Evolution and sequence similarity

Bartek Wilczyński

March 3^{rd} , 2020

Tree of life?

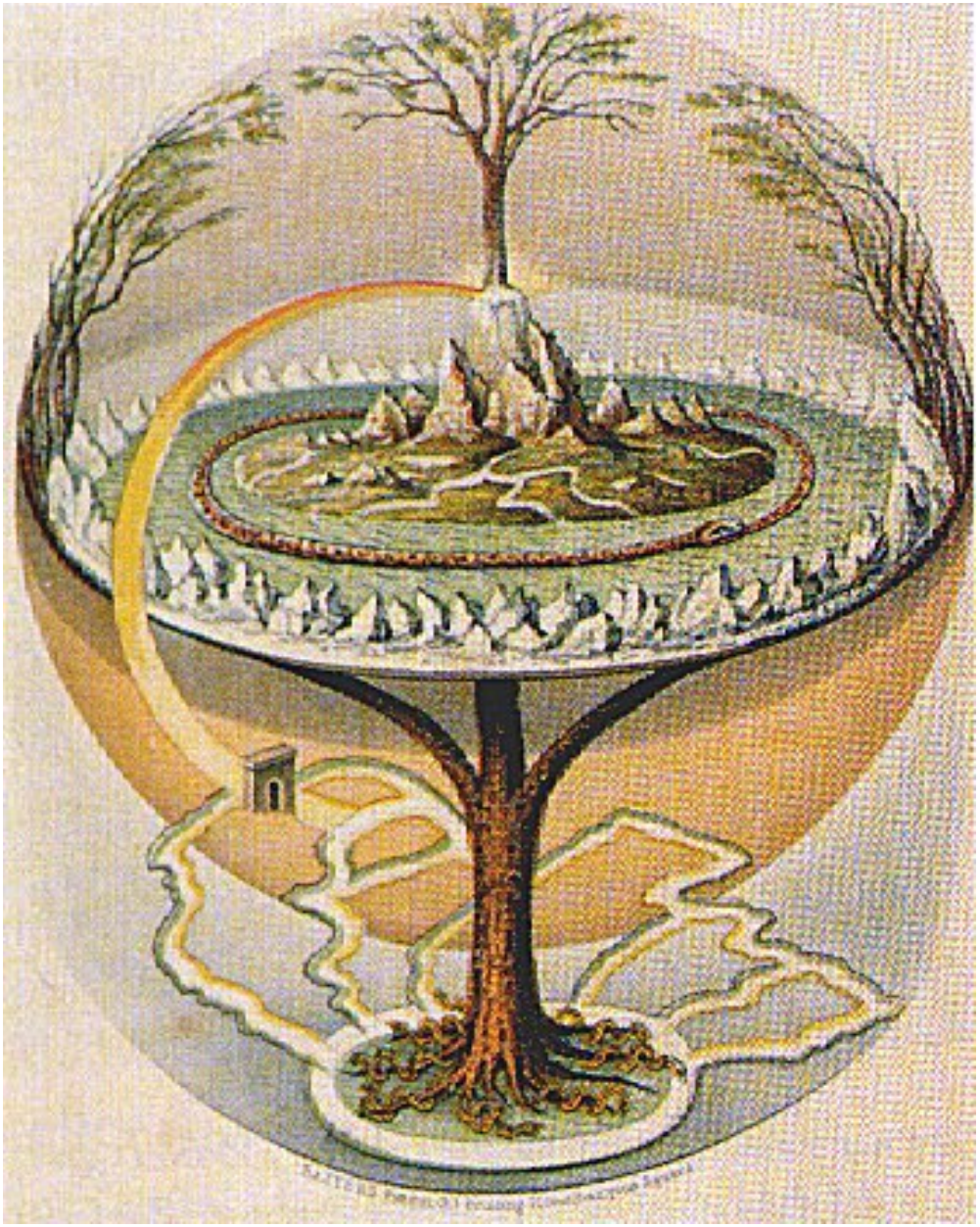


image (c) Wikimedia

Molecular tree of life

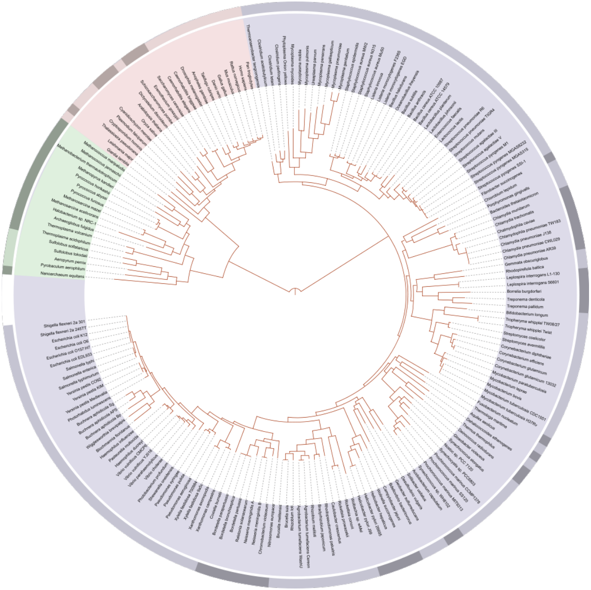
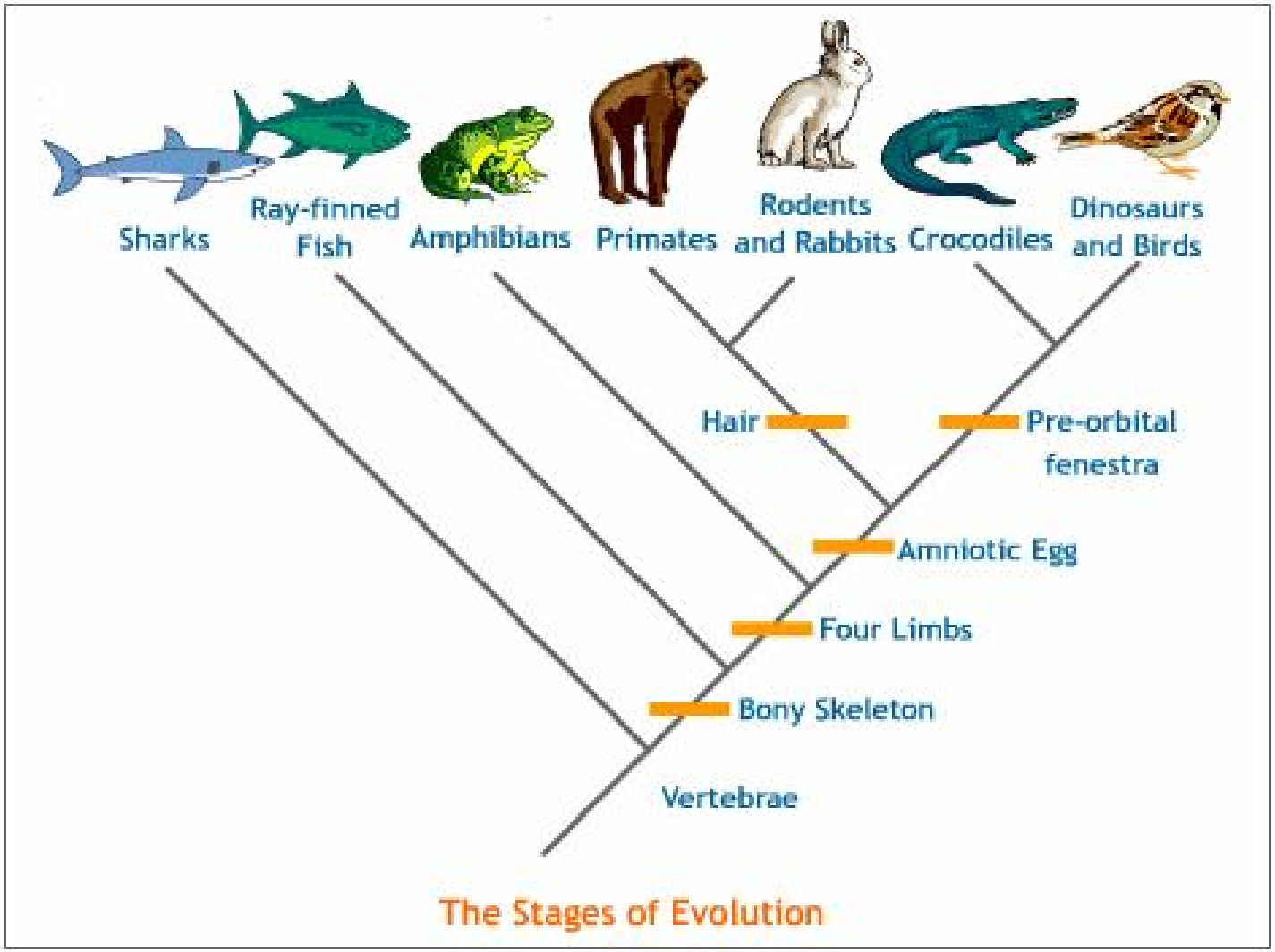


image (c) I. Letunic – itol.org

Stages of evolution



We are not the "most" evolved species

image (c) wistatutor.com

DNA replication enables inheritance

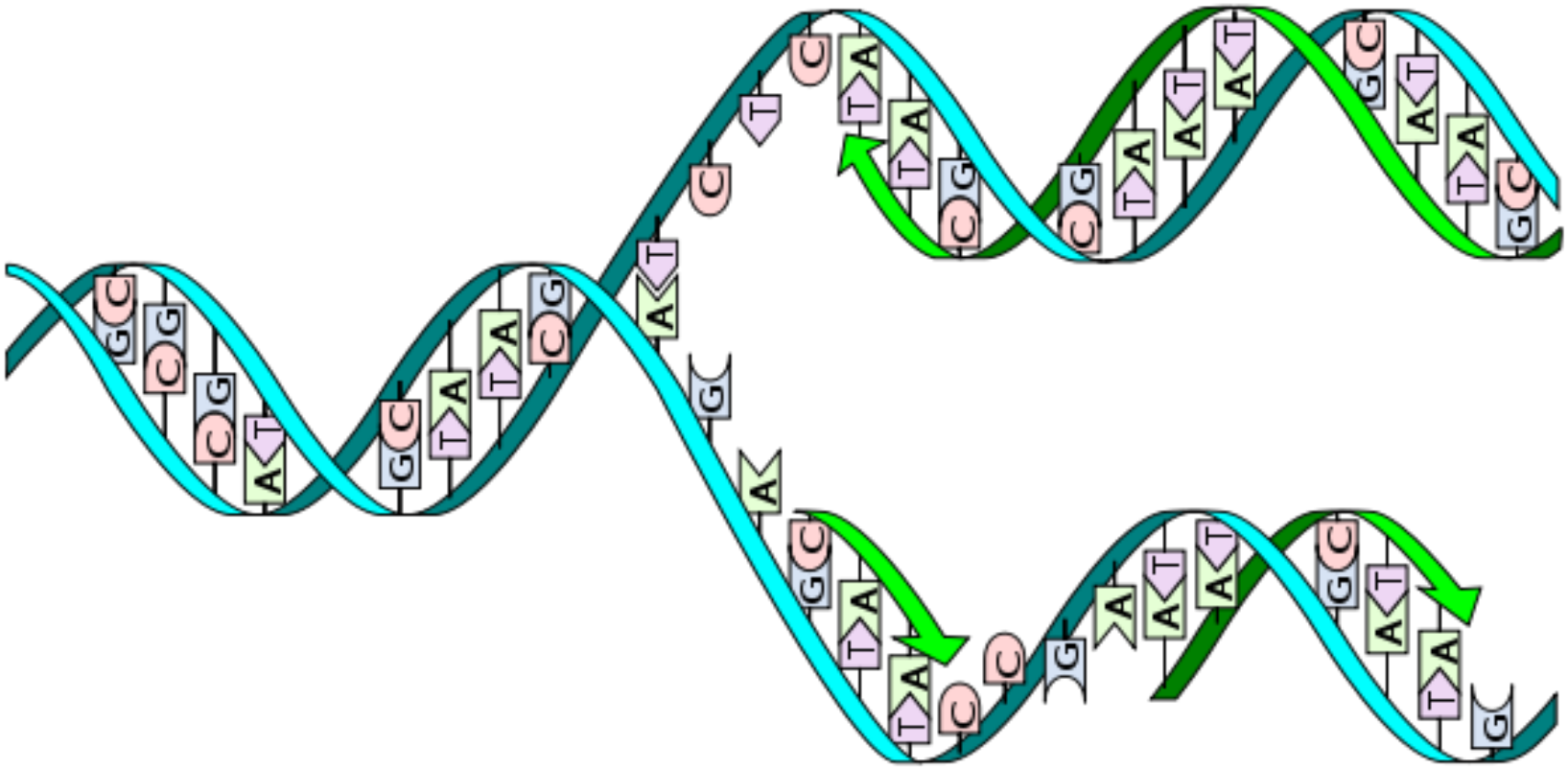
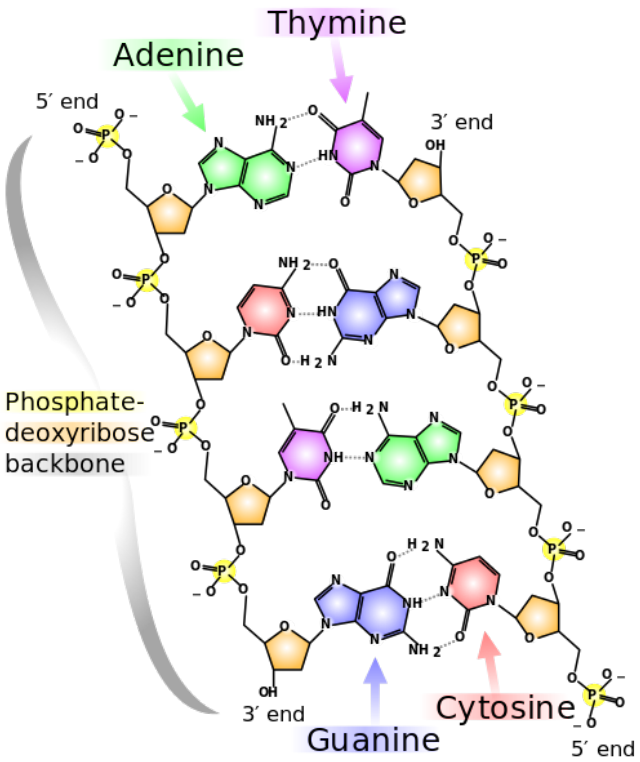
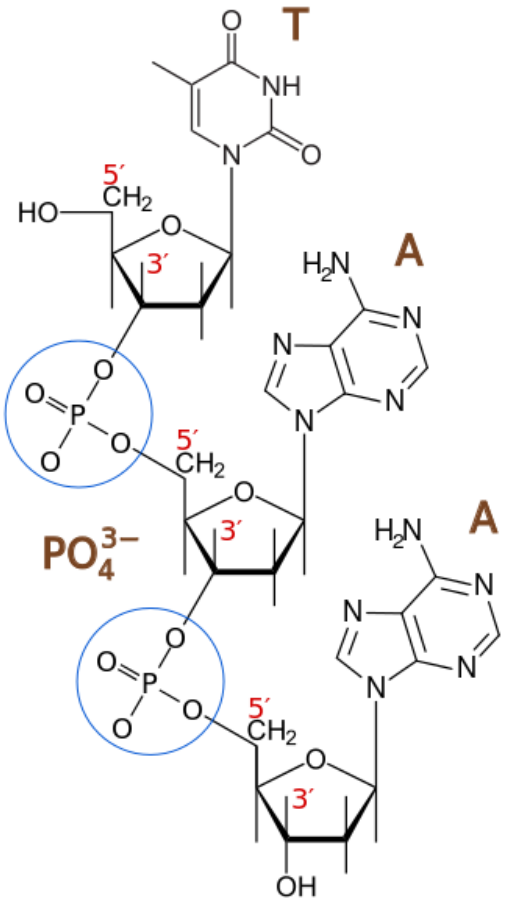


image (c) wikimedia

DNA structrue





images (c) wikimedia

DNA replication – mechanism

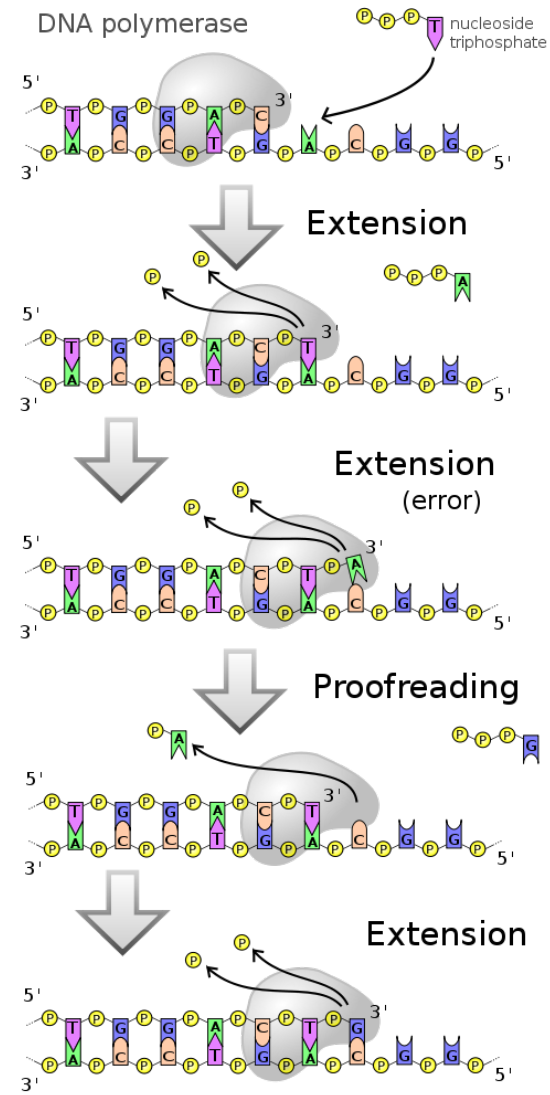


image (c) wikimedia

- DNA polymerase is the key enzyme for DNA replication
- During replication, helper enzymes carry out "proof-reading" of the replicated strand
- error rate (under no stress) $< 10^{-7}$ nucleotides

How sequences evolve?

GTCTGTAGTA

image (c) BW

How sequences evolve?

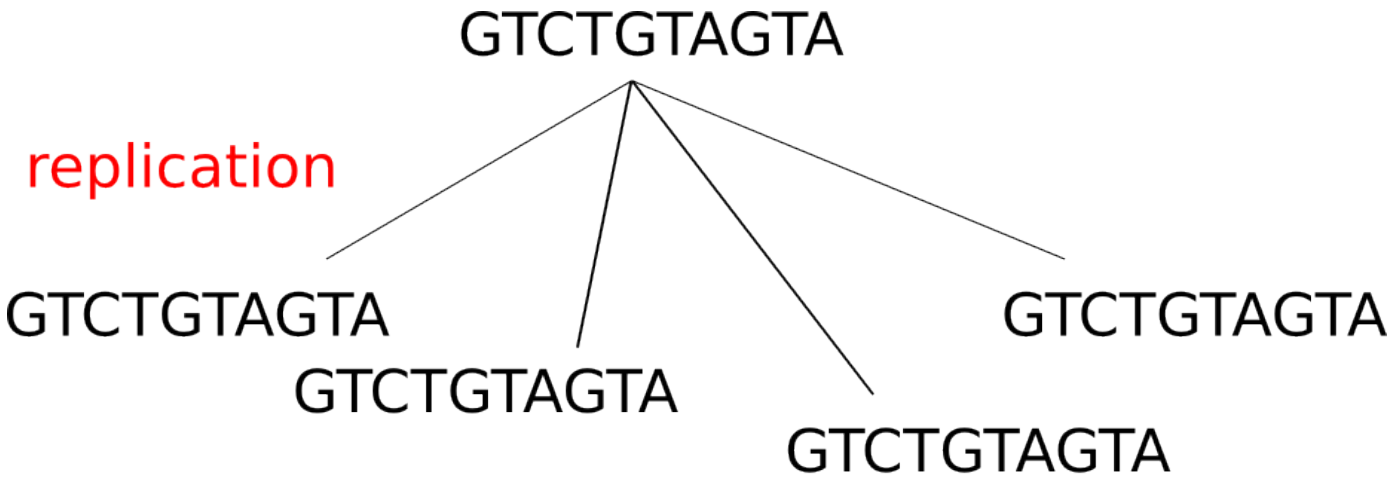


image (c) BW

How sequences evolve?

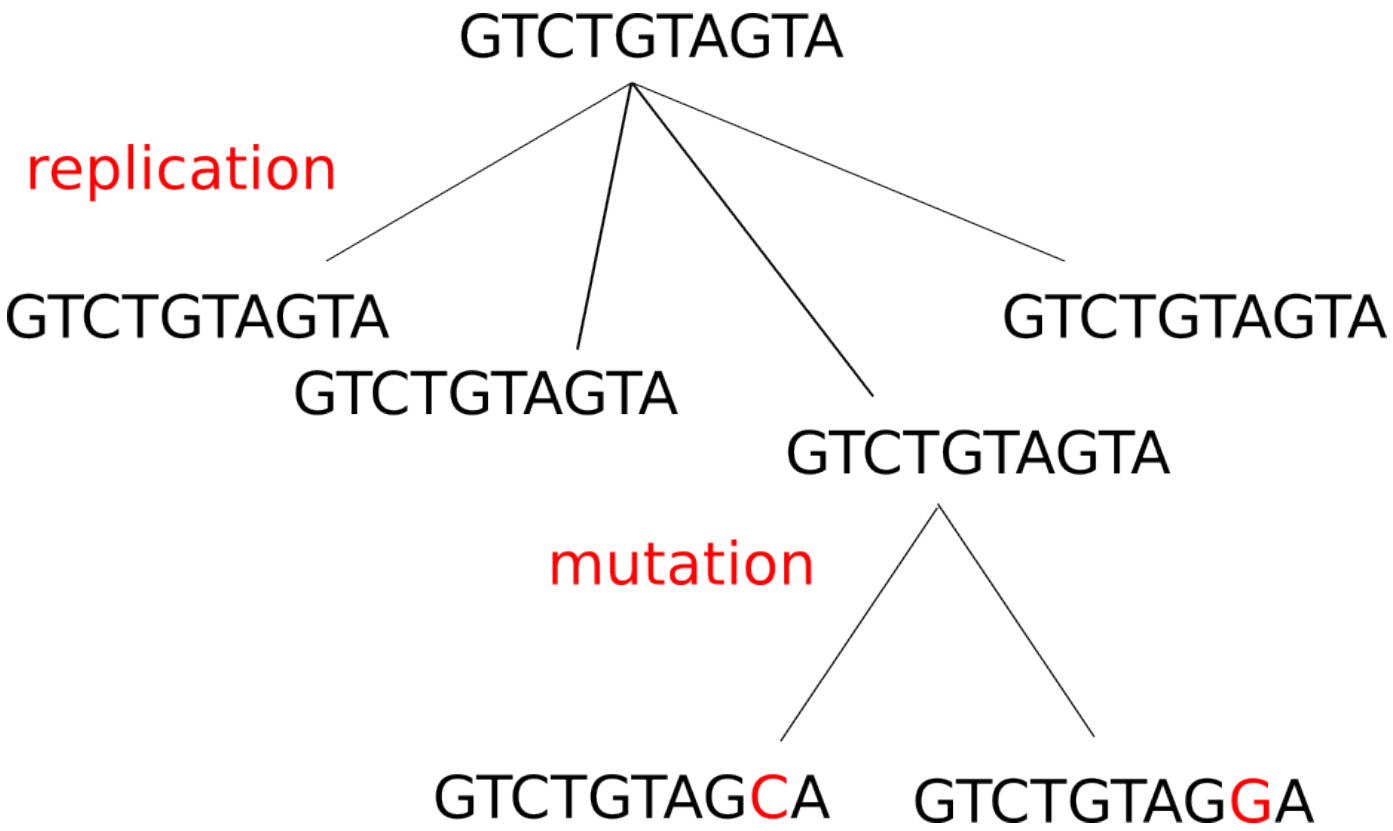


image (c) BW

How sequences evolve?

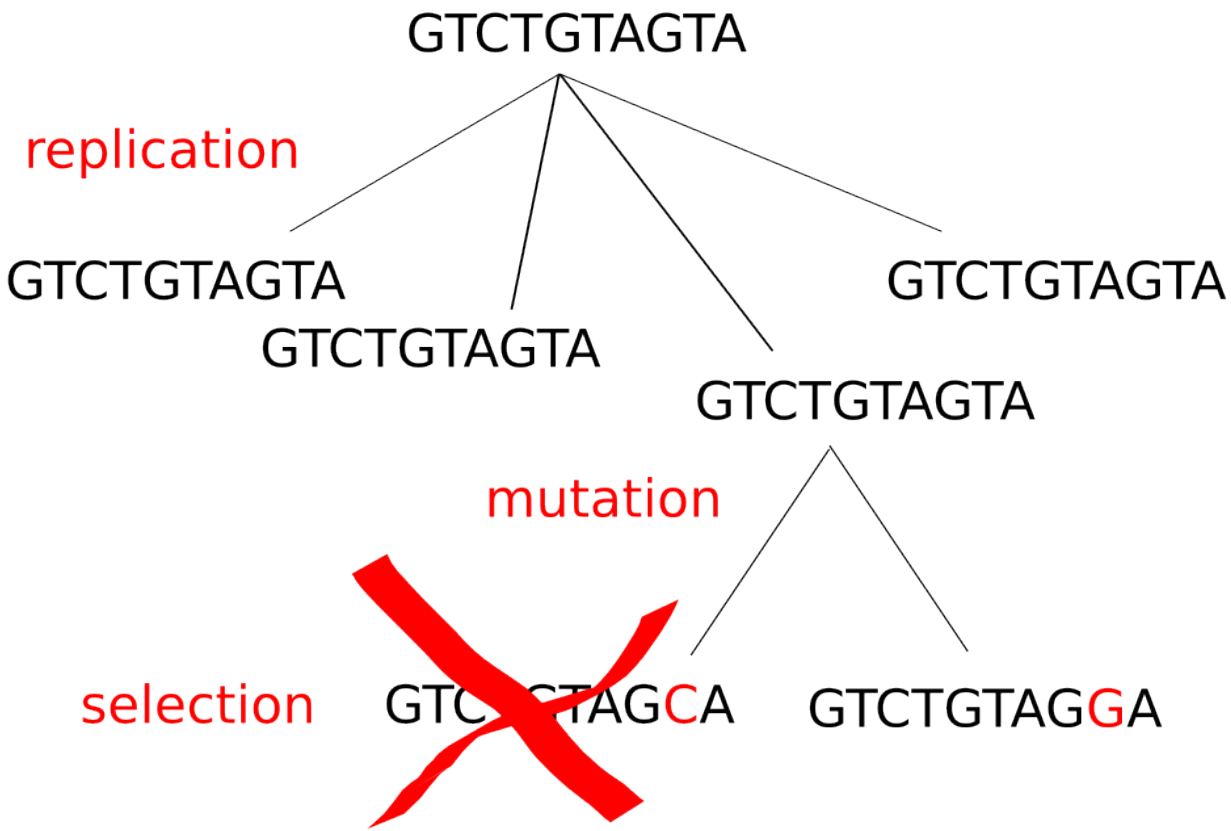


image (c) BW

Fundamental problem

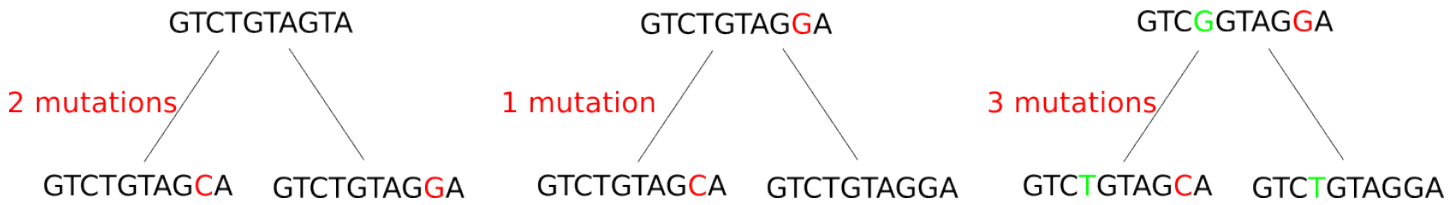
- How far in evolution are the sequences that we can observe in different living species?
- More formally: Can we define a measure of sequence similarity

$$d : \Sigma^* \times \Sigma^* \rightarrow \mathcal{R}^+$$

  approximating the true evolutionary distance?
- Hint: We should count the number of mutations leading to the observed divergence.

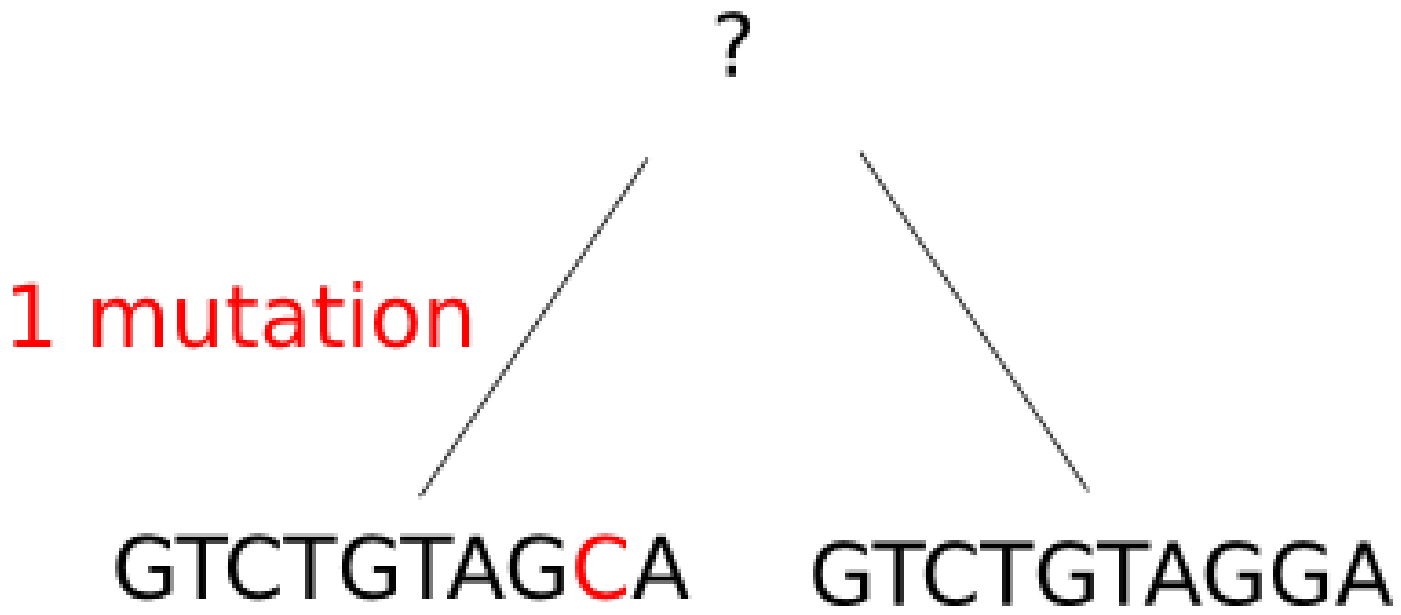
Subproblem 1: multiple scenarios

We can observe only the current situation. What about ancestral sequences?



Solution: *Parsimony* – In case of lack of evidence for a more complex situation, take the simplest possible explanation.

Subproblem 2: Time reversibility



Technically, in order to estimate the ancestral sequence, we need to assume that the process is "time-reversible", i.e. In the stable state, the rates of mutating the sequence s_1 into s_2 are the same as s_2 into s_1. This is a reasonable simplification for "short" evolutionary time-scales.

Approach 1: Markov chain

- Time-reversible Markov Chain*
- Sequences from Σ^k are states (How many of them?)
- Transition probabilities assume independent base substitution
- We need to define a symmetric base *substitution matrix*
- (*) In fact, we should consider a continuous-time Markov chain, to avoid problems with exact generation times...

Trasition and rate matrices

- We need to define the transition matrix

$$P(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AG}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GG}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CG}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TG}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) \end{pmatrix},$$

- From the Markov property, we know that $P(t+\tau) = P(t)P(\tau)$
- Converting to an additive model, we have $\mathbf{p}(t+\Delta t) = \mathbf{p}(t) + \mathbf{p}(t)Q\Delta t$,
- where the rate matrix
- $$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}.$$
- The dynamics is described by $\mathbf{p}'(t) = \mathbf{p}(t)Q\cdot$ and $P(t) = \exp(tQ)$,

The simplest model JC69 (Jukes-Cantor, 1969)

Only one parameter: μ

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Solution for continuous time t:

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Kimura 1980 (K80) model

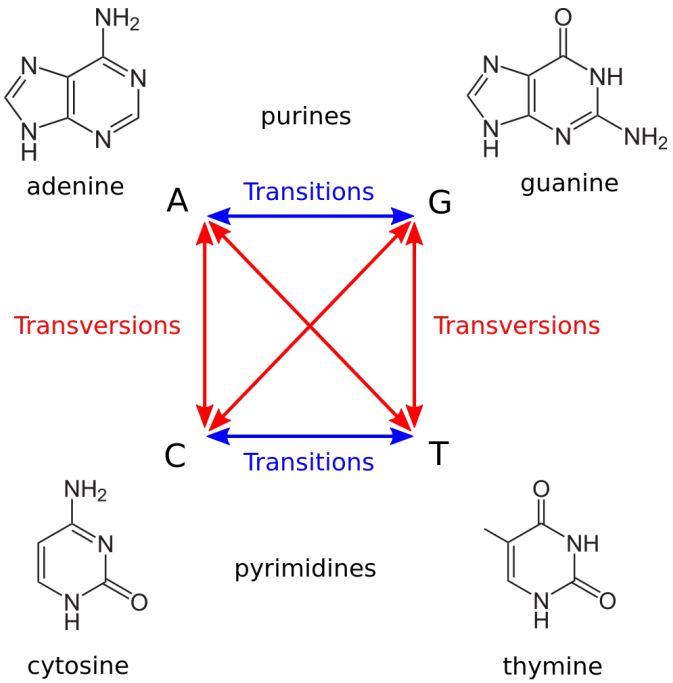


image (c) wikimedia

- We can observe that transitions are different than transversions. This leads to the Kimura model (with p,q being the probability of transition, transversion).
- $$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$
- $$K = -\frac{1}{2}\ln((1-2p-q)\sqrt{1-2q})$$

Felsenstein model F81 (Felsenstein, 1981)

We do not assume equal probability of nucleotides, but a distribution, with

$$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$$

Then the mutation rate matrix may look like the following

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

Problems with DNA evolution models

- Mutations occur on DNA level, but selection acts much higher: on the phenotype level.
- This makes the assumption of base independence invalid
- Long evolutionary times violate time-reversibility
- Multiplicative measure not too convenient in practice
- We can only account for substitutions, not for insertions or deletions

Suggested solutions:

- Use protein sequences for comparisons
- Define additive substitution matrices

mRNA translation into proteins

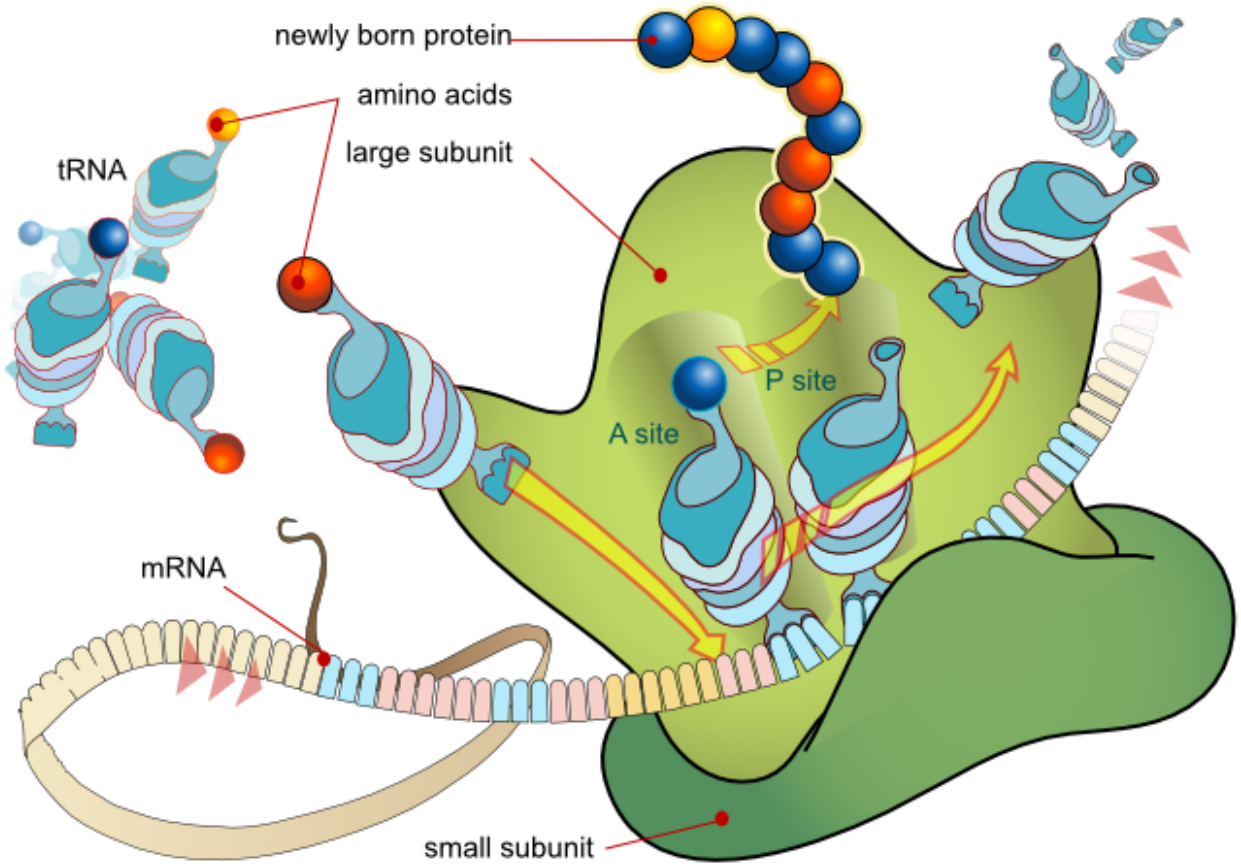


image (c) wikimedia.org

Protein codon table

Second base

First base	U	C	A	G	Third base
U	UUU, UUC — Phenylalanine F UUA, UUG — Leucine L	UCU, UCC, UCA, UCG — Serine S	UAU, UAC — Tyrosine Y UAA Stop codon UAG Stop codon	UGU, UGC — Cysteine C UGA Stop codon UGG Tryptophan W	U C A G
C	CUU, CUC, CUA, CUG — Leucine L	CCU, CCC, CCA, CCG — Proline P	CAU, CAC — Histidine H CAA, CAG — Glutamine Q	CGU, CGC, CGA, CGG — Arginine R	U C A G
A	AUU, AUC, AUA — Isoleucine I AUG Methionine start codon M	ACU, ACC, ACA, ACG — Threonine T	AAU, AAC — Asparagine N AAA, AAG — Lysine K	AGU, AGC — Serine S AGA, AGG — Arginine R	U C A G
G	GUU, GUC, GUA, GUG — Valine V	GCU, GCC, GCA, GCG — Alanine A	GAU, GAC — Aspartic acid D GAA, GAG — Glutamic acid E	GGU, GGC, GGA, GGG — Glycine G	U C A G

image (c) biogem.org

Protein substitution matrices

- We are still assuming time-reversible Markov chain, but now in space of protein sequences.
- Matrix entries contain log-probabilities, leading to additive measures of similarity
- PAM (Point accepted mutations) matrices (Dayhoff, 1978) describe observed probabilities of occurence of point mutations for a given average divergence (PAM1 = one mutation/100 bases, mostly used PAM250)
- BLOSUM (BLOcks Substitution Matrix) (Henikoff, Henikoff 1992) were constructed using short protein alignments (Blocks) of given sequence identity. e.g.BLOSUM80 was derived from sequences of $\geqslant 80\%$ identity