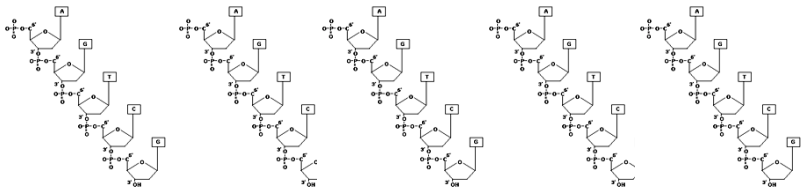
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Molecular Phylogeny and Evolution

10 – 14 February 2020

Juan I. MONTOYA BURGOS



UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Title of the course: **Molecular Phylogeny and Molecular Evolution**

Evolutionary relationships among organisms = tree topology

First phylogenetic methods did not make use of models of molecular evolution
(UPGMA, Maximum Parsimony)

A better understanding of molecular evolution improves:

- topology and branch length reconstruction (=phylogenetic tree)

Better phylogenetic trees improve:

-the understanding of evolutionary processes

=> Models of molecular evolution

Lab of Molecular Phylogeny and Evolution in Vertebrates

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Why should we care about phylogenies?

Are you using phylogenetics ?

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Current phylogenetic methods allow:

- reconstruction of evolutionary relationships

But also the analysis of:

Gene/genome duplication
Recombination
Evolutionary rates
Divergence time among lineages
Selective pressure
Demography
Biodiversity
Genetic variability
Conservation
Spread of contagious disease
Biogeography
Discovery of biomedical compounds
Adaptive evolution
Protein-protein co-evolution
Biomonitoring

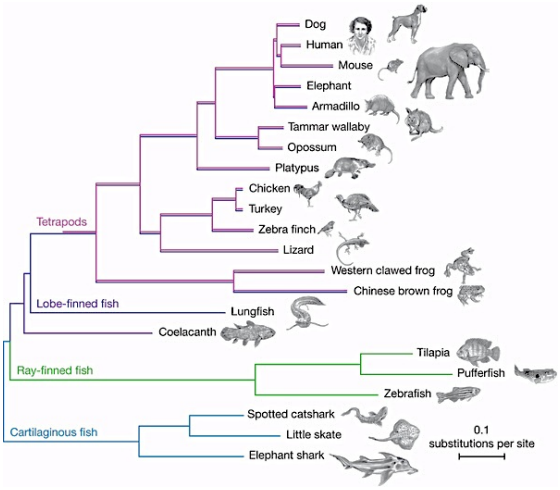
And more ….

Phylogenetic analyses: many uses

Understanding evolutionary relationships

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Determining the closest extant relative to tetrapods

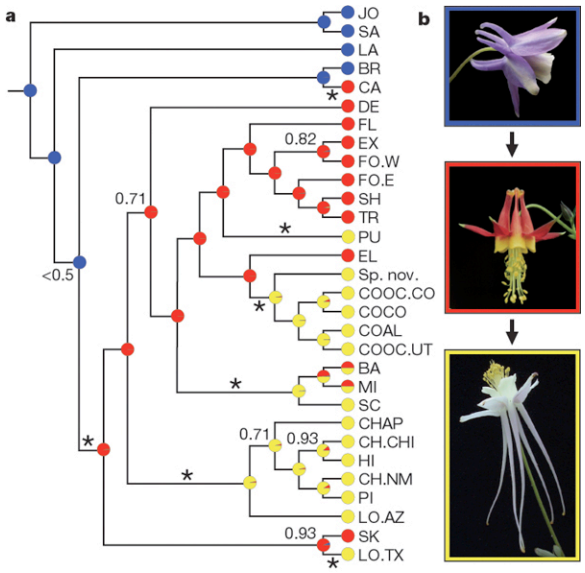


Amemiya et al., 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496.

Phylogenetic analyses: many uses

Understanding morphological evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES



Using a species-level phylogeny of *Aquilegia*, Whittall and Hodges showed a significant evolutionary trend for **longer nectar spur** during directional shifts to **pollinators with longer tongues** (moths).

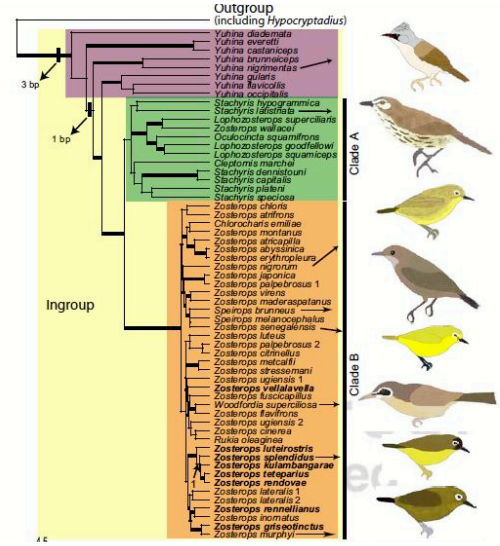
Whittall and Hodges, 2007. Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nature* **447**, 706-709

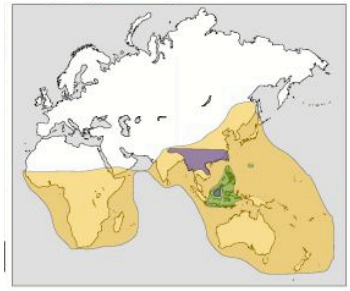
Phylogenetic analyses: many uses

Understanding bio-diversification

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Calibrated phylogeny of *Zosterops* and relatives





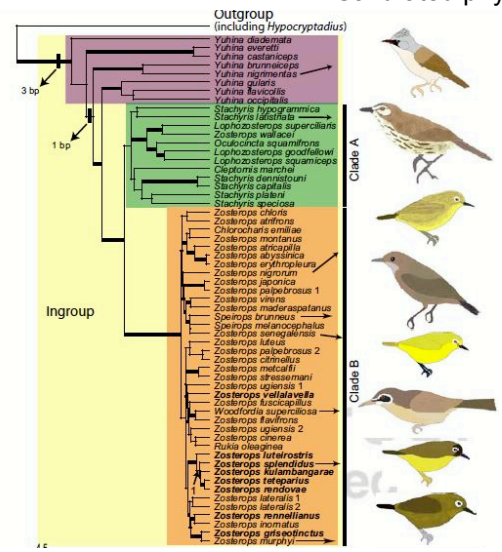
Moyle et al. 2009. Explosive Pleistocene diversification and hemispheric expansion of a "great speciator." PNAS

Phylogenetic analyses: many uses

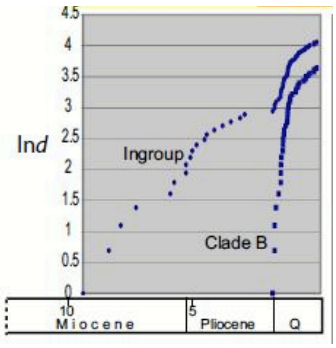
Understanding bio-diversification

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Calibrated phylogeny of *Zosterops* and relatives



Lineage-through-time plots (LTT)



Moyle et al. 2009. Explosive Pleistocene diversification and hemispheric expansion of a "great speciator." PNAS

Lab of Molecular Phylogeny and Evolution in Vertebrates

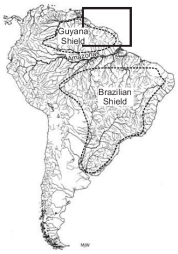
Phylogenetic analyses: many uses

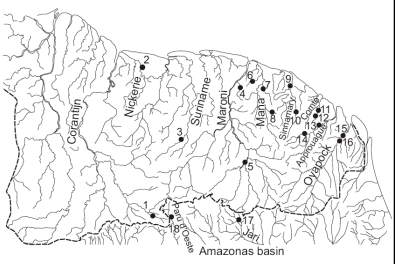
New biodiversity

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Unexpected diversity in the catfish *Pseudancistrus brevispinis* reveals dispersal routes in a Neotropical center of endemism: the Guyanas Region

YAMILA P. CARDOSO[1] and JUAN I. MONTOYA-BURGOS [1]*

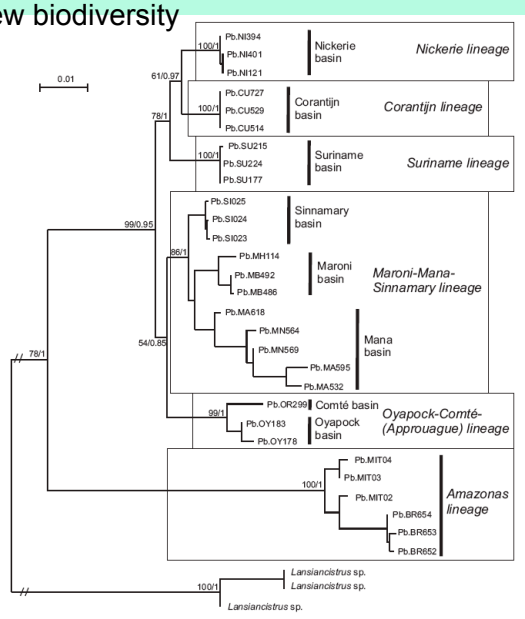




Phylogenetic analyses: many uses

New biodiversity

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES



1 species → 6 species

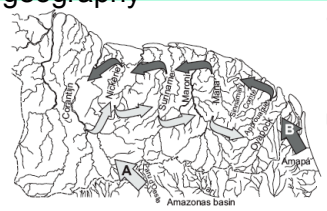
Morphology		
Plates between adipose and caudal fins	Head profile	Body shape
Three small, one large	Ascending from eyes to dorsal fin	Intermediate
Four small	Rounded from eyes to dorsal fin	Massif
Three small, one large	Horizontal and straight from eyes to dorsal fin	Widest
Two small, two large	Shortest and pointed snout. Smallest interorbital distance	Highest
Three small, one large	Ascending from eyes to dorsal fin	Longest and thinnest
Three small, two large	Descending straight from eyes to dorsal fin	Flattest

Lab of Molecular Phylogeny and Evolution in Vertebrates

Phylogenetic analyses: many uses

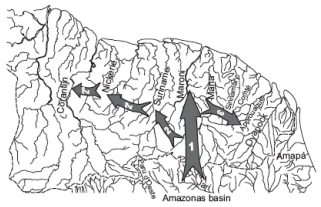
Biogeography

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

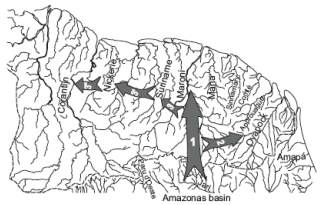


A.
- Oyapock-Comté-Approuague basin
- Maroni-Mana-Sinnamary basin
- Suriname basin
- Nickerie basin
- Corantijn basin
- Amazonas basin

B.
- Corantijn basin
- Nickerie basin
- Suriname basin
- Maroni-Mana-Sinnamary basin
- Oyapock-Comté-Approuague basin
- Amazonas basin



C.
- Oyapock-Comté-Approuague basin
- Maroni-Mana-Sinnamary basin
- Corantijn basin
- Nickerie basin
- Suriname basin
- Amazonas basin



D.
- Corantijn basin
- Nickerie basin
- Suriname basin
- Maroni-Mana-Sinnamary basin
- Oyapock-Comté-Approuague basin
- Amazonas basin

Biogeography: four possible entrance scenarios

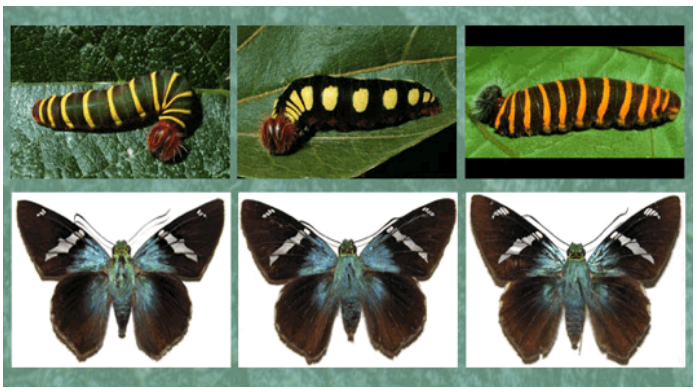
Topological tests:
-SH test
-AU test

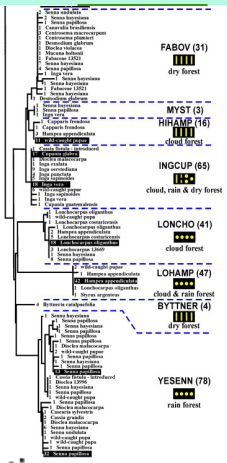
Lab of Molecular Phylogeny and Evolution in Vertebrates

Phylogenetic analyses: many uses

DNA taxonomy of Animals: species identification

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES





Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*

Paul D. N. Hebert*†, Erin H. Penton*, John M. Burns‡, Daniel H. Janzen§, and Winnie Hallwachs§

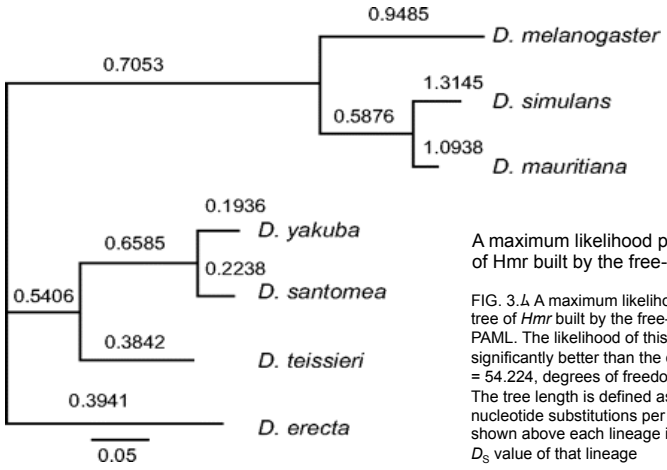
Hebert et al. PNAS 101:14812 (2004)

Lab of Molecular Phylogeny and Evolution in Vertebrates

Phylogenetic analyses: many uses

Evolutionary rates and selective pressure acting on genes

The evolutionary histories of genes bear the marks of the functional demands to which they have been subjected.



A maximum likelihood phylogenetic tree of Hmr built by the free-ratio model in PAML

FIG. 3.4 A maximum likelihood phylogenetic tree of *Hmr* built by the free-ratio model in PAML. The likelihood of this model was significantly better than the one-ratio model ($2l = 54.224$, degrees of freedom = 11, $P < 10^{-7}$). The tree length is defined as the number of nucleotide substitutions per codon. The number shown above each lineage is the estimated D_N/D_S value of that lineage

Maheshwari, S. et al. Mol Biol Evol 2008 25:2421-2430; doi:10.1093/molbev/msn190

Lab of Molecular Phylogeny and Evolution in Vertebrates

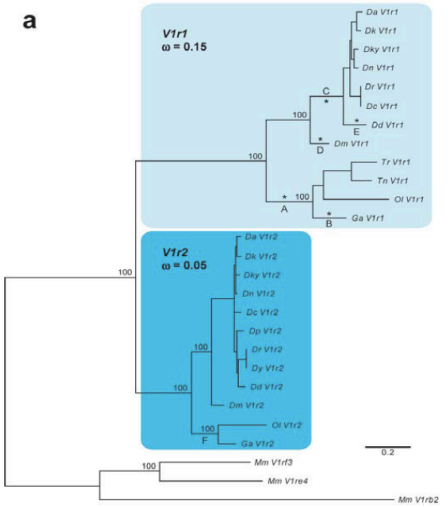
Phylogenetic analyses: many uses

OPEN ACCESS Freely available online — PLoS one

Divergent Evolution among Teleost V1r Receptor Genes

Patrick Pfister[1,2], Jerome Randall[1,2], Juan I. Montoya-Burgos[1], Ivan Rodriguez[1,2]*

1 Department of Zoology and Animal Biology, University of Geneva, Geneva, Switzerland, 2 National Center of Competence in Research (NCCR) Frontiers in Genetics, University of Geneva, Geneva, Switzerland



Analysis of:

- duplicated genes
- gene families
- genome duplication / polyploidisation

Lab of Molecular Phylogeny and Evolution in Vertebrates

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Evolution and selection on proteins, protein interactions and protein networks

Protein evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

REVIEWS

COMPUTATIONAL TOOLS

Emerging methods in protein co-evolution

David de Juan[1], Florencio Pazos[2] and Alfonso Valencia[1]

NATURE REVIEWS | GENETICS VOLUME 14 | APRIL 2013 | 249

Lab of Molecular Phylogeny and Evolution in Vertebrates

Phylogenetic analyses: many uses

Protein evolution

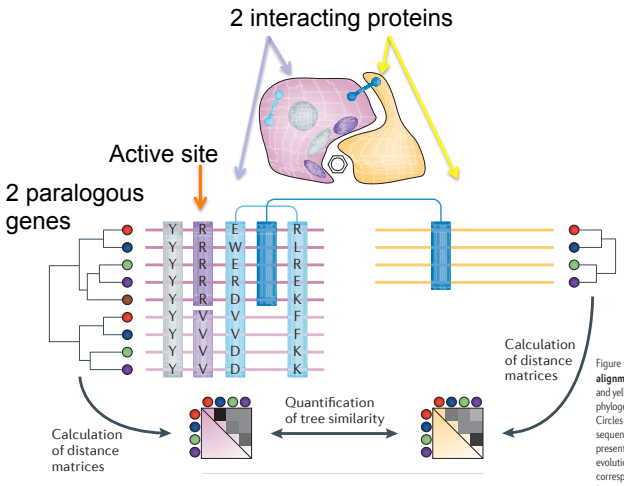


Figure 1 | **Co-evolutionary features extracted from protein multiple sequence alignments.** The three-dimensional structures of two interacting proteins (purple and yellow) are schematized as well as their multiple sequence alignments (MSAs) and phylogenetic trees based on orthologous sequences from a number of organisms. Circles of different colours represent different species from which the protein sequences are derived. The purple MSA includes a family of paralogues that are present in some of these organisms (sequences in dark purple). A number of evolutionary and co-evolutionary features are schematized in the alignments and the corresponding three-dimensional structures. Fully conserved positions (grey) tend to form a part of the protein core and are also in functional regions (such as protein interaction sites and catalytic sites). Specificity-determining positions (SDPs; purple) tend to be in functional sites conferring specificity. Intra-protein correlated mutations (light blue) are related to residue spatial proximity, whereas inter-protein correlated mutations (dark blue) reflect in many cases proximity between residues in different protein chains. Additionally, the protein–protein co-evolution can be evaluated: the phylogenetic trees of the two families can be compared in different ways, as shown below the MSAs. For example, the trees can be converted to distance matrices to

Phylogenetic analyses: many uses

Protein evolution

eLIFE elifesciences.org

RESEARCH ARTICLE

Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity

Conor J Howard[1†], Victor Hanson-Smith[2†], Kristopher J Kennedy[1], Chad J Miller[3], Hua Jane Lou[3], Alexander D Johnson[2], Benjamin E Turk[3*], Liam J Holt[1*]

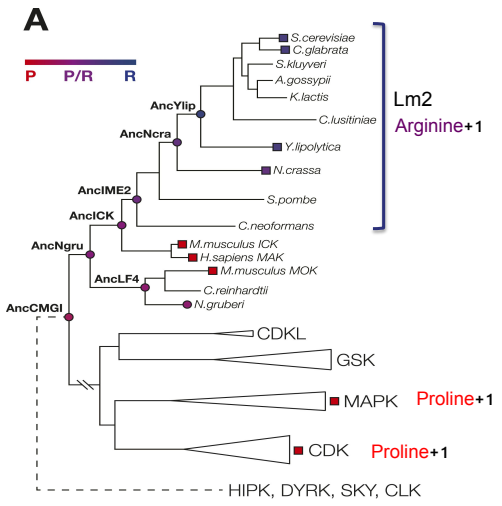
DOI: http://dx.doi.org/10.7554/eLife.04126
Cite as eLife 2014;3:e04126

Lab of Molecular Phylogeny and Evolution in Vertebrates

Phylogenetic analyses: many uses

Protein evolution

Ancestral sequence reconstruction using Maximum Likelihood inference.
Ancestral sequence synthesis and enzymatic activity essays.

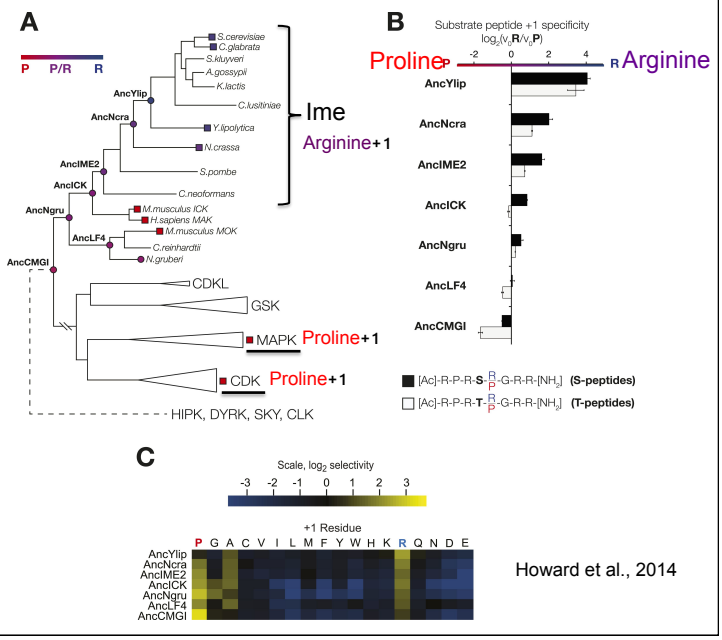


The cyclin dependent kinases (CDKs) and mitogen activated protein kinases (MAPKs) require proline (P) at the +1 position of their substrates,

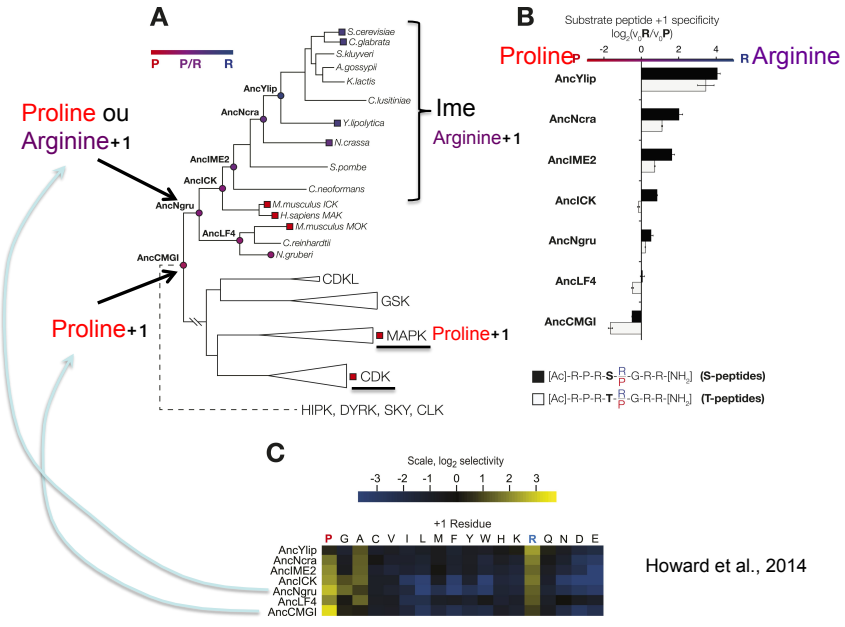
while Ime2 prefers arginine (R).

Lab of Molecular Phylogeny and Evolution in Vertebrates

Howard et al., 2014



Howard et al., 2014



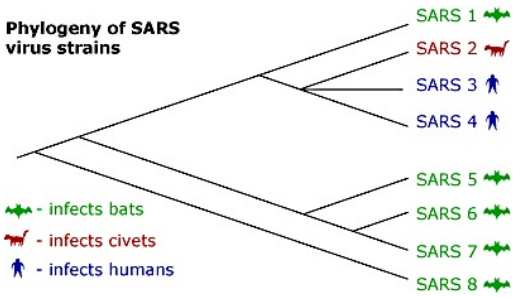
Howard et al., 2014

Phylogenetic analyses: many uses

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Origin of viruses, disease phylogeny

In 2002-2003, the previously unknown SARS virus generated widespread panic causing 774 deaths and more than 8000 cases of illness.
Where did this virus comes from?



Source of Severe Acute Respiratory Syndrome coronavirus (SARS) was unknown… phylogenetic analyses traced it to civets and bats.

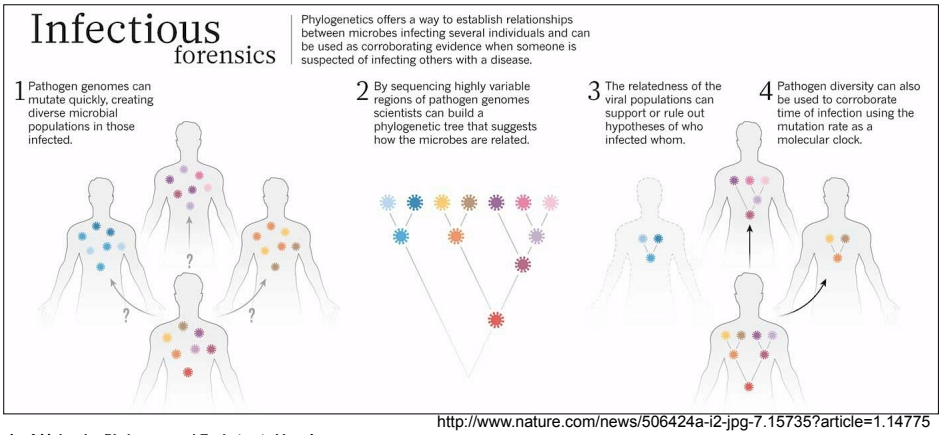
Normile, Dennis. (2005). Researchers tie deadly SARS virus to bats. Science 309:2154-2155.

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Phylogenetics as a tool in various fields

Phylogenetic is used as a tool:

- in historical linguistics,
- in security applications for networks and computers such as **artificial immune systems** for computers,
- in forensic studies, for instance in infectious forensics.



http://www.nature.com/news/506424a-i2-jpg-7.15735?article=1.14775

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Todays's theory refreshment

Molecular evolution

Molecular evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

- **Molecular basis of evolution.**
- Evolution of protein sequences
- Evolution of DNA sequences

Molecular evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Molecular basis of evolution

1) Evolutionary processess
2) Genome structure
3) Gene structure and fonctions
4) Mutational changes in DNA
5) Codon Usage in coding sequences

Lab of Molecular Phylogeny and Evolution in Vertebrates

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Causal factor of evolution:
mutational changes in DNA, particularly in coding sequences or regulatory regions

Main mutational changes:

- nucleotide substitution
- insertion/deletion of nucleotides (indels) / mobile genetic elements
- non-homologous recombination
- gene conversion
- duplications of genes / chromosomal region / genomes

Mutation Fixation in a species:

A new mutation can spread in a population over generations and can eventually be fixed (or lost) in the species by genetic drift and/or by natural selection.

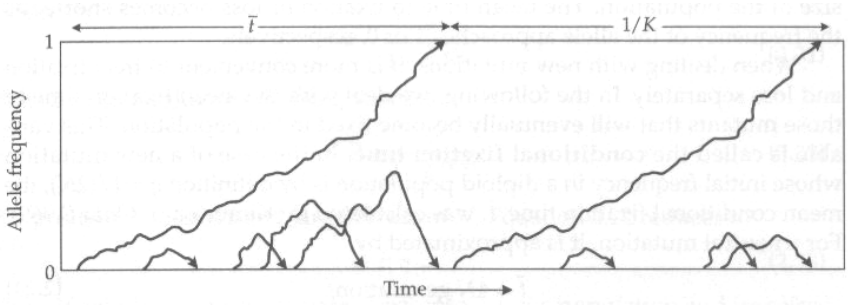
Natural selection:…- positive Darwinian selection promotes fixation
- negative selection (purifying selection) leads to mutation loss

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Fixation (or loss) of a mutation by genetic drift or natural selection.

A mutation is **fixed** when it is present in **all members of a species**.



Lab of Molecular Phylogeny and Evolution in Vertebrates

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Fixation of a mutation

Neutral mutations
A fraction μ of all copies of a gene mutate (μ= mutation rate).
Of these 1/2N (equal to the initial frequency of the mutant) succeed in drifting to fixation for the mutant (=probability of fixation).
There are in all 2N copies of the gene available to mutate.

The resulting rate of substitution is

$r = \mu \times 2N \times 1/2N = \mu$

So the rate of substitution of neutral mutations is equal to the mutation rate.

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Fixation of a mutation

Selected mutations
A fraction μ of copies of the gene mutate. There are in all 2N copies available.
A fraction 2s succeed in fixing (=probability of fixation for a selected mutation).
The resulting rate of substitution is:

$r = \mu \times 2N \times 2s = 4Ns\mu$

s = selective coefficient.

Note that this is 4Ns times as high as for neutral mutants, if the mutation rate in both categories were equal.

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Fixation of a mutation

Selected mutations

Fixation rate (***r***) depends on:

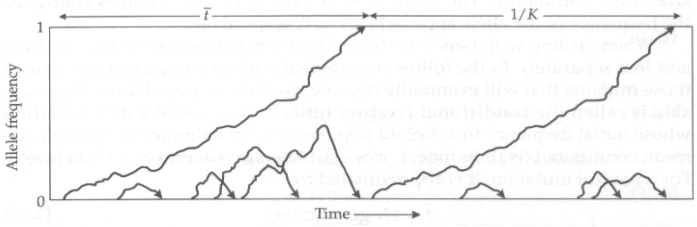
- population size (*N*),
- mutation fitness (*s*)
- mutation rate (μ).

$$r = 4Ns\mu$$

1) **Evolutionary processess**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

The fitness (*s*) of a mutation influences the fixation rate



fitness (*s*)	time to fixation
Neutral mutation:	t_{neu}
Positively selected mutation:	$t_{pos} < t_{neu}$
Negatively selected mutation:	$t_{neg} > t_{neu}$, or $t_{neg} = \infty$

Lab of Molecular Phylogeny and Evolution in Vertebrates

Molecular evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Molecular basis of evolution

1) Evolutionary processess

2) Genome structure

3) Gene structure and fonctions

4) Mutational changes in DNA

5) Codon Usage in coding sequences

2) Genome Structure

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Coding and non-coding regions ❖

Non-coding regions (3):
- intergenic DNA or "junk" DNA
- regulatory regions (promoters, enhancers, …)
- non-coding repetitive elements (Alu, SINE, telomeric repeats)

Coding regions (4): - protein-coding genes with messenger RNAs (mRNAs)

- genes coding for structural RNAs:
  ribosomal RNAs (rRNA),
  transfer RNAs (tRNA),
  small nuclear RNAs (snRNA)

- genes coding for small regulatory RNAs: short interfering RNAs, microRNAs,…

- genes coding for long non-coding RNAs (> 200 bp) : function?

Lab of Molecular Phylogeny and Evolution in Vertebrates

Molecular evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Molecular basis of evolution

1) Evolutionary processess

2) Genome structure

3) Gene structure and functions

4) Mutational changes in DNA

5) Codon Usage in coding sequences

2) Gene structure and function

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Structure of protein-coding genes

Transcribed regions: succession of exons and introns
(except: mono-exonic genes)

-Untranslated 5' region (= 5' UTR) in the first (or few first) exon(s).
-Start codon (AUG) usually in the first exon.
-Stop codon usually in the last exon.
-Untranslated 3' region (= 3' UTR) the last (or few last) exon(s).

The pre-mRNA still contains introns.

The **splicing** process discards introns, which leads to a messenger RNA (mRNA)

mRNA still contains **5' and 3' UTRs** in addition to the **coding sequence (CDS)**

The CDS starts at the **Start codon** (AUG) and ends at the codon just before the **Stop codon**.

2) Gene structure and function

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

The genetic code

Every triplet of the CDS encodes for an amino acid = genetic code

The genetic code is …. **universal** (or almost universal).
It is essentially the same in Prokaryotes, Eukaryotes, choloroplasts and in mitochondria.

The genetic code is … **degenerated**.
There are 64 possible codons (4^3), yet only 20 amino acids. Some codons encode for the same amino acid, they are **synonymous**.

Start and Stop codons:

The standard Start codon is AUG which encodes a methionin (Met).
When several methionins are candidates, the "Kozack methionin" is the Start codon, also referred to as **the Translation Initiation Site (TIS)**,.
Usually, there are three termination codons (STOP codons): UAA, UAG et UGA.

2) Gene structure and function

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

universal genetic code

2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Lab of Molecular Phylogeny and Evolution in Vertebrates

2) Gene structure and function

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

mitochondrial genetic codes

Vertebrate mtDNA.

Codon	New translation
AGA	Stop
AGG	Stop
ATA	Met
TGA	Trp

Invertebrate mtDNA.

Codon	New translation
AGA	Ser
AGG	Ser
ATA	Met
TGA	Trp

Yeast mtDNA.

Codon	New translation
ATA	Met
CTA	Thr
CTC	Thr
CTG	Thr
CTT	Thr
TGA	Trp

2) Gene structure and function

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

mitochondrial genetic codes

Mold.Protozoan and Coelenterate mtDNA.

Codon	New translation
TGA	Trp

Echinoderm mtDNA.

Codon	New translation
AAA	Asn
AGA	Ser
AGG	Ser
TGA	Trp

Ascidian mtDNA.

Codon	New translation
AGA	Gly
AGG	Gly
AGG	Met
TGA	Trp

Flatworm mtDNA.

Codon	New translation
AAA	Asn
AGA	Ser
AGG	Ser
TAA	Tyr
TGA	Trp

Lab of Molecular Phylogeny and Evolution in Vertebrates

2) Gene structure and function

Different genetic codes in the **nucleus**

Ciliate Nuclear Code.

Codon	New translation
TAA	Gln
TAG	Gln

Alternative Yeast Nuclear.

Codon	New translation
CTG	Ser

Molecular evolution

Molecular basis of evolution

1) Evolutionary processess

2) Genome structure

3) Gene structure and functions

4) Mutational changes in DNA

5) Codon Usage in coding sequences

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

❖

Morphological and physiological heritable characters are encoded in the DNA.

➡ Every heritable character change is caused by a change in the DNA.

Four types of changes at the nucleotide level:

1. Substitution of a nucleotide by another nucleotide
2. Insertion
3. Deletion
4. Inversion

Insertions, deletions and inversions may involve more than a single nucleotide.

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

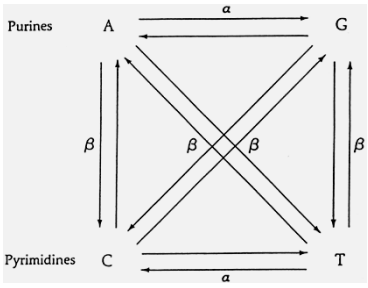
Two types of substitutions: **transitions** and **transversions**

Transitions: substitution between nucleotides belonging to the same family:

a change from a purin (A, G) to a purin
or
from a pyrimidin (C, T) to a pyrimidin (C, T)

Transversions: substitution betwen nucleotides belonging to different families.

Substitution scheme



Lab of Molecular Phylogeny and Evolution in Vertebrates

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Consequences at the level of coding regions:

Due to the degenerated genetic code, the impact of a substitution will vary:

A ..... will not change the encoded amino acid.

A ..... will change the encoded amino acid.

A ..... will generate a STOP codon.

Due to the degenerated genetic code:

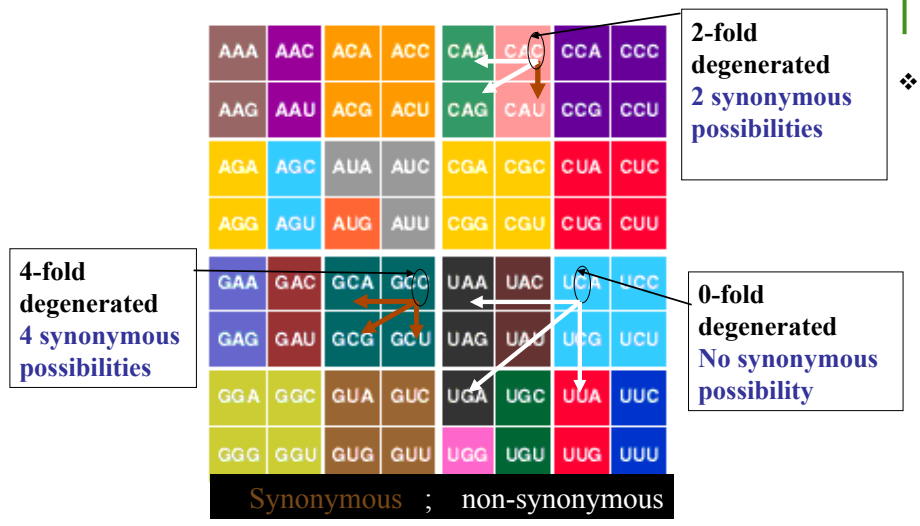
- most of the **synonymous** mutations take place at the **3rd** codon position,
- a few **synonymous** mutations take place at the **1st** codon position.

Mutations occurring at the 2nd codon position will be either **non-synonymous** or **nonsense** mutations.

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Sites are classified into 3 categories according to their degeneration level



2-fold degenerated
2 synonymous possibilities

4-fold degenerated
4 synonymous possibilities

0-fold degenerated
No synonymous possibility

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

4-fold degenerated: **32 possibilities in 3rd codon position over 61 codons**

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

2-fold degenerated: **25 possibilities in 3rd codon position 8 possibilities in 1st codon position.**

Lab of Molecular Phylogeny and Evolution in Vertebrates

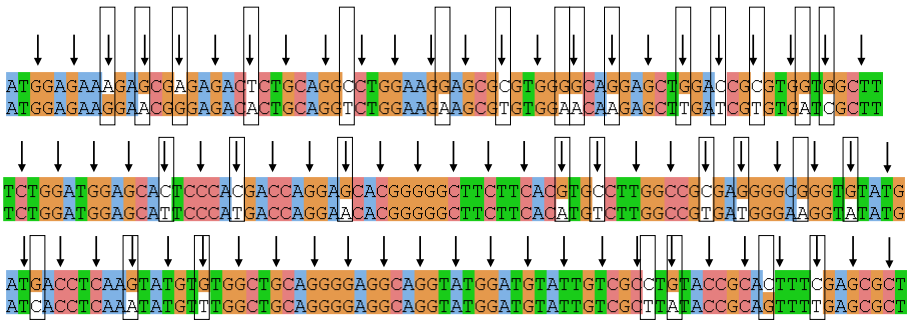
3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

0-fold degenerated: **all 2nd codon positions (61)**
53 1st codon positions

Most substitutions are located at the 3rd codon position

An example



First 220 nucleotide positions of the gene encoding the renin binding protein of human and mouse

Arrows ↓ indicate 3rd codon positions

31 changes:
4 - 1st codon position
4 - 2nd codon position
23 - 3rd codon position

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

❖

Consequences at the level of coding regions:

Theoretical proportions of the three substitution types.

If we assume that:
- codons are used with the same frequency
- nucleotides can be substituted by any other nucleotide with the same frequency

then:
non-synonymous substitutions = …… **71%**
synonymous substitutions = …… **25%**
non-sense substitutions = …… 4%

Is this found in real data ?

These theoretical proportions are not found in real data !
(the assumptions are not valid)

3) Mutational changes in DNA

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Insertions / deletions : main causes

Small indels (involving few nucleotide sites) are generally due to **DNA replication errors** during meiosis
(as for most nucleotide substitutions).

Large indels may result from:
- **unequal crossing overs**
- **transposition** via transposons or other transposable elements (retrotransposons, LINEs, SINEs, Alu, ...).

Some insertions may be due to **horizontal gene transfers,**

= integration into the genome of a species of a piece a DNA coming from a different species (via viruses, plasmids, transposons...?).

Lab of Molecular Phylogeny and Evolution in Vertebrates

Molecular evolution

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Molecular basis of evolution

1) Evolutionary processess

2) Genome structure

3) Gene structure and functions

4) Mutational changes in DNA

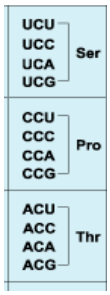
5) Codon Usage in coding sequences

4) Codon Usage in Coding Sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Are **synonymous codons** used with **equal frequency**?

UCU = UCC = UCA = UCG ??? (Serine)



If we assume:
-absence of selective pressure
-no bias in nucleotide substitution frequency

then **synonymous** codons should be present at the same frequency across all protein-coding genes in a genome.

Real data:
Synonymous codons are not equally frequent (assumptions above are not realistic).
The bias in synonymous codon usage is highly variable across organisms and genes.

Example: *E.coli* RNA polymerase genes (in parenthesis: *RSCU*)
Relative Synonymous Codon Usage

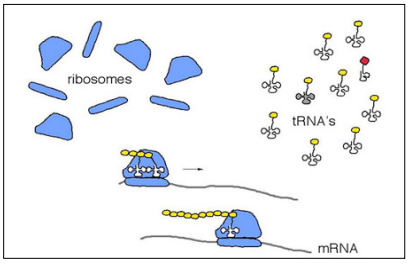
Val	GUU: 55 (1.53)	Pro	CCU: 9 (0.48)
	GUC: 21 (0.58)		CCC: 0 (0.00)
	GUA: 34 (0.94)		CCA: 11 (0.59)
	GUG: 34 (0.94)		CCG: 55 (2.93)

Lab of Molecular Phylogeny and Evolution in Vertebrates

4) Codon Usage in Coding Sequences

What can explain the bias in synonymous codons usage?

Observations in highly transcribed genes:
- the bias is high
- the usage frequency is high for codons with abundant tRNA in the sites of protein synthesis



Why is it so?

Explanation 1: For genes that need to be **highly transcribed**, alleles that are composed of codons that possess aboundant tRNAs can be rapidly transcribed and are favored by natural selection.
In other words, **purifying selection** eliminates alleles bearing codons for which tRNAs are rare.

4) Codon Usage in Coding Sequences

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Biased gene conversion drives codon usage in human and precludes selection on translation efficiency

Fanny Pouyet, Dominique Mouchiroud, Laurent Duret, Marie Semon
doi: https://doi.org/10.1101/086447
This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Info/History | Metrics | Supplementary material | Preview PDF

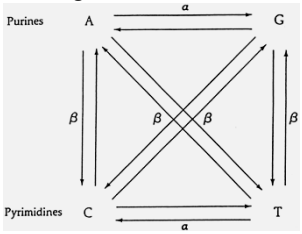
Abstract

In humans, as in other mammals, synonymous codon usage (SCU) varies widely among genes. In particular, genes involved in cell differentiation or in proliferation display a distinct codon usage, suggesting that SCU is adaptively constrained to optimize translation efficiency in distinct cellular states. However, in mammals, SCU is known to correlate with large-scale fluctuations of GC-content along chromosomes, caused by meiotic recombination, via the non-adaptive process of GC-biased gene conversion (gBGC). To disentangle and to quantify the different factors driving SCU in humans, we analyzed the relationships between functional categories, base composition, recombination, and gene expression. We first demonstrate that SCU is predominantly driven by large-scale variation in GC-content and is not linked to constraints on tRNA abundance, which excludes an effect of translational selection. In agreement with the gBGC model, we show that differences in SCU among functional categories are explained by variation in intragenic recombination rate, which, in turn, is strongly negatively correlated to gene expression levels during meiosis. Our results indicate that variation in SCU among expression levels during meiosis. Our results indicate that variation in SCU among functional categories (including variation associated to differentiation or proliferation) result from differences in levels of meiotic transcription, which interferes with the formation of crossovers and thereby affects gBGC intensity within genes. Overall, the gBGC model explains 81.3% of the variance in SCU among genes. We argue that the strong heterogeneity of SCU induced by gBGC in mammalian genomes precludes any optimization of the tRNA pool to the demand in codon usage.

Lab of Molecular Phylogeny and Evolution in Vertebrates

4) Codon Usage in Coding Sequences

What can explain the bias in synonymous codons usage?



Explanation 2:
Bias in the substitution rate among nucleotides.

Examples:

1. In some bacteria (*Mycoplasma* sp.),
the substitution rate is biased in favor of A or T
(Muto at Osawa, 1987)

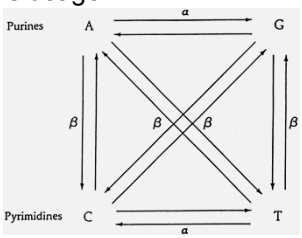
G, C ⇀↽ A, T

Consequence on synonymous codon usage?

All 3rd codon positions are A or T, leading to a bias in the use of synonymous codons

4) Codon Usage in Coding Sequences

What can explain the bias in synonymous codons usage?

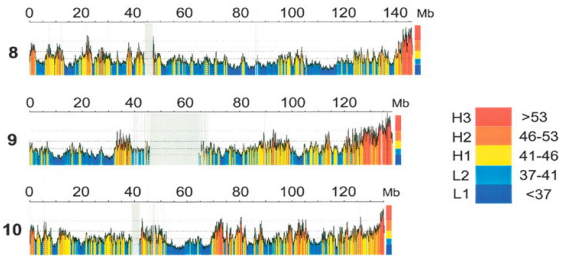


Explanation 2:
Bias in the substitution rate among nuceotides.

Examples:

2. Isochores in mammals:
alternating chromosomal regions with
high G+C content (60%) and low G+C content (30%).

The substitution bias can thus **vary within a genome**!



An isochore map of human chromosomes, *Constantini et al., 2006*

4) Codon Usage in Coding Sequences

Statistical measure of the synonymous codon usage bias.

The ***RSCU*** (Relative Synonymous Codon Usage) **characterizes a codon** and gives its usage level as compared to the other synonymous codons in a given gene (or set of genes).

$$RSCU_i = \frac{Obs_i}{Exp_i}$$

where:
Obs*i* is the observed number of codons *i* in the gene
Exp*i* is the expected number of codons *i* in the gene.

Exp*i* is obtained as follows:

$$Exp_i = \frac{\sum aa_i}{\sum syn_i}$$

occurrence of the aa encoded by the codon *i* in the gene
---
number of **synonymous codons** encoding that aa

4) Codon Usage in Coding Sequences

Statistical measure of the synonymous codon usage bias.

The **CAI** (Codon Adaptation Index) **characterizes a gene** Sharp and Li (1987).

The CAI uses a set of highly expressed genes as a reference for the usage of synonymous codons. The CAI varies from 0 to 1.

High CAI (0.6 to 1) indicates that a gene "behaves" like a highly expressed gene (i.e. has a similar bias in synonymous codon usage).

This Index is used for:
- estimating the expression level of a gene
- comparing the synonymous codon usage bias among organisms

$$\ln(\mathrm{CAI}) = \sum_{i=1}^{61} f_i \ln W_i$$

fi is the relative frequency of codon *i* in the analyzed sequence

Wi is the ratio of the frequency of the codon *i* over the frequency of the most frequently used synonymous codon in the set of highly expressed genes.

4) Codon Usage in Coding Sequences

CAI (Codon Adaptation Index) Sharp and Li (1987)

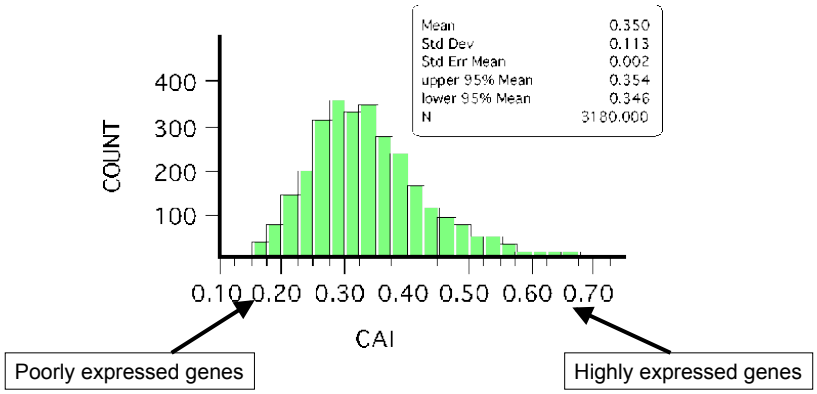


Figure: Distribution of CAI for 3180 *E. coli* genes

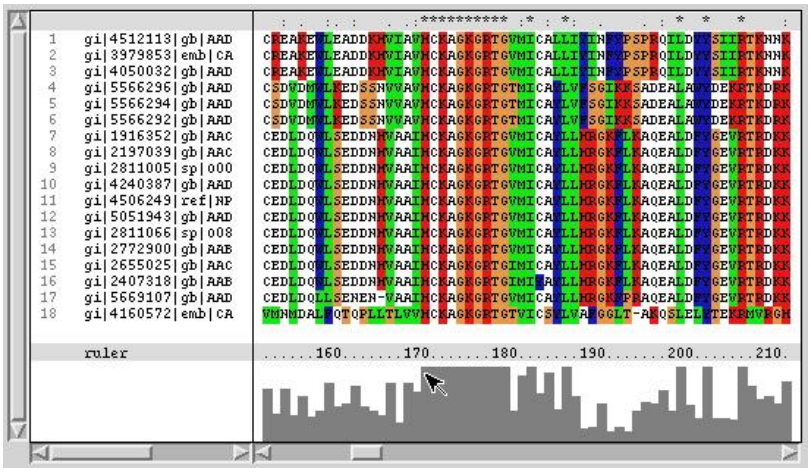
Molecular evolution

- Molecular basis of evolution.
- **Evolution of protein sequences**
- Evolution of DNA sequences

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

How to measure the difference between two sequences ?



ClustalX window

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

How to measure the difference between two sequences ?

A simple measure is to count the number of different amino acids and divide it by the total number of positions compared = ***p* distance**

$$p = n_d / n$$

p = 3/30 = 0.1

n_d – number of amino acid differences between two sequences; n – number of aligned amino acids.

The ***p* distance** is an exact measure of divergence if:
- all the changes that occurred during the independent evolution of the two sequences can be observed today.

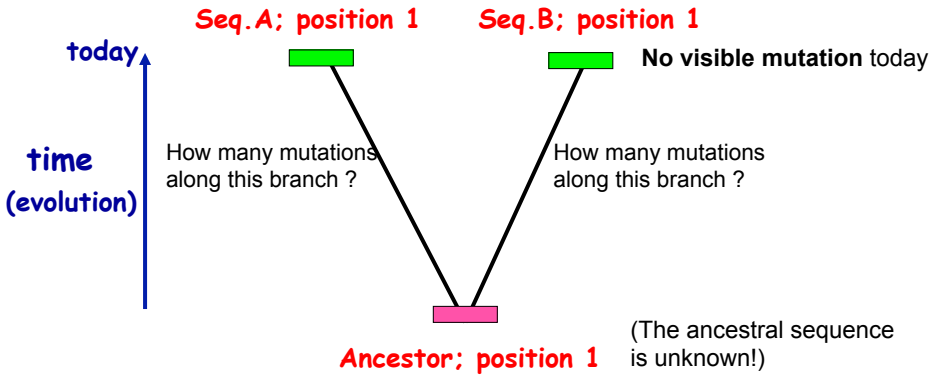
Is this always the case ???

Lab of Molecular Phylogeny and Evolution in Vertebrates

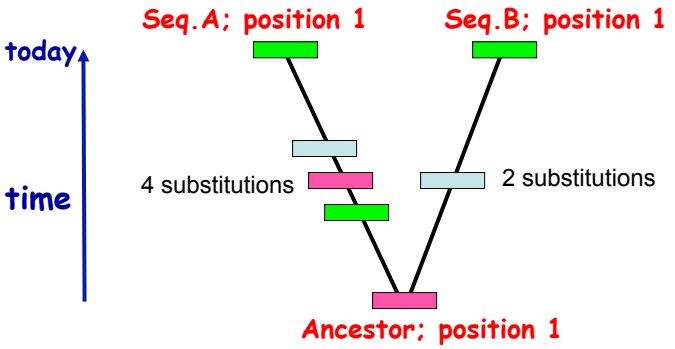
Evolution of protein sequences

How to measure the difference between two sequences ?

Are all mutations visible today ??



Evolution of protein sequences



The number of hidden substitutions increases with time.

→ The ***p* distance** is not proportional to time.

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

The ***p* distance** must be "corrected" to give an accurate measure of divergence.

How to correct the ***p* distance**?

Two ways:

- Theoretical correction: Poisson corrected distance (PC)

- Empirical correction: substitution matrices (PAM, JTT,...)

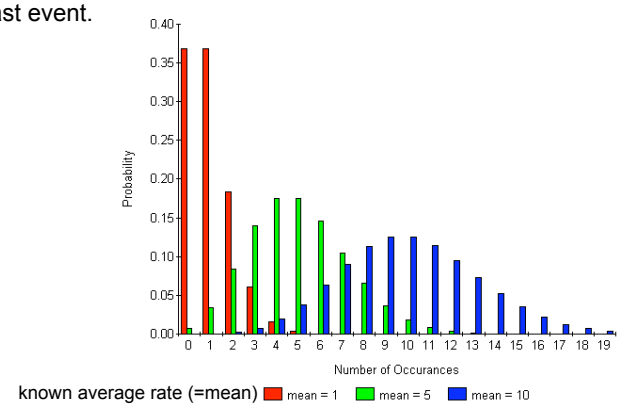
Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Theoretical correction: Poisson corrected distance (PC)

The Poisson distribution is a discrete probability distribution.

It expresses the probability of a number of events occurring in a fixed period of time if these events occur with a **known average rate** (λ) and independently of the time since the last event.



Lab of Molecular Phylogeny and Evolution in Vertebrates

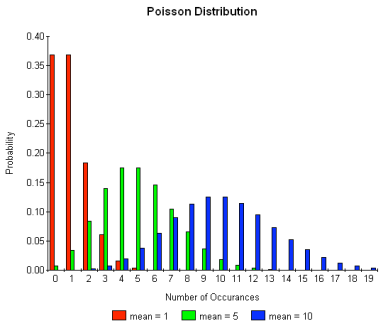
Evolution of protein sequences

Theoretical correction: Poisson corrected distance (PC)

The probability of having exactly ***k*** events (***k*** = 0, 1, 2, 3,...) is given by:

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!},$$

where:
e is the base of the natural logarithm (*e*=2.71828….)
k! is the factorial of *k*
λ is the average number of occurrences that occur during the given time interval

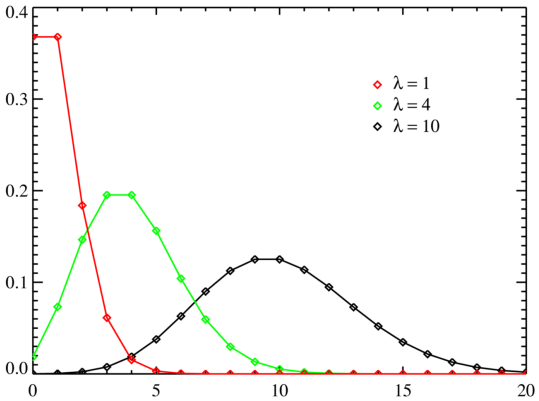


λ is the only variable that determines the distribution.

Evolution of protein sequences

Poisson distribution as a fonction of λ.

(λ = average number of occurrences that occurs during the given time interval)



Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Theoretical correction: Poisson corrected distance (PC)

For us, what is the value of λ ?

(λ = mean number of occurrences that occur during the given interval)

Let's suppose that ***r*** is the substitution rate of any amino acid to any another amino acid during a year, and that this rate is the same for all sites
r = mean number of events per year

The mean substitution rate per site and for a period of time ***t*** years is: ***rt***
rt = mean number of events over a period of ***t*** years

For us, **λ** is thus equal to ***rt***

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Theoretical correction: Poisson corrected distance (PC)

$$f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!},$$

The probability of no change (***k*** = 0) along time ***t*** for a given site is:

$$P(0;t) = f(0;rt) = e^{-rt}$$

(remember that $n^0 = 1$ and that 0! = 1)

This equation is not useful because the ancestral sequence at ***t* = 0** is unknown.

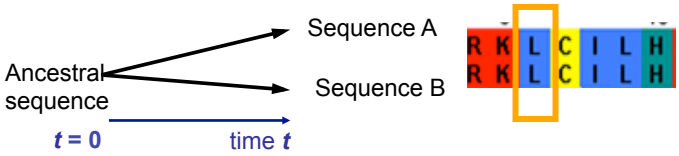
Solution: the proportion of **unchanged** amino acids **between two sequences** along time ***t*** is empirically obtained.

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Theoretical correction: Poisson corrected distance (PC)

The amount of unchanged amino acids is calculated by comparing two sequences that diverged since a time ***t*** .



The probability (***q***) that none of the two aa of an homologous site in sequences A and B has been substituted along time ***t*** is given by:

$$q = (e^{-rt}) \times (e^{-rt}) = e^{-2rt}$$

The probability ***q*** can be estimated by ***q*** = 1- ***p***

p* = *p distance = proportion of observed differences between sequences A and B

Evolution of protein sequences

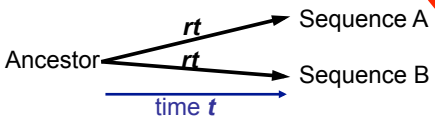
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Theoretical correction: Poisson corrected distance (PC)

Using equation ***q*** = 1- ***p***
and ***q*** = e^{-2rt}

we can derive 1- ***p*** = e^{-2rt} and by extracting the exponent $2rt = -\ln(1-p)$

Also, the "true" distance separating sequences A and B is: ***d*** = 2 ***rt***
(r = substitution rate)



Thus, $d = -\ln(1-p)$ This is the Poisson corrected distance !

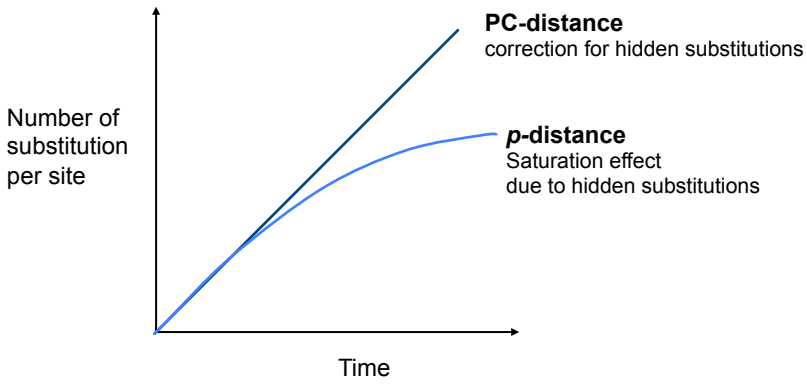
The variance is: $V(d) = p/[(1-p)n]$

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of protein sequences

Theoretical correction: Poisson corrected distance (PC)

Relationship between ***p*-distance** and **Poisson corrected-distance** (PC)



Evolution of protein sequences

The **Poisson-corrected distance** method assumes that:

-the substitution probability is equal across all sites of the polypeptide sequence.

Is this assumption generally correct ? **NO**

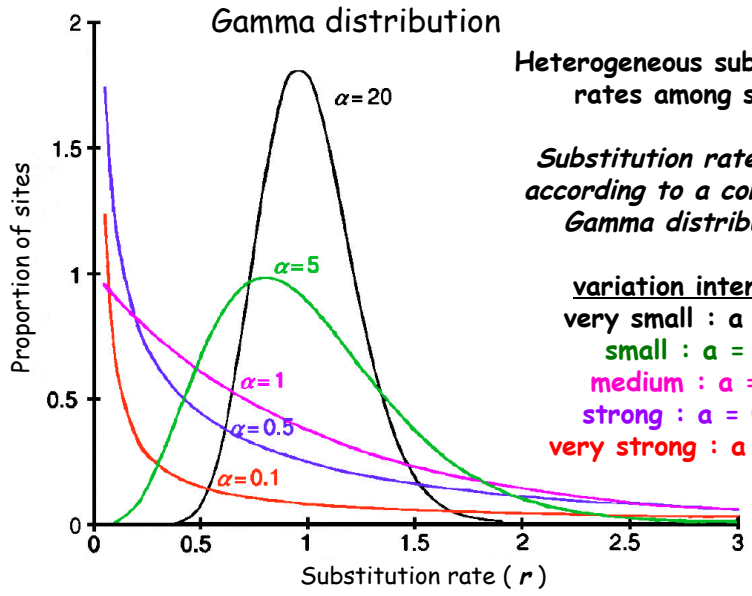
The distribution of the number of substitution per site **(*k*)** has a variance larger than the variance of the Poisson distribution (its an overdispersed Poisson distribution).

The distribution of ***k*** corresponds better to a **negative binomial distribution,** which has one more parameter than the Poisson distribution (Uzzell and Corbin, 1971).

→ This implies that the substitution rate (***r***) varies among sites according to a **gamma distribution**.

Gamma distribution: a single parameter (α) defines the shape of the distribution.

Evolution of protein sequences



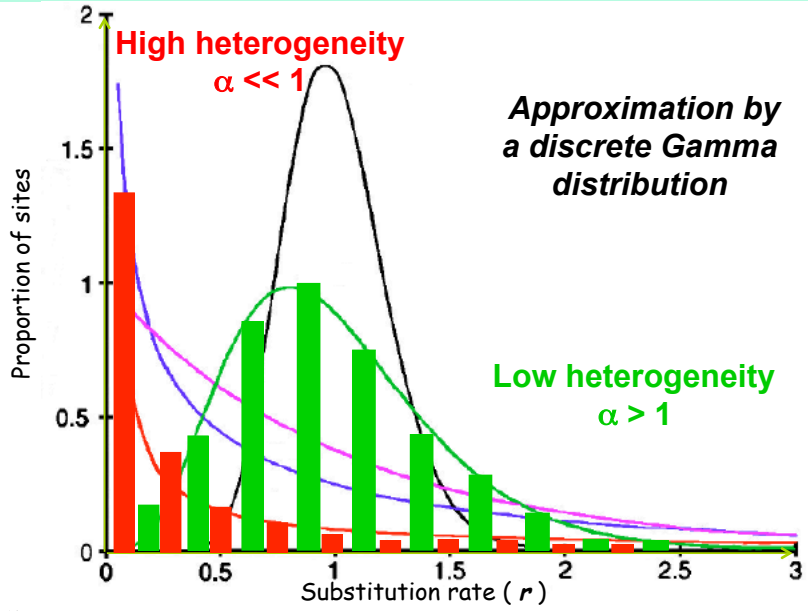
Heterogeneous substitution rates among sites

Substitution rates vary according to a continuous Gamma distribution

variation intensity
very small : a = 20
small : a = 5
medium : a = 1
strong : a = 0,5
very strong : a = 0,1

Lab of Molecular Phylogeny and Evolution in Vertebrates

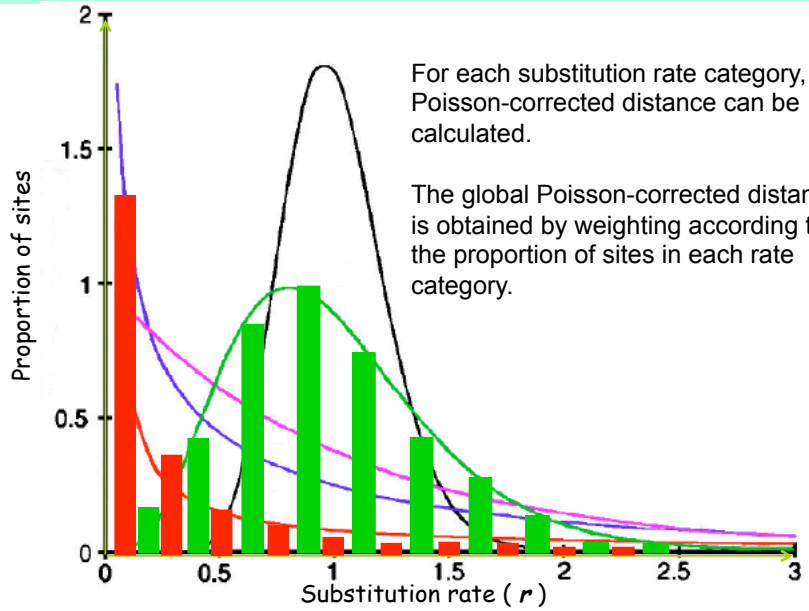
Evolution of protein sequences



Approximation by a discrete Gamma distribution

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of protein sequences



For each substitution rate category, a Poisson-corrected distance can be calculated.

The global Poisson-corrected distance is obtained by weighting according to the proportion of sites in each rate category.

Evolution of protein sequences

Examples of Poisson-corrected distance without / with incorporation of substitution rate heterogeneity across sites.

hemoglobin alpha chain

	P-distance	PC-distance	PC + Gamma-distance
Human/cow	0.121	0.129	0.134
Human/kangaroo	0.186	0.205	0.216
Human/carp	0.486	0.665	0.789

However:

The **Poisson-corrected + Gamma** model of aa sequence evolution is still to simplistic.

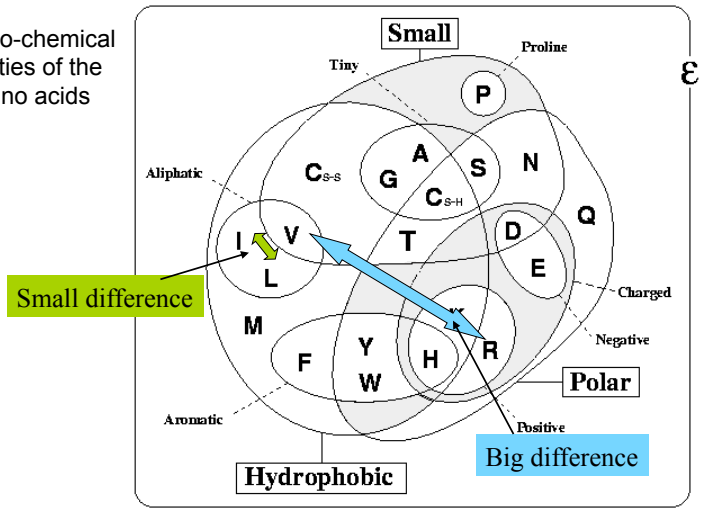
Empirical corrections fit better to the complexity of the evolution of aa sequences.

Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Empirical distance correction: substitution matrices (PAM, JTT,...)

Physico-chemical properties of the 20 amino acids



Evolution of protein sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Empirical distance correction: substitution matrices (PAM, JTT,...)

Empirical matrix of substitution probability of one aa to another aa, based on the analysis of a large amount of comparisons.
The matrix gives "relatedness" scores.

C	Cys	12																
S	Ser	0	2															
T	Thr	-2	1	3														
P	Pro	-3	1	0	6													
A	Ala	-2	1	1	1	2												
G	Gly	-3	1	0	-1	1	5											
N	Asn	-4	1	0	-1	0	0	2										
D	Asp	-5	0	0	-1	0	1	2	4									
E	Glu	-5	0	0	-1	0	0	1	3	4								
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4							
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6						
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6					
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5				
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6			
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5		
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4
F	Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

PAM 250 (Point Accepted Mutation) substitution matrix

PAM 250:

$M_{ij} = 10(\log_{10} R_{ij})$,

where M_{ij} is the matrix element and R_{ij} is the probability of that substitution as observed in the database, divided by the normalized frequency of occurence for amino acid i. All of the number are rounded to the nearest integer.

Lab of Molecular Phylogeny and Evolution in Vertebrates

Molecular evolution

- Molecular basis of evolution.

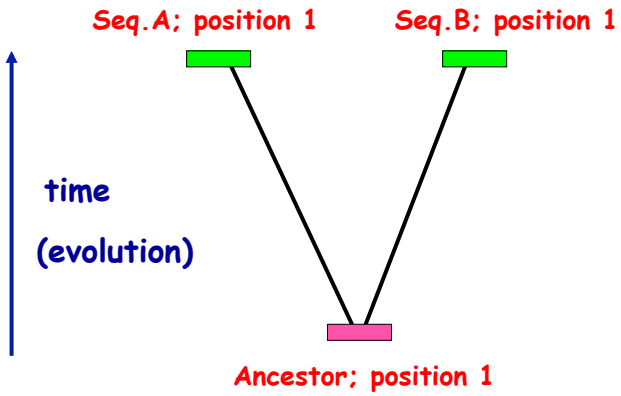
- Evolution of protein sequences

- Evolution of DNA sequences

Evolution of DNA sequences

Observed distance or ***p***-distance:

Number of different nucleotides / Number of compared nucleotides



Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Calculating *p*-distance

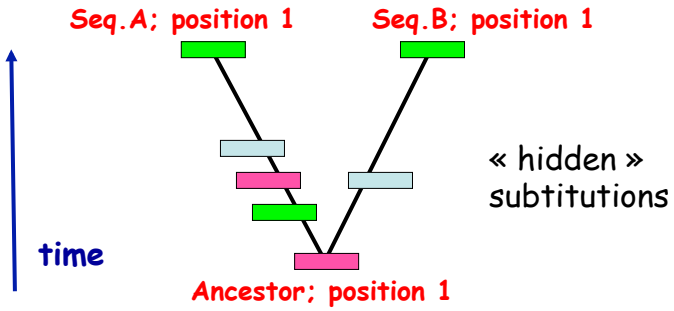
```
Séq X   C G A T G A T A A C
Séq Y   C G A A A A C A G C
            * *   *     *
```

$$D_{xy} = \frac{\text{Nb of different sites (k)}}{\text{Nb of sites compared (N)}} = 4 / 10 = 0.4$$

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

p-distance vs evolutionary distance



The amount of hidden subtitutions increases with evolutionary time.

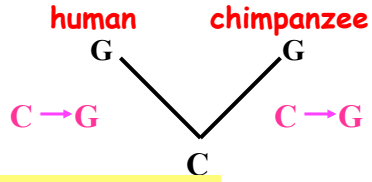
This problem is greater for DNA sequences (4 possible states per site) than for amino acid sequences (20 possible states per site).

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

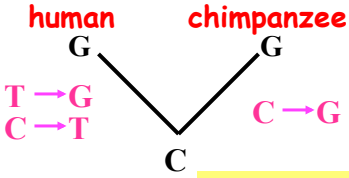
Hidden substitutions

Parallele ubstitution



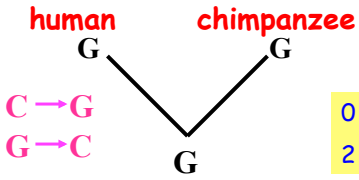
0 observed subt.
2 real subt.

convergente substitution



0 observed subt.
3 real subt.

Reverse subtitution



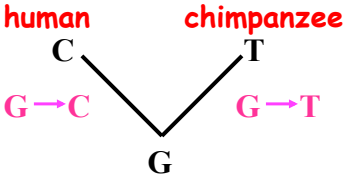
0 observed subt.
2 real subt.

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

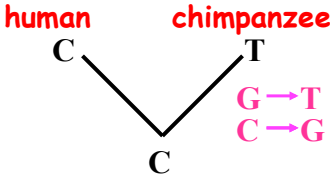
Hidden substitutions

Coincidence Substitutions



1 observed subt.
2 real subt.

Multiple Substitutions



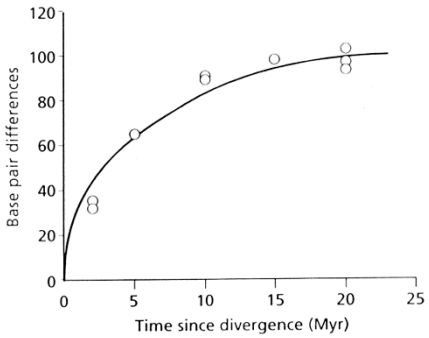
1 observed subt.
2 real subt.

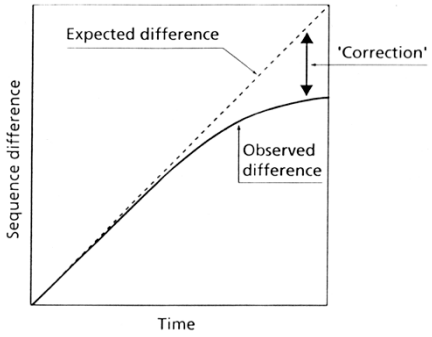
Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Correcting *p*-distance





Page & Holmes, Molecular Evolution, 1998

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

To correct the *p*-distance, we need models of DNA sequence evolution.

First and simplest model of DNA sequence evolution:

the model of Jukes and Cantor, 1969 (JC69)

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Jukes-Cantor (JC69) model

Substitution Matrix $P_t =$

	A	C	G	T	
	.	α	α	α	A
	α	.	α	α	C
	α	α	.	α	G
	α	α	α	.	T

a single parameter: α

Parameters: $pA = pT = pG = pC$ and $\alpha = \beta$

α : transistions
β : transversions

Formula: corrected distance **$d_{xy} = -3/4 \ln (1-4/3\ D)$**

where **D** is the *p*-distance

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Example

Two sequences of 500 nucleotides differing by 50 substitutions have a *p*-distance of:

$D = 50 / 500 = 0.1$

The evolutionary distance calculated according to the Jukes and Cantor model is:

$d_{xy} = 0.1073$

The corrected number of substitutions is 53.66,

that is 3-4 substitutions that where not observed

Evolution of DNA sequences

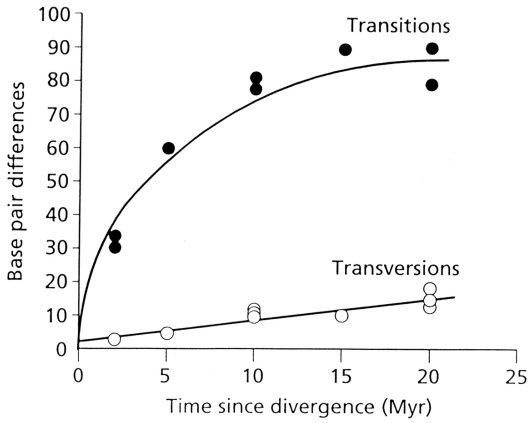
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

More realistic and complex models have been developed, with the help of the increase of computation power.

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Transitions are generally more frequent than transversions



	Ts/Tv ratio
mt DNA	9.0
12 S rRNA	1.75
α- et β-globines	0.66

Accumulation of transitions and transversions in COI sequences of bovids

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Kimura 2 parameters model (K2P)

$P_t =$

	A	C	G	T	
	.	β	α	β	A
	β	.	β	α	C
	α	β	.	β	G
	β	α	β	.	T

2 free parameters:

α = **transitions**

β = **transversions**

Parameters: **pA = pT = pG = pC**

- **α ≠ β**

- **Formula:** **d xy = - 1/2 ln (1-2P-Q) + 1/4 ln (1-2Q)**
- **where P = proportion of transitions; Q = proportion of transversions**

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

More complex models of sequence evolution

F81 Model (Felsenstein, 1981)

$\pi A \neq \pi C \neq \pi G \neq \pi T$ $\quad \alpha = \beta$

4 free parameters

HKY85 Model (Hasegawa, Kishino, Yano, 1985)

$\pi A \neq \pi C \neq \pi G \neq \pi T$ $\quad \alpha \neq \beta$

5 free parameters

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

The more complex model of sequence evolution

GTR Model
General Time Reversible
(Simon and Tavaré,1986)

$$P_t = \begin{vmatrix} . & \mathbf{a}(pAC) & \mathbf{b}(pAG) & \mathbf{c}(pAT) \\ \mathbf{a}(pCA) & . & \mathbf{d}(pCG) & \mathbf{e}(pCT) \\ \mathbf{b}(pGA) & \mathbf{d}(pGC) & . & \mathbf{f}(pGT) \\ \mathbf{c}(pTA) & \mathbf{e}(pTC) & \mathbf{f}(pTG) & . \end{vmatrix}$$

Parameters: $\pi A \neq \pi C \neq \pi G \neq \pi T$

Every subtitution has its own probability

a = A ↔ C **b = A ↔ G** **c = A ↔ T**
d = C ↔ G **e = C ↔ T** **f = T ↔ G**

9 free parameters

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Most phylogenetic methods assume that:

1. Substitutions are independent from one another.
2. The substitution rate is constant in time and among lineages.
3. Base frequency is homogeneous among lineages.
4. The substitution rate is equal among sites and constant in time.

These assumptions are not always satisfied!

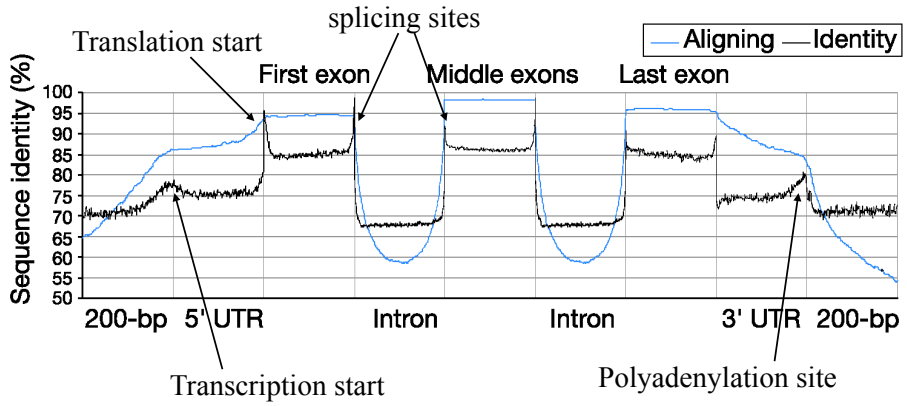
Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Substitution rate along a 'model' gene

Mean sequence identity of 3,165 human-mouse comparisons

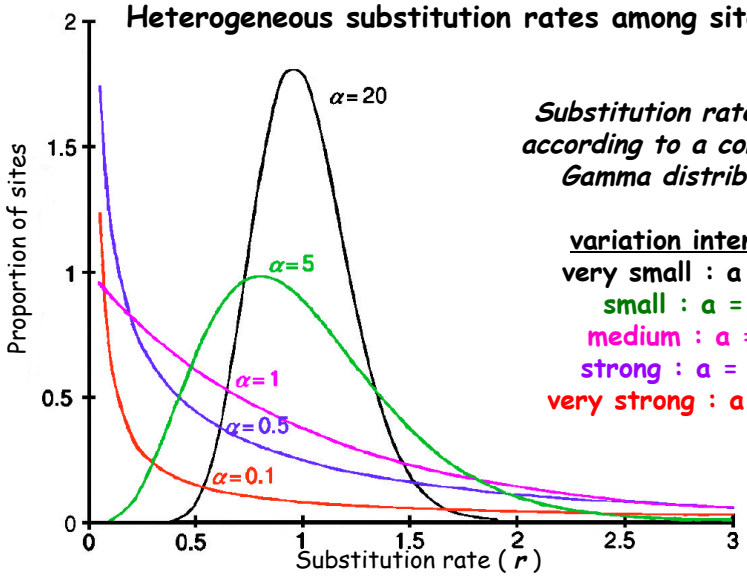


MGSC Nature (2002)

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Heterogeneous substitution rates among sites



Substitution rates vary according to a continuous Gamma distribution

<u>variation intensity</u>

very small : a = 20

small : a = 5

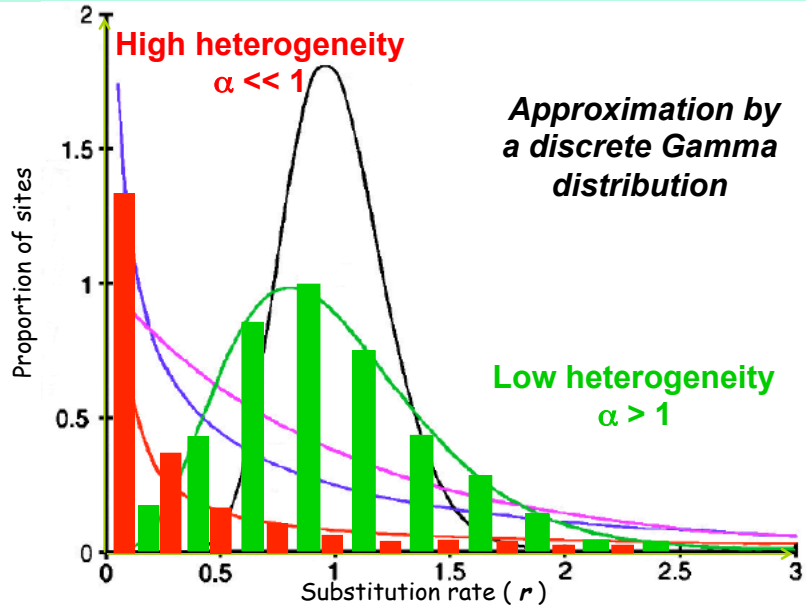
medium : a = 1

strong : a = 0,5

very strong : a = 0,1

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences



Evolution of DNA sequences

Estimation of the parameter α (Yang, 1996):

• albumin	1.05
• insulin	0.40
• prolactin	1.37
• 16S rRNA (stem)	0.29
• 16S rRNA (loop)	0.58
• 12S rRNA mt	0.16
• D-loop	0.17

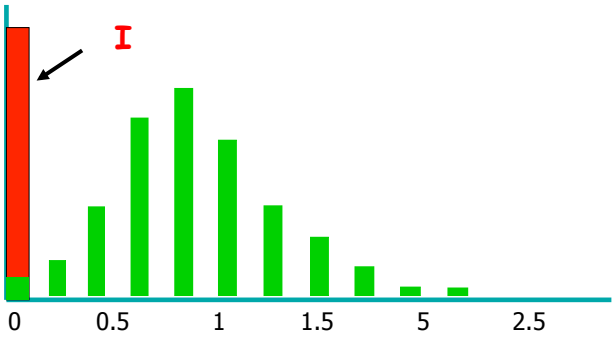
Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Invariant sites

Sometimes, the category of invariant sites does not fit to the distibution.

The category of invariant sites can then be treated as an additional parameter: **I**



The Gamma distribution is then used only for the remaining categories.

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

ModelTest: 56 models

Posada and Crandall,1998

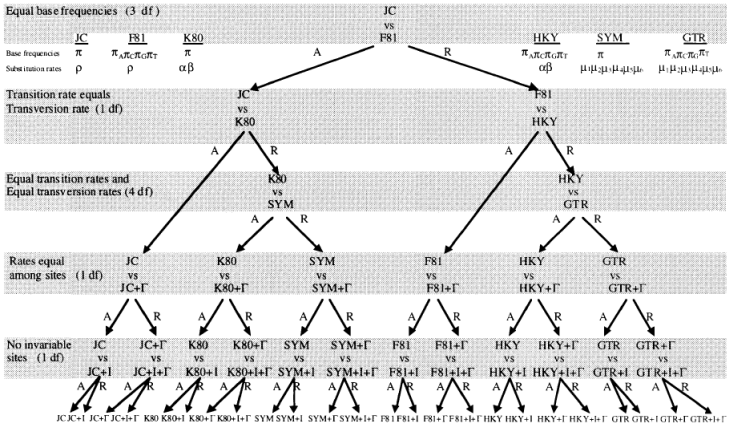


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. ^{1}: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

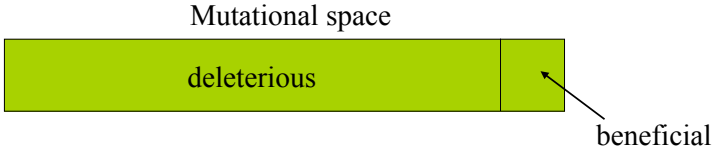
Selective pressure

acting on DNA sequences

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Functional view of mutations before the Neutral Theory of Evolution (Kimura 1968)



- Most mutations were thought to be deleterious, reducing the fitness of the organism
  - the fate of such mutations is to disappear via purifying selection.
- Some mutations might be beneficial, increasing the fitness of the organism
  - such mutations are kept via positive Darwinian selection (adaptive)

The fitness is a function of survival and fecundity (number of offspring)

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Motoo Kimura

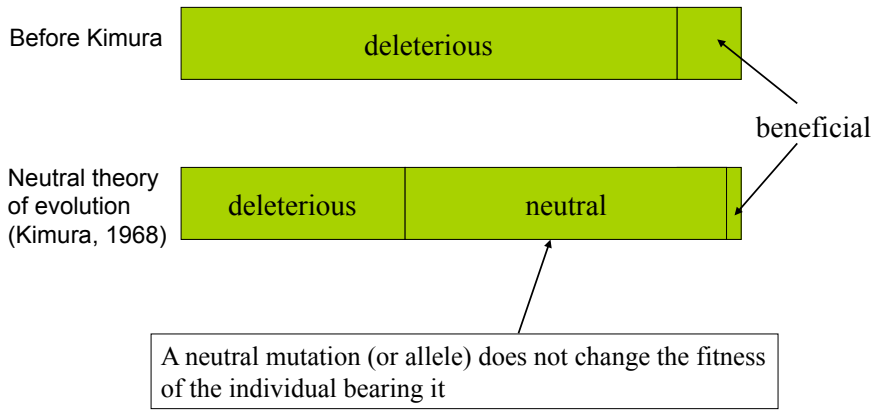
In 1968 Kimura analyzed the patterns of substitution in several genes (hemoglobin, cytochrom c and triosephosphate dehydrogenase)

Kimura (1968) *Nature* 217 624-626

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Kimura proposed that most of the substitutions are neutral (no change in fitness) and that very few have been kept due to positive Darwinian selection.



Kimura (1968) *Nature* 217 624-626

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Synonymous mutations can be neutral mutations

Predictions:

- If most changes in gene sequence were due to positive selection then we would expect that most changes would occur in 1st or 2nd codon positions (as they generally change the amino acid)
- If changes in gene sequence include neutral mutations, then 3rd codon positions would change at a higher rate because they are mostly synonymous mutations without any effect on fitness (=neutral mutations)

Which prediction is correct ?

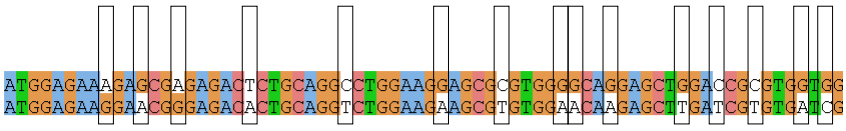
King, J. L., and Jukes, T. H. 1969. Non-Darwinian evolution, Science 164, 788-798.

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

Synonymous and non-synonymous substitution rates

- K_S = number of Synonymous substitutions per synonymous site
- K_A = number of non-synonymous substitutions (Altering) per non-synonymous site



It is possible to estimate K_a and K_s empirically, by examining the alignments.

However, appropriate methods have been developed to estimate K_a and K_s.

Lab of Molecular Phylogeny and Evolution in Vertebrates

Evolution of DNA sequences

Synonymous substitution rates are generally higher than non-synonymous substitution rates

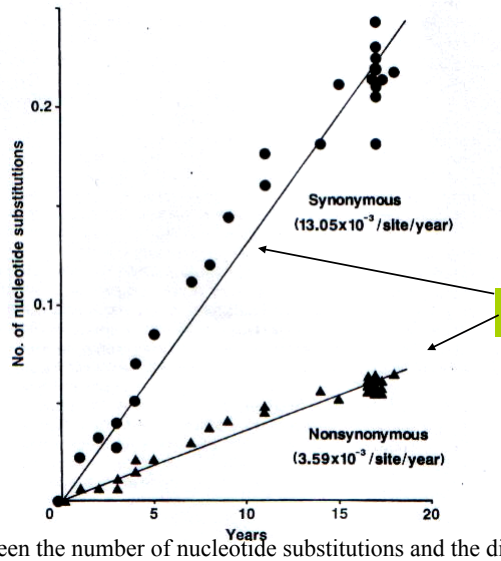
TABLE 8.1 Numbers of nucleotide substitutions per 100 sites between species[a]

	Synonymous sites		Nonsynonymous sites	
Species pair	K_S	L^b	K_A	L^b
Mouse–rat	18.0 ± 0.7	4,229	1.8 ± 0.1	15,217
Mouse–hamster	30.3 ± 1.0	4,229	2.9 ± 0.1	15,217
Rat–hamster	31.3 ± 1.0	4,229	2.7 ± 0.1	15,217
Mouse–human	53.4 ± 1.5	4,229	5.2 ± 0.2	15,217
Rat–human	51.6 ± 1.5	4,229	5.0 ± 0.2	15,217
Hamster–human	52.3 ± 1.5	4,229	5.1 ± 0.1	15,217

From O'hUigin and Li (1992).
[a]Computed by Li et al.'s (1985b) method.
[b]Number of sites compared.

From Li, W-H Molecular Evolution who took it from:
Oh' Uigin and Li. JME 1992 35: 377-384

The Molecular Clock of Viral Evolution



Different rates

Relationship between the number of nucleotide substitutions and the difference in the year of isolation for the H3 hemagglutinin gene of human influenza A viruses. All sequence comparisons were made with the strain isolated in 1968.

Gojobori et al. 1990 PNAS 87 10015-10018

Evolution of DNA sequences

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département de zoologie et biologie animale

The ratio Ka/Ks is used to assess the selective pressure acting on coding regions.

$K_A/K_s \approx 1$ -> neutral evolution, no selection.

$K_A/K_s > 1$ -> positive selection

$K_A/K_s << 1$ -> negative selection (purifying selection)