MOLECULAR EVOLUTION (CC-14, unit-3)

BY : Sriparna Ray
ZOOLOGY DEPARTMENT,
BIDHAN CHANDRA COLLEGE 2020

What is Molecular Evolution ?

Molecular evolution address two broad range of questions:

- 1. Use **DNA** to study the evolution of **organism**s, e.g. population structure, geographic variation and phylogeny
- 2. Use different **organisms** to study the evolution process of **DNA**

Molecular evolution

•The increasing available completely sequenced organisms and the importance of evolutionary processes that affect the species history, have stressed the interest in studying the molecular evolution events at the sequence level.

Molecular evolution

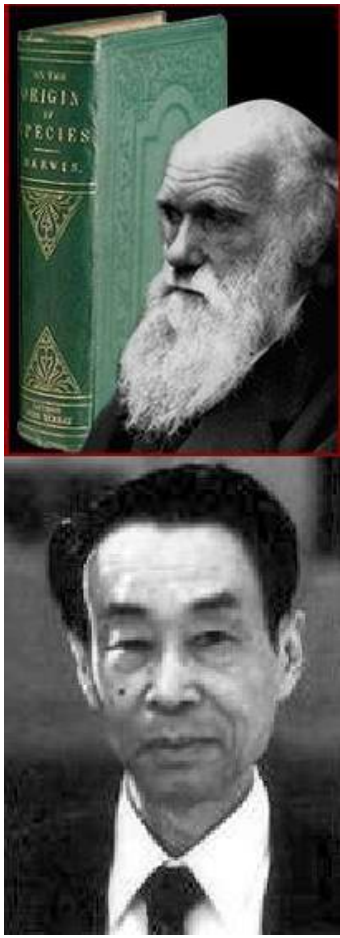
GACGACCATAGACCAGCATAG

GACTACCATAGA-CTGCAAAG

- **Mutations arise due to inheritable changes in genomic DNA sequence;**
- **Mechanisms which govern changes at the protein level are most likely due to nucleotide substitution or insertions/deletions;**
- **Changes may give rise to new genes which become fixed if they give the organism an advantage in selection;**

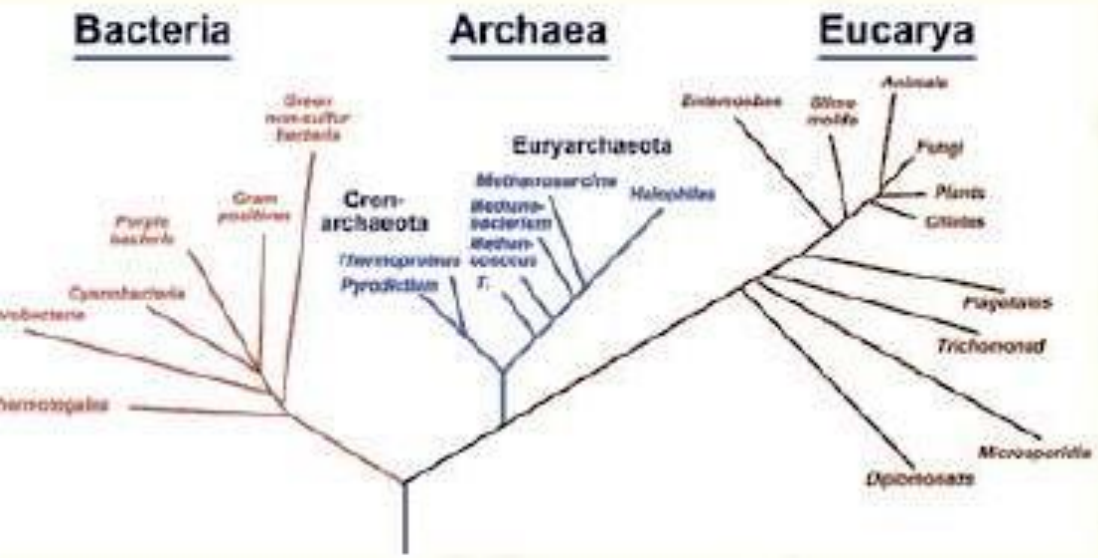
A brief historical perspective

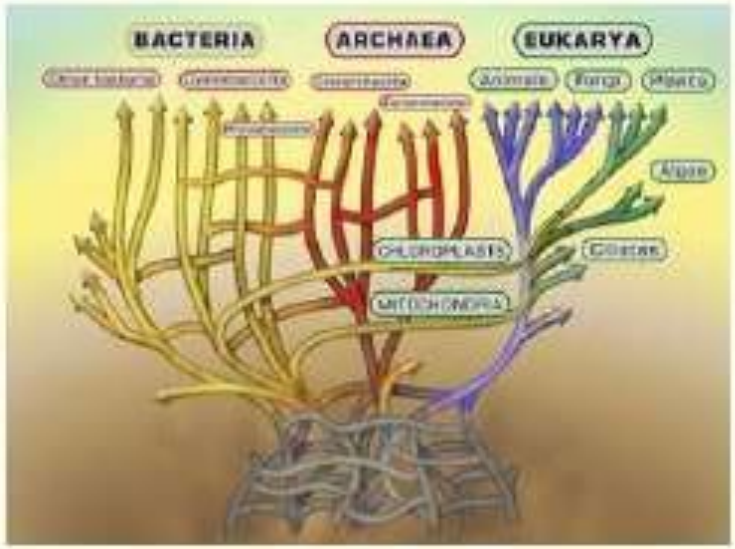
- Darwin first came up with the idea that living organisms are evolutionarily related
- Molecular evolution became a science following discovery of DNA and crack of genetic code
- Insulin: first protein sequenced (Sanger, 1955), and sequence compared across species.
- Neutral theory: Motoo Kimura, Thomas Jukes (1968,69)
- Effect of population size: Michael Lynch (2000s)



A brief historical perspective

- Until 1970s, cellular organisms were divided into eukaryotes (have nucleus) and prokaryotes (no nucleus)
- Using 16S rRNA gene sequence, Carl Woese redefined three domains





Ford Doolittle

- To recover evolutionary relationships from amino acid or nucleotide sequences, rigorous models of molecular evolution are needed.

Functional versus Evolutionary biology: "The molecular war"

- In 1961, Ernst Mayr argued for a clear distinction between two "*distinct and complementary*" pillars of biology:
- Functional biology, which considered proximate causes and asked "how" questions;
- Evolutionary biology, which considered ultimate causes and asked "why" questions;
- This reflects a "culture change" in biology after the emergence of molecular biology and biochemistry. It was in that context that Dobzhansky first wrote in 1964, "**nothing in biology makes sense except in the light of evolution**".

Similar statements …

- "Nothing in Evolution Makes Sense Except in the Light of Biology"

- "Nothing in Evolution Makes Sense Except in the Light of Domestication"

- "Nothing in Evolution Makes Sense Except in the Light of Population Genetics (in relation to population size)" – Michael Lynch

Mutations in DNA and protein

- **Synonymous mutation** -> do not change amino acid
- **Nonsynonymous mutation** -> change amino acid
- **Nonsense** mutation: point mutation resulting in a pre-mature stop codon
- **Missense** mutation: resulting in a different amino acid
- **Frameshift** mutation: insertion / deletion of 1 or 2 nucleotides
- **Silent** mutation: the same as nonsynonymous mutation

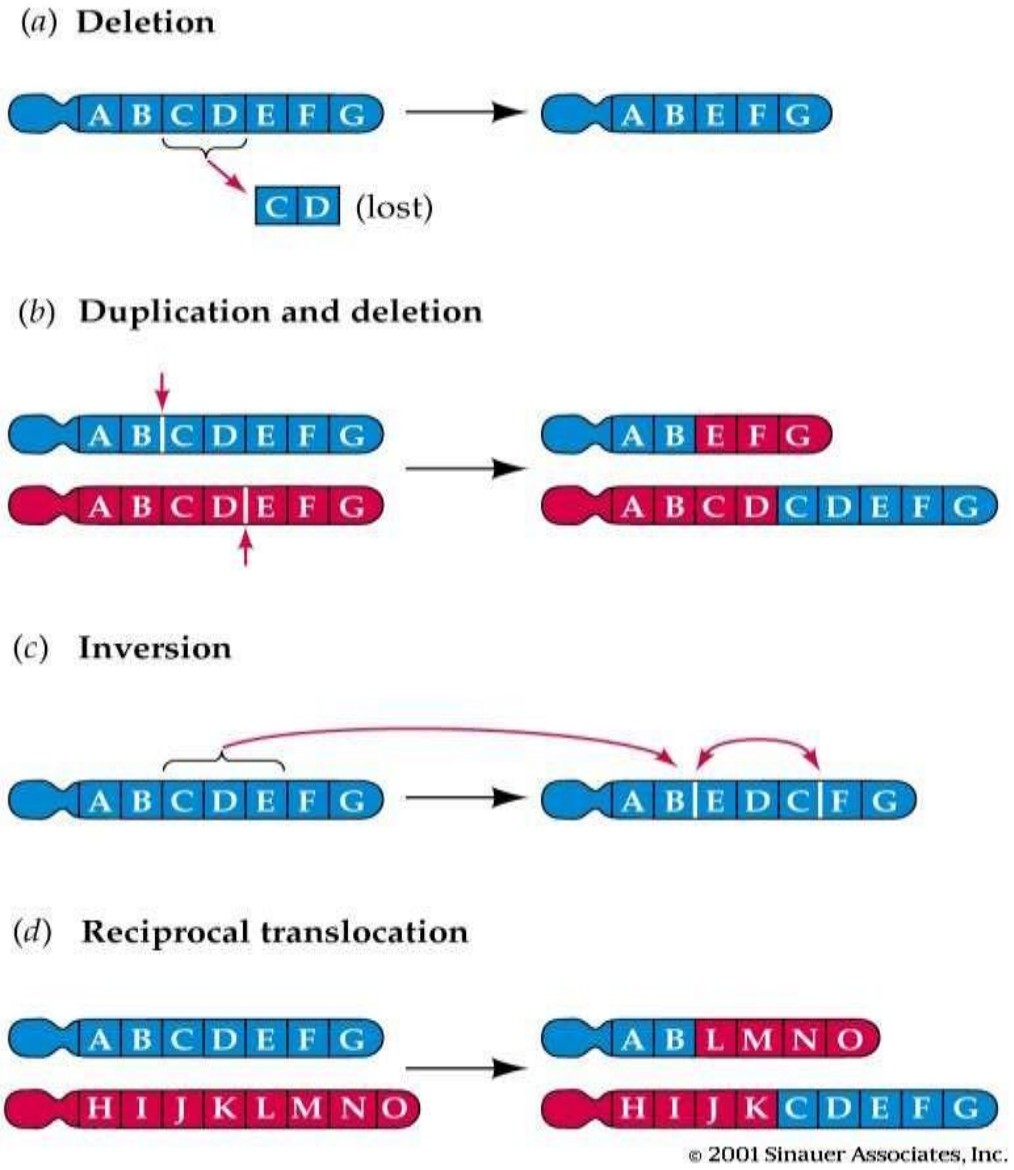
- Neutral mutation: mutation has no fitness effects, invisible to evolution (neutrality usually hard to confirm)
- Deleterious mutation: has detrimental fitness effect
- Beneficial mutation:

Fitness = ability to survive and reproduce

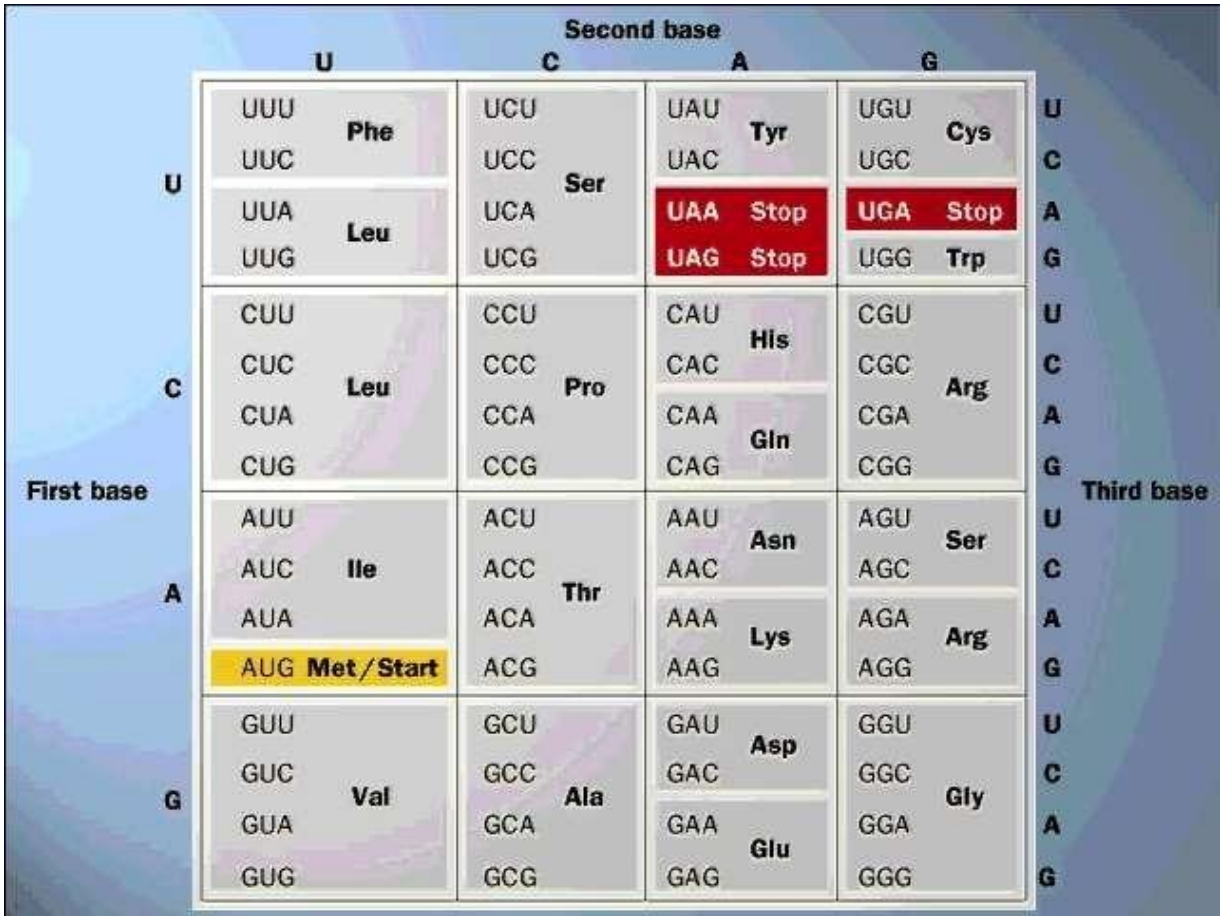
Consider molecular evolutionary changes at two levels

- Changes in ***DNA***;
  - Point mutation; mutations of single genes; small alterations in sequence or number of nucleotides
  - Chromsomal mutations; alterations that are more extensive than point mutations; four types – deletions duplications, inversions, translocations
  - scope extends from point mutations in introns or exons, to changes in the size and composition of genomes
- Changes in ***gene products***;
  - RNA
  - Proteins…polypeptide chains…amino acid sequences

Overview of four classes of chromosomal mutations



Degeneracy of genetic code



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

- **Because there are only 20 amino acids, but 64 possible codons, the same amino acid is often encoded by a number of different codons, which usually differin the third base of the triplet.**

- **Because of this repetition the genetic code is said to be degenerate and codons which produce the same amino acid are called synonymous codons.**

Molecular Evolution

	Morphological Similarity	Genetic Similarity
4.	low	high

***Pan,* Chimp**

***Homo,* Human**

-distance between humans and chimpanzees is less than between sibling species of *Drosophila.*

-for example, from a sample of 11 proteins representing 1271 amino acids, only 5 differ between humans and chimps.

-the other six proteins are identical in primary structure.

- most proteins that have been sequenced exhibit no amino acid differences - e.g., alphaglobin

Negative Selection and Positive Selection

- **Negative selection (purifying selection)**

– Selective removal of **deleterious mutations** (alleles)

– Result in **conservation** of functionally important amino acids

– Examples: ribosomal proteins, RNA polymerase, histones

- **Positive selection (adaptive selection, Darwinian selection)**

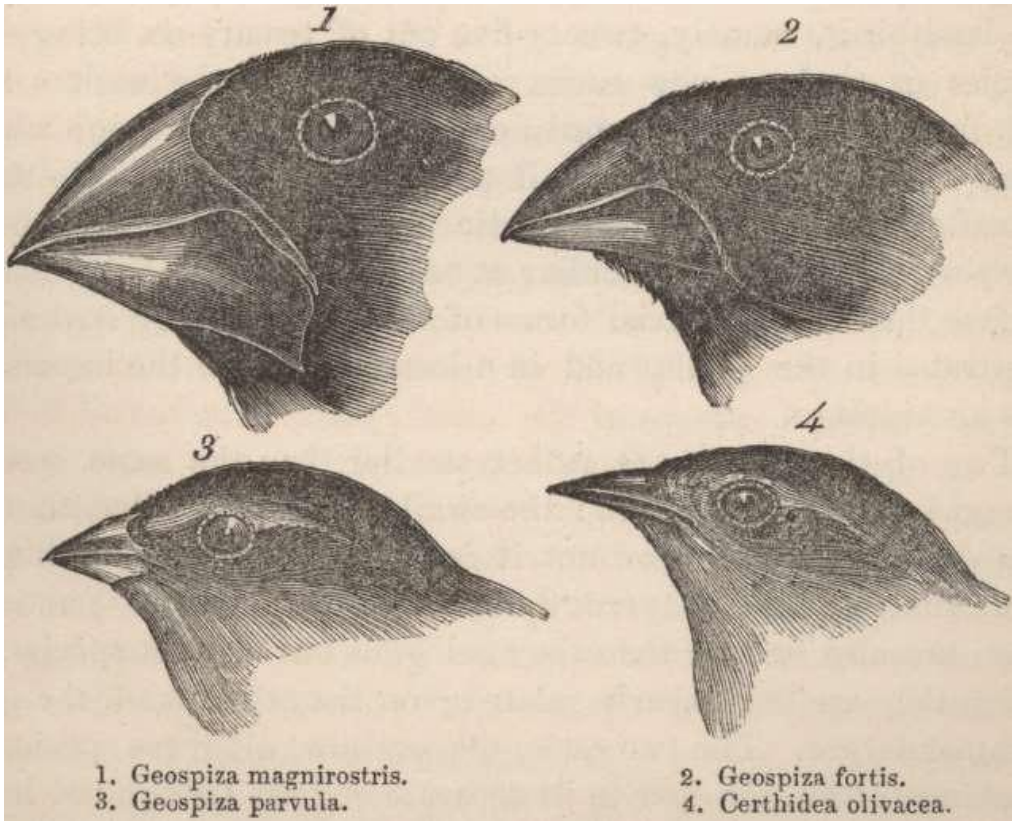
–Increase the frequency of **beneficial mutations** (alleles) that increase **fitness** (success in reproduction)

–Examples: male seminal proteins involved in sperm competition, membrane receptors on the surface of innate immune system

–Classic examples: Darwin's finch, rock pocket mice in Arizona (however the expression level of these genes instead of their protein sequence are targeted by selection)

The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches

Arhat Abzhanov[1]†, Winston P. Kuo[1,2,3]†, Christine Hartmann[4], B. Rosemary Grant[5], Peter R. Grant[5] & Clifford J. Tabin[1]



"They show that **calmodulin** (CaM), a molecule involved in mediating Ca^{2+} signalling, is expressed at **higher levels** in the long and pointed beaks of cactus finches than in more robust beak types of other species."

Nature 2006

The genetic basis of adaptive melanism in pocket mice

Michael W. Nachman*, Hopi E. Hoekstra, and Susan L. D'Agostino

The Developmental Role of Agouti in Color Pattern Evolution

Marie Manceau,[1,2] Vera S. Domingues,[1,2] Ricardo Mallarino,[1] Hopi E. Hoekstra[1,2]*

Nachman et al PNAS 2003
Manceau Science 2011

Purifying (negative) Selection

Seq1	AAG	ACT	GCC	GGG	CGT	ATT
Seq2	AAA	ACA	GCA	GGA	CGA	ATC

Seq1	K	T	A	G	R	I
Seq2	K	T	A	G	R	I

 # Synonymous substitutions = 6
 # Non-synonymous substitutions = 0

Ka / Ks
= Non-synonymous / Synonymous substitutions
= 0

Neutral Selection

Seq1	AAG	ACT	GCC	GGG	CGT	ATT
Seq2	AAA	ACA	GAC	GGA	CAT	ATG

Seq1	K	T	A	G	R	I
Seq2	K	T	D	G	H	M

 # Synonymous substitutions = 3
 # Non-synonymous substitutions = 3

Ka / Ks
= Non-synonymous/Synonymous substitutions
= 1

Positive Selection

Seq1	AAG	ACT	GCC	GGG	CGT	ATT
Seq2	AAA	ATT	GAC	GAG	CAT	ATG

Seq1	K	T	A	G	R	I
Seq2	K	I	D	E	H	M

 # Synonymous substitutions = 1
 # Non-synonymous substitutions = 5

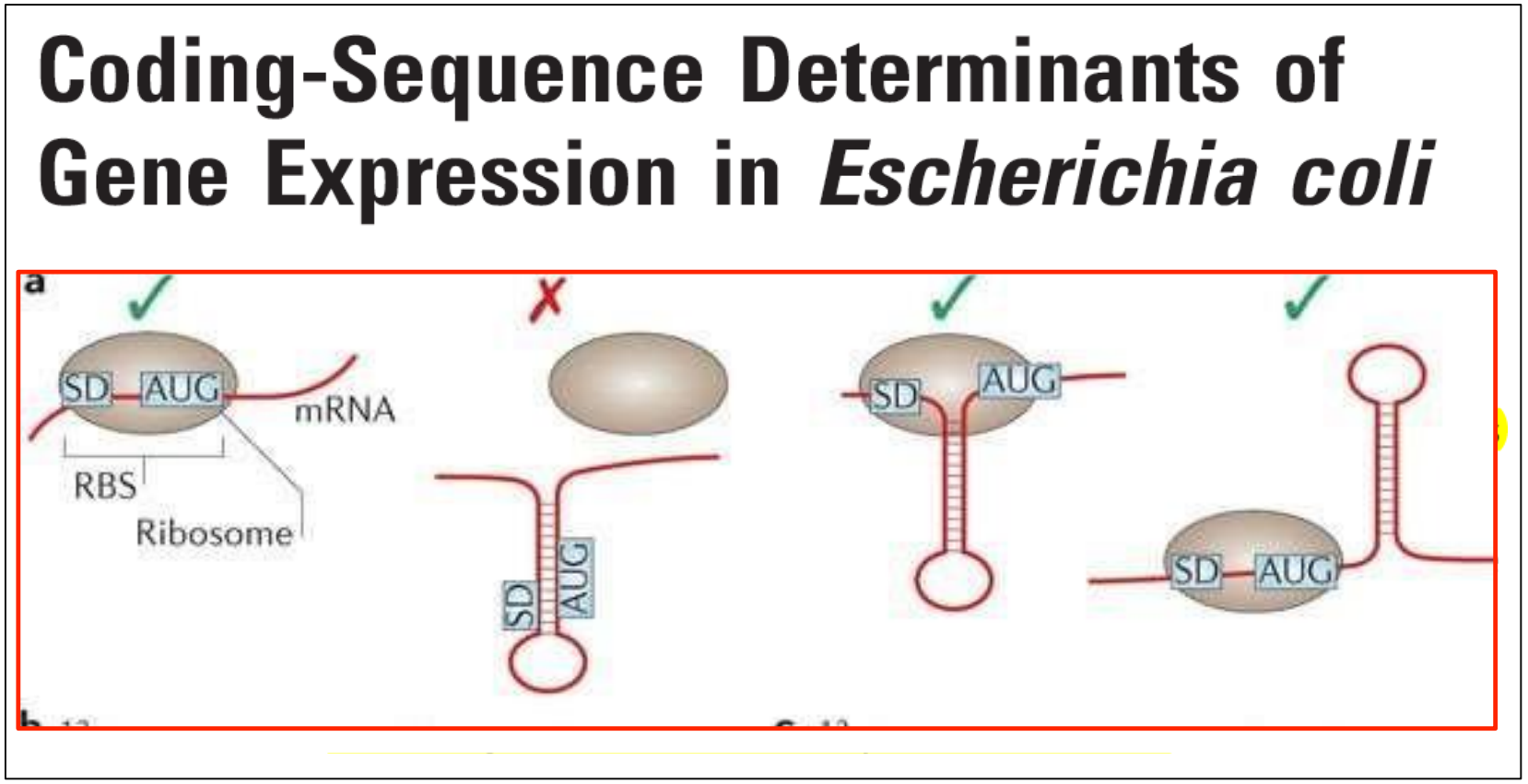
Ka / Ks
= Non-synonymous/Synonymous substitutions
=5

Synonymous substitutions are NOT always neutral

Different codons for the same amino acid may have different functional constraints and fitness effects

- Translational efficiency: codon usage bias
- RNA stability and correct folding of secondary structures
- RNA editing
- Protein folding
- Exon splicing regulatory motifs
- Binding sites for microRNA and RNA binding proteins (RBP)

Synonymous codons influence mRNA secondary structure and gene expression



Slotkin Kudla Nat Rev Gen, 2011

Synonymous codons influence mRNA secondary structure and gene expression

Coding-Sequence Determinants of Gene Expression in *Escherichia coli*

Grzegorz Kudla,[1*] Andrew W. Murray,[2] David Tollervey,[3] Joshua B. Plotkin[1†]

Synonymous mutations do not alter the encoded protein, but they can influence gene expression. To investigate how, we engineered a synthetic library of 154 genes that varied randomly at synonymous sites, but all encoded the same green fluorescent protein (GFP). When expressed in *Escherichia coli*, GFP protein levels varied 250-fold across the library. GFP messenger RNA (mRNA) levels, mRNA degradation patterns, and bacterial growth rates also varied, but codon bias did not correlate with gene expression. Rather, the stability of mRNA folding near the ribosomal binding site explained more than half the variation in protein levels. In our analysis, mRNA folding and associated rates of translation initiation play a predominant role in shaping expression levels of individual genes, whereas codon bias influences global translation efficiency and cellular fitness.

Slotkin Kudla Nat Rev Gen, 2011

"Rare codons" can influence protein structure

A "Silent" Polymorphism in the *MDR*1 Gene Changes Substrate Specificity

Chava Kimchi-Sarfaty,*† Jung Mi Oh,†‡ In-Wha Kim, Zuben E. Sauna,
Anna Maria Calcagno, Suresh V. Ambudkar, Michael M. Gottesman†

Synonymous single-nucleotide polymorphisms (SNPs) do not produce altered coding sequences, and therefore they are not expected to change the function of the protein in which they occur. We report that a synonymous SNP in the *Multidrug Resistance* 1 (*MDR*1) gene, part of a haplotype previously linked to altered function of the *MDR*1 gene product P-glycoprotein (P-gp), nonetheless results in P-gp with altered drug and inhibitor interactions. Similar mRNA and protein levels, but altered conformations, were found for wild-type and polymorphic P-gp. We hypothesize that the presence of a rare codon, marked by the synonymous polymorphism, affects the timing of cotranslational folding and insertion of P-gp into the membrane, thereby altering the structure of substrate and inhibitor interaction sites.

The Neutral Theory of Molecular Evolution

- **Motoo Kimura advanced the Neutral Theory of Molecular Evolution in 1968. Two observations underlie the theory**
- 1. Most natural populations harbor high levels of genetic variation *higher than would be expected* if natural selection were the evolutionary force primarily responsible for influencing the level of genetic variation in populations
- 2. Many mutations in sequences of genes do not alter the proteins encoded by those genes
  - virtually always true for synonomous substitutions
  - sometimes true for non-synonomous substitutions
  - If protein function is not altered by a mutation, the allelic variant that results from that mutation is unlikely to be influenced by natural selection…

The Neutral Theory of Molecular Evolution

The Neutral Theory holds that, because most mutations are selectively neutral at the molecular level..

- the majority of evolutionary change that macromolecules undergo results from random genetic drift

- much of the variation within species results from random genetic drift (variation in the relative frequency of different genotypes in a small population, owing to the chance disappearance of particular genes as individuals die or do not reproduce).

- Kimura developed a mathematical model showing that the rates at which neutral substitutions accumulate is a function of the mutation rate (not to selection forces)

- by this theory, levels of molecular variation in genomes are strongly influenced by a balance between mutation, which generates variations, and genetic drift, which can eliminate it.

Neutral theory of evolution

- Using sequence data of hemoglobin, insulin, cytochrome *c* from many vertebrates, Motoo Kimura calculated on average sequence evolution in mammals had been very rapid: 1 amino acid change every 1.8 years

- Such a high mutation frequency suggest the majority of substitutions have no fitness effects, i.e. selectively neutral, and are created by genetic drift.

- Rate of molecular evolution is equal to the neutral mutation rate, this gives rise to the concept of "molecular clock"

“The Neutralist-Selectionist debate”

- **Agree:**
  - Most mutations are deleterious and are removed.
  - Some mutations are favourable and are fixed.

- **Neutral theory**
  - Advantageous (adaptive) mutations are very rare
  - Most of the amino acid changes and polymorphisms are neutral, and created by genetic drift.
  - The concept of Molecular clock

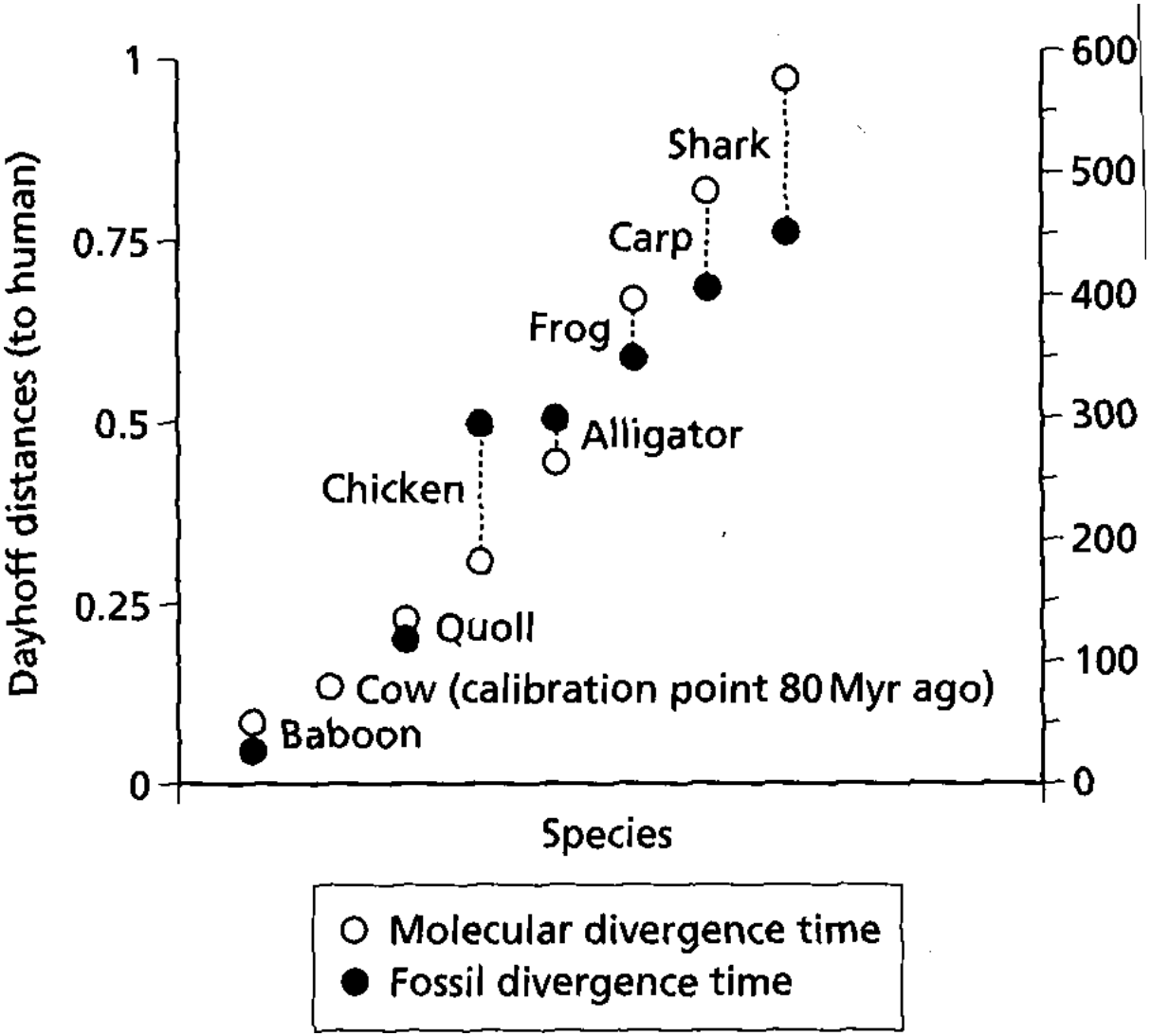
- **Selectionist theory**
  - Advantageous mutations are more common
  - Molecular evolution will are dominated by selection
  - No Molecular clock

Evidence supporting neutral evolution

- Pseudogenes (dead genes that have no function and no fitness effect) evolve very fast.
- Synonymous codon positions (3-fold, 4-fold degenerate sites) evolve faster than non-synonymous sites, and should evolve with a constant rate. (not always true, see previous slides)
- Genes that have important functions should evolve slower.

Evidence for Rate Constancy

in Hemoglobin



Large carniverous marsupial

Page and Holmes p229

The Molecular Clock Hypothesis

- The molecular clock is figurative term for a technique that uses the mutation rate of biomolecules to deduce the time in prehistory when two or more life forms diverged.
- Amount of genetic difference between sequences is a function of time since separation.
- Rate of molecular change is constant (enough) to predict times of divergence
- It is sometimes called a **gene clock** or an **evolutionary clock**.

Outline

- Methods for estimating time under a molecular clock
  - Estimating genetic distance
  - Determining and using calibration points
  - Sources of error
- Rate heterogeneity
  - reasons for variation
  - how its taken into account when estimating times
- Reliability of time estimates
- Estimating gene duplication times

Measuring Evolutionary time with a molecular clock

1. Estimate genetic distance

   d = number amino acid replacements

2. Use paleontological data to determine date of common ancestor

   T = time since divergence

3. Estimate calibration rate (number of genetic changes expected per unit time)

   $r = d / 2T$

4. Calculate time of divergence for novel sequences

   $T = d / 2r$

Five factors combine to limit the application of molecular clock models:

- Changing generation times (If the rate of new mutations depends at least partly on the number of generations rather than the number of years)
- Population size (Genetic drift is stronger in small populations, and so more mutations are effectively neutral)
- Species-specific differences (due to differing metabolism, ecology, evolutionary history, ...)
- Change in function of the protein studied (can be avoided in closely related species by utilizing non-coding DNA sequences or emphasizing silent mutations)
- Changes in the intensity of natural selection.

Applying a Molecular Clock: The Origin of HIV

- Phylogenetic analysis shows that HIV is descended from viruses that infect chimpanzees and other primates
- HIV spread to humans more than once
- Comparison of HIV samples shows that the virus evolved in a very clocklike way
- Application of a molecular clock to one strain of HIV suggests that that strain spread to humans during the 1930s

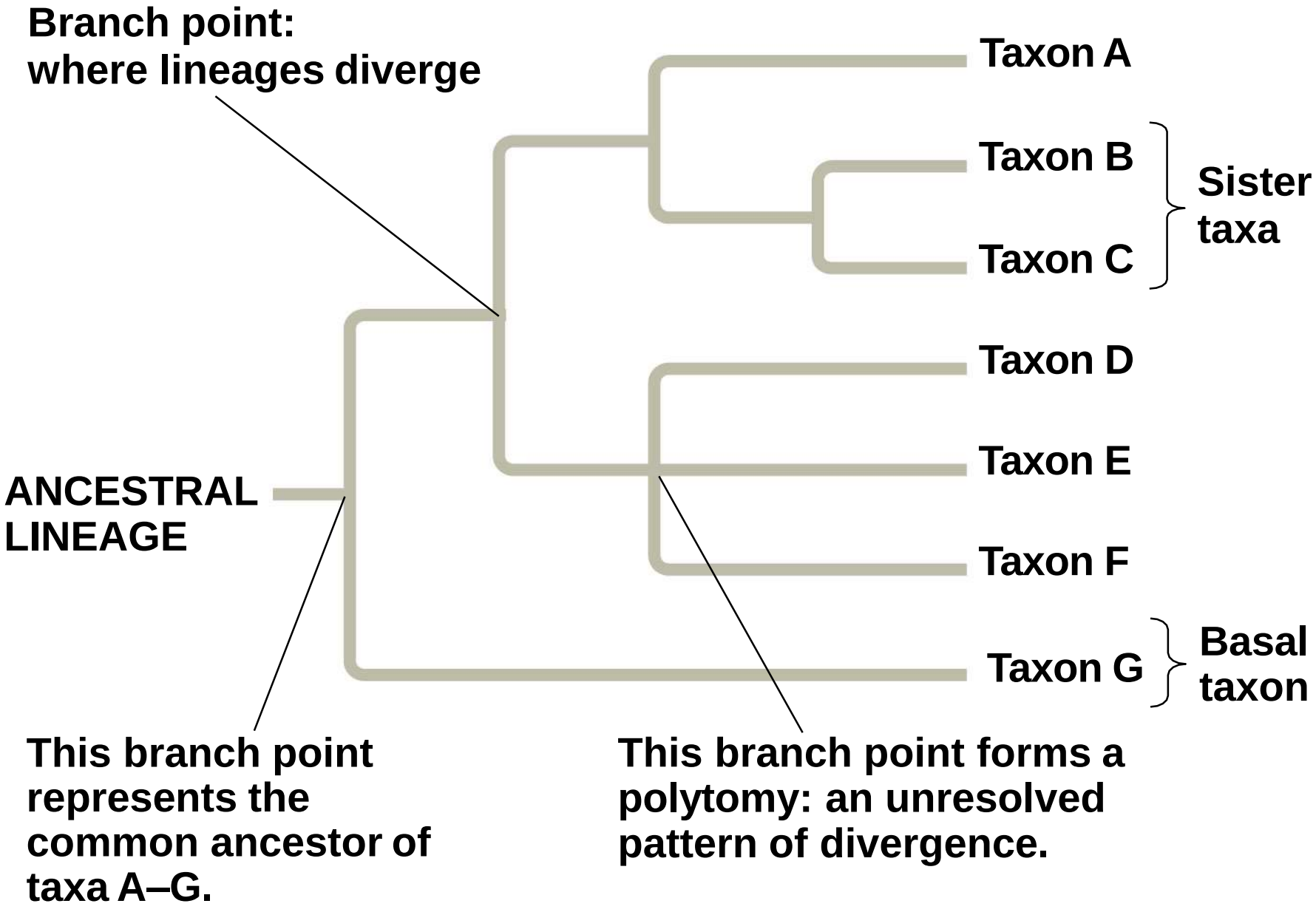
Genome Evolution

- Orthologous genes are widespread and extend across many widely varied species
  - For example, humans and mice diverged about 65 million years ago, and 99% of our genes are orthologous
- Gene number and the complexity of an organism are not strongly linked
  - For example, humans have only four times as many genes as yeast, a single-celled eukaryote
- Genes in complex organisms appear to be very versatile, and each gene can perform many functions

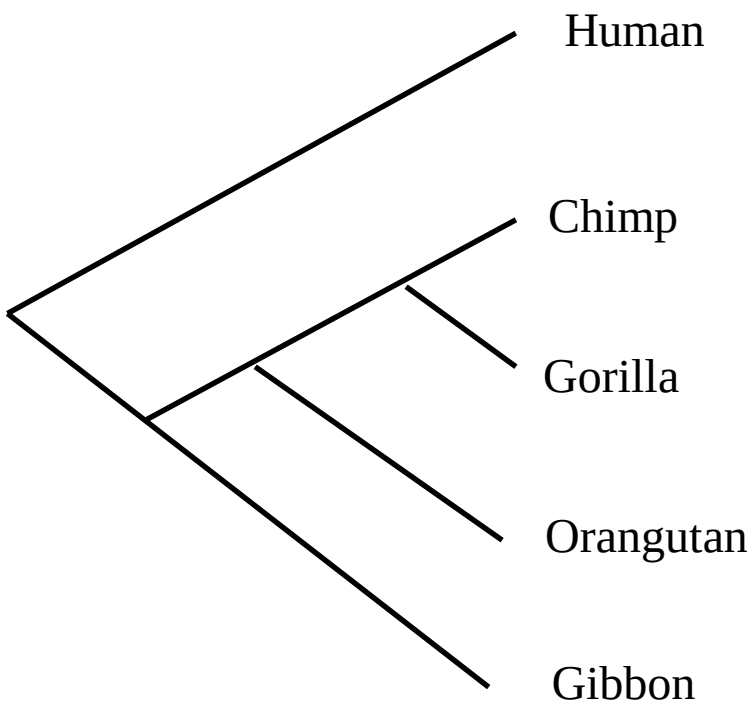
Phylogenetic trees

- **Phylogeny** is the evolutionary history of a species or group of related species
- **Taxonomy** is the ordered division and naming of organisms
- A taxonomic unit at any level of hierarchy is called a **taxon**
- Systematists depict evolutionary relationships in branching **phylogenetic trees**
- A **phylogenetic tree** represents a hypothesis about evolutionary relationships

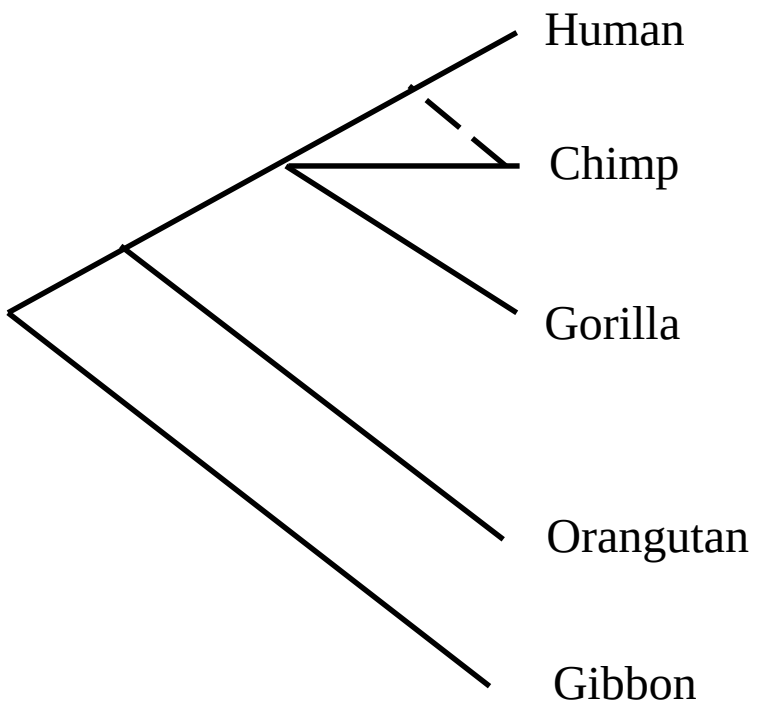
- Each **branch point** represents the divergence of two species
- **Sister taxa** are groups that share an immediate common ancestor
- A **rooted** tree includes a branch to represent the last common ancestor of all taxa in the tree
- A **basal taxon** diverges early in the history of a group and originates near the common ancestor of the group
- A **polytomy** is a branch from which more than two groups emerge



Phylogenetic analysis using DNA sequence



Traditional



Molecular

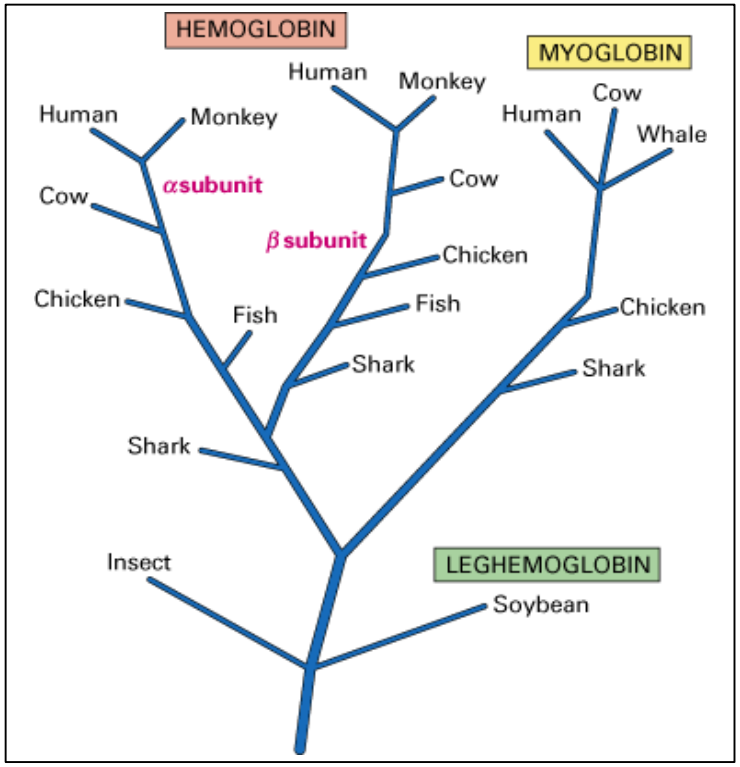
Two Areas in Phylogenetic analysis

- Phylogenetic inference or "**tree building**":
  - To infer the branching orders and lengths between "taxa" (or genes, populations, species etc).
  - For example, can DNA tell us giant panda more similar to bear or to dog, and when did they diverge ?

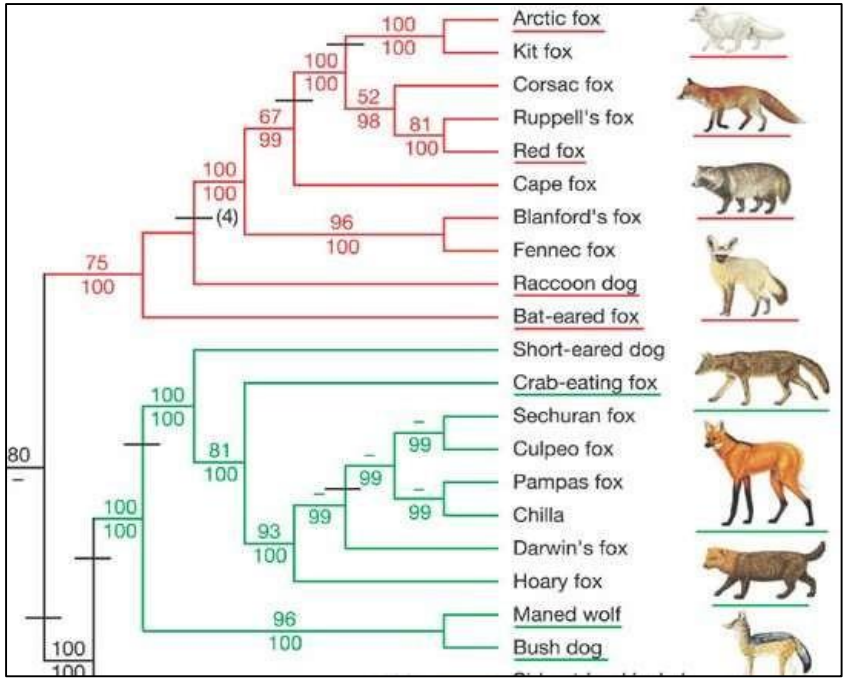
- **Character and rate** analysis:
  - Using phylogeny as a framework to understand the evolution of traits or genes.
  - For example, is gene X under positive or purifying selection ?

Phylogenetic Tree

Gene Tree



Species Tree



Lindblad-Toh Nature 2005

http://www.muhlenberg.edu/depts/biology/courses/bio152

Methods of Tree reconstruction

- **Maximum Parsimony methods**
- **Distance based methods**
- **Maximum Likelihood methods**
- **Bayesian methods**

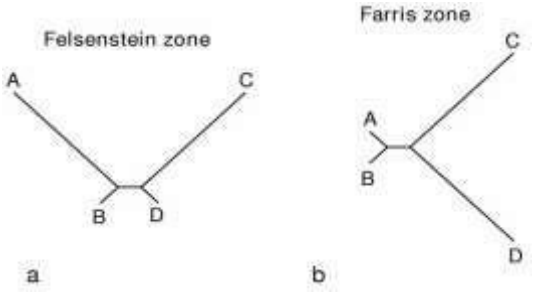
(Don’t worry, there are software programs that are easy and fun to use)

Parsimony Methods

- Optimality criterion: The "most-parsimonious" tree is the one that requires the fewest number of evolutionary events (e.g. nucleotide substitutions, amino acid replacements) to explain the observed sequences.

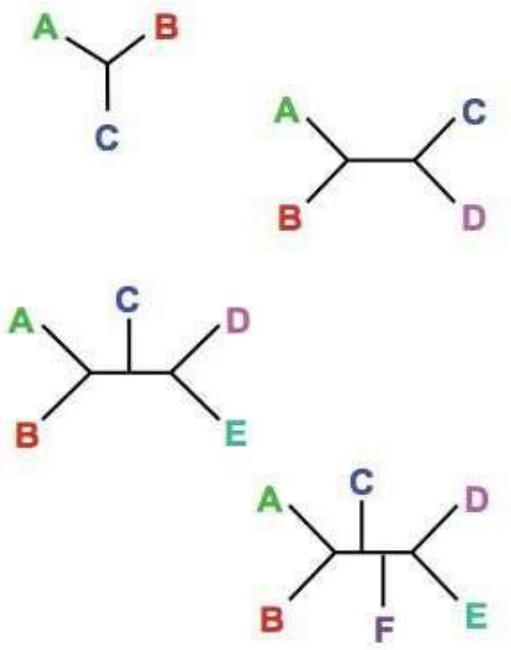
- Advantages:
  - Intuitive, logical and simple (can be done with pencil-and paper)
  - Can be used on molecular and other (morphological, language) data.
  - Can be used to infer the sequences of extinct (hypothetical) ancestors

- Disadvantages
  - Can be fooled by high levels of homoplasy ("same events")
  - Can be problematic when the real tree is mixed with very short and long branches, e.g. long-branch attraction



Number of Possible Trees Increases With the Number of Taxa

Exact searches become increasingly difficult, and eventually impossible, as the number of taxa increases:



# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	3.58×10^{36}

Number of unrooted trees for *n* taxa

$N_u=(2n-5)*(2n-7)*...*3*1=(2n-5)!/[2^{n-3}*(n-3)!]$

Distance based methods

- Estimate the number of substitutions between each pair of sequences in a group of sequences.
- Try to build a tree so that the branch lengths represent the pair-distances.
- What are these "distances" ? Example: sequence identity between two protein and DNA sequences

Distance based methods

What distance to use ?

?

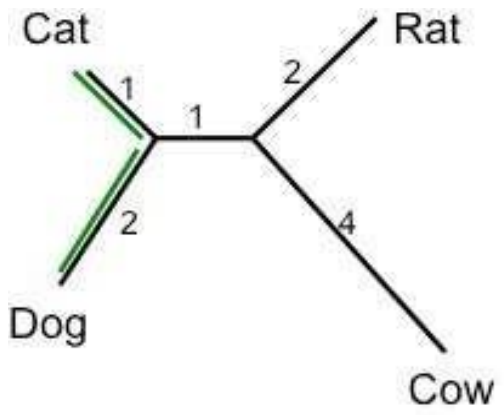
Cat	ATTTGCGGTA
Dog	ATCTGCGATA
Rat	ATTGCCGTTT
Cow	TTCGCTGTTT

Number of different nucleotides

	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6

•The observed differences do not always represent the actual evolutionary events that occurred, e.g. multiple substitutions at the same site.

•Substitution rates are different between different types of nucleotides



Molecular evolution

Applications:

Molecular evolution analysis has clarified:

- **the evolutionary relationships between humans and other primates;**
- **the origins of AIDS;**
- **the origin of modern humans and population migration;**
- **speciation events;**
- **genetic material exchange between species.**
- **origin of some deseases (cancer, etc...)**