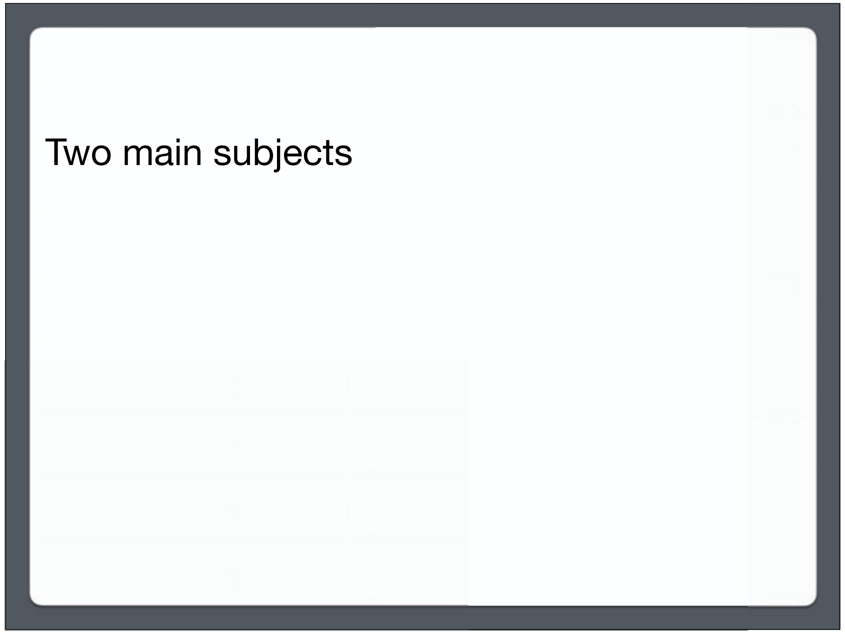
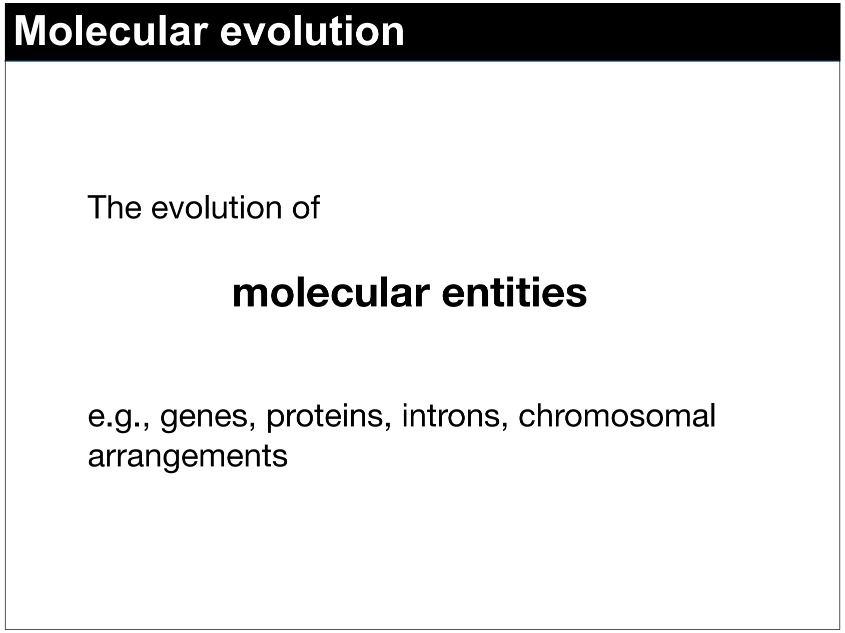
WHY

Molecular evolution matters?

Information, graphics and texts mostly after:
Dan Graur, Department of Zoology, Tel Aviv University, Israel
Itai Yanai, Molecular Genetics, Weizmann Institute of Science, Israel
Rose Hoberman, Carnegie Mellon University, USA



Two main subjects



Molecular evolution

The evolution of

molecular entities

e.g., genes, proteins, introns, chromosomal arrangements

Molecular evolution

The evolution of

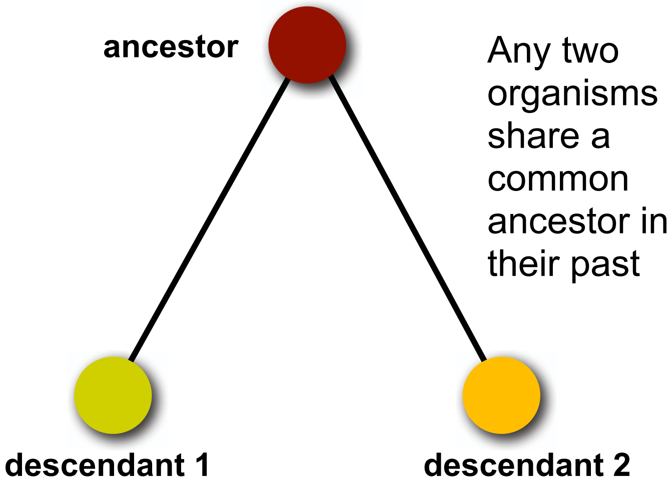
organisms and biological complexes

e.g., species, higher taxa, coevolutionary systems, ecological niches, and migratory patterns, by using molecular data

Molecular evolution

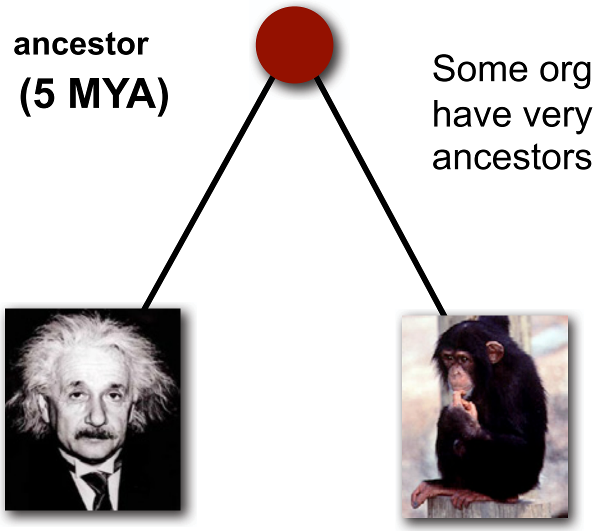
Assumption: Life is monophyletic

Molecular evolution

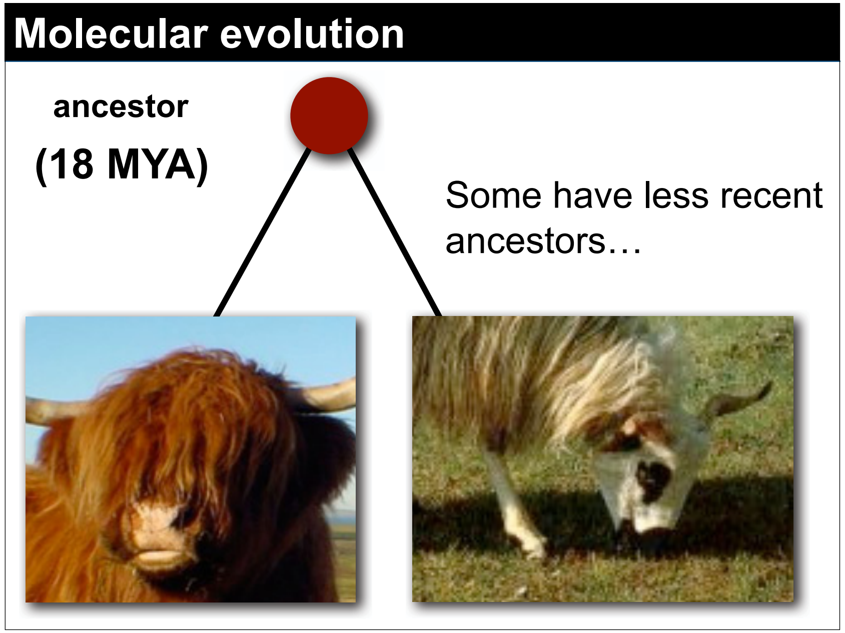


Any two organisms share a common ancestor in their past

Molecular evolution



Some organisms have very recent ancestors.

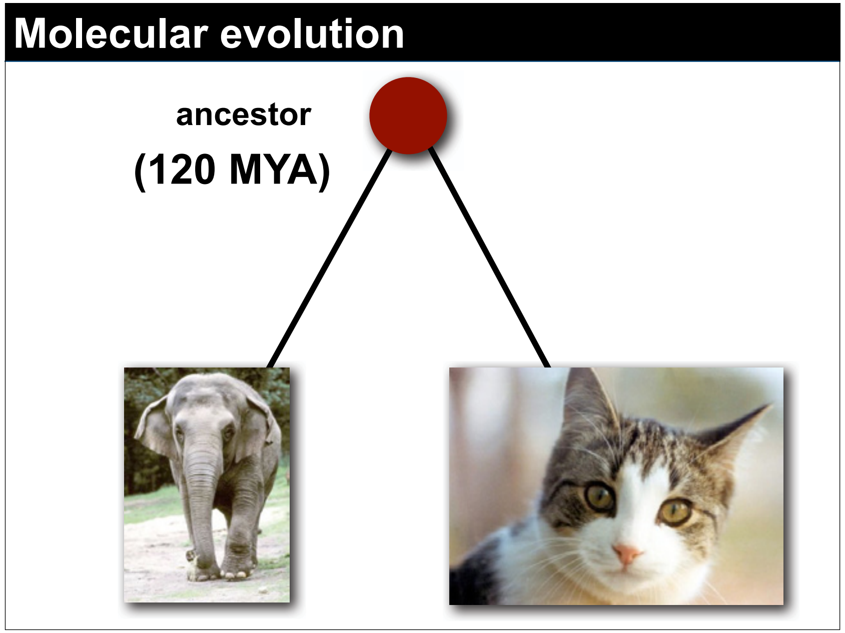


Molecular evolution

ancestor

(18 MYA)

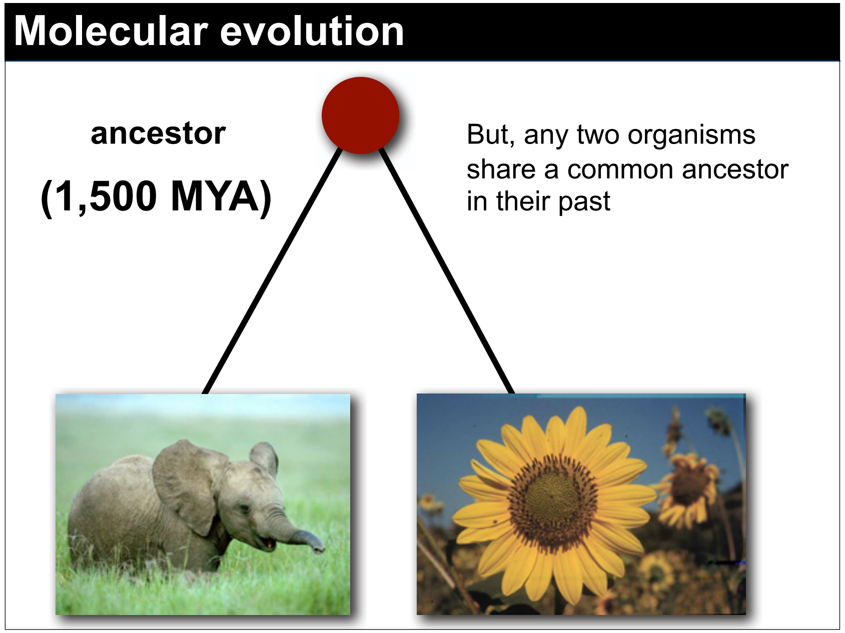
Some have less recent ancestors…



Molecular evolution

ancestor

(120 MYA)

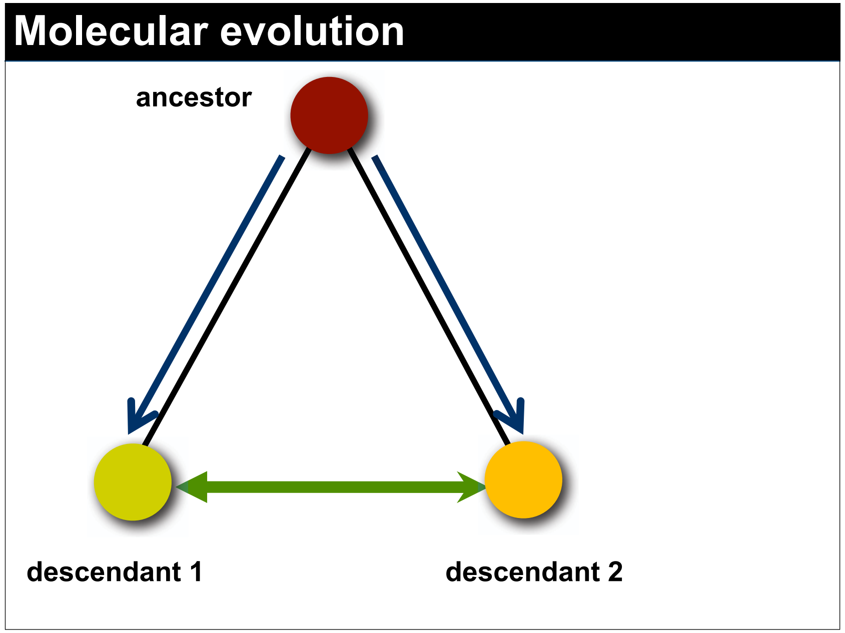


Molecular evolution

ancestor

(1,500 MYA)

But, any two organisms share a common ancestor in their past



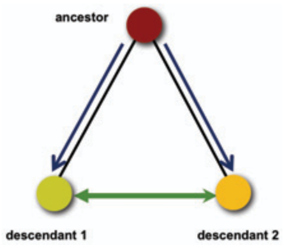
Molecular evolution

ancestor

descendant 1

descendant 2

Molecular evolution

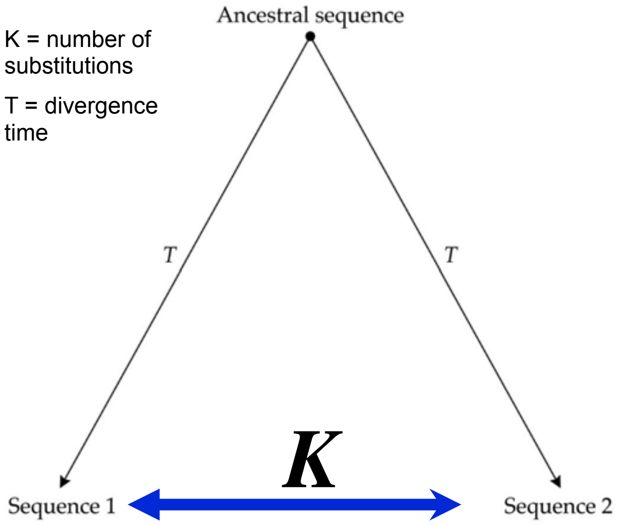


The differences between 1 and 2 are the result of **changes** on the lineage leading to descendant 1 **+** those on the lineage leading to descendant 2.

Molecular evolution

K = number of substitutions

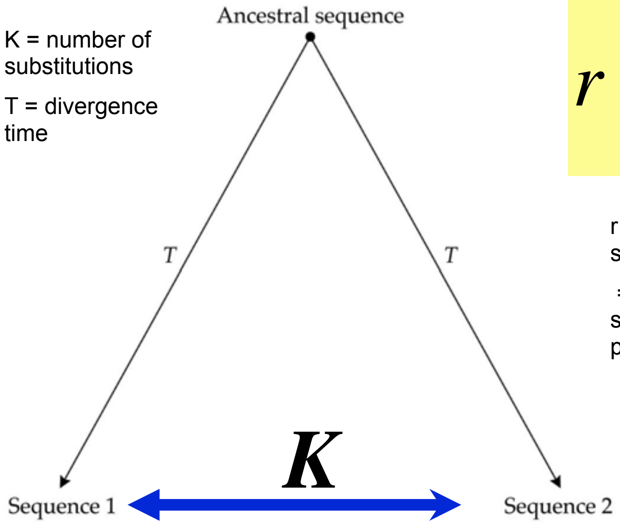
T = divergence time



Molecular evolution

K = number of substitutions

T = divergence time



$$r =$$

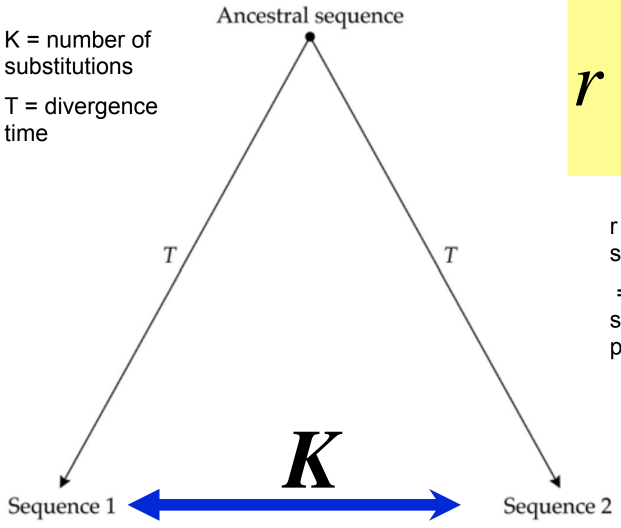
r = Nucleotide substitution rate

= número de substituições por posição por ano

Molecular evolution

K = number of substitutions

T = divergence time

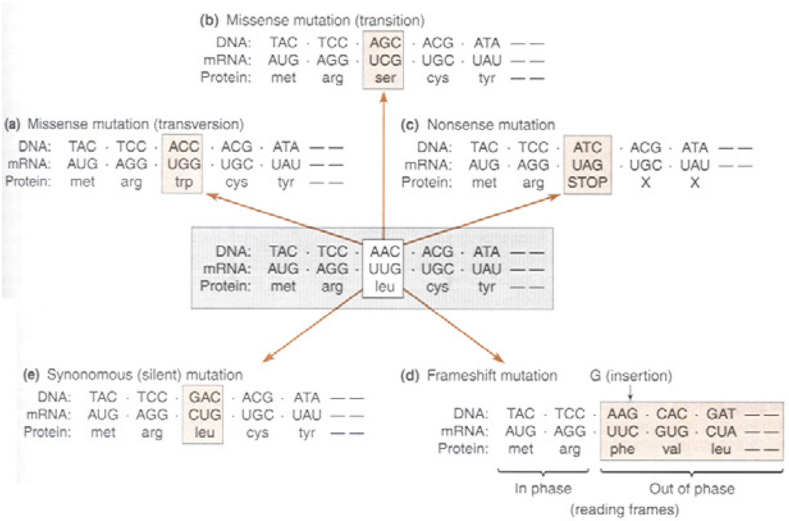


$$r = \frac{K}{2T}$$

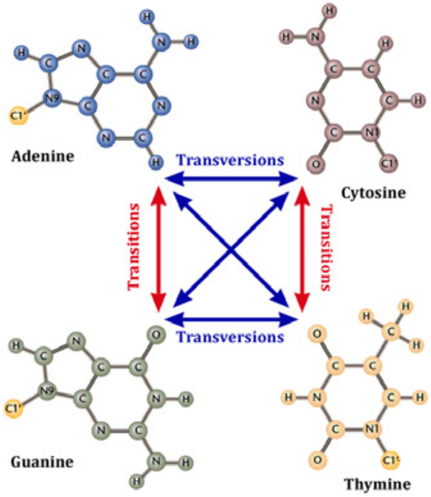
r = Nucleotide substitution rate

= número de substituições por posição por ano

Molecular evolution

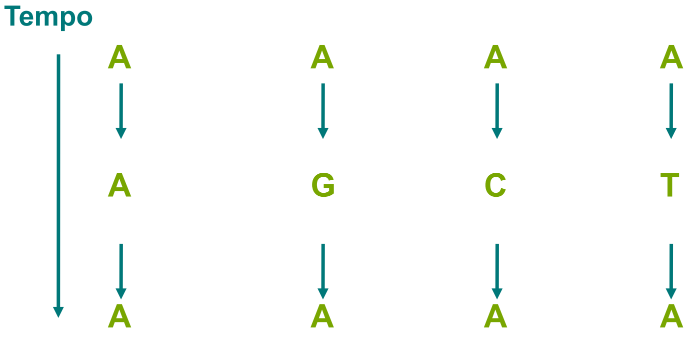

(b) Missense mutation (transition)
DNA: TAC · TCC · AGC · ACG · ATA — —
mRNA: AUG · AGG · UCG · UGC · UAU — —
Protein: met arg ser cys tyr — —
(a) Missense mutation (transversion)
DNA: TAC · TCC · ACC · ACG · ATA — —
mRNA: AUG · AGG · UGG · UGC · UAU — —
Protein: met arg trp cys tyr — —
(c) Nonsense mutation
DNA: TAC · TCC · ATC · ACG · ATA — —
mRNA: AUG · AGG · UAG · UGC · UAU — —
Protein: met arg STOP X X
DNA: TAC · TCC · AAC · ACG · ATA — —
mRNA: AUG · AGG · UUG · UGC · UAU — —
Protein: met arg leu cys tyr — —
(e) Synonomous (silent) mutation
DNA: TAC · TCC · GAC · ACG · ATA — —
mRNA: AUG · AGG · CUG · UGC · UAU — —
Protein: met arg leu cys tyr — —
(d) Frameshift mutation
G (insertion)
DNA: TAC · TCC · AAG · CAC · GAT — —
mRNA: AUG · AGG · UUC · GUG · CUA — —
Protein: met arg phe val leu — —
In phase
Out of phase
(reading frames)

Molecular evolution



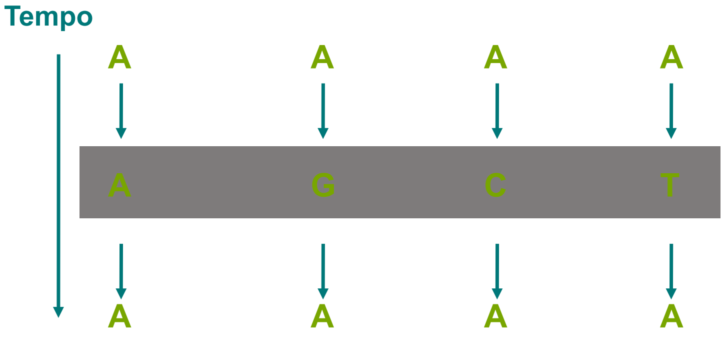
Molecular evolution

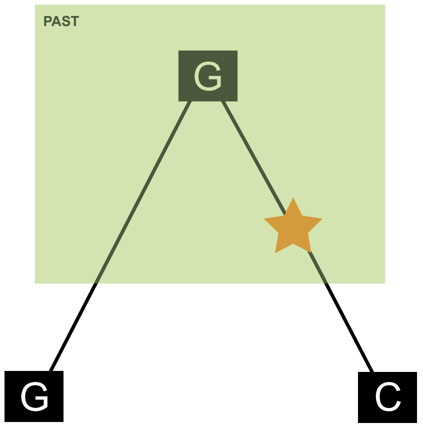
Homoplasias



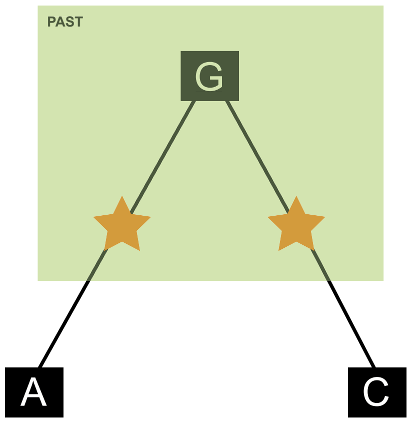
Molecular evolution

Homoplasias

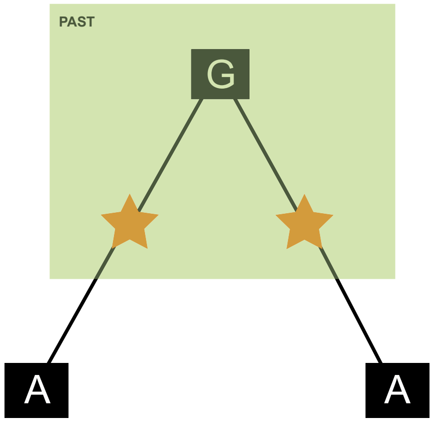




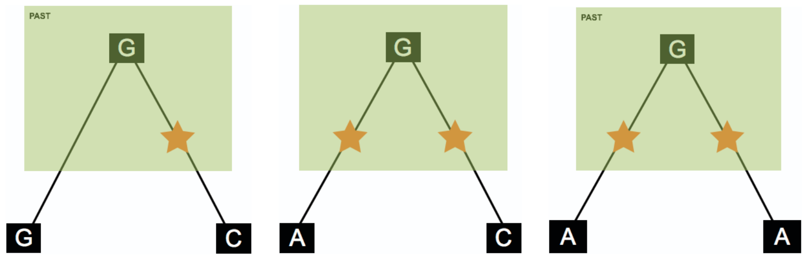
One substitutions happened - one substitution is visible



Two substitutions happened - only one substitution is visible



Two substitutions happened - no visible substitution



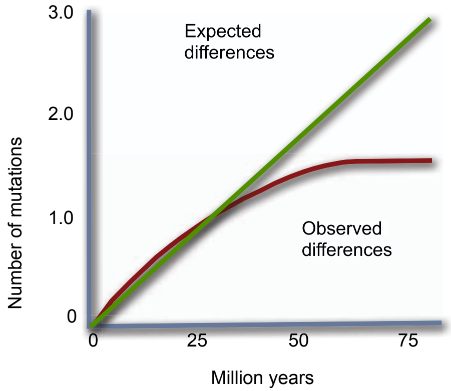
One substitutions happened
one substitution is visible

Two substitutions happened
no visible substitution

Two substitutions happened
only one substitution is visible

● Molecular Evolution ●

Estimating Genetic Differences



If all nucleotides are equally likely, the observed difference would plateau at 0.75

Therefore, simply counting differences underestimates distances, because it

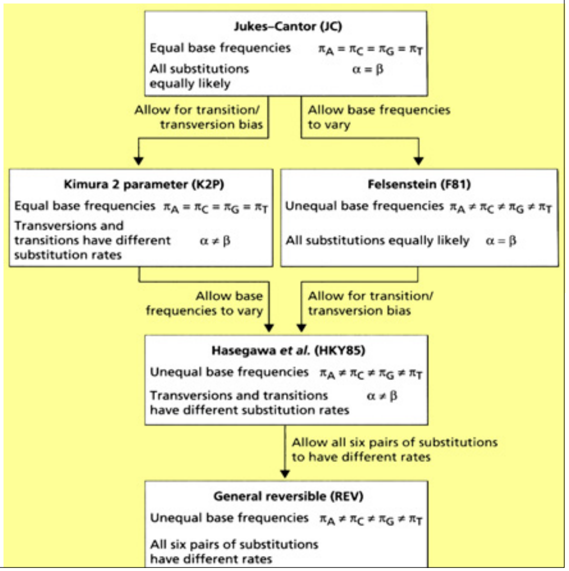
fails to count for multiple hits

Molecular evolution

Models

Molecular evolution

Models of evolution



Jukes–Cantor (JC)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
All substitutions equally likely $\alpha = \beta$

Allow for transition/transversion bias

Allow base frequencies to vary

Kimura 2 parameter (K2P)
Equal base frequencies $\pi_A = \pi_C = \pi_G = \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$

Felsenstein (F81)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All substitutions equally likely $\alpha = \beta$

Allow base frequencies to vary

Allow for transition/transversion bias

Hasegawa *et al.* (HKY85)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
Transversions and transitions have different substitution rates $\alpha \neq \beta$

Allow all six pairs of substitutions to have different rates

General reversible (REV)
Unequal base frequencies $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$
All six pairs of substitutions have different rates

Page RDM, Holmes EC (1998) Molecular Evolution: a phylogenetic approach Blackwell Science, Oxford.

Impact of models: 3 sequences

AGC
AAC
ACC

Sequences 1 and 2 differs at 1 out of 3 positions = 1/3
Sequences 1 and 3 differs at 1 out of 3 positions = 1/3
Sequences 2 and 3 differs at 1 out of 3 positions = 1/3

	1	2	3
1	-		
2	0.333	-	
3	0.333	0.333	-

http://artedi.ebc.uu.se/course/X3-2004/Phylogeny/Exercises/nj.html

JC69 model (Jukes-Cantor, 1969)

AGC
AAC
ACC

$$d = -\frac{3}{4}\ln\left[1-\frac{4P}{3}\right]$$

	1	2	3
1	-		
2	0.333	-	
3	0.333	0.333	-

Where P is the proportion of nucleotides that are different (the observed differences above) in the two sequences and ln is the natural log function. To calculate the JC distances from the observed differences above·

$$d = -\frac{3}{4}\ln\left[1-\frac{4(1/3)}{3}\right] = -\frac{3}{4}\ln\left[1-\frac{4}{9}\right] = -\frac{3}{4}\ln\left[\frac{5}{9}\right] \approx 0.441$$

	1	2	3
1	-		
2	0.441	-	
3	0.441	0.441	-

http://www.bioinf.manchester.ac.uk/resources/phase/manual

K80 model (Kimura, 1980) or Kimura 2P

AGC
AAC
ACC

Kimura's Two Parameter model (K2P) incorporates the observation that the rate of transitions per site (a) may differ from the rate of transversions (b), giving a total rate of substitiutions per site of (a + 2b)(there are three possible substitutions: one transition and two transversions).
The transition:transversion ratio a/b is often represented by the letter kappa (k).

In the K2P model the number of nucleotide substitutions per site is given by:

$$d = \frac{1}{2}\ln\left[\frac{1}{1-2P-Q}\right] + \frac{1}{4}\ln\left[\frac{1}{1-2Q}\right]$$

where:
P the proportional differences between the two sequences due to transitions
Q are the proportional differences between the two sequences due to transitions and transversions respectively.

K80 model (Kimura, 1980) or Kimura 2P

AGC
AAC

AGC
ACC

AAC
ACC

Sequences 1 and 2 differ one transition

$$d = \frac{1}{2}\ln\left[\frac{1}{1-2(1/3)-0}\right] + \frac{1}{4}\ln\left[\frac{1}{1-2\cdot 0}\right] = \frac{1}{2}\ln[3] + \frac{1}{4}\ln[1] = \frac{1}{2}\ln[3] \approx 0.549$$

Sequences 1 and 3 differ one transversion
Sequences 2 and 3 differ one transversion

$$d = \frac{1}{2}\ln\left[\frac{1}{1-2\cdot 0-(1/3)}\right] + \frac{1}{4}\ln\left[\frac{1}{1-2(1/3)}\right] = \frac{1}{2}\ln\left[\frac{3}{2}\right] + \frac{1}{4}\ln[3] \approx 0.477$$

	1	2	3
1	-		
2	0.549	-	
3	0.477	0.549	-

Note how the differences caused by the application of different models give different distances

Observed differences

	1	2	3
1	-		
2	0.333	-	
3	0.333	0.333	-

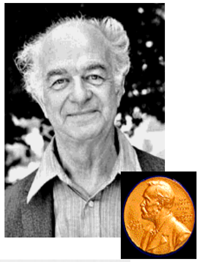
Jukes-Cantor model

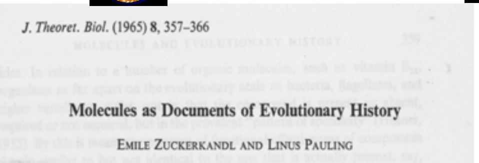
	1	2	3
1	-		
2	0.441	-	
3	0.441	0.441	-

Kimura 2P

	1	2	3
1	-		
2	0.549	-	
3	0.477	0.549	-

Molecular evolution





J. Theoret. Biol. (1965) 8, 357–366

Molecules as Documents of Evolutionary History

Emile Zuckerkandl and Linus Pauling

“the rate of molecular evolution is approximately constant over time in all lineages”

Molecular evolution

Gene sequences accumulate substitutions at a constant rate, therefore we can use genes sequences to time divergences.

This is referred to as a ‘Molecular Clock’

Molecular evolution

Molecular divergence is

ROUGHLY correlated

with divergence of time

Molecular evolution

The idea of a molecular clock was initially suggested by **Zuckerkandl and Pauling in 1962**.

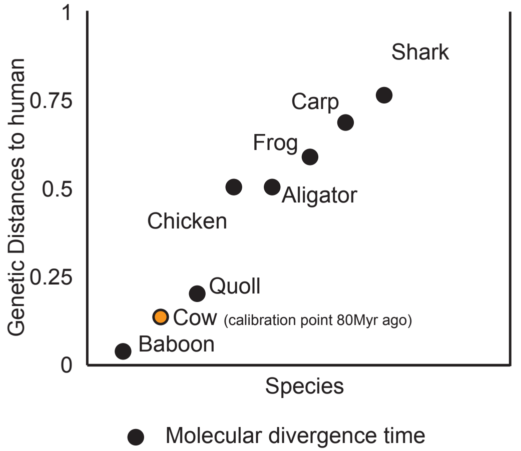
They noted that rates of amino acid replacements in animal haemoglobin were roughly proportional to real time, as judged against the fossil record.

Molecular evolution

The “constancy” of the molecular clock is particularly striking when compared to the obvious variation in the rates of morphological evolution (e.g. the existence of “living fossils”).

Molecular evolution

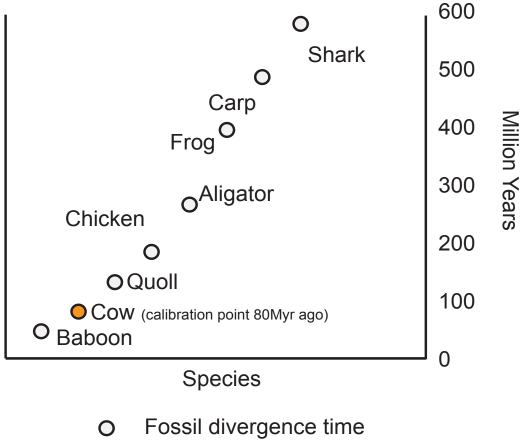
Evidence for rate constancy in haemoglobin



from Zuckerkandl and Pauling (1965)

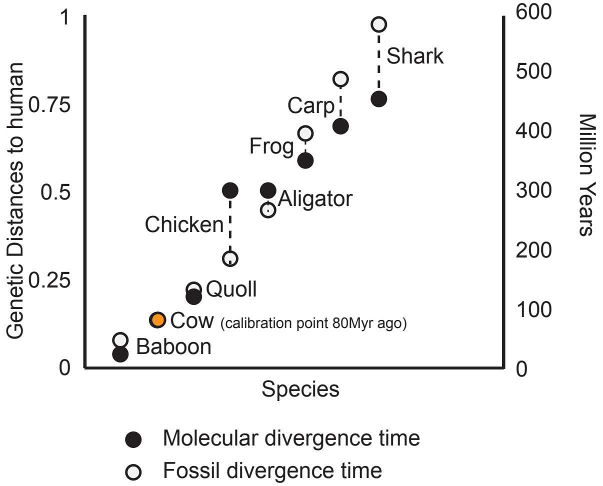
Molecular evolution

Evidence for rate constancy in haemoglobin



Molecular evolution

Evidence for rate constancy in haemoglobin



Molecular evolution

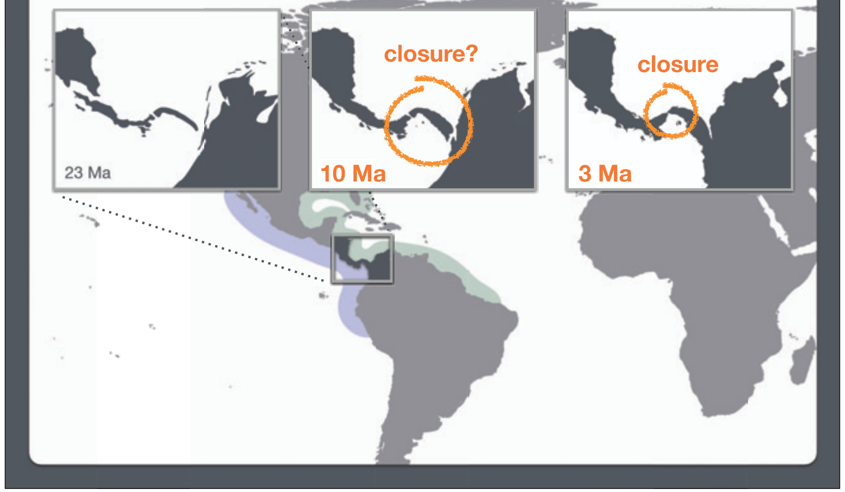
A Hipótese do Relógio Molecular

- A quantidade de diferenças genéticas entre sequências é função do tempo desde a separação.
- A taxa de mutação é (suficientemente) constante para estimar tempos de divergência

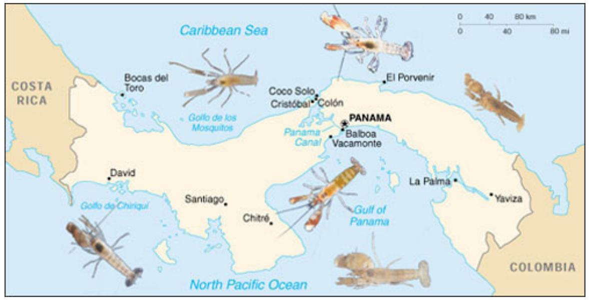
Molecular evolution

Calibrations

Isthmus of Panama



Molecular evolution



Molecular evolution

ATLANTIC

closure of Isthmus of Panama ≈ 3.1 MY

PACIFIC

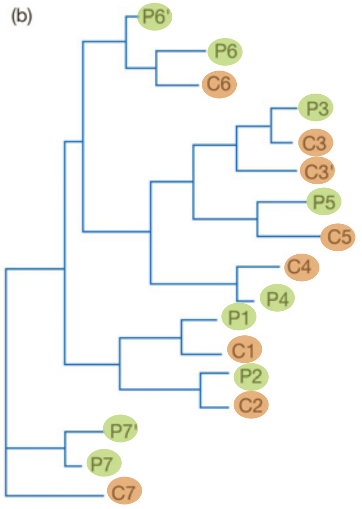
3.1% genetic distance

MUTATION RATE??

• Calibrating the molecular clock •

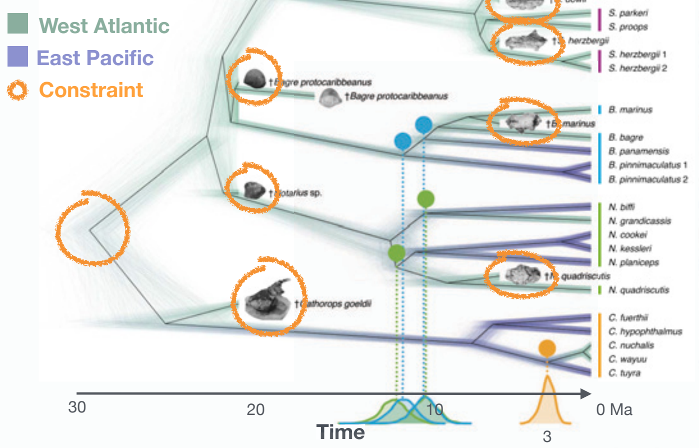
Phylogeny of **Pacific** (P) and **Caribbean** (C) species pairs of *Alpheus.*

In 6 out of 7 cases, the closest relative of a species was on the other side of the Isthmus

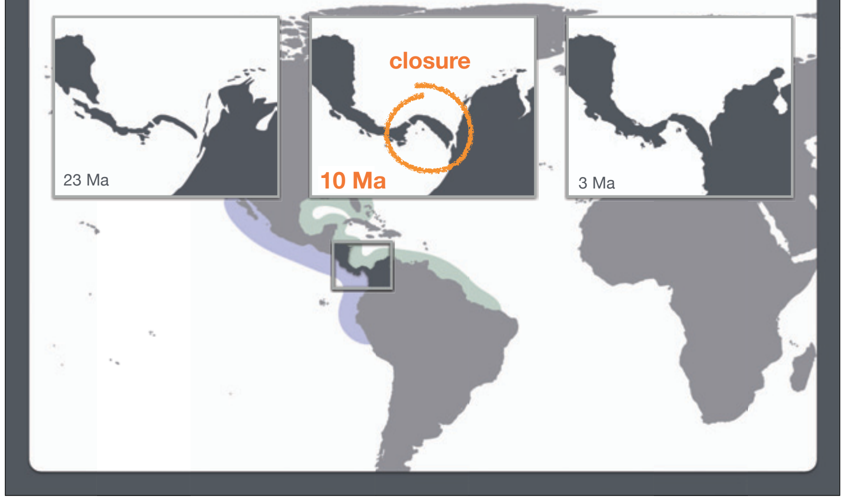


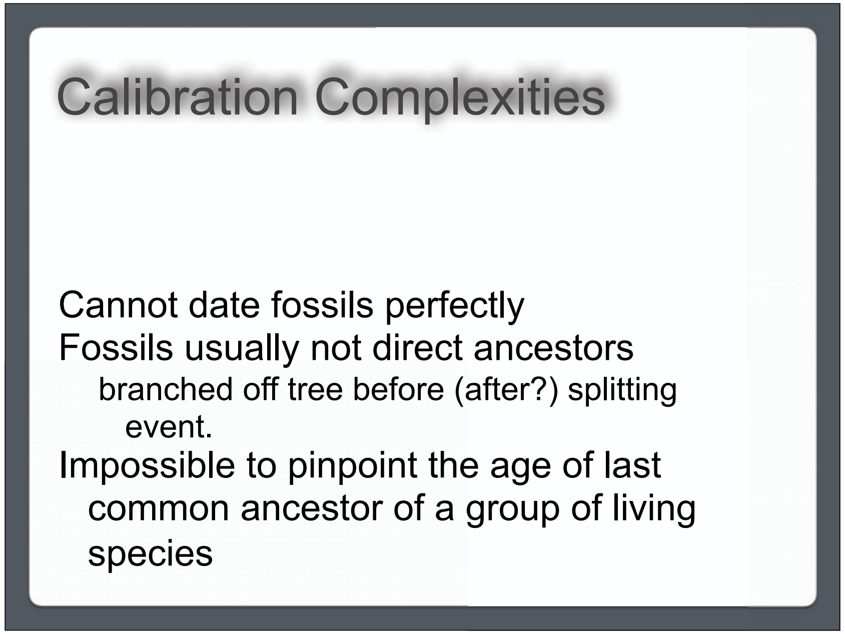
Knowlton, N., Weigt, L., Solorzano, L., Mills, D., & Bermingham, E. (1993). *Science*, *260* (5114), 1629.

Results



Isthmus of Panama





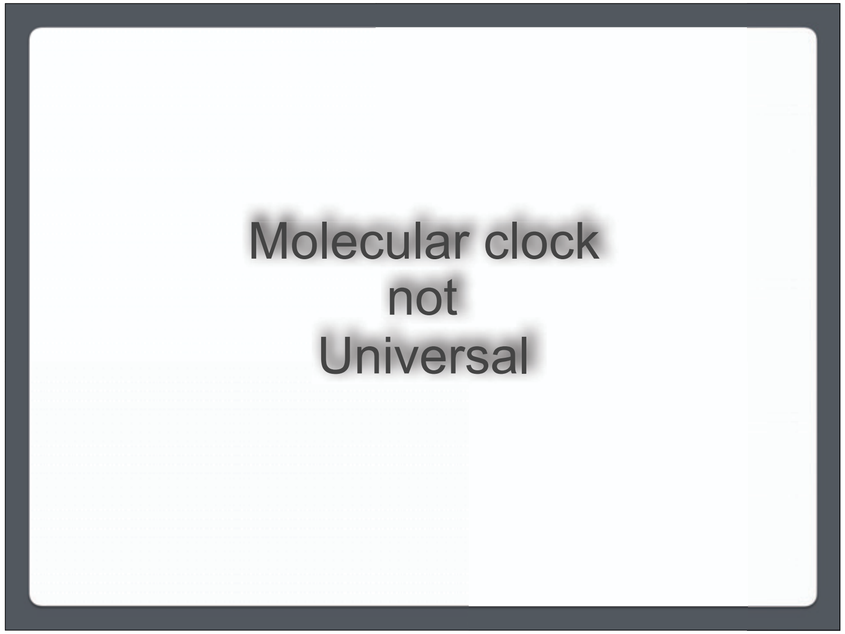
Calibration Complexities

Cannot date fossils perfectly

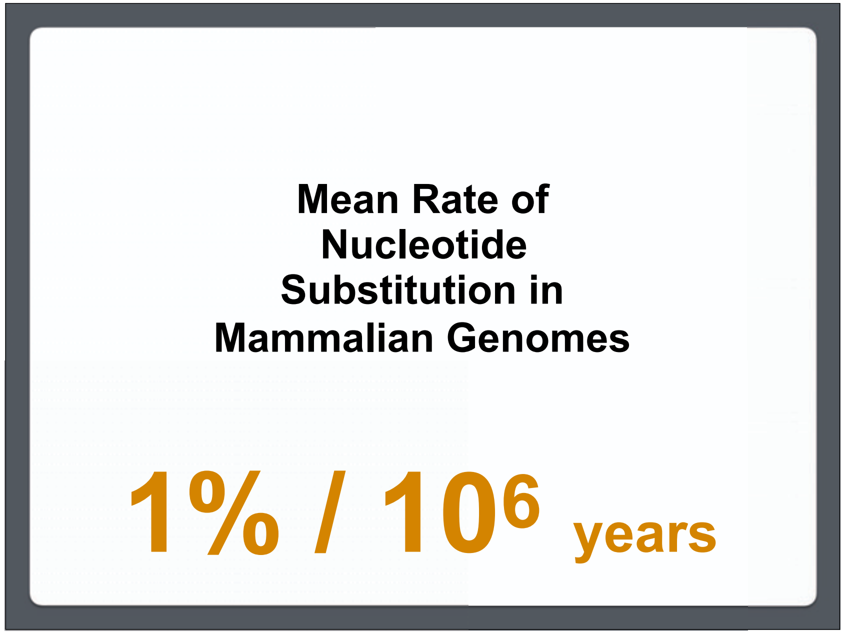
Fossils usually not direct ancestors

- branched off tree before (after?) splitting event.

Impossible to pinpoint the age of last common ancestor of a group of living species

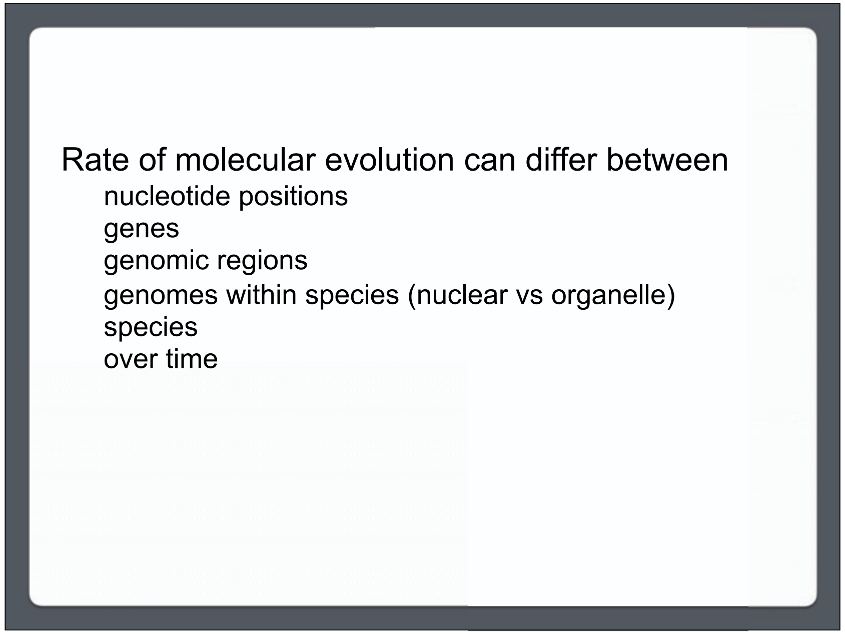


Molecular clock
not
Universal



Mean Rate of Nucleotide Substitution in Mammalian Genomes

1% / 10^6 years



Rate of molecular evolution can differ between

- nucleotide positions
- genes
- genomic regions
- genomes within species (nuclear vs organelle)
- species
- over time

Rate of molecular evolution can differ between

- nucleotide positions
- genes
- genomic regions
- genomes within species (nuclear vs organelle)
- species
- over time

If not considered, introduces bias into time estimates

Rate Heterogeneity among lineages

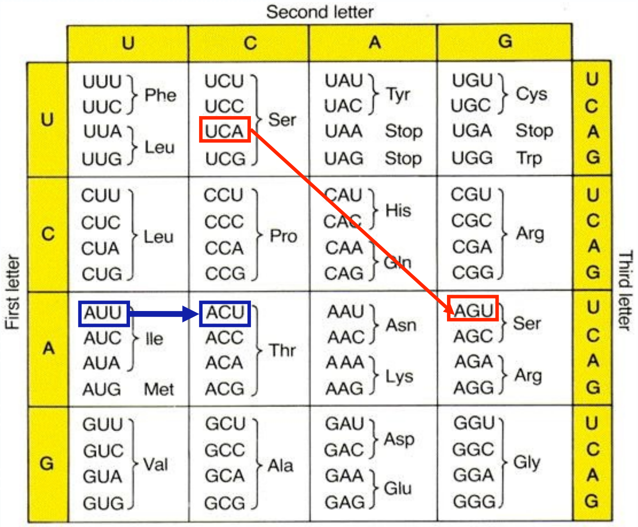
Cause	Reason
Repair mechanisms	e.g. RNA viruses have error-prone polymerases
Metabolic rate	More free radicals
Generation time	Copies DNA more frequently
Population size	Effects mutation fixation rate

How different regions of the genome may vary?

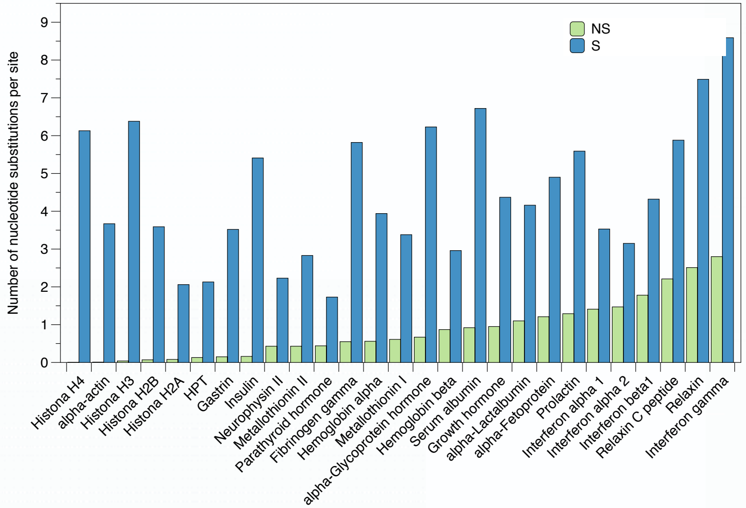
Evolution is a very <u>slow</u> process at the molecular level

Rates of Substitutions in Protein-Coding regions

Synonymous vs non-synonymous
Functional vs non-functional



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU UUC Phe UUA UUG Leu	UCU UCC UCA UCG Ser	UAU UAC Tyr UAA Stop UAG Stop	UGU UGC Cys UGA Stop UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU GAC Asp GAA GAG Glu	GGU GGC GGA GGG Gly	U C A G



Page & Holmes p240

Mean non-synonymous rate	0.84 ± 0.66 × 10^{-9}
Mean synonymous rate	4.44 ± 1.36 × 10^{-9}

substitutions per site per year

The rate of synonymous substitution is much larger than the **non-synonymous** rate.

Functional constraint
=
Degree of intolerance towards mutations

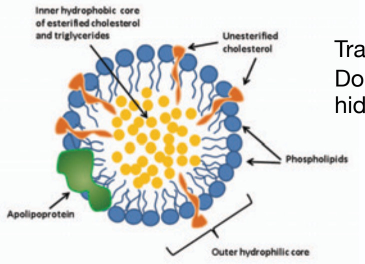
The functional constraint defines the range of alternative residues that are acceptable at a site without affecting negatively the function or structure of the gene or the gene product.

Two different examples:

Apolipoproteins

Histones 3

Apolipoproteins

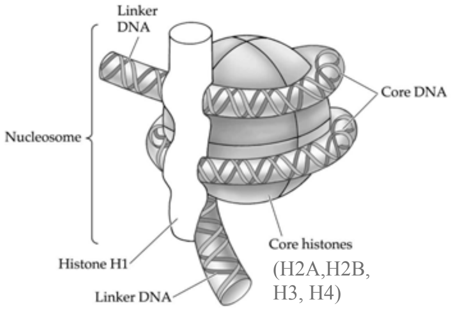


Transportadores de lípidos no sangue.
Domínios constituídos por resíduos hidrofóbicos

Alterações entre aminoácidos hidrofóbicos (valina – leucina) permitidas em muitas posições.

Histones

As histonas interagem directamente com outras histonas ou com o DNA para a formação do nucleossoma.

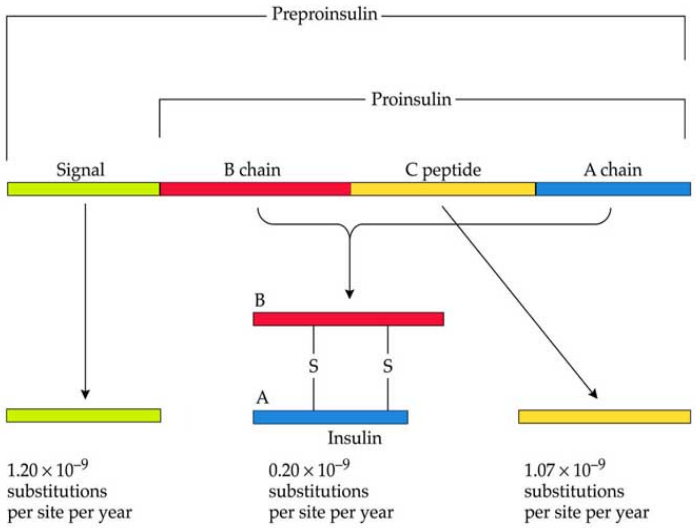


Manutenção da compactação e alcalinidade necessárias = poucas substituições.

Histonas mutam 1000 vezes mais lentamente do que as apolipoproteínas.

Apolipoproteins	Função compatível com substituições entre aminoácidos hidrofóbicos
Histones	Manutenção da compactação e alcalinidade necessárias = poucas substituições.

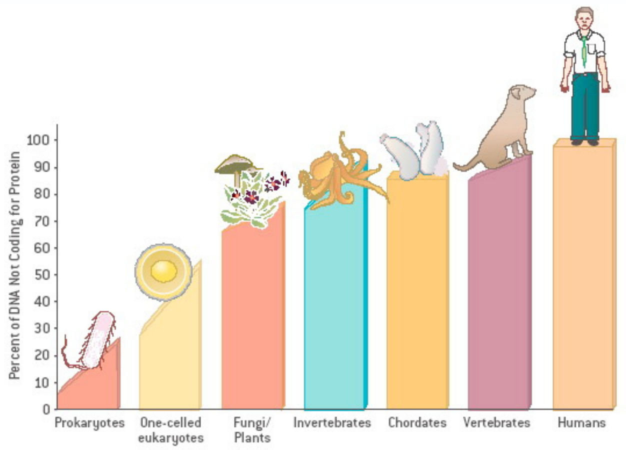
Histonas mutam 1000 vezes mais lentamente do que as apolipoproteínas.


Preproinsulin
Proinsulin
Signal
B chain
C peptide
A chain
B
S
S
A
Insulin
1.20 × 10⁻⁹ substitutions per site per year
0.20 × 10⁻⁹ substitutions per site per year
1.07 × 10⁻⁹ substitutions per site per year

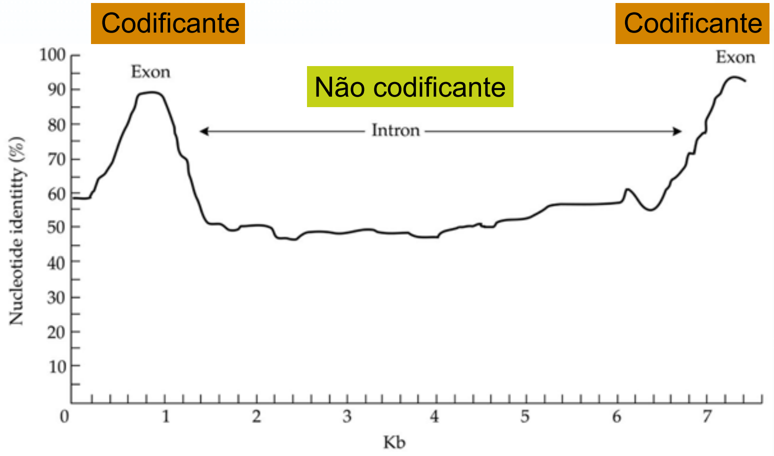
Functional regions evolve slower than nonfunctional regions.

Important proteins evolve slower than unimportant ones.

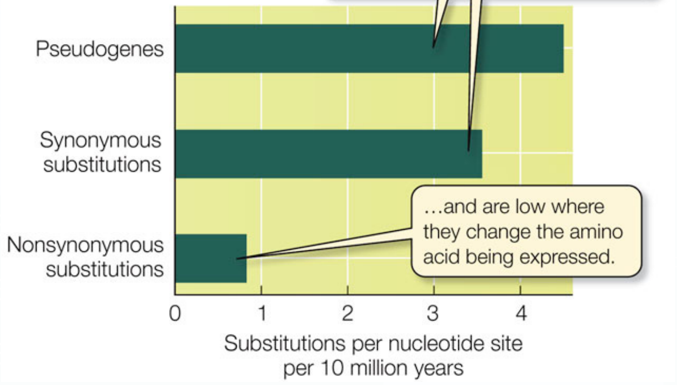
Rates of Substitutions in Non-Coding regions



NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.



Perfil de semelhança de duas sequências de DNA.



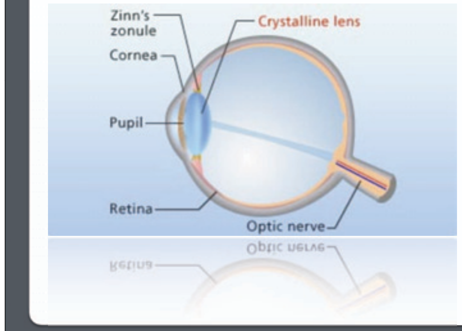
Rates of substitution are high where they do not affect functioning…

…and are low where they change the amino acid being expressed.

Coding regions evolve slower than noncoding regions.

Spalax ehrenberghi

FACT: *S. ehrenberghi* aA-crystallin lost its functional role



Water-soluble structural protein found in the lens and the cornea of the eye accounting for the transparency of the structure

The main function of crystallins at least in the lens of the eye is probably to increase the refractive index while not obstructing light.

WHEN:

more than 25 MA ago

(when the mole rat became subterranean and presumably gradually lost use of its eyes)

FACT:

The aA-crystallin of *S. ehrenberghi* **evolves 20 times faster than the aA-crystallins in other rodents, such as rats, mice, hamsters, gerbils and squirrel.**

FACT:
The aA-crystallin of *S. ehrenberghi* **possess all the prerequisites for normal function and expression, including the proper signals for alternative splicing.**

S. ehrenberghi **aA-crystallin lost its functional role but it could function...**

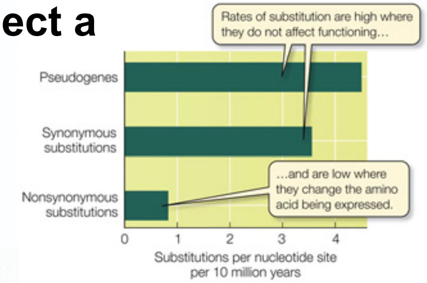
The functional role was lost a long time ago: over 25 MA

The gene evolves 20 times faster than the aA-crystallins in other rodents

We would expect a

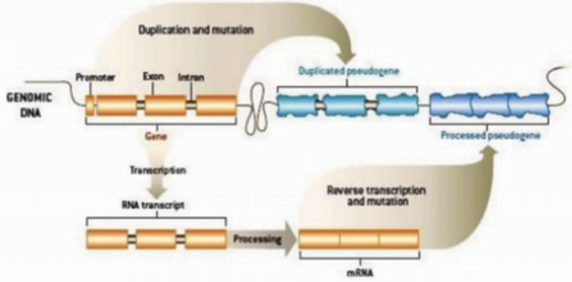
larger
equal
lower

mutational rate than the one from pseudogenes?



Pseudogenes

They are dysfunctional relatives of known genes in the genome that never become proteins



Notwithstanding…

The aA-crystallin of *S. ehrenberghi* evolves <u>slower</u> than pseudogenes.

Several explanations...

The genes are functional for the vision?

Was the loss of vision more recent (than 25 MY)?

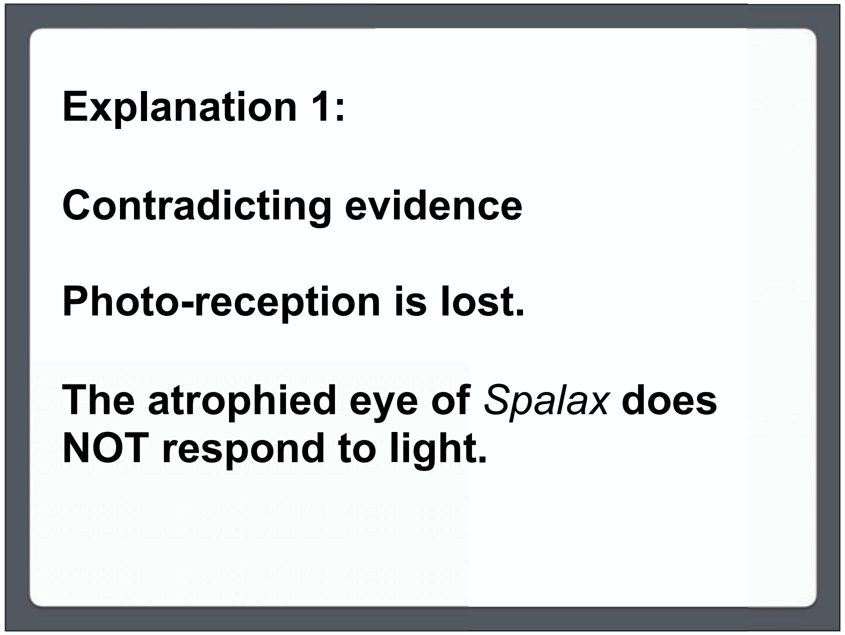
The gene has another function?

Explanation 1:

Are the genes functional?

Maybe not all function is lost …

(e.g.photoperiod perception)

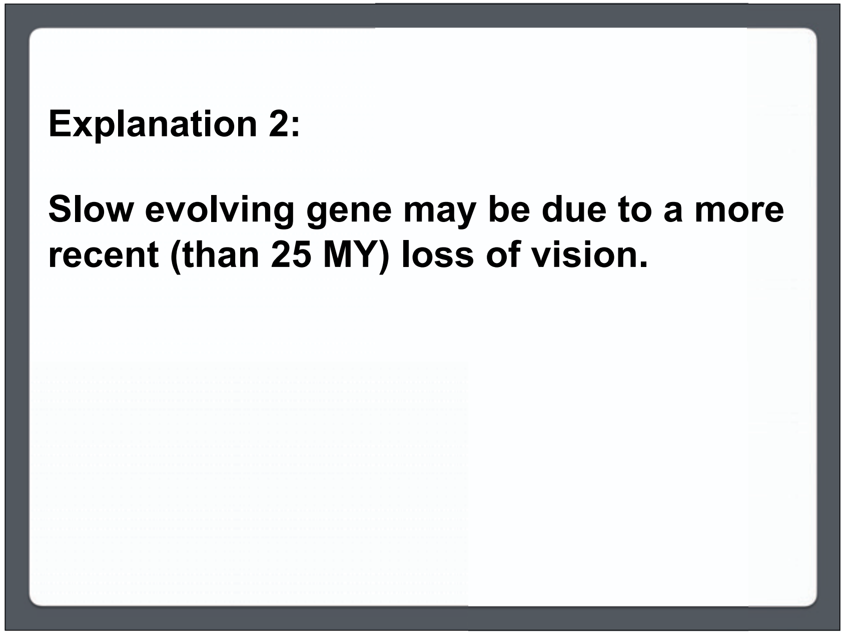


Explanation 1:

Contradicting evidence

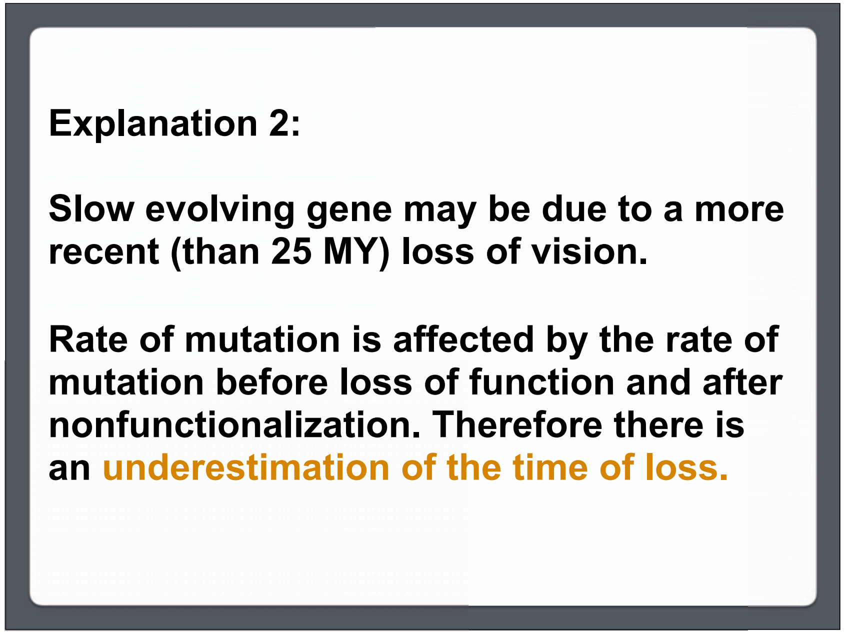
Photo-reception is lost.

The atrophied eye of *Spalax* does NOT respond to light.



Explanation 2:

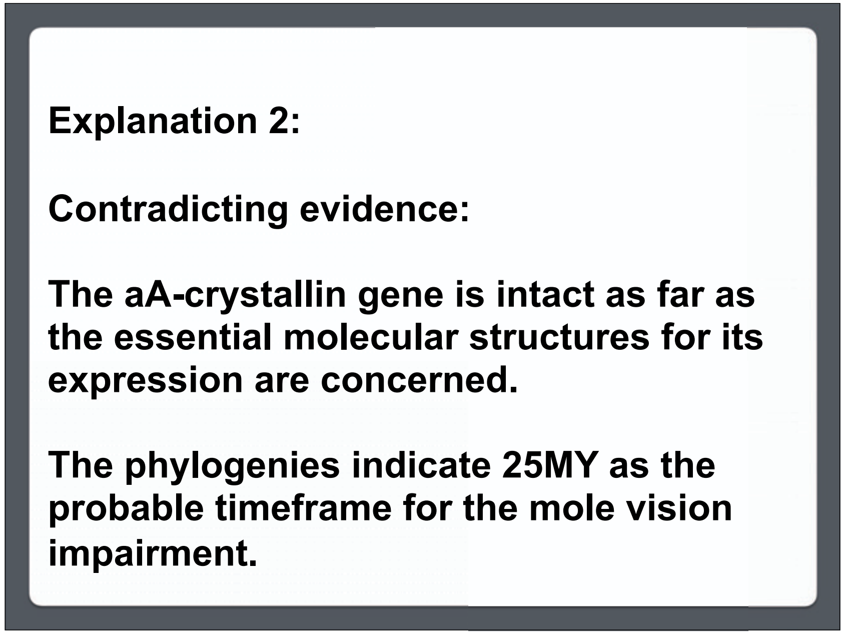
Slow evolving gene may be due to a more recent (than 25 MY) loss of vision.



Explanation 2:

Slow evolving gene may be due to a more recent (than 25 MY) loss of vision.

Rate of mutation is affected by the rate of mutation before loss of function and after nonfunctionalization. Therefore there is an underestimation of the time of loss.

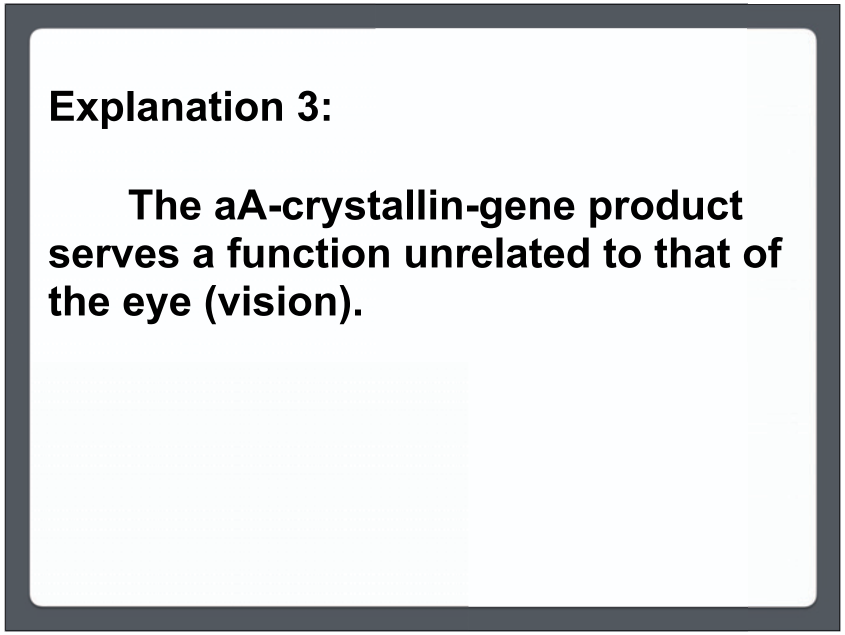


Explanation 2:

Contradicting evidence:

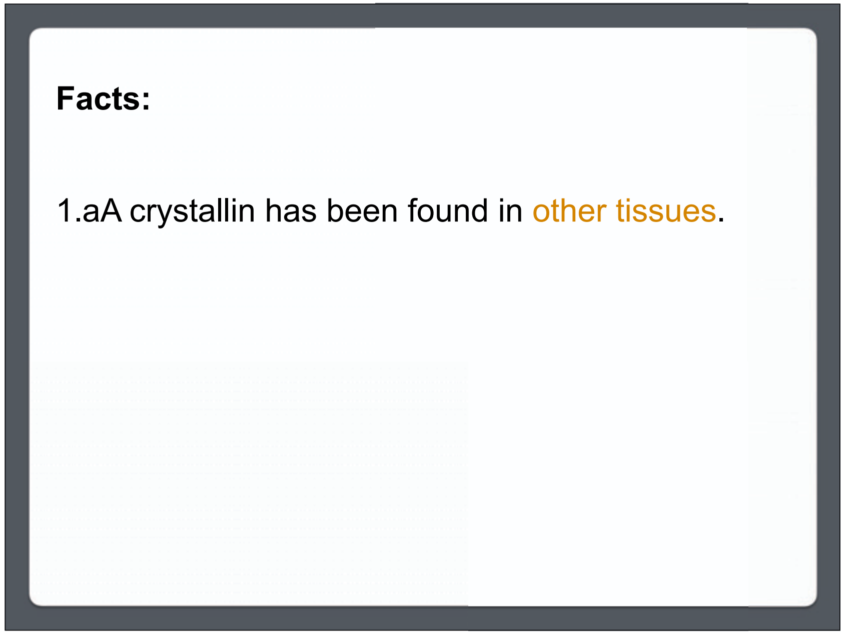
The aA-crystallin gene is intact as far as the essential molecular structures for its expression are concerned.

The phylogenies indicate 25MY as the probable timeframe for the mole vision impairment.



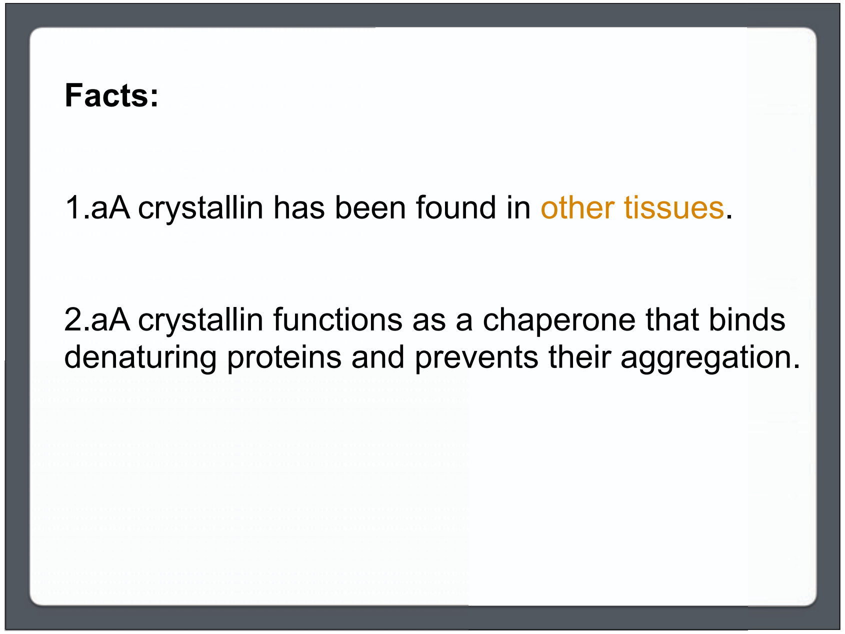
Explanation 3:

The aA-crystallin-gene product serves a function unrelated to that of the eye (vision).



Facts:

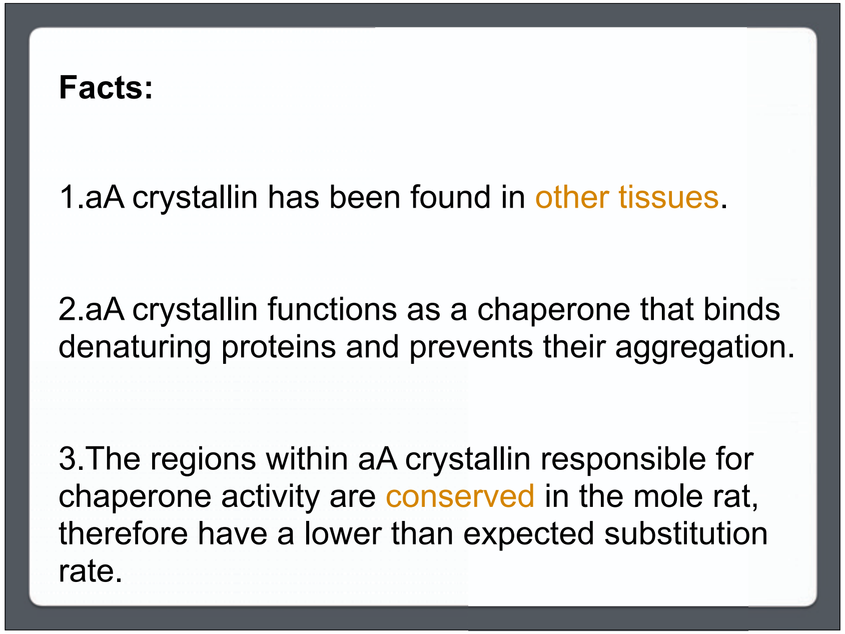
1.aA crystallin has been found in other tissues.



Facts:

1.aA crystallin has been found in other tissues.

2.aA crystallin functions as a chaperone that binds denaturing proteins and prevents their aggregation.



Facts:

1.aA crystallin has been found in other tissues.

2.aA crystallin functions as a chaperone that binds denaturing proteins and prevents their aggregation.

3.The regions within aA crystallin responsible for chaperone activity are conserved in the mole rat, therefore have a lower than expected substitution rate.