Substitution patterns

"Organic life, we are told, has developed gradually from the protozoon to the philosopher, and this development, we are assured, is indubitably an advance. Unfortunately, it is the philosopher, not the protozoon, who gives us this assurance."

(B. Russel, *Mysticism and Logic*, 1918)

Table of contents

- Molecular evolution
- Substitution patterns in genes
- Estimation of the number of substitutions
- Differences in the gene evolutionary speed
- Molecular clocks
- The evolution in organelles

Introduction

- Comparisons among nucleotide sequences of two or more organisms often reveal that changes have been accumulated, at the DNA level, even if all the sequences come from functionally equivalent regions
- Actually, it is not uncommon that, during the evolution, homologous sequences have become so different as to make it very difficult to obtain reliable alignments
- The analysis of both the number and the type of substitutions, that have been occurred during time, are of central importance for the study of molecular evolution

Why do we use molecular evolution? – 1

- *DNA molecules are not only the key to heredity, but they are "document of evolutionary history"* (Emile Zuckerkandl)
- Molecular evolution integrates evolutionary biology, molecular biology, and population genetics
  - It describes the process of evolution (changes in time, Being vs Becoming) of DNA, RNA and proteins
  - It includes the study of the rate of changes of genetic sequences, of the relative importance of adaptive and neutral changes, and of changes in genome structure
  - It deals with patterns (diagrams, models) and studies the evolution of…
    - …molecular entities, like genes, genomes, proteins, introns, chromosomal arrangements
    - …organisms and biological systems, i.e. species, systems that co–evolve, ecological niches, migration patterns

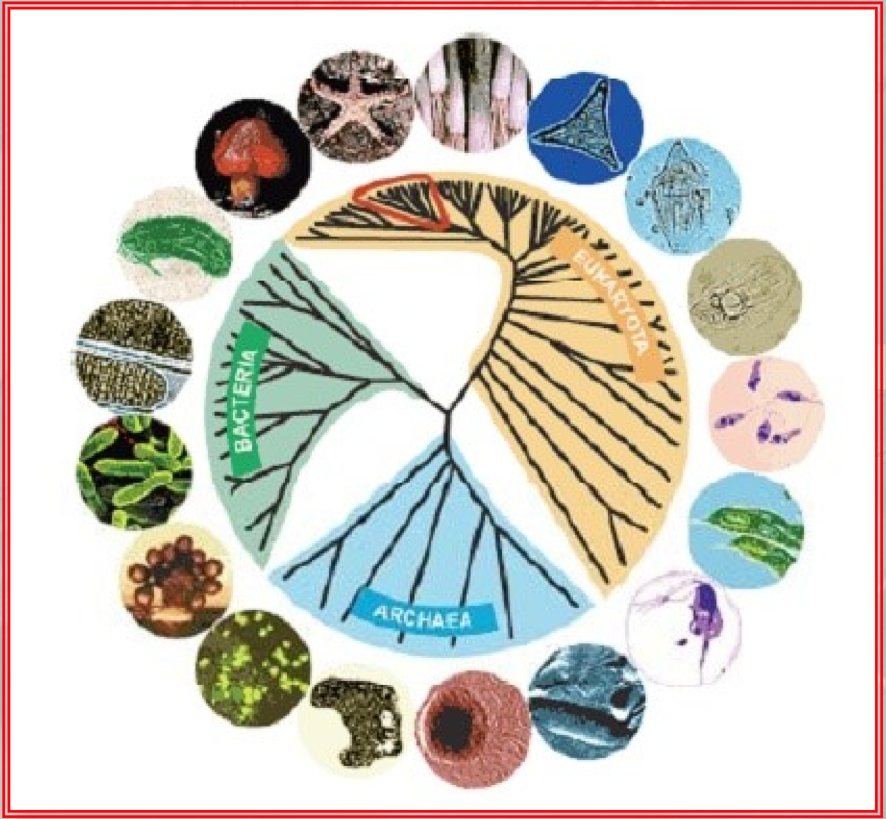
  using molecular data

Why do we use molecular evolution? – 2

- In order to understand the basis of biological diversity

Why do we use molecular evolution? – 3

- In order to understand the evolutionary history of the life on the Earth, which is written in our molecules

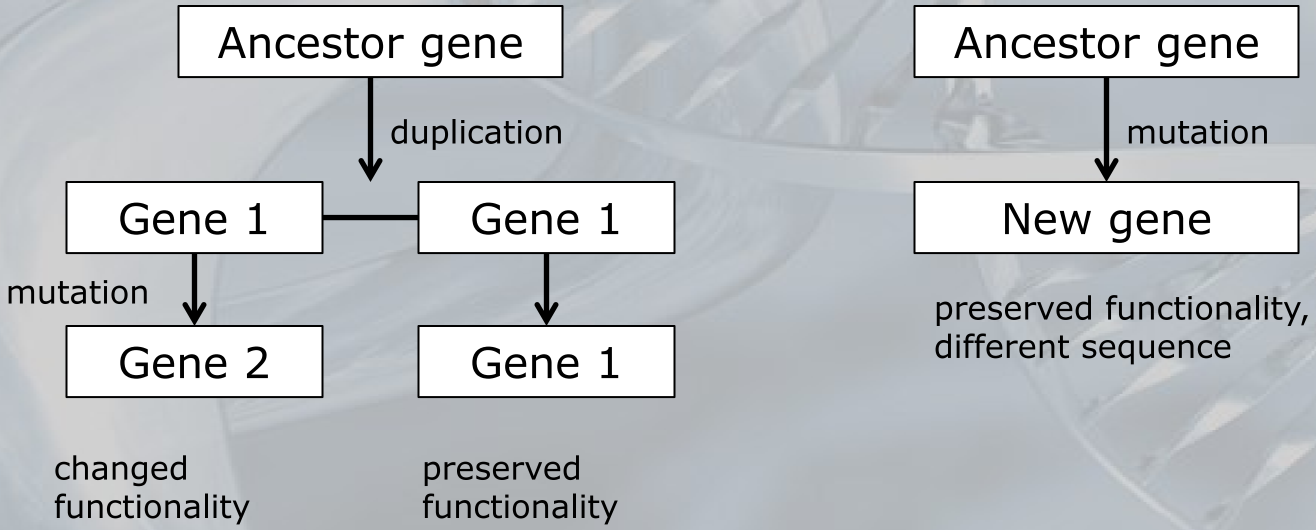


Why do we use molecular evolution? – 4

- Since the process of natural selection is truly effective in removing harmful changes, molecular evolution also serves to recognize and characterize the genome portions that are more important from the functional point of view
- …or, in other words, to detect how the frequency of the nucleotide replacements is different in different areas of the same gene, for different genes, and across species, and may be used as a measure of the functional significance of a particular sequence (and, therefore, it accounts for the need of its "conservation")

Genes and proteins – 1

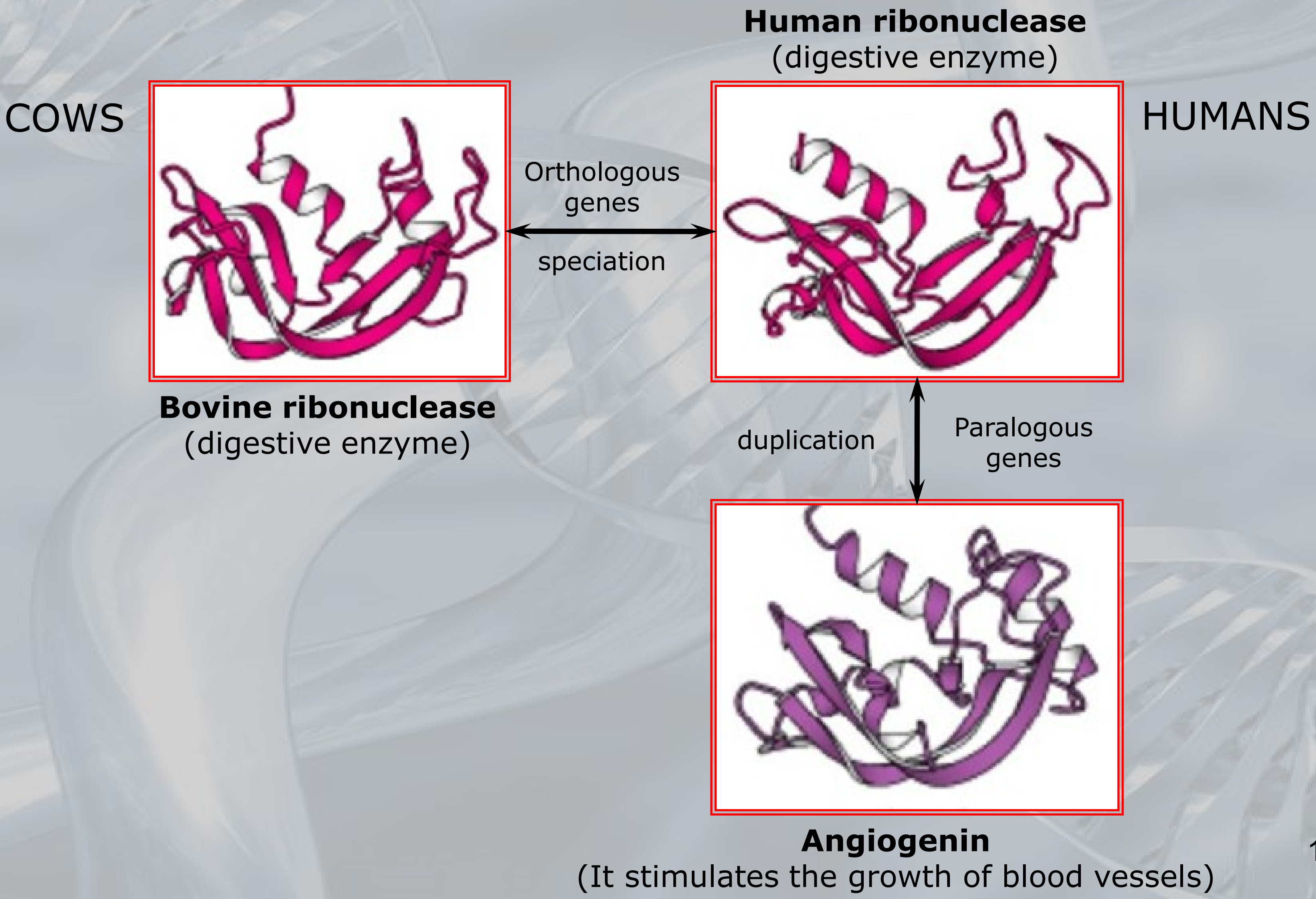
- Why do proteins change?
  - Because there are many proteins that perform the same or similar function(s) (within the organisms), so if a particular protein changes, the function is still preserved
  - Because such change does not affect neither the structure (destabilization) nor the function of the protein



Genes and proteins – 2

- Gene duplication, or gene amplification, is an important mechanism by which new genetic material is generated during molecular evolution; it can be defined as any duplication of a DNA region that contains a gene and can arise due to different types of errors in DNA replication and repair machinery
- Orthologous genes: similar genes, found in organisms related to each other
  - The speciation phenomenon leads to the divergence of genes and, therefore, of the proteins that they encode
  - Example: Human and mouse β–globins started to diverge about 80 million years ago, when the evolutionary event, that gave rise to primates and rodents, took place
- Paralogous genes: genes originated from the duplication of a single gene in the same organism
  - Example: Human α–globins and β–globins began to diverge due to the duplication of an ancestral globin gene
- In both cases, there is homology

Genes and proteins – 3



COWS

HUMANS

Human ribonuclease
(digestive enzyme)

Orthologous genes

speciation

Bovine ribonuclease
(digestive enzyme)

duplication

Paralogous genes

Angiogenin
(It stimulates the growth of blood vessels)

How proteins can change – 1

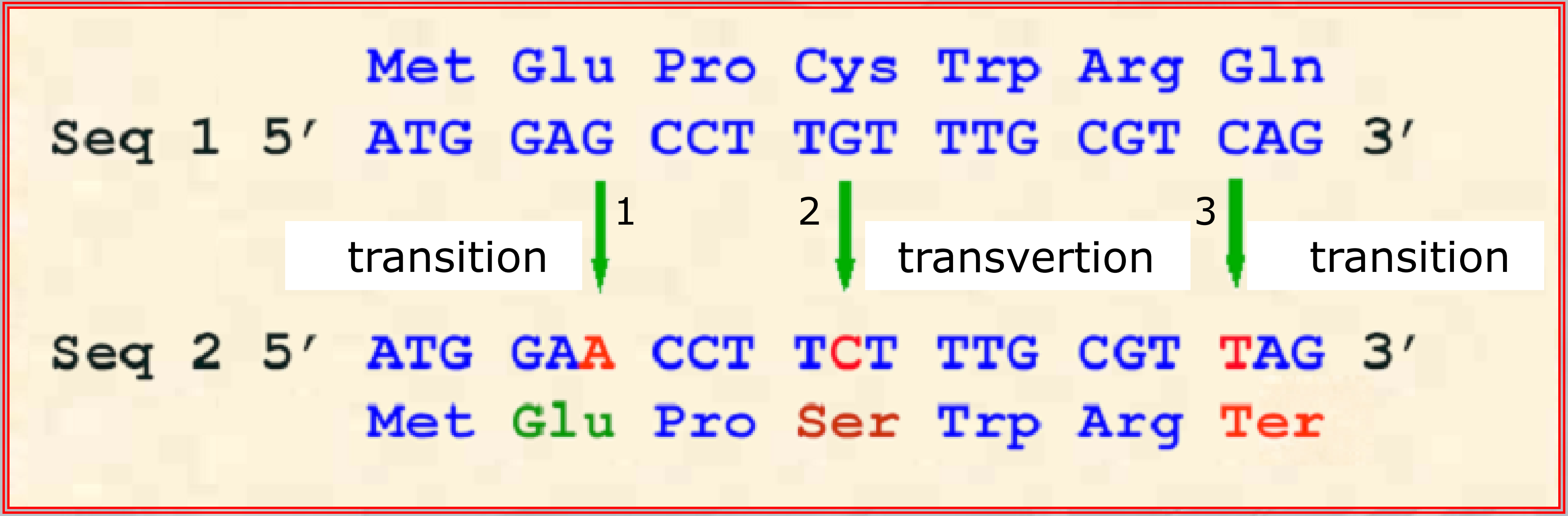
- A protein present in a particular organism can change as a result of some mutations in its coding sequence
- Mutations can be point–like or frame–shift
  - Point mutations ⇨ substitution of a single nucleotide
  - Insertion ⇨ one or more nucleotides are inserted
  - Deletion ⇨ one or more nucleotides are removed
  - Inversion ⇨ a DNA stretch is reversed

How proteins can change – 2

- The genetic code is redundant and, therefore, a substitution does not always lead to a change of an amino acid
  - A silent mutation occurs if the protein remains (functionally) unchanged
- In other cases, from the mutation point onwards, all the amino acids change, and the protein can become “unrecognizable” and definitely lose its functionality

How proteins can change – 3

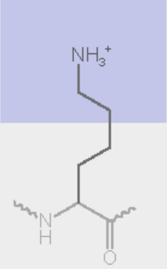
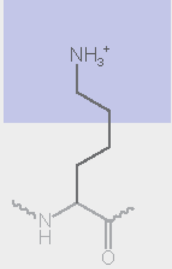
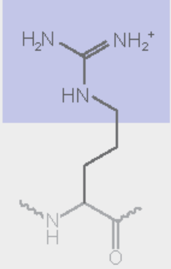
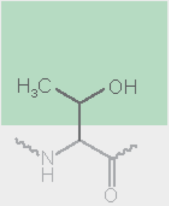
Point mutations



```
               Met Glu Pro Cys Trp Arg Gln
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′
                   1           2              3
      transition      transvertion     transition
Seq 2 5′ ATG GAA CCT TCT TTG CGT TAG 3′
               Met Glu Pro Ser Trp Arg Ter
```

(1) Glutamic acid → Glutamic acid
(2) Cysteine → Serine (amino acids with a polar, chiral molecule)
(3) Glutamine → Stop codon

How proteins can change – 4

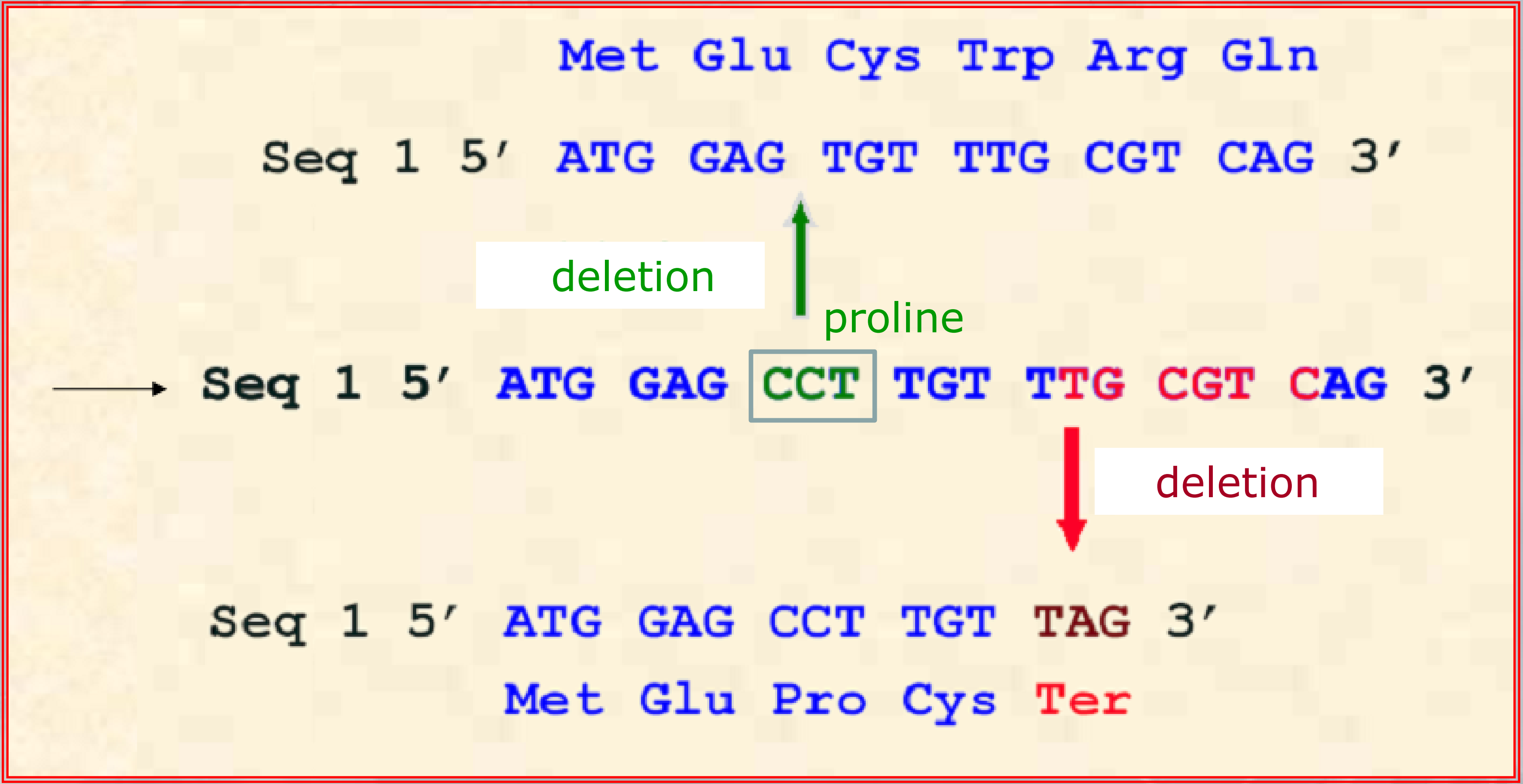
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr
					

basic 🟦
polar 🟩

Arginine and lysine are both basic amino acids (positively charged), while threonine is a polar amino acid (hydrophilic)

How proteins can change – 5

Deletions



Met Glu Cys Trp Arg Gln

Seq 1 5′ ATG GAG TGT TTG CGT CAG 3′

deletion

proline

Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′

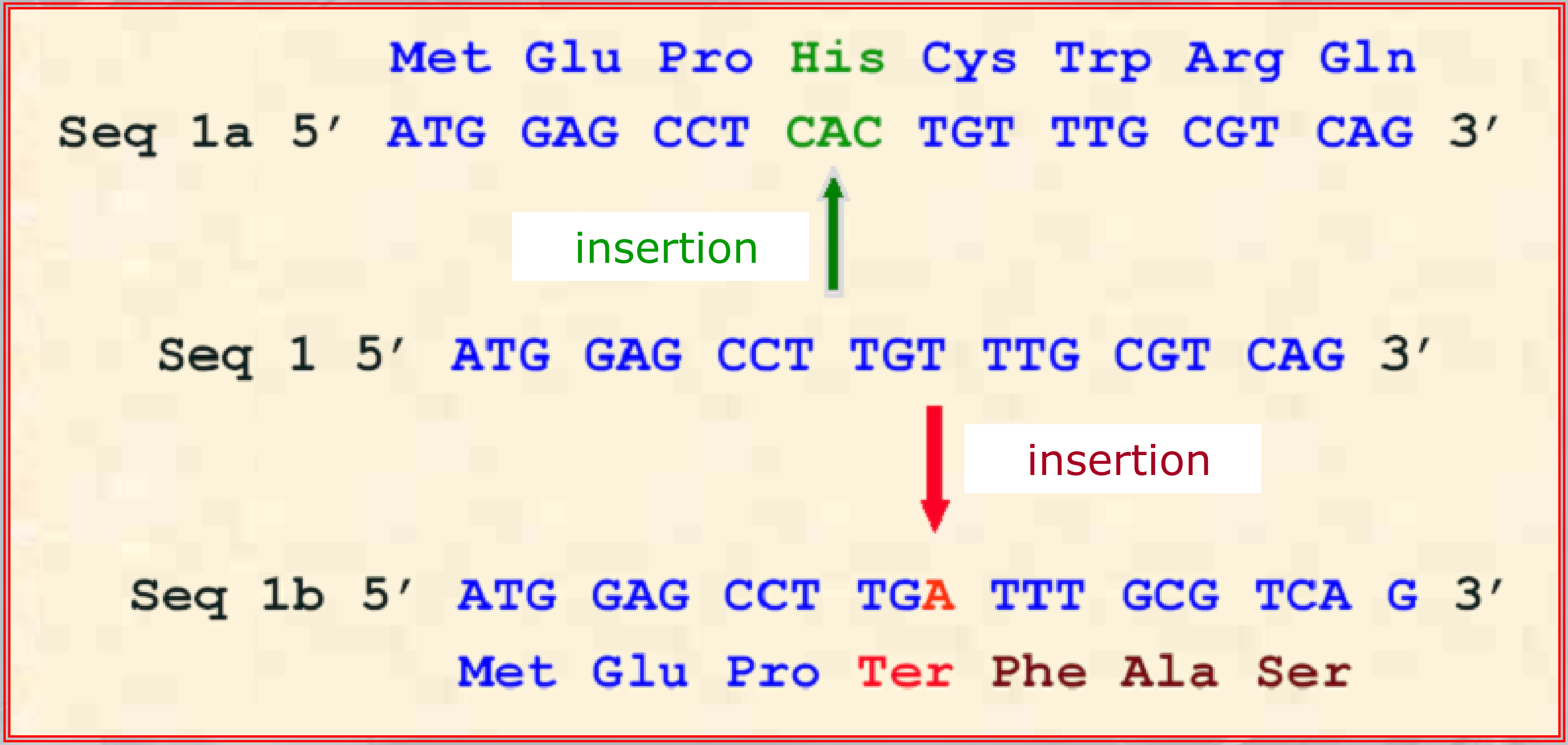
deletion

Seq 1 5′ ATG GAG CCT TGT TAG 3′

Met Glu Pro Cys Ter

How proteins can change – 6

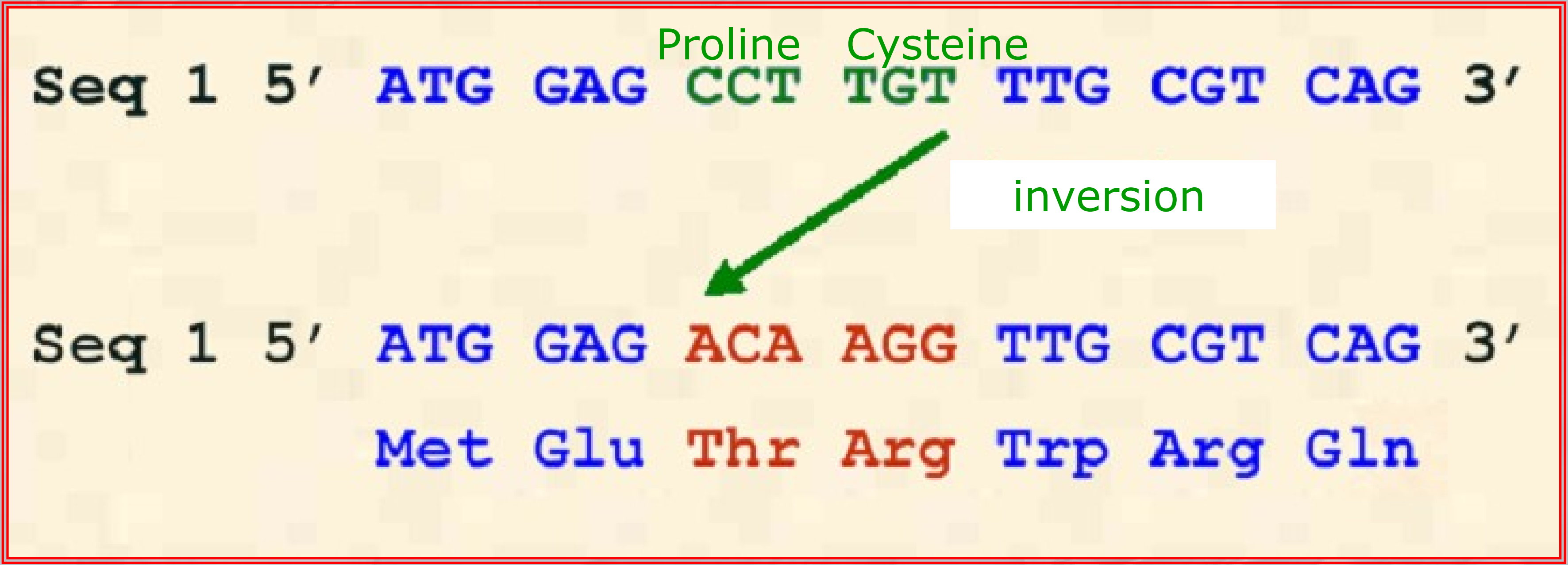
Insertions



```
              Met Glu Pro His Cys Trp Arg Gln
Seq 1a  5′  ATG GAG CCT CAC TGT TTG CGT CAG 3′

                    insertion

   Seq 1  5′  ATG GAG CCT TGT TTG CGT CAG 3′

                              insertion

  Seq 1b  5′  ATG GAG CCT TGA TTT GCG TCA G 3′
              Met Glu Pro Ter Phe Ala Ser
```

How proteins can change – 7

Inversions



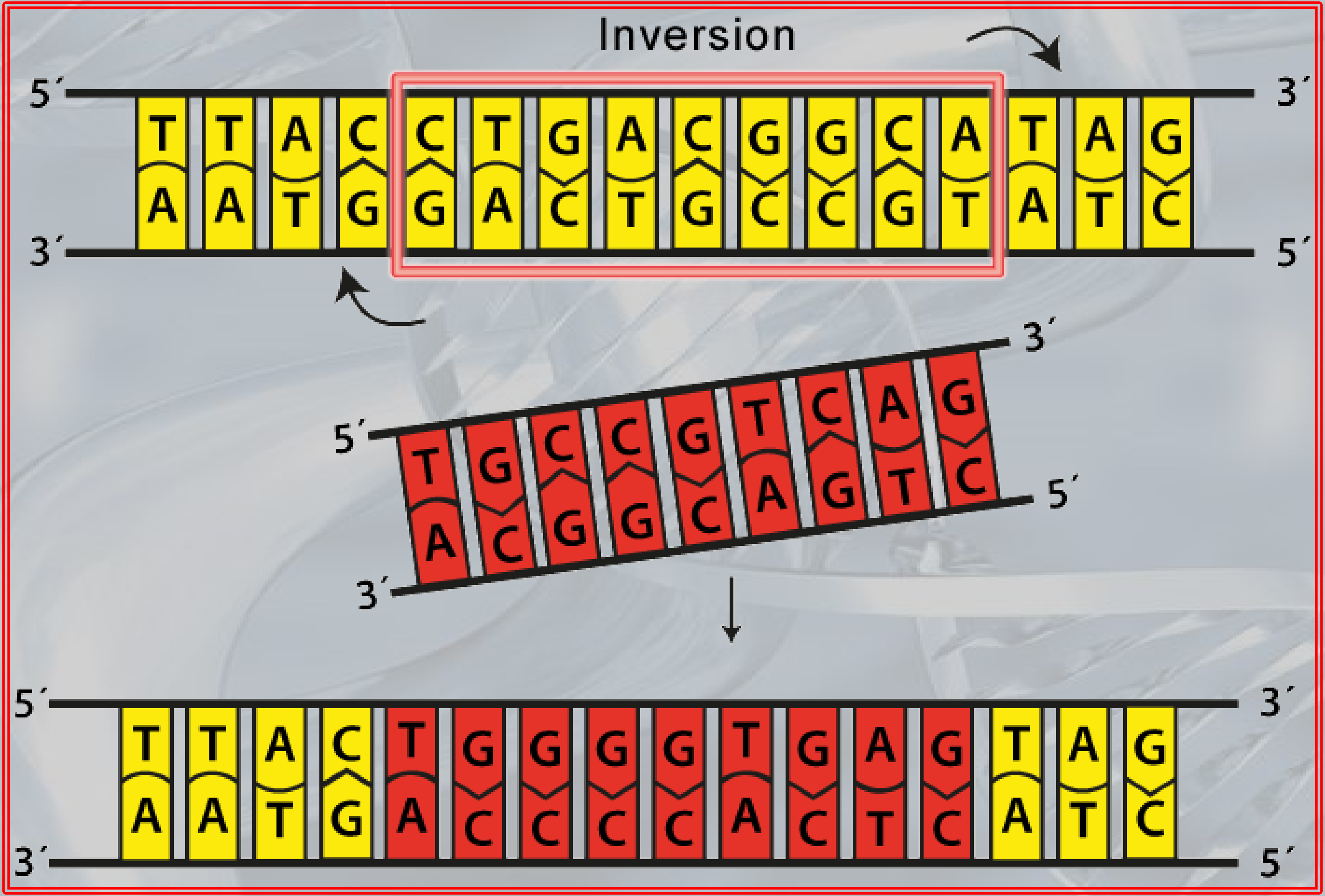
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′

Seq 1 5′ ATG GAG ACA AGG TTG CGT CAG 3′

Met Glu Thr Arg Trp Arg Gln

If two breaks occur in a sequence, sometimes the region between the breaks rotates 180° before rejoining with the two end fragments. Unlike indels, inversions do not change the overall amount of the genetic material, so inversions are generally viable and show no particular abnormalities at the phenotypic level. In some cases, breaks happen within a gene of essential function, and then that breakpoint acts as a lethal gene mutation.

How proteins can change – 8


Inversion
5´
T T A C C T G A C G G C A T A G
A A T G G A C T G C C G T A T C
3´
3´
5´
5´
T G C C G T C A G
A C G G C A G T C
3´
3´
5´
5´
T T A C T G G G G T G A G T A G
A A T G A C C C C A C T C A T C
3´
3´
5´

How proteins can change – 9

- Biological similarity is often due to homology, but can also randomly occur, or be due to adaptive convergence phenomena, both morphological (analogy) and at the molecular level
- Adaptive convergence: Different organisms may adopt similar “technical solutions” to fit similar environments, sometimes even starting from a very different organ or apparatus
- Example: Bird wings and bat wings have evolved independently and, therefore, they are not homologous
- For DNA sequences, it is more correct to use the term *similarity*, as it is always possible to establish if (and how much) two sequences are similar, whereas if the similarity is due to homology, to adaptive convergence, or to chance cannot always be established

shutterstock

How proteins can change – 10

- If two sequences have a significant degree of similarity for all their length, it is very likely that this is due to a a sort of "memory" of their evolutionary relationship
- Two sequences that do not show a strong similarity, however, can still be homologous (sharing a very remote common ancestor, or having subdue to a very rapid evolutionary dynamics)
- Note that… Similarity ≠ Homology

It is a quantitative information, based on the chosen metrics, and it is independent from assumptions about the cause of the similarity itself

It represents a qualitative information, that stands for the common phylogenetic origin of two sequences

Substitution patterns in genes – 1

- Alterations in the DNA sequences can have drastic consequences for living cells
  - Mutations: Nucleotide changes or indel events
  - Errors can be deleterious, beneficial or neutral
  - In addition:
    - Beneficial changes usually occur with a lower frequency
    - Some changes in nucleotide sequences have more significant consequences, which further differ in relation to different genes and different organisms
- However, for an organism in its typical environment, most of the genes are very close to the optimal status
  - Cells have developed complex mechanisms that ensure the accuracy of the DNA replication and repair

Substitution patterns in genes – 2

- The DNA replication is the molecular mechanism that produces a copy of the cellular DNA
  - Each time a cell divides itself, in fact, the entire genome must be replicated, in order to be passed on to the offspring
- Each strand of the original DNA molecule serves as a template for the production of its counterpart, in a semi–conservative replication process
  - The new helix will be composed of an original DNA strand as well as a newly synthesized strand
  - Cellular proofreading and error–checking mechanisms

Substitution patterns in genes – 3

- The process of DNA repair is a set of mechanisms by which a cell identifies and corrects damages to its DNA molecules
- The DNA repair is essential to the cell survival, since it protects the genome from permanent and harmful mutations
  - It is a process that takes place continuously
  - Example: in human cells, both normal metabolic activities and environmental factors determine at least 500,000 (but up to one million) individual molecular lesions per cell per day

Substitution patterns in genes – 4

- When the cells get older, the rate of DNA replication/ repair decreases, until no longer being able to keep up with the damage events
  - Senescence (irreversible dormancy) indicates the process by which, during cell replication, some cells gradually lose their ability to divide themselves
  - Apoptosis (programmed cell death) is a sophisticated mechanism in which the cellular evolution has acted as the sieve to defend the body from virus–infected cells, from autoreactive cells of the immune system, from cells in which DNA damages occurred, and from tumor cells
  - Carcinogenesis is the process that transforms normal cells into cancer cells

Mutation frequencies – 1

- The number of substitutions K that two homologous sequences have been undergone since their last common ancestor can be estimated by counting their (measurable) differences
- When K is expressed in terms of the number of substitutions per site and it is coupled with a divergence time T
  - The replacement frequency r can esplicitly be evaluated

Mutation frequencies – 2

- Assuming that substitutions accumulate simultaneously and independently in both sequences, the frequency of replacement is equal to

$$r = K/(2T)$$

- The replacement frequency evaluation is effective if the evolutionary rates, in different species, are similar
  - Time estimation for evolutionary events (supposing r almost constant in a certain region)
- Comparisons among replacement frequencies within the same gene, and among genes, are useful to determine the role of different genomic regions

Functional constraints – 1

- Changes in genes that decrease the life expectancy of an organism are "stemmed" by the process of natural selection
- Since proteins are responsible for the cell functionality, it is not surprising that those changes in the nucleotide sequence which cause the structural or catalytic properties of the encoded proteins to be varied are subject to natural selection
  - Those portions of genes, to which a particular importance is recognized, are defined under functional constraints and tend to vary little (or to change very slowly) over the course of evolution

Functional constraints – 2

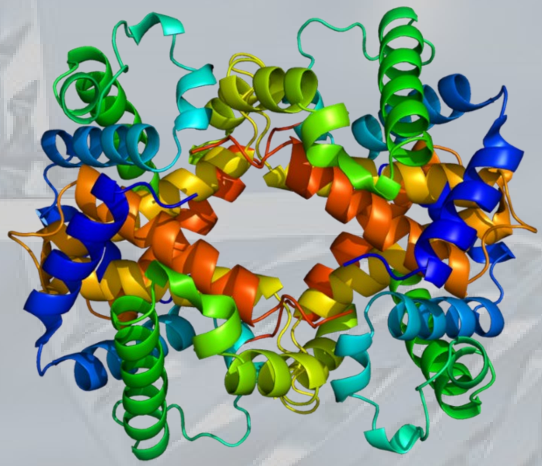
- Conversely, many changes in the nucleotide sequence of a gene do not have effect on the coding of the amino acid sequence or on the expression level of the codified protein
  - This type of changes are less subject to natural selection and rapidly accumulate during the evolutionary process

Functional constraints – 3

- Example: Accumulated changes in the genes for β–globins of four different mammals (human, mouse, rabbit and cow, who have had a common ancestor ~100 million years ago)

Region	Length (in base pairs)	Substitution frequency
Non coding sequences	913	3.33
Coding sequences	441	1.58
5′–*flanking* sequence	300	3.39
Untranslated 5′ sequence	50	1.86
Intron 1	131	3.48
Untranslated 3′ sequence	132	3.00
3′–*flanking* sequence	300	3.60

Substitution frequency calculated as the number of substitution per site over one billion years



Functional constraints – 4

- Example (cont.)
  - A typical eukaryotic gene (and adjacency) is composed both by nucleotides that specify the amino acid sequence of a protein (coding sequences), and by non–coding sequences
  - The rate of changes is about twice in the non–coding sequences of the genes for the β–globins (3.33×10^{-9} substitutions/site/year against 1.58×10^{-9} substitutions/site/year)
  - The non–coding sequences are divided into:
    - Introns
    - *Leader* regions, transcribed but not translated (upstream w.r.t. the gene starting site)
    - *Trailer* regions, transcribed but not translated (downstream w.r.t. the gene end site)
    - Adjacent sequences w.r.t. 5′ and 3′ terminations

Functional constraints – 5

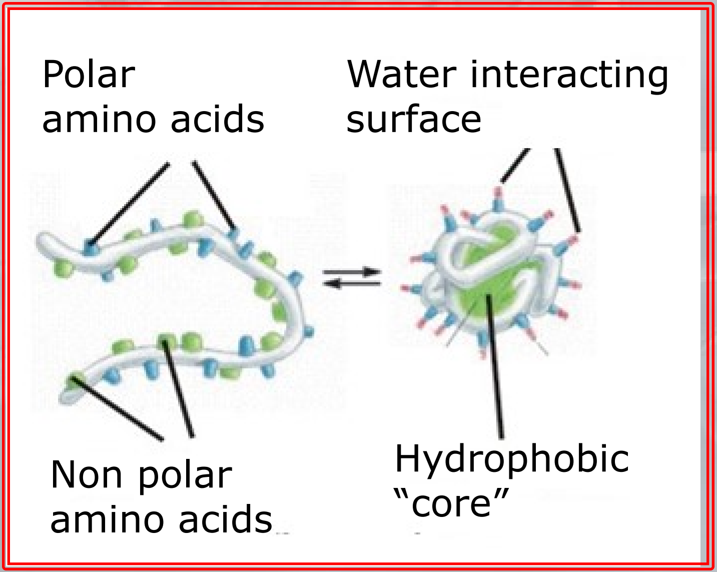
- Example (cont.)
  - Each region tends to accumulate changes at different rates, based on the strength of the functional constraints on its nucleotides
  - In addition, it is logical to expect that different genes accumulate substitutions at different frequencies, as well as the genes for the β–globins underlie different levels of functional constraints for distinct species
  - However, in general:
    - Changes accumulate more rapidly within introns and flanking sequences…
    - … then in the regions that are transcribed but not translated (with the only exception of the sequence at the 5′ termination, which is functionally important for the sub-sequent phase of gene translation)
    - … and, finally, less rapidly within coding sequences

Functional constraints – 6

- Example
  - The β–globin data provide an estimate of the times at which nucleotide changes occur
  - While for a nucleotide sequence, a change of 0.35% per million years (the approximate frequency of changes within introns and flanking sequences) may seem extremely slow from a human perspective, it proves relatively fast from the molecular evolution point of view

Functional constraints – 7

- From the structural point of view:
  - Most mutations occur on the protein surface, while the *core* amino acids are more conserved, so as to allow the same folding
  - In the evolution, the sequence similarity is less preserved than the tertiary structure



Synonimous and non-synonymous substitutions – 1

- 18 out of 20 amino acids are encoded by more than one codon
  - For instance, **GGG**, **GGA**, **GGU**, **GGC** codify all for glycine
  - Every change in the third position of a codon for glycine leads to a codon that ribosomes interpret equivalently for the construction of the primary structure of the protein
- Changes at the nucleotide level that do not vary the amino acid sequence are called synonymous substitutions
- Conversely, changes in the second position of the glycine codon can cause changes in the resulting amino acid sequence (for example, **GCG** codify for alanine) and represent a non–synonymous substitution (conservative missense)

Synonimous and non-synonymous substitutions – 2

- If it is true that natural selection performs a clear distinction between functional and dysfunctional proteins, the synonymous substitutions should be observed more frequently than non–synonymous ones (at least, in coding sequences)
- Moreover, not all the positions within the nucleotide triplet representing a codon give rise to non–synonymous substitutions with the same probability

Synonimous and non-synonymous substitutions – 3

- The positions within a codon actually belong to three different categories:
  - Non degenerate sites: codon positions where mutations always result in some amino acid substitutions (e.g., **UUU** codifies for phenylalanine, **CUU** for leucine, **AUU** for isoleucine, and **GUU** for valine)
  - Doubly degenerate sites: codon positions where two different nucleotides lead to the translation of the same amino acid, while the other two encode for a different amino acid (e.g., **GAU** and **GAC** codify for aspartic acid, whereas **GAA** and **GAG** for glutamic acid)
  - Four times degenerate sites: codon positions in which the change of a nucleotide with each of the other three alternatives has no effect on the amino acid that ribosomes translate into the protein (e.g., the third position of the glycine codon)

Synonimous and non-synonymous substitutions – 4

- Natural selection “contrasts” primarily those substitutions that alter the protein function
  - Nucleotide changes can accumulate more rapidly in the four times degenerate sites and less quickly in non degenerate sites
- The described situation is easily observable in Nature
  - The substitutions that have been accumulated in the genes encoding for human and rabbit β–globins are found mainly at four times degenerate sites (the replacement frequency is very similar to that of the 3’–flanking sequences and, in general, of the regions free from selective constraints)
  - In fact, out of the 47 substitutions that have accumulated in human and rabbit β–**globin** genes, over the last 100 million years, 27 are synonymous and 20 non–synonymous, although the possibility of non–synonymous substitutions is about three times greater

Region	Length (bp)	No. of changes	Substitution frequency
Non degenerate	302	17	0.56
2–degenerate	60	10	1.67
4–degenerate	85	20	2.35

Indel and pseudogenes – 1

- In the case of expressed genes, a strong propensity exists in Nature to counteract insertion and deletion events, because of their tendency to alter the reading frame used by ribosomes
- This trend, which is contrary to the mutations in the coding regions, is so strong that enzymes involved in DNA replication and repair seem to have evolved in such a way as to make the indel events approximately ten times less likely than substitutions, in every region of the genome
- On the other hand, in the case of gene duplication, it may happen that genes, which were originally subject to selective constraints, have become transcriptionally inactive

Indel and pseudogenes – 2

- The genes with new functions commonly arise from existing genes with useful features
- The duplication of an entire gene can allow for a copy of the gene maintaining the original function, while the other is able to disengage from functional constraints and accumulate mutations (in the coding region or in the promoter)
- Sometimes, the mutated copy of the gene is subject to changes that allow it to acquire a new function, crucial to the health of the organism
- More often, however, a copy becomes a pseudogene, which is transcriptionally inactive
- The genomes of mammals are rich in pseudogenes, and their sequences tend to accumulate substitutions at a very high rate, with an average of ~4 substitutions per site per 1 billion years, slightly faster than that of the 3’ flanking regions of the expressed genes

Mutations and substitutions

- The natural selection has an insidious effect on the data available for bioinformatic analyses
- With rare exceptions, in fact, in the populations of organisms found in Nature, the only available alleles (variants of a gene) are those which have not had a detrimental effect on the health of the organisms
  - Changes in the nucleotide sequence of a gene are all possible, but not all are "observable"
  - Difference between the concepts of mutation and substitution
    - Substitutions are changes in the nucleotide sequence which accidentally occur during the process of DNA replication/repair
    - Instead, mutations are substitutions that have already "passed the filter" of natural selection
  - The number of mutations is "easy" to calculate, whereas it is rather difficult to obtain a reliable estimate of the occurred substitutions

Genetic drift and fixation – 1

- Most of the populations of organisms currently present in Nature show a large number of genetic variations
- Humans, for instance, differ from each other, on average, for a pair of bases out of 200
- Different versions of a gene within organisms of a given species are called alleles
- Differences among alleles can…
  - …be relatively harmless (for instance, a single nucleotide mutation in a 3'–flanking sequence)
  - …have dramatic consequences (for example, the presence of a premature stop codon that causes the production of a truncated, nonfunctional protein)
- The change in the relative frequencies of different alleles is just the essence of evolution

Genetic drift and fixation – 2

- With the exception of those alleles introduced by migration or transfer between species (horizontal transport of DNA, not due to inheritance), new alleles come from substitutions that occur in one existent allele of a single member in a population
  - The new versions of the genes initially occur with a very low frequency

    $$q=1/(2N)$$

    being N the number of diploid organisms actively reproductive within the population
  - A neutral allele just arisen because of a replacement in a population of N individuals has a probability $1/(2N)$ to be fixed, and a probability $(2N–1)/(2N)$ to be eliminated

Genetic drift and fixation – 3

- Since the replacement frequencies are generally low and those changes which are crucial for the health of an individual can quickly reach a frequency equal to 0 or 1, how can we explain the relatively high levels of variations found within populations of organisms?
  - Most of the observed variations among individuals has a negligible effect (beneficial or harmful), that tends to be selectively neutral
  - In fact, the genetic drift can lead to the fixation of neutral alleles appeared because of random substitutions

Genetic drift and fixation – 4

- The probability P, that any neutral variant of a gene was eventually lost within a population represents a random event and equals $1-q$, where q is the relative frequency of the allele in the population
- For the same principle, the probability that a particular neutral allele was fixed is q, being q the current frequency of the gene in the population

Genetic drift and fixation – 5

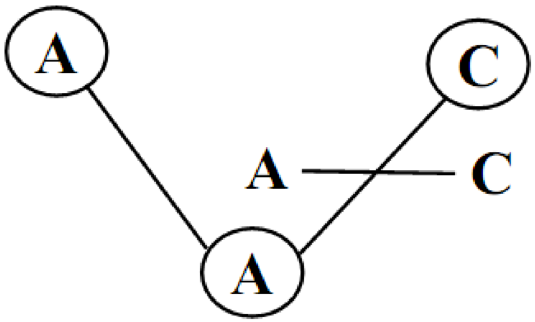
- The comparative analysis between sequences allows molecular biologists to avoid the long and tiresome process of saturation mutagenesis, through which all possible variations of the nucleotide sequence of a gene were produced to determine those capable of altering its function
- Indeed, Nature itself performs a perpetual saturation mutagenesis experiment and most of the observable variants corresponds to changes that do not alter significantly the function of genes

Estimation of the substitution number – 1

- In an alignment, the number of substitutions K between two sequences is the most important variable for the analysis of molecular evolution
- If an "optimal" alignment exists which suggests that there have been relatively few substitutions, directly counting the observable replacements p is a good estimate for K
- Nevertheless, in general, such a direct computation is an underestimate, because of multiple substitutions that may have been occurred with respect to the same nucleotide in the evolutionary path from the last common ancestor

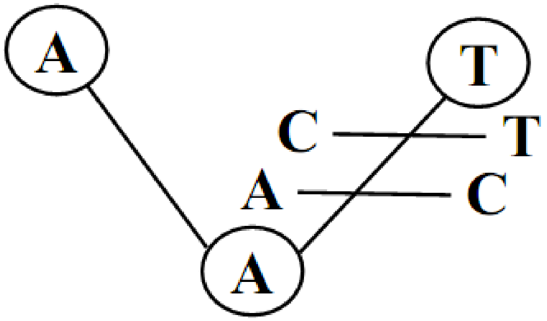
Estimation of the substitution number – 2

Single substitution



1 substitution, 1 difference

Multiple substitution

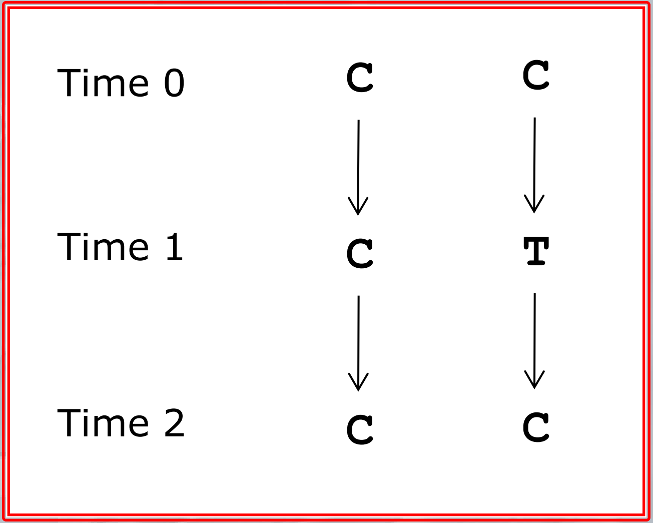


2 substitutions, 1 difference

Underestimation of the number of substitutions ⇨ due to multiple substitutions, the observed distances may underestimate the actual amount of evolutionary changes

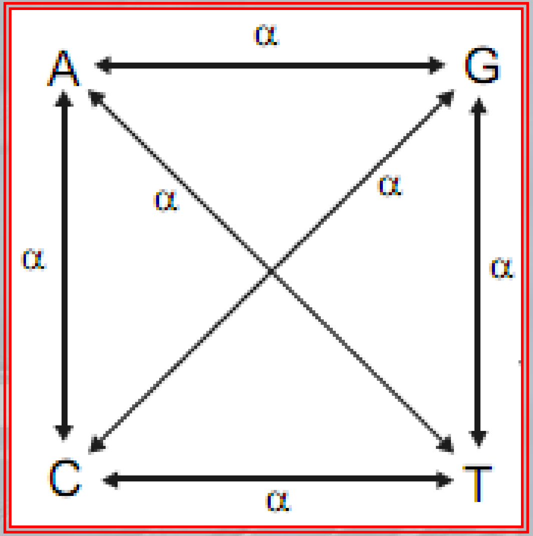
The Jukes-Cantor model – 1

- Where substitutions are common, there is no guarantee that a particular site has not been subjected to multiple changes
- To consider this possibility, T. Jukes and C. Cantor (1969) assumed that each nucleotide had the same probability of being replaced by any other
- Using this assumption, they created a mathematical model in which, if the mutation frequency of a nucleotide with respect to any other nucleotide is α, its overall frequency of replacement is 3α



The Jukes-Cantor model – 2

- In this model if, in a certain position, there is a c at time 0, then the probability $P_{c(1)}$, that the same nucleotide is still present at time 1, is $P_{c(1)}=1-3\alpha$
- Since, if the original c mutates into another nucleotide during the first time step, a reversion (or a reverse mutation) to c may occur at time 2, the probabilility $P_{c(2)}$ would be $(1-3\alpha)P_{c(1)} + \alpha(1-P_{c(1)}) = 12\alpha^2-6\alpha+1$



- Passing from discrete to continuous time, it can be shown that, at a given time t, the following relation holds:

$$P_{c(t)} = 1/4 + (3/4)e^{-4\alpha t} \quad \text{(for } t{=}1, \sim 1-3\alpha)$$

The Jukes-Cantor model – 3

- Indeed, using a formalization of the method based on the punctual substitution probability matrix, we have:

$$\mathrm{R} = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

  with r_{ij} that represents the substitution probability of nucleotide j with nucleotide i

- Let P(t) be the evolutionary matrix, where the elements p_{ij} are the probabilities of finding, in a certain site and at time t, the nucleotide i, where there was j at time 0

The Jukes-Cantor model – 4

- The evolutionary matrix P constitutes the solution of the differential equation

$$d\mathrm{P}(t)/dt = \mathrm{P}(t)\mathrm{R}$$

or, element by element,

$$dp_{ij}(t)/dt = \sum_{k=1}^{4} p_{ik}(t)r_{kj}$$

from which, it follows that:

$$\mathrm{P}(t) = exp\{\mathrm{R}t\} = \sum_{k=0}^{\infty} (\mathrm{R}t)^k/k!$$

- Therefore, the elements of P are defined by

$$p_{ij}(t) = \begin{cases} 1/4 - (1/4)e^{-4\alpha t} \text{ se } i \neq j \text{ (for } t{=}1, \sim\alpha) \\ 1/4 + (3/4)e^{-4\alpha t} \text{ se } i = j \text{ (for } t{=}1, \sim 1{-}3\alpha) \end{cases}$$

The Jukes-Cantor model – 5

- DNA data became available, for the first time, ten years after the formulation of the Jukes-Cantor (JC) model, and it was immediately apparent that the assumption of global uniformity ($\alpha=1/4$), in the substitution patterns, constituted a raw simplification
- However, their model continues to provide a useful tool for evaluating K, the number of substitutions per site, when multiple substitutions are possible

The Jukes-Cantor model – 6

- Let us observe two sequences whose speciation events dates back t; then the probability that they continue to carry the same nucleotide, C for instance, at a particular site is:

$$P_{C(2t)} = 1/4 + (3/4)e^{-8\alpha t}$$

- Therefore, the fraction of sites that differ between the two sequences is just $p = (1 - P_{C(2t)}) = 3/4(1 - e^{-8\alpha t})$
- Taking logs of both sides, we obtain $8\alpha t = -ln[1-(4/3)p]$
- If the overall frequency of replacement is 3α, the expected proportion of differences between two sequences at any time t is $K = 2t \times 3\alpha = 6t\alpha$

The Jukes-Cantor model – 7

- Then, since $K = 3/4(8\alpha t)$, the JC model allows to evaluate K by the equation

$$K = -3/4\ ln[1-(4/3)p]$$

where p is the fraction of nucleotides that a simple count shows to be different in the two sequences

  - The equation is consistent with the idea that, when two sequences have few non–correspondent sites, p is small, and also the probability that multiple substitutions have taken place in a given site is low
  - Conversely, when there is a significant number of unmatched sites, the actual number of substitutions per site will be much greater than that directly calculated
  - The terms 3/4 and 4/3 account for the presence of four nucleotides that can be replaced in three different ways, all equally probable (not related sequences should correspond for 25% just by chance)

The Jukes-Cantor model – 8

- Indeed, letting $t \to\infty$, the divergence p will approach 3/4, which makes sense: after enough time, the common ancestry of the two sequences has been erased, and 1/4 of all sites will match by chance
- Example
  - If two sequences are 95% identical, they differ for 5% or, in other words, p=0.05, and therefore

    $$K = -3/4\ ln(1-(4/3)0.05) = 0.0517$$
  - Note that the observed dissimilarities of 0.05 is slightly increased, being the estimated distance equal to 0.0517 — this makes sense because, for two very similar sequences, only a small number of multiple changes is expected, given the short divergence time
  - Anyway, if two sequences coincide only for 50%, they also differ for 50%, i.e. p=0.50, and then

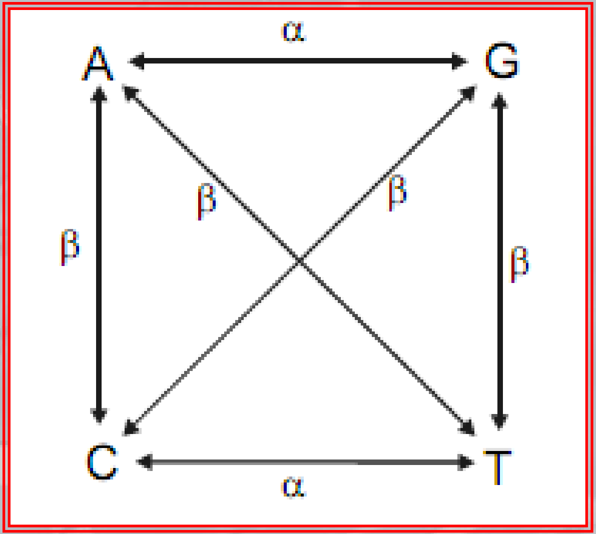
    $$K = -3/4\ ln(1-(4/3)0.5) = 0.824$$

Estimation of the substitution number (cont.)

- To increase the realism of metric models, further parameters must be considered
  - In fact, it is better to use a model that is consistent with the data rather than blindly imposing a model on the data
  - The most common way to enrich the expressiveness of the model suggests enabling different replacement rates for each type of nucleotide change

The Kimura model – 1

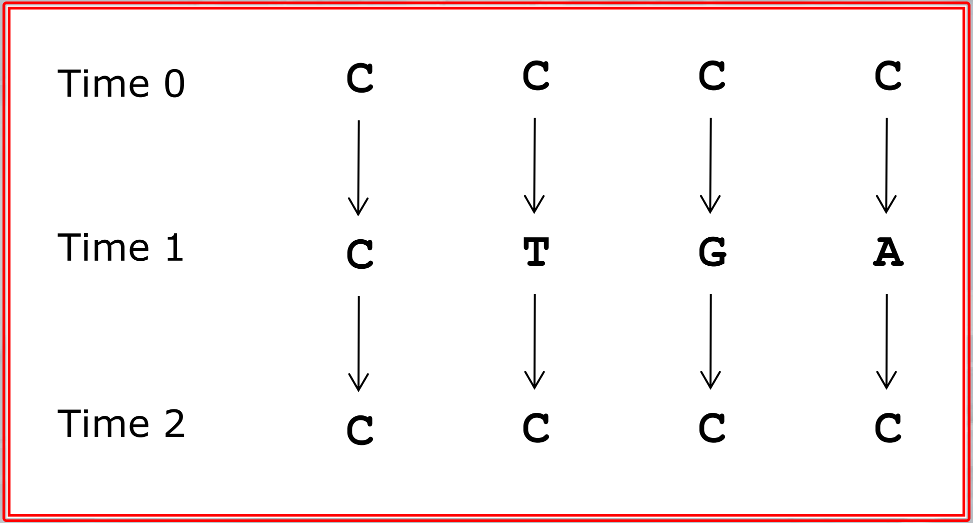
- In 1980, M. Kimura developed a model with two parameters to account for the differences in frequency of transitions and transversions
  - It is assumed that transitions occur with a constant frequency α, while transversions happen with a frequency $\beta \neq \alpha$
  - In Nature, $\alpha \approx 3\beta$
  - If a site within a gene is occupied by c at $t=0$, the probability that, at that site, the same nucleotide still remains at $t=1$ would be



$$P_{\mathrm{cc}(1)} = 1-\alpha-2\beta$$

The Kimura model – 2

- Reverse mutations may occur between t=1 and t=2, and the probability that the considered site still contains C at t=2, $P_{\mathrm{CC}(2)}$, is the sum of the probabilities associated with the four different situations:



Time 0	C	C	C	C
Time 1	C	T	G	A
Time 2	C	C	C	C

that is…

$$P_{\mathrm{CC}(2)} = (1-\alpha-2\beta)P_{\mathrm{CC}(1)} + \alpha P_{\mathrm{TC}(1)} + \beta P_{\mathrm{GC}(1)} + \beta P_{\mathrm{AC}(1)}$$

The Kimura model – 3

- As in the JC model, continuing to expand the recurrence formula for the calculation of the probability of time invariance of a given nucleotide, we obtain

$$P_{\text{cc}(t)} = 1/4 + (1/4)e^{-4\beta t} + (1/2)e^{-2(\alpha+\beta)t}$$

- Using the probability matrix, the Kimura model will be described by:

$$\mathrm{R} = \begin{array}{cc} \begin{pmatrix} 1-\alpha-2\beta & \beta & \alpha & \beta \\ \beta & 1-\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & 1-\alpha-2\beta & \beta \\ \beta & \alpha & \beta & 1-\alpha-2\beta \end{pmatrix} & \begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \\ \begin{array}{cccc} \mathbf{A} \qquad\qquad & \mathbf{C} \qquad\qquad & \mathbf{G} \qquad\qquad & \mathbf{T} \end{array} & \end{array}$$

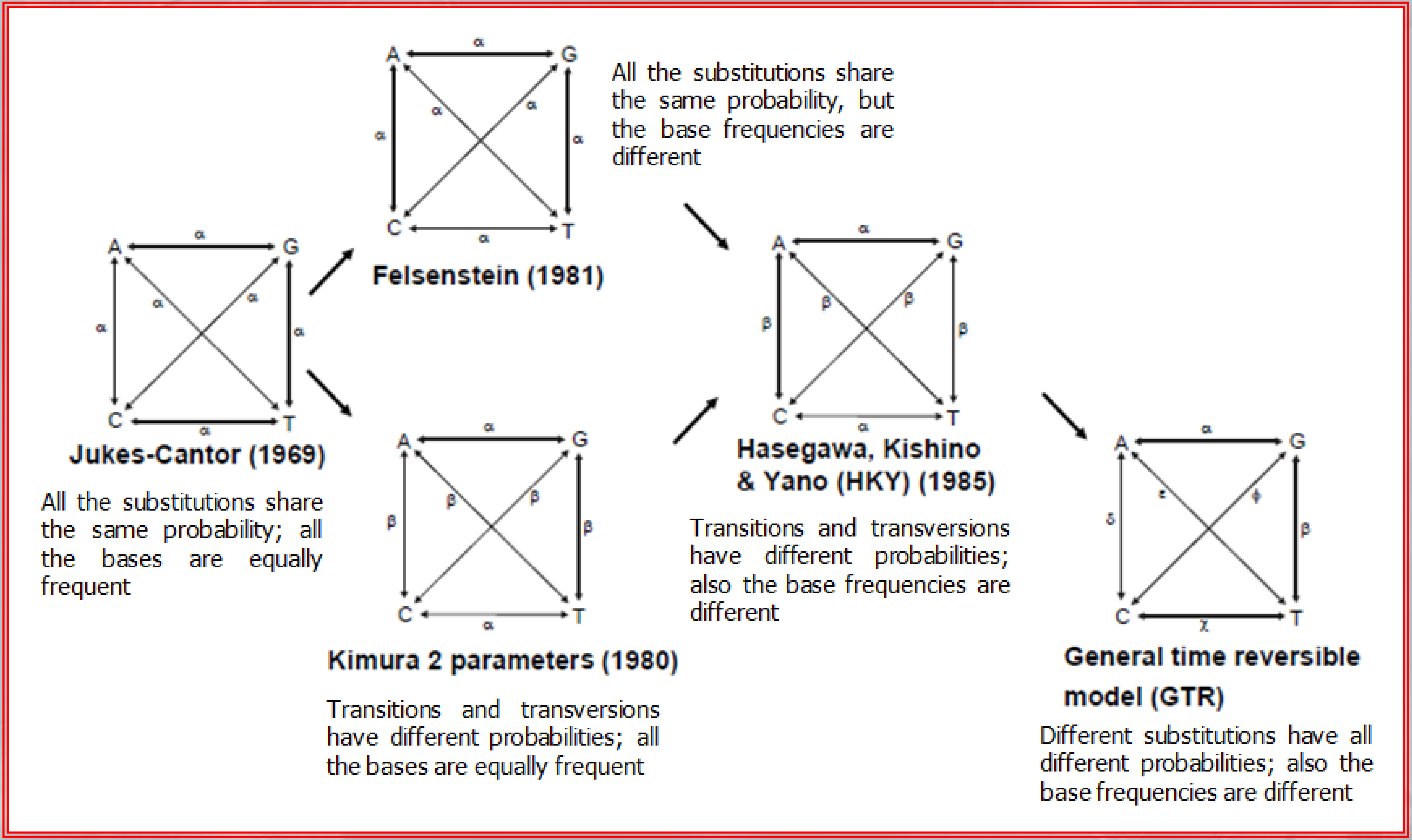
The Kimura model – 4

- The symmetry of the substitution scheme ensures that all the nucleotides share the same probability to remain *in situ* between time 0 and any time t in the future ($P_{\mathtt{GG}(t)} = P_{\mathtt{AA}(t)} = P_{\mathtt{TT}(t)} = P_{\mathtt{CC}(t)}$)
- We can derive the following estimation for K

$$K = 1/2\ ln[1/(1 - 2P - Q)] + 1/4\ ln[1/(1 - 2Q)]$$

  where P represents the fraction of nucleotides that can be directly counted as transitions, while Q counts transversions
  - If no distiction is made between transitions and transversions, placing $p = P + Q$, we obtain again the estimation given by the JC method

Evolution of K estimation models



Felsenstein (1981)

All the substitutions share the same probability, but the base frequencies are different

Jukes-Cantor (1969)

All the substitutions share the same probability; all the bases are equally frequent

Hasegawa, Kishino & Yano (HKY) (1985)

Transitions and transversions have different probabilities; also the base frequencies are different

Kimura 2 parameters (1980)

Transitions and transversions have different probabilities; all the bases are equally frequent

General time reversible model (GTR)

Different substitutions have all different probabilities; also the base frequencies are different

Multiparametric models – 1

- The large amount of DNA data generated from the '80s, revealed that the Kimura assumption, which assigns different probabilities for transitions and transversions, is still significantly far from what happens in Nature
- Since each nucleotide can in fact be replaced by any one of the other three, twelve different types of substitution are possible,

  A→C A→G A→T C→A C→G C→T

  G→A G→C G→T T→A T→C T→G

  to which different probabilities can be assigned, just producing a model with 12 parameters

Multiparametric models – 2

- An example of a scoring matrix (for the relative frequencies of nucleotide substitution in the repeated sequence of Alu–Y in the human genome) is givev by:

	A	T	C	G
A	–	4.0	4.6	9.8
T	3.3	–	10.4	2.7
C	7.2	17.0	–	6.2
G	23.6	4.6	6.0	–

The members of the Alu family are repetitive DNA, approximately 260 base pairs long; they are derived from a single or a small number of ancestral sequences, that have been duplicated almost a million times during the evolution of primates

- A thirteenth additional parameter may also be used, to compensate for the differences between what is described by the scoring matrix and the (observable) trend associated to the replacement of the regional genomic context **GC**

Multiparametric models – 3

- However… simulation studies indicate that the simpler models (with one or two parameters) often provide more reliable results with respect to multiparametric models, because…
  - …they do not require large amounts of data to estimate the relative frequencies of substitution (without the introduction of sampling errors)
  - …they are, in fact, virtually indistinguishable for closely related sequences

Substitutions in protein sequences – 1

- The proportion p of different amino acids within two protein sequences can be “observed” as well as for the nucleotide sequences (and evaluated as the ratio between the points of mismatch and the length of the sequences)
- However, exactly determining the number of substitutions that occurred in the evolution of two or more proteins is generally a more complex operation with respect to that on the corresponding DNA sequences (a single amino acid substitution can correspond to a variable number of substitutions in the encoding nucleotide sequence)
- As well as for DNA sequences, the observed mutations represent an underestimation of the substitutions actually occurred during the evolution

Substitutions in protein sequences – 2

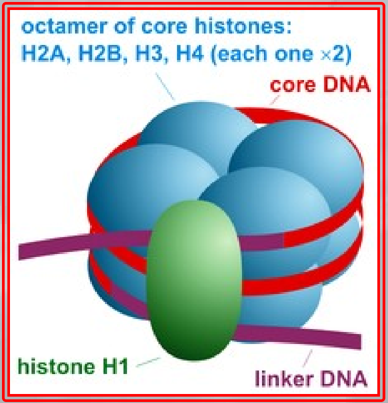
- In addition:
  - Some substitutions occur more frequently than others
  - The path that leads to the substitution of two amino acids has not always the same length
    - Example: **CCC** that codifies for proline can be converted into **CUC**, for leucine, with a single substitution, but it can be converted to **AUC**, for isoleucine, with two substitutions
  - Amino acid substitutions do not all have an equivalent effect on the protein function and, also, possible effects differ in distinct contexts
  - Weigh each amino acid substitution in a different way, according to estimates based on empirical data, using a PAM–like matrix

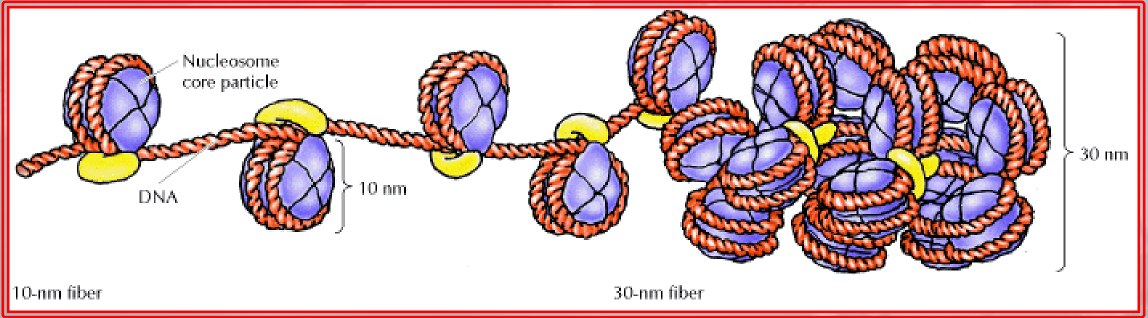
Evolutionary speed variations – 1

- Changes in evolutionary rates are visibly recognizable by comparing different regions within the same gene, as well as significant differences can be observed in the speed of evolution among different genes
- Not considering possible fluctuations due to sampling errors in small populations, differences in the evolution speed are imputable to two main factors:
  - differences in the replacement frequency
  - the effect, in quantitative terms, of natural selection on the *locus*
- Specific examples of two classes of genes, which encode for histones and apolipoproteins, illustrate the effects of different functional constraints that affect the evolutionary speed

Evolutionary speed variations – 2

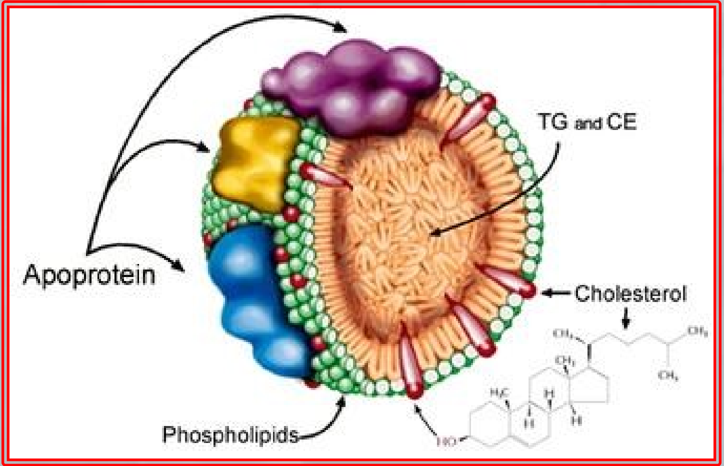
- Example 1
  - Histones are positively charged, basic proteins that bind to DNA and are present in all eukaryotes
  - The majority of amino acids belonging to a histone interacts directly with the negatively charged DNA
    - Any change in the histone amino acid sequence may affect its ability to interact with DNA
    - Histones are evolving very slowly
    - The yeast histone H2A can be substituted with its human homologue without side effects, although speciation has produced millions of years of independent evolution





Evolutionary speed variations – 3

- Example 2
  - Apolipoproteins accumulate non–synonymous substitutions at a very high frequency
  - They are responsible for the non–specific interaction with a variety of lipids and for their transport in the blood of vertebrates
  - Their binding sites with the lipid are mainly composed of hydrophobic amino acids
    - Each hydrophobic amino acid (f.i., leucine, isoleucine and valine) works equally well



Evolutionary speed variations – 4

- Although nucleotide substitutions, in many genes, are generally deleterious, in some cases, natural selection actually favors variability
- Example
  - Genes associated with the antigen (a macromolecule capable of reacting with the products of the immune system) of human leukocytes, HLA, are highly prone to evolutionary diversification
  - In the human population, about 90% of individuals receive different sets of HLA genes from their parents, and it can be estimated that, for a sample of 200 individuals, 15 to 30 different alleles exist
  - High diversity levels, in the specific case, are favored by natural selection, because the number of vulnerable individuals to a given viral infection is much smaller than in the case of a single immune system

Evolutionary speed variations – 5

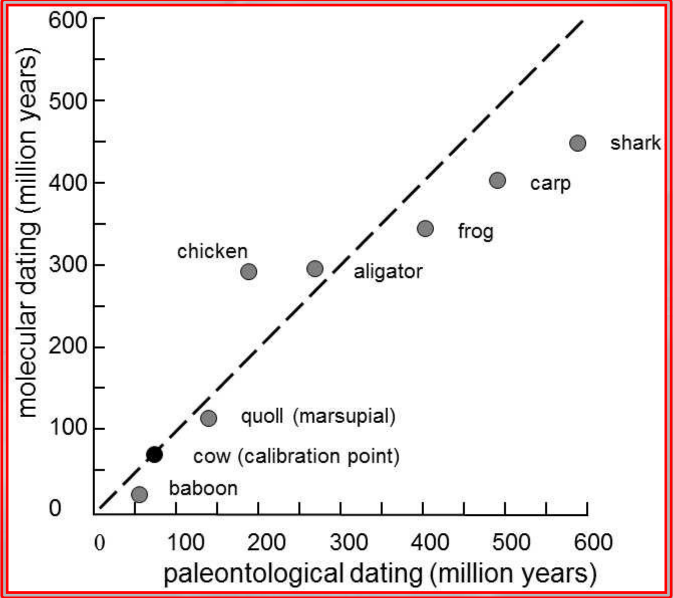
- While host populations are under evolutionary pressure to maintain their diverse immune systems, viruses are under similar pressure to evolve rapidly
- A replication that tends to errors, coupled with the natural selection which favors diversification, causes the frequency of nucleotide substitutions in the NS (non–structural) flu genes to be equal to 1.9×10^{-3} (substitutions per site per year), a million times greater that the frequency of synonymous substitutions in the representative genes of mammals

Molecular clocks – 1

- The idea of dating evolutionary events through the calibrated differences among proteins was expressed, for the first time, by E. Zuckerkandl and L. Pauling in 1962; they actually revealed that the speed of molecular evolution for $loci$ with similar functional constraints is almost constant over long time periods
- Based on some observations, made on different globins, Zuckerkandl and Pauling postulated that the genetic difference between two species, expressed by their amino acid sequences, is a “linear” function of their divergence time (from a common ancestor)
- The verification of this statement was then obtained by comparing homologous protein sequences and, therefore, their rates of amino acid substitutions, for different species, with divergence times estimated based on fossils

Molecular clocks – 2

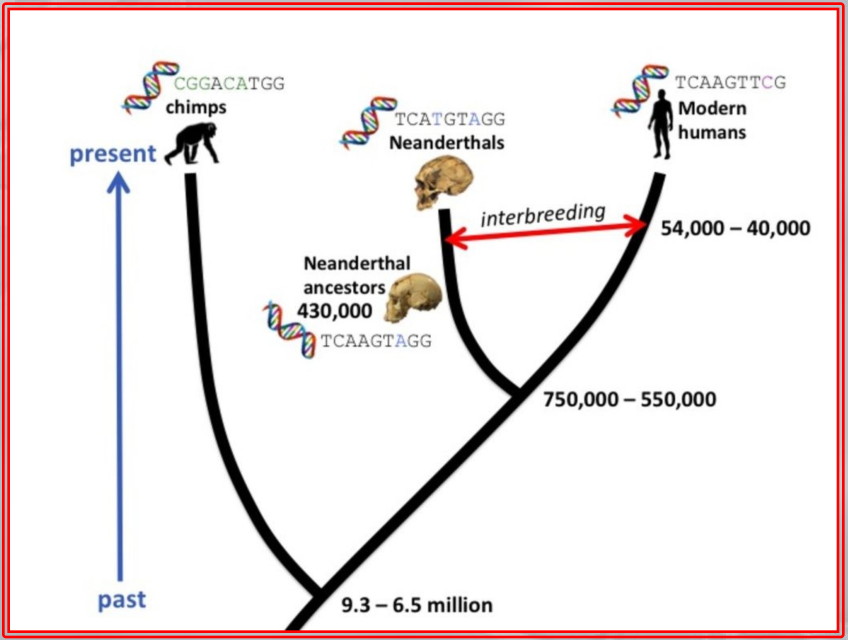
- The replacement frequencies in homologous proteins were so constant over many tens of millions of years, to suggest a direct comparison between the accumulation of amino acid changes and the constant ticking of a molecular clock
- The molecular clock can "beat" at different speeds for distinct proteins, but the number of beats between homologous proteins looks linearly correlated with the amount of time passed from the speciation event, which made them differ in their evolutionary path



The age of the branching-off of the line leading to humans from the line leading to the other vertebrates, determined on the basis of paleontological data, is plotted on the abscissa, while the same age determined on the basis of accumulation of substitutions in the gene for α–globin is plotted on the ordinate; it is apparent from the graph that, for most of the studied vertebrates, reasonable agreement exists between the dated moments of branching off of the individual species on the basis of paleontological and biomolecular data

Molecular clocks – 3

- According to the molecular clock hypothesis, therefore, both genes and gene products evolve with rates that are approximately constant over time and along the different evolutionary paths
- Then, if the genetic divergence regularly accumulates during time, it is possible to infer divergence times even in the absence of fossil evidence
- In practice, a constant frequency of variation would facilitate not only the determination of phylogenetic relationships between species, but also the calculus of the divergence time, as well as the radioactive decay of ^{14}C is used to estimate the geological timing

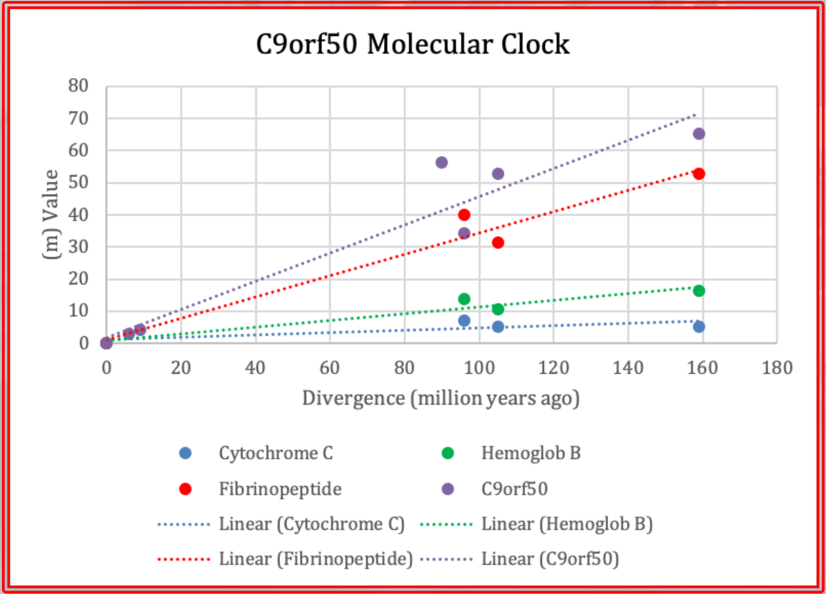


Molecular clocks – 4

- Since then, the validity of the universal molecular clock hypothesis was deeply and extensively discussed
  - In 1965, E. Mayr stated that: "*Evolution is too complex and too variable a process, connected to too many factors, for the time dependence of the evolutionary process at the molecular level to be a simple function*"…
  - …while classical evolutionists argued that the apparent irregular morphological evolution was incompatible with a constant rate of molecular changes
- Initially, a protein molecular clock was theorized, since during the '60s, DNA data were still too scarce, and intense was the debate until the '80s, which led to questioning the very essence of Zuckerkandl and Pauling idea, namely the constancy of the evolutionary speed

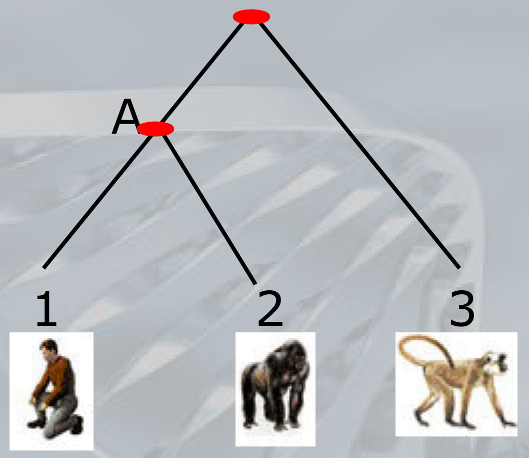
Molecular clocks – 5

- Actually, since 1971, it has become clear that different proteins evolve at widely varying rates
- As a result, the chance to observe an universal protein clock was instantly abandoned
- Statistical tests conducted by Ohta and Kimura (1971), Fitch (1976), Gillespie and Langley (1979) have yielded conflicting results, suggesting that the protein molecular clock hypothesis must be rejected for most proteins, with respect to both vertebrates and invertebrates



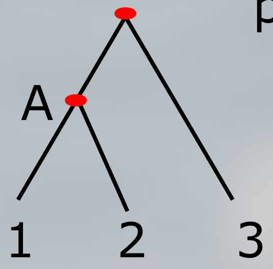
Molecular clocks – 6

- However, most of the divergence dates used in the studies of molecular evolution comes from the interpretation of fossil records, both incomplete and inaccurate
- In order to avoid any question about the dates of speciation, Sarich and Wilson (1973) proposed a method to estimate the overall rate of substitution in different lineages, regardless of the knowledge of their divergence times
- Example
  - To determine the relative substitution frequency in the lineage of species 1 and 2, a species 3 must be defined, similar but less correlated, that plays the role of the outer group
  - Humans and gorillas ⇨ *outgroup*: baboons



Molecular clocks – 7

- Example (cont.)
  - It is assumed that the number of substitutions between any two species is equal to the sum of the number of substitutions present along the branches of the related phylogenetic tree



$$d_{13} = d_{A1} + d_{A3}$$
$$d_{23} = d_{A2} + d_{A3}$$
$$d_{12} = d_{A1} + d_{A2}$$

where d_{13}, d_{23}, d_{12} are “observed” quantities that measure, respectively, the differences between species 1 and 3, 2 and 3, 1 and 2

- The divergence occurred between species 1 and 2, since they shared the last common ancestor, can then be evaluated as

$$d_{A1} = (d_{12} + d_{13} - d_{23})/2$$
$$d_{A2} = (d_{12} + d_{23} - d_{13})/2$$

- By definition, the moment in which the two species began to diverge is the same
  - ⇨ Hp. Molecular clock: d_{A1} and d_{A2} coincide

Molecular clocks – 8

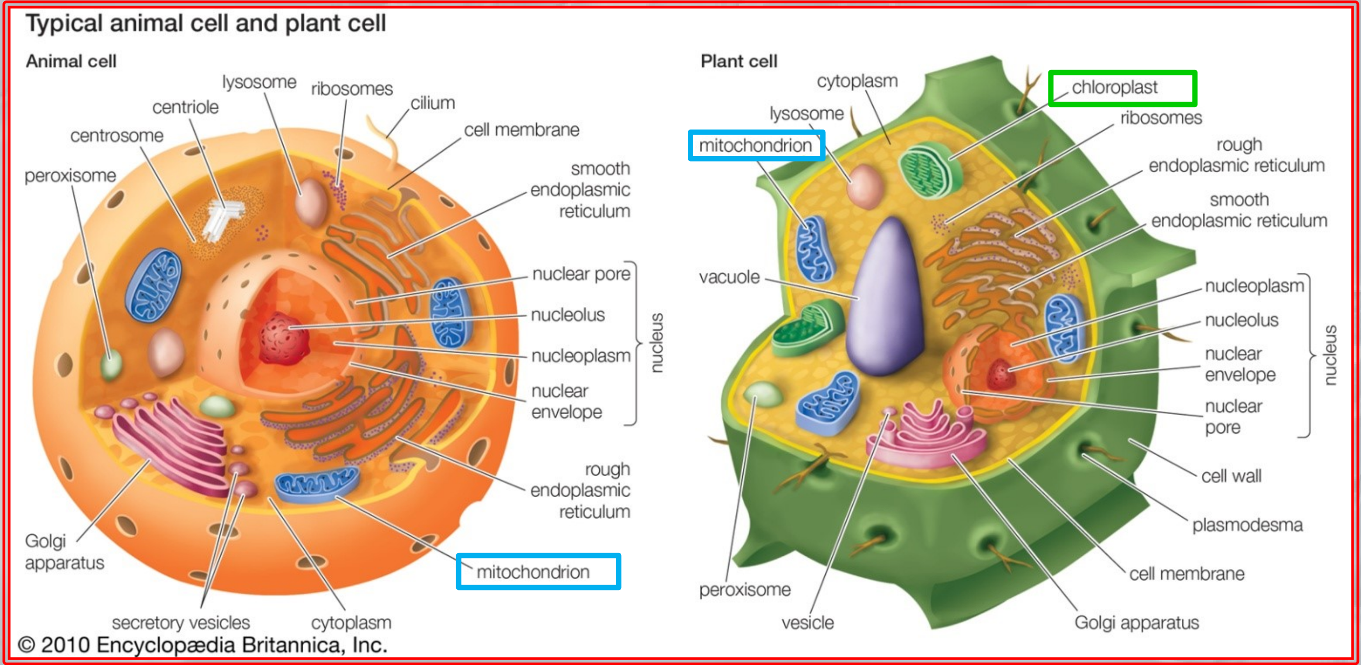
- The amount of data available to test the molecular clock hypothesis is growing exponentially
- The substitution frequencies in rats and mice were established to be very similar
- In contrast, the molecular evolution of man and ape (e.g. gorillas) shows a speed equal to a half of that relative to the Old World monkeys (e.g. baboons), since their speciation
- Moreover, some tests performed on the relative substitution frequency of homologous genes in mice and humans indicate that rodents have accumulated a number of substitutions which is double compared to that of primates, since the last common ancestor (speciation of mammals, occurred 80–100 million years ago)
  - The molecular clock is not constant: the use of molecular divergence for dating the speciation time of two species only makes sense if the species “share the clock”

Molecular clocks – 9

- Causes of the frequency changes in progeny
  - Diversity in generation times (duration of the breeding season)
  - Average repair efficiency, metabolic rate
  - Need to adapt to new ecological niches
  - Population size (genetic drift is stronger in small populations, and, therefore, more mutations are effectively neutral)
  - Changes in the intensity of natural selection
- Changes are difficult to be quantified:
  - We know the current differences
  - We know that, at the divergence time, the organisms shared similar attributes...
  - …but we have actually a little information on their differences throughout the course of evolution

The evolution in organelles – 1

- Within the eukaryotic cell, some different organelles are present, which perform diversified functions necessary for the cell survival
- The organelles, together with the cytosol, form the cytoplasm



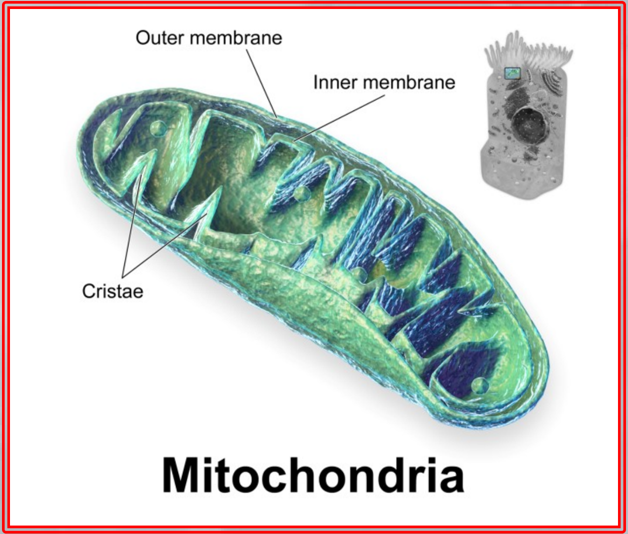
© 2010 Encyclopædia Britannica, Inc.

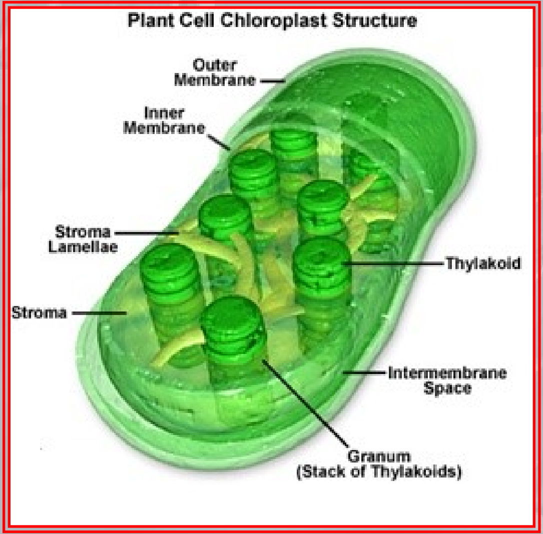
The evolution in organelles – 2

- Organelles are structures equipped with a membrane
- Theories on the birth of some organelles are various and controversial
- A highly accredited one claims that they originated from independent (prokaryotic) organisms phagocytosed by cells in prehistoric times, which began to live in symbiosis with the cells themselves until they became part of them
- In favor of this theory, many organelles, such as mitochondria and chloroplasts, are capable of dividing independently and also have their own DNA

The evolution in organelles – 3

- The average length of mitochondrial DNA (mitochondria are organelles, which serve for the production of energy, and are present, for instance, in the cytoplasm of the cells of animals having an aerobic metabolism) of mammals, abbreviated in mtDNA, is approximately 16000 base pairs
- Conversely, the DNA of chloroplasts (organelles found in plant cells and eukaryotic algae, within which the process of photosynthesis takes place) varies in length between 120000 to 220000 base pairs

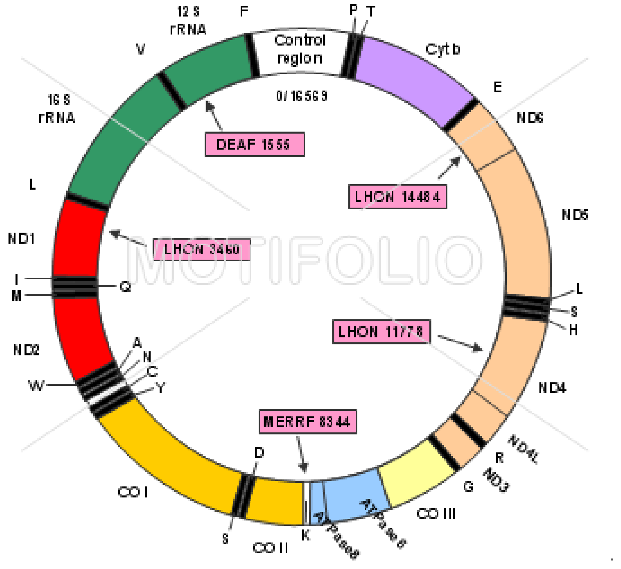




The evolution in organelles – 4

- The unique circular chromosome of both organelles contains genes encoding proteins and RNAs which are essential for their function
- The relatively small size of the chromosomes present in both mitochondria and chloroplasts and the unusual pattern of inheritance (mitochondria are a maternal contribution only in mammals) make them interesting objects for the molecular evolution studies

Human mitochondrial DNA (mtDNA) showing the location of several mutations that cause mitochondrial diseases



The evolution in organelles – 5

- The high mutagens concentration present within the mitochondria (especially oxygen free radicals) submits the mtDNA at a mutation frequency equal to ten times that of the nuclear DNA (in the same cells)
  - The mtDNA is used to study the evolutionary relationships among populations of closely related organisms
  - However, it is not very useful for species that have diverged more than 10 million years ago, because of multiple (un-observable) substitutions which have probably been occurred at each site
- In contrast, in the cpDNA, substitutions are accumulated very slowly: The number of non/synonymous substitutions represents about a fifth of the substitutions observed for the nuclear genes of the same species
- cpDNA can be used to study evolution w.r.t. slightly related organisms

Concluding… – 1

- DNA, like any other molecule, accumulates chemical damages over time
- When such damages, or an error in the DNA replication process, determine a change of the information content of a DNA molecule, it is said that a mutation is occurred
- In other words, substitutions are changes in the genetic material (DNA and more rarely RNA) of an organism
  - They can arise spontaneously or be induced by particular physical or chemical agents said, in fact, mutagenic
  - If they are not properly recognized and repaired by the DNA repair system, they can be permanently fixed in the genome and inherited by later generations
- Substitutions can have an effect, either positive or (more frequently) negative, or be neutral

Concluding… – 2

- Substitutions provide, in practice, the “raw material” on which the evolution acts
  - They create the necessary condition of genetic variability within a population, on which the processes of genetic recombination work, forming the different allelic combinations of each individual; these combinations can finally be subjected to different evolutionary processes that alter the frequencies of various alleles
- Finally, the natural selection process causes many losses in the pool of genes, and those changes that are “fixed” are called mutations
- The substitution frequency can be used as a measure of the functional importance of a gene or of a genome portion

Concluding… – 3

- The sequences are much more “stable” when a substitution may cause an unfunctional protein with deleterious consequences for the life of the organism
- To estimate the number of substitutions that led to the current divergence of two homologous sequences, several parametric models were developed, that consider the possibility of multiple substitutions at a given site
  - Models with few parameters are more robust and computationally simpler
- However, since some genes accumulate replacements faster than others, tests based on relative frequency show that different organisms can have different evolutionary characteristics, even when considering genes with similar functional constraints