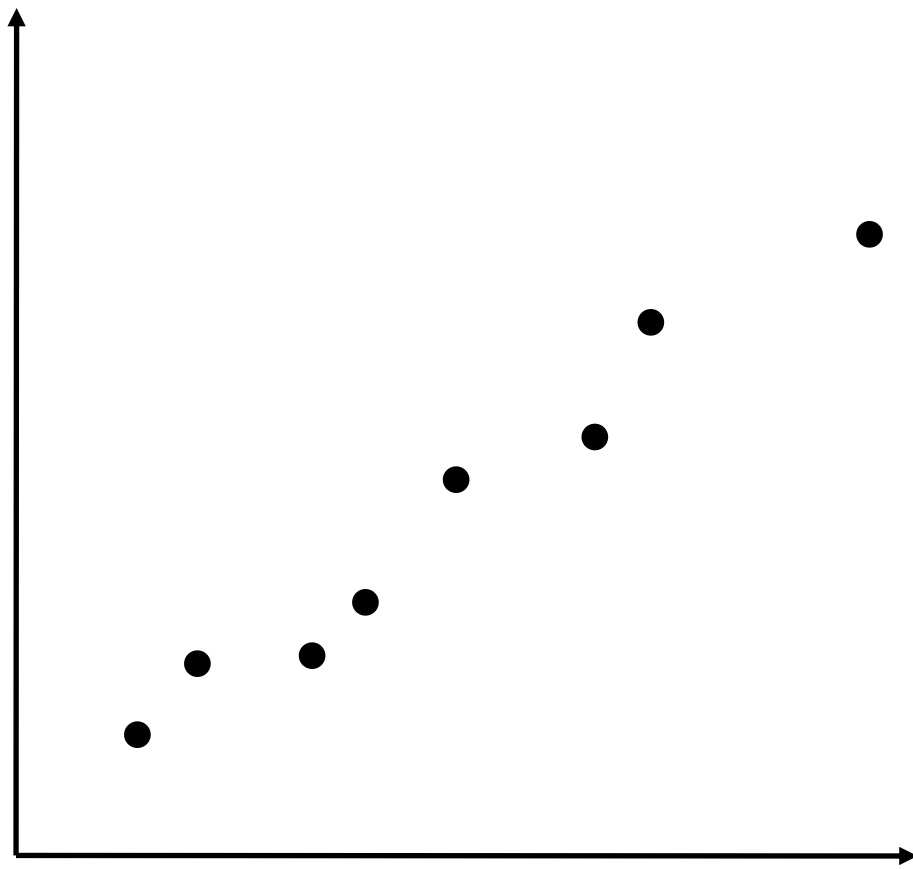


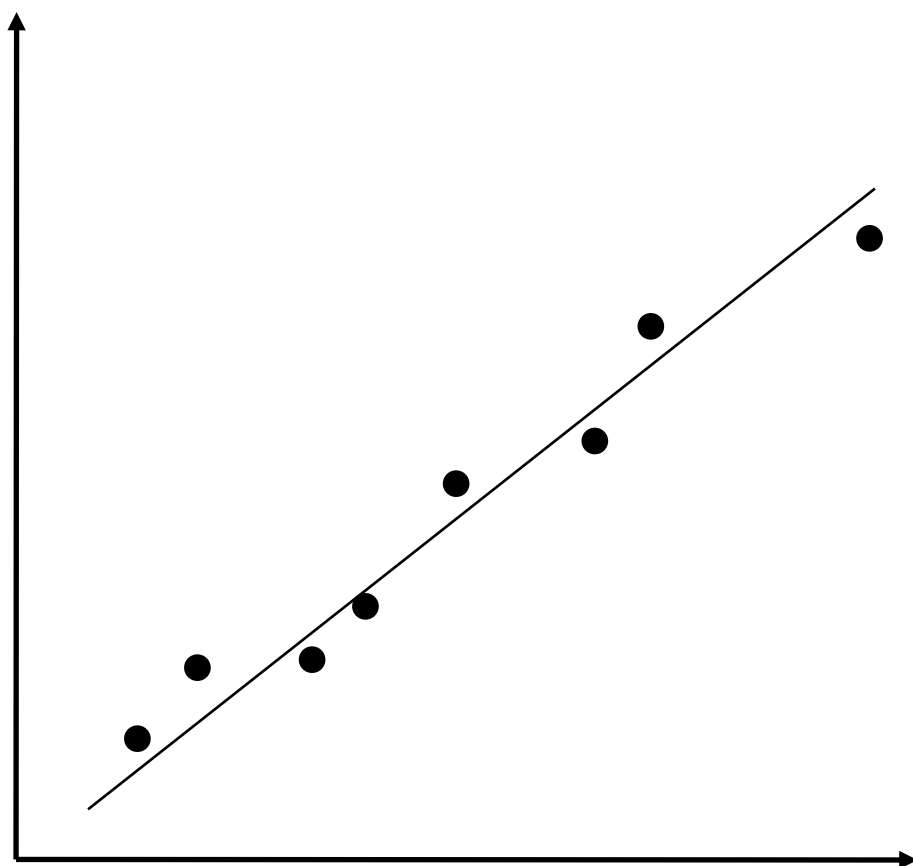
# Model Selection

# Modeling: An example

---



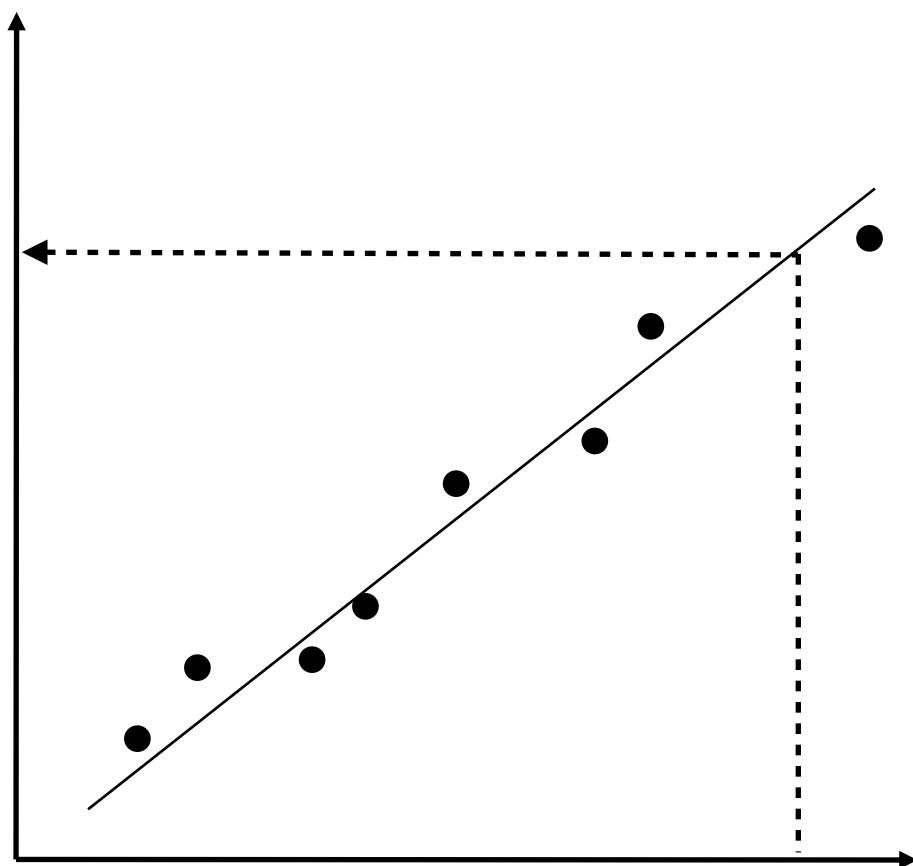
# Modeling: An example



$$y = bx + a$$

Simple 2-parameter model

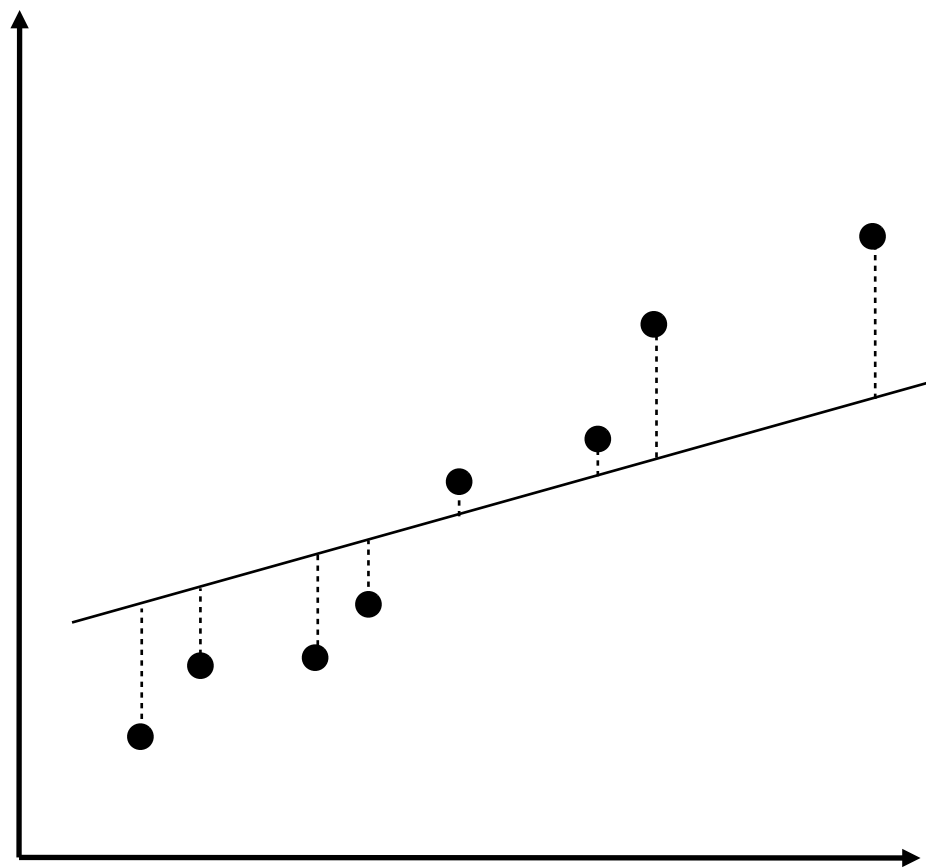
# Modeling: An example



$$y = bx + a$$

Predictions based on model

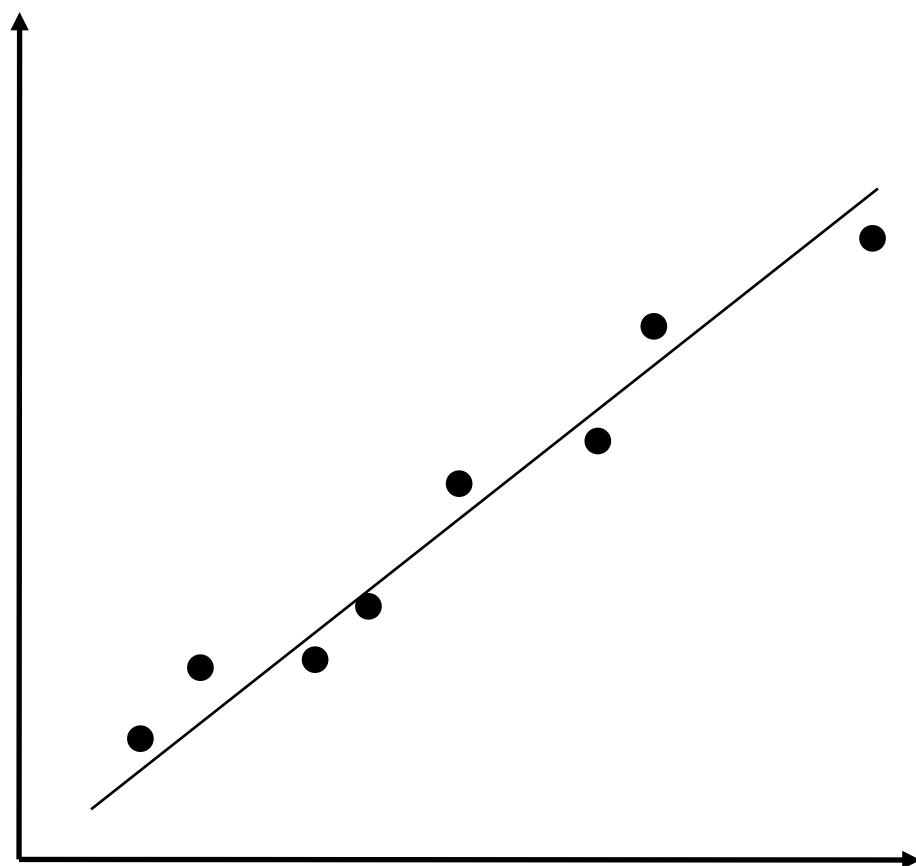
# Modeling: An example



$$y = bx + a$$

- Measure of how well the model fits the data: sum of squared errors (SSE)
- Best parameter estimates: those that give the smallest SSE (least squares model fitting)

# Modeling: An example



$$y = 0.95x - 0.26$$

- Measure of how well the model fits the data: sum of squared errors (SSE)
- Best parameter estimates: those that give the smallest SSE (least squares model fitting)

# Maximum likelihood: likelihood is a measure of model fit

---

- Likelihood (Model) = Probability (Data | Model)
- Maximum likelihood: Best estimate is the set of parameter values which gives the highest possible likelihood.

# Probabilistic modeling applied to phylogeny

- Observed data: multiple alignment of sequences

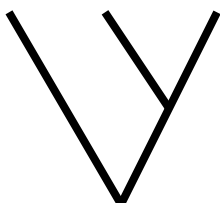
H.sapiens globin	A	G	G	G	A	T	T	C	A
M.musculus globin	A	C	G	G	T	T	T	-	A
R.rattus globin	A	C	G	G	A	T	T	-	A

- Probabilistic model:

- A model of (hypothesis about) how one ancestral sequence has evolved into the three sequences that are present in the alignment

- Probabilistic model parameters (simplest case):

- Tree topology and branch lengths
  - Nucleotide frequencies:  $\pi_A, \pi_C, \pi_G, \pi_T$
  - Nucleotide-nucleotide substitution rates (or substitution probabilities):



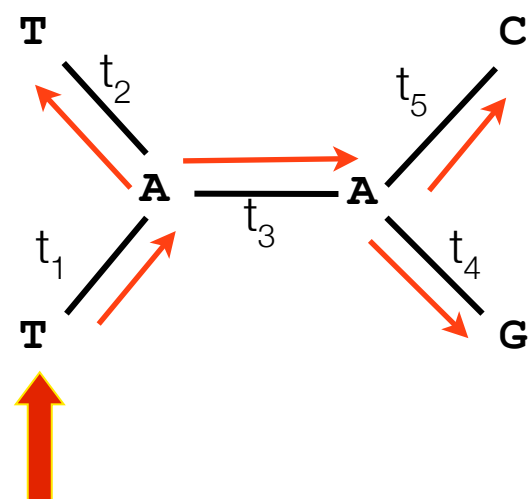
	A	C	G	T
A	$-3\alpha$	$\alpha$	$\alpha$	$\alpha$
C	$\alpha$	$-3\alpha$	$\alpha$	$\alpha$
G	$\alpha$	$\alpha$	$-3\alpha$	$\alpha$
T	$\alpha$	$\alpha$	$\alpha$	$-3\alpha$

 $\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$



# Computing the probability of one column in an alignment given tree topology and other parameters

A	T	G	G	A	T	T	C	A
A	T	G	G	T	T	T	-	A
A	C	G	G	A	T	T	-	A
A	G	G	G	T	T	T	-	A



$$\text{Pr} = \pi_T P_{TA}(t_1) P_{AT}(t_2) P_{AA}(t_3) P_{AG}(t_4) P_{AC}(t_5)$$

- Columns in alignment contain homologous nucleotides
- Assume tree topology, branch lengths, and other parameters are given. For now, assume ancestral states were A and A (we'll get to the full computation on next slide). Start computation at any internal or external node. Arrows indicate “direction” of computations (“flowing” away from the starting point).

# Computing the probability of an entire alignment given tree topology and other parameters

A	T	G	G	A	T	T	C	A
A	T	G	G	T	T	T	-	A
A	C	G	G	A	T	T	-	A
A	G	G	G	T	T	T	-	A
	$j$							

$$L_{(j)} = \text{Prob} \left( \begin{array}{c} \text{T} \\ \diagup \\ \boxed{\text{A}} \\ \diagdown \\ \text{T} \end{array} - \begin{array}{c} \text{C} \\ \diagup \\ \boxed{\text{A}} \\ \diagdown \\ \text{G} \end{array} \right) + \text{Prob} \left( \begin{array}{c} \text{T} \\ \diagup \\ \boxed{\text{C}} \\ \diagdown \\ \text{T} \end{array} - \begin{array}{c} \text{C} \\ \diagup \\ \boxed{\text{A}} \\ \diagdown \\ \text{G} \end{array} \right)$$

$$+ \dots + \text{Prob} \left( \begin{array}{c} \text{T} \\ \diagup \\ \boxed{\text{T}} \\ \diagdown \\ \text{T} \end{array} - \begin{array}{c} \text{C} \\ \diagup \\ \boxed{\text{T}} \\ \diagdown \\ \text{G} \end{array} \right)$$

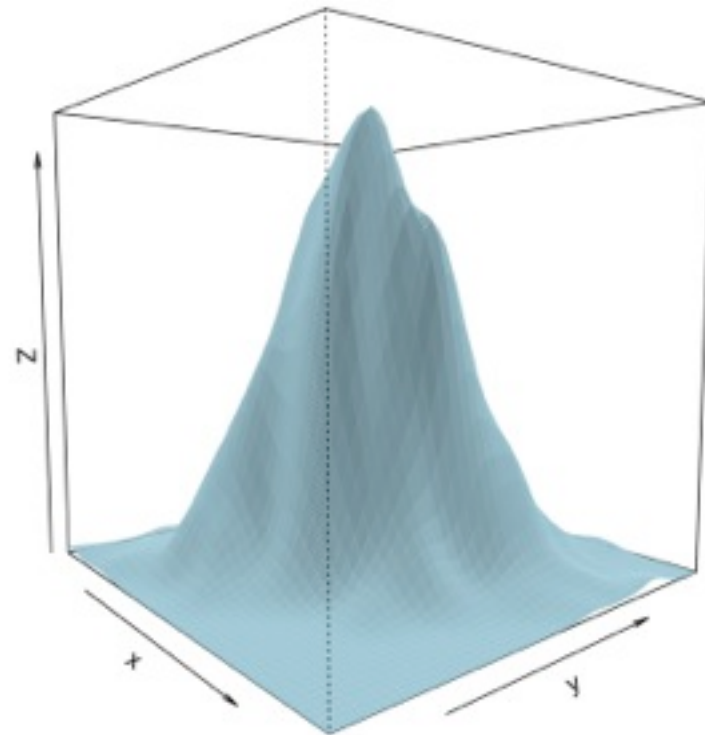
- Probability must be summed over all possible combinations of ancestral nucleotides.
- Here we have two internal nodes giving 16 possible combinations
- Probability of individual columns are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.
- In phylogeny software these computations are done using summation of the logs of the probabilities (“log likelihoods”), because multiplication of the large number of probability terms may lead to underflow (computer problems caused by very small numbers).

$$L = L_{(1)} \cdot L_{(2)} \cdots L_{(N)} = \prod_{j=1}^N L_{(j)}$$

$$\ln(L) = \ln(L_{(1)}) + \ln(L_{(2)}) + \cdots + \ln(L_{(N)}) = \sum_{j=1}^N \ln(L_{(j)})$$

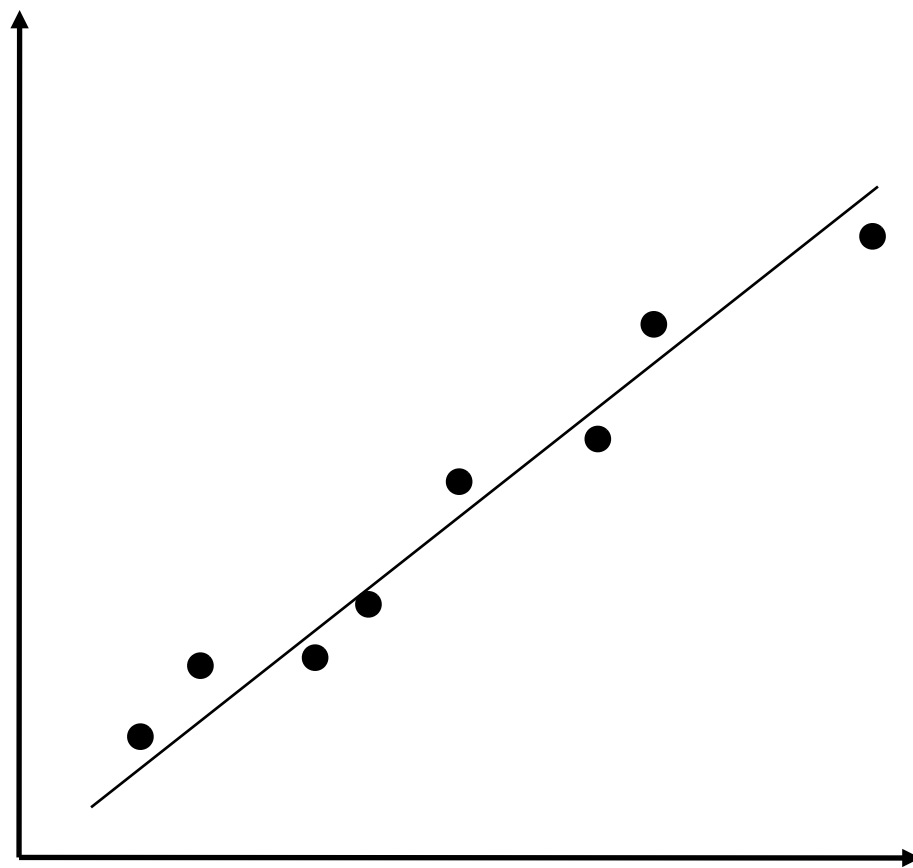
# Maximum likelihood phylogeny

- **Data:**
  - sequence alignment
- **Model parameters:**
  - nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.



- Choose random initial values for all parameters, compute likelihood
- Change parameter values slightly in a direction so likelihood improves
- Repeat until maximum found
- Results:
  - ML estimate of tree topology
  - ML estimate of branch lengths
  - ML estimate of other model parameters
  - Measure of how well model fits data (likelihood).

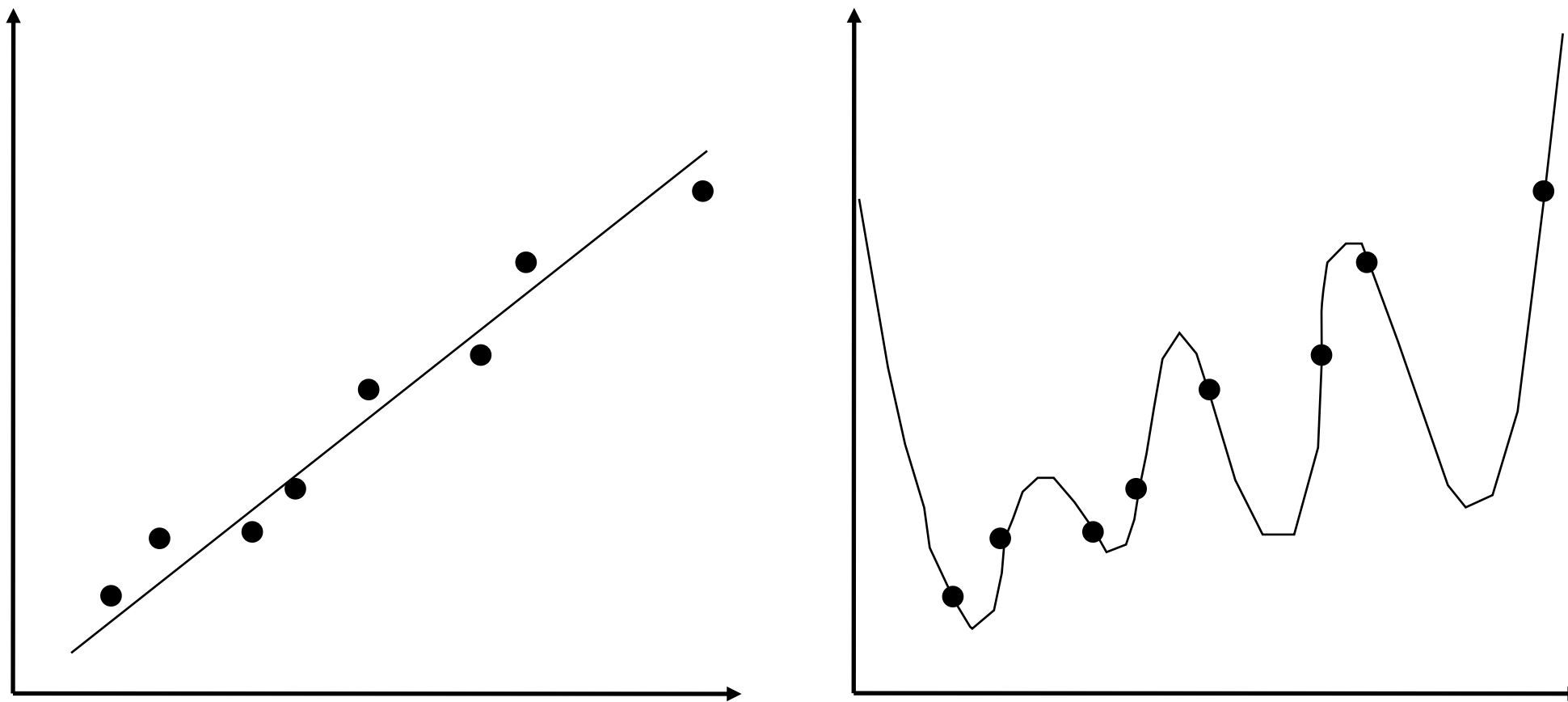
# Model Selection?



$$y = 0.95x - 0.26$$

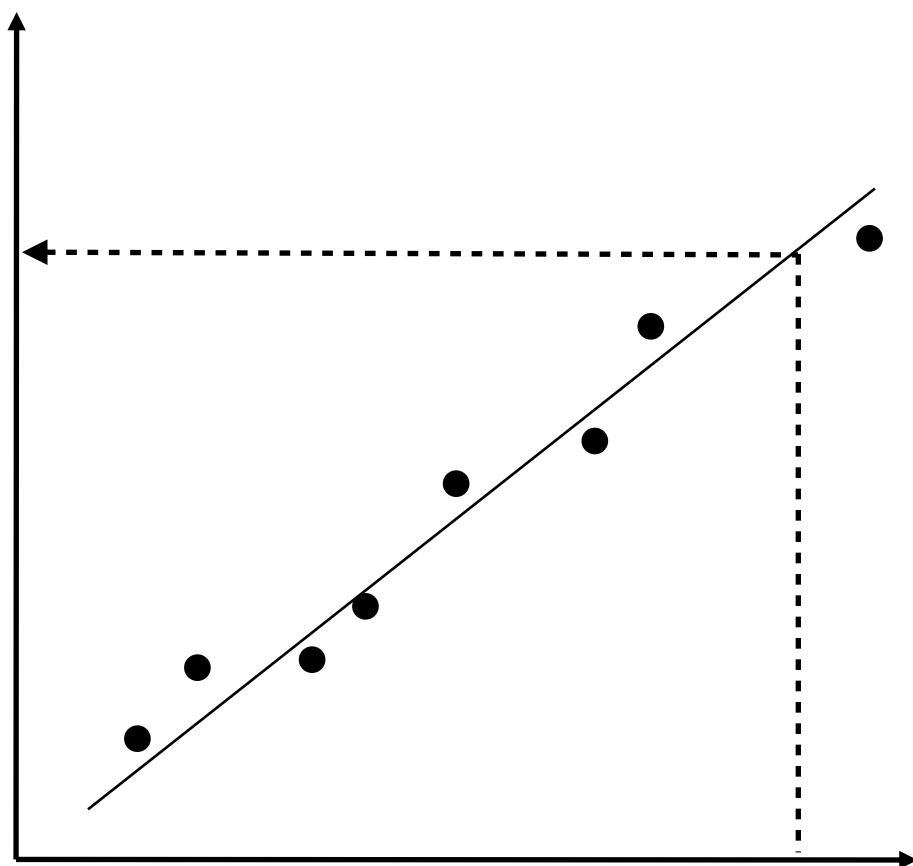
- Measure of fit between model and data (e.g., SSE, likelihood, etc.)
- How do we compare different *types* of models?

# Model Selection: How Do We Choose Between Different Types of Models?



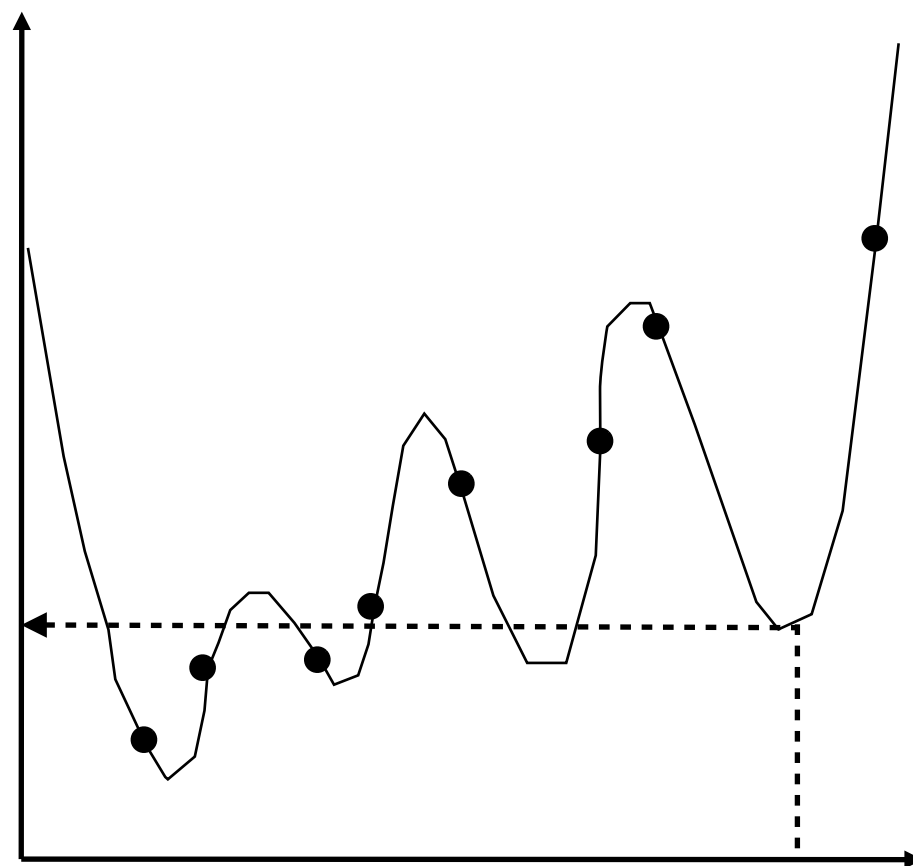
Select model with best fit?

# Over-fitting



$$y = bx + a$$

2-parameter model  
Good description, poor fit



$$y = gx^6 + fx^5 + ex^4 + dx^3 + cx^2 + bx + a$$

7-parameter model  
Poor description, good fit

For nested models, more parameters always result in a better fit to the data, but not necessarily in a better description

# Selecting the best model: the likelihood ratio test

---

- The fit of two alternative models can be compared using the ratio of their likelihoods:

$$LR = \frac{P(\text{Data} | M1)}{P(\text{Data} | M2)} = \frac{L_{M1}}{L_{M2}}$$

- Note that  $LR > 1$  if model 1 has the highest likelihood
- For nested models it can be shown that if the simplest (“null”) model is true, then

$$\Delta = \ln(LR^2) = 2 \ln(LR) = 2 (\ln L_{M1} - \ln L_{M2})$$

follows a  $\chi^2$  distribution with degrees of freedom equal to the number of extra parameters in the most complicated model.

This makes it possible to perform stringent statistical tests to determine which model (hypothesis) best describes the data

# Asking biological questions in a likelihood ratio testing framework

---

- Fit two alternative, nested models to the data.
- Record optimized likelihood and number of free parameters for each fitted model.
- Test if alternative (parameter-rich) model is significantly better than nullmodel (i.e., the simplest model), given number of additional parameters ( $n_{\text{extra}}$ ):
- Compute  $\Delta = 2 \times (\ln L_{\text{Alternative}} - \ln L_{\text{Null}})$
- Compare  $\Delta$  to  $\chi^2$  distribution with  $n_{\text{extra}}$  degrees of freedom
- Depending on models compared, different biological questions can be addressed (presence of molecular clock, presence of positive selection, difference in mutation rates among sites or branches, etc.)



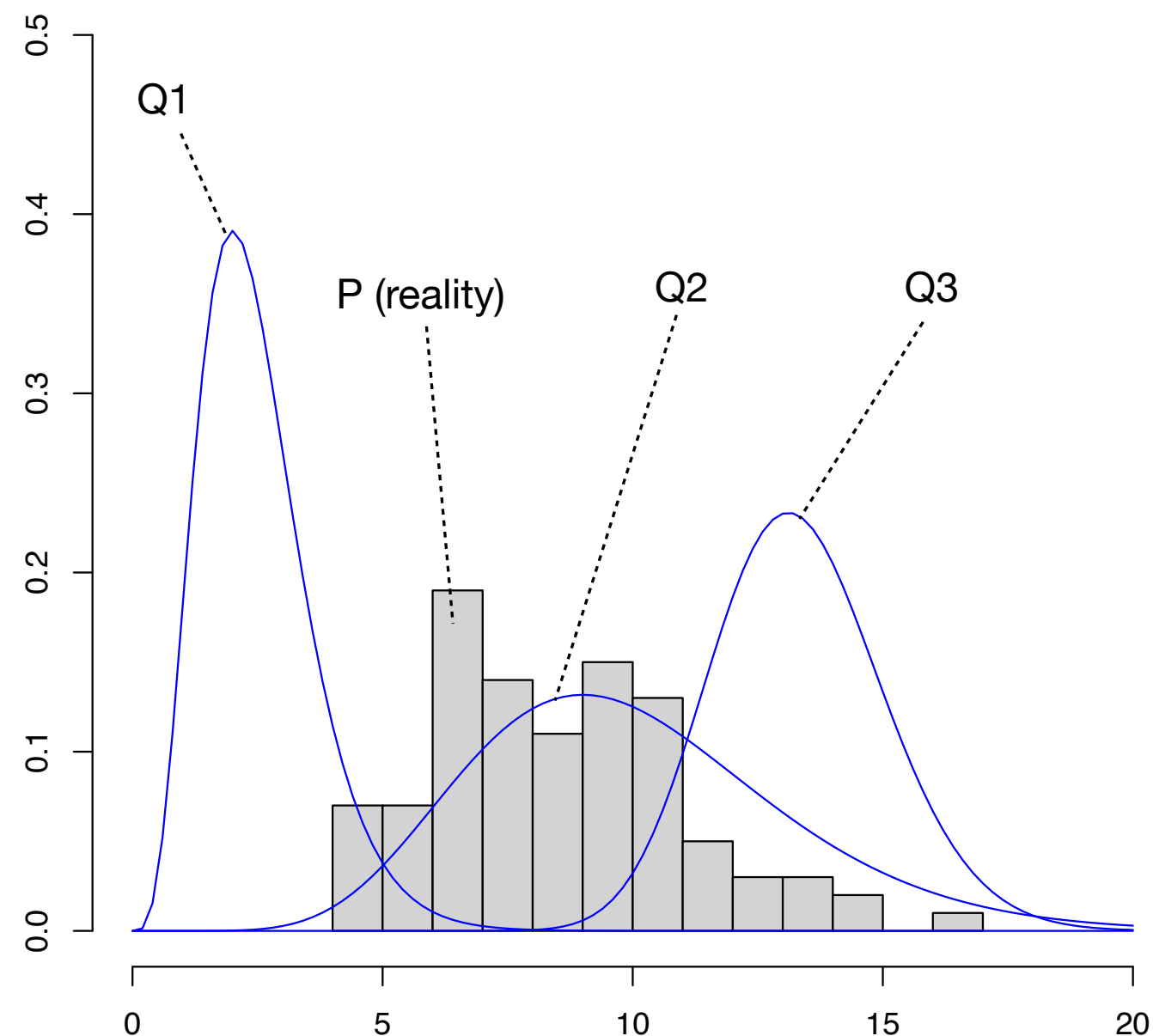
# AIC: Akaike Information Criterion

- A probabilistic model of a system defines a probability distribution over the possible outcomes (data sets)
- For instance: Probability of getting 0-10 heads when tossing a coin 10 times. Probability of getting a specific alignment for the JC model on a given phylogeny
- Kullback-Leibler (K-L) divergence is a measure of the distance between two probability distributions:

$$D(P||Q) = \sum_{i=1}^N p_i \log \left( \frac{p_i}{q_i} \right)$$

- We are interested in finding the model (probability distribution), that most closely approximates reality, i.e., the model Q that has the smallest K-L divergence from the “true probability distribution” P.
- AIC is essentially an estimate of the expected, relative Kullback-Leibler distance between the true model and the approximating model:

$$AIC = -2 \ln(L) + 2K$$



Kullback-Leibler divergence is a measure of the distance between probability distributions (models). Here, Q2 is the candidate model that has the smallest K-L distance from reality, and it is therefore the best approximation. AIC chooses the model with the smallest *expected* K-L distance

# Model Selection Using the Akaike Information Criterion (AIC)

- Fit a set of alternative models to the data.
- Record maximized log likelihood,  $\ln(L)$ , and number of free parameters,  $K$ , for each fitted model.
- For each model compute AIC according to this formula:

$$AIC = -2 \times \ln(L) + 2 \times K$$

- Models can now be ranked according to AIC: Lower AIC is better.

Model	$\ln(L)$	$K$	AIC
TVM+I+G	-3553.5002	9	7125.0004
GTR+I+G	-3553.1787	10	7126.3574
TVM+G	-3555.3269	8	7126.6538
GTR+G	-3555.0110	9	7128.0220
K81uf+I+G	-3560.2527	7	7134.5054
TIM+I+G	-3559.5247	8	7135.0494
K81uf+G	-3562.0266	6	7136.0532

# Model Selection Using the AIC: computation of model probabilities

- From the relative AIC values it is furthermore possible to compute so-called Akaike weights:

$$\Delta AIC_i = AIC_i - \min AIC$$

$$w_i = \frac{\exp(-\frac{1}{2}\Delta AIC_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta AIC_i)}$$

- Akaike weight can be interpreted as the conditional probability that a model is the K-L best one, given the data and the initial set of models.

Model	ln(L)	K	AIC	weight (w)
TVM+I+G	-3553.5002	9	7125.0004	0.45709258
GTR+I+G	-3553.1787	10	7126.3574	0.23191849
TVM+G	-3555.3269	8	7126.6538	0.19997372
GTR+G	-3555.0110	9	7128.0220	0.10089556
K81uf+I+G	-3560.2527	7	7134.5054	0.00394475
TIM+I+G	-3559.5247	8	7135.0494	0.00300533
K81uf+G	-3562.0266	6	7136.0532	0.00181936

# Probabilities as extended logic

---

- Polya, Cox, Jeffreys, Jaynes: probabilities are the only consistent basis for plausible reasoning (reasoning when there is insufficient information for deductive reasoning).
- Probabilities should form basis of all scientific inference
- Difference between probability interpretations:
  - “**Frequentist**”: probability is long-run frequency of event in repeatable experiment
  - “**Bayesian**”: probability is way of quantifying uncertainty
- Attaching probabilities to models allow us to perform multimodel inference and model averaging

# Model selection as a general strategy for answering scientific questions

---

- Construct comprehensive set of plausible alternative hypotheses for how the system under investigation works (but not too many)
- Phrase the hypotheses as mathematical models
- Assess evidence for all hypotheses by computing model probabilities
- Make conclusions, predictions, etc based on model probabilities
- Very different from null hypothesis testing approach (where you assess fit of single, implausible model that you don't believe is true...)

# Multimodel Inference:

## Basing Conclusions on More Than Just the Best Model

---

- Prediction: More robust predictions can be made by taking a weighted average of the predictions made by all models. (Weight = model probability)
- Model-averaging: More reliable estimates of parameter values can be obtained by taking a weighted average over the sub-set of fitted models that contain the parameter.
- Relative importance of parameters: the importance of a parameter can be estimated by summing the probabilities of those models that contain it.

# AIC example: Which model fits best: JC or K2P?

	A	C	G	T
A	-	$\alpha$	$\alpha$	$\alpha$
C	$\alpha$	-	$\alpha$	$\alpha$
G	$\alpha$	$\alpha$	-	$\alpha$
T	$\alpha$	$\alpha$	$\alpha$	-

## Jukes and Cantor model (JC):

All nucleotides have same frequency

All substitutions have same rate

$K = 1$  parameter

	A	C	G	T
A	-	$\beta$	$\alpha$	$\beta$
C	$\beta$	-	$\beta$	$\alpha$
G	$\alpha$	$\beta$	-	$\beta$
T	$\beta$	$\alpha$	$\beta$	-

## Kimura 2 parameter model (K2P):

All nucleotides have same frequency

Transitions and transversions have different rate

$K = 2$  parameters

Note: in principle each branch length in the tree also has an associated free parameter, but we ignore these here since they cancel out (the tree is the same in the two cases)

Note 2: depending on how you phrase the problem, JC and K2P can be said to have  $K=0$  and  $K=1$

# AIC example: Which model fits best: JC or K2P?

---

Starting point: set of DNA sequences, fit JC and K2P models to data, record likelihoods

JC:  $\ln L = -2034.3$ ,  $K = 1$

K2P:  $\ln L = -2026.2$ ,  $K = 2$

Assess evidence by computing model probabilities:

(1) Compute  $AIC = -2 \ln L + 2K$ :

JC:  $AIC = -2 \times -2034.3 + 2 \times 1 = 4070.6$

K2P:  $AIC = -2 \times -2026.2 + 2 \times 2 = 4056.4 \leq$  Best model (smallest AIC)

(2) Compute  $\Delta AIC_i = AIC_i - \min AIC$

JC:  $4070.6 - 4056.4 = 14.2$

K2P:  $4056.4 - 4056.4 = 0$

---



# AIC example: Which model fits best: JC or K2P?

---

(3) Compute model probabilities:  $w_i = \frac{\exp(-\frac{1}{2}\Delta\text{AIC}_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta\text{AIC}_i)}$

JC: numerator =  $\exp(-0.5 \times 14.2) = 0.000825$

K2P: numerator =  $\exp(-0.5 \times 0) = 1$

Sum (denominator) =  $1 + 0.000825 = 1.000825$

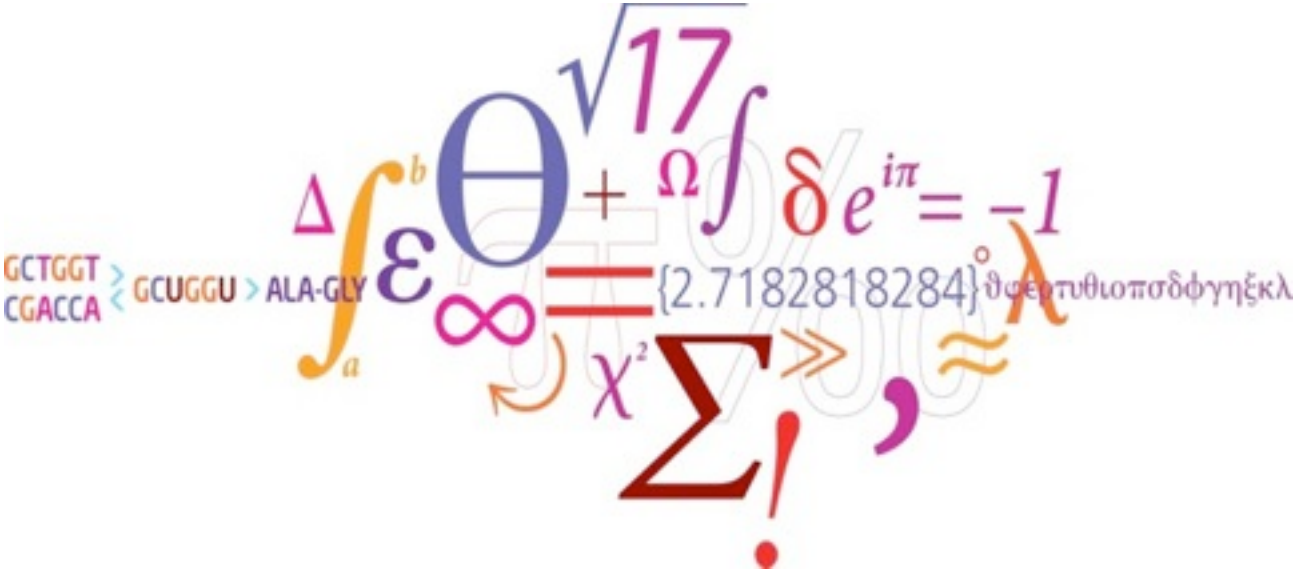
=>

$P(\text{JC}) = 0.000825 / 1.000825 = 0.0008$  (0.08 %)

$P(\text{K2P}) = 1 / 1.000825 = 0.9992$  (99.92 %) <= Strongly supported (about 1250 x stronger)

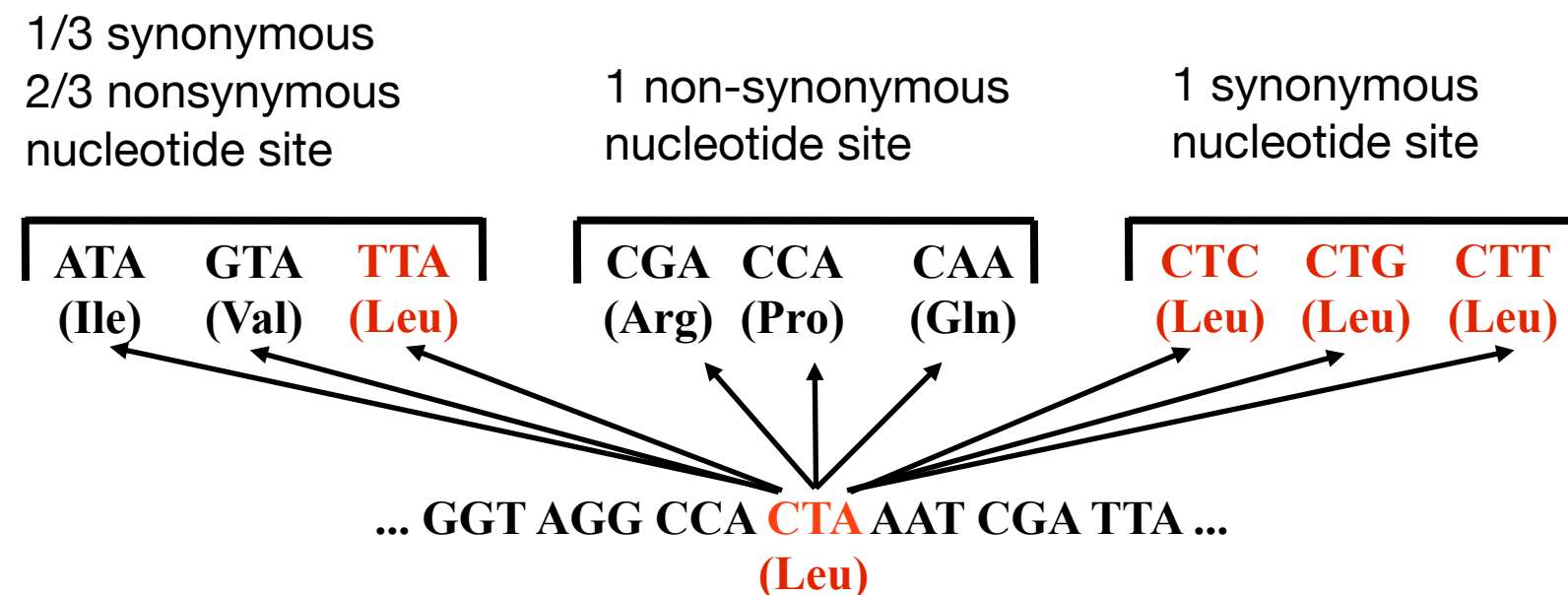
# Exercise: Detection of Selection

---



# Positive selection I: synonymous and non-synonymous mutations

- 20 amino acids, 61 codons
- Most amino acids encoded by more than one codon
  - Not all mutations lead to a change of the encoded amino acid
  - "Synonymous mutations" are rarely selected against








# Positive selection II: non-synonymous and synonymous **mutation rates** contain information about selective pressure

---

- **dN**: rate of non-synonymous mutations per non-synonymous site
  - **dS**: rate of synonymous mutations per synonymous site
  - Recall: Evolution is a two-step process:
    - (1) Mutation (random)
    - (2) Selection (non-random)
  - Randomly occurring mutations will lead to  $dN/dS=1$ .
  - Deviations from this most likely caused by subsequent selection.
  - **$dN/dS < 1$** : Higher rate of synonymous mutations: **negative (purifying) selection**
  - **$dN/dS > 1$** : Higher rate of non-synonymous mutations: **positive selection**
-

# Today's exercise: positive selection in HIV?

- Fit two alternative models to HIV data:
  - M1: two classes of codons with different dN/dS ratios:  
 $dN/dS < 1$    $dN/dS = 1$  
  - M2: three distinct classes with different dN/dS ratios:  
 $dN/dS < 1$    $dN/dS = 1$    $dN/dS > 1$  
- Compute model probabilities to assess the evidence for M2 versus M1
- If M2 much better than M1 then you have statistical evidence for positive selection.
- Most likely reason: immune escape (i.e., sites must be in epitopes)



 : Codons showing  $dN/dS > 1$ : likely epitopes