MODELS OF SEQUENCE EVOLUTION I
AND NOTES ON THE JUKES CANTOR MODEL

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@uniroma1.it

CB_23_24_L6 –L8, Rome 5th ,9th and 10th of oct 2023

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Preliminary observation

- Comparing nucleotide sequences (genetic material) 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 (suggesting a **neutral, non selective drift**)
- 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 problem of **remote homology**)
- The analysis of both the number and the type of substitutions, that have been occurred during the evolution, are of central importance for the study of **molecular evolution ----> BIOINFORMATICS & MOLECULAR EVOLUTION (higgs&attwood)**

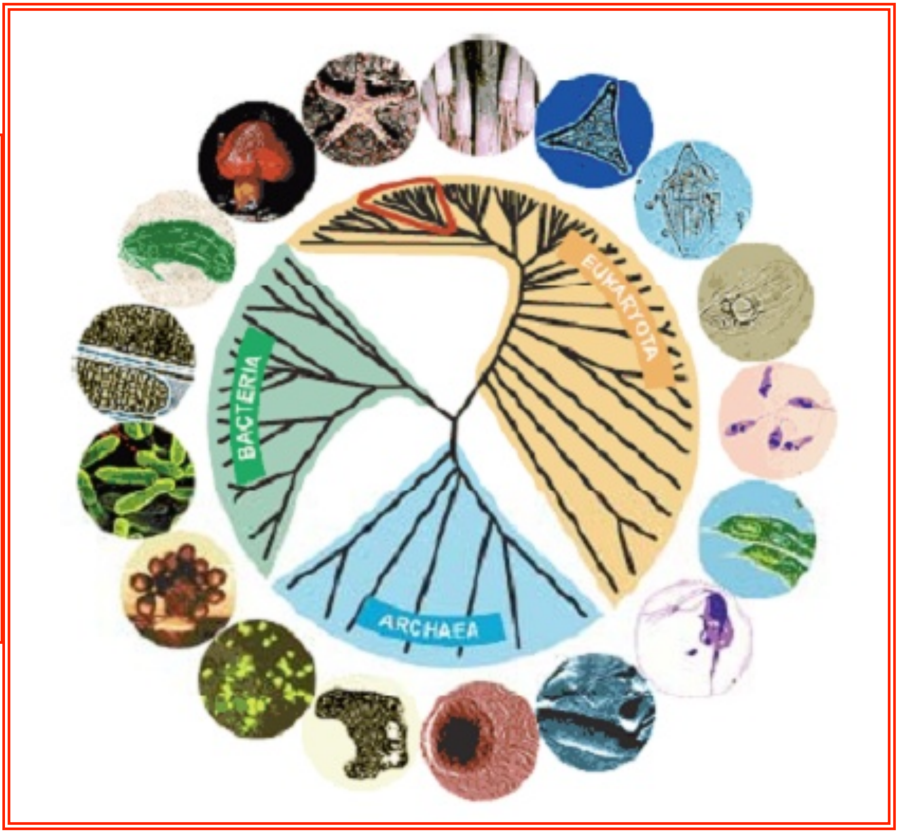
ABOUT MOLECULAR EVOLUTION

- *DNA molecules are not only the key to heredity, but they are "document of evolutionary history"* (E. Zuckerkandl)---> **DNAs, genomes are the ARCHIVES of Evolution (just fancy idea?)**
- Molecular evolution integrates evolutionary biology, molecular biology, and population genetics
  - It describes the process of evolution (changes in time) of **DNAs, mRNAs, tRNAs, ncRNAs and proteins**
  - It includes the study of **rates** of sequence change, relative importance of **adaptive** and **neutral** changes, and changes in **genome structure (e.g. chromatin structure, Hi-C maps)**
  - 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 (the pioneer has been **Carl Woese, 1928-2012**). See the movie **by Nigel Goldenfeld**

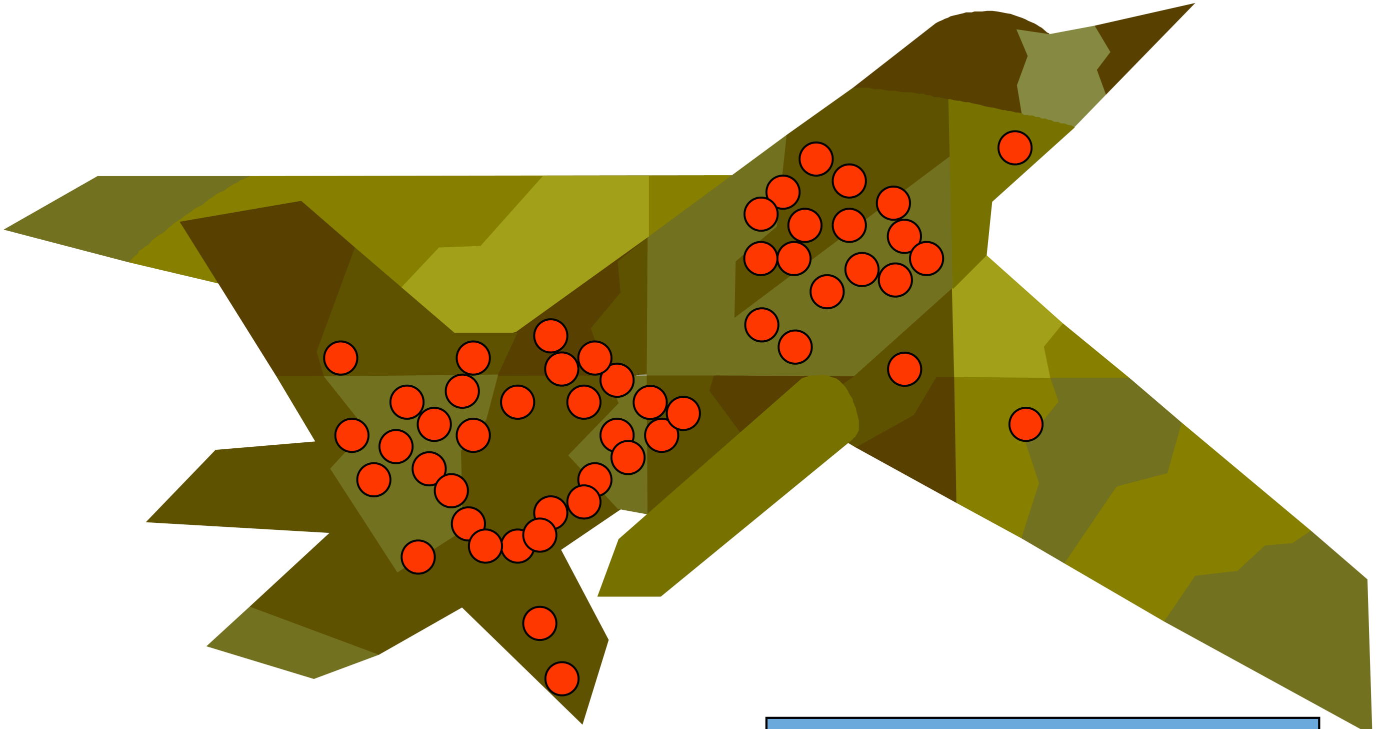
https://www.bing.com/videos/search?q=carl+woese&docid=608031837571450660&mid=B9395F71053978B504BFB9395F71053978B504BF&view=detail&FORM=VIRE

Molecular evolution and biological diversity
The **tree** (graph, network) **of life**



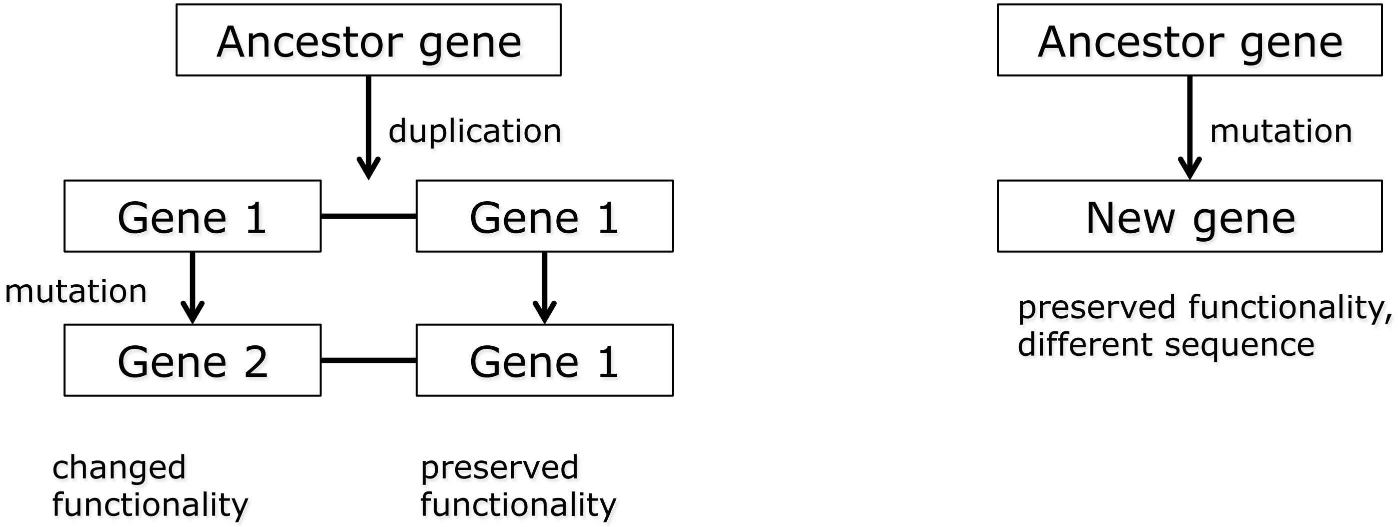
- The **process** of natural selection is truly effective in removing harmful (**fitness** reducing) changes, molecular evolution also serves to recognize and characterize the genome portions that are **more important (conserved, invariant, subject to purifying selection)** from the functional point of view
- …the rates (see then the Jukes–Cantor model) of nucleotide substitutions are 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”)

Evolution in a nutshell (rememberingAnna Tramontano)



Manguel M, Samaniego F.J.,
Abraham Wald's Work on Aircraft Suvivability,
J. American Statistical Association. 79, 259-270, (1984)

Genes and proteins - 1



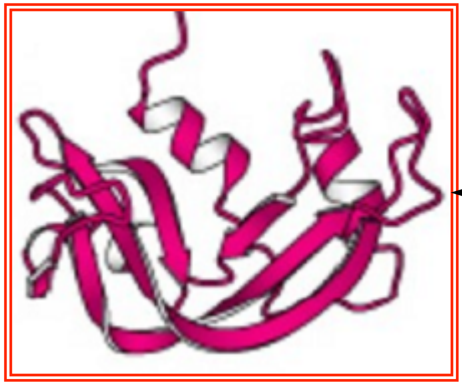
Paralogy/orthology/synteny (topology)
Define these terms

Genes and proteins - 2

- 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

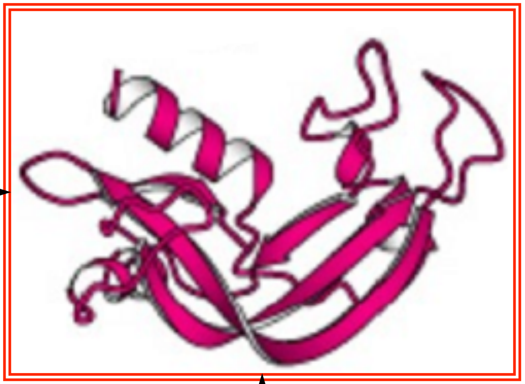


Bovine ribonuclease
(digestive enzyme)

Orthologous genes

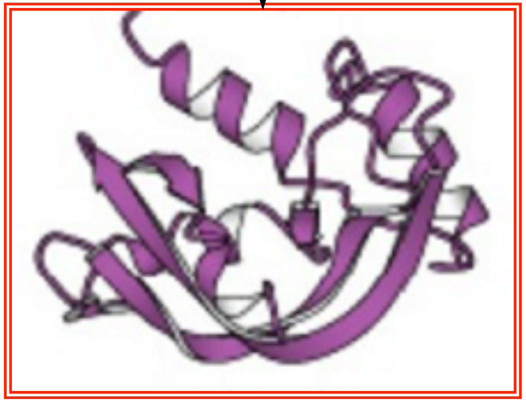
speciation

Human ribonuclease
(digestive enzyme)



duplication

Paralogous genes



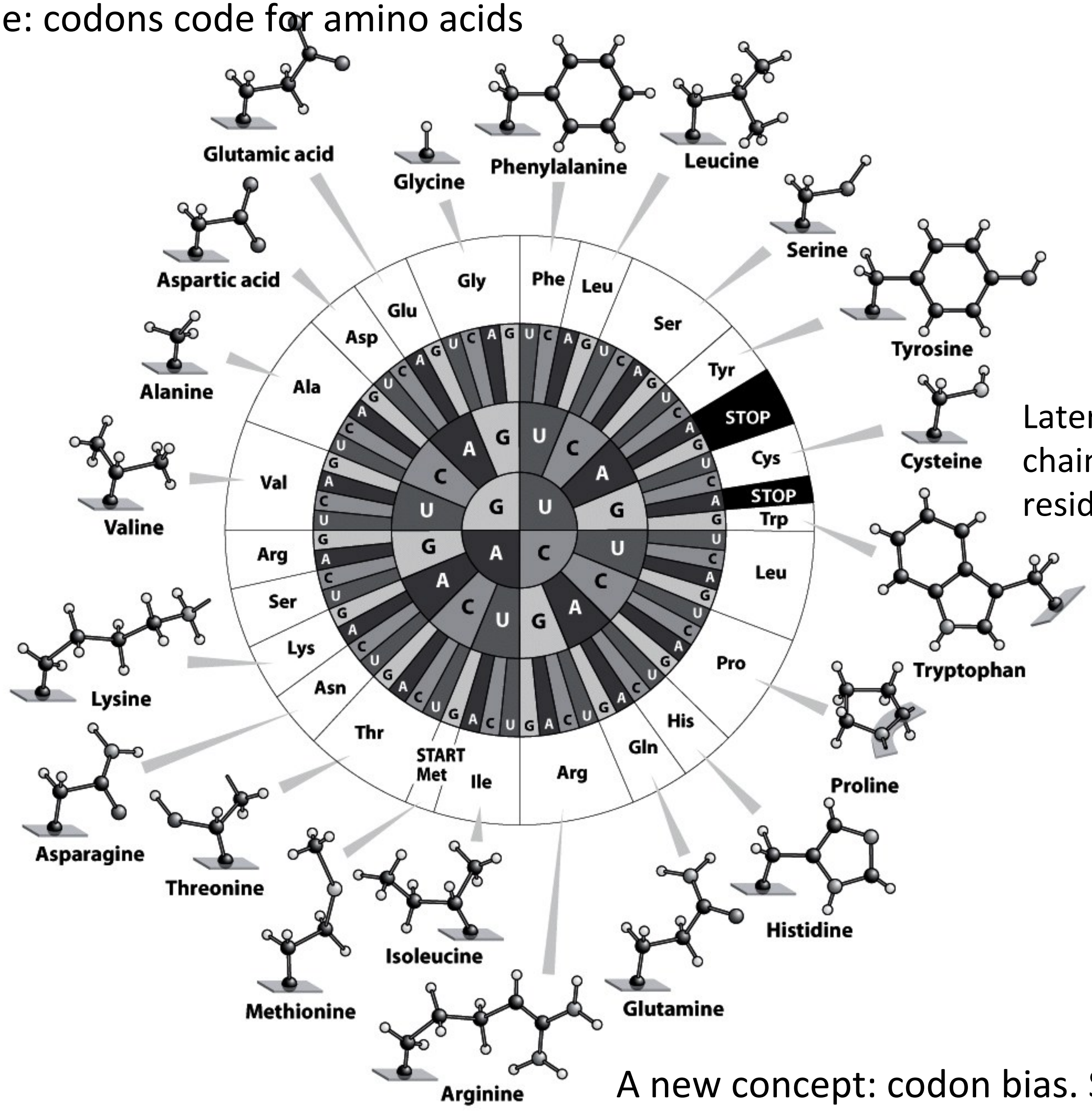
Angiogenin
(It stimulates the growth of blood vessels)

How proteins (phenotypes) change

- 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
  - What about **transposons?**

- The genetic code is **redundant (-->codon bias)** 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, the amino acids change, and the protein can become "unrecognizable" and definitely loses its function

The genetic code: codons code for amino acids



Lateral chains: or residues

A new concept: codon bias. Slow/fast codon

Figure 1.4 Physical Biology of the Cell (© Garland Science 2009)

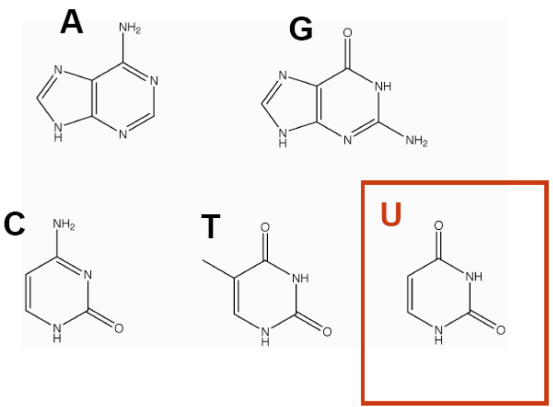
TYPES OF MUTATIONS

(A and G purines; C and T/U are pyrimidines)

Causes of mutations: DNA damage, errors in the replication

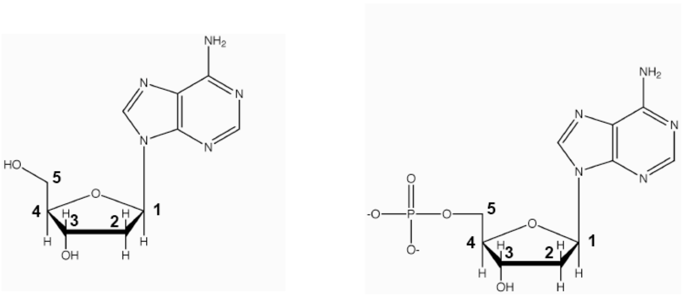
Transitions/ transversions

DNA structure



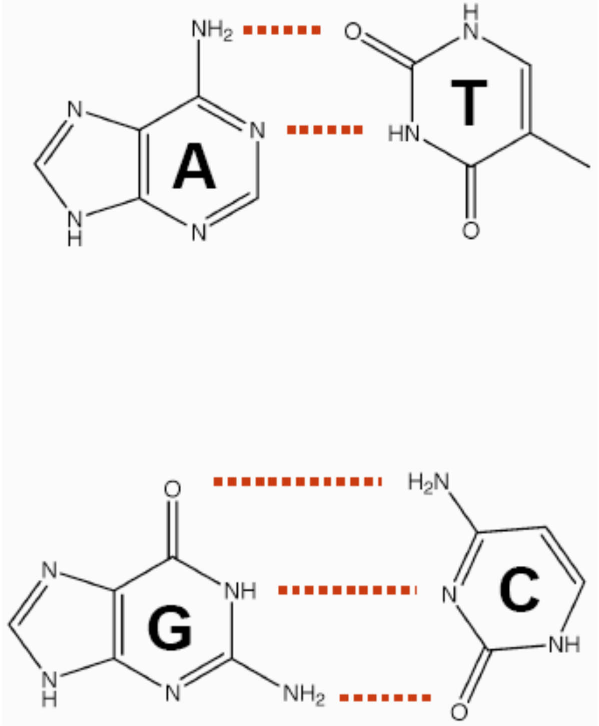
Bases

DNA structure



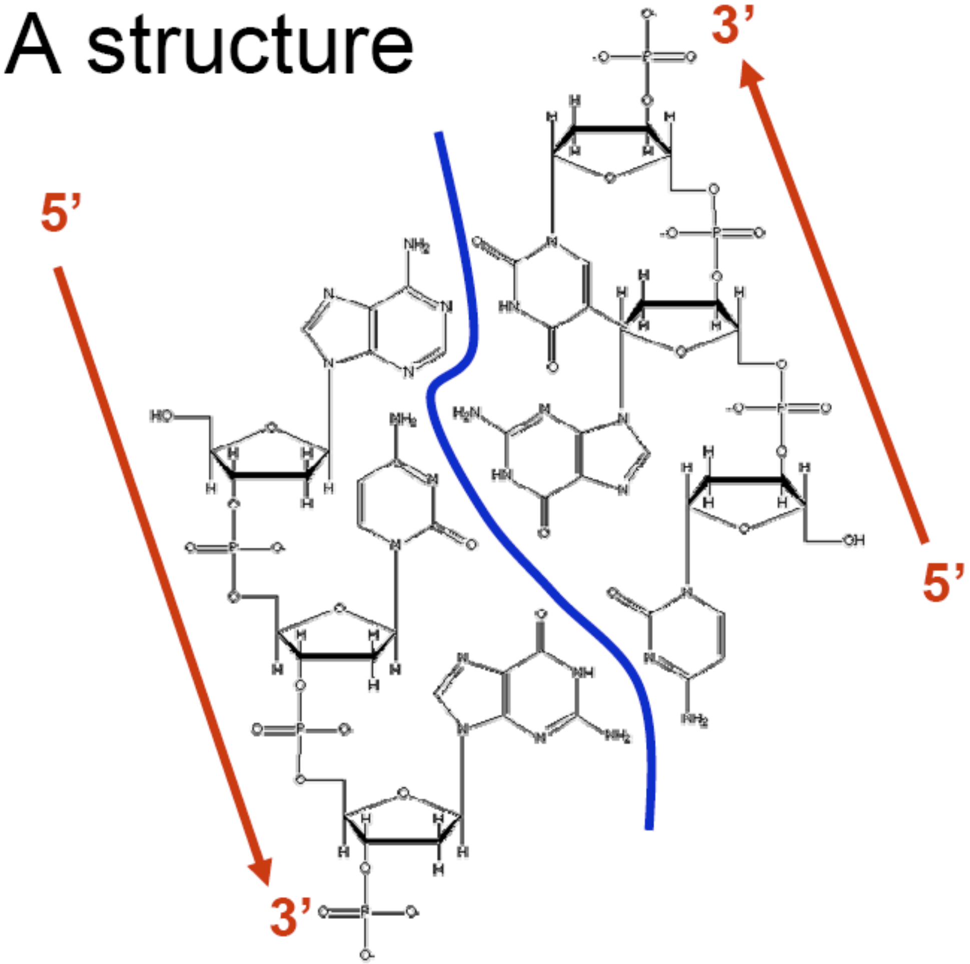
Nucleoside **Nucleotide**

DNA structure

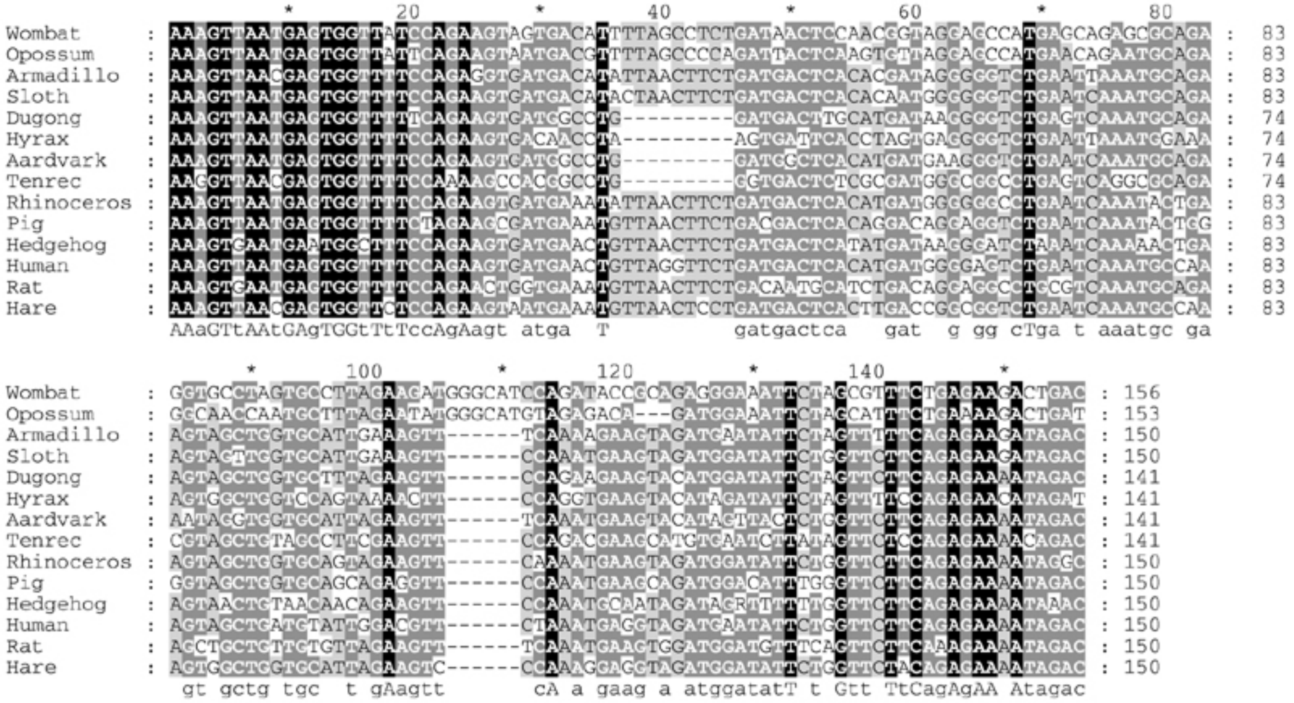


Base pairing

DNA structure



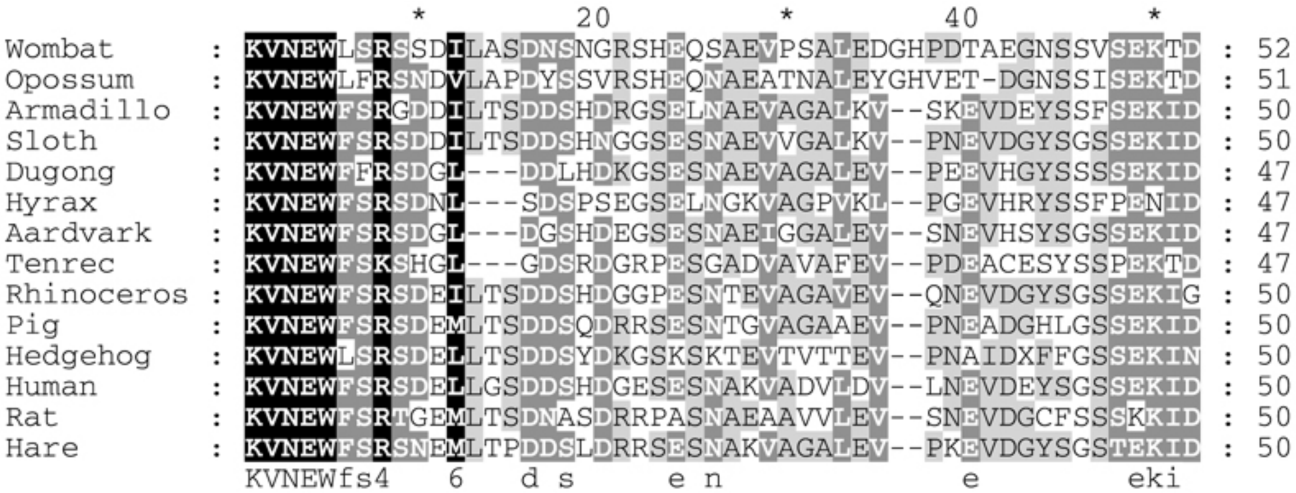
Part of the alignment of the DNA sequencees of the BRAC1 gene (fig.3.1)



```
                      *        20         *        40         *        60         *        80
Wombat     : AAAGTTAATGAGTGGTTATCCAGAAGTAGTGACATTTTAGCCTCTGATAACTCCAACGGTAGCACCCATGAGCAGAGCGCAGA :  83
Opossum    : AAAGTTAATGAGTGGTTATTCAGAAGTAATGACGTTTTAGCCCCAGATTACTCAAGTGTTAGCACCCATGAACAGAATGCAGA :  83
Armadillo  : AAAGTTAACGAGTGGTTTTCCAGAGGTGATGACATATTAACTTCTGATGACTCACACGATAGGGGTCTGAATTAAATGCAGA :  83
Sloth      : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGACATACTAACTTCTGATGACTCACACAATGGGGGTCTGAATCAAATGCAGA :  83
Dugong     : AAAGTTAATGAGTGGTTTTTCAGAAGTGATGGCCTG---------GATGACTTGCATGATAACGGGTCTGAGTCAAATGCAGA :  74
Hyrax      : AAAGTTAATGAGTGGTTTTCCAGAAGTGACAACCTA---------AGTGATTCACCTAGTGAGGGGTCTGAATTAAATGGAAA :  74
Aardvark   : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGGCCTG---------GATGGCTCACATGATGAAGGGTCTGAATCAAATGCAGA :  74
Tenrec     : AAGGTTAACGAGTGGTTTTCCAAAAGCCACGGCCTG---------GGTGACTCTCGCGATGGCCGGCCTGAGTCAGGCGCAGA :  74
Rhinoceros : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGAAATATTAACTTCTGATGACTCACATGATGGCGGCCTGAATCAAATACTGA :  83
Pig        : AAAGTTAATGAGTGGTTTTCTAGAAGCGATGAAATGTTAACTTCTGACGACTCACAGGACAGCAGGTCTGAATCAAATACTGG :  83
Hedgehog   : AAAGTGAATGAATGGCTTTCCAGAAGTGATGAACTGTTAACTTCTGATGACTCATATGATAACGGATCTAAATCAAAAACTGA :  83
Human      : AAAGTTAATGAGTGGTTTTCCAGAAGTGATGAACTGTTAGGTTCTGATGACTCACATGATGGCGAGTCTGAATCAAATGCCAA :  83
Rat        : AAAGTGAATGAGTGGTTTTCCAGAACTGGTGAAATGTTAACTTCTGACAATGCATCTGACAGCAGGCCTGCGTCAAATGCAGA :  83
Hare       : AAAGTTAACGAGTGGTTCTCCAGAAGTAATGAAATGTTAACTCCTGATGACTCACTTGACCGGCGGTCTGAATCAAATGCCAA :  83
             AAaGTtAAtGAgTGGtTtTccAgAagt atga  T          gatgactca   gat  g gg cTga t aaatgc ga

                   *       100         *       120         *       140         *
Wombat     : GGTGCCTAGTGCCTTAGAAGATGGGCATCCAGATACCGCAGAGGGAAATTCTAGCGTTTCTGAGAAGACTGAC : 156
Opossum    : GCCAACCAATGCTTTAGAATATGGGCATGTAGAGACA---GATGGAAATTCTAGCATTTCTGAAAAGACTGAT : 153
Armadillo  : AGTAGCTGGTGCATTGAAAGTT------TCAAAAGAAGTAGATGAATATTCTAGTTTTTCAGAGAAGATAGAC : 150
Sloth      : AGTAGTTGGTGCATTGAAAGTT------CCAAATGAAGTAGATGGATATTCTGGTTCTTCAGAGAAGATAGAC : 150
Dugong     : AGTAGCTGGTGCTTTAGAAGTT------CCAGAAGAAGTACATGGATATTCTAGTTCTTCAGAGAAAATAGAC : 141
Hyrax      : AGTGGCTGGTCCAGTAAAACTT------CCAGGTGAAGTACATAGATATTCTAGTTTTCCAGAGAACATAGAT : 141
Aardvark   : AATAGGTGGTGCATTAGAAGTT------TCAAATGAAGTACATAGTTACTCTGGTTCTTCAGAGAAAATAGAC : 141
Tenrec     : CGTAGCTGTAGCCTTCGAAGTT------CCAGACGAAGCATGTGAATCTTATAGTTCTCCAGAGAAAACAGAC : 141
Rhinoceros : AGTAGCTGGTGCAGTAGAAGTT------CAAAATGAAGTAGATGGATATTCTGGTTCTTCAGAGAAAATAGGC : 150
Pig        : GGTAGCTGGTGCAGCAGAGGTT------CCAAATGAAGCAGATGGACATTTGGGTTCTTCAGAGAAAATAGAC : 150
Hedgehog   : AGTAACTGTAACAACAGAAGTT------CCAAATGCAATAGATAGRTTTTTGGTTCTTCAGAGAAAATAAAC : 150
Human      : AGTAGCTGATGTATTGGACGTT------CTAAATGAGGTAGATGAATATTCTGGTTCTTCAGAGAAAATAGAC : 150
Rat        : ACCTGCTGTTGTGTTAGAAGTT------TCAAATGAAGTGGATGGATGTTTCAGTTCTTCAAAGAAAATAGAC : 150
Hare       : AGTGGCTGGTGCATTAGAAGTC------CCAAAGGAGCTAGATGGATATTCTGGTTCTACAGAGAAAATAGAC : 150
              gt gctg tgc  t gAagtt       cA a gaag a atggatatT t Gtt TtCagAgAA Atagac
```

Alignment of the Brca1 protein sequences from the same region of the gene as in fig. 3.1



```
                      *        20         *        40         *
Wombat     : KVNEWLSRSSDILASDNSNGRSHEQSAEVPSALEDGHPDTAEGNSSVSEKTD : 52
Opossum    : KVNEWLFRSNDVLAPDYSSVRSHEQNAEATNALEYGHVET-DGNSSISEKTD : 51
Armadillo  : KVNEWFSRGDDILTSDDSHDRGSELNAEVAGALKV--SKEVDEYSSFSEKID : 50
Sloth      : KVNEWFSRSDDILTSDDSHNGGSESNAEVVGALKV--PNEVDGYSGSSEKID : 50
Dugong     : KVNEWFFRSDGL---DDLHDKGSESNAEVAGALEV--PEEVHGYSSSSEKID : 47
Hyrax      : KVNEWFSRSDNL---SDSPSEGSELNGKVAGPVKL--PGEVHRYSSFPENID : 47
Aardvark   : KVNEWFSRSDGL---DGSHDEGSESNAEIGGALEV--SNEVHSYSGSSEKID : 47
Tenrec     : KVNEWFSKSHGL---GDSRDGRPESGADVAVAFEV--PDEACESYSSPEKTD : 47
Rhinoceros : KVNEWFSRSDEILTSDDSHDGGPESNTEVAGAVEV--QNEVDGYSGSSEKIG : 50
Pig        : KVNEWFSRSDEMLTSDDSQDRRSESNTGVAGAAEV--PNEADGHLGSSEKID : 50
Hedgehog   : KVNEWLSRSDELLTSDDSYDKGSKSKTEVTVTTEV--PNAIDXFFGSSEKIN : 50
Human      : KVNEWFSRSDELLGSDDSHDGESESNAKVADVLDV--LNEVDEYSGSSEKID : 50
Rat        : KVNEWFSRTGEMLTSDNASDRRPASNAEAAVVLEV--SNEVDGCFSSSKKID : 50
Hare       : KVNEWFSRSNEMLTPDDSLDRRSESNAKVAGALEV--PKEVDGYSGSTEKID : 50
             KVNEWfs4   6   d s     e n             e        eki
```

Why align sequences?

- Functional predictions based on identifying homologues.

Assumes:

conservation of sequence ↔ conservation of function

BUT: Function carried out at level of proteins, i.e.
3-D structure
Sequence conservation carried out at level of DNA
1-D sequence

BASIC CONCEPTS UP TO THIS POINT: **HOMOLOGY, ORTHOLOGY, PARALOGY**

EVOLUTION/SEQUENCES/STRUCTURES (see PBC chap. 18)

THE CHOTHIA LESK DIAGRAM

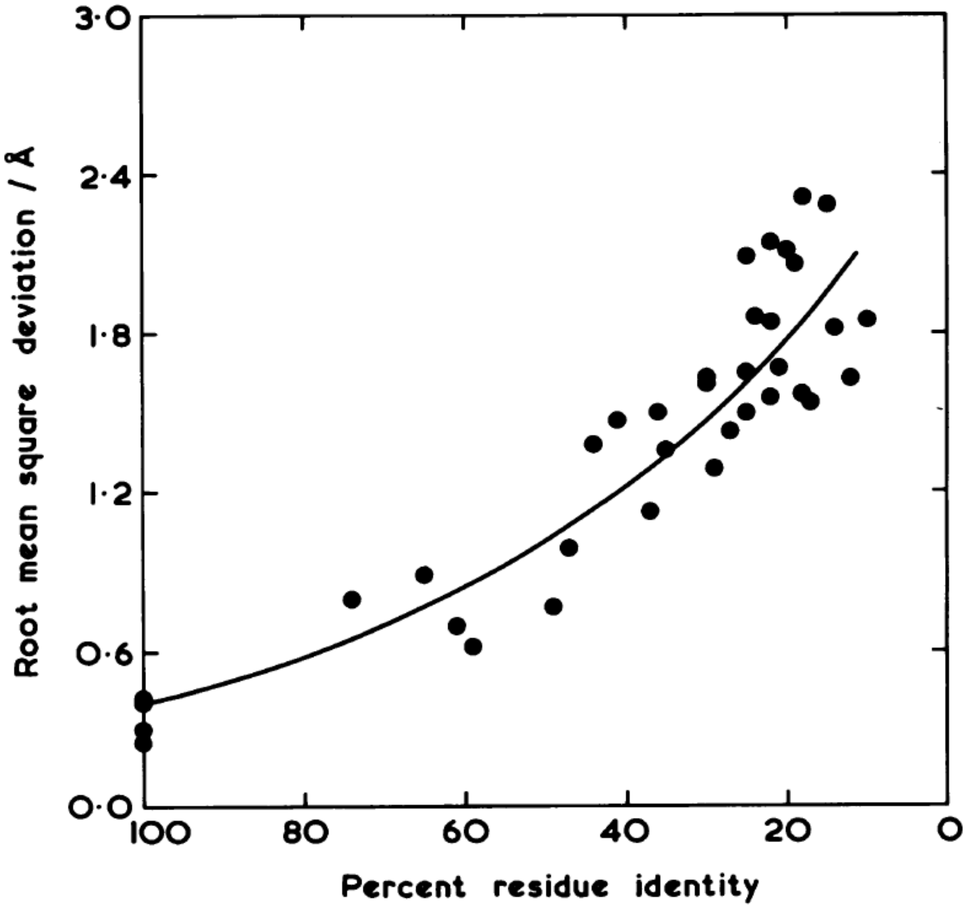


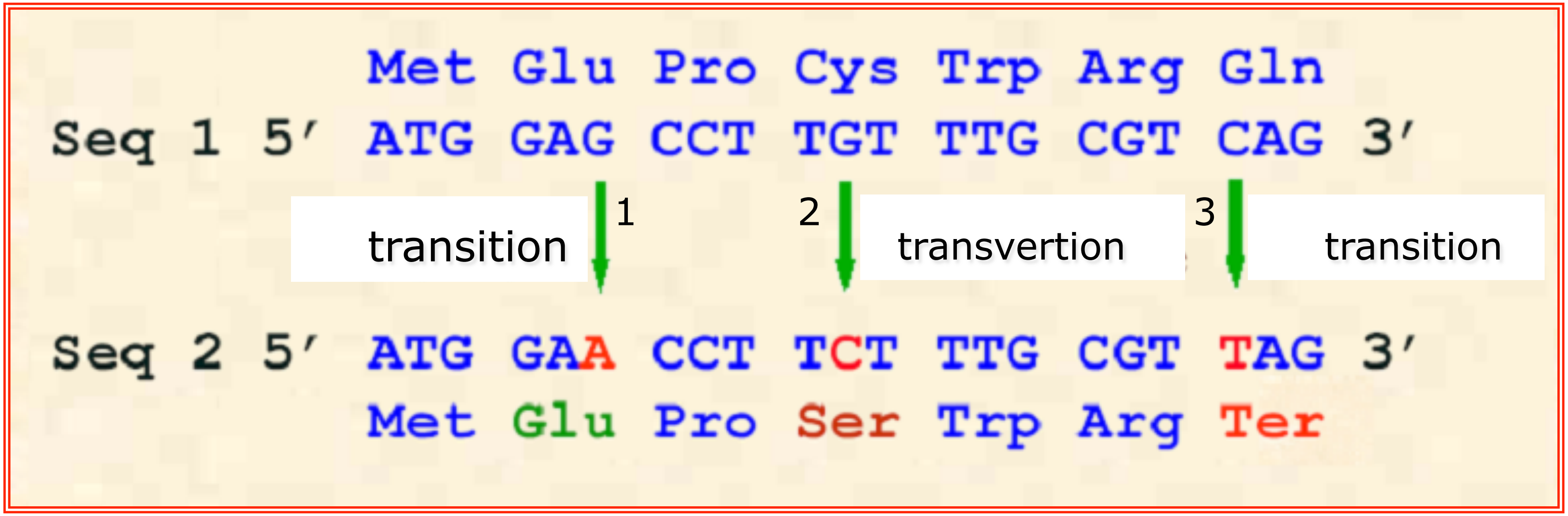
Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

From: Chothia and Lesk

The EMBO Journal vol.5 no.4 pp.823–826, 1986

How proteins can change

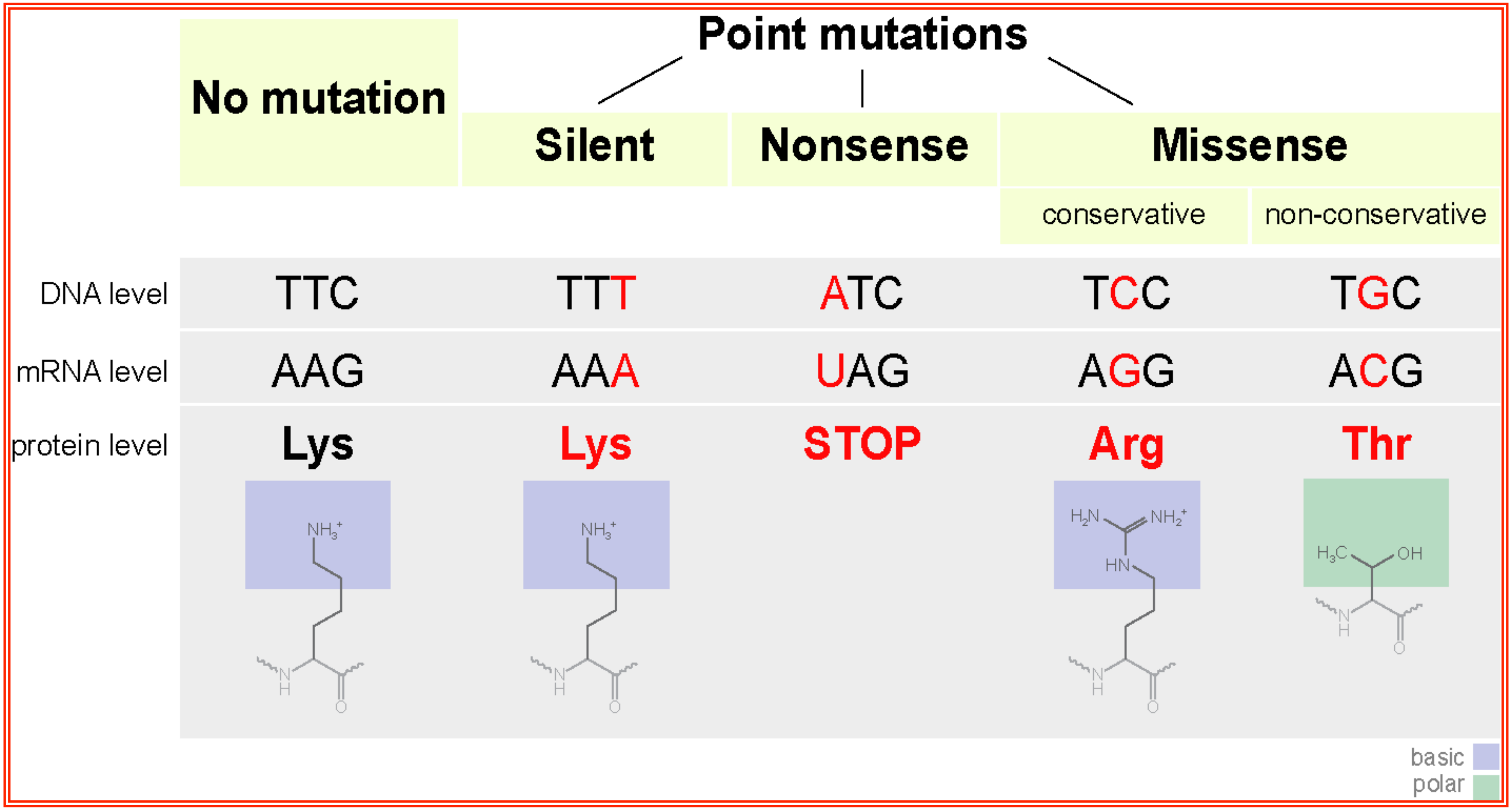
Look at point mutations (purines vs pirimidines)



		Met	Glu	Pro	Cys	Trp	Arg	Gln	
Seq 1	5′	ATG	GAG	CCT	TGT	TTG	CGT	CAG	3′
		transition	1		2	transvertion		3	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



		Point mutations			
	No mutation	**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**

NH_3^+
NH_3^+
H_2N
NH_2^+
HN
H_3C
OH
N
H
O

basic ▨
polar ▨

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

- The Jukes-Cantor model
- Time reversibility (Detailed Balance)
- Variability of rates between sites

…AT THE BLACKBOARD

- If two sequences have a significant **degree of similarity (how to measure? Hamming distance? Distance in amino acid composition)** for all their length, it is very likely that this is due to a sort of "memory" of their evolutionary relationship (do evolutionarily related proteins have a memory in amino acidic composition? **E.G. chitinases**)
- 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

Sequence Similarity is a quantitative information, based on the chosen metric, and it is independent from assumptions about the cause of the similarity itself

Sequence Homology is a qualitative information related to the ontology, that stands for the common phylogenetic origin of two sequences. The **evolutionary distance** is related to homology. There is a twilight zone: elaborate on that.

EVOLUTION SEQUENCES STRUCTURES (see PBC chap. 18)

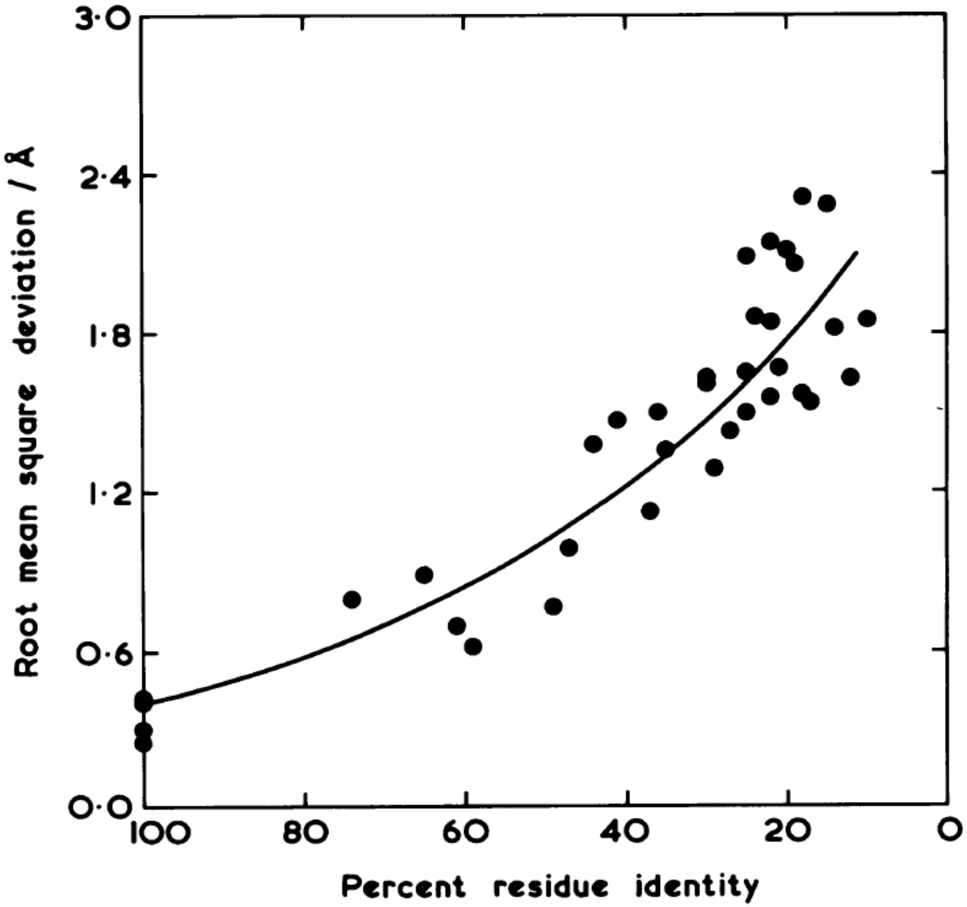


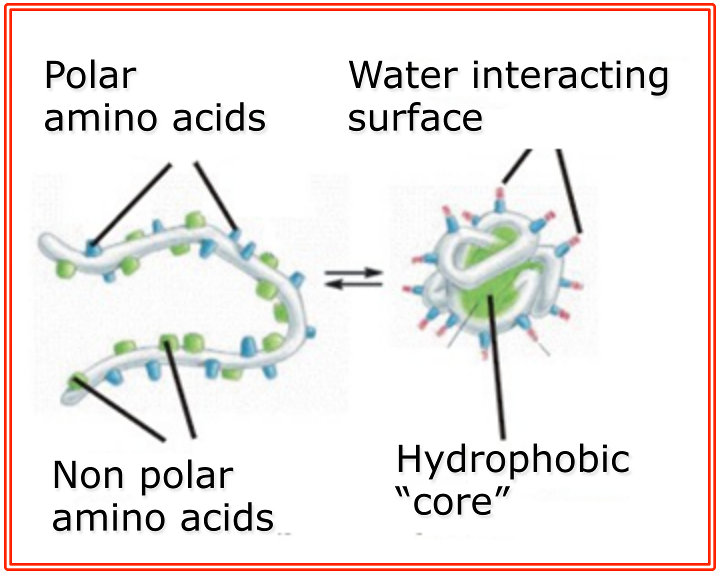
Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

From: Chothia and Lesk

The EMBO Journal vol.5 no.4 pp.823–826, 1986

Functional constraints

- From the structural point of view:
  - Most of the 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

- 18 out of 20 amino acids are encoded by more than one codon (on average degeneracy 3)
    - 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
- 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

Mutations and substitutions

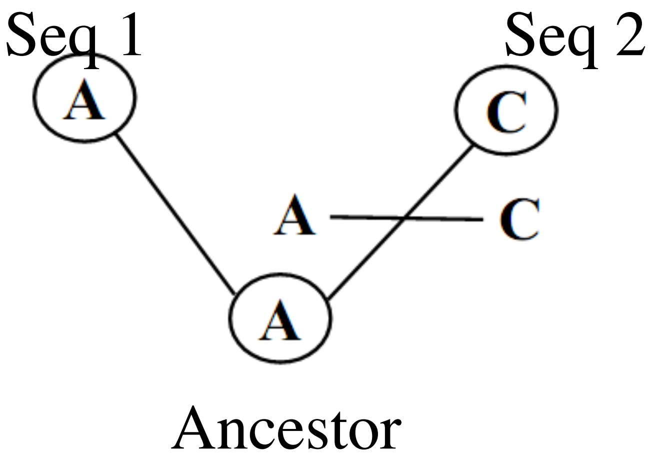
- **Remark sequence effects due to natural selection**
- In the populations of organisms found in Nature, the only available alleles (variants of a gene) are those which have not had a detrimental effect on the health of the organisms **detrimental allelles are non-observable**
  - Changes in the nucleotide 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 just "passed the filter" of natural selection, That were fixed in the **population** and **species**
  - The number of mutations is "easy" to calculate, whereas it is rather difficult to obtain a reliable estimate of the substitution frequency

Estimate of the number of substitutions

- 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 mutations, 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

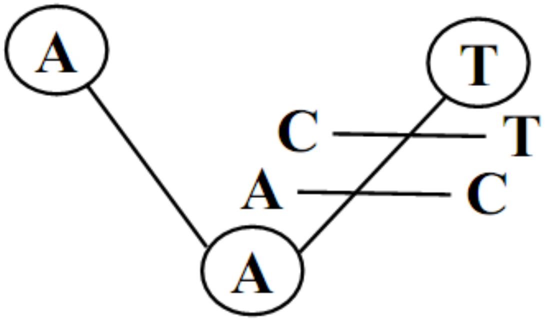
Estimation of the number of substitutions

Single substitution



1 substitution, 1 difference

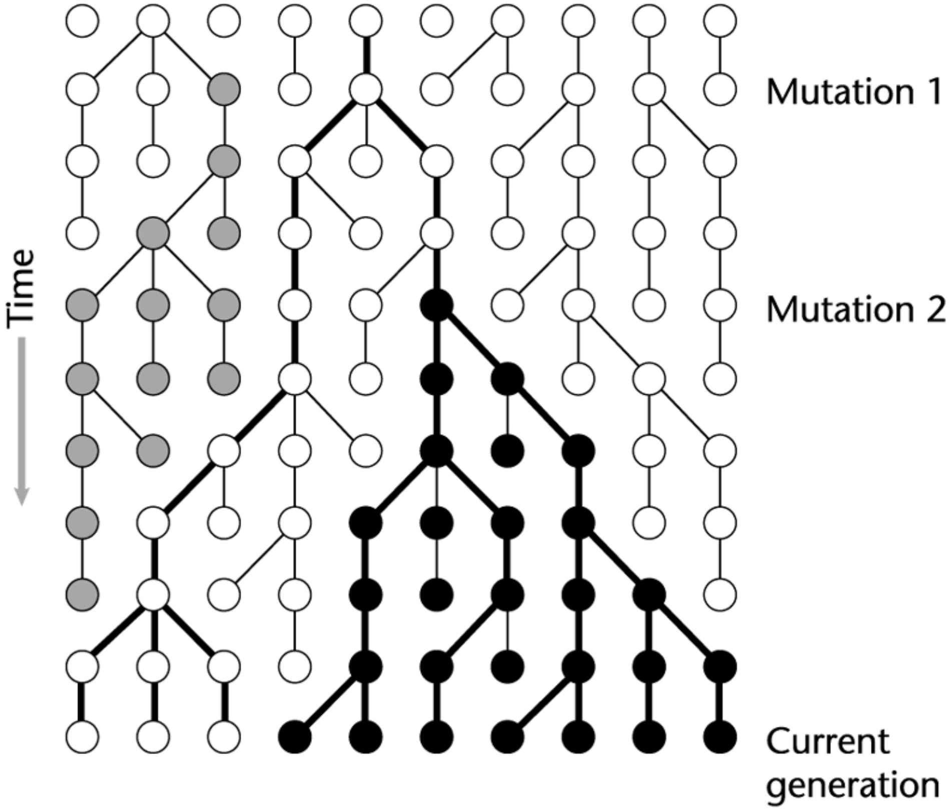
Multiple substitution



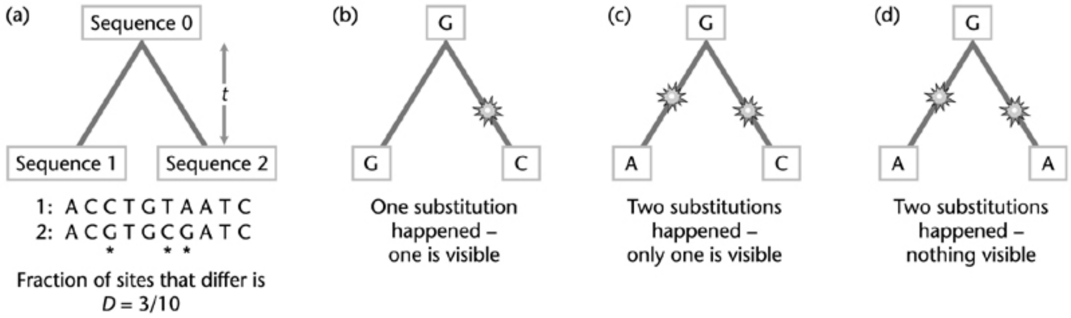
2 substitutions, 1 difference

Underestimation of the number of substitutions due to multiple substitutions, the observed differences underestimate the actual amount of evolutionary changes (accumulation of mutation/substitutions)

“Family tree” of a gene over a population (uniparental inheritance) and its generations:
time is the background elusive concept behind evolution



The accumulation of substitutions in two sequences descending from a common ancestor


(a)
Sequence 0
t
Sequence 1
Sequence 2
1: A C C T G T A A T C
2: A C G T G C G A T C
* * *
Fraction of sites that differ is
D = 3/10
(b)
G
G
C
One substitution happened – one is visible
(c)
G
A
C
Two substitutions happened – only one is visible
(d)
G
A
A
Two substitutions happened – nothing visible

Q. Why evolutionary models?

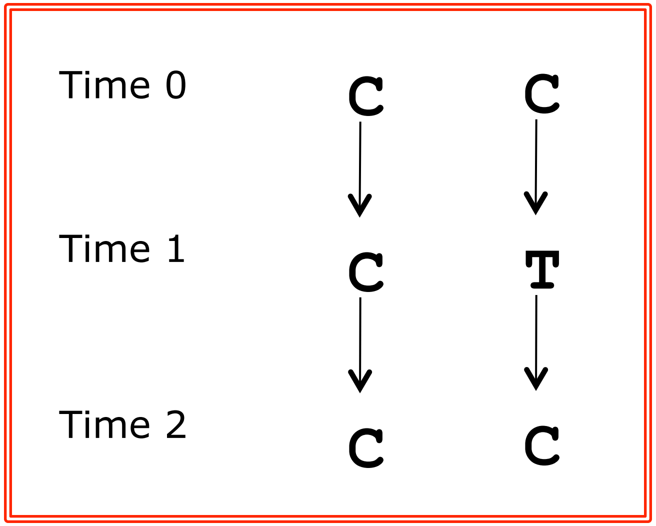
A. To infer d(A,B) from D(A,B)
B. Through an evolutionary (probabilistic model)

Note: D is not linear in time (see above) and is not Additive

$$D_{12} \neq D_{01}+D_{02}$$

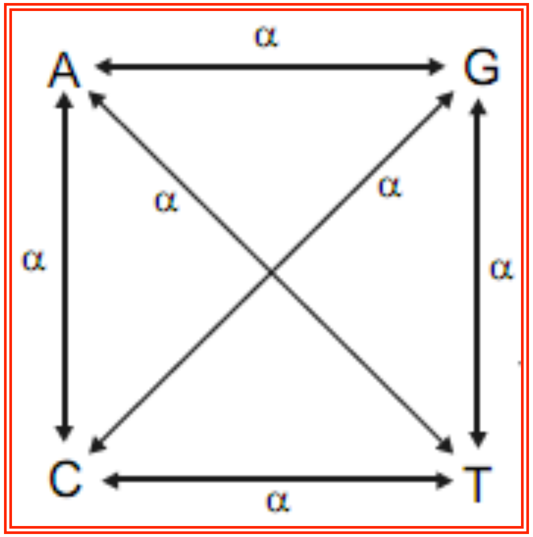
The Jukes-Cantor model - 1

- 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 probabilistic 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 $(1-3\alpha)P_{c(1)} + \alpha(1-P_{c(1)})$



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

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

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 rate of substitutions between nucleotides j and 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:

$$\mathrm{P}(t) = exp\{\mathrm{R}t\} = \sum_{k=1}^{\infty} (\mathrm{R}t)^k/k!$$

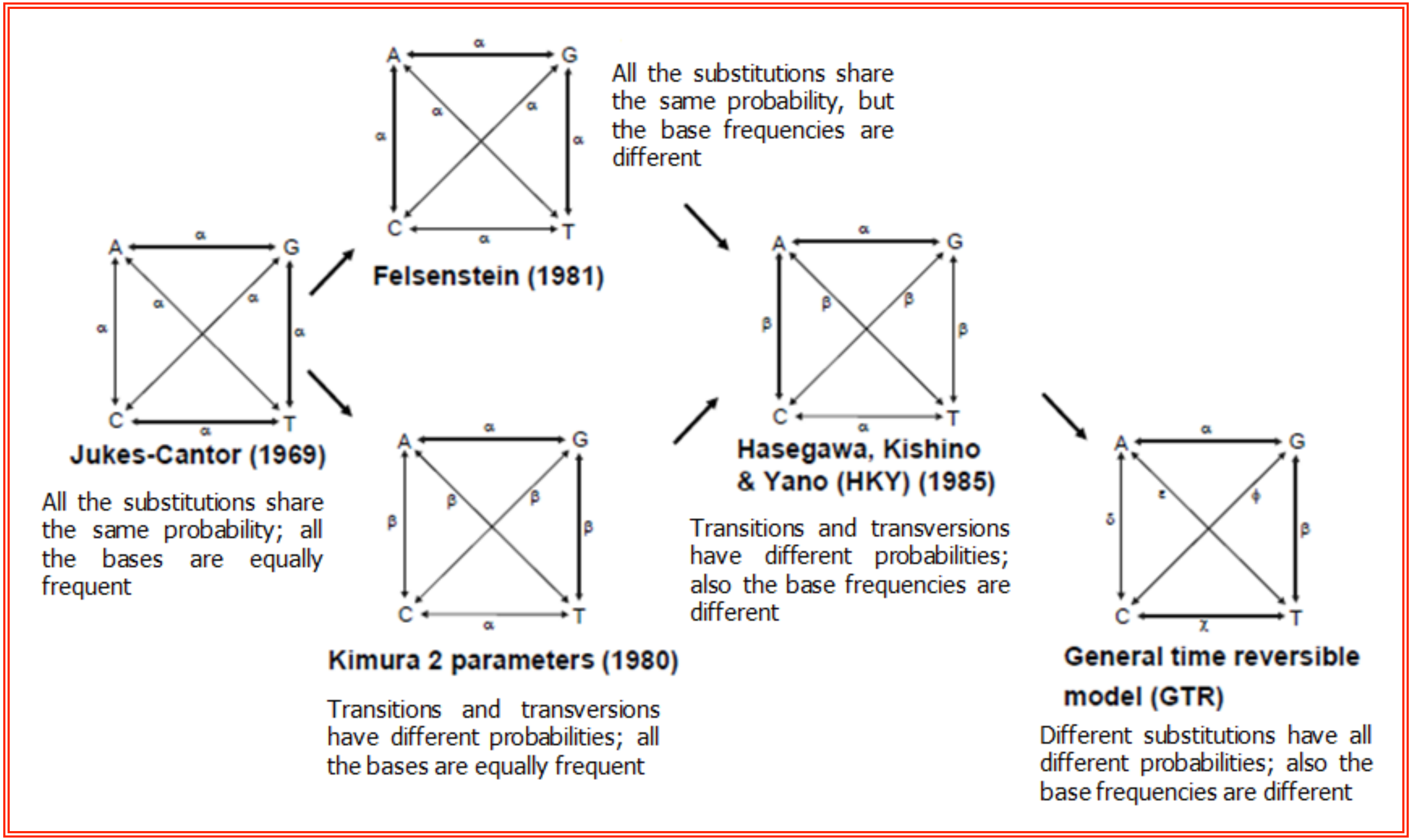
- Therefore, the elements of P are defined by

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

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 (α=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

Evolution of *K* estimation models



Jukes-Cantor (1969)

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

Felsenstein (1981)

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

Kimura 2 parameters (1980)

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

Hasegawa, Kishino & Yano (HKY) (1985)

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

General time reversible model (GTR)

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