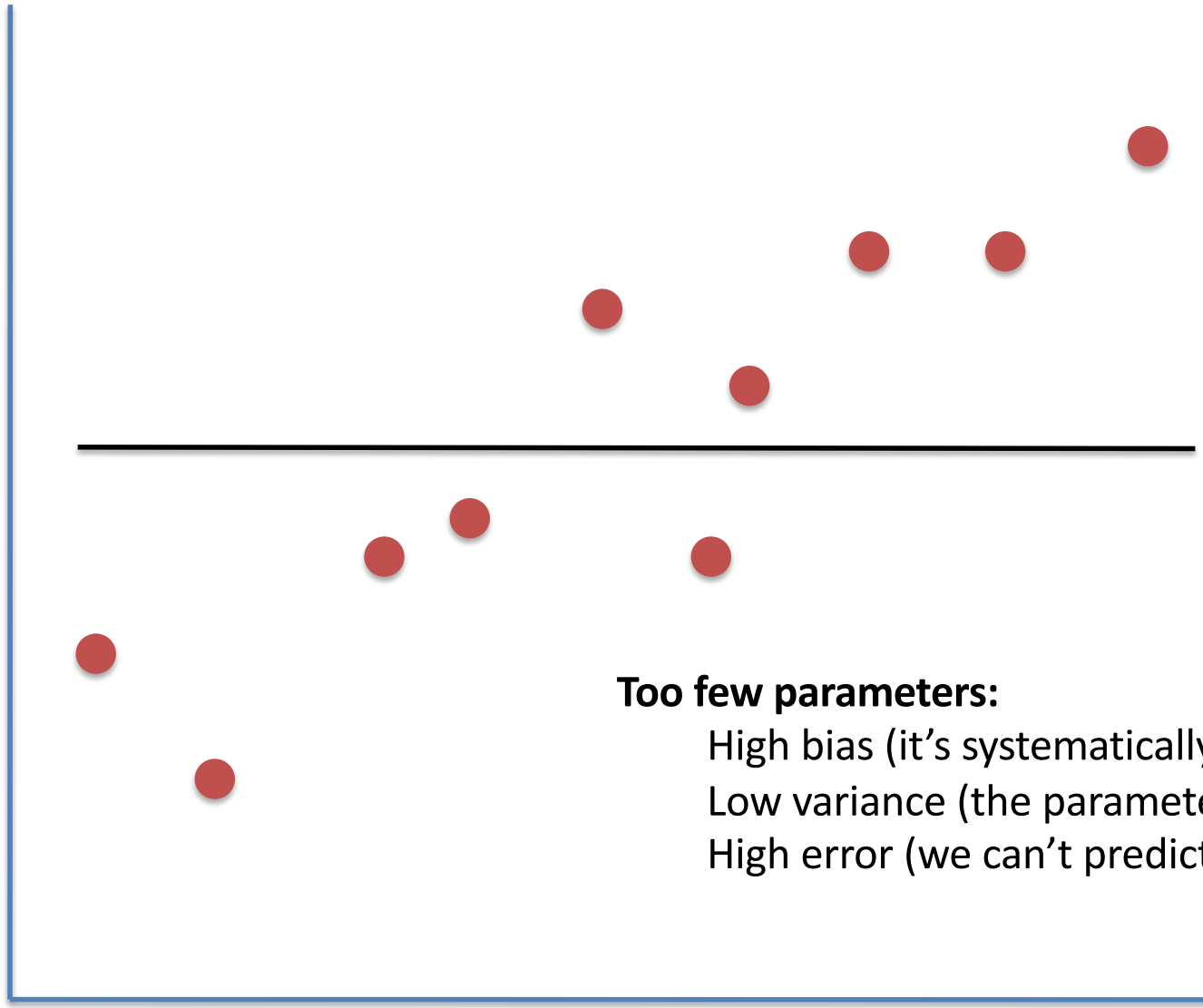
Too many parameters:

Low bias (it's not systematically wrong)

High variance (the parameters have to change a lot)

High error (we can't predict the truth accurately)



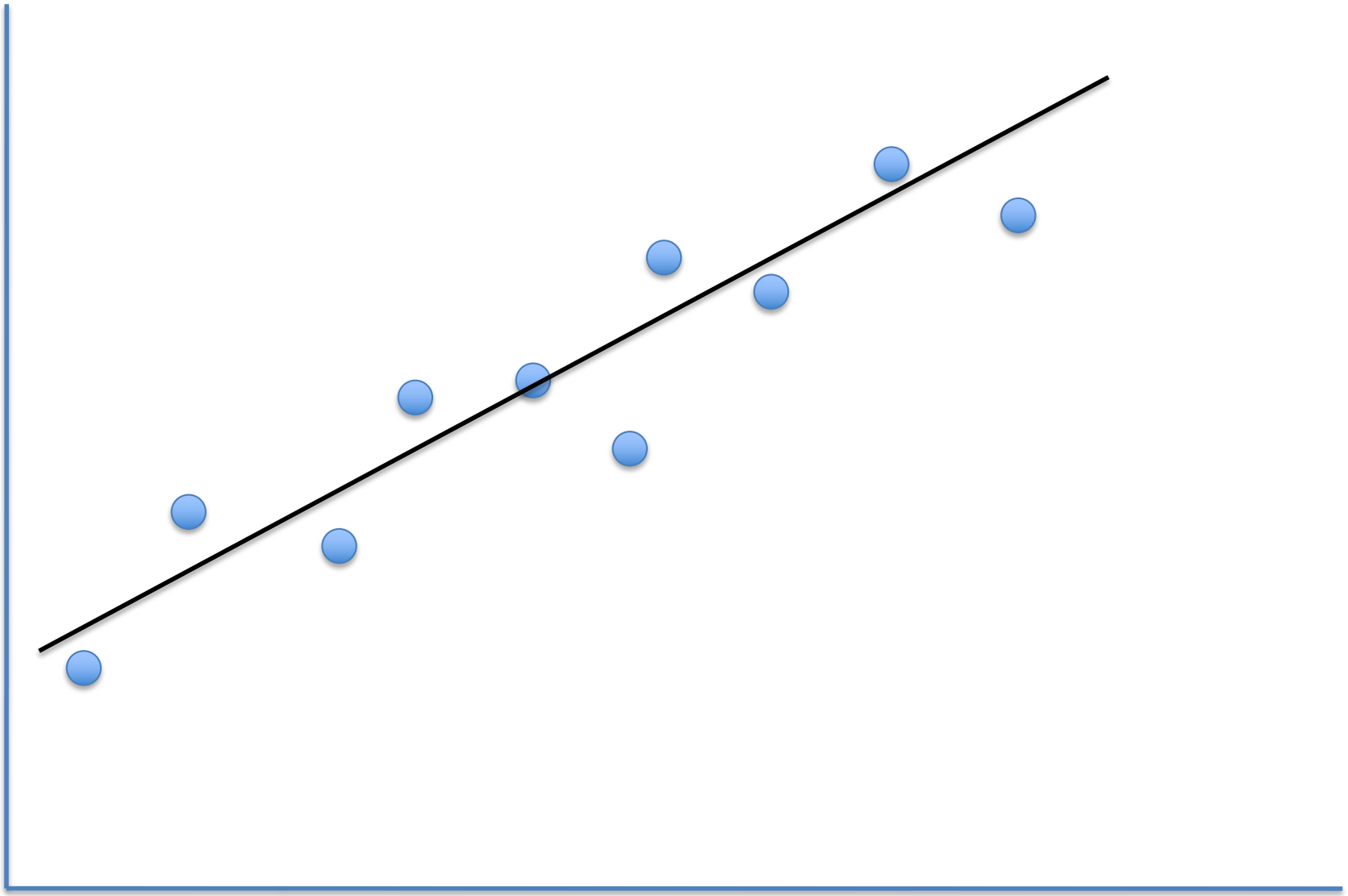


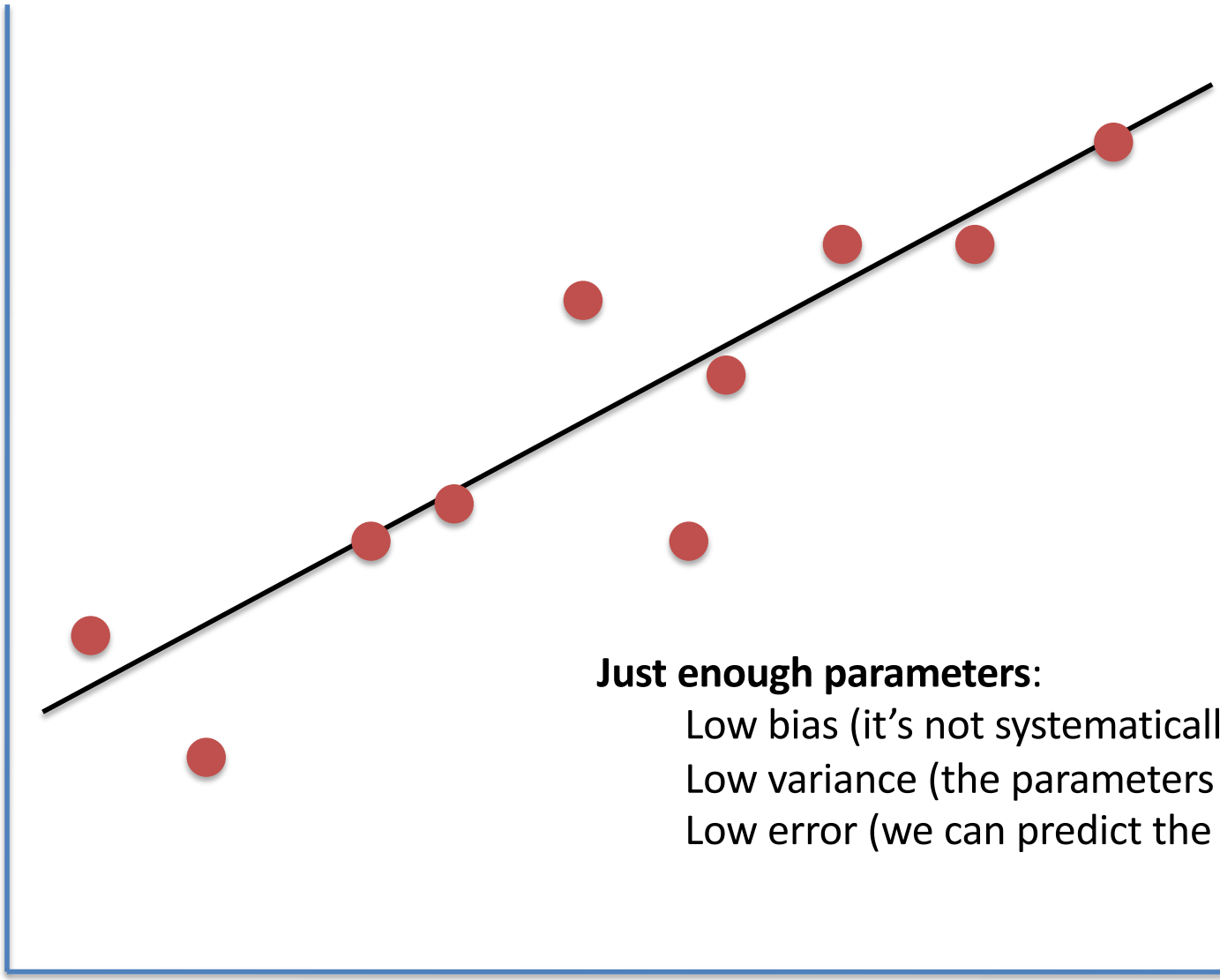
Too few parameters:

High bias (it's systematically wrong)

Low variance (the parameter doesn't change)

High error (we can't predict the truth accurately)





Just enough parameters:

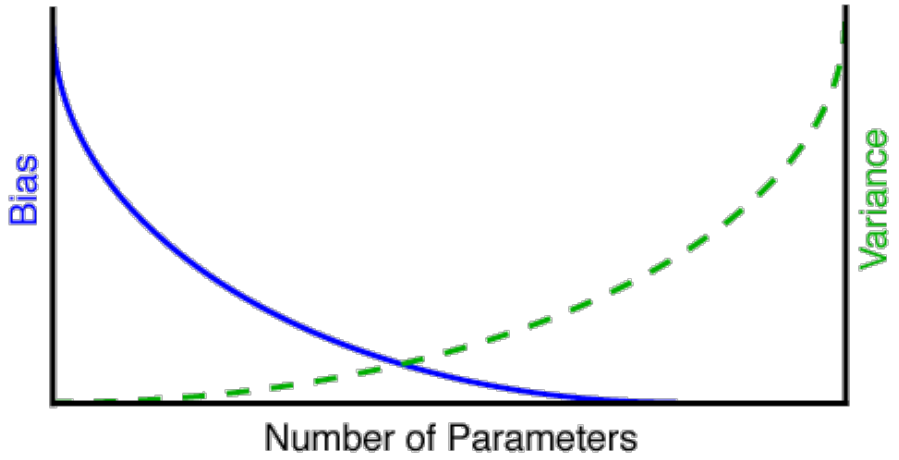
Low bias (it's not systematically wrong)

Low variance (the parameters don't have to change)

Low error (we can predict the truth OK)

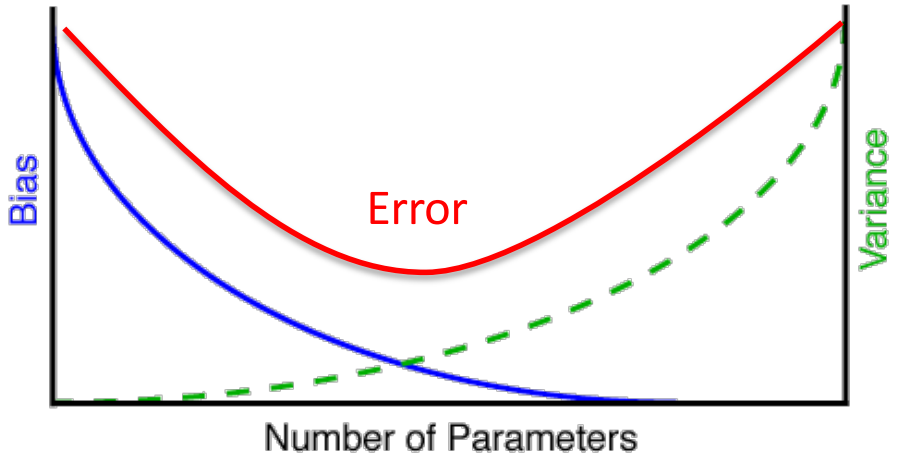
Key concepts

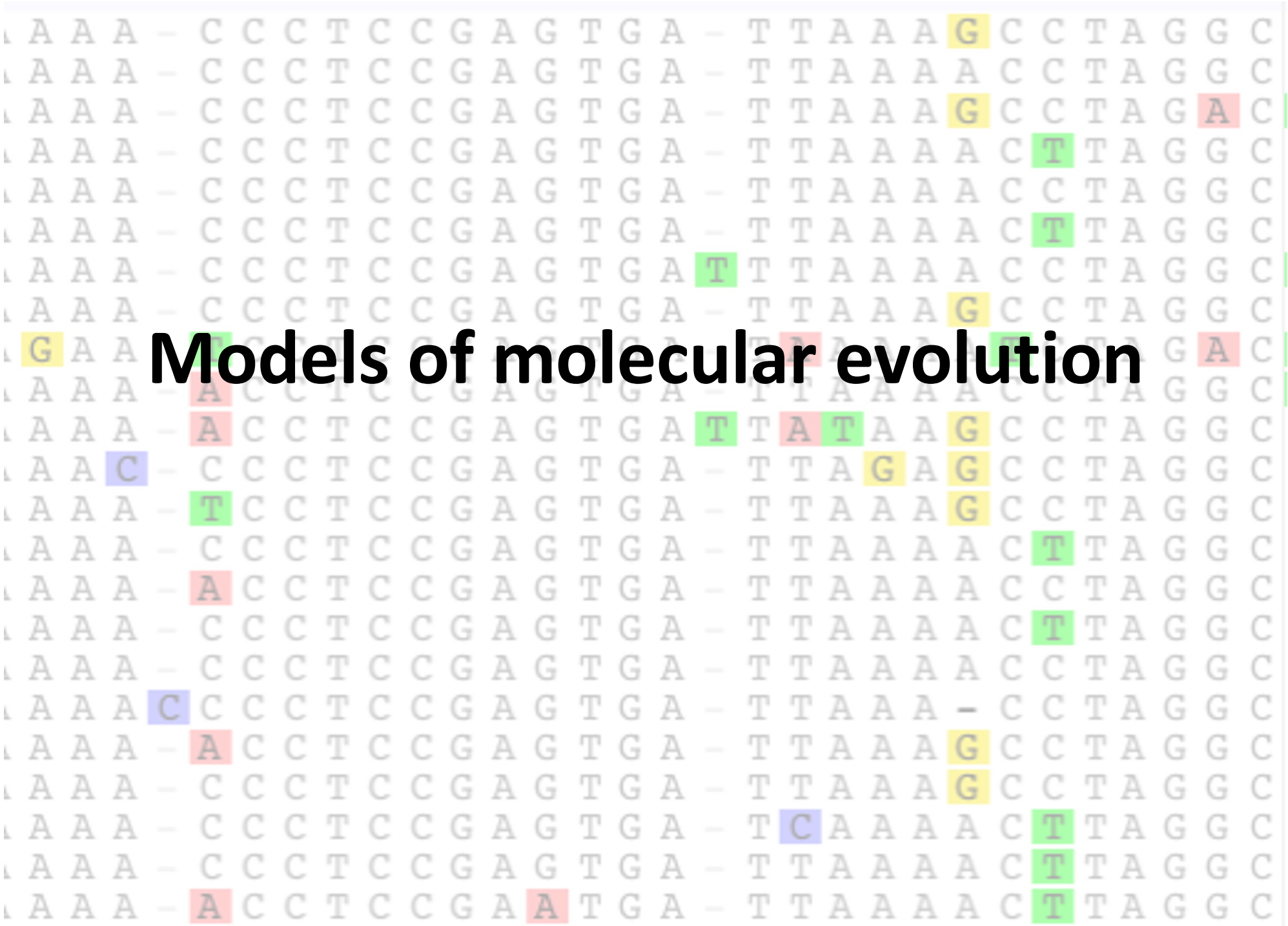
- Adding more parameters ALWAYS improves the fit of the model to the observed data
- E.g. more parameters -> higher R^2
- E.g. more parameters -> better likelihood
- But it doesn't necessarily improve the model!



Key concepts

- The goal of model selection is to balance bias and variance, so minimise the error of the model

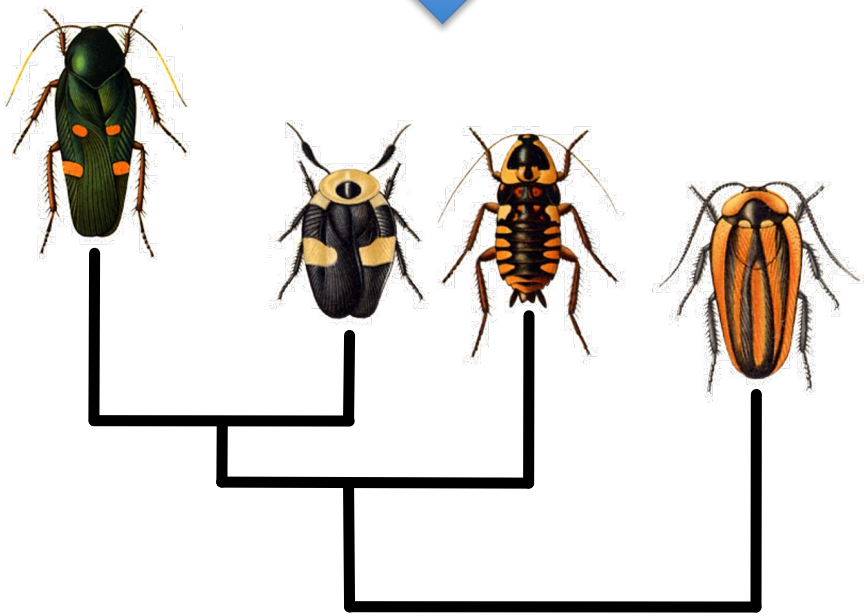




Models of molecular evolution

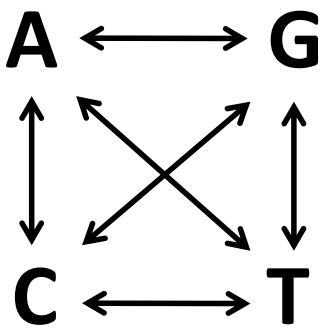
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac

Molecular evolution



```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

Rate Matrix



Base Frequencies

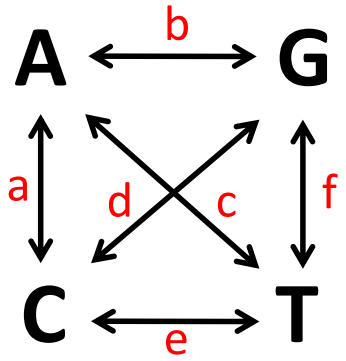
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Site Rates

$$+ I + G + R$$

```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

Rate Matrix



Base Frequencies

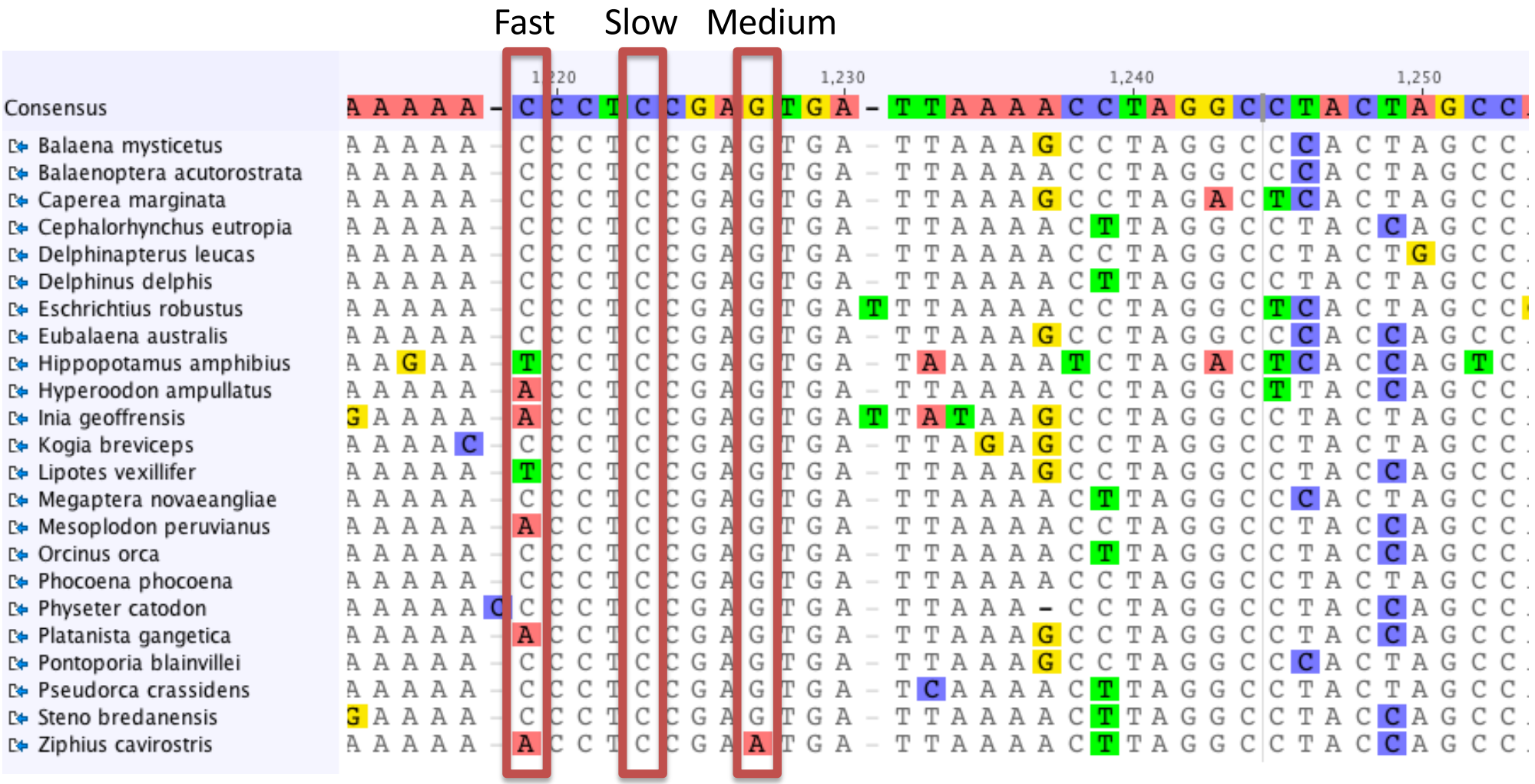
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Site Rates

+ I + G + R

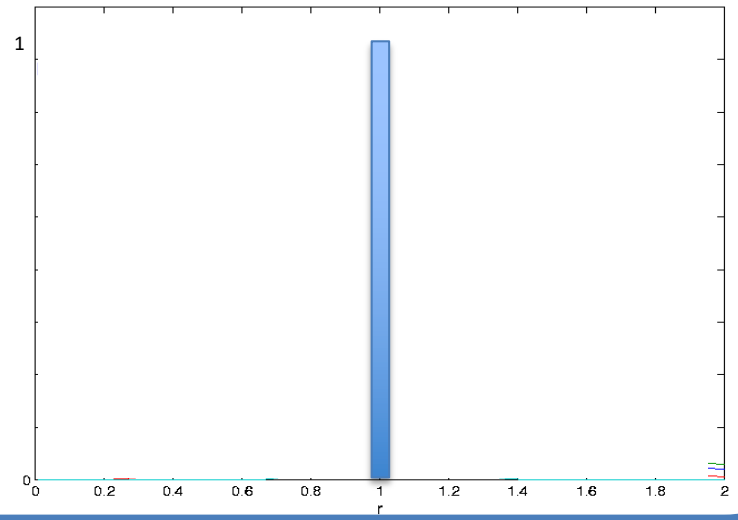
JC	HKY	GTR	GTR+I+G
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$	$\pi_A, \pi_C, \pi_G, \pi_T$
No I or G	No I or G	No I or G	I, G
0 free parameters	4 free parameters	8 free parameters	10 free parameters

How do we model variation in rates of evolution at different sites?



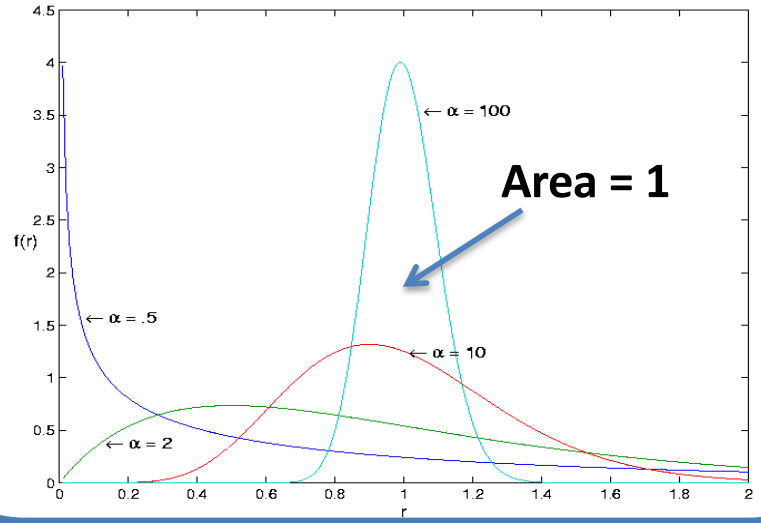
GTR

- All sites evolve at equal rates



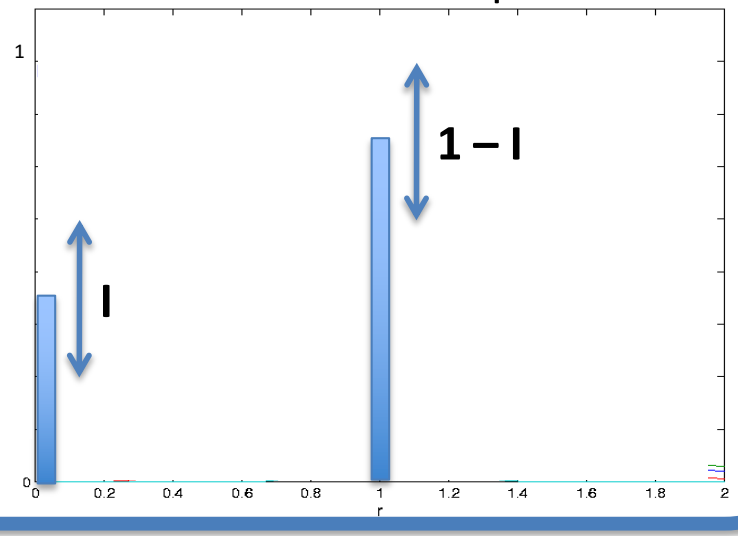
GTR+G

- Rates vary according to gamma dist.



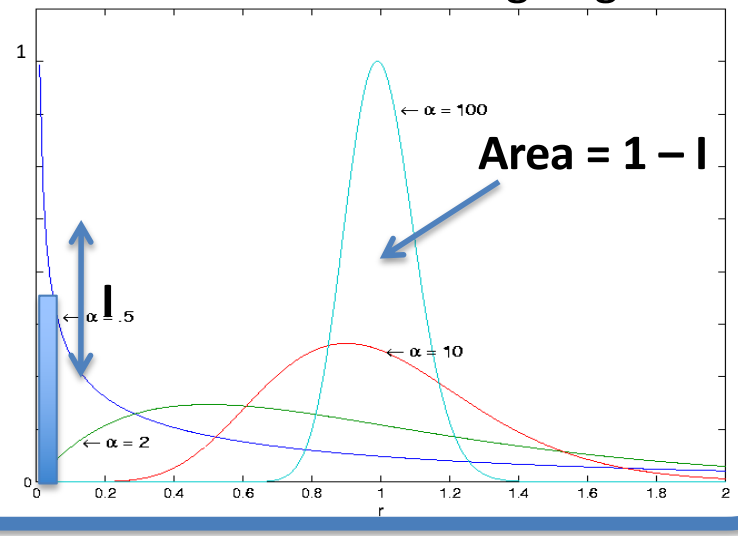
GTR+I

- A proportion of sites don't change
  - The rest evolve at equal rates



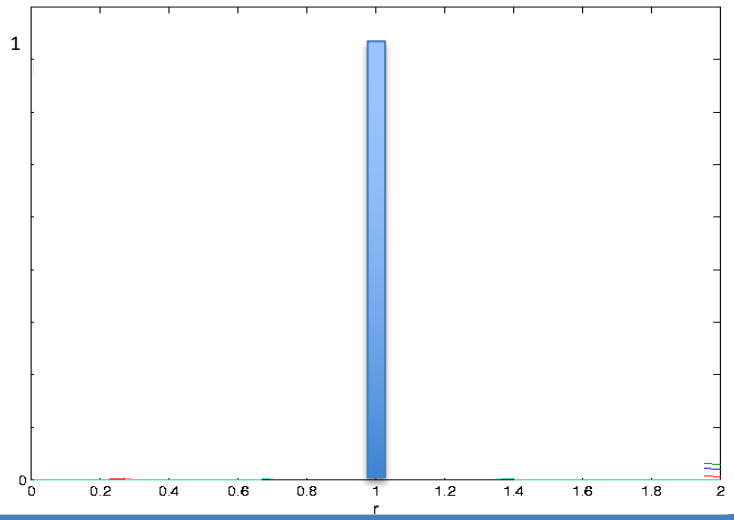
GTR+I+G

- A proportion of sites don't change
- The rest evolve according to gamma



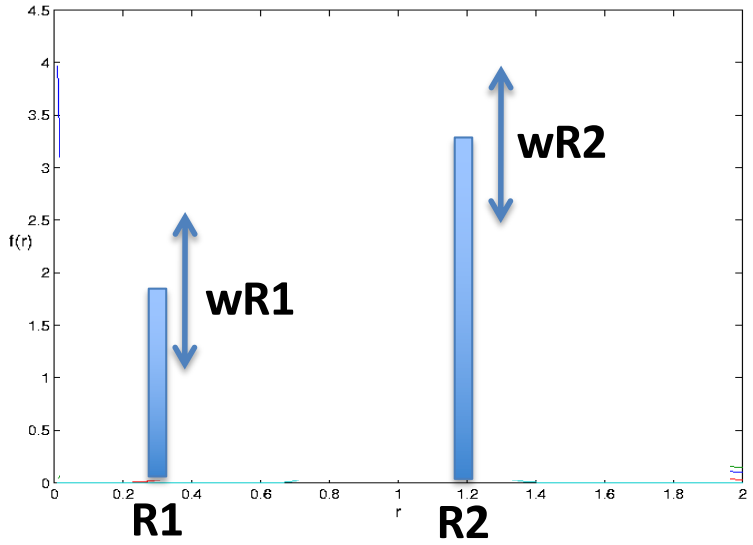
GTR

- All sites evolve at equal rates



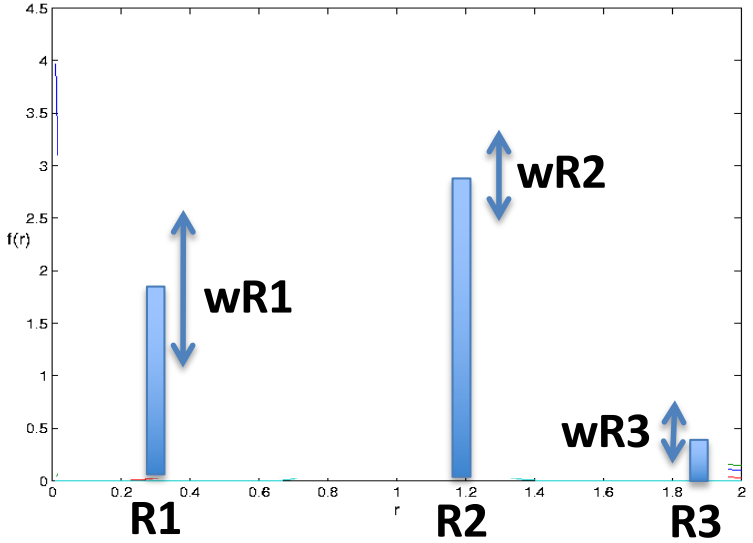
GTR+R2

- Two free rate categories



GTR+R2

- Two free rate categories

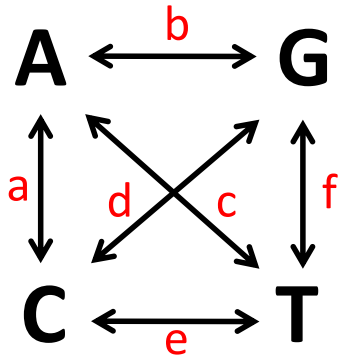


GTR+RX

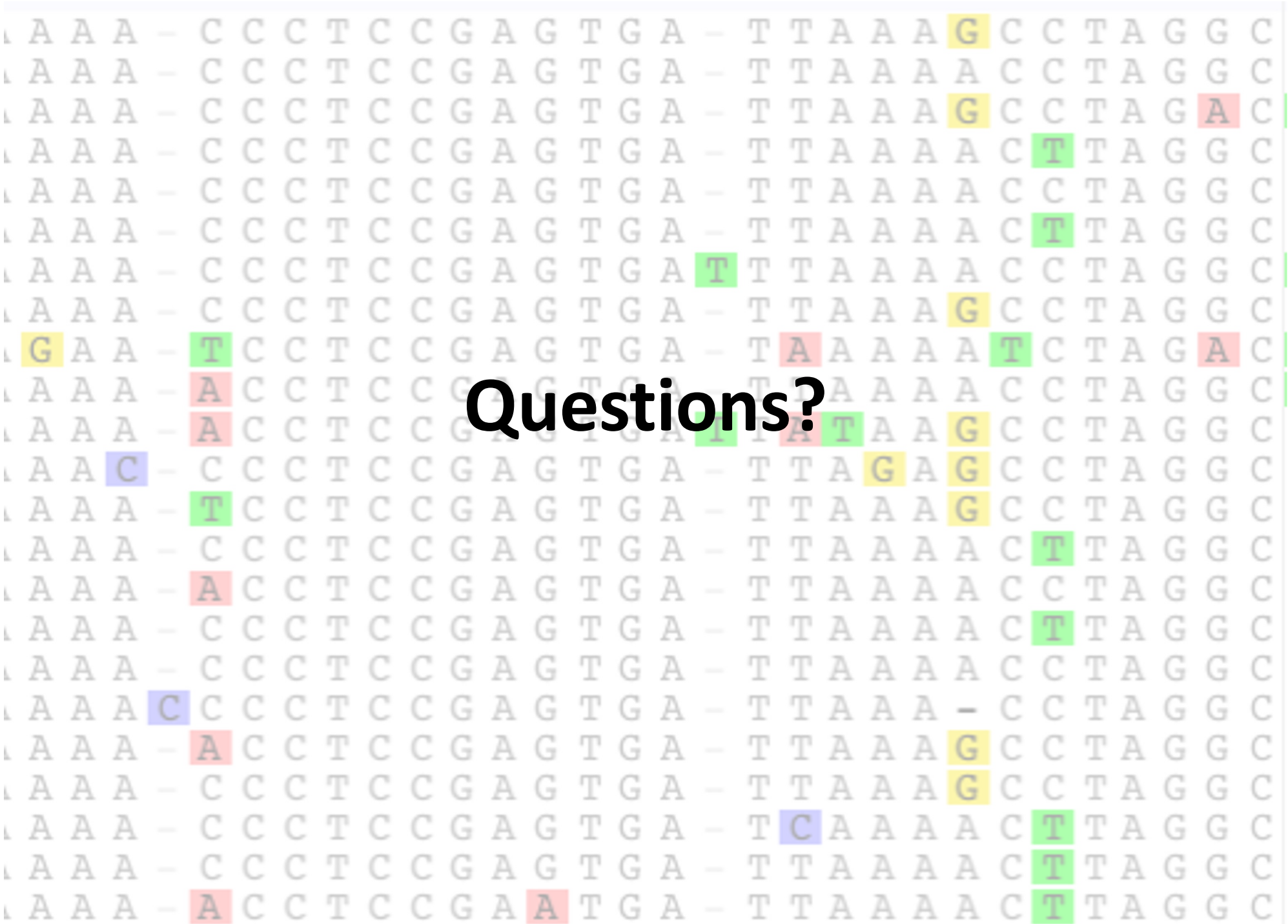
- Add as many rate categories as you like
- More data: more categories
- More rate variation: more categories

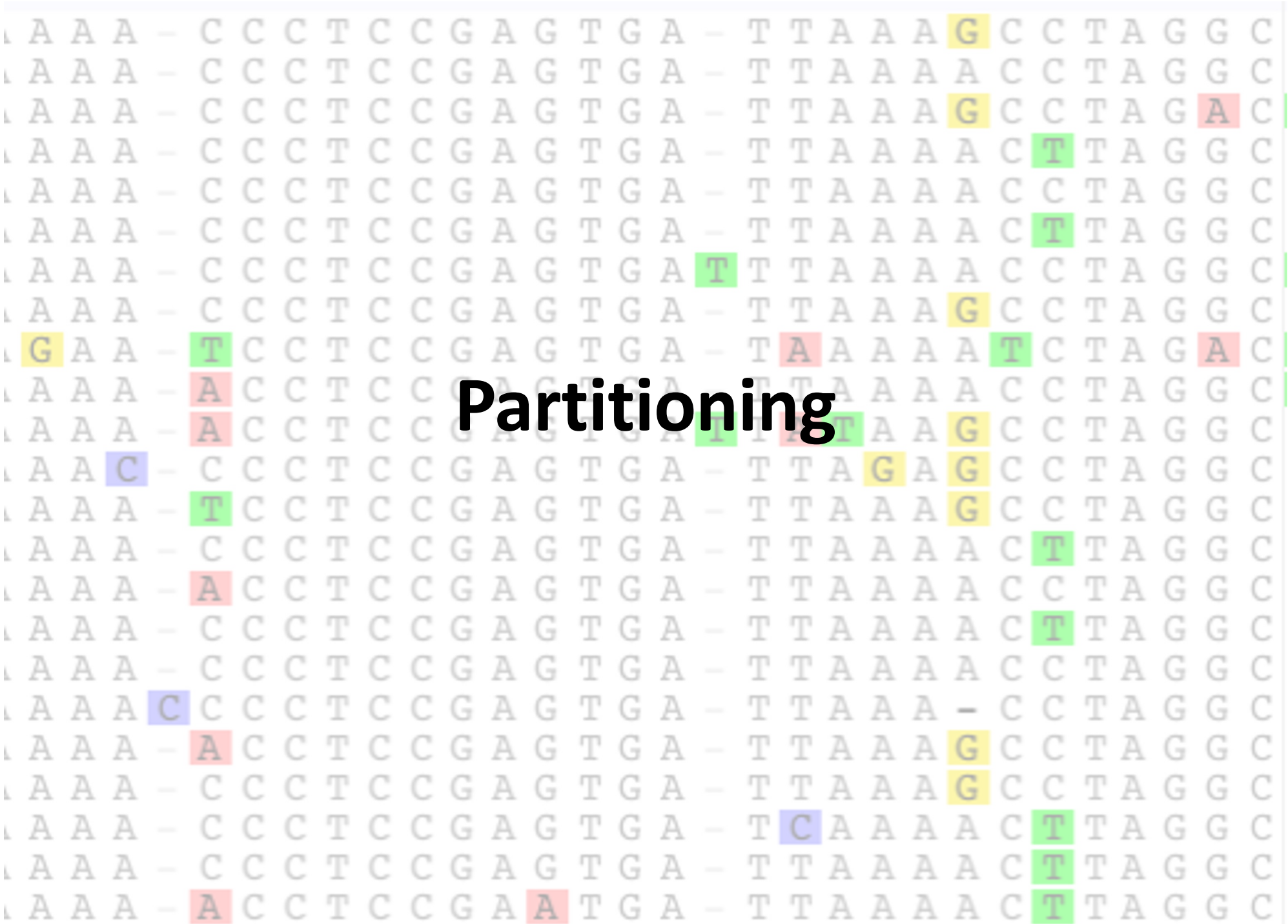
How many models are there in the GTR+I+G+R8 family?

```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

Rate Matrix	Base Frequencies	Site Rates
	$\pi_A + \pi_C + \pi_G + \pi_T = 1$	+ I + G + R8

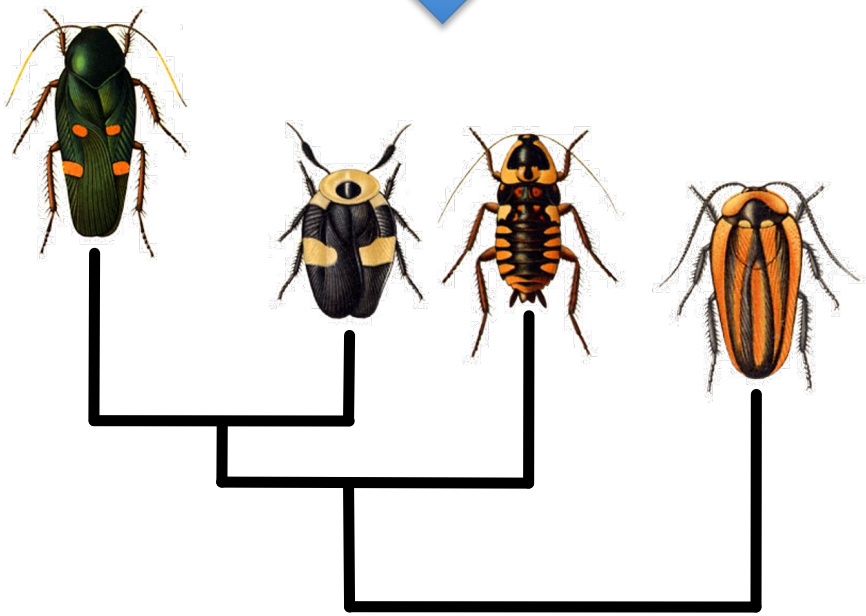
#Models **203** x **15** x **12** = **36,540**


Questions?


Partitioning

actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac

Molecular evolution



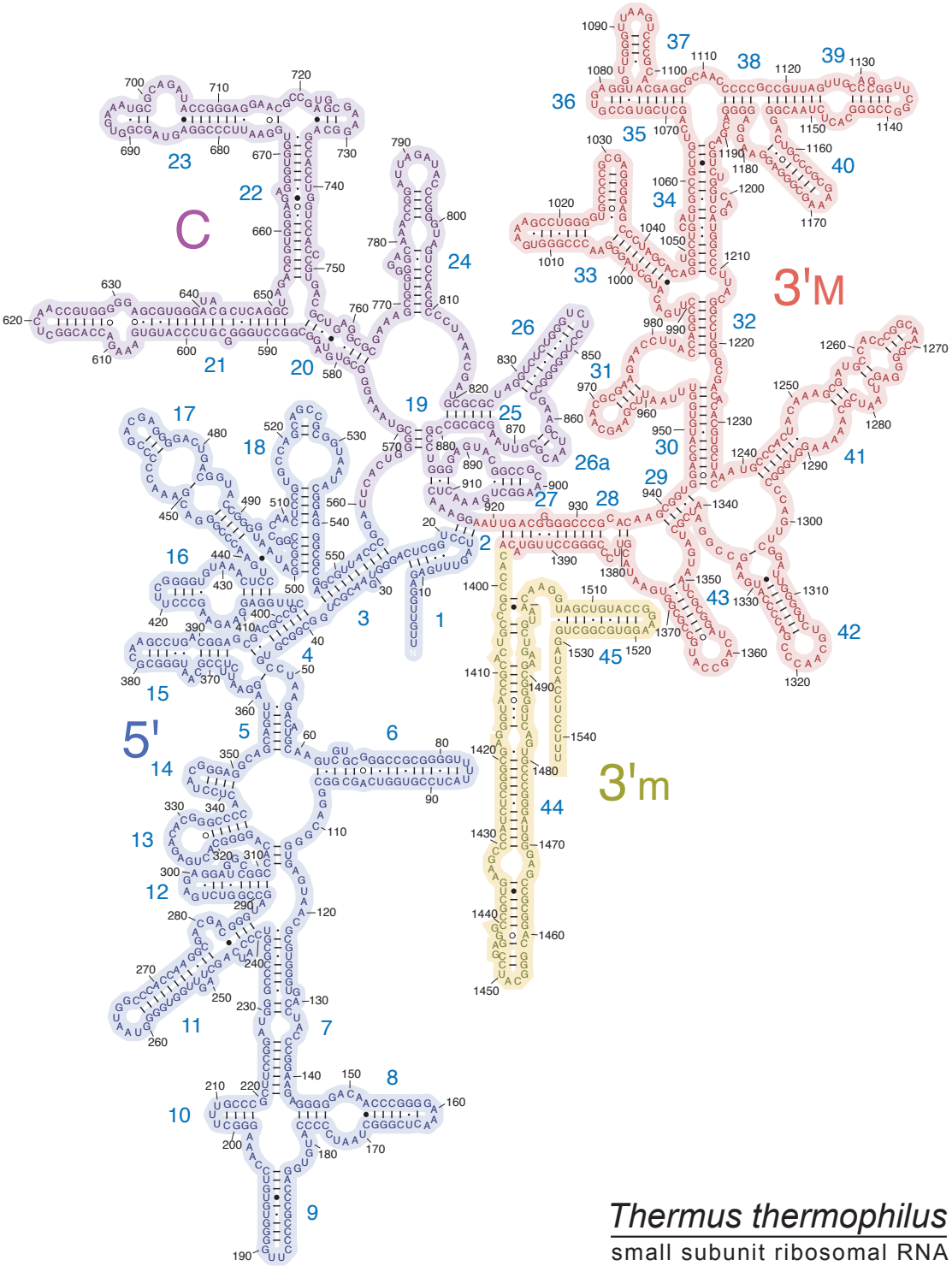
GTR+I+G

a, b, c, d, e, f

π_A, π_C, π_G, π_T

I, G

10 free parameters



Thermus thermophilus

small subunit ribosomal RNA

```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

Molecular evolution

Stems+Loops

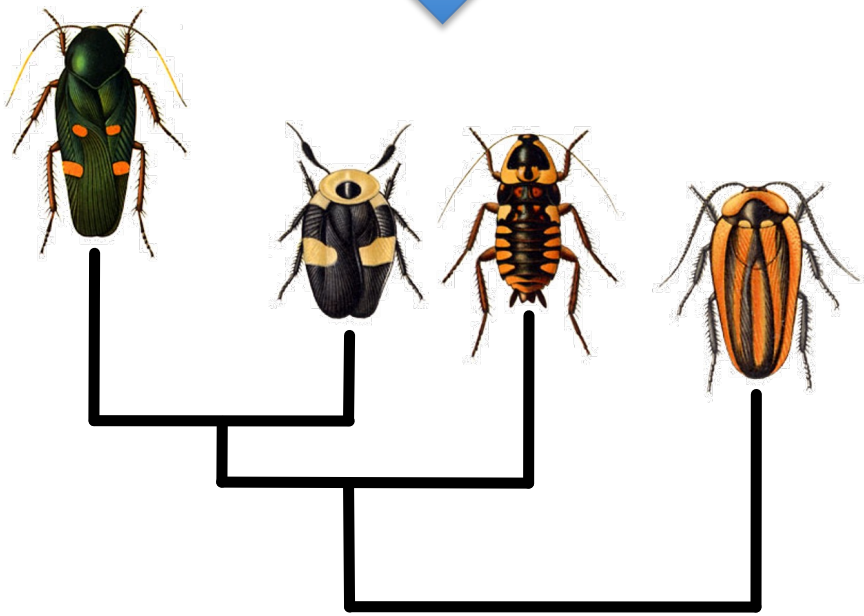
GTR+I+G

a, b, c, d, e, f

π_A, π_C, π_G, π_T

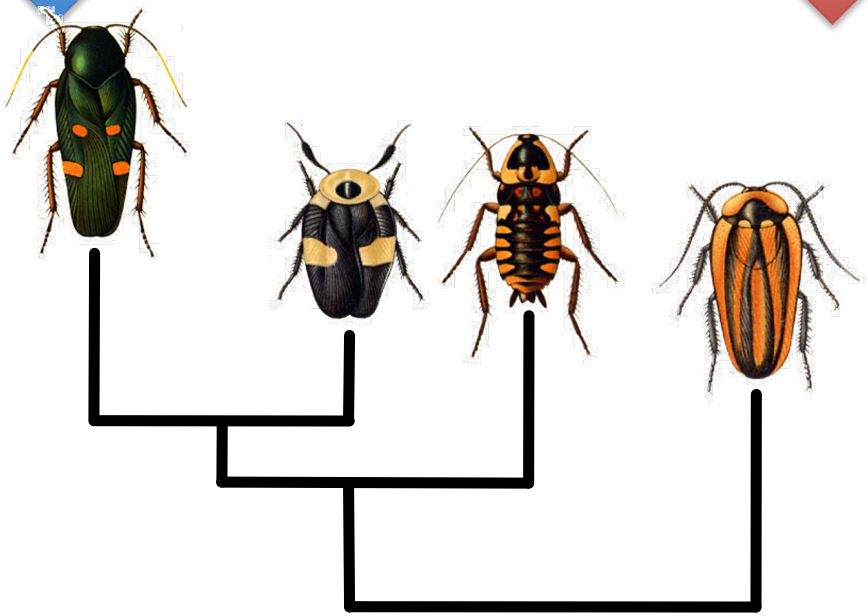
I, G

10 free parameters



actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac

Molecular evolution **Molecular evolution**



Stems

GTR+I+G

a, b, c, d, e, f

π_A, π_C, π_G, π_T

I, G

10 free parameters

Loops

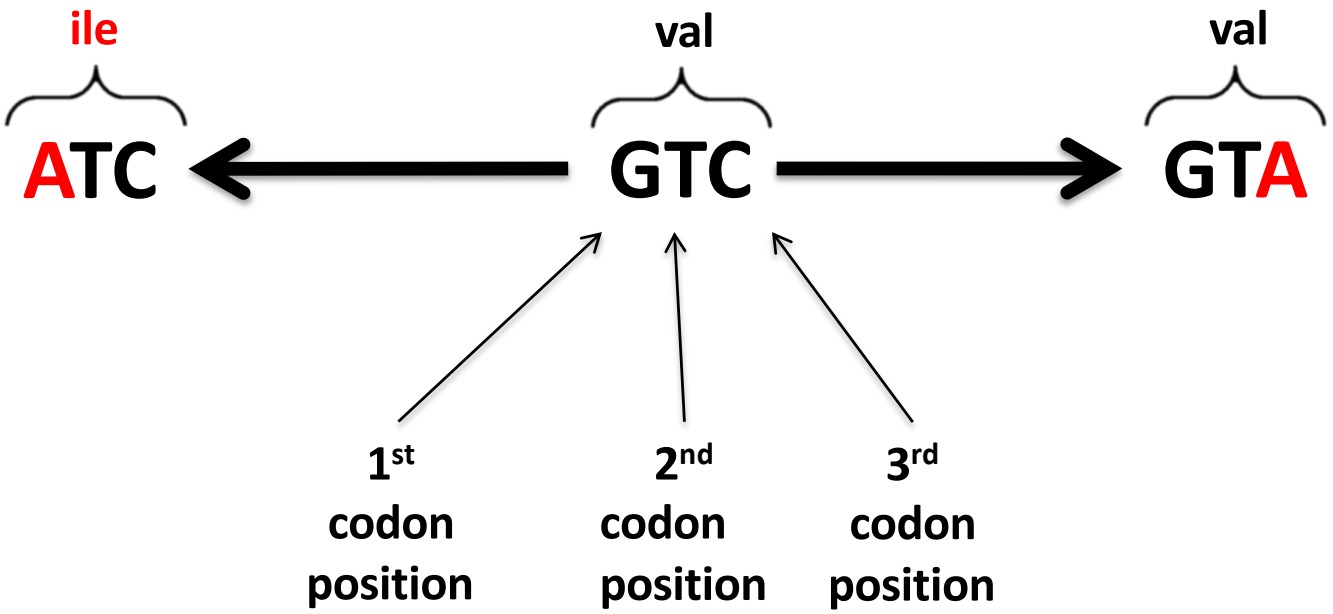
GTR+I+G

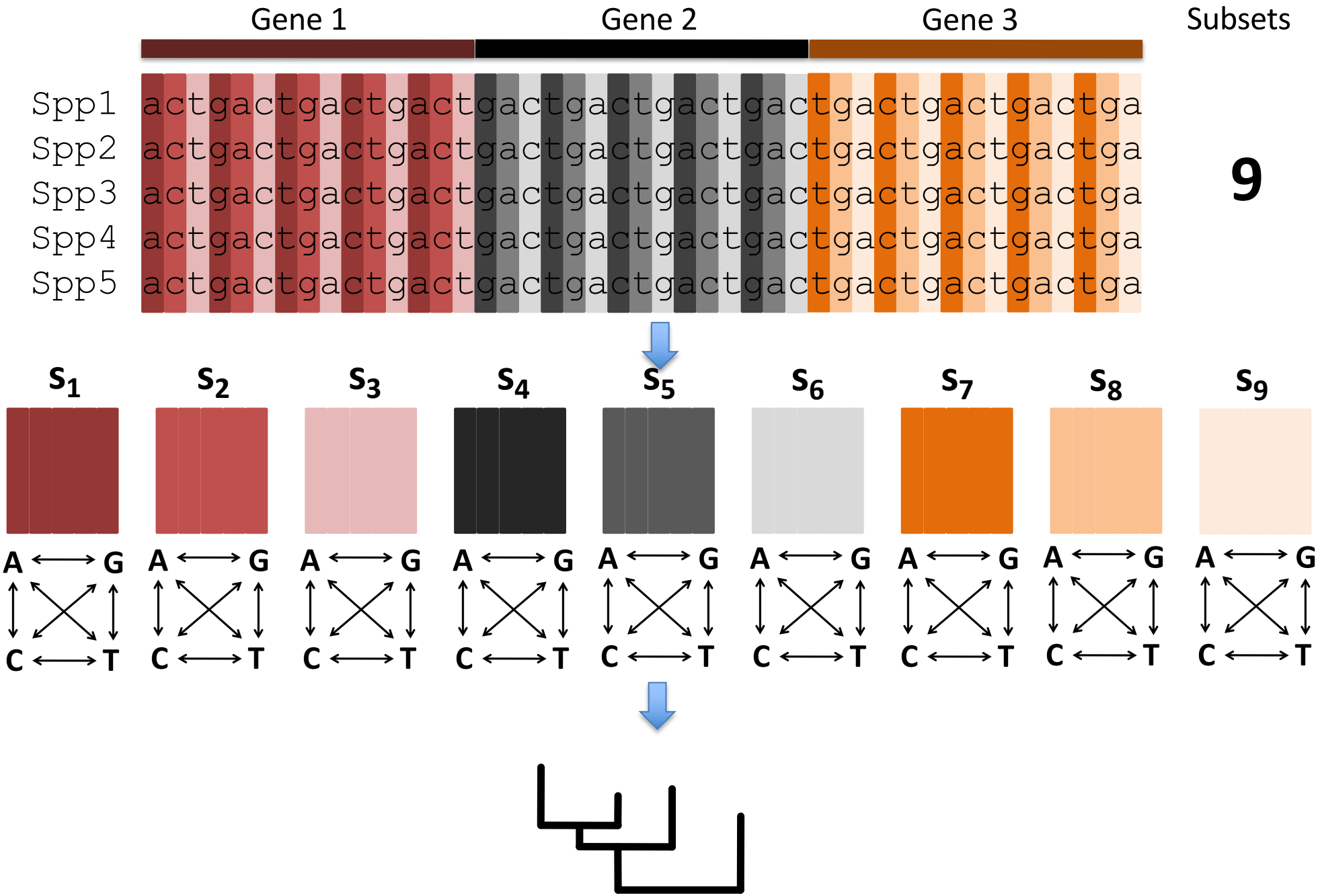
a, b, c, d, e, f

π_A, π_C, π_G, π_T

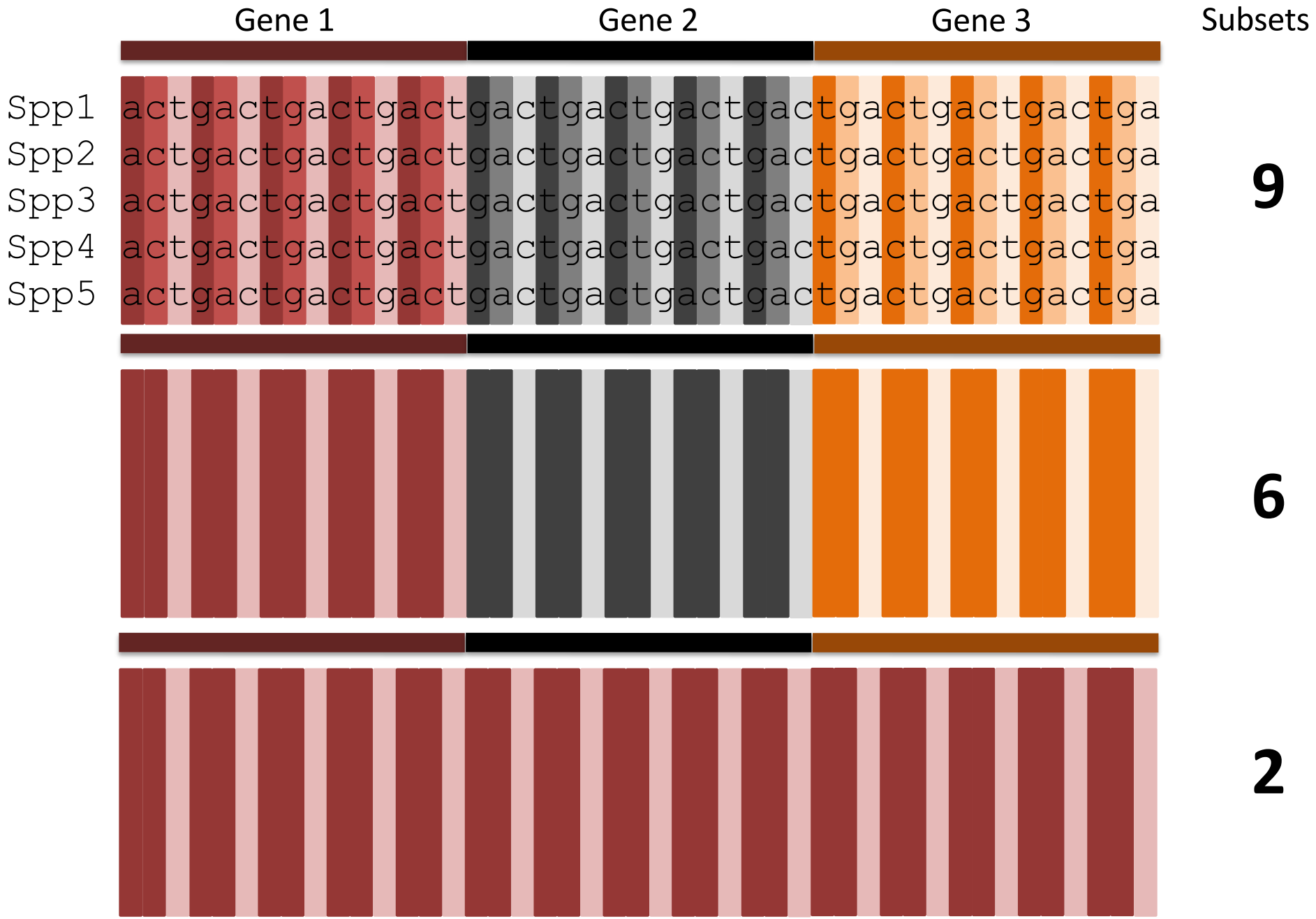
I, G

10 free parameters

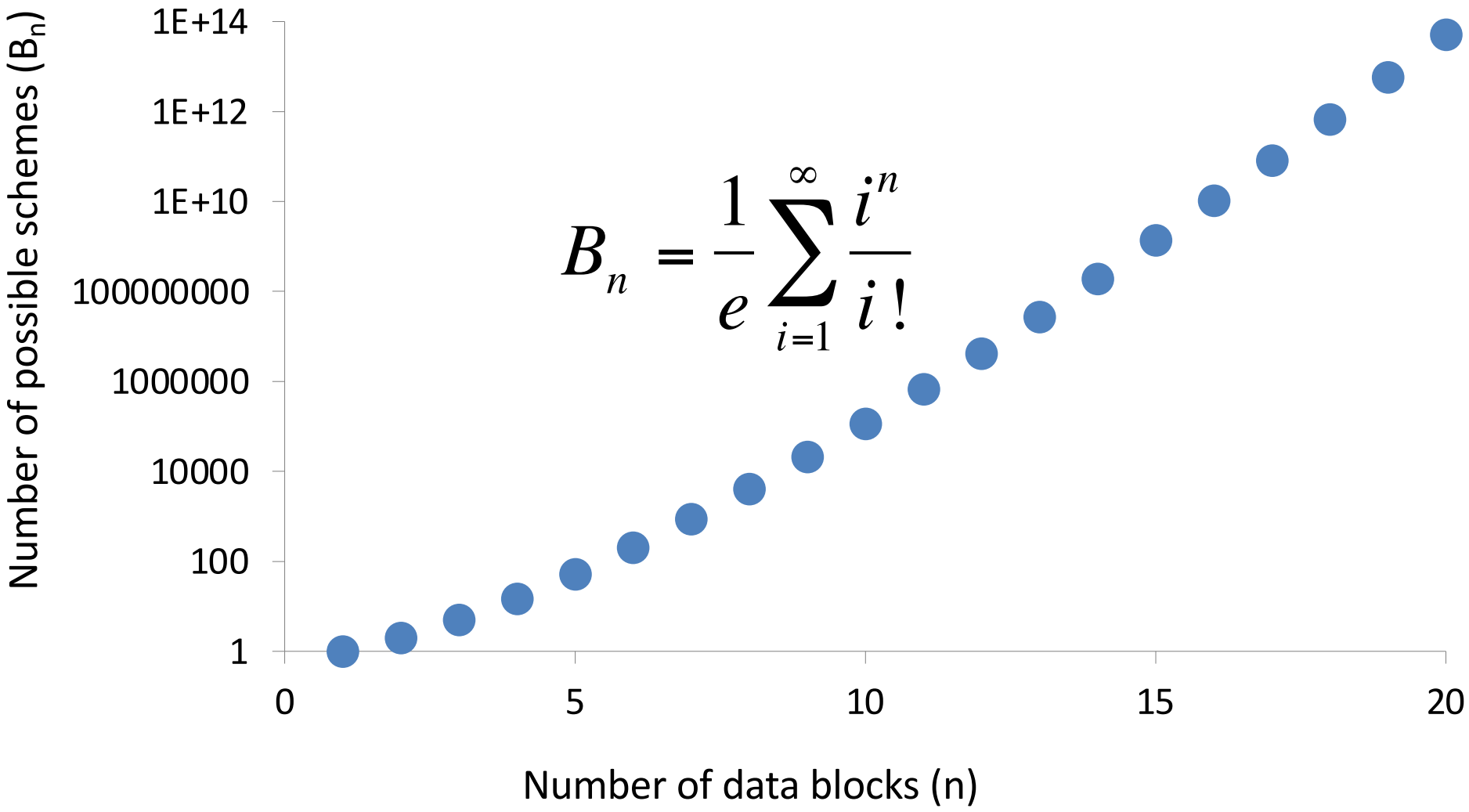

ile
val
val
ATC
GTC
GTA
1st codon position
2nd codon position
3rd codon position


Gene 1
Gene 2
Gene 3
Subsets
Spp1 actgactgactgactgactgactgactgactgactgactgactgactga
Spp2 actgactgactgactgactgactgactgactgactgactgactgactga
Spp3 actgactgactgactgactgactgactgactgactgactgactgactga
Spp4 actgactgactgactgactgactgactgactgactgactgactgactga
Spp5 actgactgactgactgactgactgactgactgactgactgactgactga
9
S1
S2
S3
S4
S5
S6
S7
S8
S9
A G
C T

The Problem


Gene 1
Gene 2
Gene 3
Subsets
Spp1 actgactgactgactgactgactgactgactgactgactgactgactga
Spp2 actgactgactgactgactgactgactgactgactgactgactgactga
Spp3 actgactgactgactgactgactgactgactgactgactgactgactga
Spp4 actgactgactgactgactgactgactgactgactgactgactgactga
Spp5 actgactgactgactgactgactgactgactgactgactgactgactga
9
6
2

How many schemes are there?



The central problem

- For any reasonably-sized dataset, there are more possible partitioning schemes than atoms in the known universe
- For each data block in each partitioning scheme, there are more than 36,000 models to choose from
- Our job is to choose a good model from this large collection of possibilities…

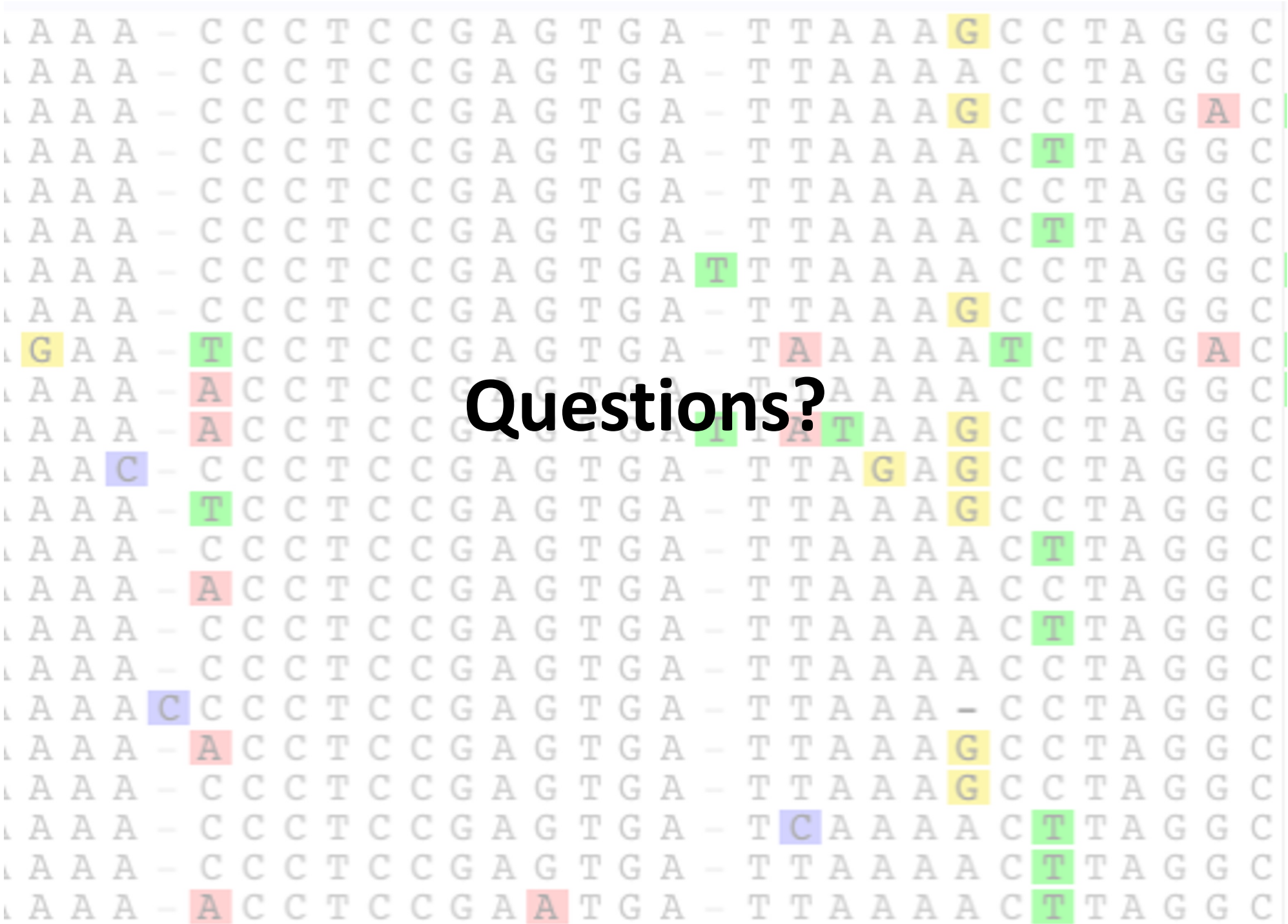
Two solutions

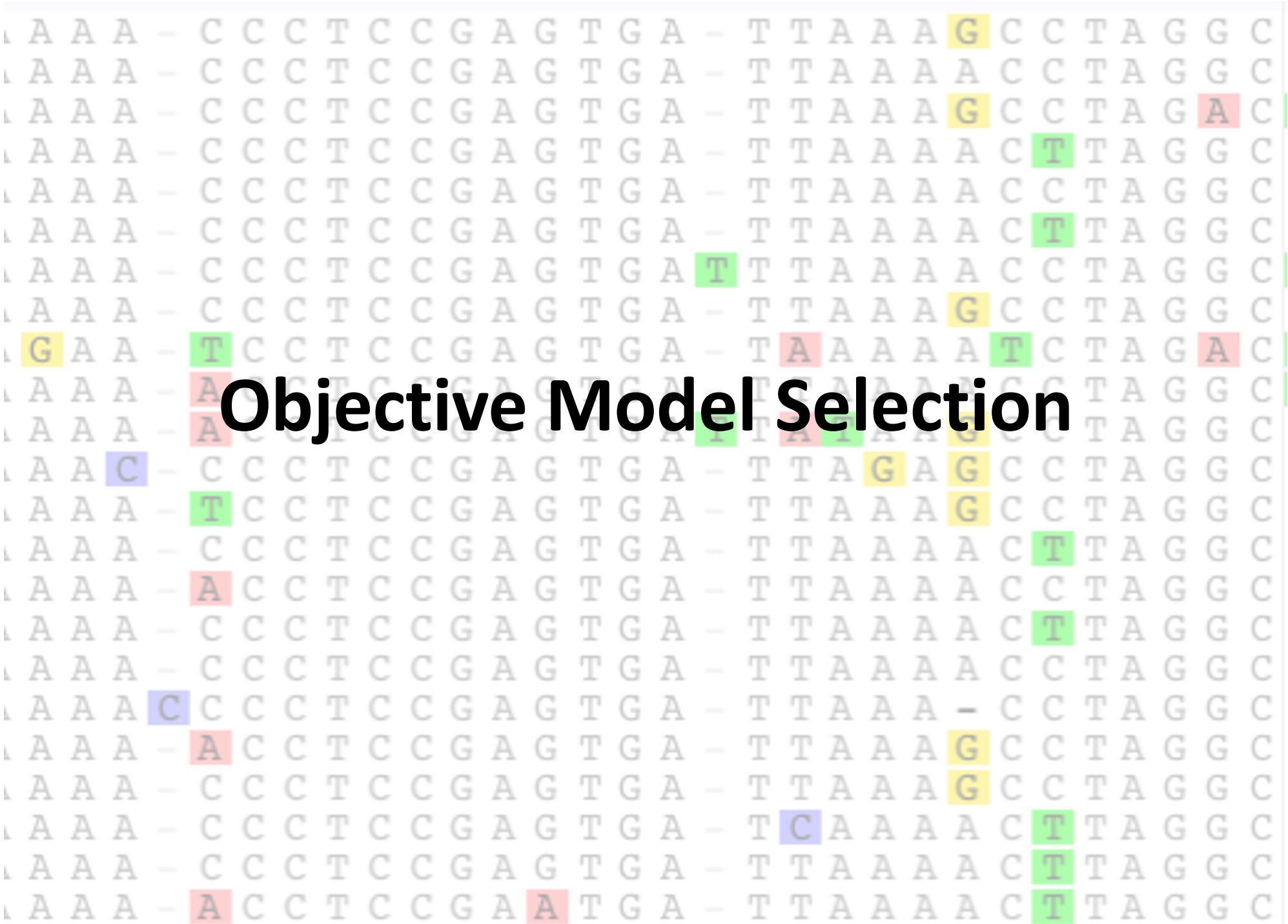
Subjective model selection

- pick a model that seems sensible, and balances the number of partitions, the amount of information in each, and the number of parameters in each of your models

Objective model selection

- Use information theory and let a computer do it for you


Questions?



Objective Model Selection

Model Selection Criteria

AIC

- Akaike's Information Criterion. Tries to find the model which minimises the loss of information when modelling the truth.

AICc

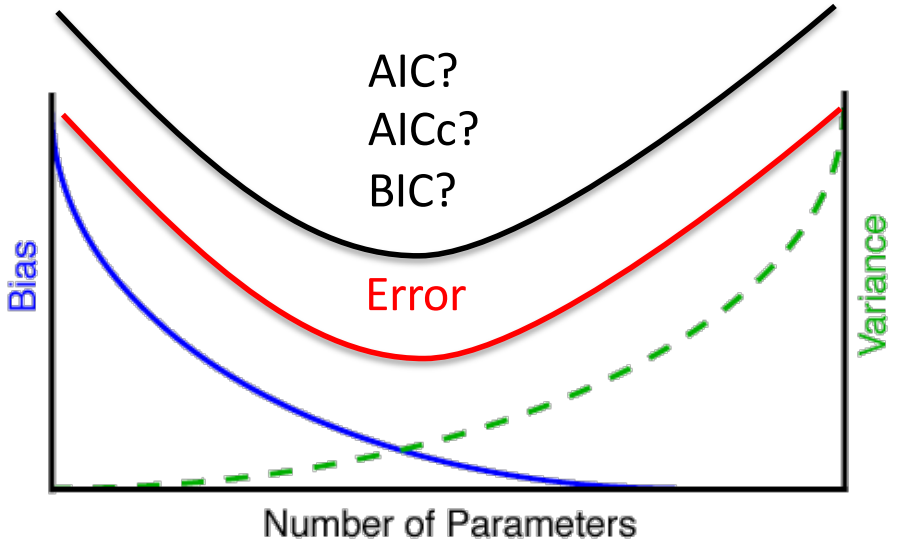
- The AIC, corrected for sample size. In theory, this should always be preferred to the AIC, because it's asymptotic to the AIC when you have large samples. **What is the sample size in molecular phylogenetics though?**

BIC

- Bayesian Information Criterion. Similar to AIC, but motivated from a Bayesian perspective. (Some would say it's completely made up!)
- Assumes the true model is in the set

Model Selection Criteria

- Each criterion makes subtly different assumptions about how the world works
- In practice, most perform fine in most situations*



Selecting a model from the GTR family

```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

JC …about 36,000 other models… GTR+I+G; +R

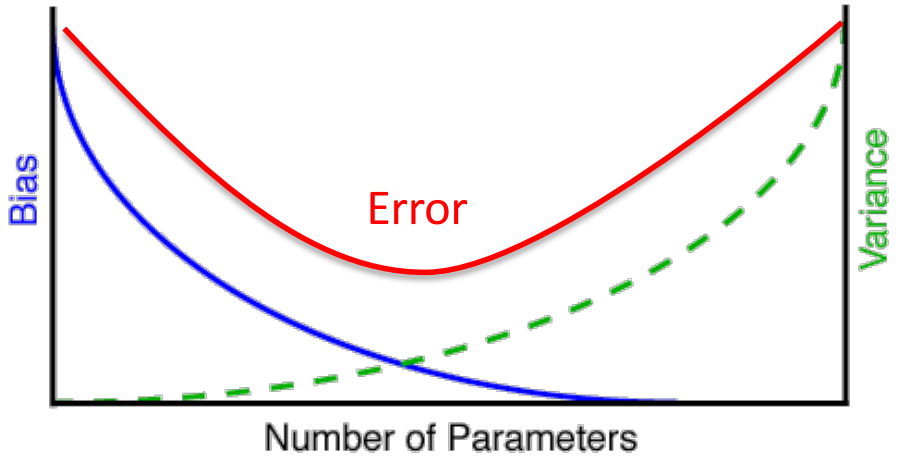
Optimise model parameters + brlens with ML

Calculate Score, e.g. $AIC = 2(k - \ln L)$

Choose ‘best’ model (lowest AIC score)

A new problem!

- We select a model from a candidate set
- So the candidate set limits how good our answers can be

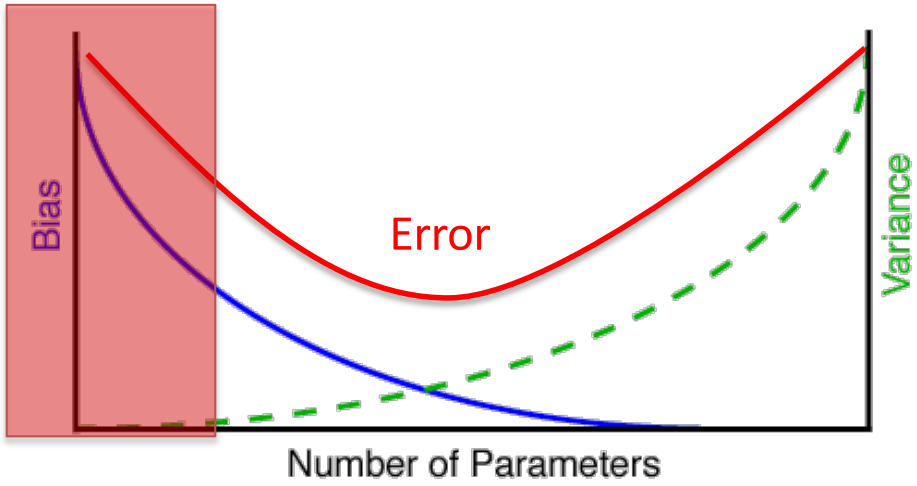


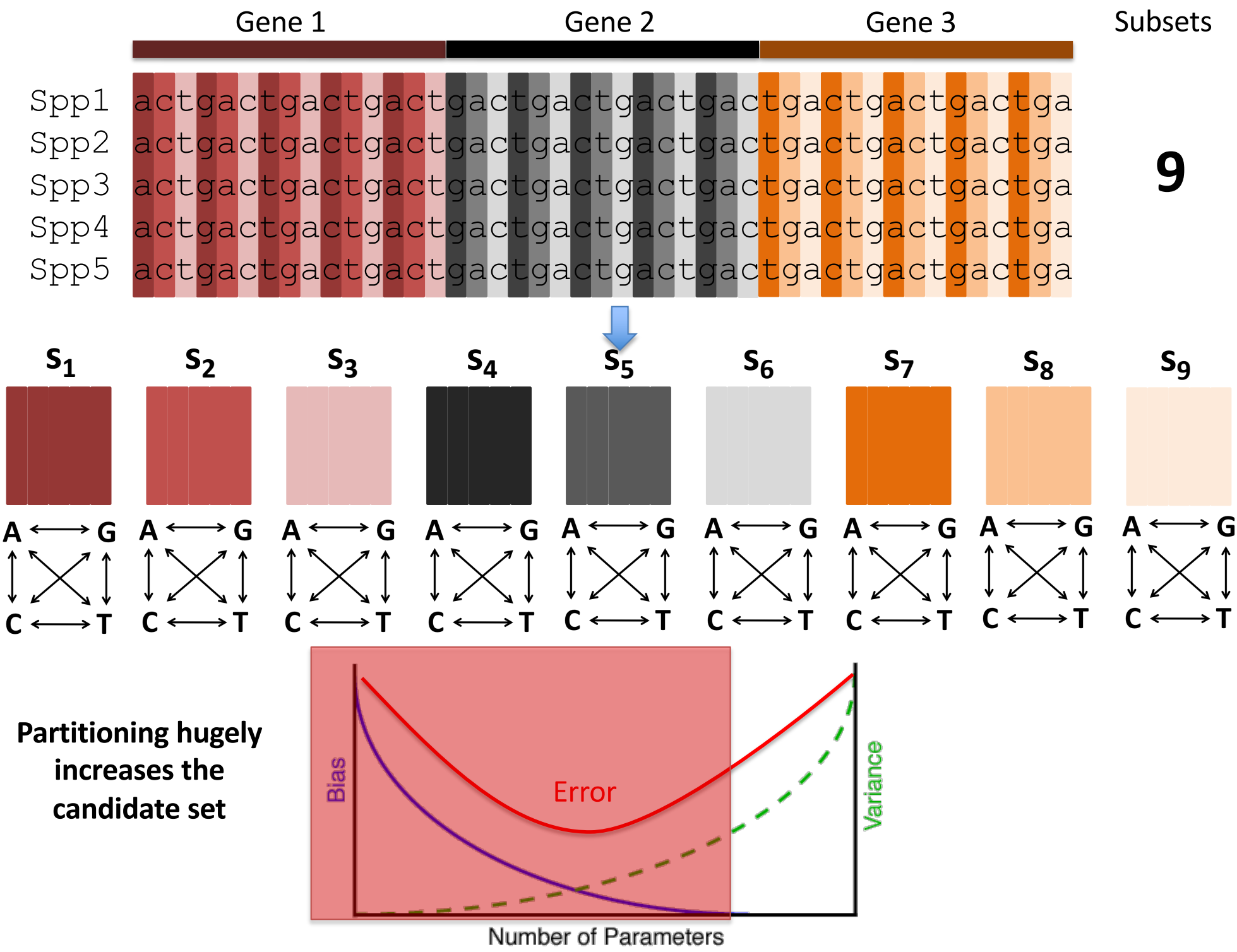
The Problem

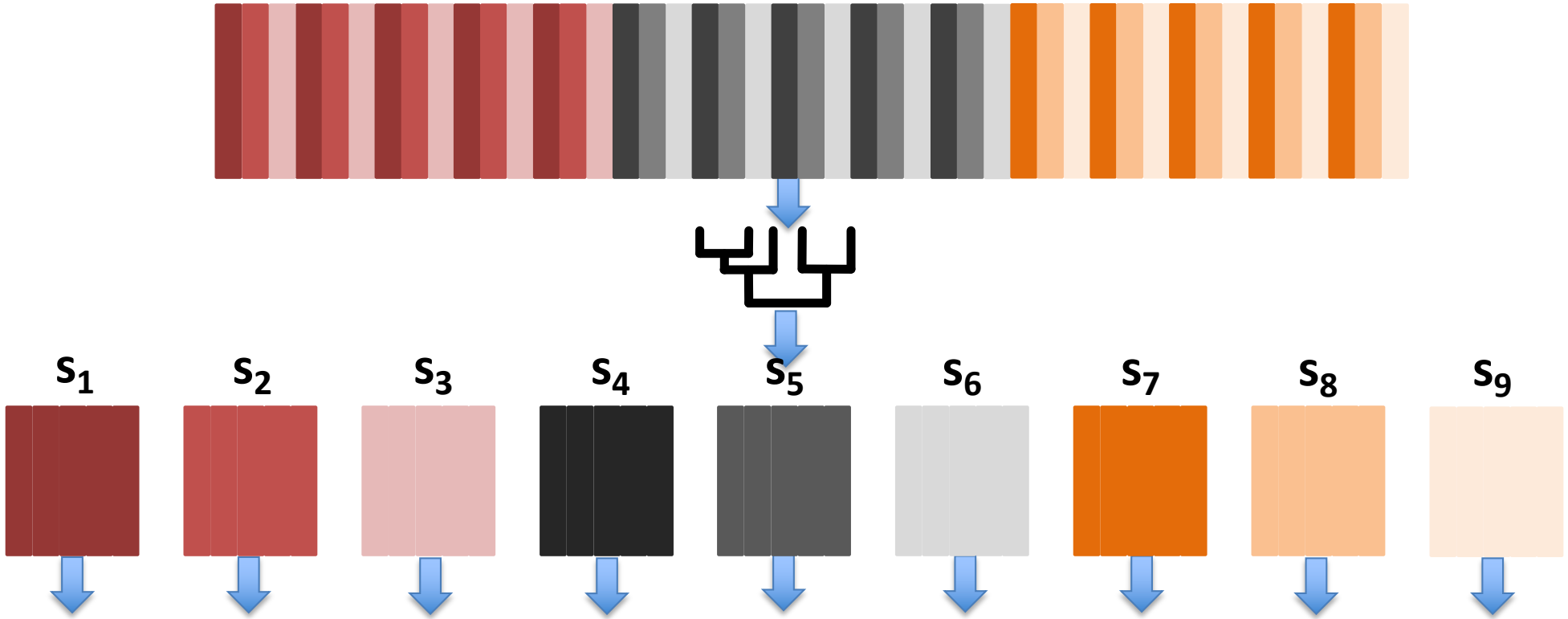
```
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
actgactgactgactgactgactgactgactgactgactgactgac
```

Almost always select **GTR+I+G**
(the most complex model)

Our candidate set is too small!



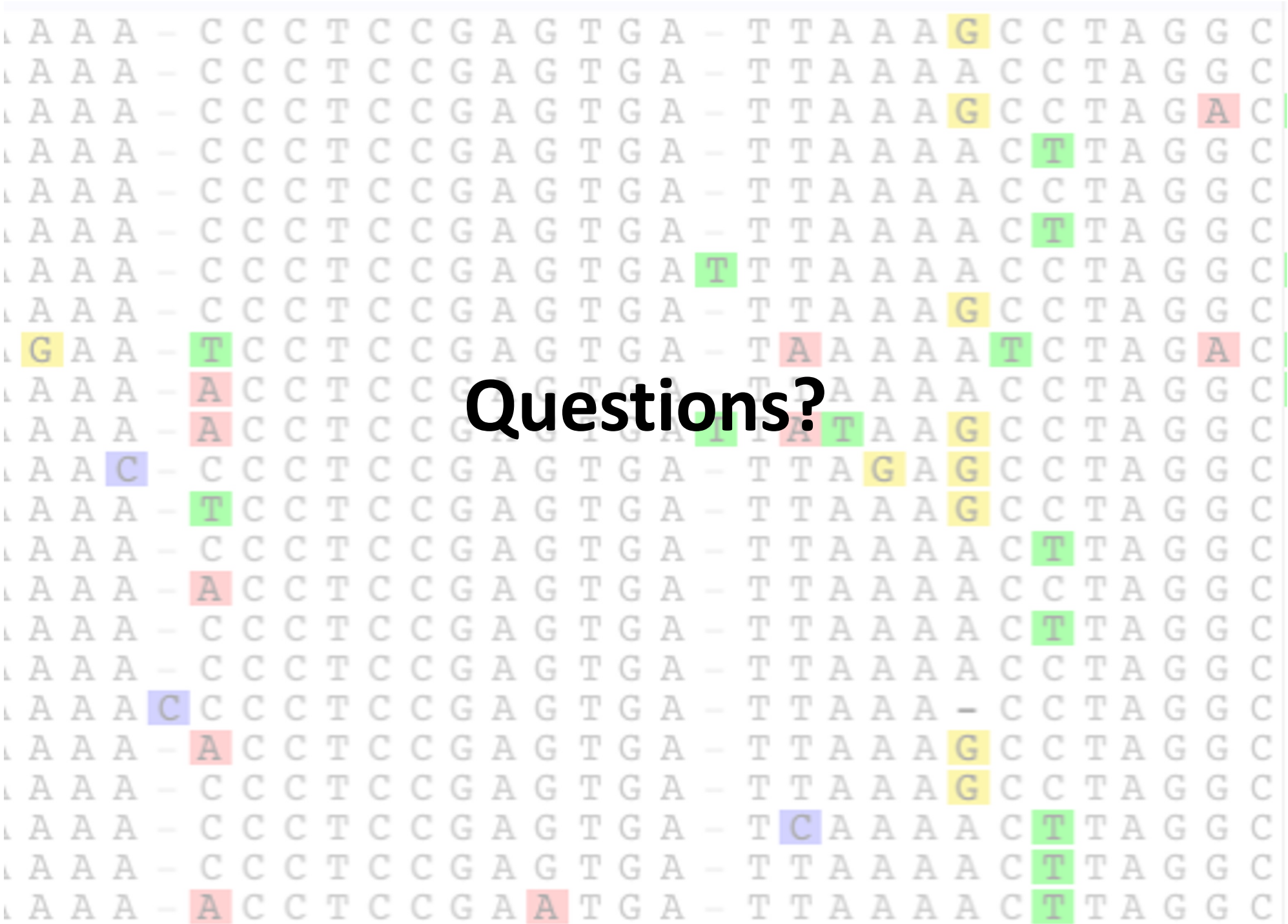

Gene 1
Gene 2
Gene 3
Subsets
Spp1 actgactgactgactgactgactgactgactgactgactgactgactga
Spp2 actgactgactgactgactgactgactgactgactgactgactgactga
Spp3 actgactgactgactgactgactgactgactgactgactgactgactga
Spp4 actgactgactgactgactgactgactgactgactgactgactgactga
Spp5 actgactgactgactgactgactgactgactgactgactgactgactga
9
S1 S2 S3 S4 S5 S6 S7 S8 S9
A G
C T
Partitioning hugely increases the candidate set
Bias
Error
Variance
Number of Parameters

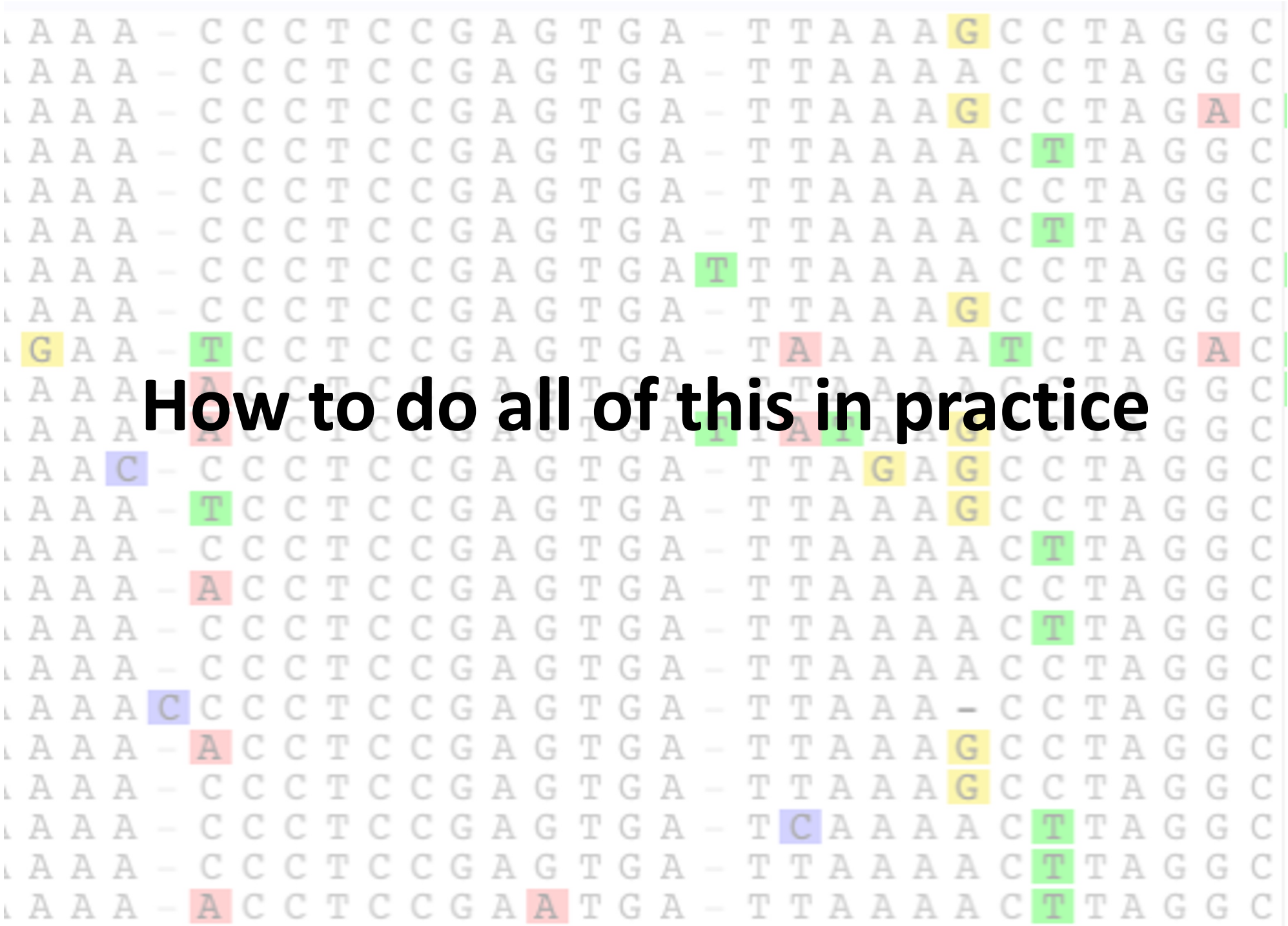


Estimate best model for each subset, conditioned on tree and relative branchlengths
Calculate AIC/AICc/BIC

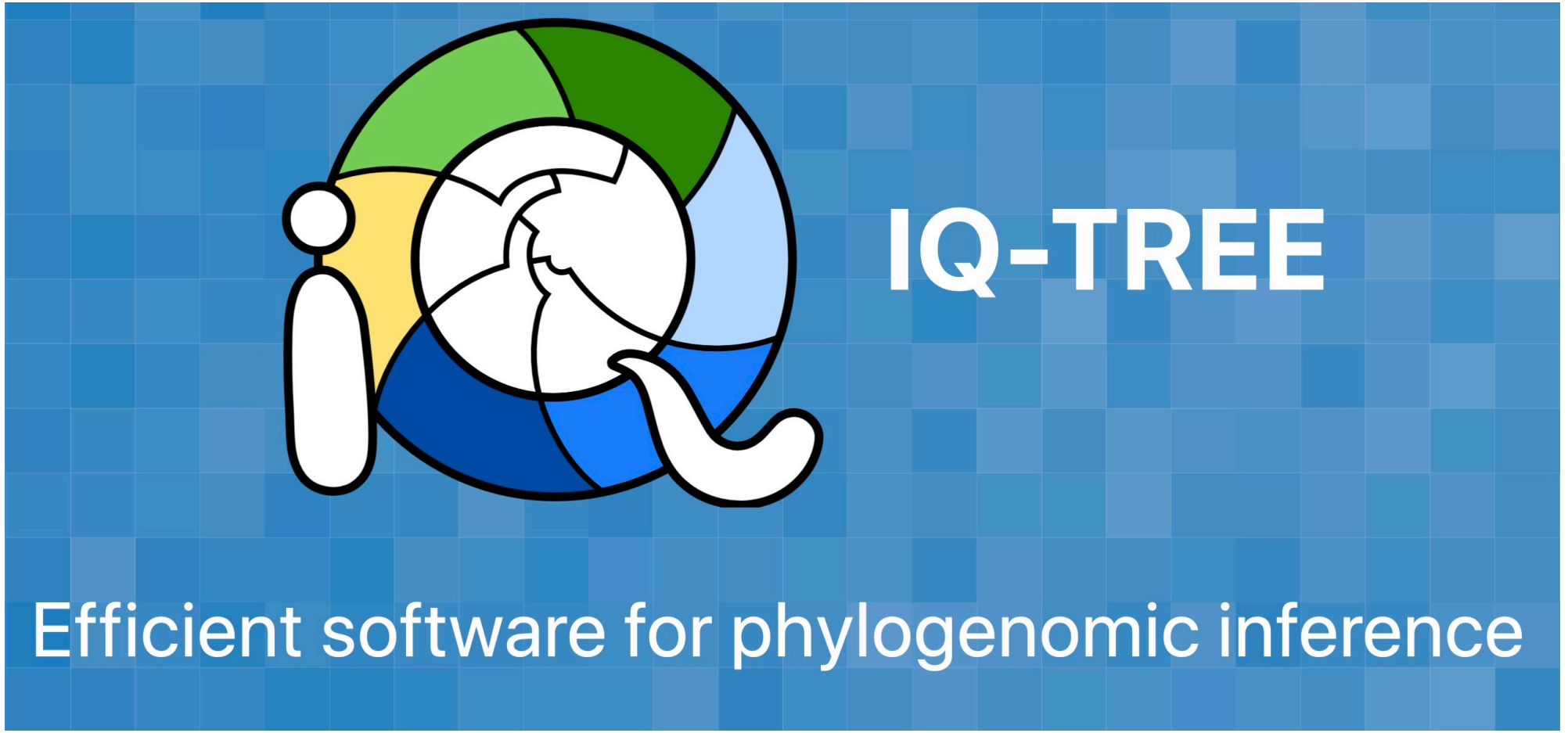
Heuristic search for Partitioning Schemes with better AIC/AICc/BIC
This works by trying out different ways to merge subsets together and continuing until you can't find any more subsets to merge that improve the BIC

Estimate of optimal partitoining scheme, with models of molecular evolution for each partition


Questions?



How to do all of this in practice



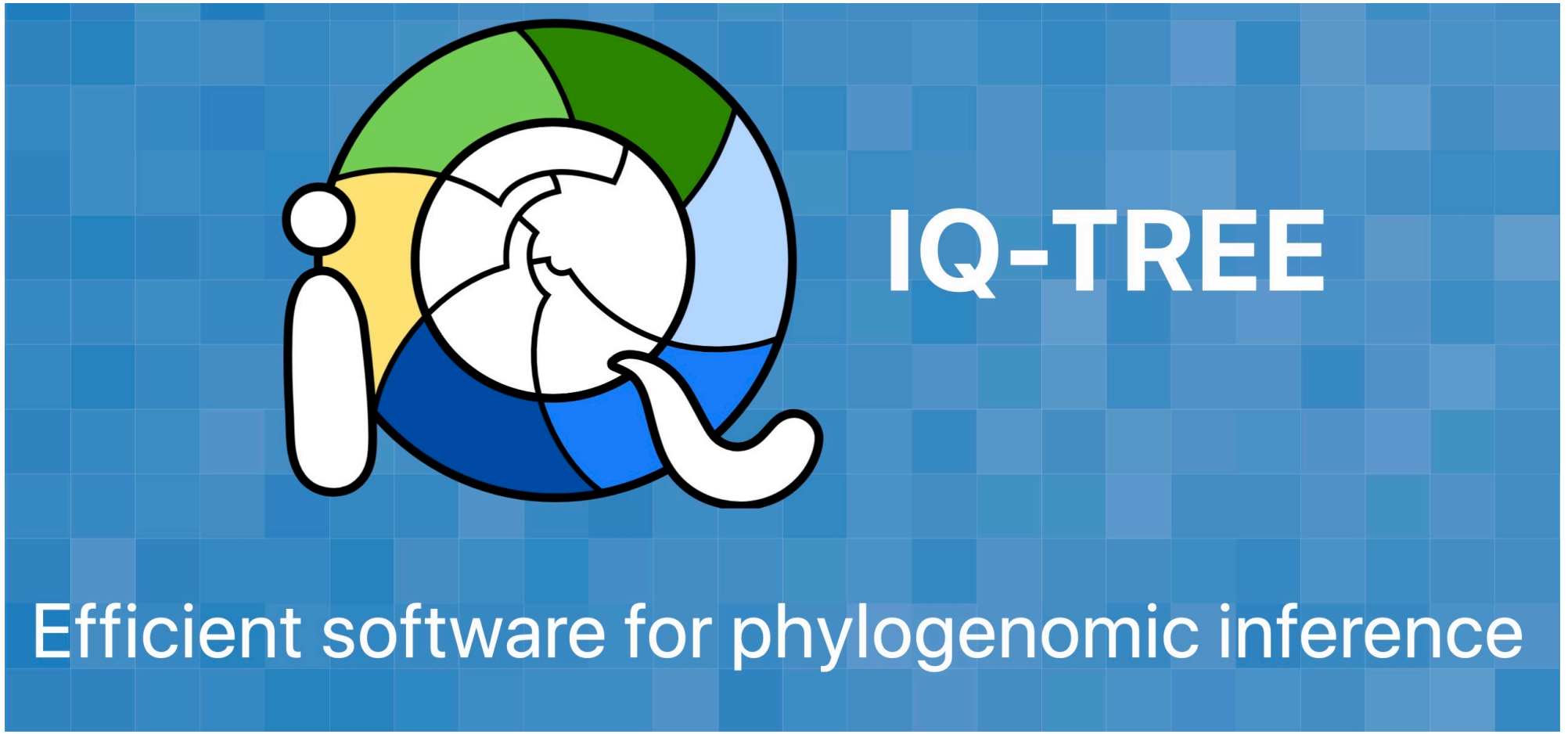
https://iqtree.github.io/doc/Advanced-Tutorial

```
iqtree -s example.phy -p example.nex -m MFP+MERGE
```

your concatenated alignment

your partition file

Model**F**inder2 + **P**lus tree search + **MERGE** subsets

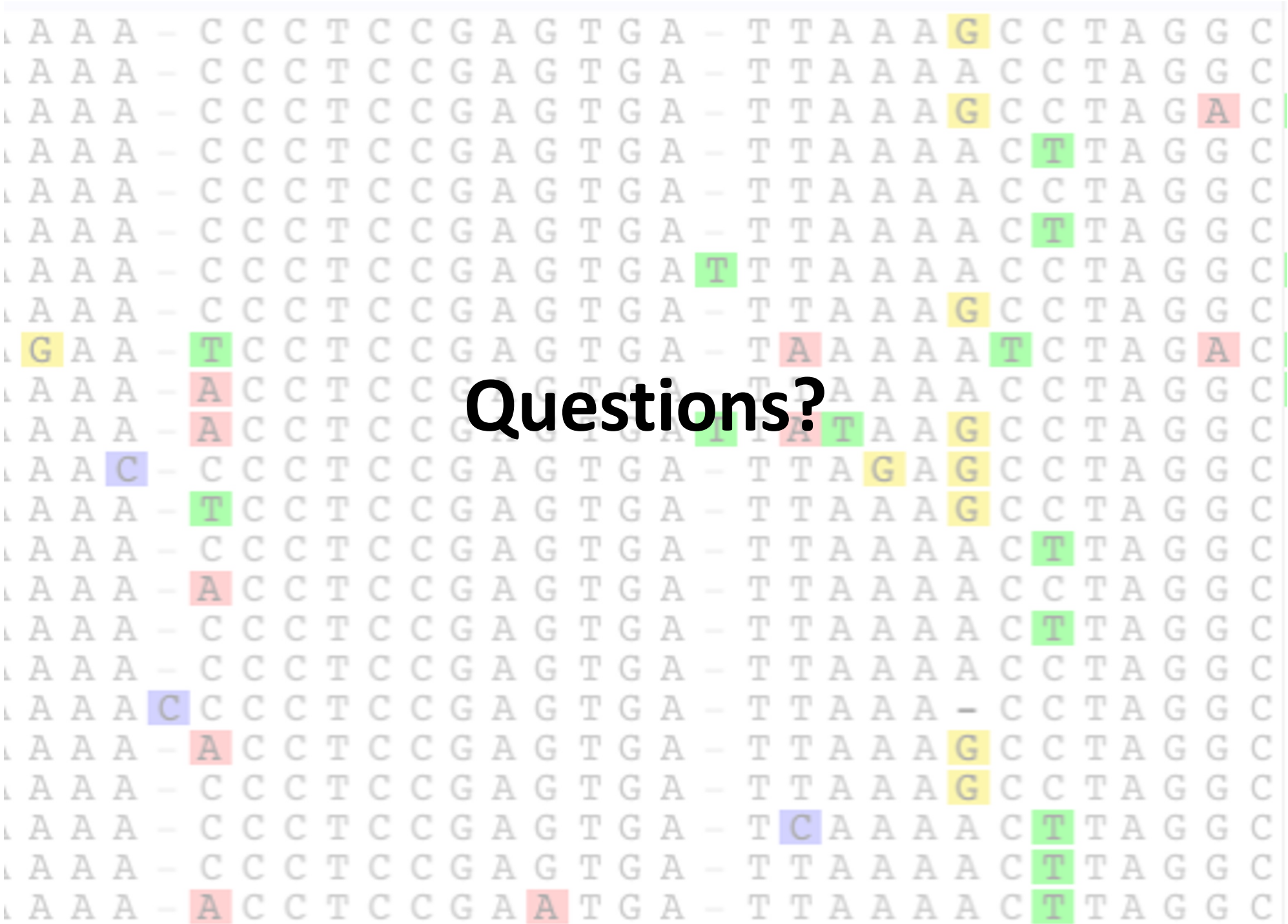


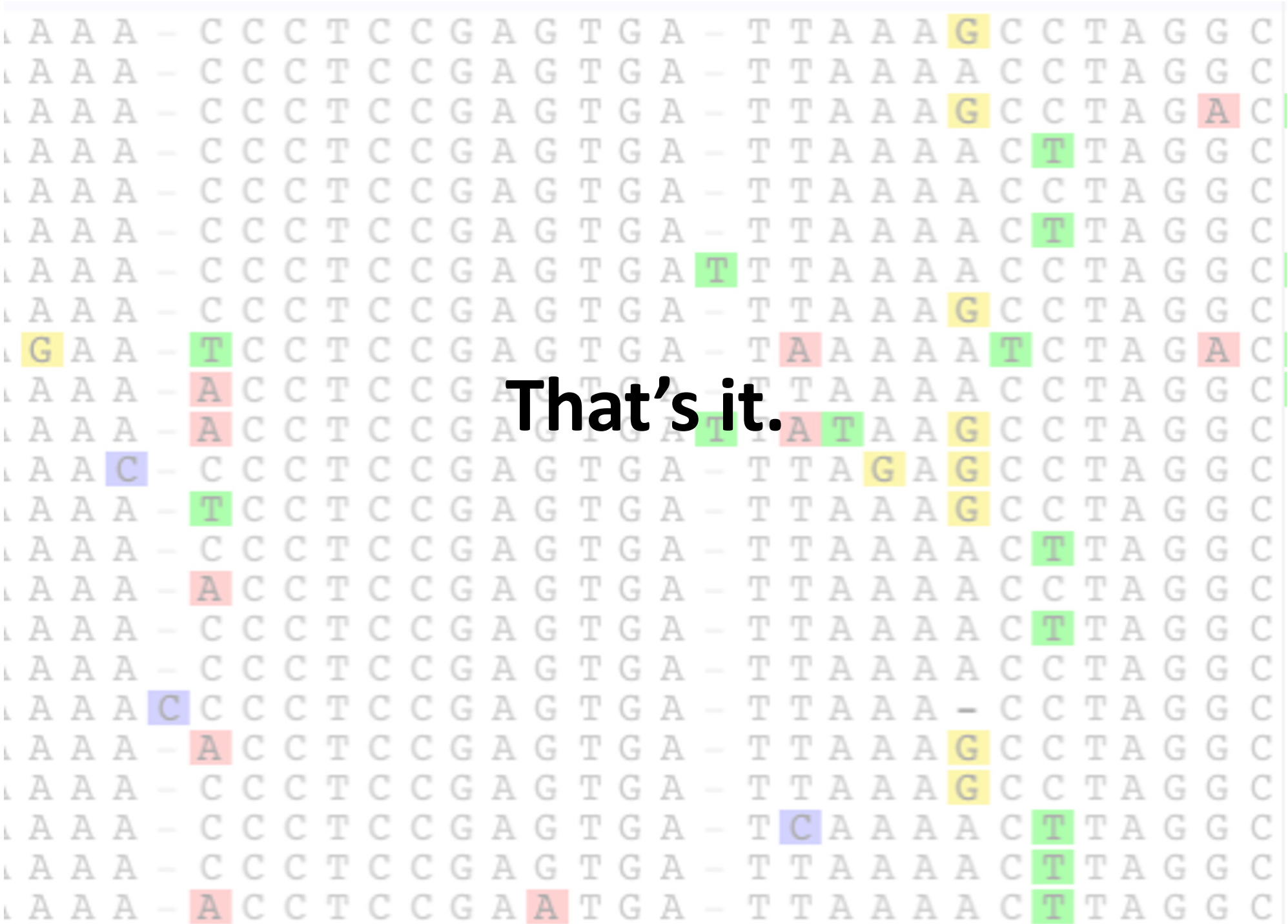
https://iqtree.github.io/doc/Advanced-Tutorial

```
iqtree -p alignments/ -m MFP+MERGE
```

your folder of alignments

Model**F**inder2 + **P**lus tree search + **MERGE** subsets


Questions?


That’s it.