

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

Rob Lanfear

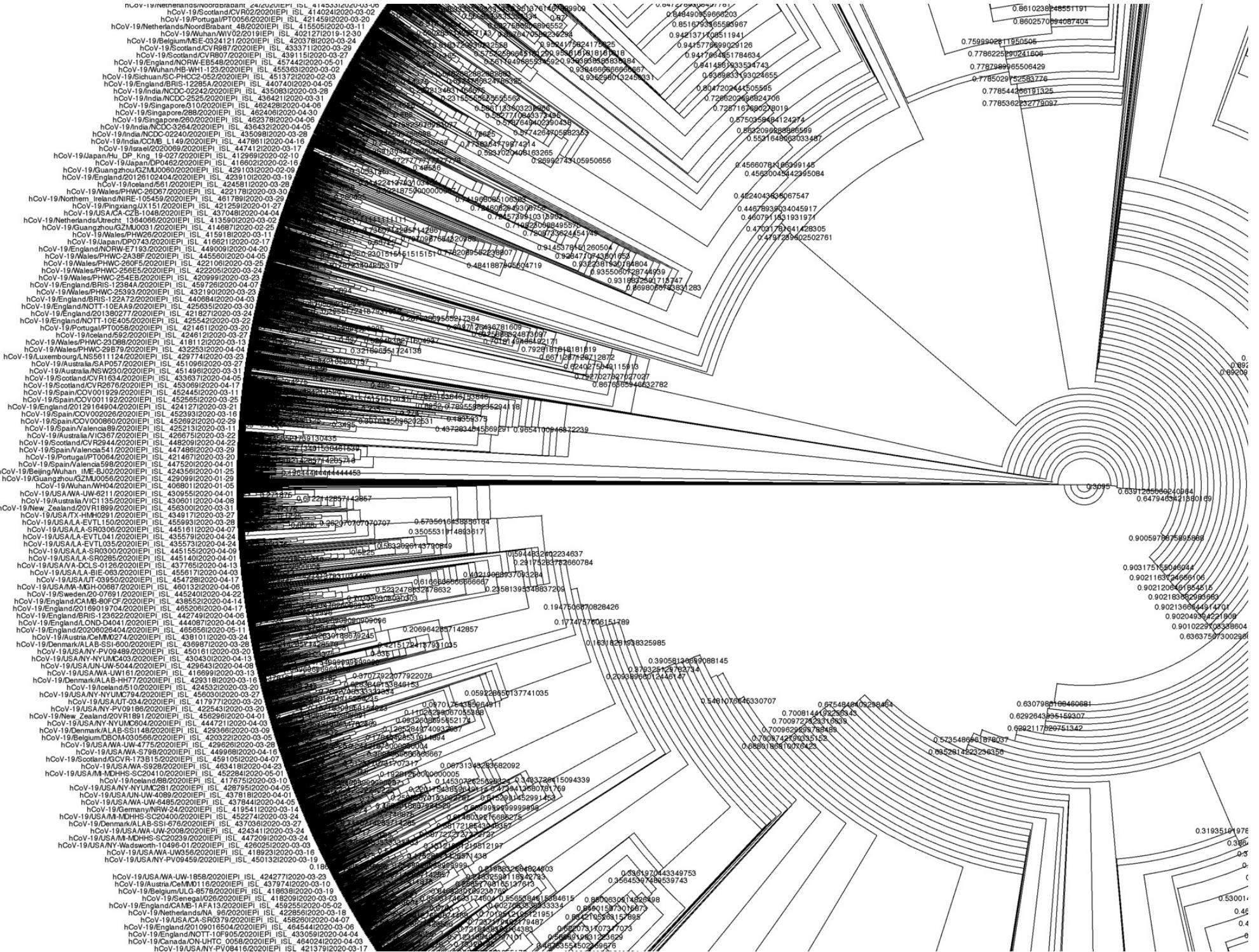
Ecology and Evolution

Australian National University

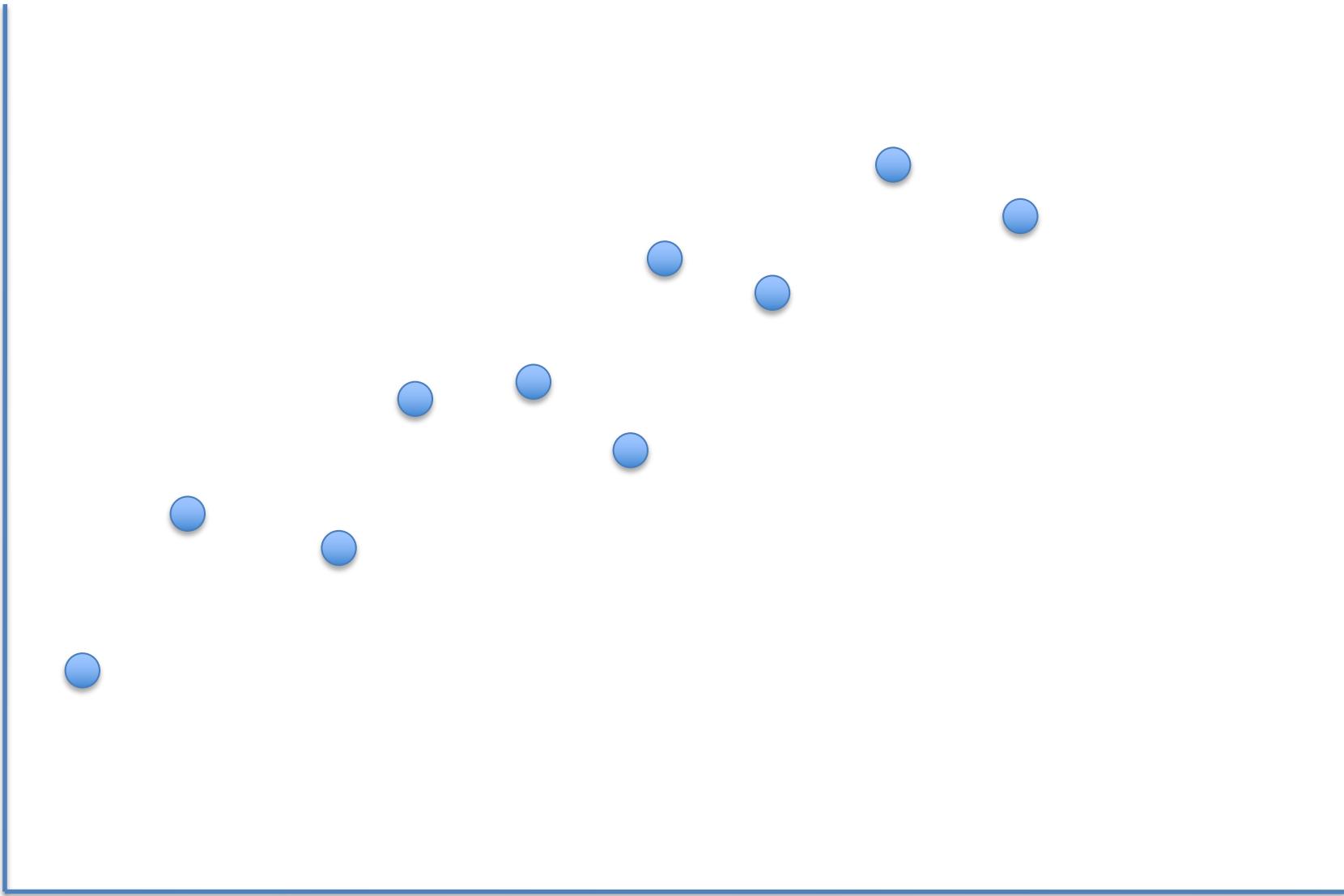
# Why Care?

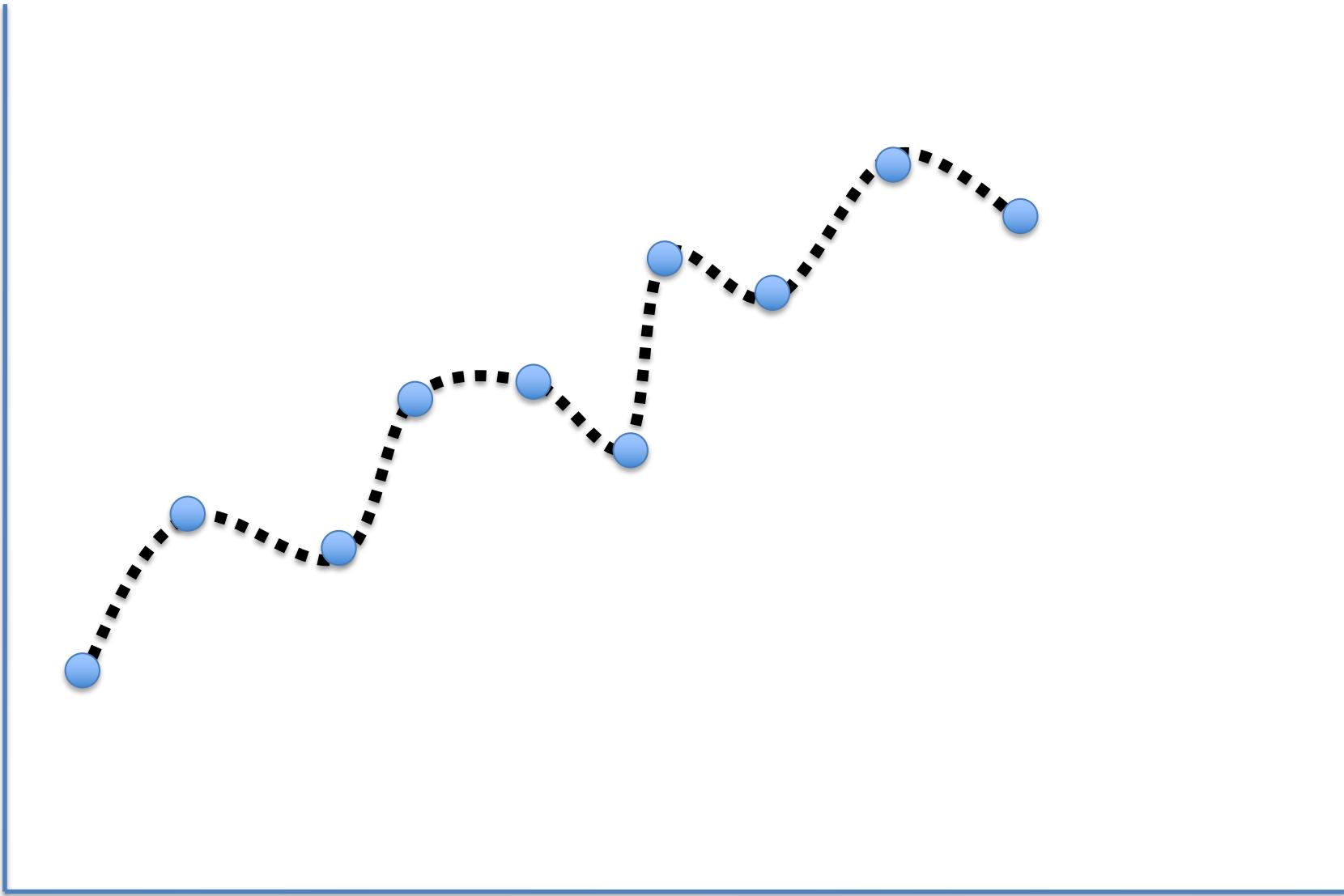
The figure displays a grid of DNA sequence data. The sequences are arranged in rows and columns. A specific sequence is highlighted in yellow and red, and other mutations are highlighted in green and blue. The text "Why Care?" is overlaid on the grid.

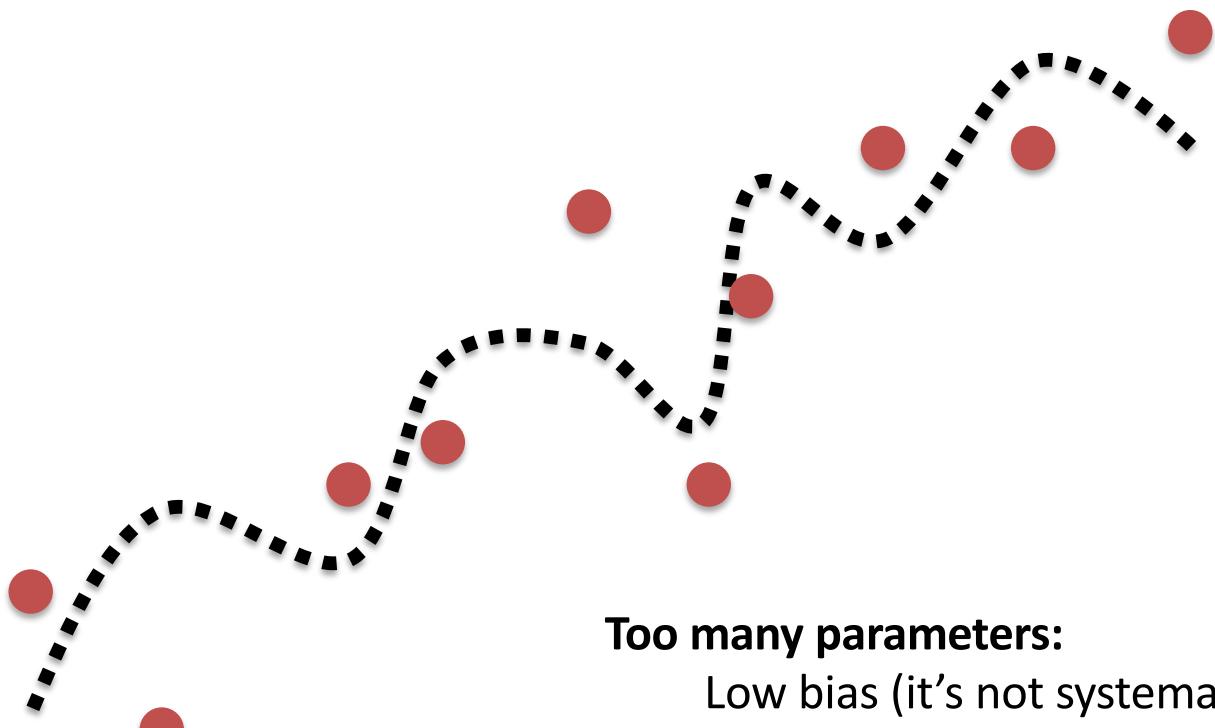
Why Care?



# Why we need model selection

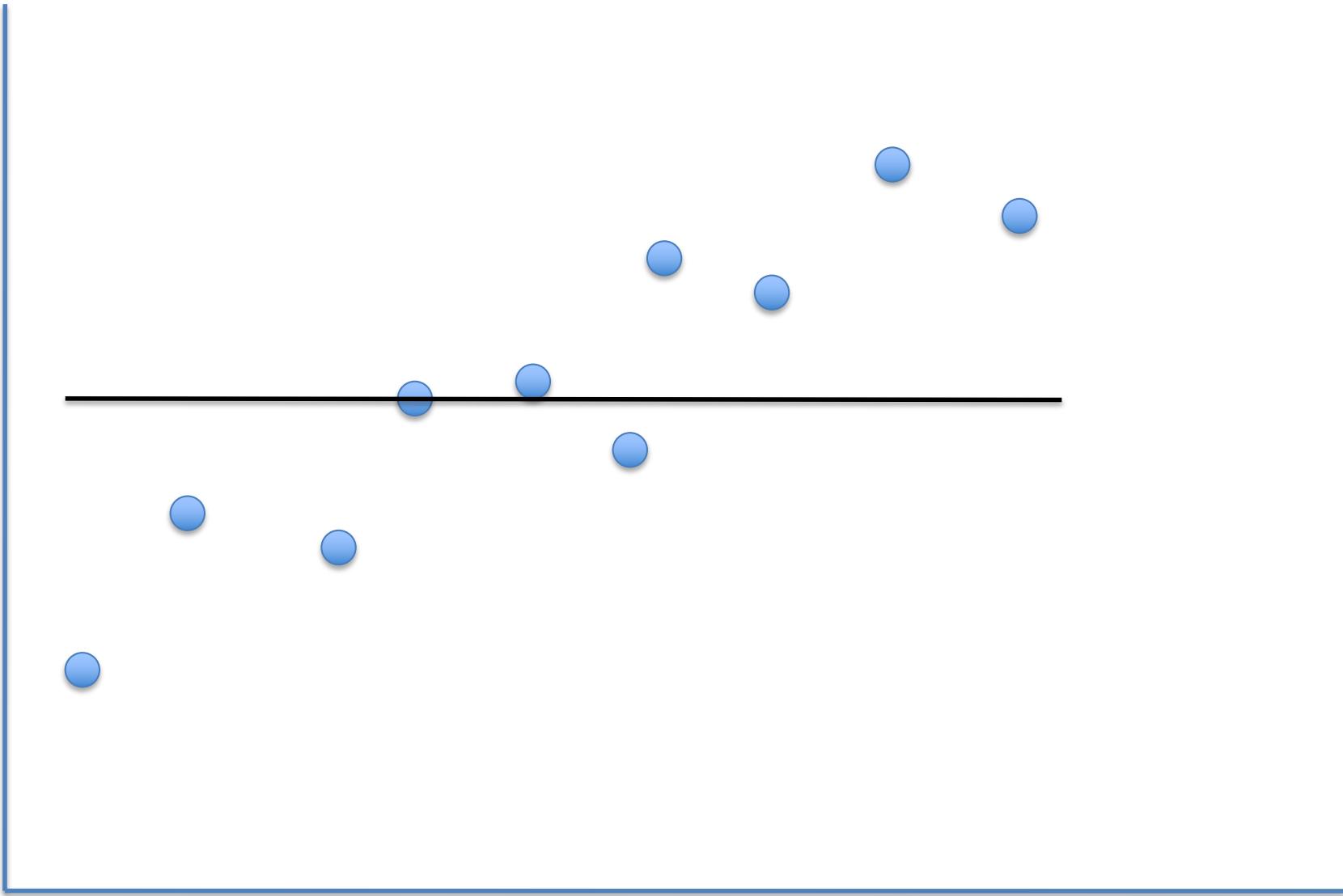


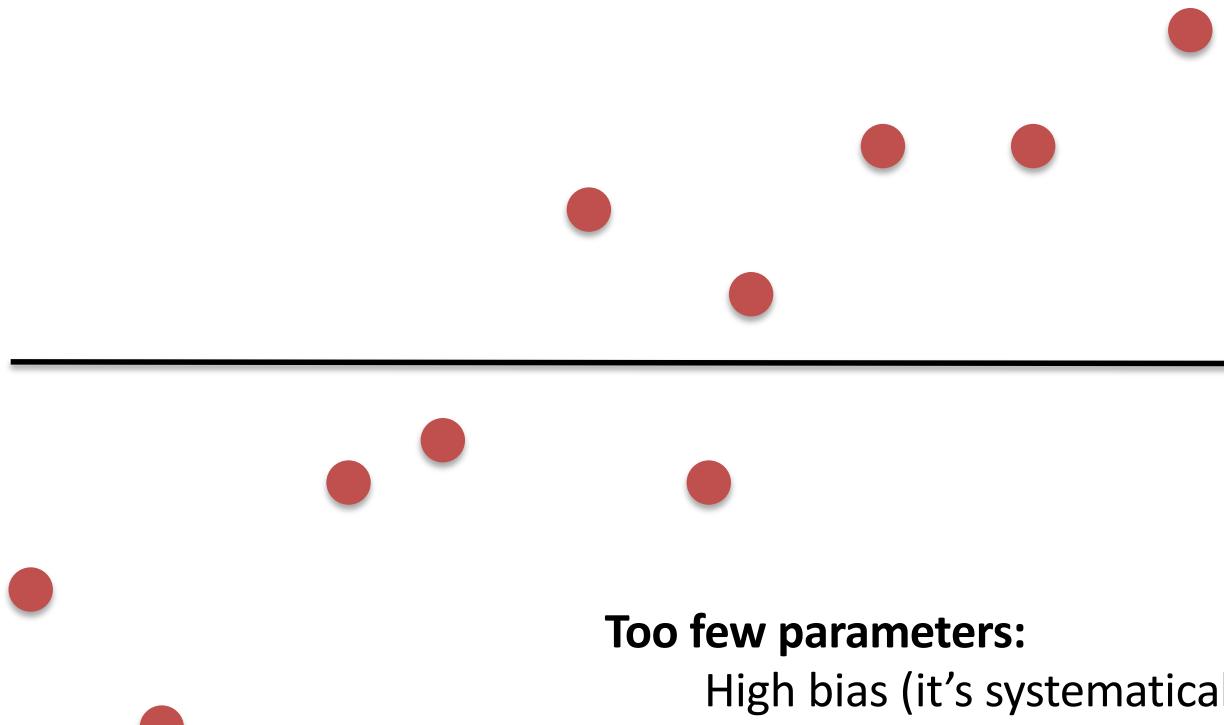




**Too many parameters:**

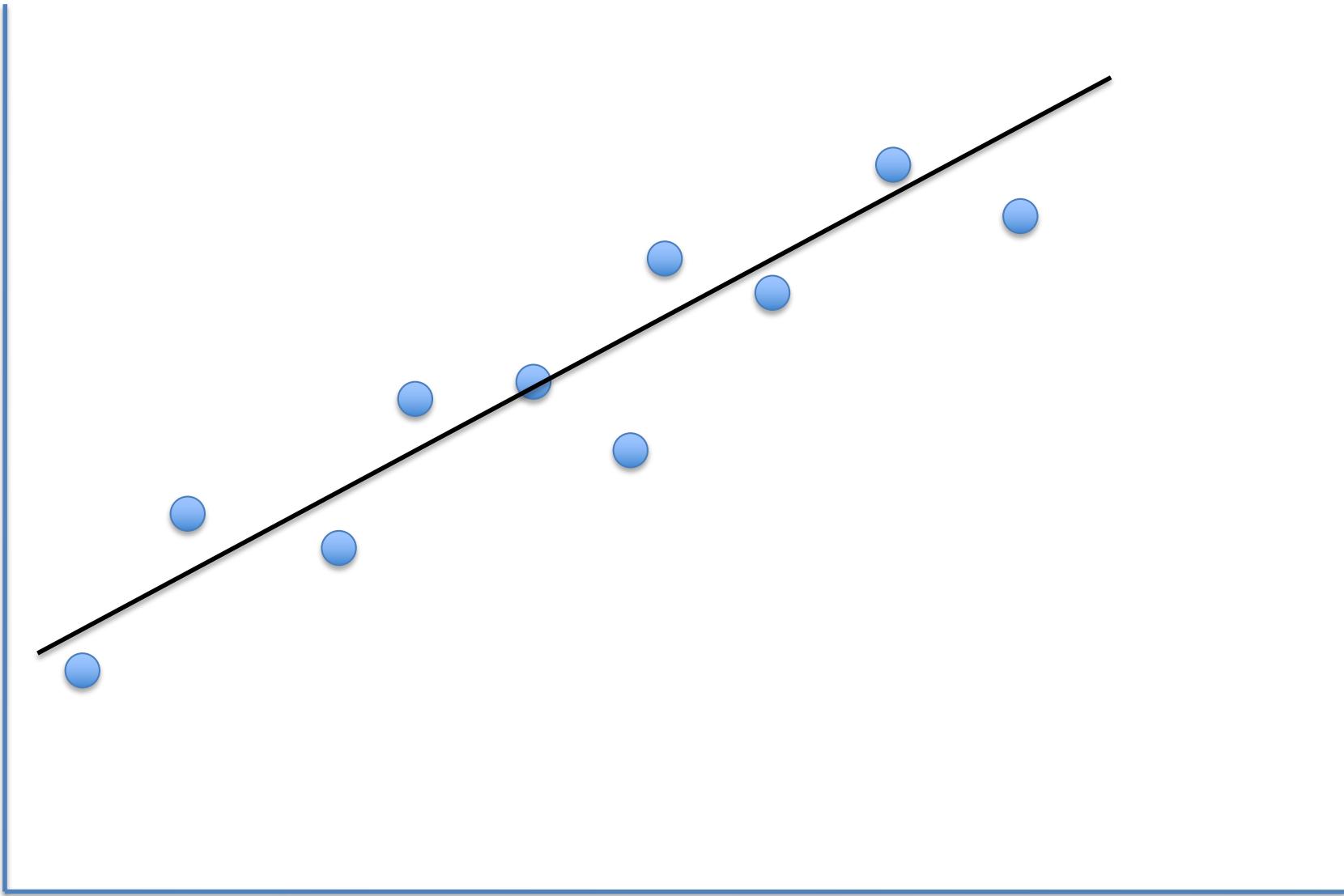
- Low bias (it's not systematically wrong)
- High variance (the parameters have to change a lot)
- High error (we can't predict the truth accurately)

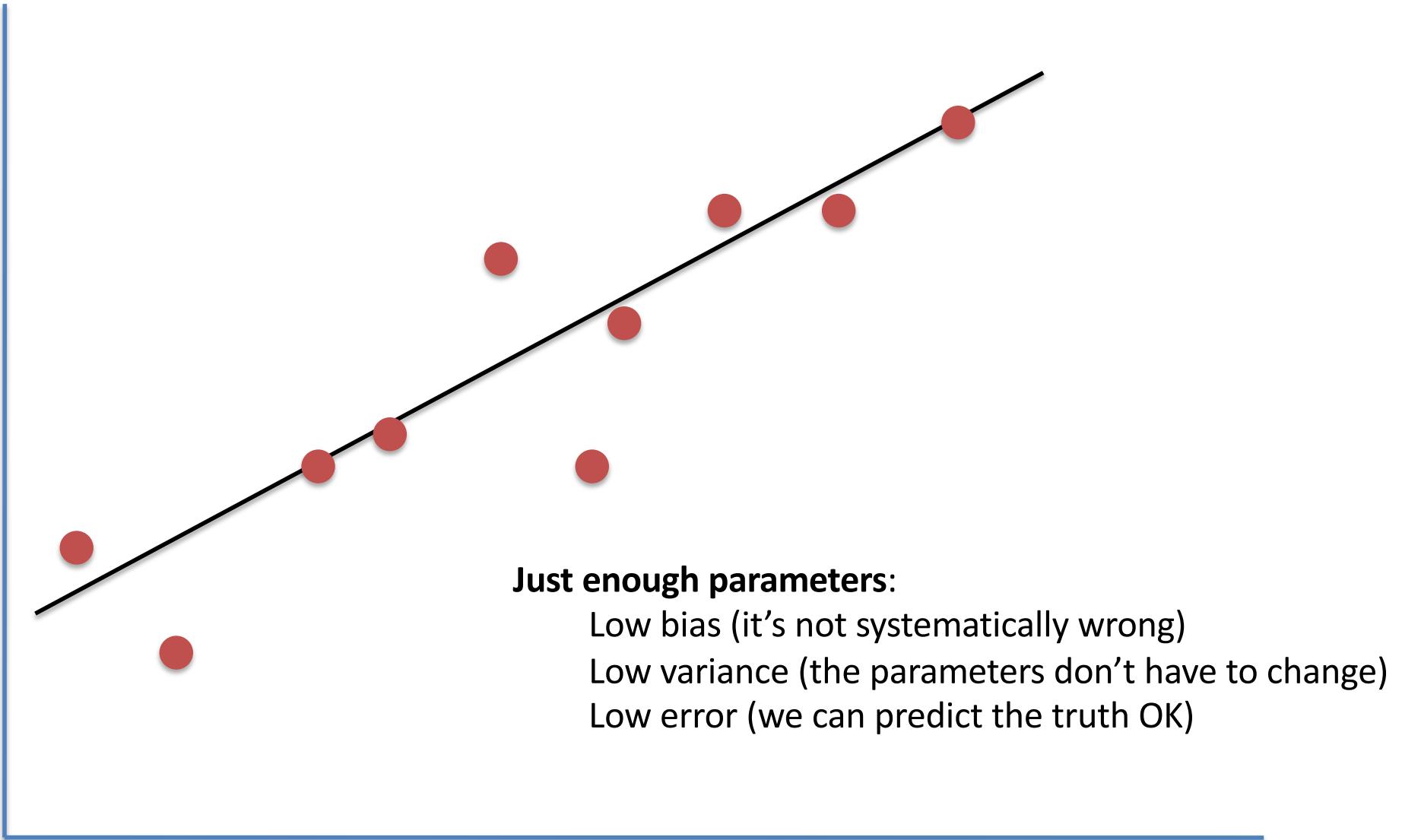




### Too few parameters:

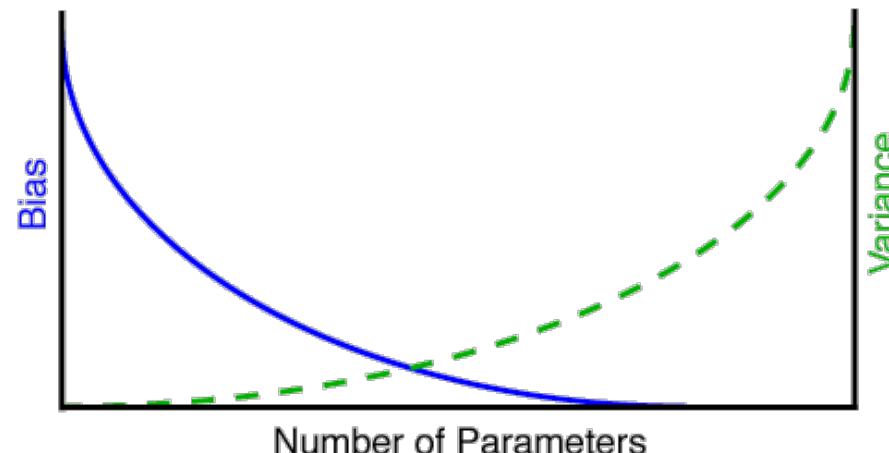
- High bias (it's systematically wrong)
- Low variance (the parameter doesn't change)
- High error (we can't predict the truth accurately)





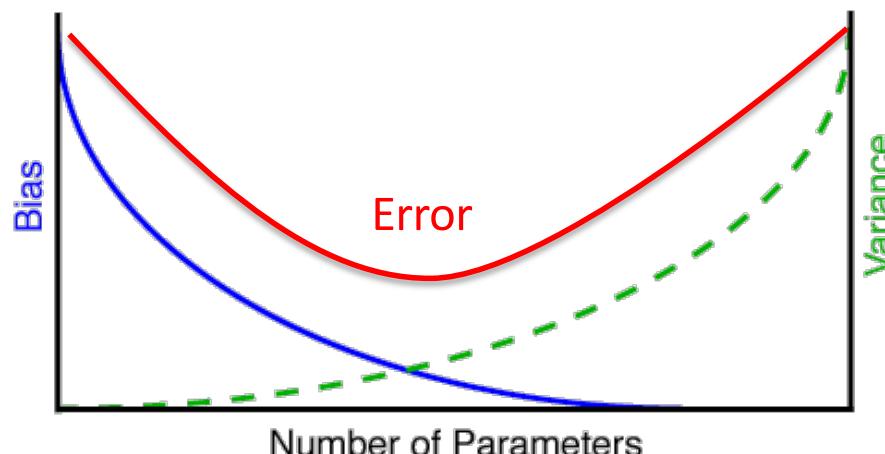
# Key concepts

- 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

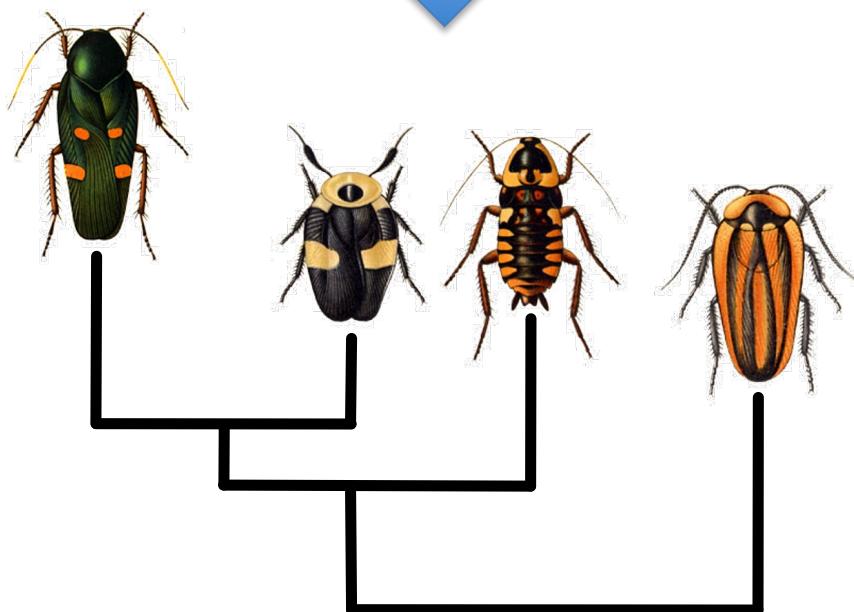


# Models of molecular evolution



actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac

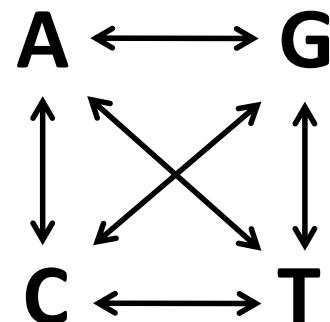
## Molecular evolution





actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgac

### Rate Matrix      Base Frequencies      Site Rates



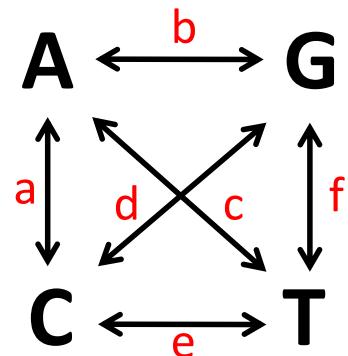
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

$$+ I + G + R$$



actgactgactgactgactgactgactgactgactgactgactgactgac  
 actgactgactgactgactgactgactgactgactgactgactgac  
 actgactgactgactgactgactgactgactgactgactgac  
 actgactgactgactgactgactgactgactgactgac

### Rate Matrix      Base Frequencies      Site Rates



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

$$+ I + G + R$$

#### JC

$$a=b=c=d=e=f$$

$$\pi_A = \pi_C = \pi_G = \pi_T$$

No I or G

0 free parameters

#### HKY

$$a=c=d=f, b=e$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

No I or G

4 free parameters

#### GTR

$$a, b, c, d, e, f$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

No I or G

8 free parameters

#### GTR+I+G

$$a, b, c, d, e, f$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

I, G

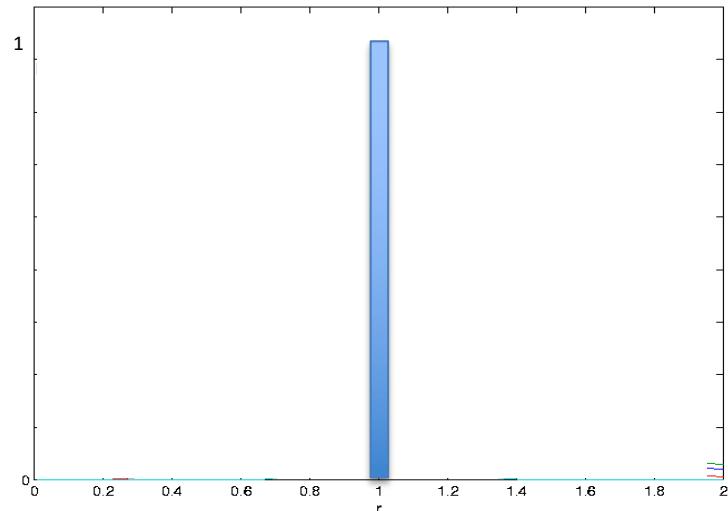
10 free parameters

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

	Fast	Slow	Medium	
Consensus	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C T A G C C .			
Balaena mysticetus	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G G C C C A C T A G C C .	1,220		
Balaenoptera acutorostrata	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C C A C T A G C C .			
Caperea marginata	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G A C T C A C T A G C C .			
Cephalorhynchus eutropia	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C A G C C .			
Delphinapterus leucas	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C T G G C C .			
Delphinus delphis	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C T A G C C .			
Eschrichtius robustus	A A A A A - C C C T C C G A G T G A T T A A A A C C T A G G C T C A C T A G C C .			
Eubalaena australis	A A A A A - C C C T C C G A G T G A - T T A A A A G C C T A G G C C C A C C A G C C .			
Hippopotamus amphibius	A A G A A - C C C T C C G A G T G A - T A A A A A T C T A G A C T C A C C A G T C .			
Hyperoodon ampullatus	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C T T A C C A G G C .			
Inia geoffrensis	G A A A A - C C C T C C G A G T G A T T A T A A G C C T A G G C C T A C T A G C C .			
Kogia breviceps	A A A A A C - C C C T C C G A G T G A - T T A G A G G C C T A G G C C T A C T A G C C .			
Lipotes vexillifer	A A A A A T - C C C T C C G A G T G A - T T A A A G C C T A G G C C T A C C C A G C C .			
Megaptera novaeangliae	A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C C A C T A G C C .			
Mesoplodon peruvianus	A A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C C C A G C C .			
Orcinus orca	A A A A A C - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C C A G C C .			
Phocoena phocoena	A A A A A - C C C T C C G A G T G A - T T A A A A C C T A G G C C T A C T A G C C .			
Physeter catodon	A A A A A C - C C C T C C G A G T G A - T T A A A - C C T A G G C C T A C C C A G C C .			
Platanista gangetica	A A A A A A - C C C T C C G A G T G A - T T A A A G C C T A G G C C T A C C C A G C C .			
Pontoporia blainvillei	A A A A A C - C C C T C C G A G T G A - T T A A A G C C T A G G C C C A C T A G C C .			
Pseudorca crassidens	A A A A A - C C C T C C G A G T G A - T C A A A A C T T A G G C C T A C T A G C C .			
Steno bredanensis	G A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C C A G C C .			
Ziphius cavirostris	A A A A A A - C C C T C C G A G T G A - T T A A A A C T T A G G C C T A C C C A G C C .			

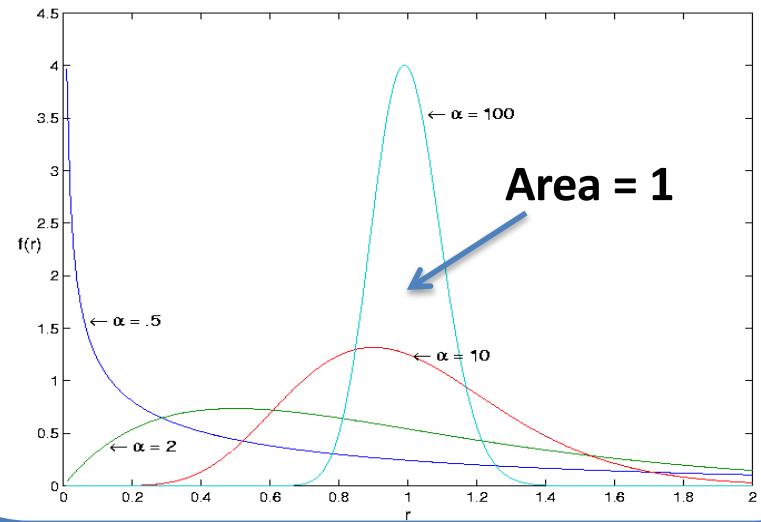
## GTR

- All sites evolve at equal rates



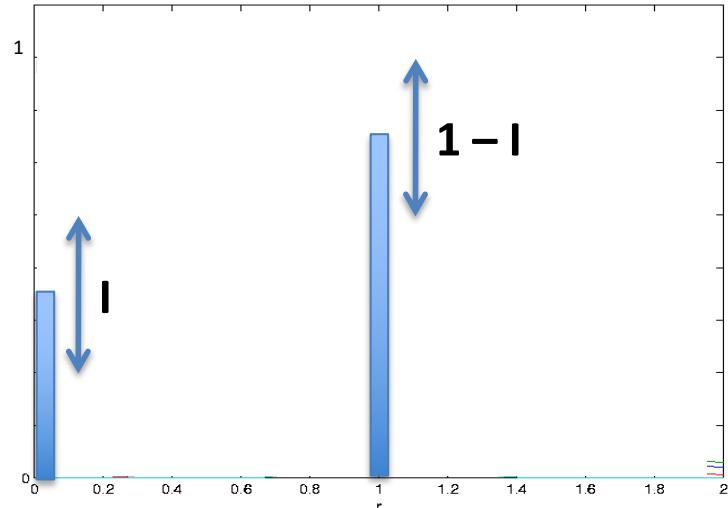
## GTR+G

- Rates vary according to gamma dist.



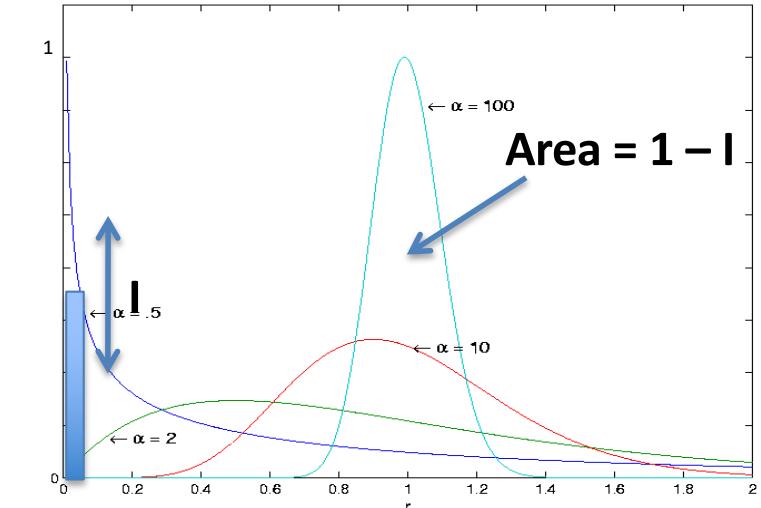
## GTR+I

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



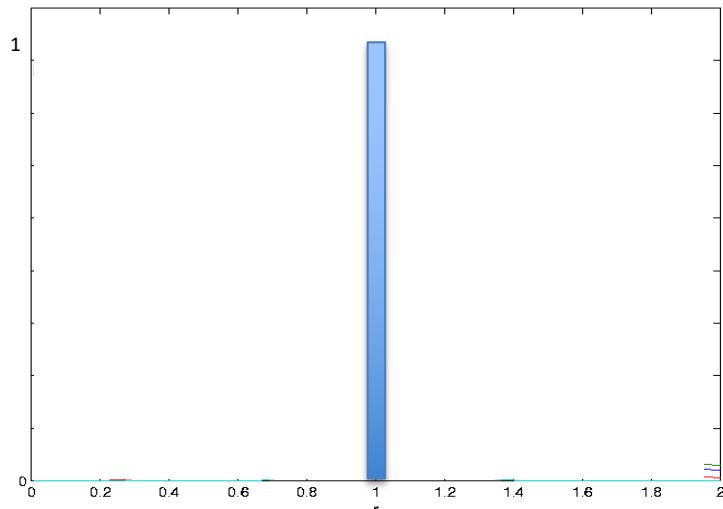
## GTR+I+G

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



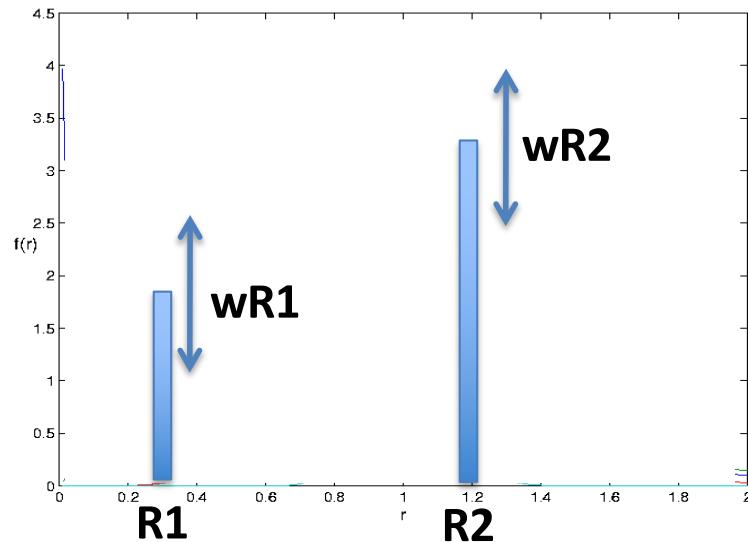
## GTR

- All sites evolve at equal rates



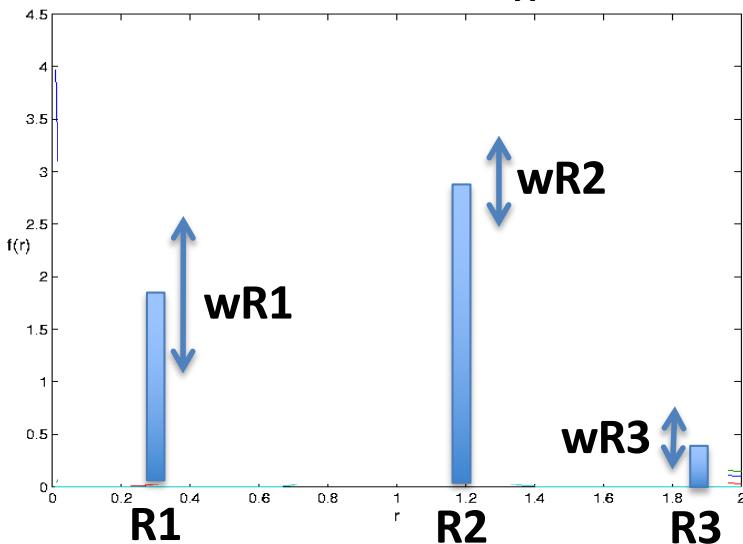
## GTR+R2

- Two free rate categories



## GTR+R2

- Two free rate categories



## GTR+RX

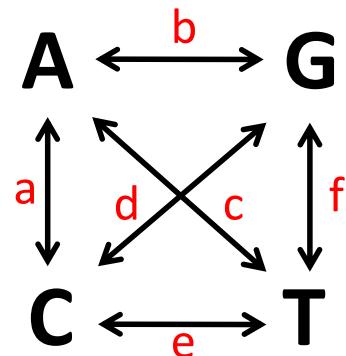
- Add as many rate categories as you like
- More data: more categories
- More rate variation: more categories

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



actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgac

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

The figure displays a grid of DNA sequence data. The sequences are arranged in rows and columns. Some positions in the grid contain colored boxes (yellow, green, red, blue) over specific letters, indicating mutations or specific bases of interest. A large, bold, black word "Partitioning" is centered in the middle of the grid.

Partitioning



actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgac

## Molecular evolution

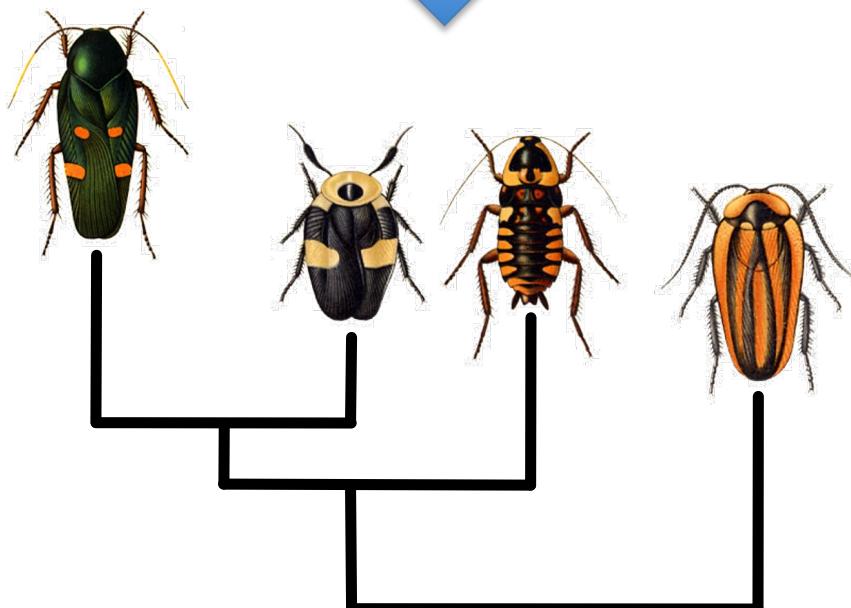
**GTR+I+G**

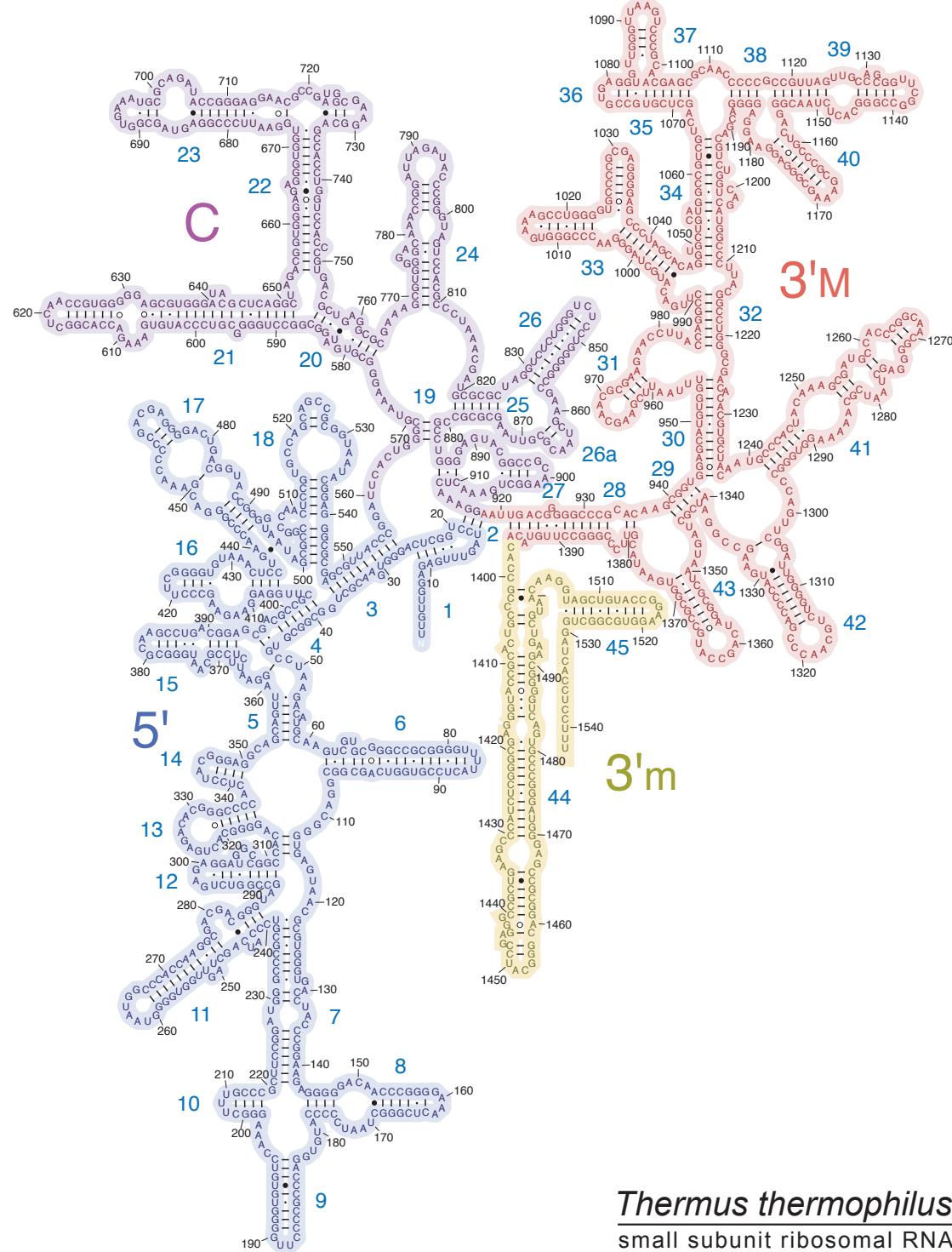
a, b, c, d, e, f

$\pi_A, \pi_C, \pi_G, \pi_T$

I, G

10 free parameters







actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgac

Molecular evolution

Stems+Loops

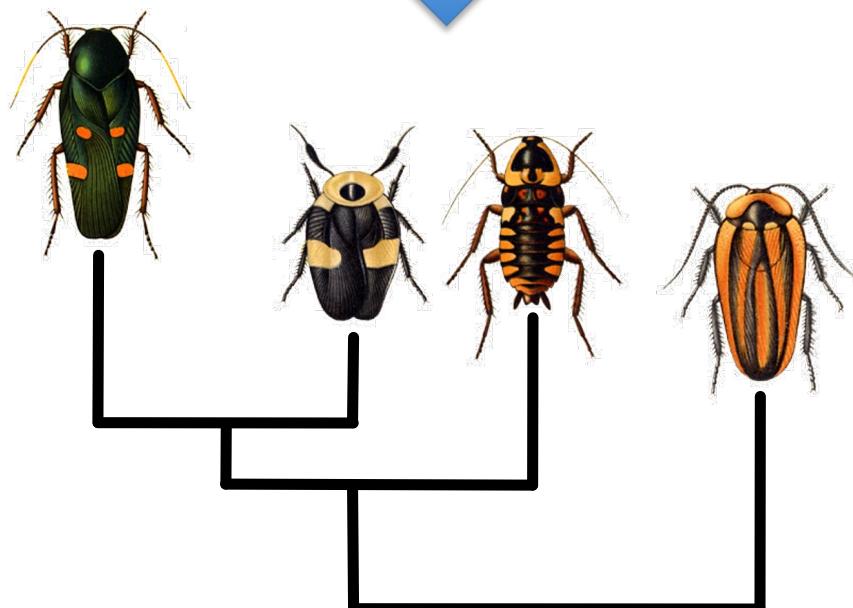
GTR+I+G

a, b, c, d, e, f

$\pi_A, \pi_C, \pi_G, \pi_T$

I, G

10 free parameters





actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac

Molecular evolution Molecular evolution

Stems

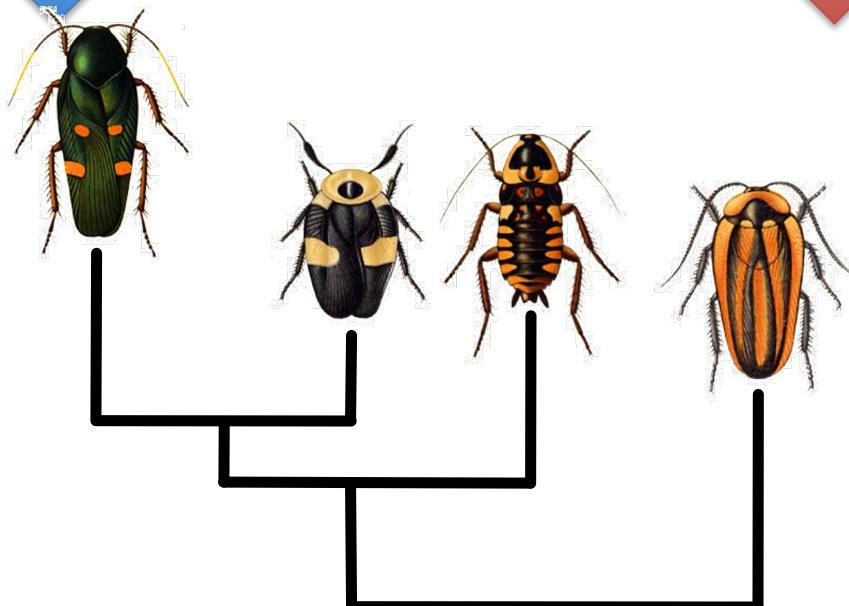
GTR+I+G

a, b, c, d, e, f

$\pi_A, \pi_C, \pi_G, \pi_T$

I, G

10 free parameters



Loops

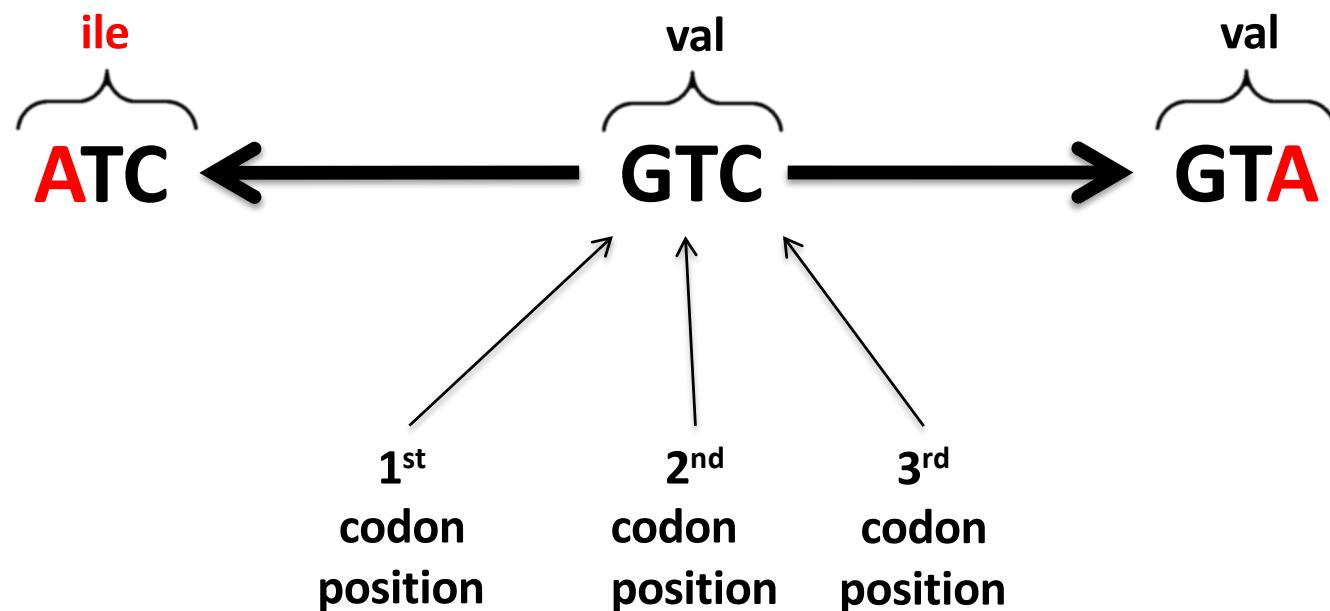
GTR+I+G

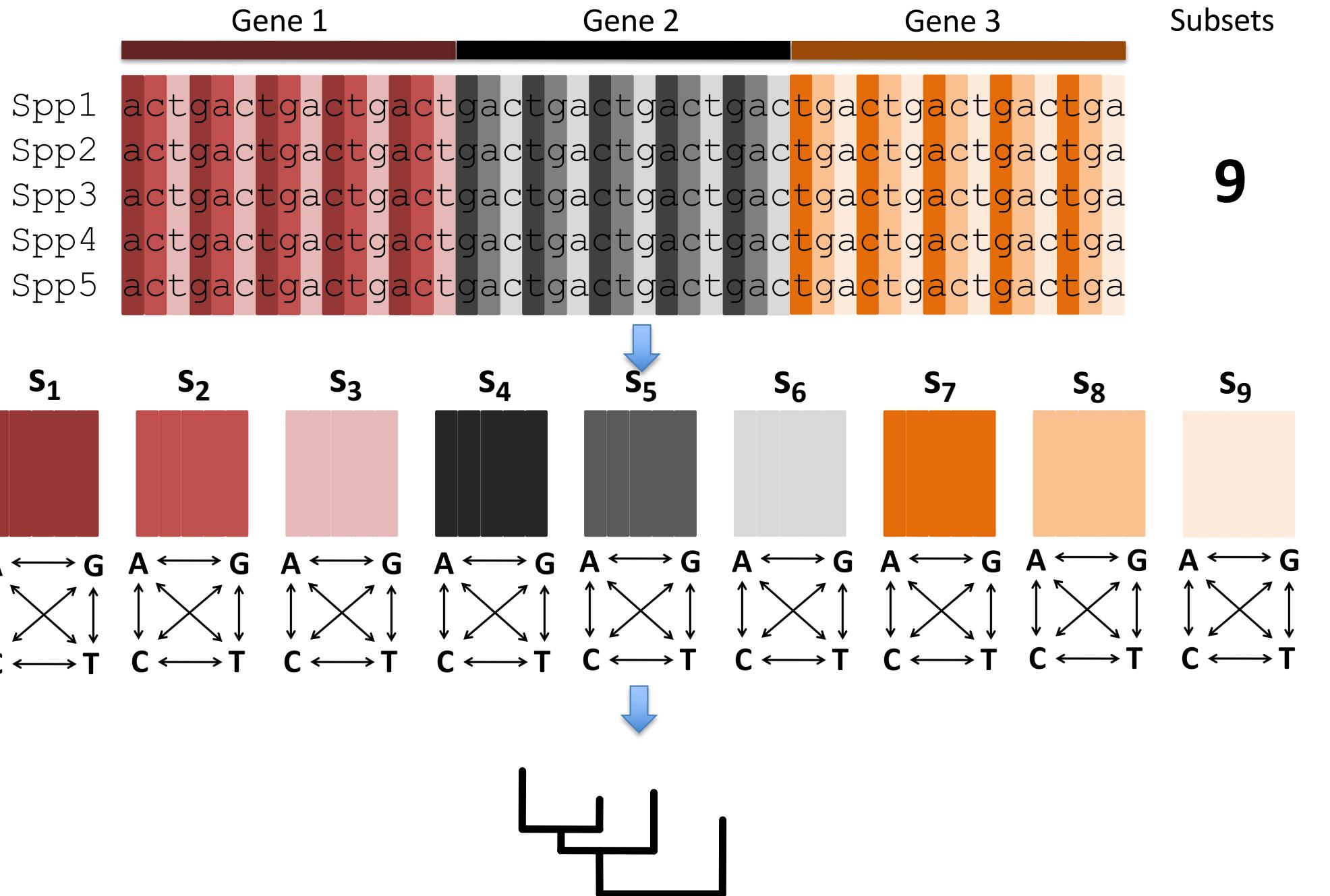
a, b, c, d, e, f

$\pi_A, \pi_C, \pi_G, \pi_T$

I, G

10 free parameters





# The Problem

## Gene 1

## Gene 2

## Gene 3

## Subsets

9

6

2

Spp1

Spp2

Spp3

Spp4

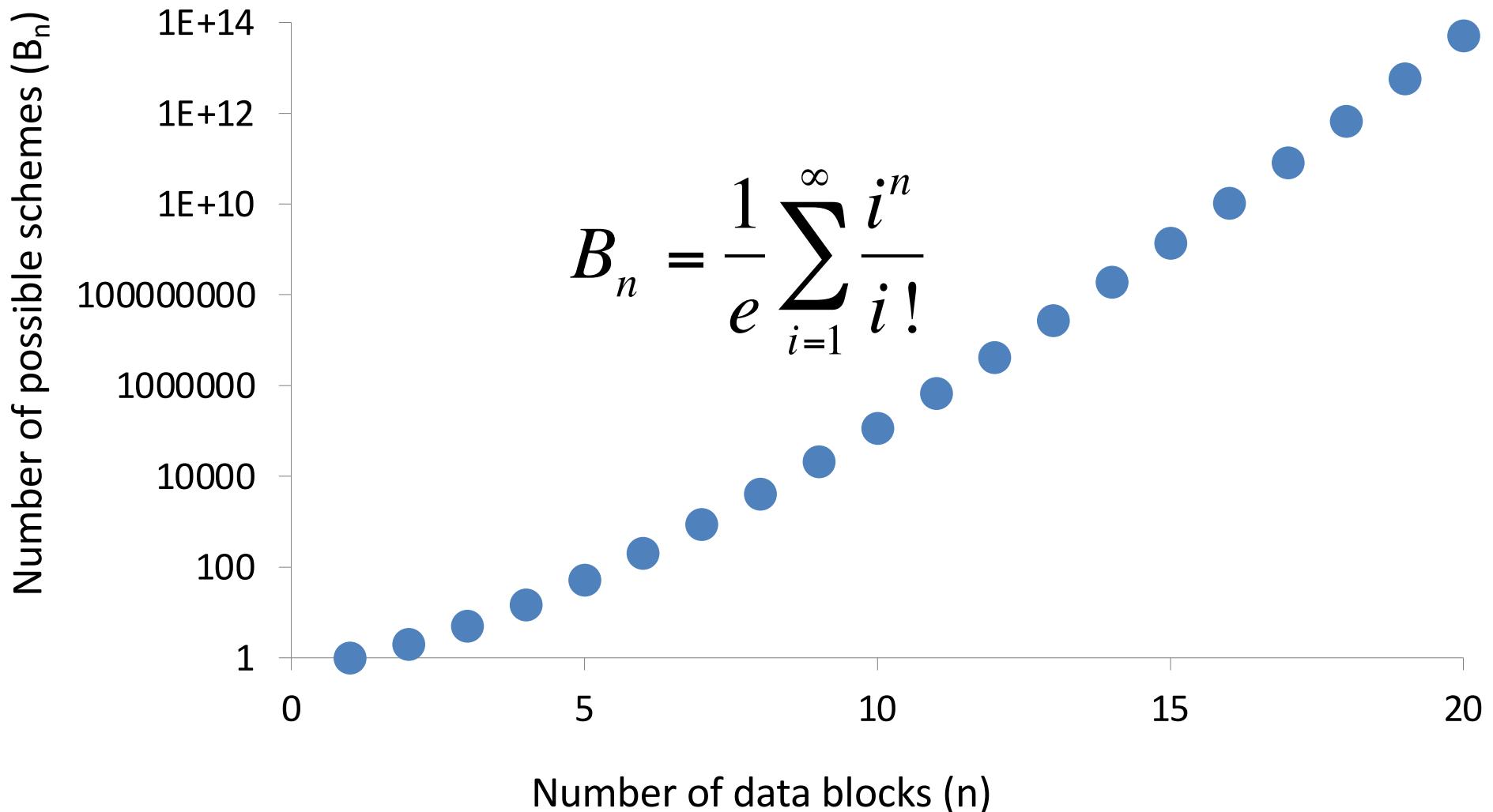
Spp5

actgactgactgactgactgactgactgactgactgactgactgactgactgactgactga  
actgactgactgactgactgactgactgactgactgactgactgactgactgactgactga  
actgactgactgactgactgactgactgactgactgactgactgactgactgactgactga  
actgactgactgactgactgactgactgactgactgactgactgactgactgactgactga  
actgactgactgactgactgactgactgactgactgactgactgactgactgactgactga

A horizontal bar chart with 15 bars. The bars are colored in a gradient: dark red, light pink, medium pink, dark red, light pink, medium pink, dark grey, dark grey, dark grey, dark grey, orange, light orange, medium orange, light orange, and orange. The bars are of varying heights, indicating the magnitude of the variable for each category.

A horizontal bar chart illustrating the frequency of 18 distinct categories. The categories are represented by vertical bars of three colors: dark red, light pink, and black. The black bars are positioned at the far left and right ends of the chart. The remaining 16 categories are composed of dark red bars, with one category appearing 14 times and another appearing 2 times.

# 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?

The figure displays a grid of DNA sequence data. The sequences are arranged in rows, each starting with 'AAA' and ending with 'TGA'. The data is color-coded to highlight specific mutations:

- Row 1:** The first 'C' in the first column is highlighted in green.
- Row 2:** The second 'C' in the first column is highlighted in red.
- Row 3:** The third 'C' in the first column is highlighted in blue.
- Row 4:** The fourth 'C' in the first column is highlighted in green.
- Row 5:** The fifth 'C' in the first column is highlighted in red.
- Row 6:** The sixth 'C' in the first column is highlighted in blue.
- Row 7:** The seventh 'C' in the first column is highlighted in green.
- Row 8:** The eighth 'C' in the first column is highlighted in red.
- Row 9:** The ninth 'C' in the first column is highlighted in blue.
- Row 10:** The tenth 'C' in the first column is highlighted in green.
- Row 11:** The eleventh 'C' in the first column is highlighted in red.
- Row 12:** The twelfth 'C' in the first column is highlighted in blue.
- Row 13:** The thirteenth 'C' in the first column is highlighted in green.
- Row 14:** The fourteenth 'C' in the first column is highlighted in red.
- Row 15:** The fifteenth 'C' in the first column is highlighted in blue.
- Row 16:** The sixteenth 'C' in the first column is highlighted in green.
- Row 17:** The seventeenth 'C' in the first column is highlighted in red.
- Row 18:** The eighteenth 'C' in the first column is highlighted in blue.
- Row 19:** The nineteenth 'C' in the first column is highlighted in green.
- Row 20:** The twentieth 'C' in the first column is highlighted in red.
- Row 21:** The twenty-first 'C' in the first column is highlighted in blue.
- Row 22:** The twenty-second 'C' in the first column is highlighted in green.
- Row 23:** The twenty-third 'C' in the first column is highlighted in red.
- Row 24:** The twenty-fourth 'C' in the first column is highlighted in blue.
- Row 25:** The twenty-fifth 'C' in the first column is highlighted in green.
- Row 26:** The twenty-sixth 'C' in the first column is highlighted in red.
- Row 27:** The twenty-seventh 'C' in the first column is highlighted in blue.
- Row 28:** The twenty-eighth 'C' in the first column is highlighted in green.
- Row 29:** The twenty-ninth 'C' in the first column is highlighted in red.
- Row 30:** The thirtieth 'C' in the first column is highlighted in blue.
- Row 31:** The thirty-first 'C' in the first column is highlighted in green.
- Row 32:** The thirty-second 'C' in the first column is highlighted in red.
- Row 33:** The thirty-third 'C' in the first column is highlighted in blue.
- Row 34:** The thirty-fourth 'C' in the first column is highlighted in green.
- Row 35:** The thirty-fifth 'C' in the first column is highlighted in red.
- Row 36:** The thirty-sixth 'C' in the first column is highlighted in blue.
- Row 37:** The thirty-seventh 'C' in the first column is highlighted in green.
- Row 38:** The thirty-eighth 'C' in the first column is highlighted in red.
- Row 39:** The thirty-ninth 'C' in the first column is highlighted in blue.
- Row 40:** The forty-thousandth 'C' in the first column is highlighted in green.

**Questions?**

# Objective Model Selection

# 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

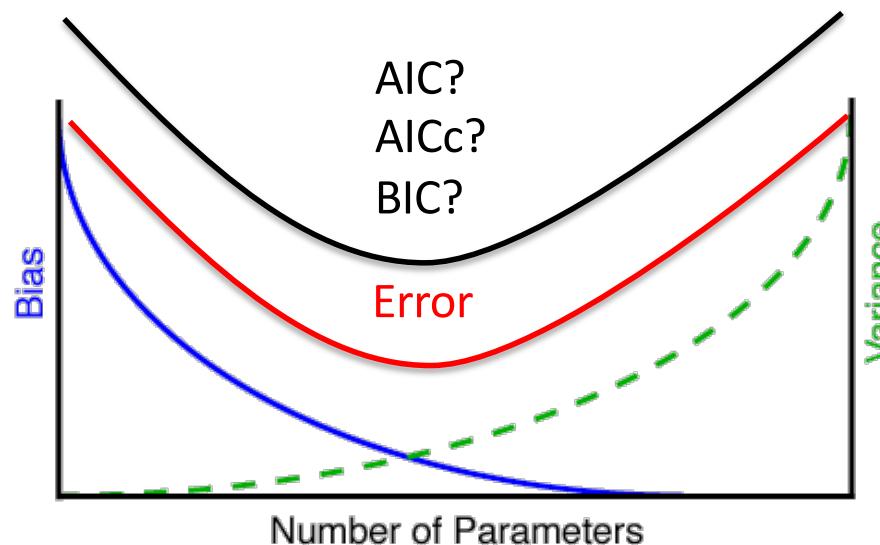
- The AIC, corrected for sample size. In theory, this should always be preferred to the AIC, because it's asymptotic to the AIC when you have large samples. **What is the sample size in molecular phylogenetics though?**

## BIC

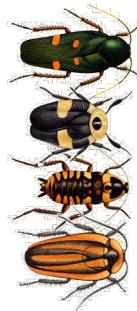
- Bayesian Information Criterion. Similar to AIC, but motivated from a Bayesian perspective. (Some would say it's completely made up!)
- Assumes the true model is in the set

# Model Selection Criteria

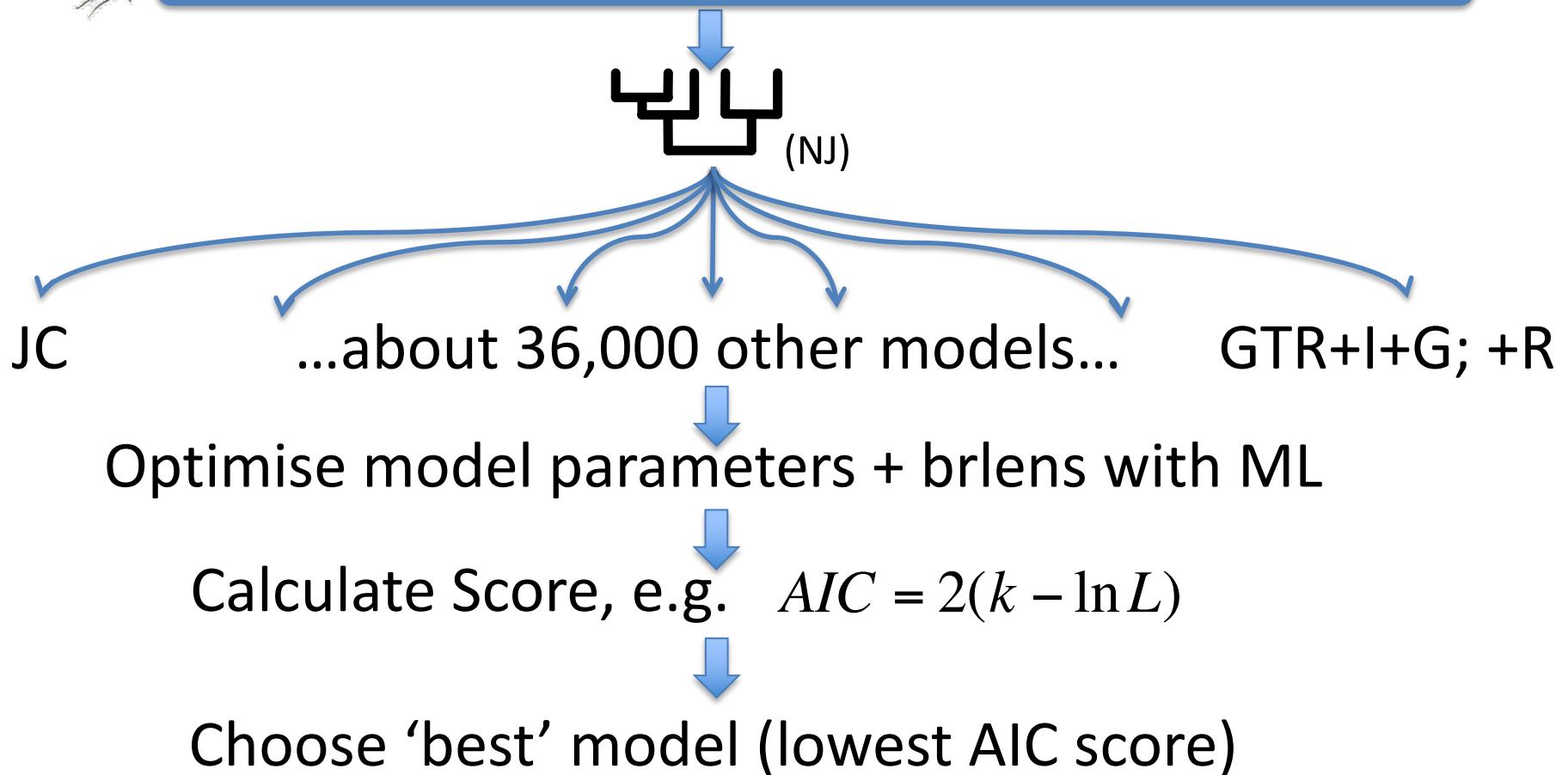
- Each criterion makes subtly different assumptions about how the world works
- In practice, most perform fine in most situations\*



# Selecting a model from the GTR family

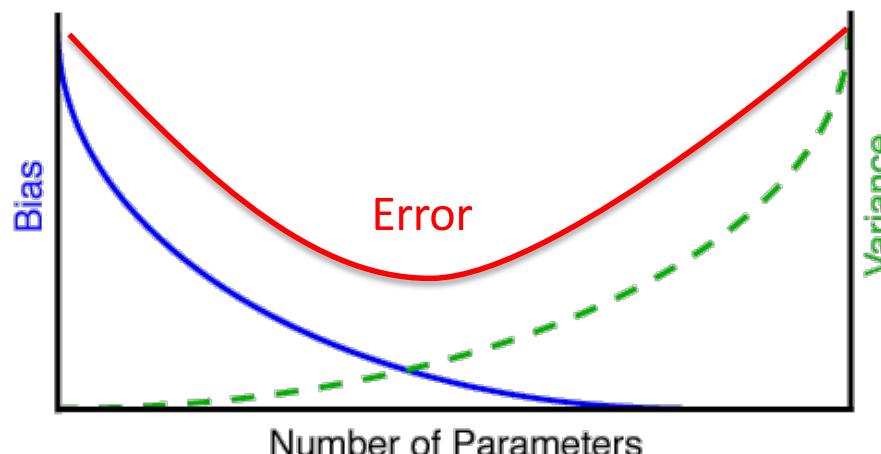


actgactgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac



# A new problem!

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



# The Problem

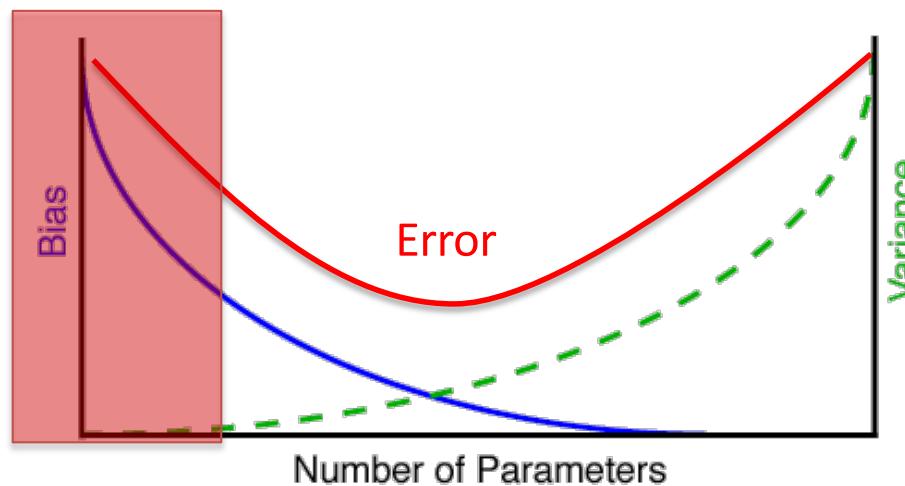


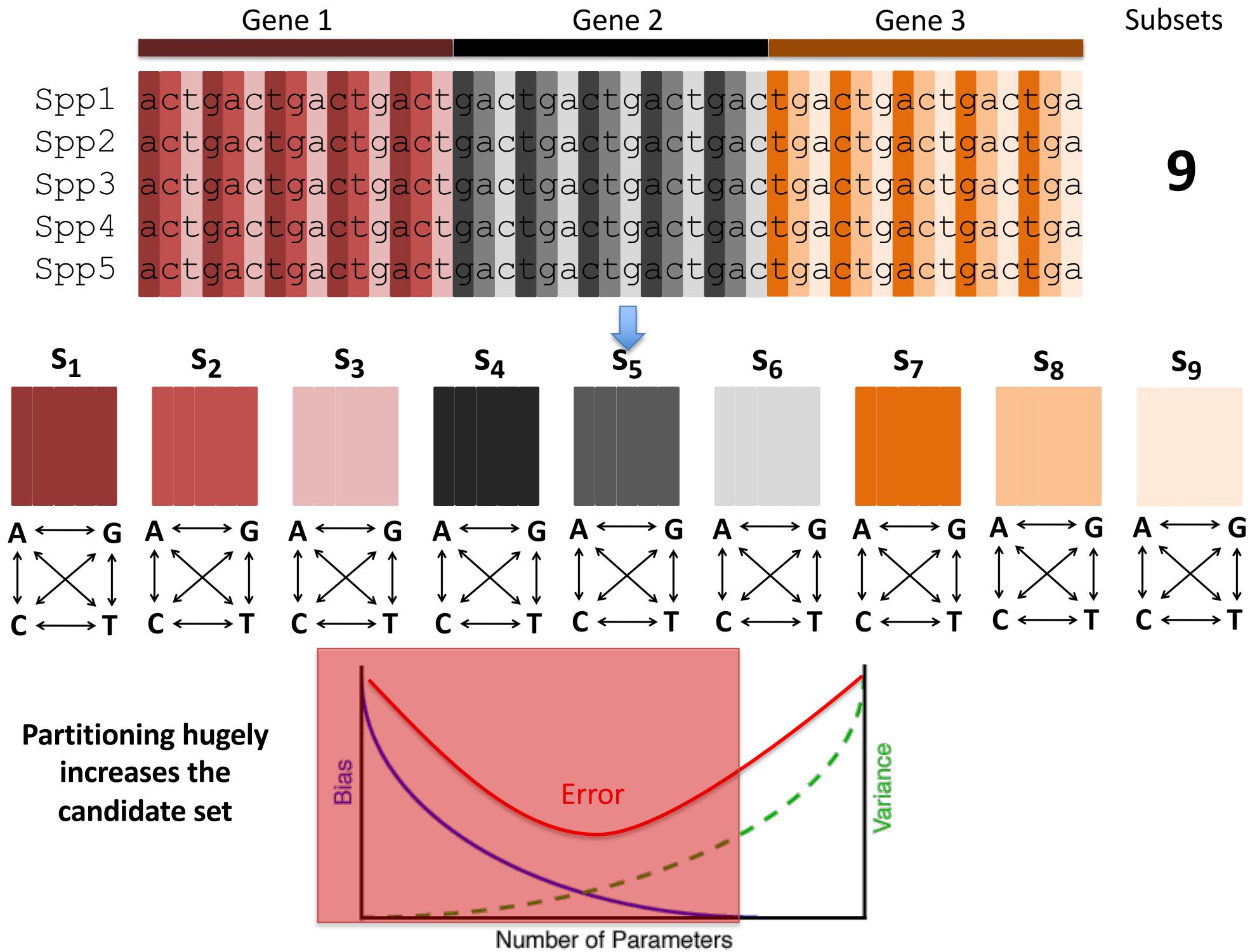
```
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgactgac  
actgactgactgactgactgactgactgactgactgactgac
```

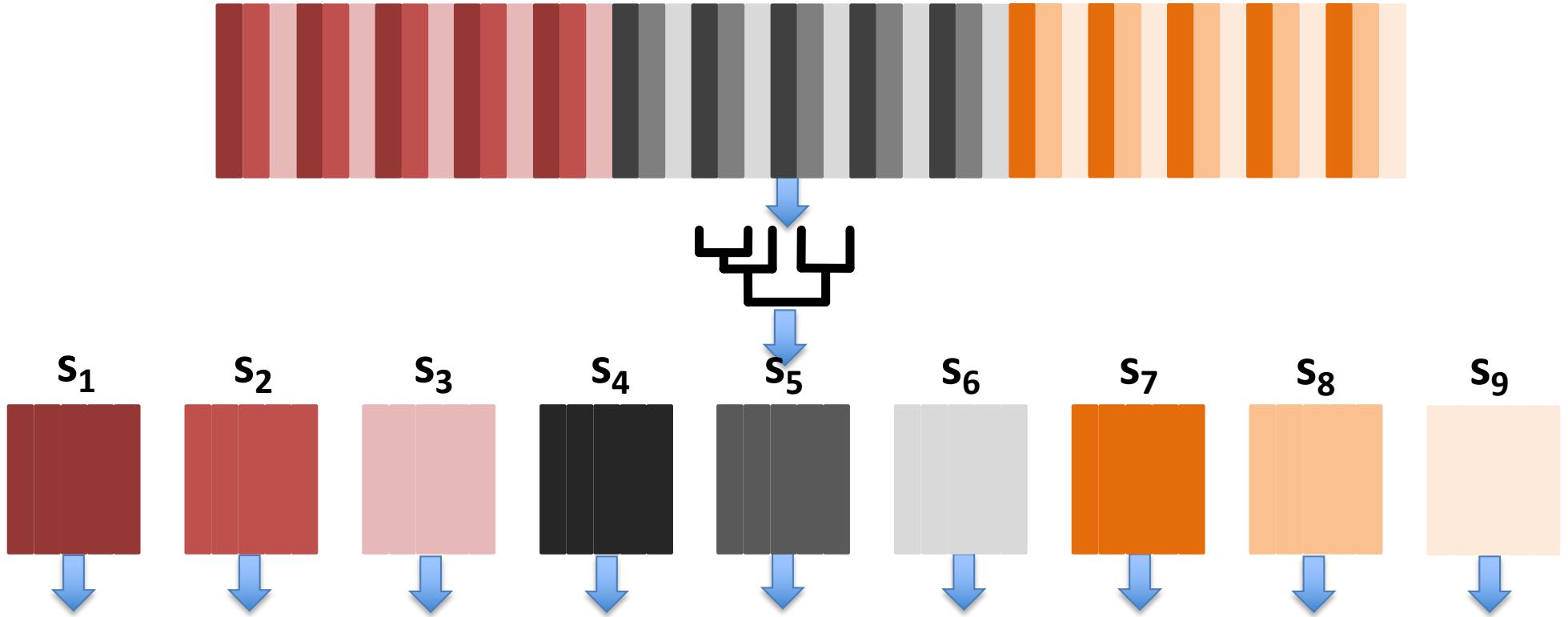


Almost always select **GTR+I+G**  
(the most complex model)

Our candidate set is too small!







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 partitioning scheme, with models of molecular evolution for each partition

# Questions?

How to do all of this in practice



# IQ-TREE

## Efficient software for phylogenomic inference

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

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

your  
concatenated  
alignment

your  
partition  
file

**ModelFinder2 +**  
**Plus tree search +**  
**MERGE subsets**



# IQ-TREE

## Efficient software for phylogenomic inference

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

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

your  
folder of  
alignments

ModelFinder2 +  
Plus tree search +  
**MERGE** subsets

# Questions?

# That's it.

AAA - C C C T C C G A G T G A - T T A A A G C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A G C C T A G G A  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
GAA - T C C T C C G A G T G A - T A A A A A T C T A G A C  
AAA - A C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - A C C T C C G A G T G A - T A T A A G C C T A G G C  
AAC C - C C C T C C G A G T G A - T T A G G A G C C T A G G C  
AAA - T C C T C C G A G T G A - T T A A A G C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - A C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - A C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - A C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T T A A A A C C T A G G C  
AAA - C C C T C C G A G T G A - T C A A A A C C T A G G C

**That's it.**