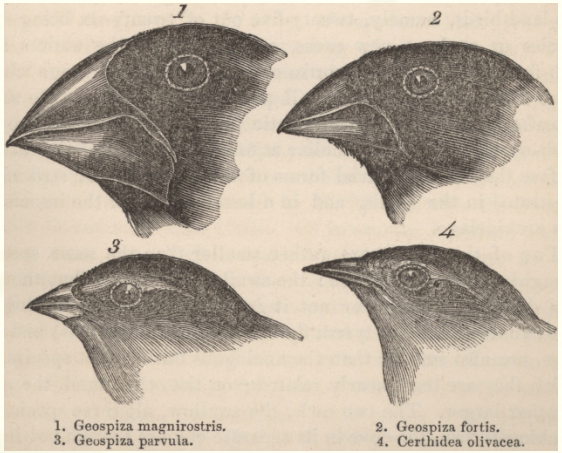
Lecture 2: Fundamentals in Molecular Evolution



Lecture 2

Outline of lecture

- Introduction and historical background
- Mutations and substitutions
  - Positive, negative, neutral selection, synonymous and nonsynonymous substitutions
- Codon bias
- Neutral theory of evolution
- Phylogenetic trees

What is Molecular Evolution ?

- Molecular evolution address two broad range of questions:
  1. Use **DNA** to study the evolution of **organisms**, e.g. population structure, geographic variation and phylogeny
  2. Use different **organisms** to study the evolution process of **DNA**

What is Molecular Evolution ?

- How and when were genes and proteins created ? How “old” is a gene ? How can we calculate the “age” of a gene ?

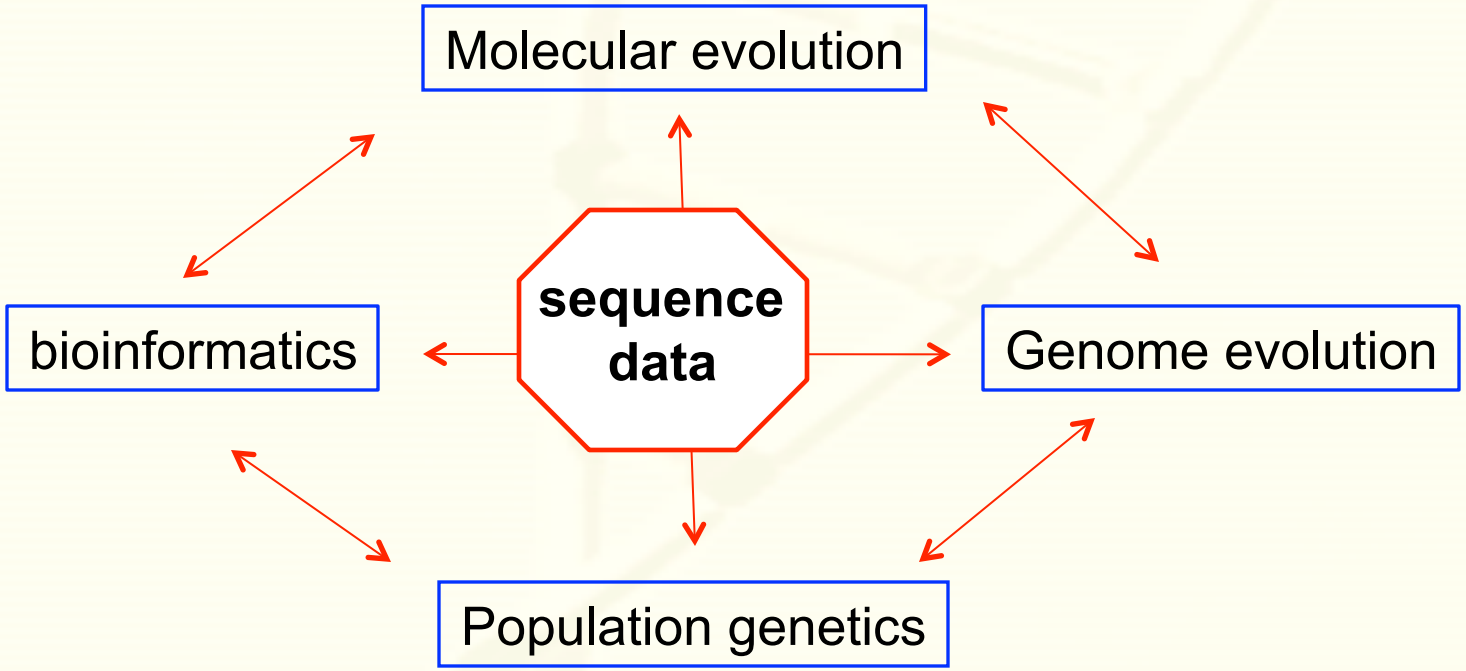
- How did the gene evolve to the present form ? What selective forces (if any) influence the evolution of a gene sequence and expression ? Are these changes in sequence adaptive or neutral ?

- How variable is a gene’s sequence or expression level among individuals within a species and between species (or individuals), and what does such information tell us about the functional role of this gene ?

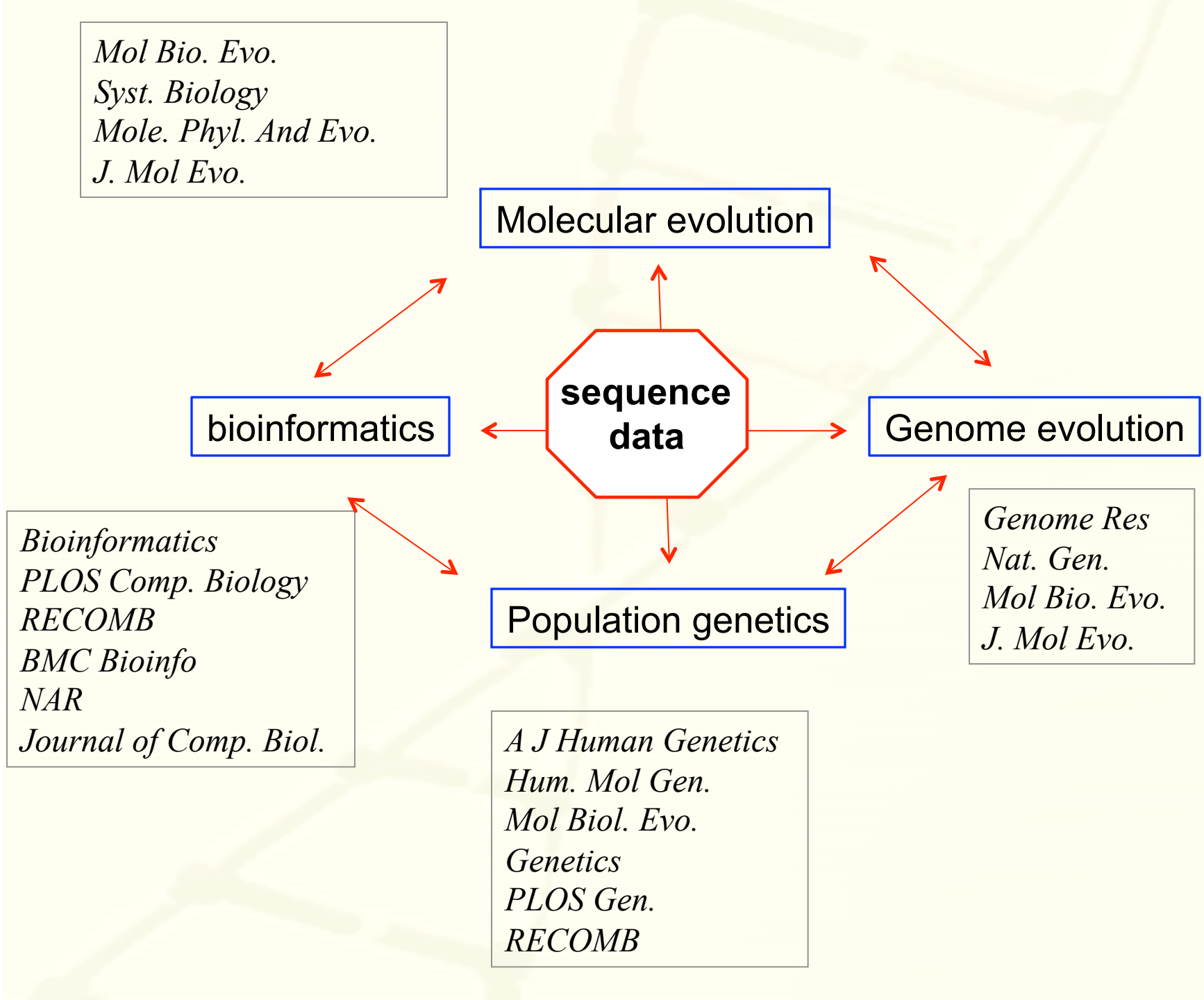
- How do species evolve? How can evolution of a gene tell us about the evolutionary relationship of species ?

The Genomic Revolution

Genomic sequencing, high-throughput biology, and computational biology / bioinformatics have provided new data to analyze, and posed new questions to address.

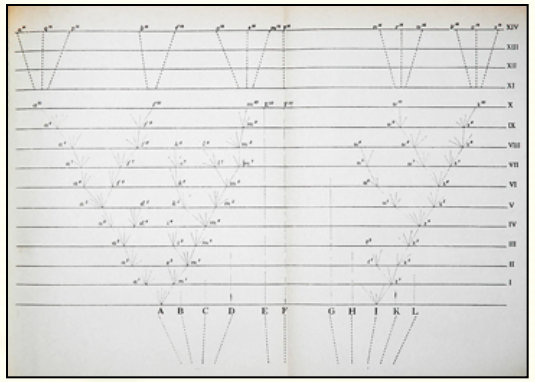


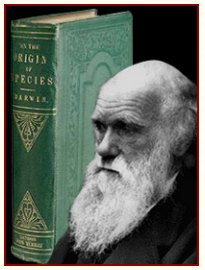
These are overlapping disciplines but they do have their own conferences and journals

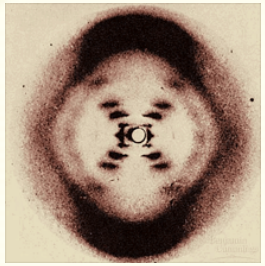

Mol Bio. Evo.
Syst. Biology
Mole. Phyl. And Evo.
J. Mol Evo.
Molecular evolution
bioinformatics
sequence data
Genome evolution
Bioinformatics
PLOS Comp. Biology
RECOMB
BMC Bioinfo
NAR
Journal of Comp. Biol.
Population genetics
Genome Res
Nat. Gen.
Mol Bio. Evo.
J. Mol Evo.
A J Human Genetics
Hum. Mol Gen.
Mol Biol. Evo.
Genetics
PLOS Gen.
RECOMB

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)



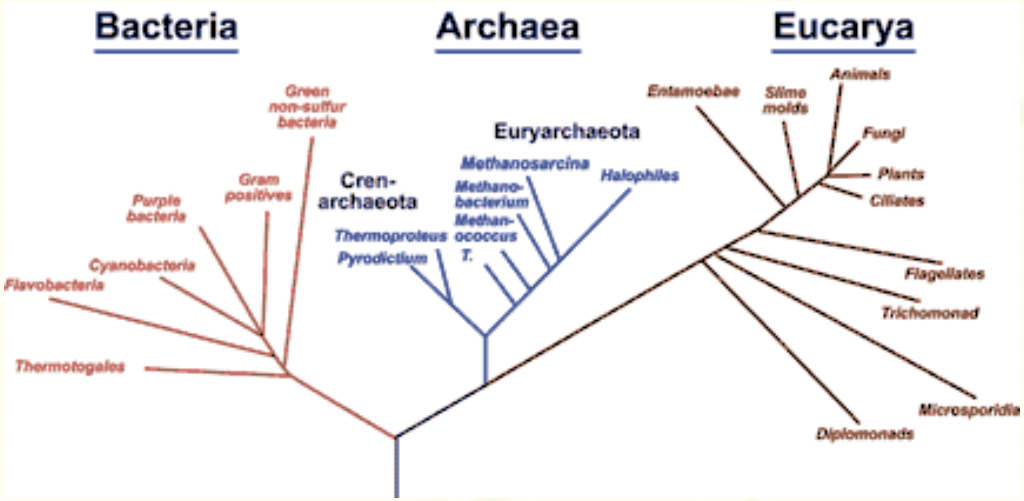


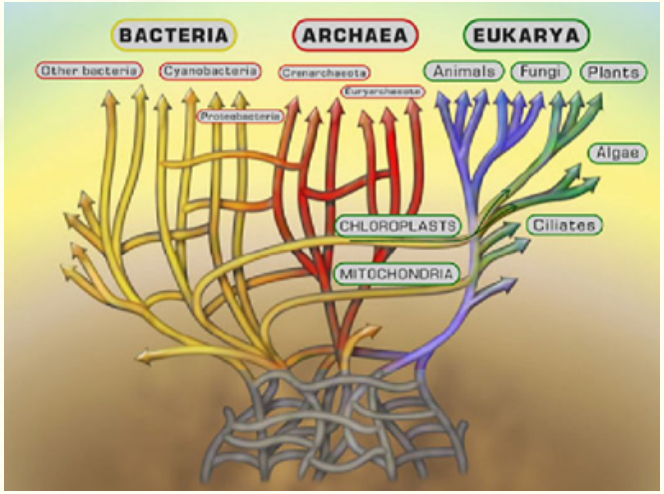


A brief historical perspective

- Until 1970s, cellular organisms were divided into eukaryotes (have nucleus) and prokaryotes (no nucleus)

- Using 16S rRNA gene sequence, Carl Woese redefined three domains



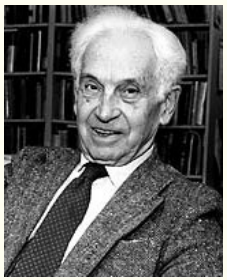


Ford Doolittle

- 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 …

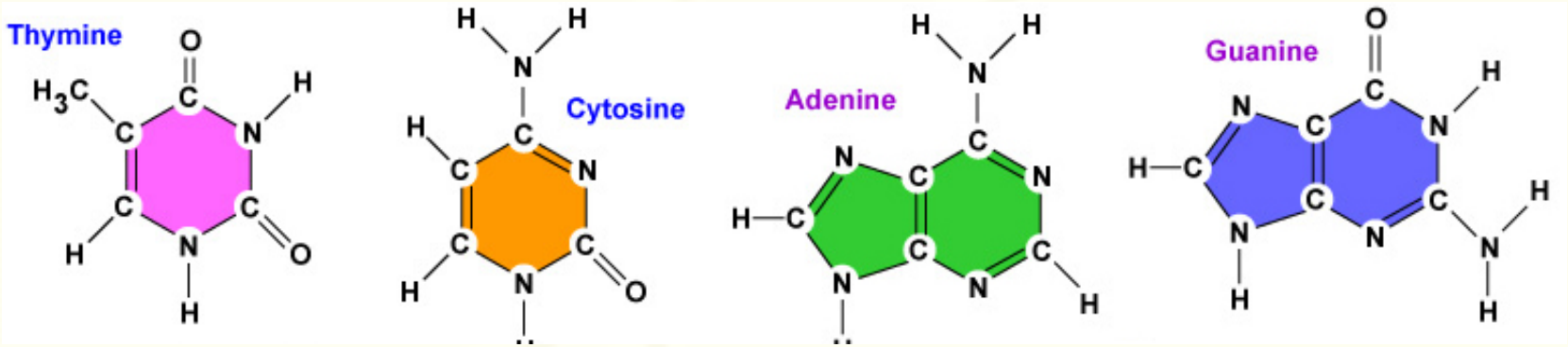
- "Nothing in Evolution Makes Sense Except in the Light of Biology"
- "Nothing in Evolution Makes Sense Except in the Light of Domestication"
- "Nothing in Evolution Makes Sense Except in the Light of Population Genetics (in relation to population size)" – Michael Lynch

Mutations in DNA and protein

- Mechanism of molecular evolution: mutation, insertion, and deletion

GACGACCATAGACCAGCATAG

GACTACCATAGA-CTGCAAAG



- **Transition**: A<->G, C<->T
- **Transversion:** purine <-> pyrimidine

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

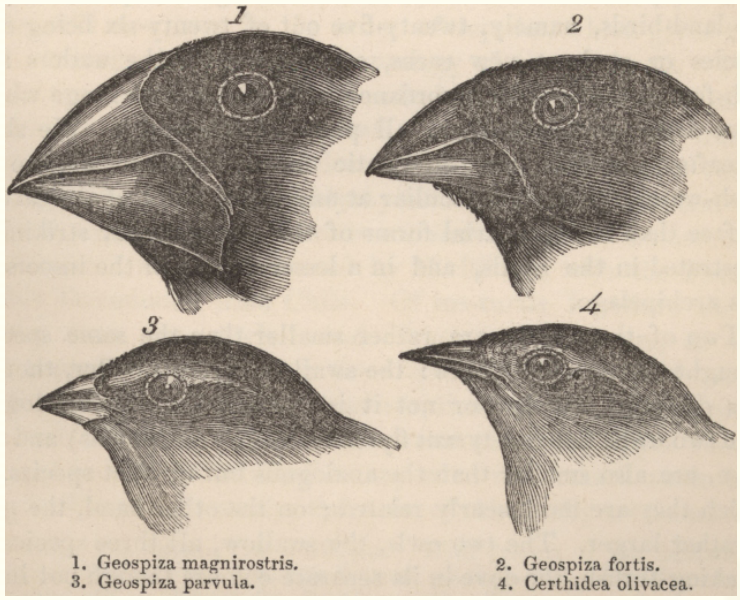
Degeneracy of genetic code

UUU, UUC — phenylalanine UUA, UUG — leucine	UCU, UCC, UCA, UCG — serine	UAU, UAC — tyrosine UAA, UAG — stop	UGU, UGC — cysteine UGA — stop UGG — tryptophan
CUU, CUC, CUA, CUG — leucine	CCU, CCC, CCA, CCG — proline	CAU, CAC — histidine CAA, CAG — glutamine	CGU, CGC, CGA, CGG — arginine
AUU, AUC, AUA — isoleucine AUG — methionine	ACU, ACC, ACA, ACG — threonine	AAU, AAC — asparagine AAA, AAG — lysine	AGU, AGC — serine AGA, AGG — arginine
GUU, GUC, GUA, GUG — valine	GCU, GCC, GCA, GCG — alanine	GAU, GAC — aspartic acid GAA, GAG — glutamic acid	GGU, GGC, GGA, GGG — glycine

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]

"We 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

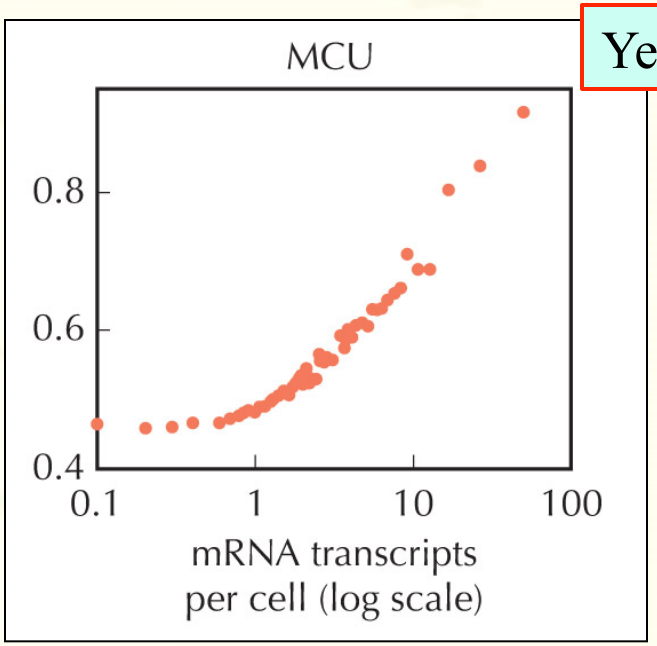
Ka / Ks
= Non-synonymous/Synonymous substitutions
=5

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)

Highly expressed genes tend to use optimal codons



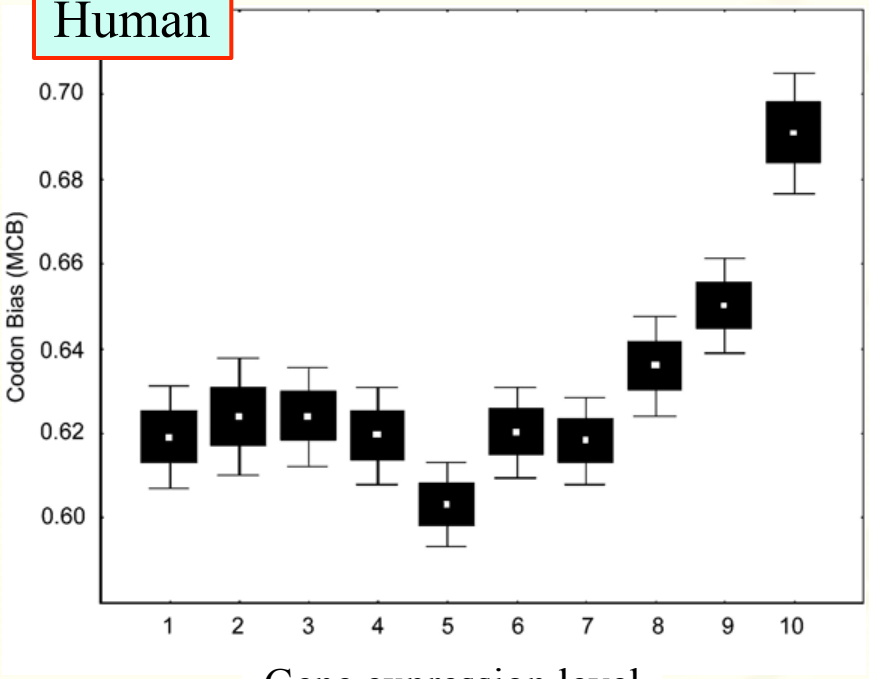
Yeast

Gene expression and molecular evolution
Hiroshi Akashi

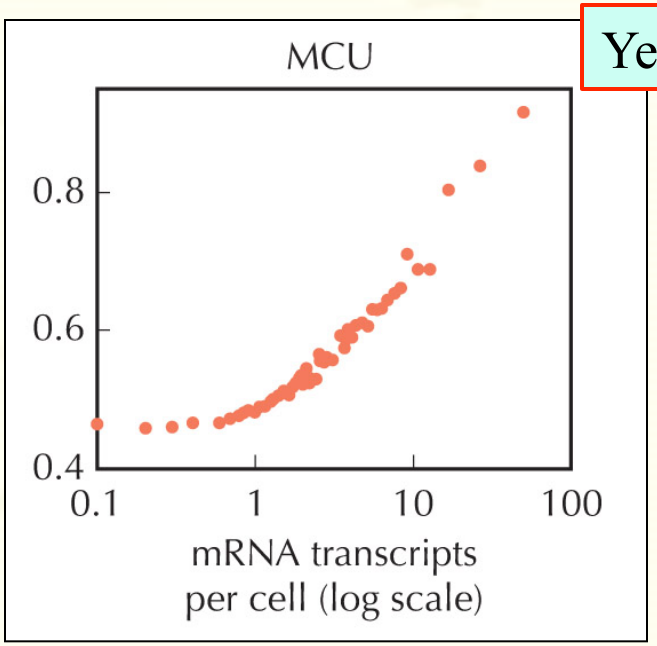
CAI (Codon Adaptation Index) measures how optimal a gene's codons are, relative to the tRNA pool in the cell.

Highly expressed genes tend to use optimal codons

Human



Yeast



Gene expression and molecular evolution
Hiroshi Akashi

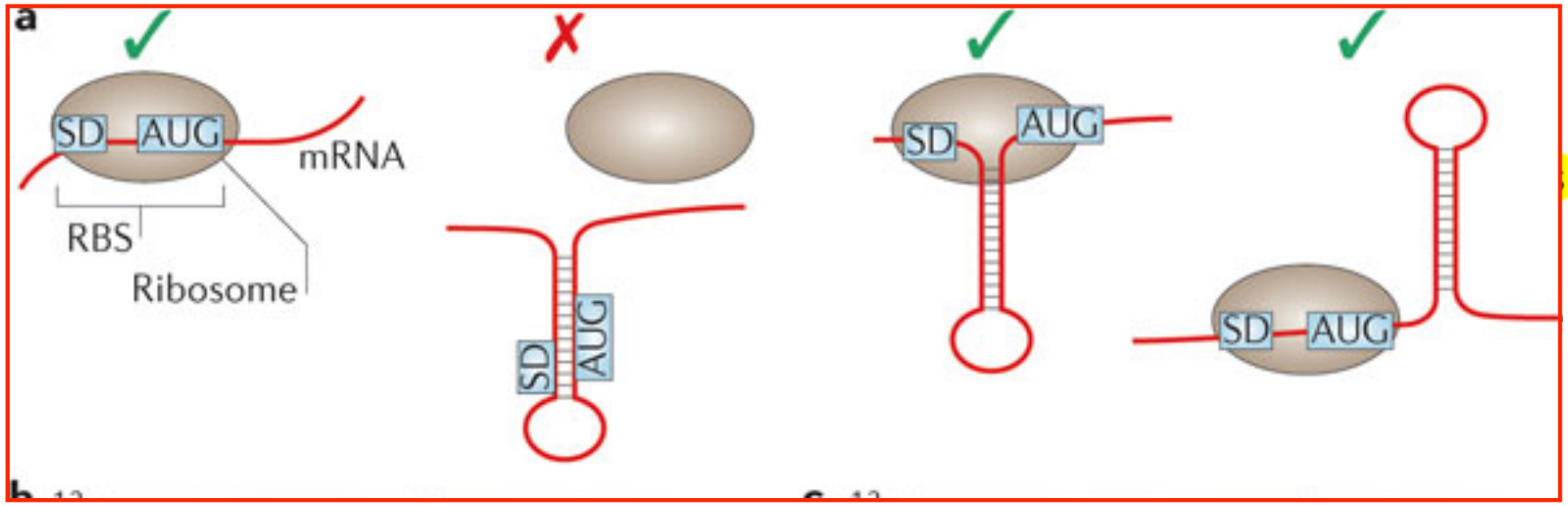
The Signature of Selection Mediated by Expression on Human Genes

Araxi O. Urrutia and Laurence D. Hurst[1]

CAI (Codon Adaptation Index) measures how optimal a gene's codons are, relative to the tRNA pool in the cell.

Synonymous codons influence mRNA secondary structure and gene expression

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



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

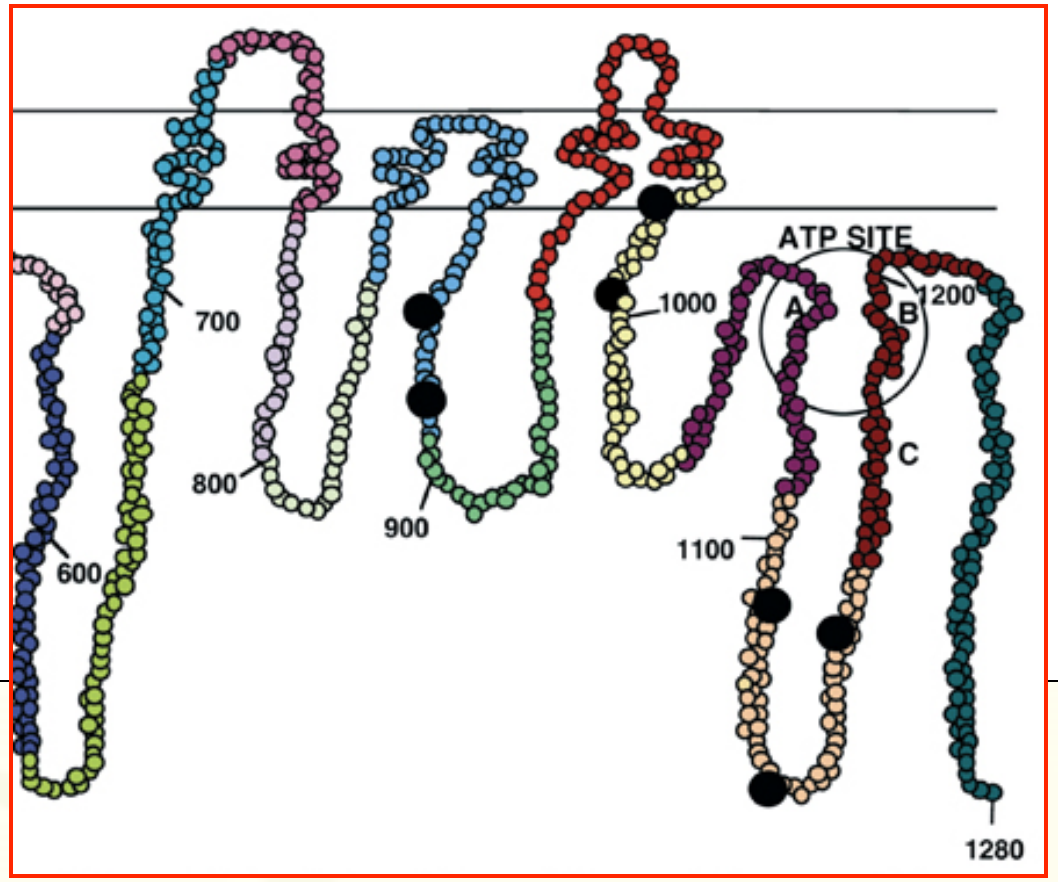
A "Silent" Polymorphism in the *MDR*1 Gene Changes Substrate Specificity

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.

“Rare codons” can influence protein structure

A “Silent” Polymorphism in the *MDR1* Gene Changes Substrate Specificity



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"

Evolutionary Rate at the Molecular Level

by
MOTOO KIMURA

National Institute of Genetics,
Mishima, Japan

Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations involved must be neutral ones.

Darwinism is so well established that it is difficult to think of evolution except in terms of selection for desirable characteristics and advantageous genes. New technical developments and new knowledge, such as the sequential analysis of proteins and the deciphering of the genetic code, have made a much closer examination of evolutionary processes possible, and therefore necessary. Patterns of evolutionary change that have been observed at the phenotypic level do not necessarily apply at the genotypic and molecular levels. We need new rules in order to understand the patterns and dynamics of molecular evolution.

Kimura Science 1968

Non-Darwinian Evolution

Most evolutionary change in proteins may be due to neutral mutations and genetic drift.

Jack Lester King and Thomas H. Jukes

King & Jukes Nature 1969

“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.

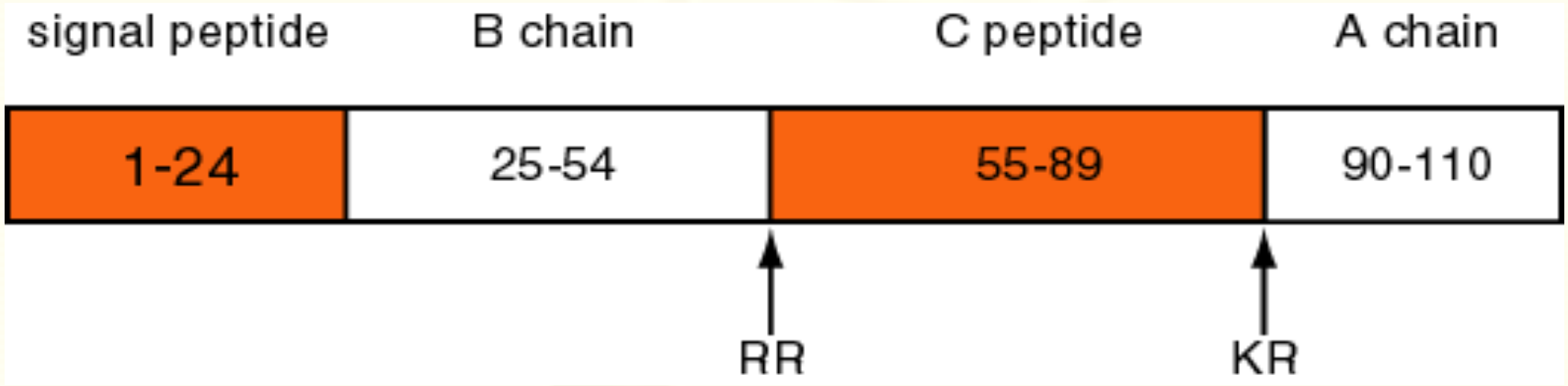
Genes evolve at different rates

Rates of nucleotide substitution (per site per billion years)

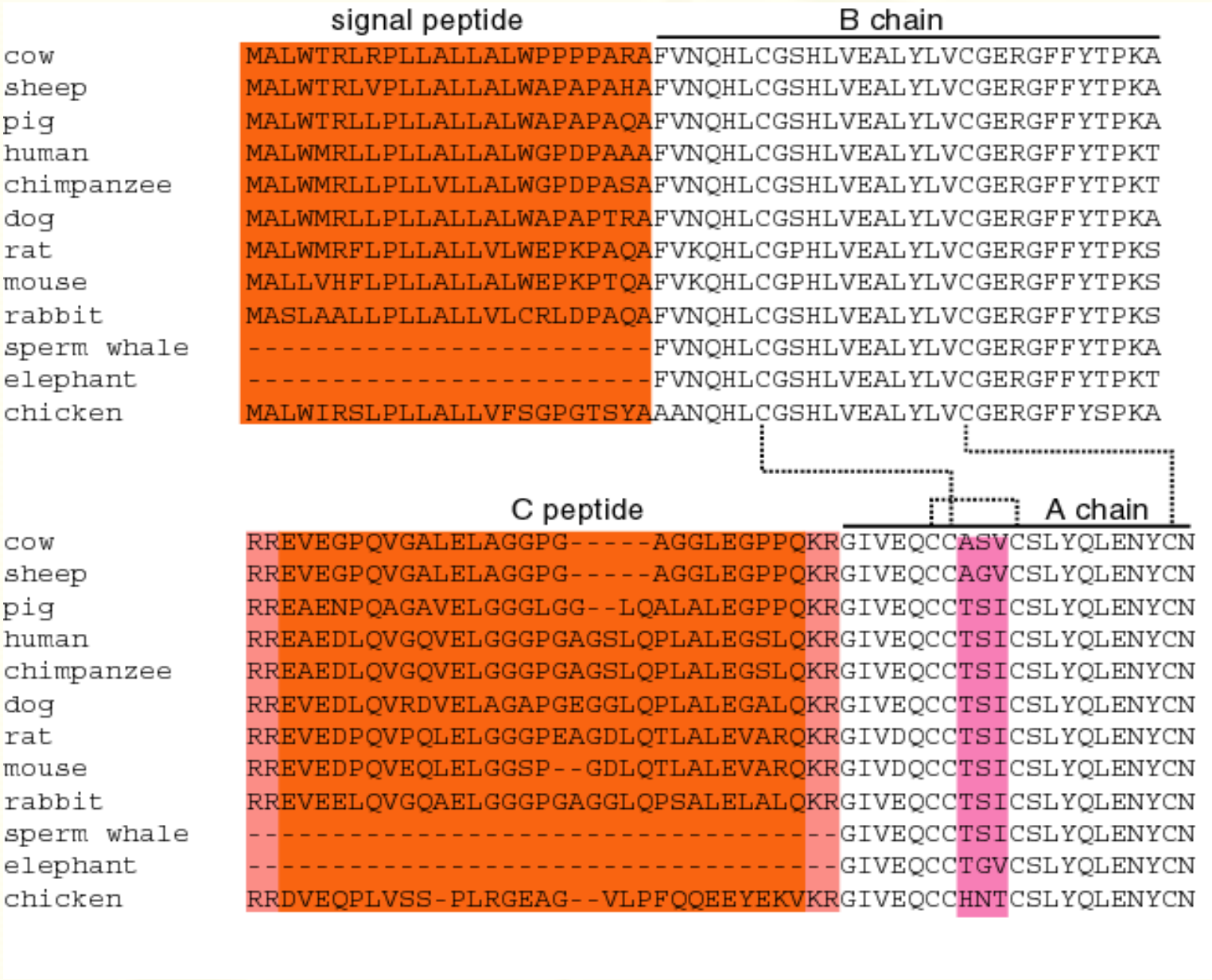
Gene	Non-synonymous rate	Synonymous rate
Histone H4	0.00	3.94
Histone H2	0.00	4.52
Actin a	0.01	3.68
Ribosomal protein S14	0.02	2.16
Insulin	0.13	4.02
a-globin	0.78	2.58
Myoglobin	0.57	4.10
β-Interferon	3.06	5.50
MHC (HLA-A)	13.30	3.5

Different domains of a protein evolve at different rate: insulin as an example.

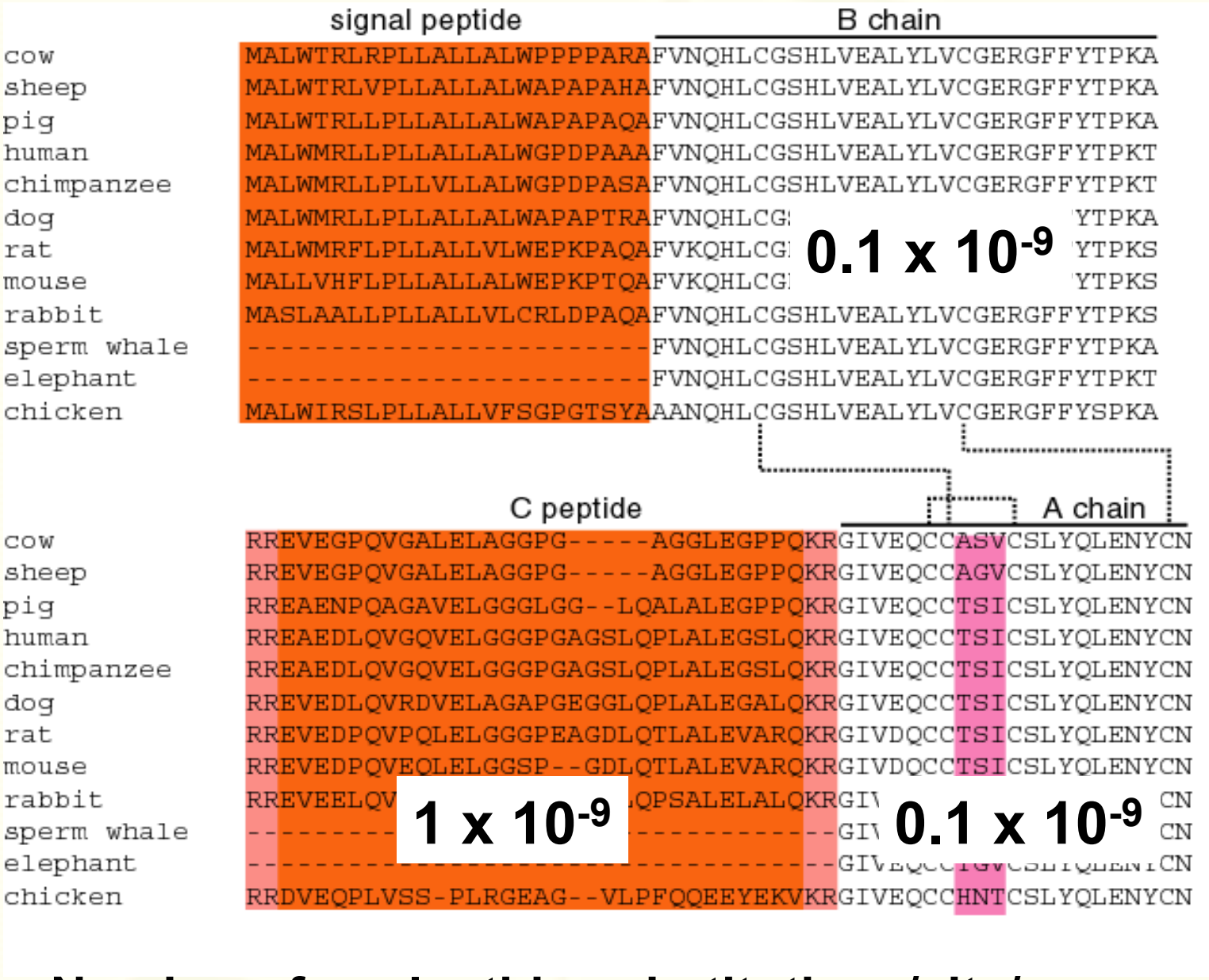
Mature insulin consists of an A chain and B chain heterodimer connected by disulphide bridge



The signal peptide and C peptide are cleaved,
and their sequences display fewer functional constraints.



http://www.bioinfbook.org



Number of nucleotide substitutions/site/year

http://www.bioinfbook.org

Guinea pig insulin have undergone an extremely rapid rate of evolutionary change

Arrows indicate positions at which guinea pig insulin (A chain and B chain) differs from both human and mouse

http://www.bioinfbook.org

Molecular clock

- Different proteins have different rates
- Different domains of the same protein may have different rate
- Same protein in different organisms may have different rates
- Are the substitution rates constant at the different geological time period, e.g. different oxygen content in the atmosphere, different radiation level? before or after mass extinction ? The role of chaperones on protein sequence evolution ?

More on neutral theory

- Probably correct for some fraction of the genome
- What fraction of the proteins evolves neutrally and how much is under selection?
- What fraction of the genome evolves neutrally and how much is under selection?
- What about gene expression and regulation ? How much of the difference in expression level between species is due to nature selection or genetic drift ?

Methods to detect positive selection

- **Ka / Ks test:** suitable for between species
- **McDonald-Kreitman (MK) test**
  - Compare between species and within species
- **Fixation index** (Fst)
  - Testing difference in allele frequency between population
- **Linkage disequilibrium (LD)**
  - Look for nonrandom association of alleles at linked loci

All these methods take neutrality as null hypothesis

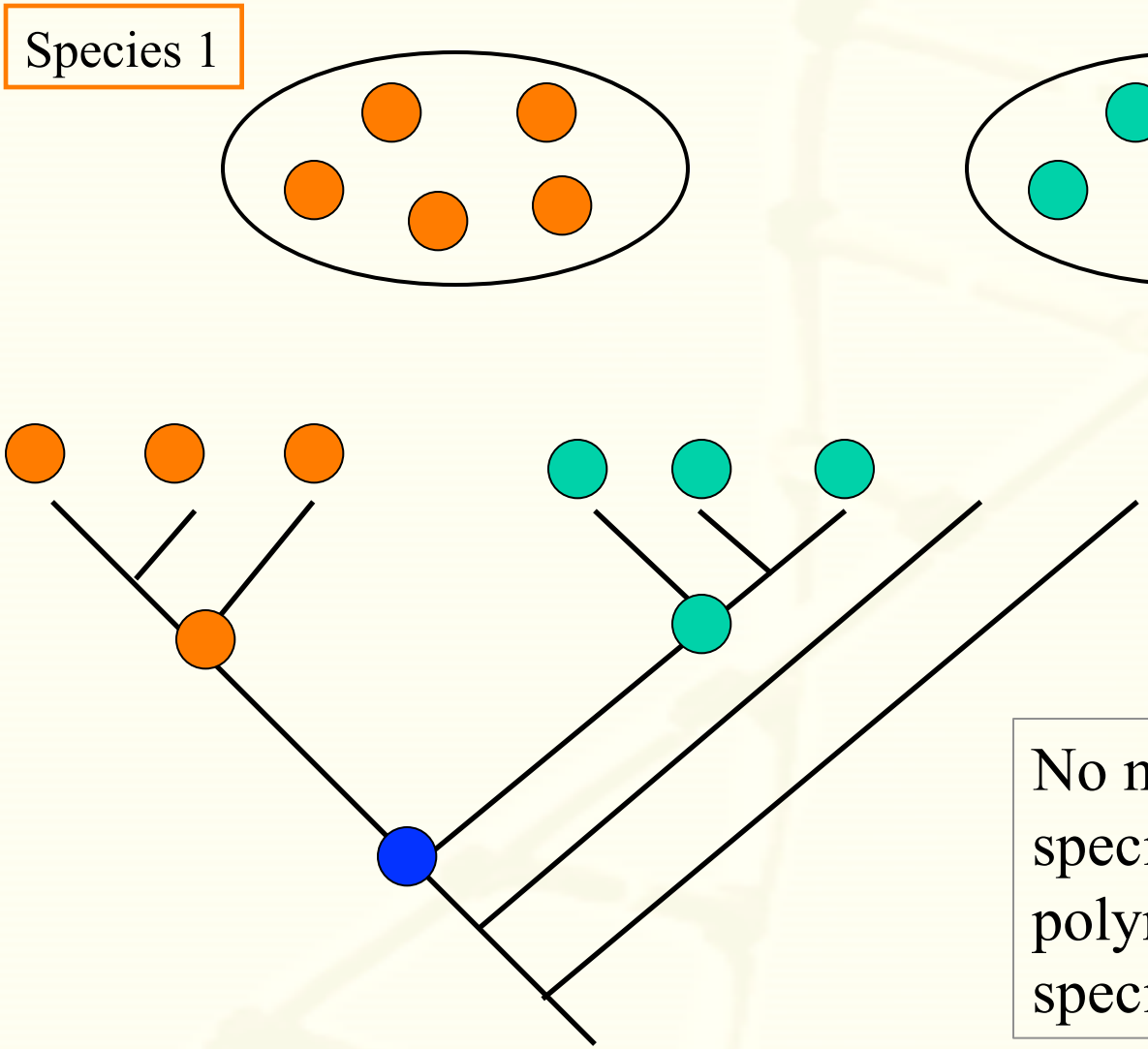
McDonald-Kreitman (MK) test

- McDonald-Kreitman (MK) Test compares divergence between two species with polymorphism within each species.

- If a gene evolves neutrally, i.e. the DNA substitutions follow random drift, then the polymorphism within each species should follow the same pattern as divergence between species.

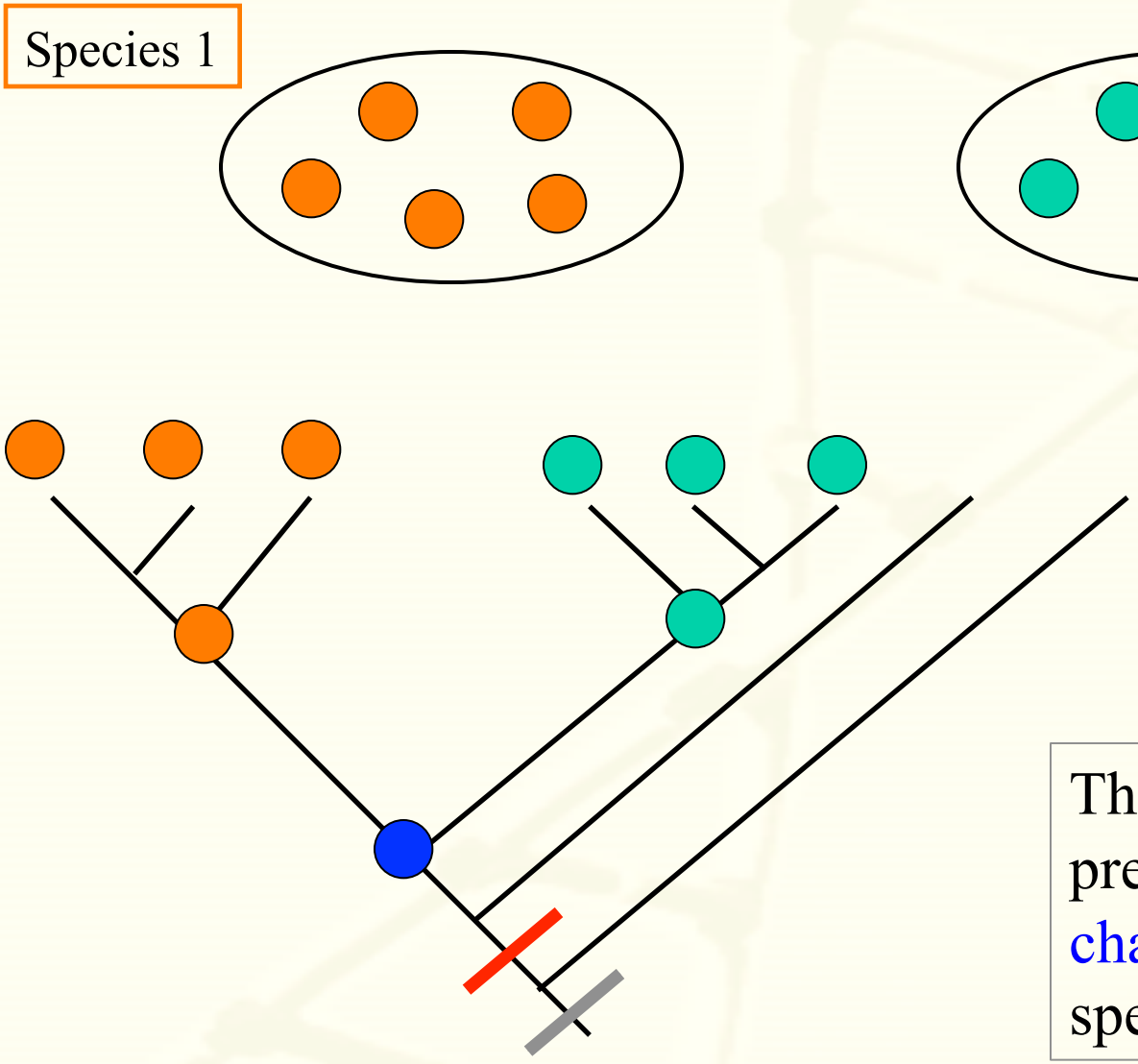
- This predicts similar ratio of synonymous and non-synonymous substitutions between and within species.

McDonald-Kreitman (MK) test



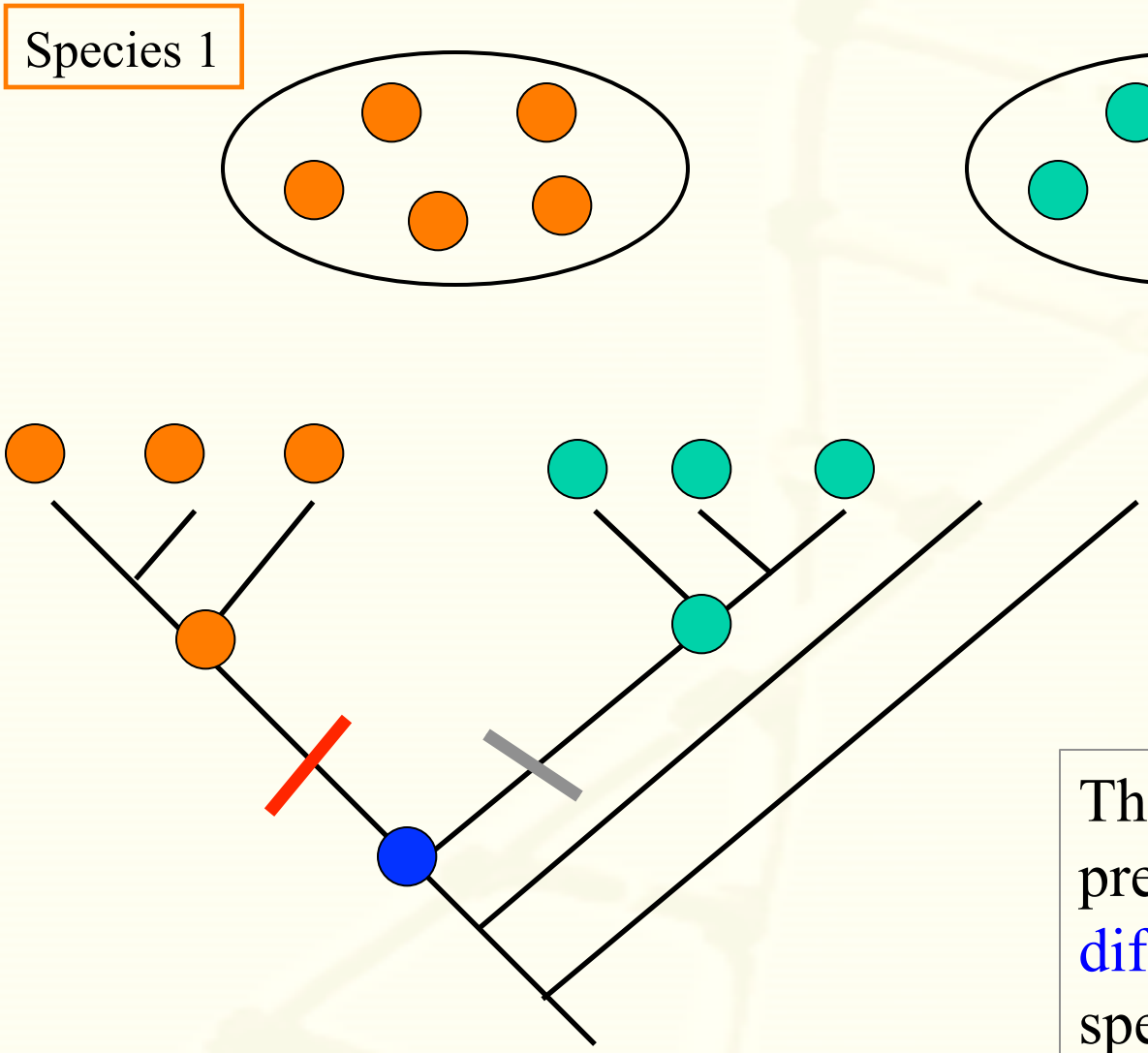
No mutations between species and polymorphisms within species

McDonald-Kreitman (MK) test



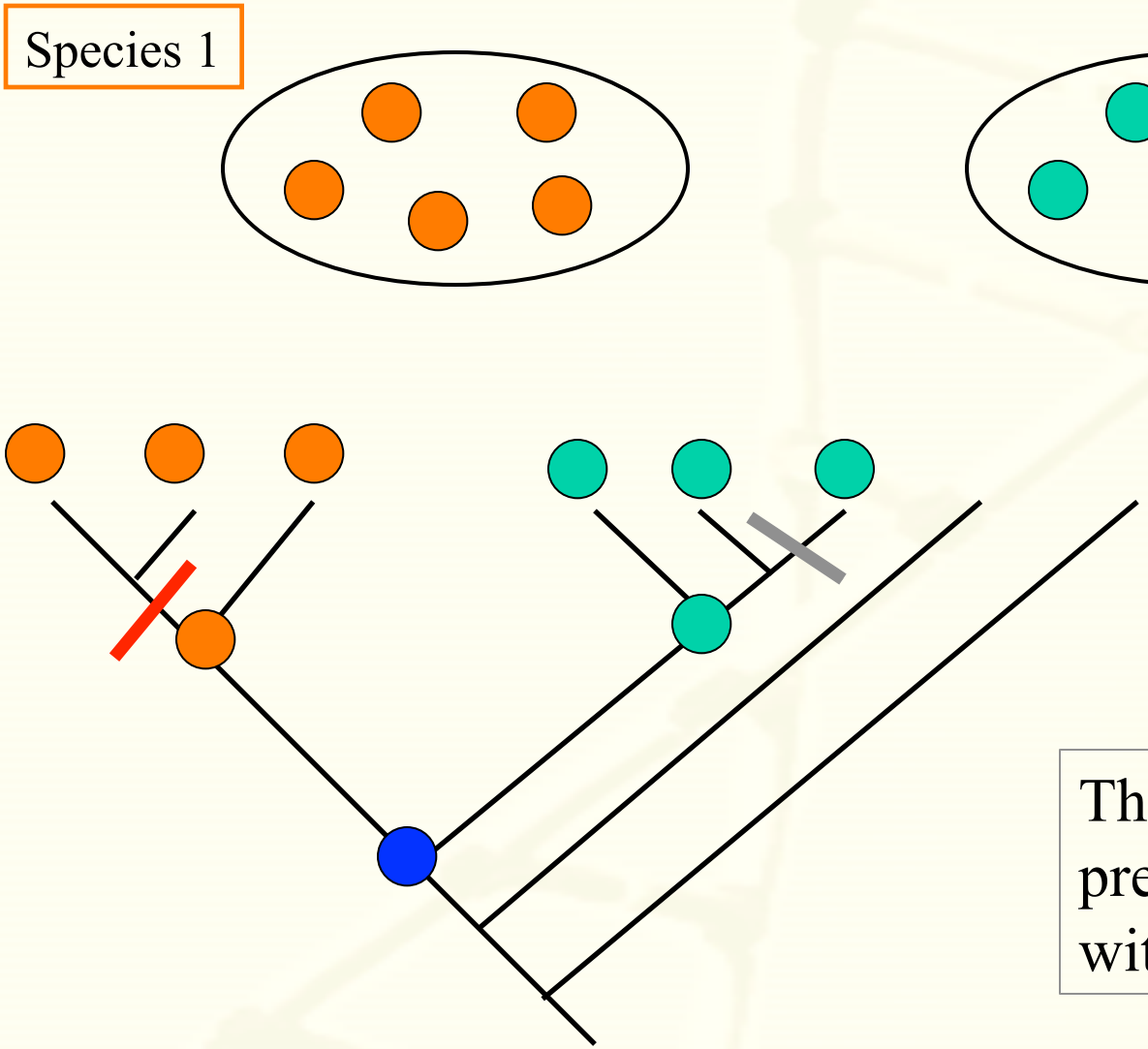
These mutations are present as shared characters in both species

McDonald-Kreitman (MK) test



These mutations are present as fixed differences between species

McDonald-Kreitman (MK) test



These mutations are present as polymorphism within each species

McDonald-Kreitman (MK) test

A C G A T T C A C G G

T C G A G T C A C C G

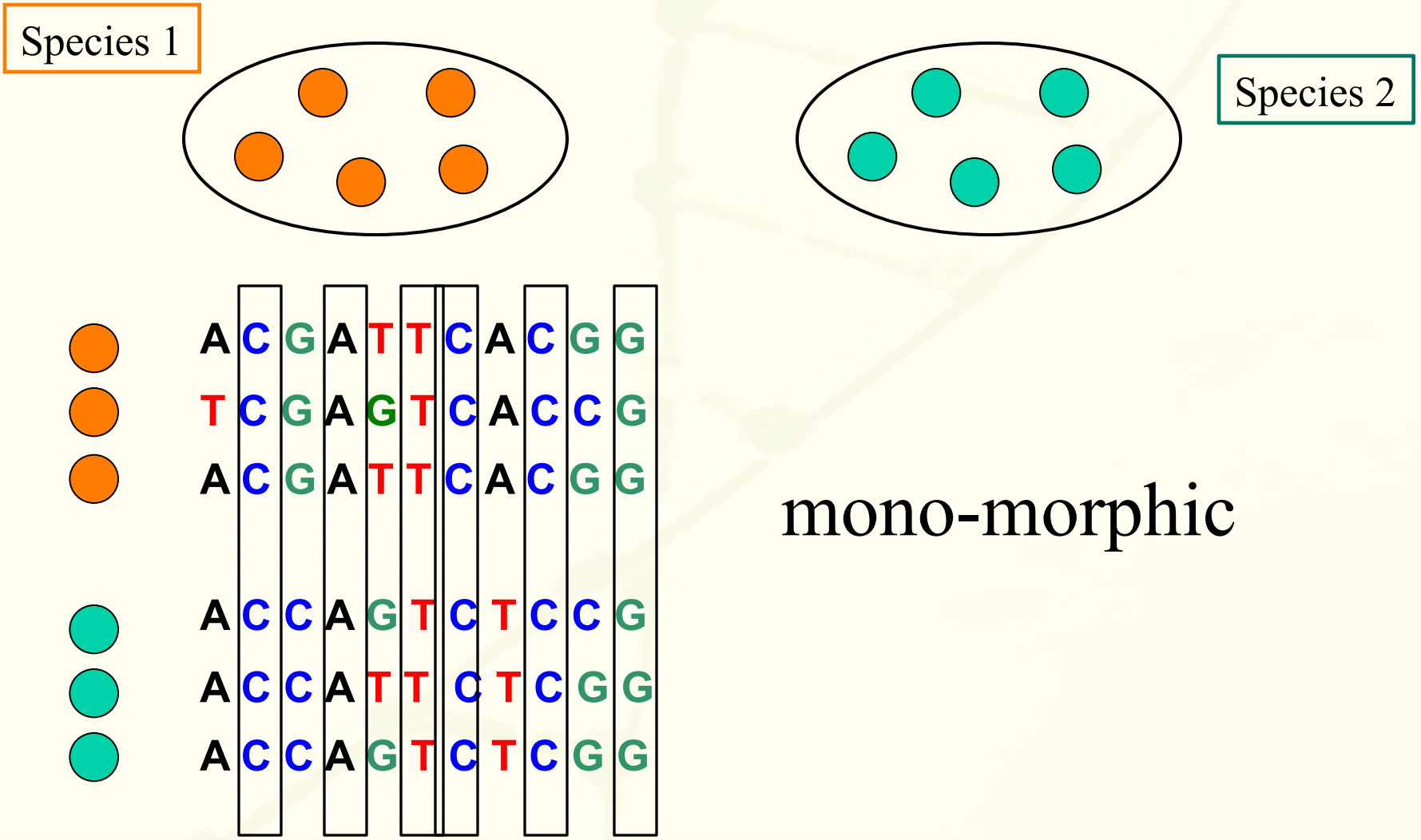
A C G A T T C A C G G

A C C A G T C T C C G

A C C A T T C T C G G

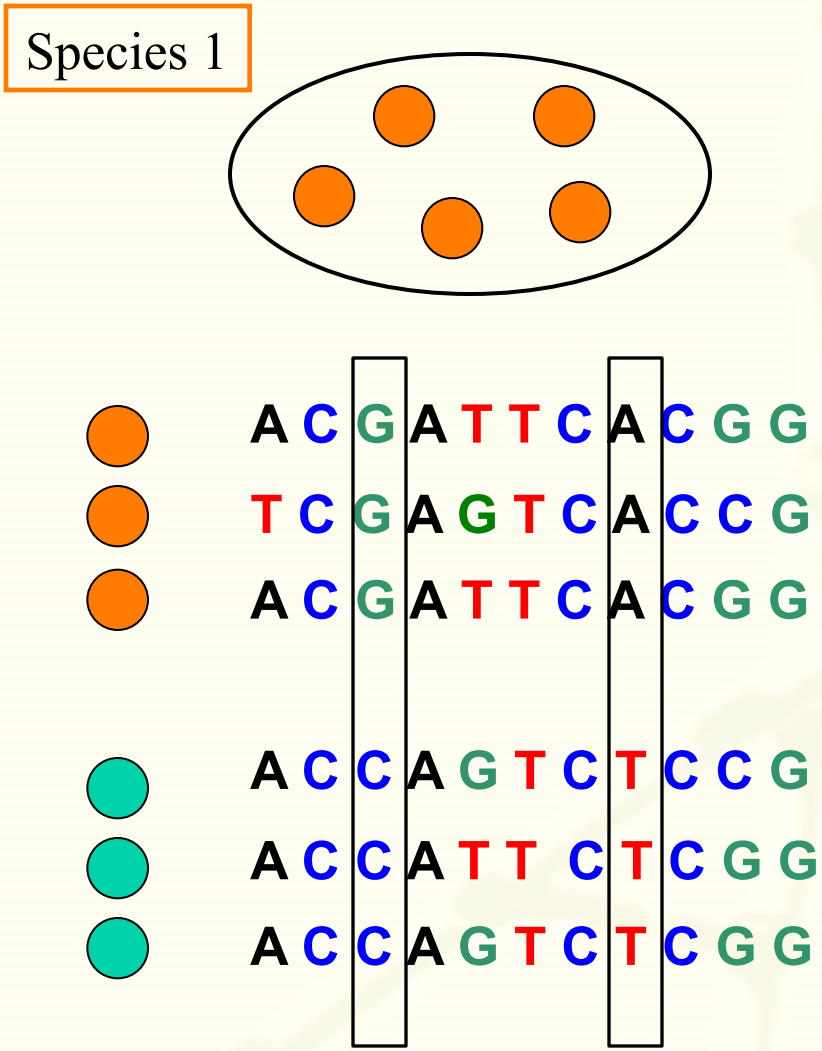
A C C A G T C T C G G

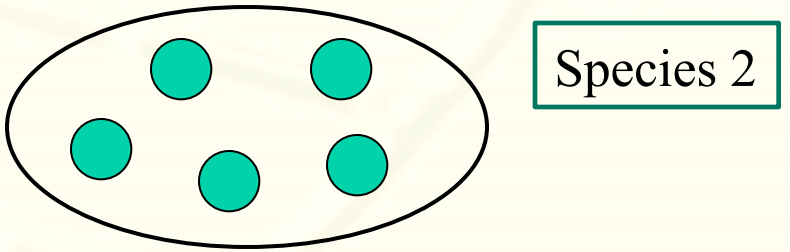
McDonald-Kreitman (MK) test



mono-morphic

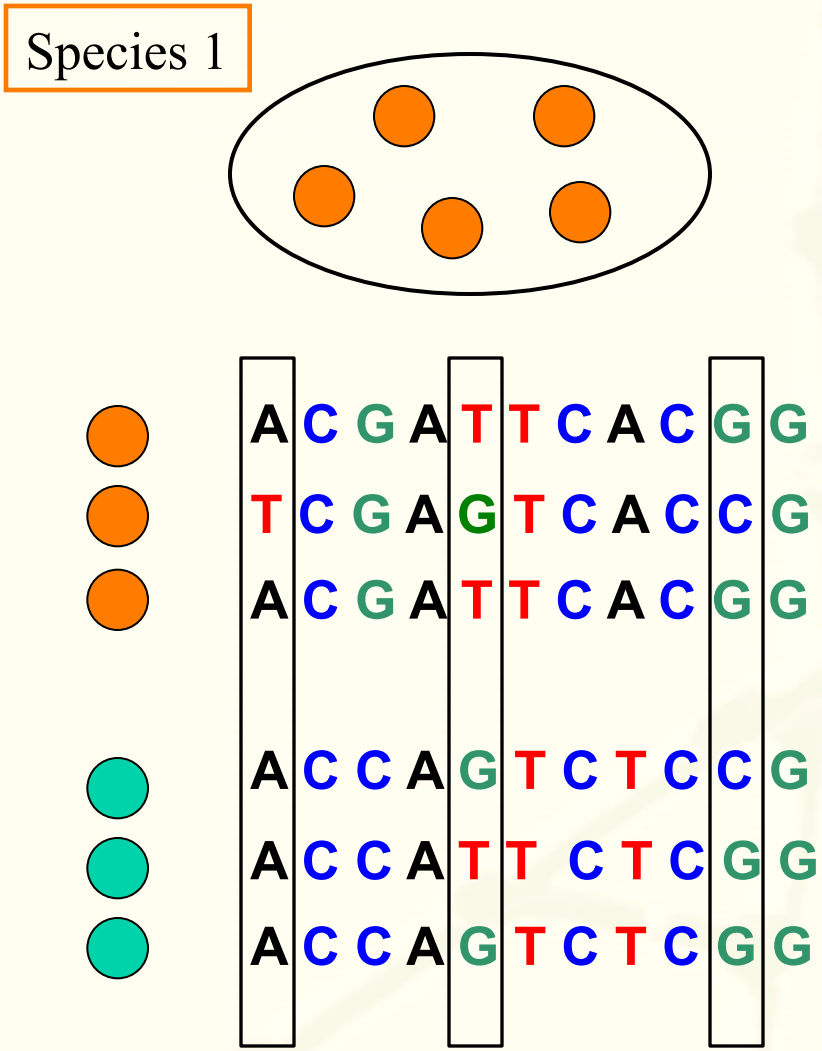
McDonald-Kreitman (MK) test

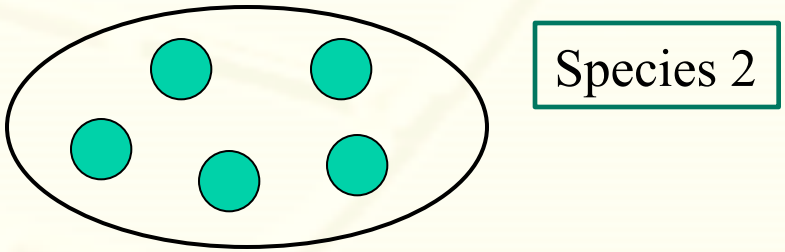




Fixed difference

McDonald-Kreitman (MK) test





Polymorphic

McDonald-Kreitman (MK) test

	Fixed difference	Polymorphism
Synonymous		
Non-synonymous		

McDonald-Kreitman (MK) test

	Fixed difference	Polymorphism
Synonymous	W	Y
Non-synonymous	X	Z

Under neutrality: W / X = Y / Z

Statistically significant deviation from such null hypothesis can be tested by Chi-square test

letters to nature

Nature **351**, 652 - 654 (20 June 1991); doi:10.1038/351652a0

Adaptive protein evolution at the *Adh* locus in *Drosophila*

JOHN H. MCDONALD & MARTIN KREITMAN

	D. melanogaster												*D. simulans*						*D. yakuba*													
Con.	a	b	c	d	e	f	g	h	i	j	k	l	a	b	c	d	e	f	a	b	c	d	e	f	g	h	i	j	k	l		
G	T	T	T	T	T	T	T	T	T	T	T	T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	Repl.	Fixed
T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	C	C	C	C	C	C	C	C	C	C	C	C	Syn.	Fixed
A	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	G	G	G	G	G	G	G	G	G	G	G	G	Repl.	Fixed
G	T	T	T	T	–	–	–	–	–	–	–	T	T	T	T	T	T	T	–	–	–	–	–	–	–	–	–	–	–	–	Syn.	Poly.
T	–	–	–	–	–	–	–	–	–	–	–	–	C	C	–	–	–	C	–	–	–	–	–	–	–	–	–	–	–	–	Syn.	Poly.
C	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	G	C	G	C	G	C	G	C	G	C	G	C	Repl.	Fixed
C	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	G	G	G	G	G	A	G	G	G	G	G	G	Syn.	2 Poly
C	T	T	T	T	T	T	T	T	T	T	T	T	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	Syn.	Fixed
G	–	–	–	–	–	–	–	–	–	–	–	–	–	A	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	Syn.	Poly.
G	–	–	–	–	–	–	–	–	–	–	–	–	T	–	T	T	T	T	–	–	–	–	–	–	–	–	–	–	–	–	Syn.	Poly.

They analyzed polymorphism at the Alcohol Dehydrogenase gene in three Drosophila species: *D. melanogaster, D. simulans, D. yakuba.*

McDonald-Kreitman (MK) test

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

Non-synonymous substitutions among polymorphisms:
2 / (2+42) = 4.5%,
Non-synonymous substitutions among fixed differences:
7 / (7+17) = 29%

This suggests positive selections for adaptive alleles in different species. P-value = 0.4%

Potential issues with MK test

- Ignores multiple substitutions
- Ignores selection against synonymous substitutions,

SIR — McDonald and Kreitman[1] claim that adaptive mutations are largely responsible for the evolution of alcohol dehydrogenase (Adh) because, according to their calculations, in the Adh gene the ratio of nonsynonymous to synonymous substitutions between three *Drosophila* species (7:17) is much larger than the ratio (2:42) within species. However, their test has at least the following problems.

In conclusion, it is not clear as to whether the ADH data can be taken as evidence against the neutral hypothesis.

SIR – Comparing nucleotide sequences of the alcohol dehydrogenase (Adh) gene within and between three species of *Drosophila*, McDonald and Kreitman[1] concluded that the number of non-

We believe that there are subtle but serious problems in McDonald and Kreitman's reasoning.

Thus, these results do not support the conclusion that there is a significant excess of nonsynonymous substitutions resulting from adaptive fixation of mutations.

Adaptive protein evolution in *Drosophila*

Nick G. C. Smith*† & Adam Eyre-Walker*

* *Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK*

For over 30 years a central question in molecular evolution has been whether natural selection plays a substantial role in evolution at the DNA sequence level[1,2]. Evidence has accumulated over the last decade that adaptive evolution does occur at the protein level[3,4], but it has remained unclear how prevalent adaptive evolution is. Here we present a simple method by which the number of adaptive substitutions can be estimated and apply it to data from *Drosophila simulans* and *D. yakuba*. We estimate that 45% of all amino-acid substitutions have been fixed by natural selection, and that on average one adaptive substitution occurs every 45 years in these species.

MK test on real data

This is in contradictory with the neutral theory

Smith, Eyre-Walker, Nature, 2002

Positive selection among human genes

Nature **437**, 1153-1157 (20 October 2005) | doi:10.1038/nature04240; Received 24 April 2005; Accepted 14 September 2005

Natural selection on protein-coding genes in the human genome

Carlos D. Bustamante[1], Adi Fledel-Alon[1], Scott Williamson[1], Rasmus Nielsen[1,2], Melissa Todd Hubisz[1], Stephen Glanowski[3], David M. Tanenbaum[3], Thomas J. White[4], John J. Sninsky[4], Ryan D. Hernandez[1], Daniel Civello[4], Mark D. Adams[5], Michele Cargill[4,7] & Andrew G. Clark[6,7]

. Here we contrast patterns of coding sequence polymorphism identified by direct sequencing of 39 humans for over 11,000 genes to divergence between humans and chimpanzees, and find strong evidence that natural selection has shaped the recent molecular evolution of our species. Our analysis discovered 304 (9.0%) out of 3,377 potentially informative loci showing evidence of rapid amino acid evolution.

Positive selection among human genes

% of loci (%)	Locus type	Outgroup species	Method	Study
20%	Protein	Chimpanzee	MK	Zhang and Li 2005
6%	Protein	Chimpanzee	MK	Bustamante et al. 2005
0-9%	Protein	Chimpanzee	MK	Chimpanzee Sequencing and Analysis Consortium 2005
10-20%	Protein	Chimpanzee	MK	Boyko et al. 2008
9.8%	Protein	Chimpanzee	dn/ds	Nielsen et al. 2005a
1.1%	Protein	Chimpanzee	dn/ds	Bakewell et al. 2007
35%	Protein	Old-world monkey	MK	Fay et al. 2001
0%	Protein	Old-world monkey	MK	Zhang and Li 2005
0%	Protein	Old-world monkey	MK	Eyre-Walker and Keightley 2009
0.4%	Protein	Old-world monkey	dn/ds	Nielsen et al. 2005b
0%	Protein	Mouse	MK	Zhang and Li 2005

More examples of Positive Selection

Adaptive evolution of non-coding DNA in *Drosophila*

Peter Andolfatto[1] Nature 2005

Expression profiling in primates reveals a rapid evolution of human transcription factors

Yoav Gilad[1]†, Alicia Oshlack[2], Gordon K. Smyth[2], Terence P. Speed[2,3] & Kevin P. White[1] **Nature 2004**

Diet and the evolution of human amylase gene copy number variation

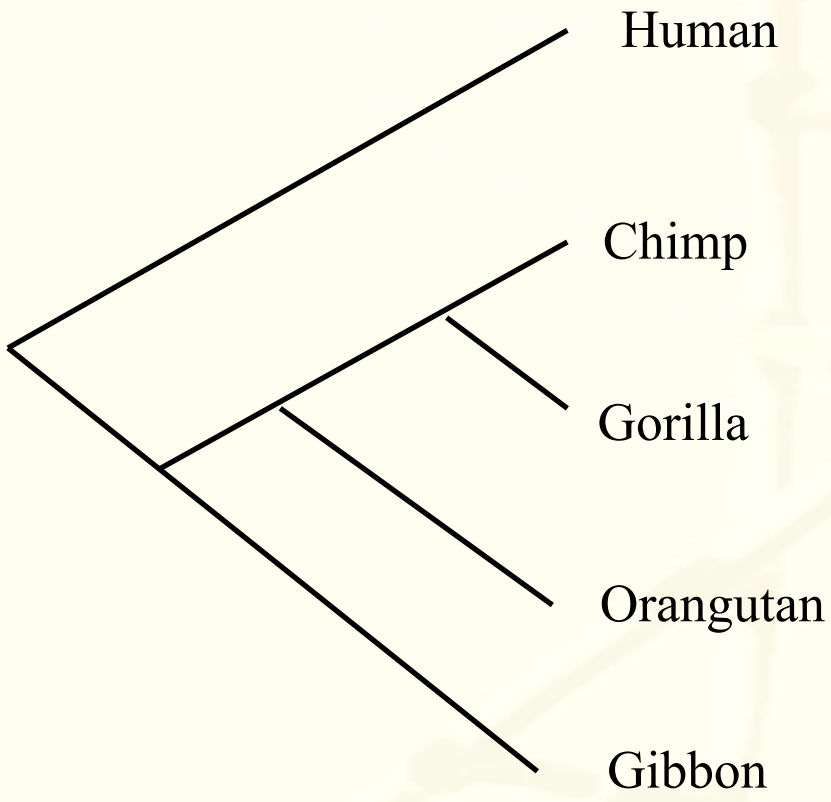
George H Perry[1,2], Nathaniel J Dominy[3], Katrina G Claw[1,4], Arthur S Lee[2], Heike Fiegler[5], Richard Redon[5], John Werner[4], Fernando A Villanea[3], Joanna L Mountain[6], Rajeev Misra[4], Nigel P Carter[5], Charles Lee[2,7,8] & Anne C Stone[1,8]

Be careful about confounding factors: population history, migration, and **population size**

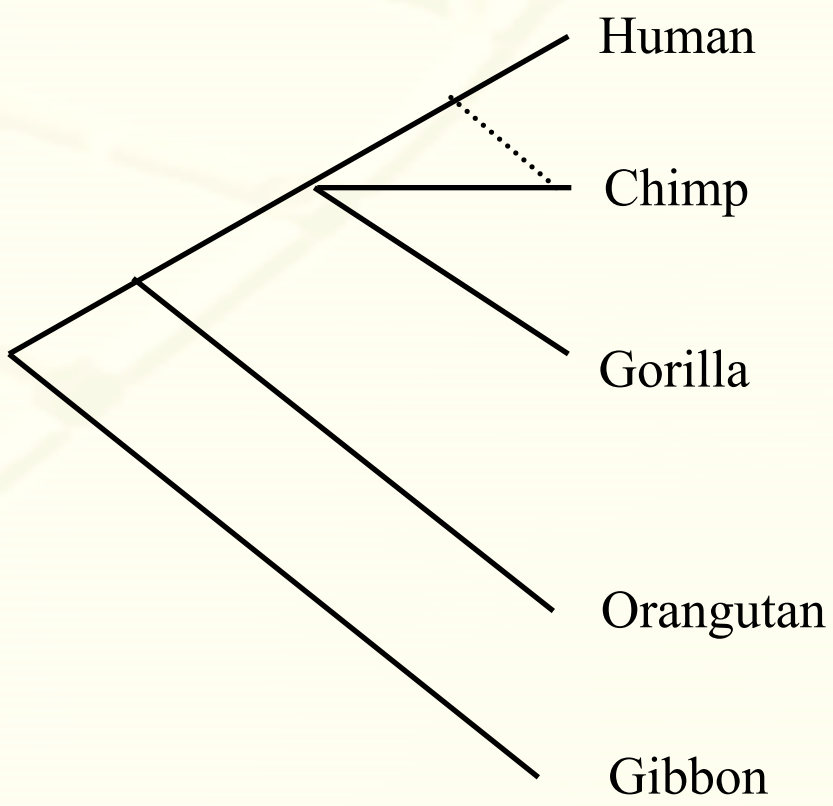
Coffee Break ?

Phylogenetic analysis using DNA sequence

Phylogenetic analysis using DNA sequence



Traditional



Molecular

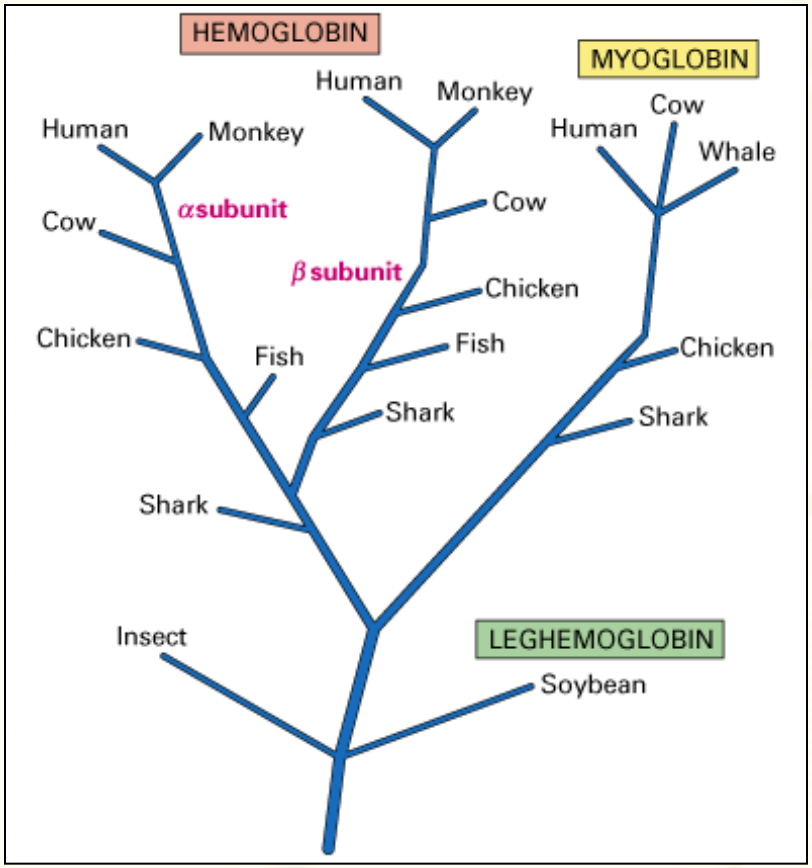
Two Areas in Phylogenetic analysis

- 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 ?

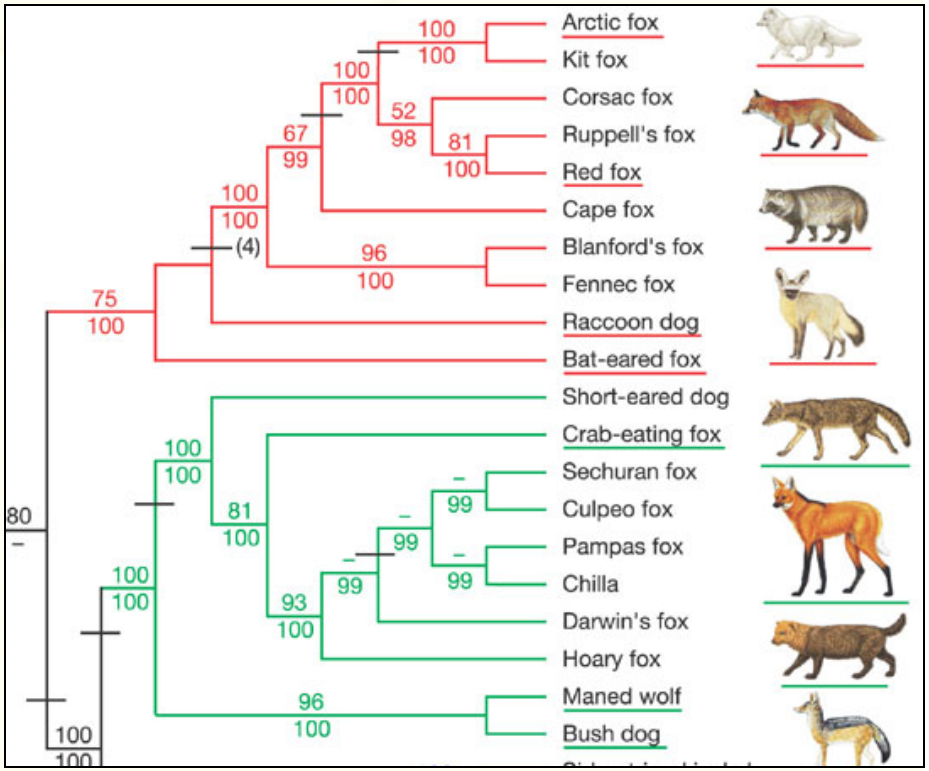
- **Character and rate** analysis:
  - Using phylogeny as a framework to understand the evolution of traits or genes.
  - For example, is gene X under positive or purifying selection ?

Phylogenetic Tree

Gene Tree

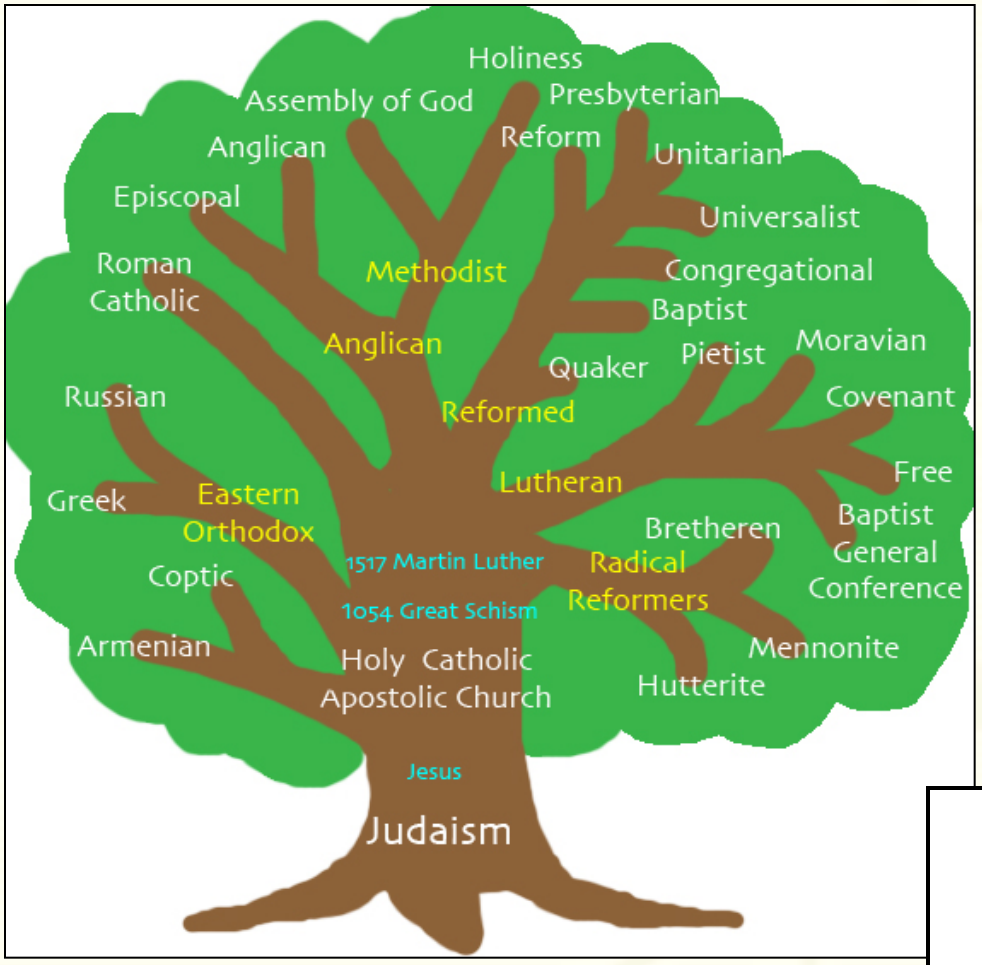


Species Tree



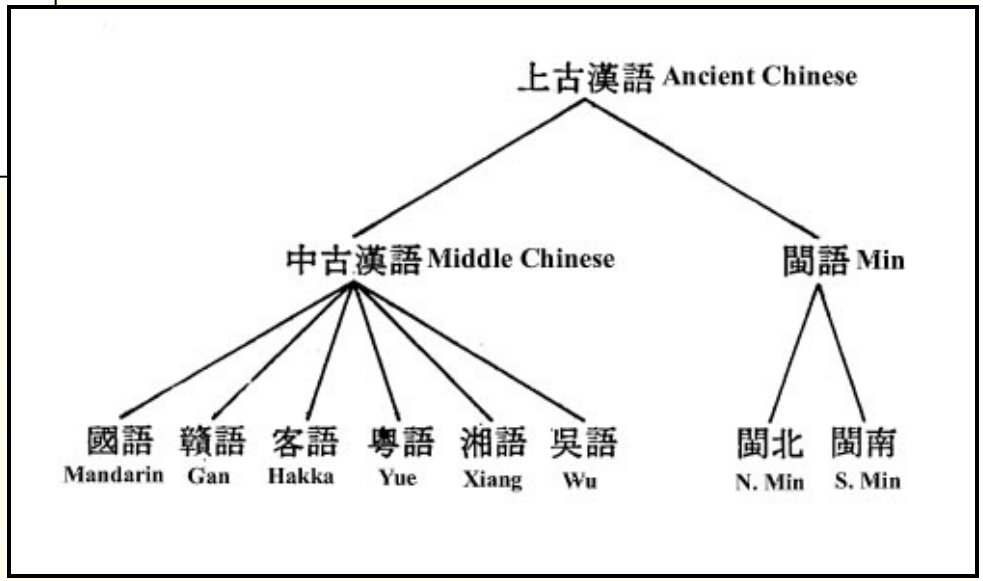
Lindblad-Toh Nature 2005

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

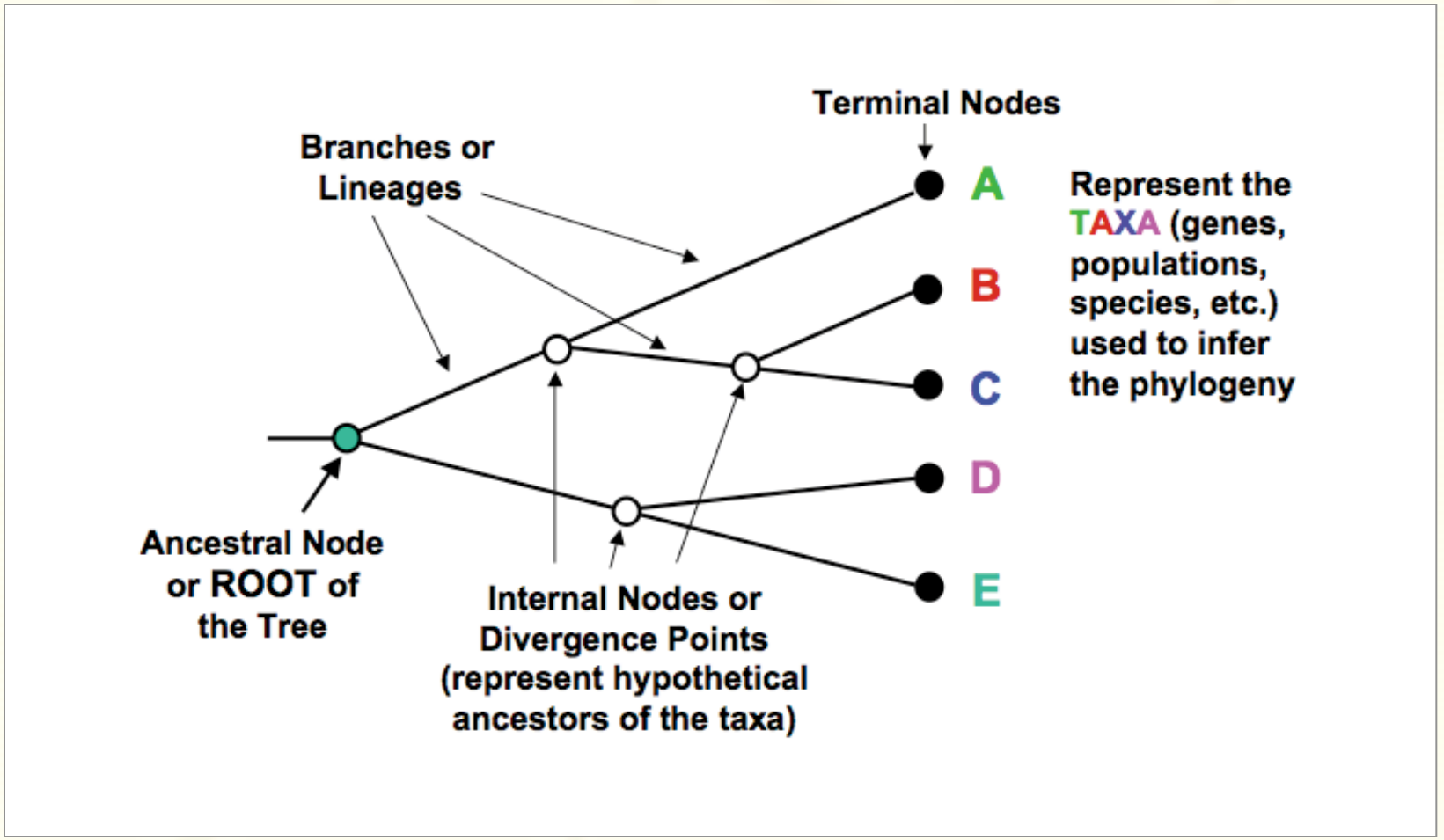


Tree of world religions

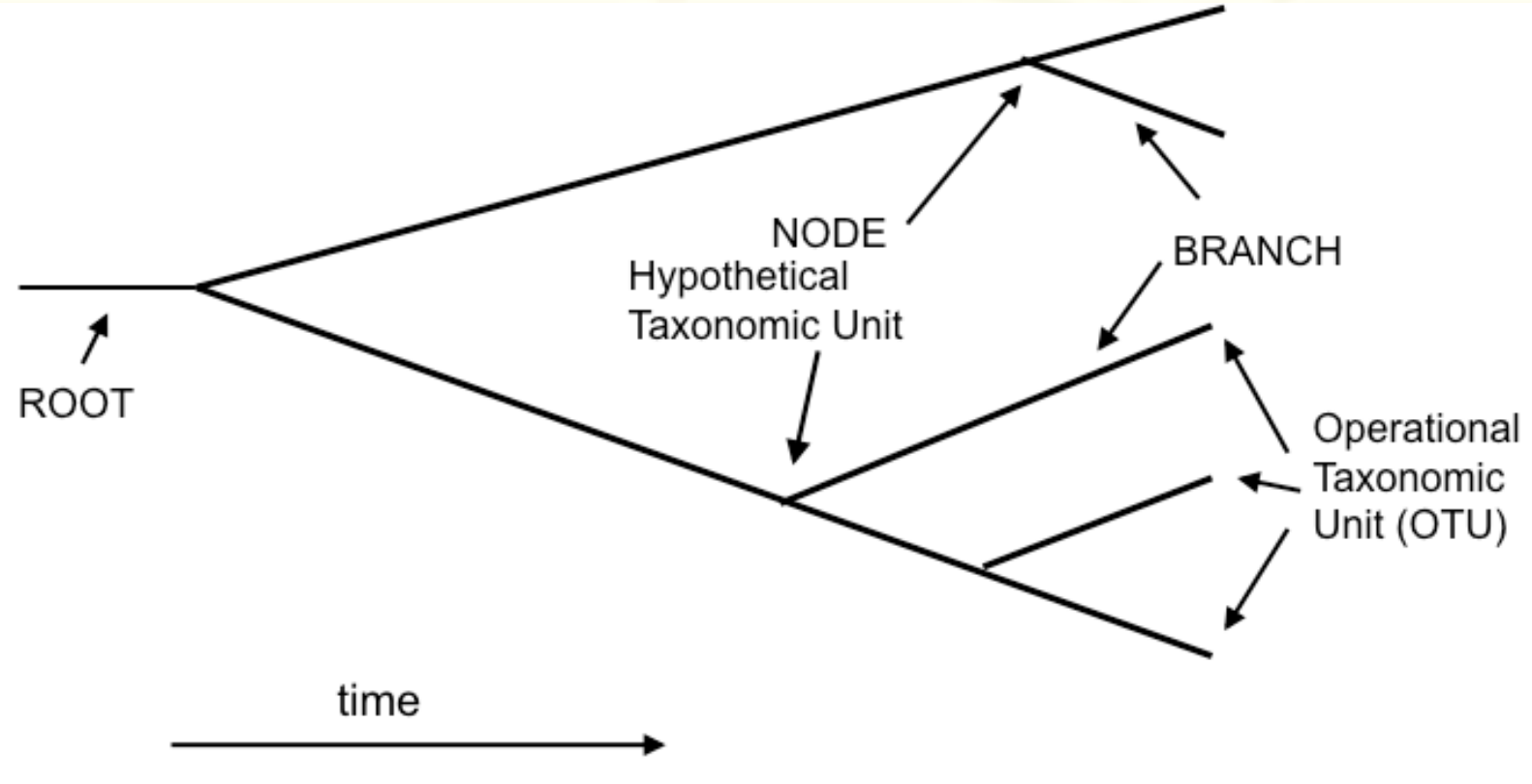
Tree of languages



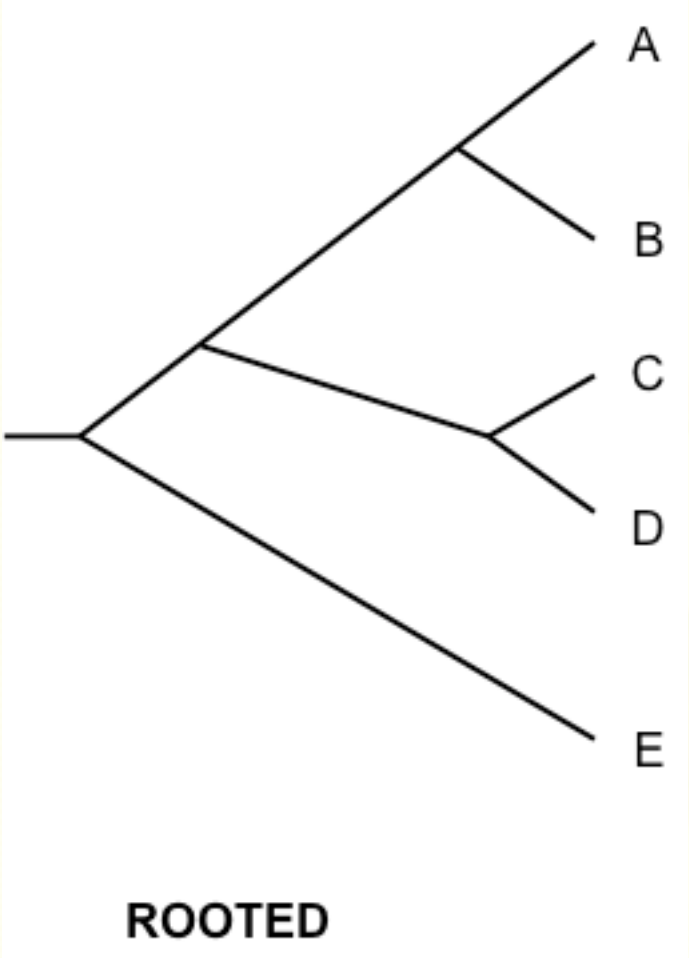
Phylogenetic Tree Terminology

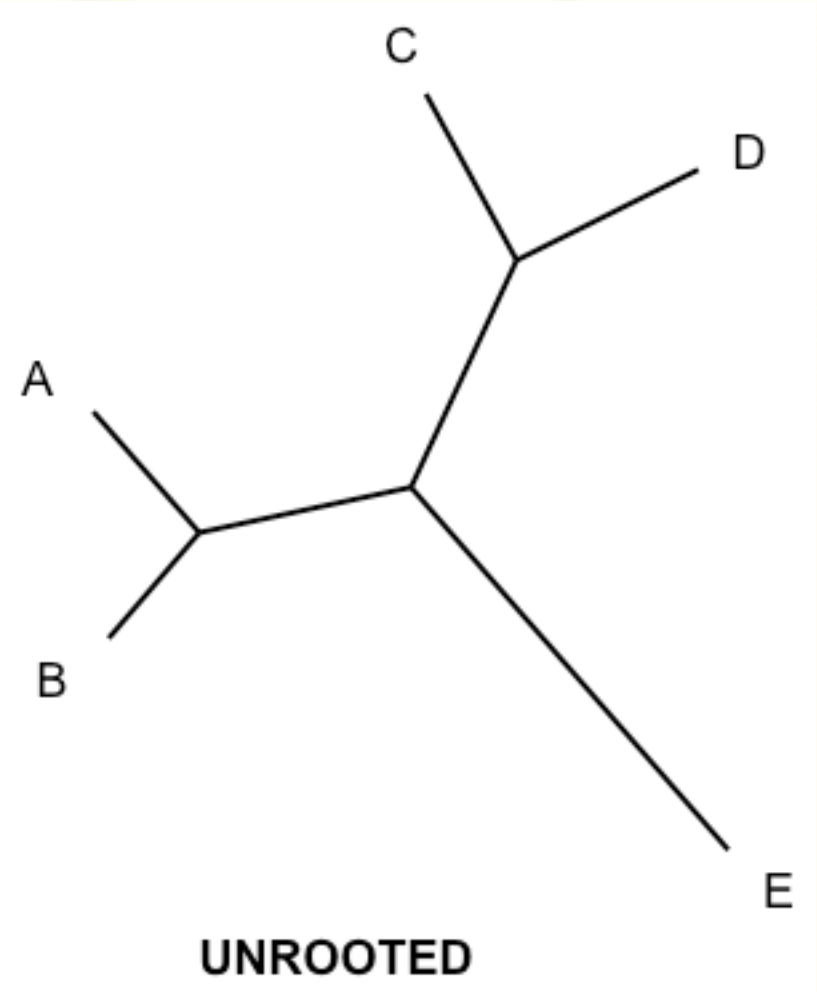


Phylogenetic Tree Terminology

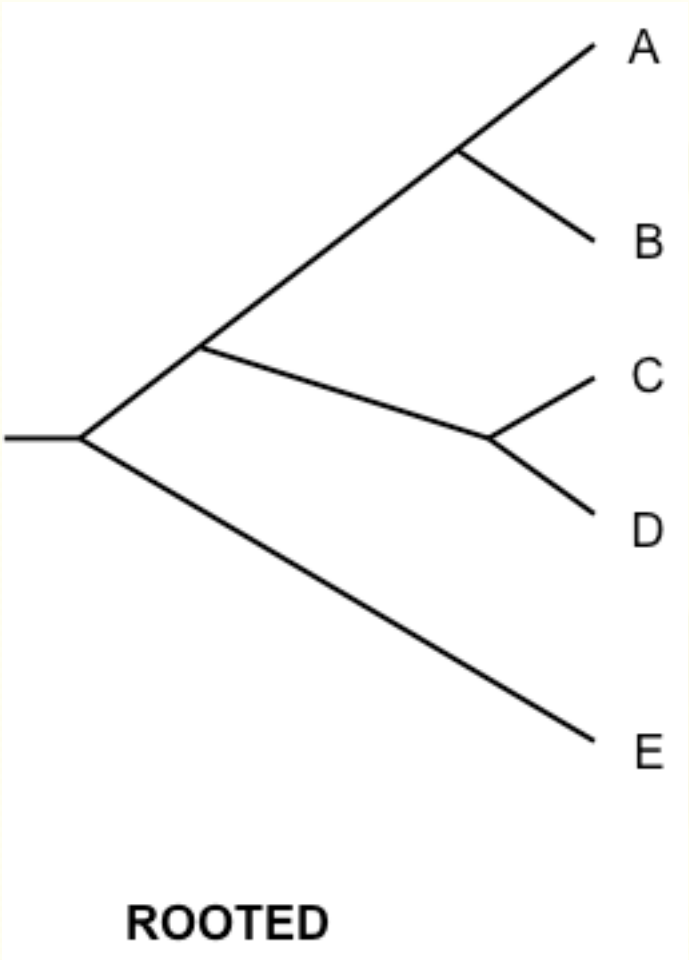


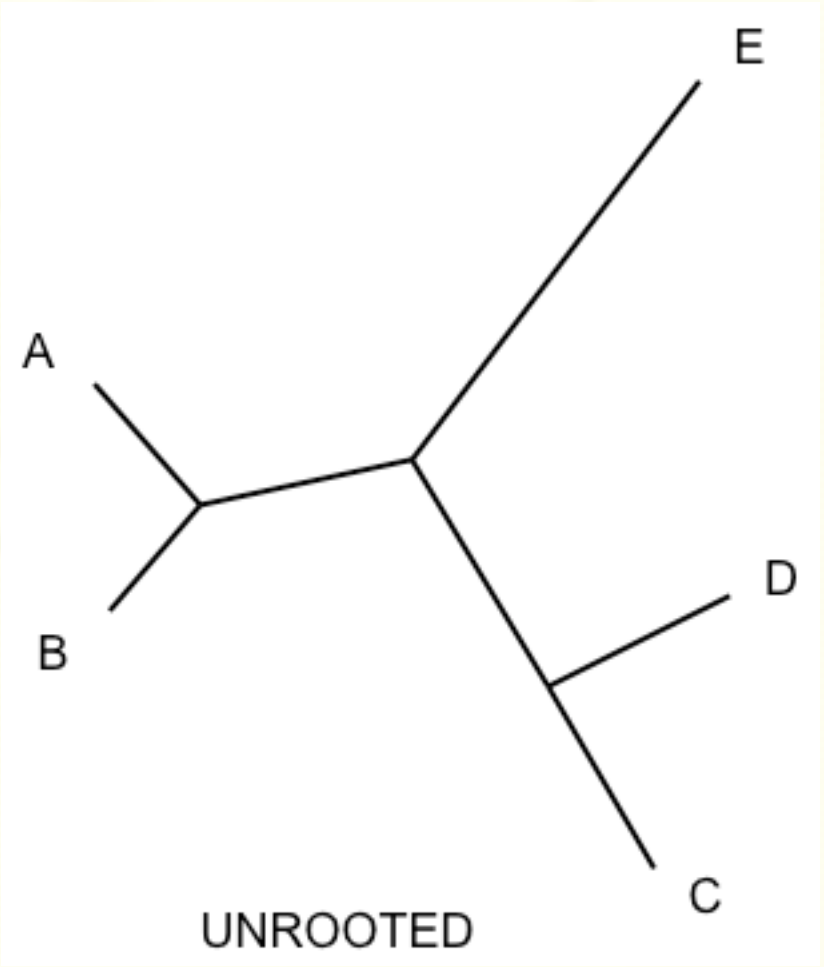
Rooted and unrooted trees



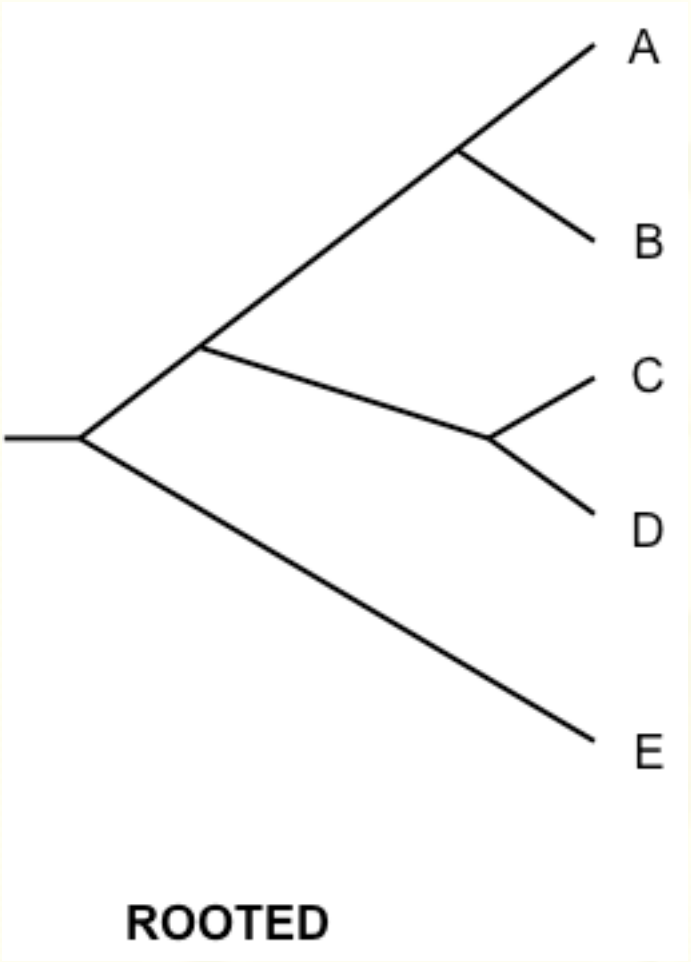


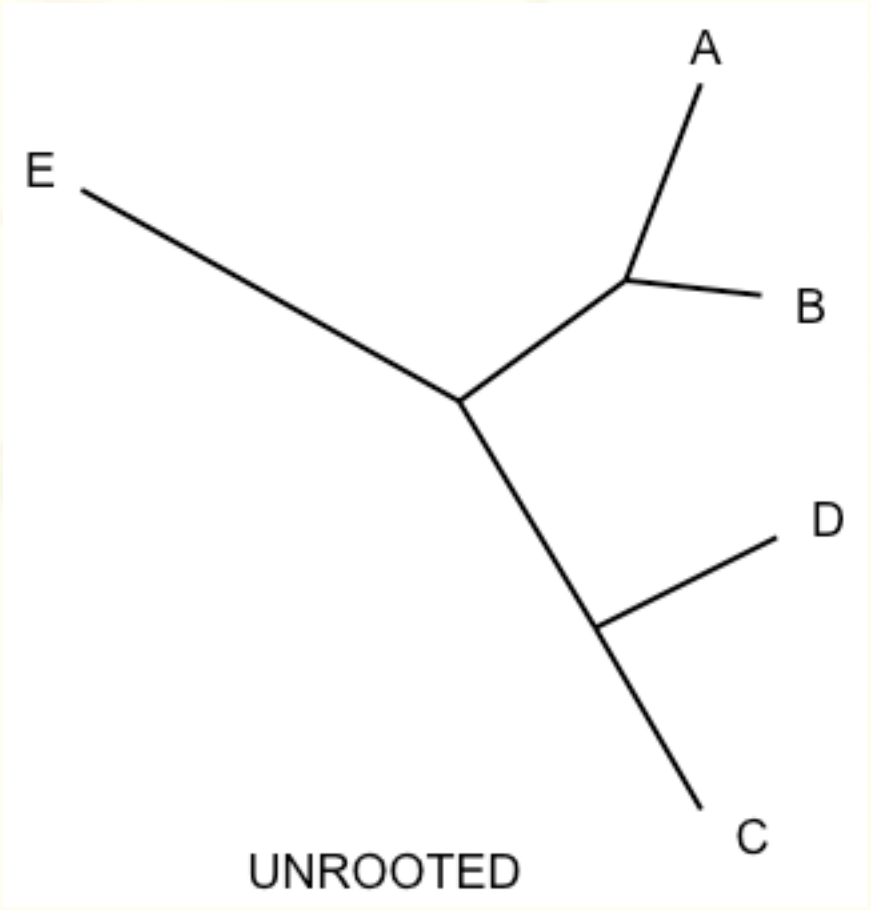
Rooted and unrooted trees

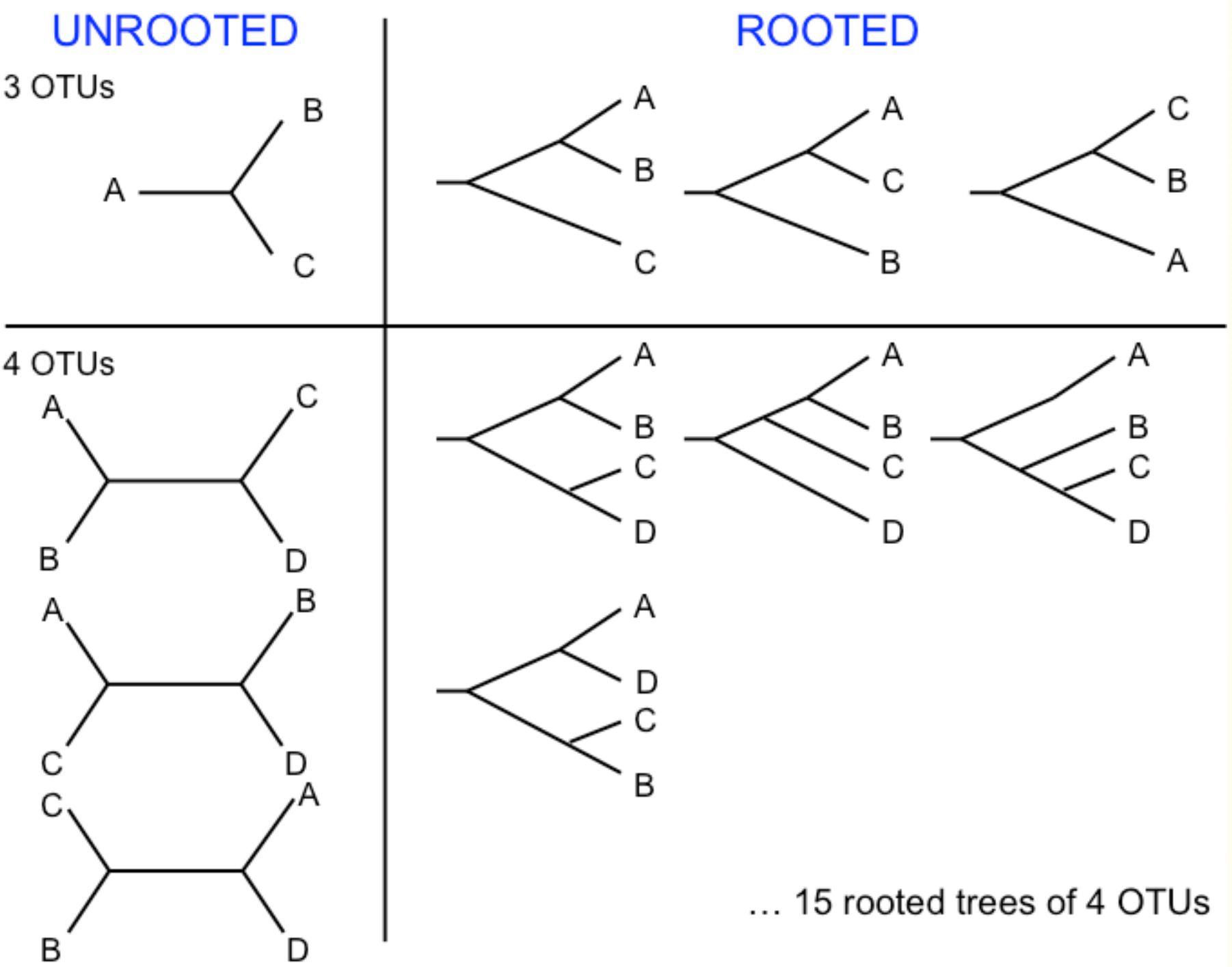




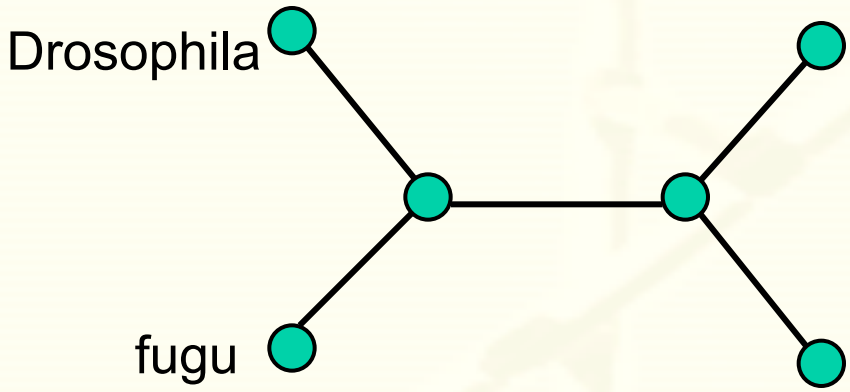
Rooted and unrooted trees

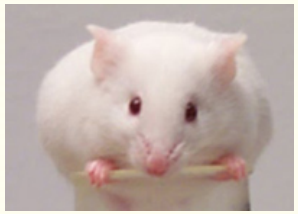




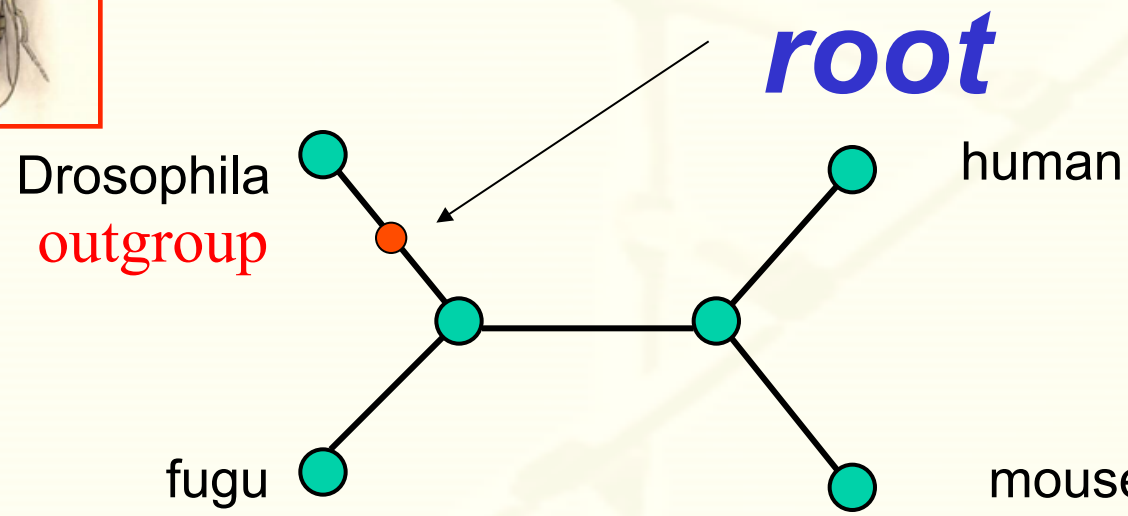

UNROOTED
ROOTED
3 OTUs
A
B
C
A
B
C
A
C
B
C
B
A
4 OTUs
A
B
C
D
A
B
C
D
A
B
C
D
A
B
C
D
A
B
C
D
A
B
C
D
… 15 rooted trees of 4 OTUs

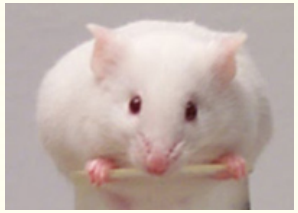
Root a tree using an outgroup



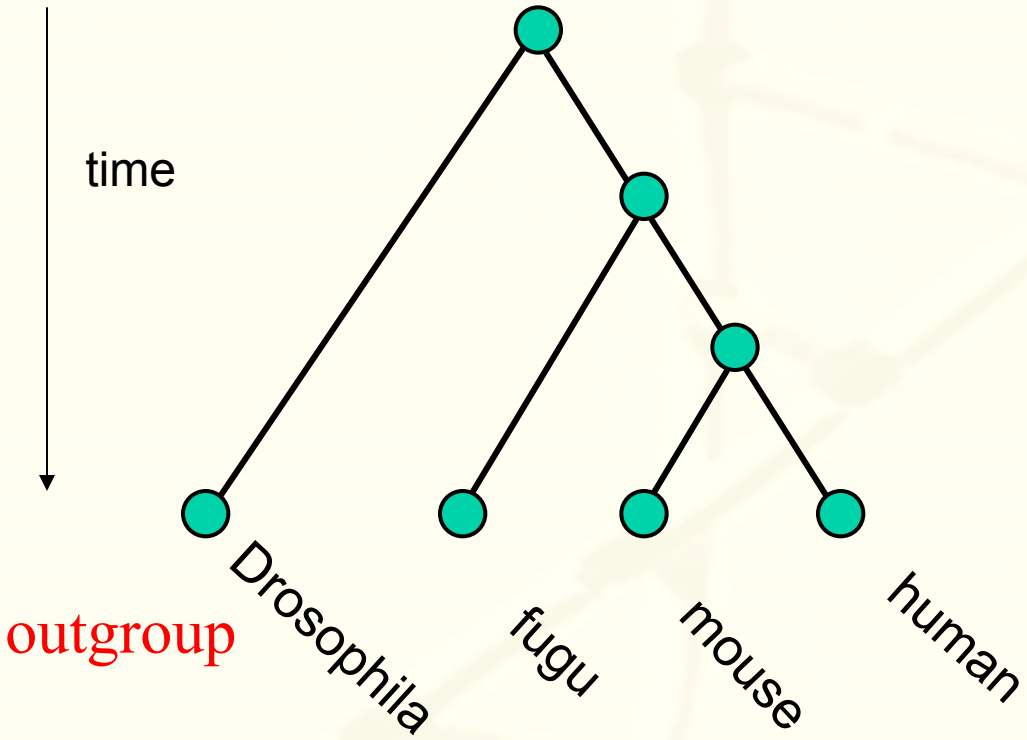


Root a tree using an outgroup

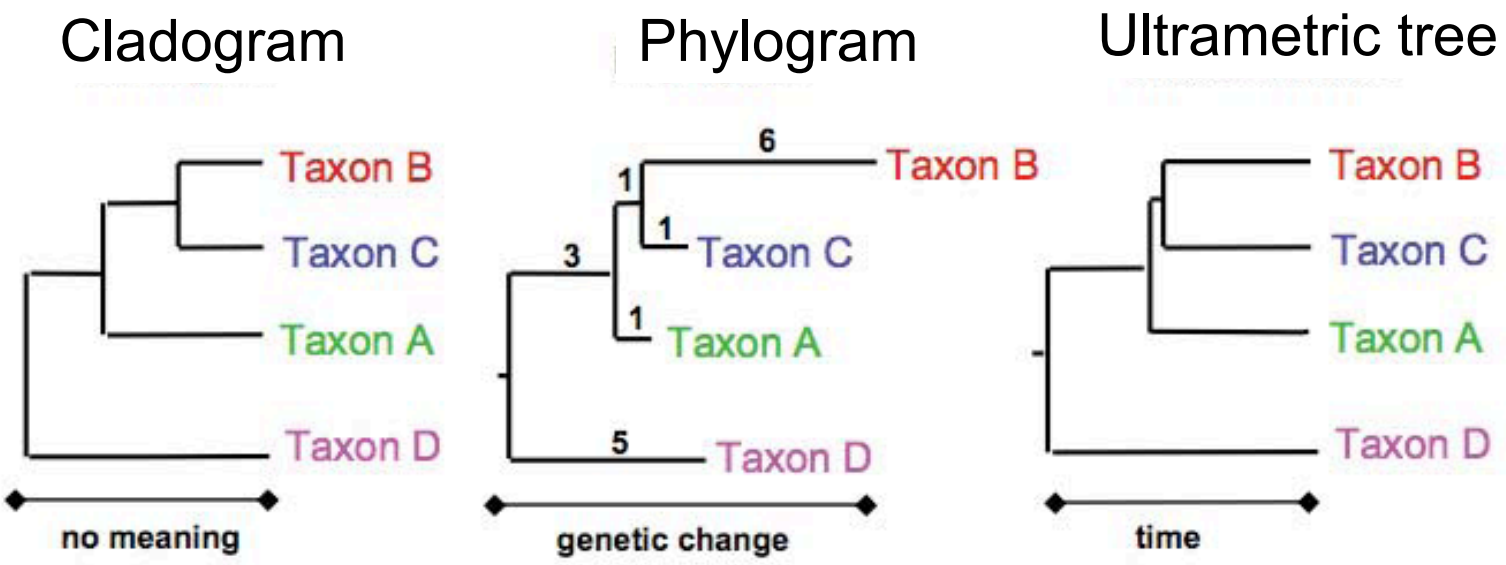




Root a tree using an outgroup



Three Types of Trees



All show the same evolutionary relationships, or branching orders, between the taxa.

Reconstruct phylogeny from molecular data

Methods of Tree reconstruction

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

Methods of Tree reconstruction

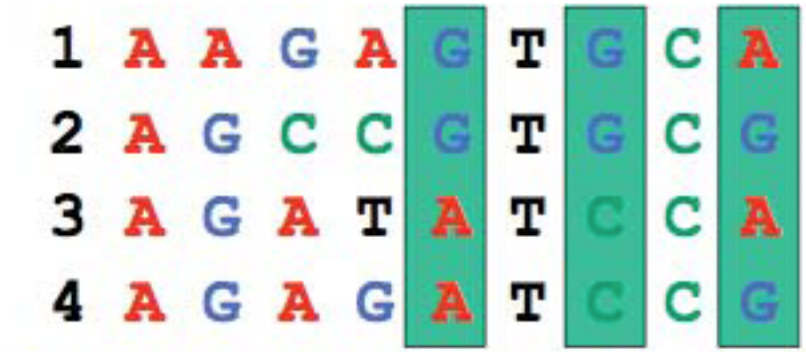
- **Maximum Parsimony methods**
- **Distance based methods**
- **Maximum Likelihood methods**
- **Bayesian methods**

(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.

Maximum Parsimony Example



- four sequences, three possible unrooted trees
- Some sites are informative, others are not
- Informative site has same sequence character in at least two different sequences
- Only informative sites are considered

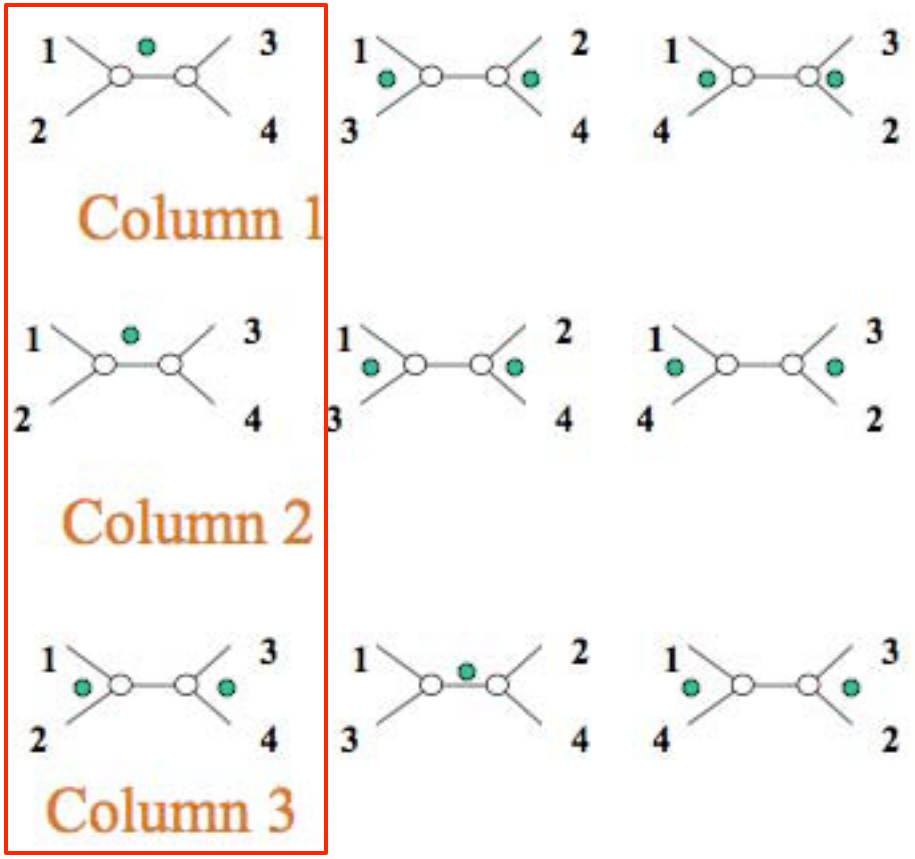
Maximum Parsimony Example

1 G G A

2 G G G

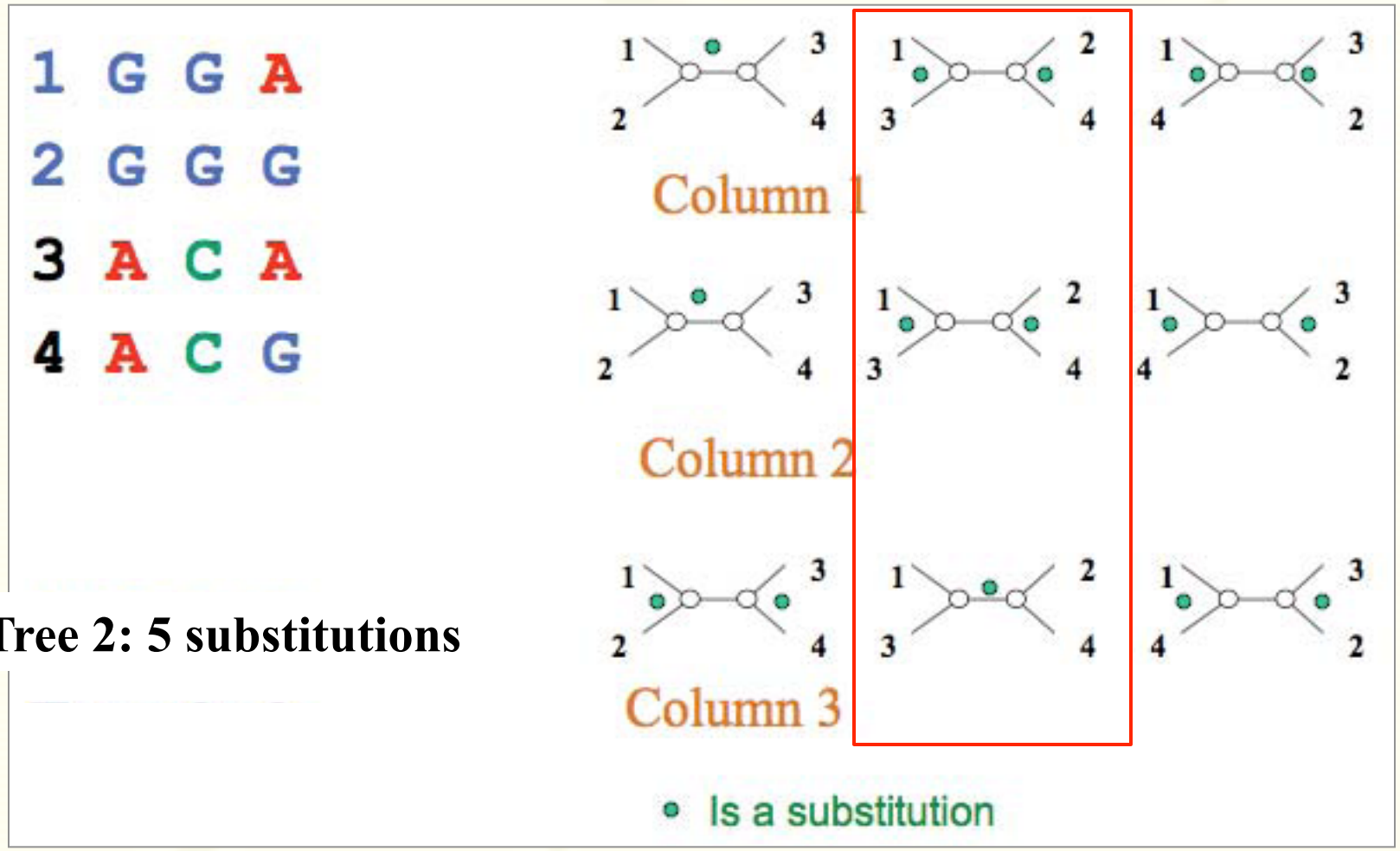
3 A C A

4 A C G



Tree 1: 4 substitutions

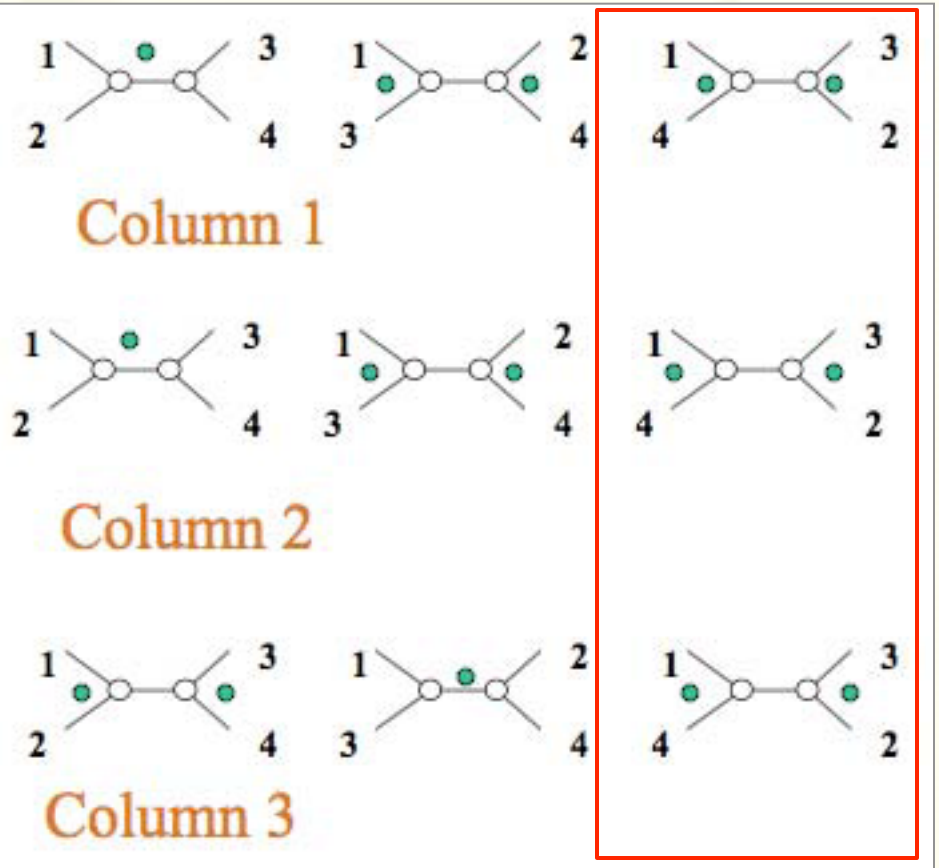
Maximum Parsimony Example



Tree 2: 5 substitutions

Maximum Parsimony Example

1	G	G	A
2	G	G	G
3	A	C	A
4	A	C	G



Tree 3: 6 substitutions

Maximum Parsimony Example

1 G G A

2 G G G

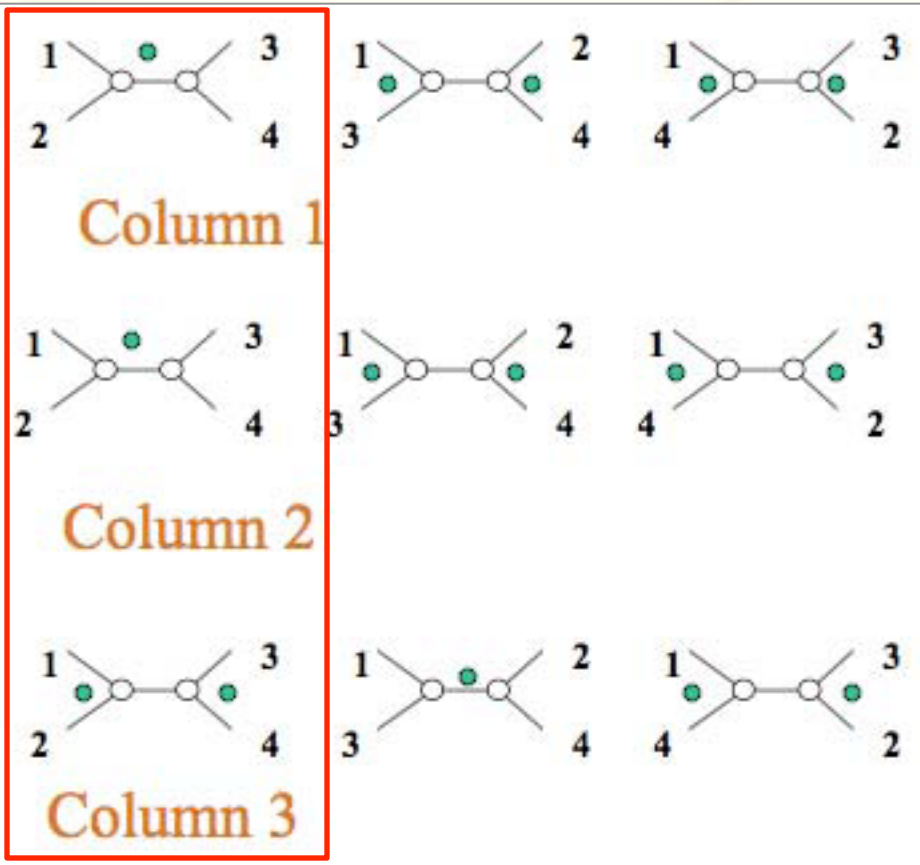
3 A C A

4 A C G

Tree 1: 4

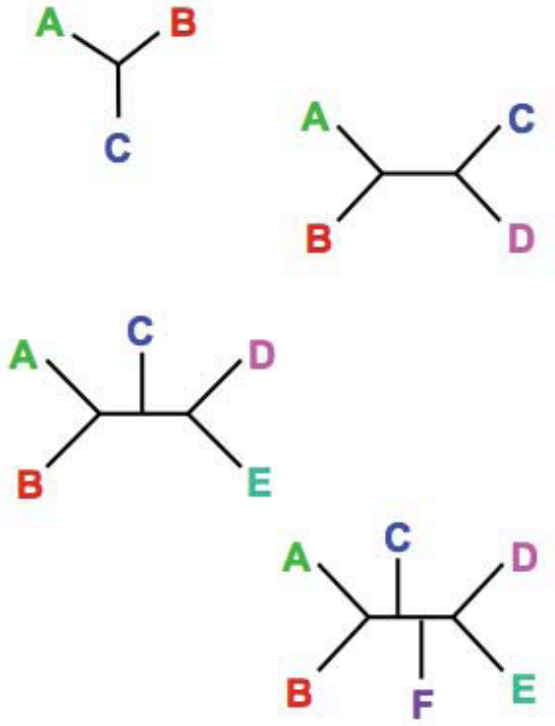
Tree 2: 5

Tree 3: 6



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

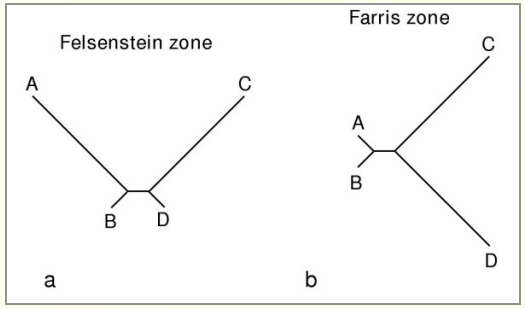
$N_u=(2n-5)*(2n-7)*...*3*1=(2n-5)!/[2^{n-3}*(n-3)!]$

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

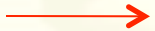


Distance based methods

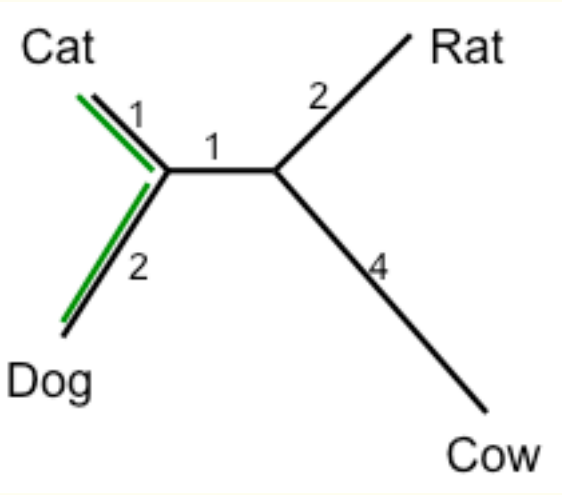
- Estimate the number of substitutions between each pair of sequences in a group of sequences.
- Try to build a tree so that the branch lengths represent the pair-distances.
- What are these “distances” ? Example: sequence identity between two protein and DNA sequences

Distance based methods

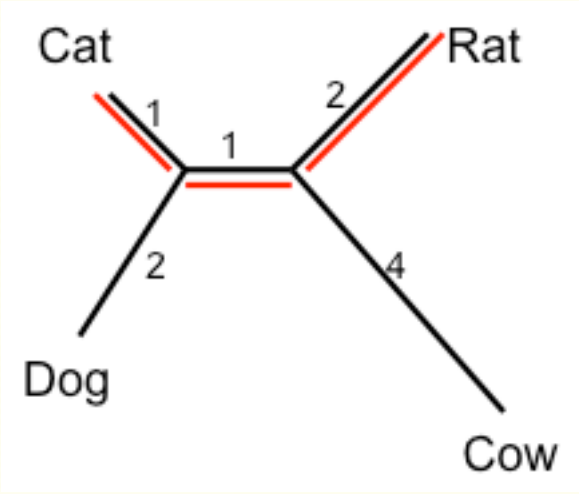
Cat	ATTTGCGGTA
Dog	ATCTGCGATA
Rat	ATTGCCGTTT
Cow	TTCGCTGTTT



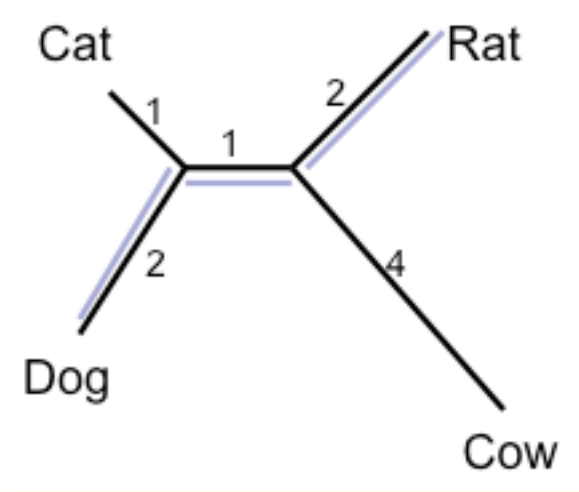
	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



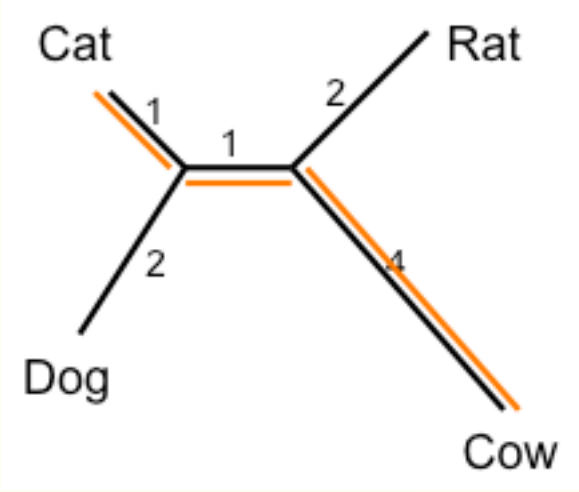
	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



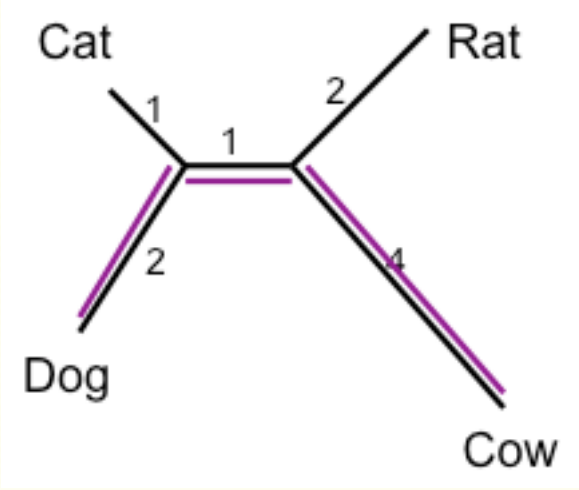
	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



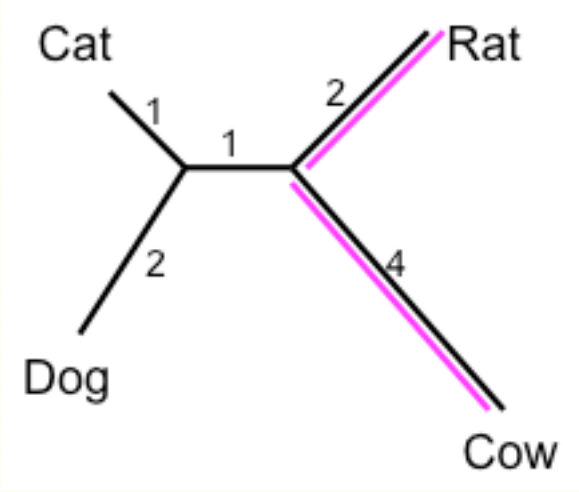
	Cat	Dog	Rat
Dog	3		
Rat	4	5	
Cow	6	7	6



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



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



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.

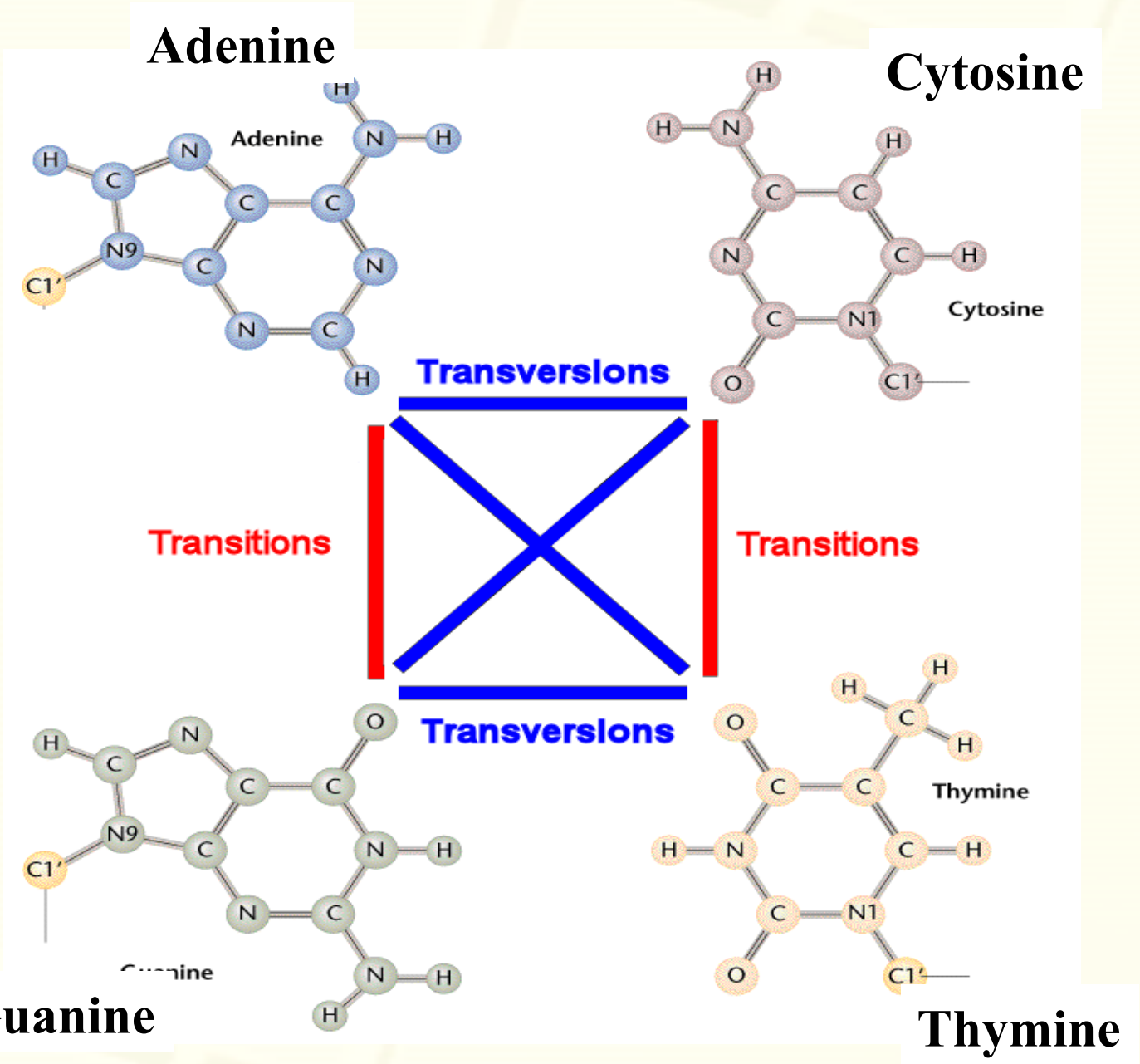
- Substitution rates are different between different types of nucleotides

Substitution models

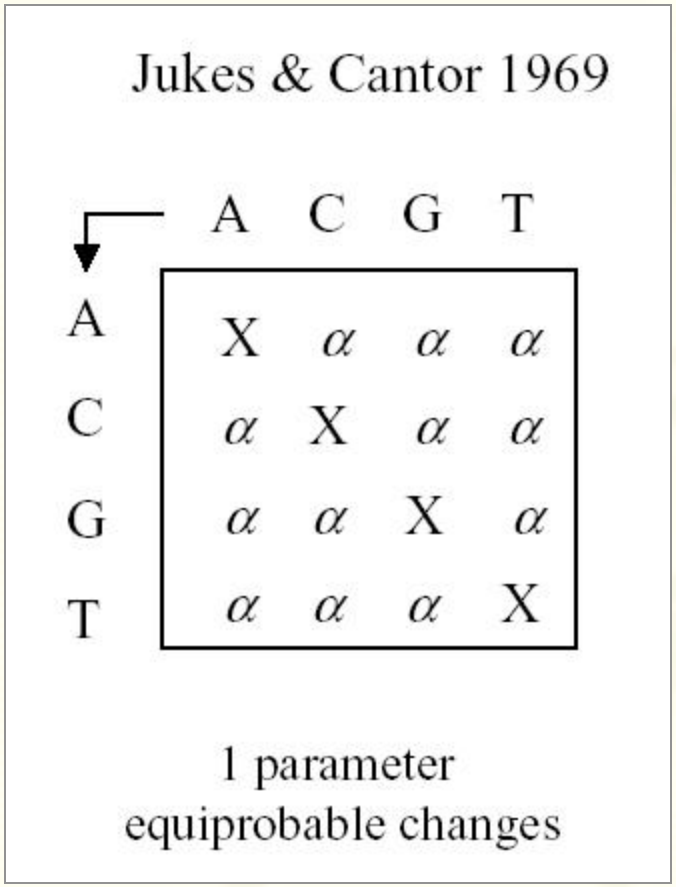
- Substitution model: given the observed number of changes we estimate the actual number of changes that have happened. Some assumptions are needed regarding the probability of substitution of a nucleotide by another.

- Some are naïve, while others are mathematically complex.
  - Jukes-Kantor one parameter model (1969)
  - Kimura Two-parameter model (1980)
  - F81 model (Felsenstein 1981), considers equilibrium frequency.
  - HKY85 6-parameter model (Hasegawa, Kishino and Yano 1985)
  - Tamura92 model (Tamura 1992)
  - TN93 model (Tamura and Nei 1993)

- These models become less accurate for highly divergent sequences.


Adenine
Cytosine
Adenine
Cytosine
Transversions
Transitions
Transitions
Transversions
Thymine
Guanine
Thymine

Jukes & Cantor's one-parameter model



Assumption: substitutions occur with equal probabilities *α* among the four nucleotide types.

Kimura's 2-parameter model

Kimura 1980

	A	C	G	T
A	X	α	$\kappa.\alpha$	α
C	α	X	α	$\kappa.\alpha$
G	$\kappa.\alpha$	α	X	α
T	α	$\kappa.\alpha$	α	X

2 parameters
transition rate ≠
transversion rate

Assumption: The rate of transitions and transversions are different; the ratio between transition and transversion is *k*

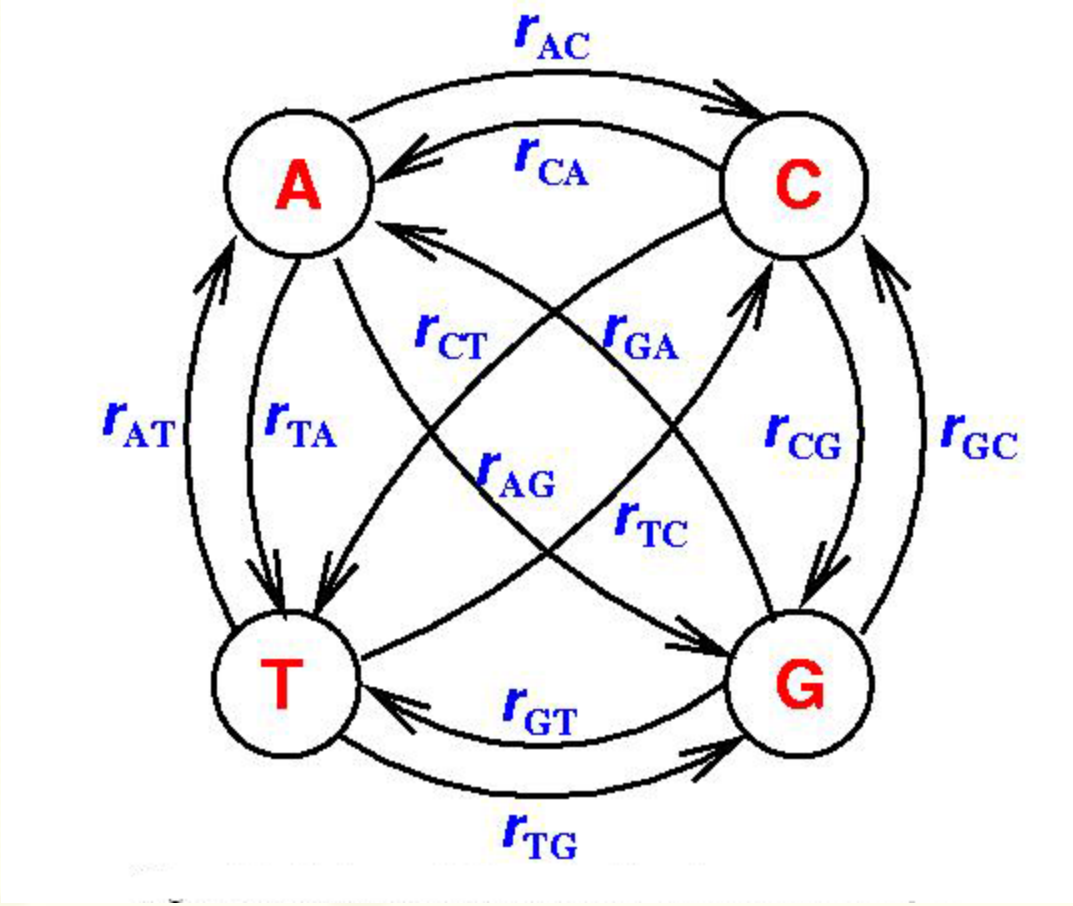
Hasegawa-Kishino-Yano (HKY85) 5-parameter model

	A	C	G	T
A	–	$\pi_C\beta$	$\pi_G\kappa\beta$	$\pi_T\beta$
C	$\pi_A\beta$	–	$\pi_G\beta$	$\pi_T\kappa\beta$
G	$\pi_A\kappa\beta$	$\pi_C\beta$	–	$\pi_T\beta$
T	$\pi_A\beta$	$\pi_C\kappa\beta$	$\pi_G\beta$	–

Assumption: On the basis of Kimura model, added equilibrium frequencies for 4 nucleotides: πA, πG, πC, πT.

$$\pi A+\pi G+\pi C+\pi T=1$$

The extreme – 12 parameter model



Protein substitution models

- Amino acids substitution models are usually empirically estimated from homolog sequences.
  - PAM: *Percent Accepted Mutation*: Dayhoff, 1970s,
  - BLOSUM model: *BLOck SUbstitution Matrix*
  - JTT model: Jones DT, Taylor WR, Thornton JM (1992).

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**																			
S	-1	**4**																		
T	-1	1	**4**																	
P	-3	-1	1	**7**																
A	0	1	0	-1	**4**															
G	-3	0	-2	-2	0	**6**														
N	-3	1	0	-2	-2	0	**6**													
D	-3	0	-1	-1	-2	-1	1	**6**												
E	-4	0	-1	-1	-1	-2	0	2	**5**											
Q	-3	0	-1	-1	-1	-2	0	0	2	**5**										
H	-3	-1	0	-2	-2	-2	1	-1	0	0	**8**									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	**5**							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**1**

Make trees from pair-wide distances

- **Neighboring joining**
  - Pair with the smallest branch lengths chosen to be joined
  - A new distance table is created with joint sequences entered as a composite.
  - Repeat process to select next pair to join.
  - Repeat process until correctly branched tree and distances identified

- UPGMA
  - Unweighted Pair Group Method with Arithmetic Mean

More advanced methods

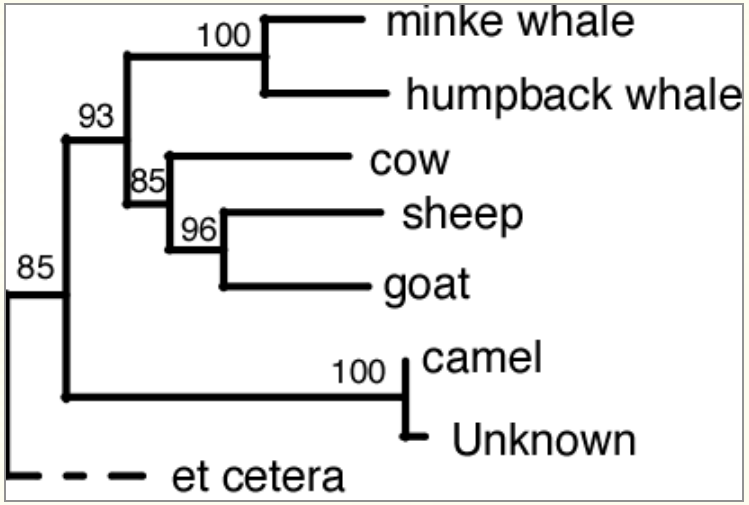
- **Maximum likelihood methods:**
  - ML methods evaluate phylogenetic hypothesis in terms of the **probability** that a proposed model and the parameters gave rise to the observed data. The tree found to have the highest likelihood is considered to be the optimal tree.

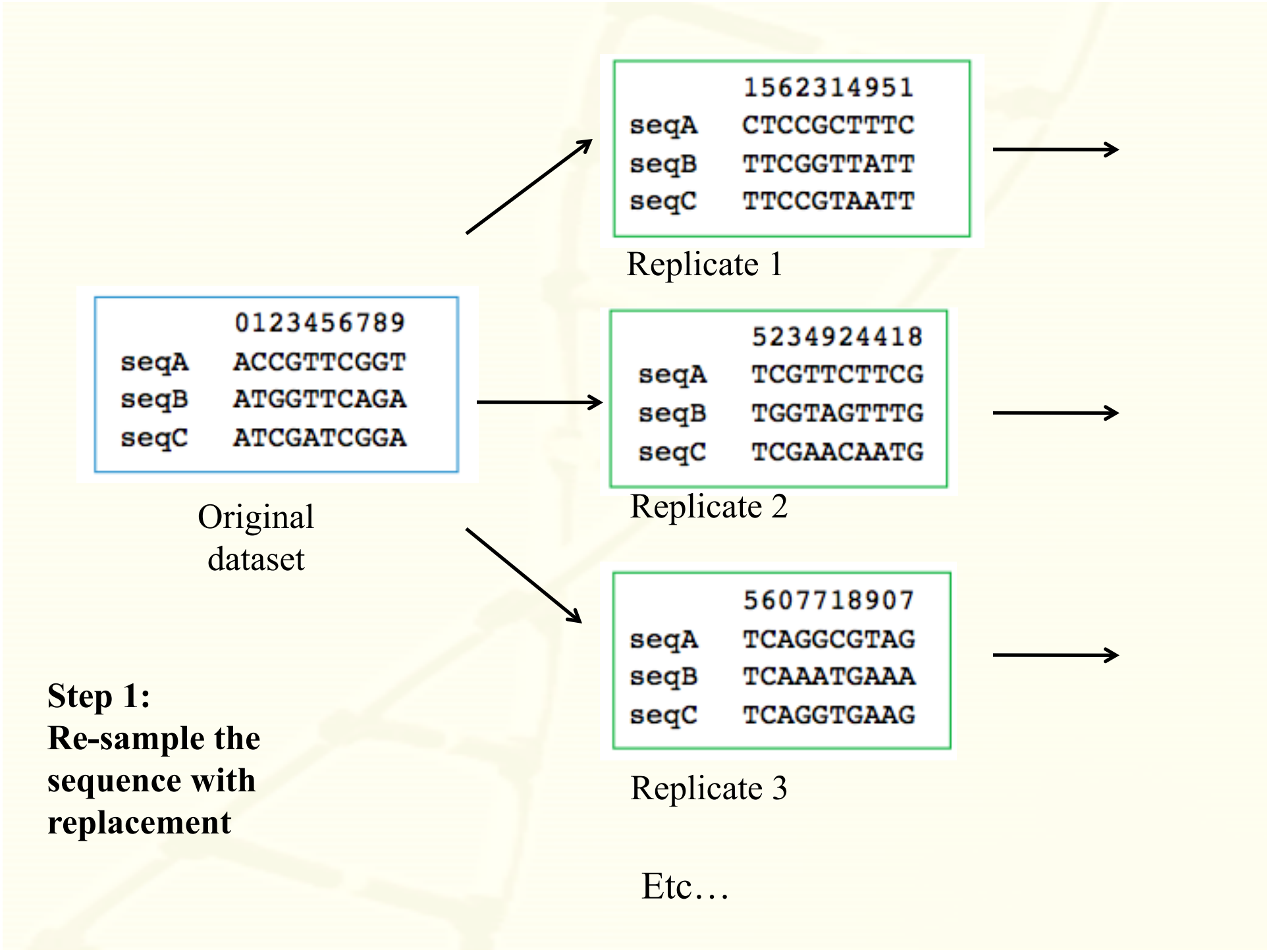
- **Bayesian Markov chain Monte Carlo methods**
  - Generate a large population of trees, then take a random walk through the “tree space” until a perfect tree is found.

Bootstrapping

- How robust is the tree ? How much does the data support the tree ? How confident are we about a particular branch point ?

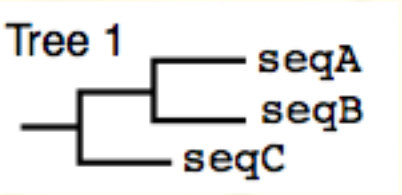
- To test this, we repeatedly re-sampled the data with the replacement and re-calculate the tree, and ask how many times do we still see the original tree or branch point.




0123456789
seqA ACCGTTCGGT
seqB ATGGTTCAGA
seqC ATCGATCGGA
Original dataset
1562314951
seqA CTCCGCTTTC
seqB TTCGGTTATT
seqC TTCCGTAATT
Replicate 1
5234924418
seqA TCGTTCTTCG
seqB TGGTAGTTTG
seqC TCGAACAATG
Replicate 2
5607718907
seqA TCAGGCGTAG
seqB TCAAATGAAA
seqC TCAGGTGAAG
Replicate 3
Step 1:
Re-sample the sequence with replacement
Etc…

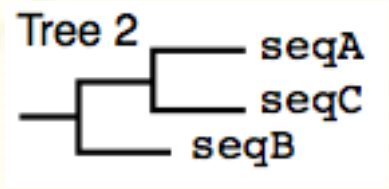
Step 2:
Build trees

```
          1562314951
seqA      CTCCGCTTTC
seqB      TTCGGTTATT
seqC      TTCCGTAATT
```



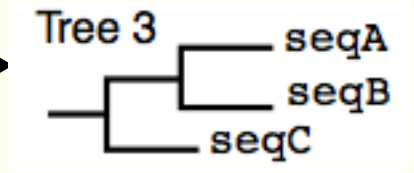
Replicate 1

```
          5234924418
seqA      TCGTTCTTCG
seqB      TGGTAGTTTG
seqC      TCGAACAATG
```



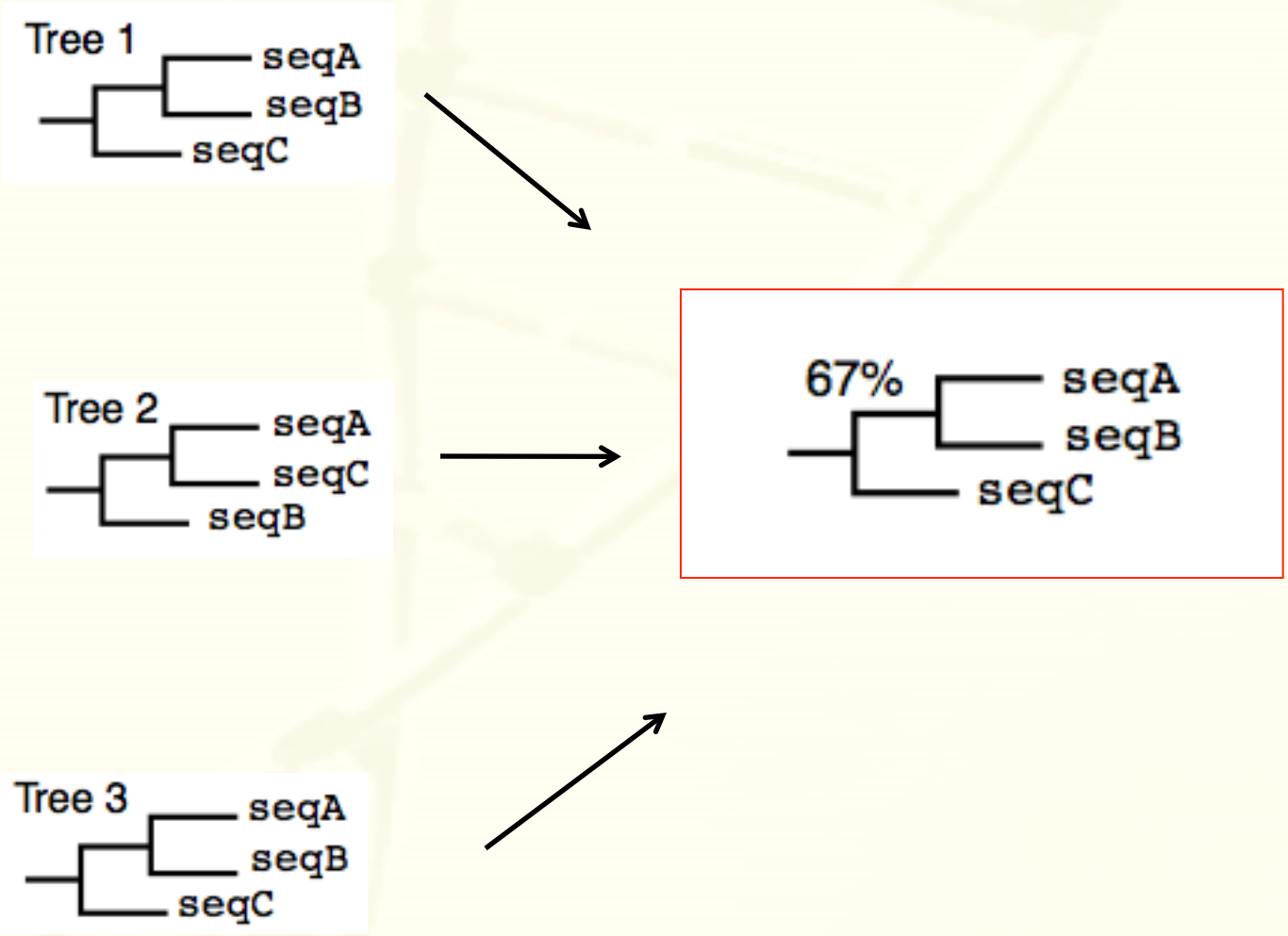
Replicate 2

```
          5607718907
seqA      TCAGGCGTAG
seqB      TCAAATGAAA
seqC      TCAGGTGAAG
```



Replicate 3

Etc …

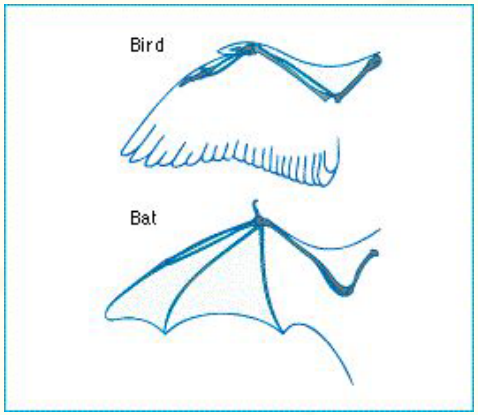


Step 2:
Build consensus tree with bootstrapping value

Homoplasy vs Homology

- A **homology** is a character shared between two species that was present in their common ancestor; a **homoplasy** is a character shared between two species that was not present in their common ancestor bur caused by parallel or convergent evolution.

- Homologous similarity reveals a phylogenetic relationship; homoplasious similarity does not.



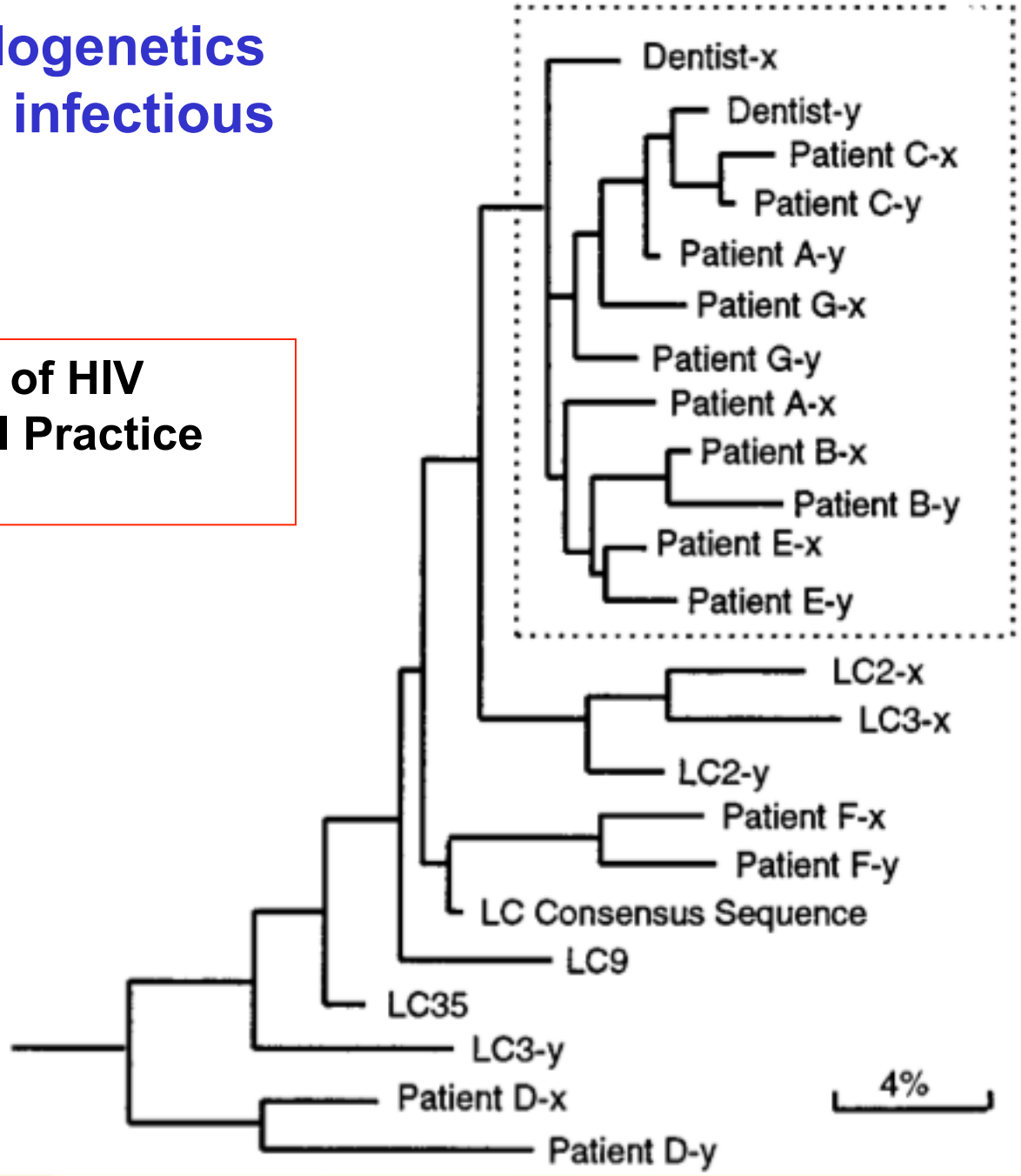
Constructing organism phylogeny from specific genes

- The gene must be present in all organisms
- The gene cannot be subject to horizontal transfer
- The gene must display an appropriate level of sequence conservation for the divergences of interest, i.e. evolving not too fast and not too slow.
- The gene must be sufficiently large to carry a record of the historical information.

16s rRNA

Application of phylogenetics in epidemiology of infectious diseases

Molecular Epidemiology of HIV Transmission in a Dental Practice
Ou et al Science 1992



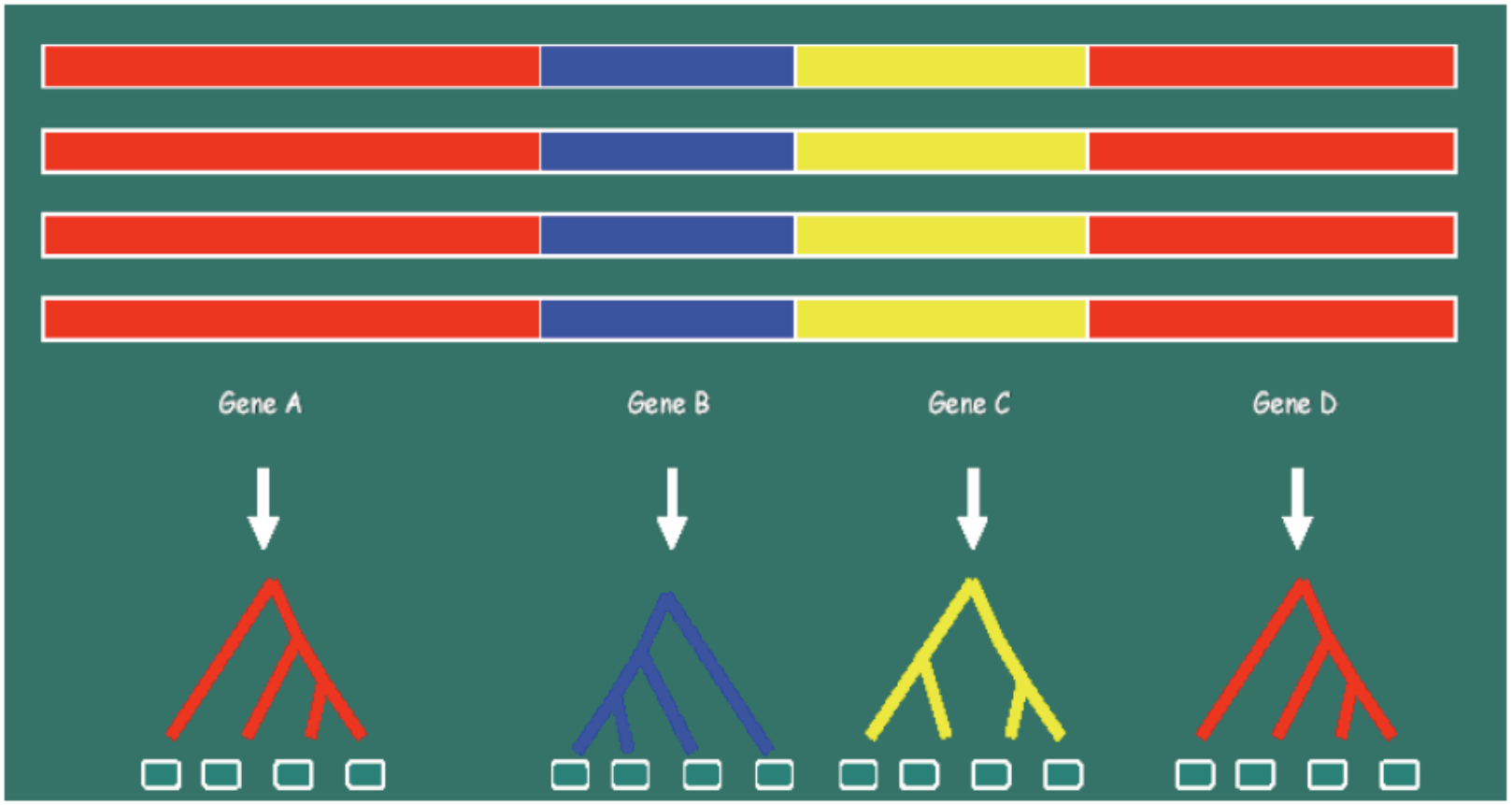
Phylogeny on the genomic scale: what to do with many genes ?

- Combined gene phylogenies
  - concatenated sequences, build a super gene
  - consensus trees: build individual genes from a set of genes and then look for consensus tree

- Gene order phylogeny: the spatial order of the genes on the chromosomes

- Gene content phylogeny: presence and absence of genes

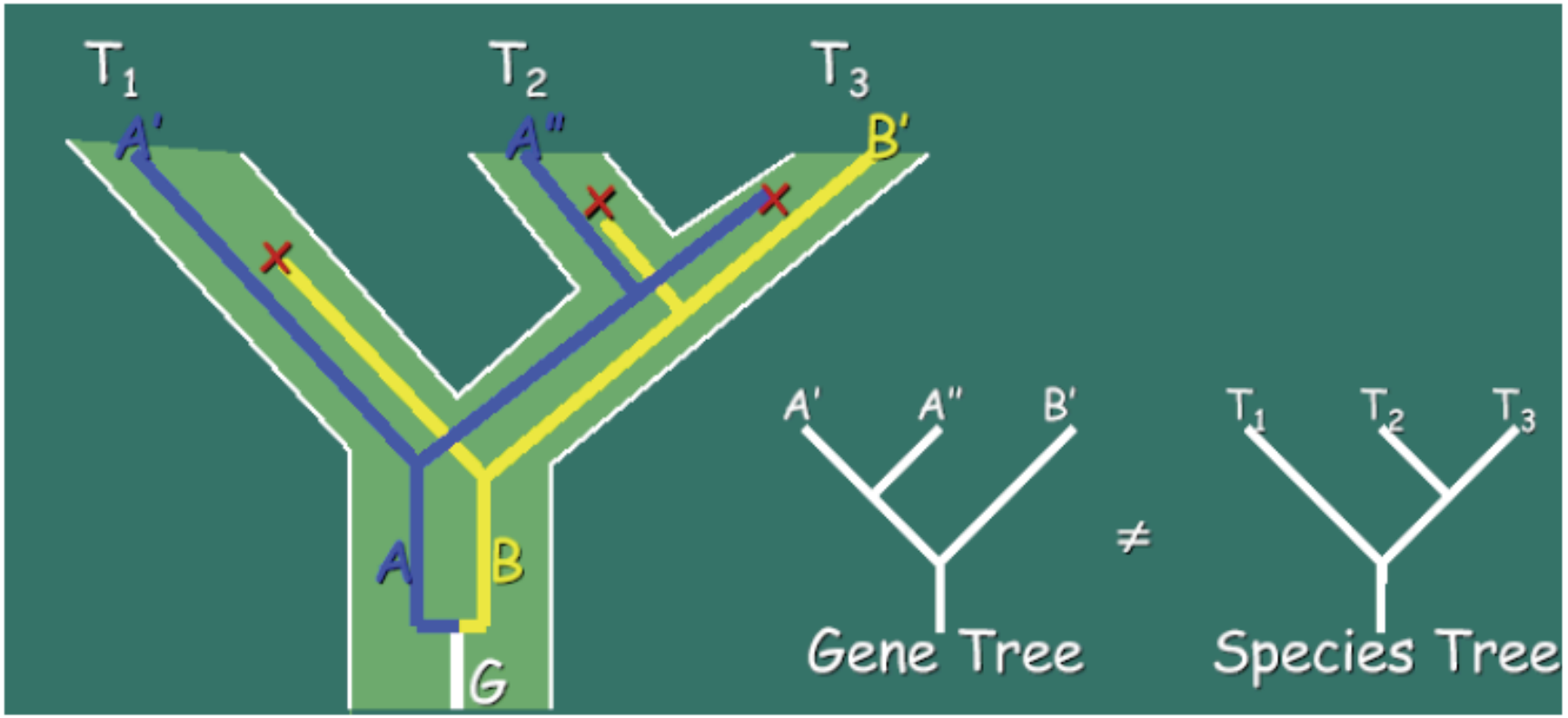
Concatenated Gene Trees



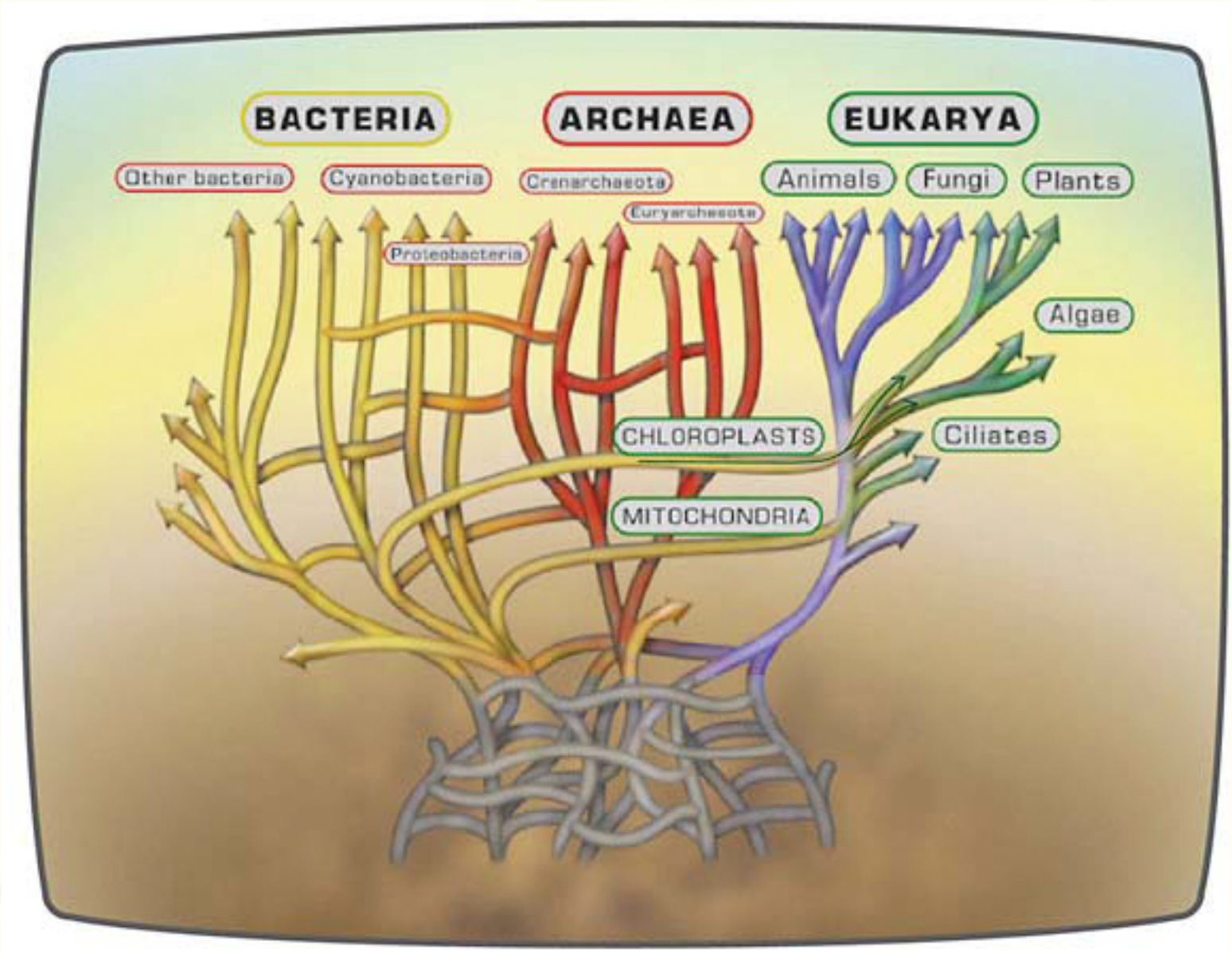
Potential problems: sensitive to ortholog assignment, horizontal gene transfer, sampling errors

Potential issue: Gene tree and species tree are not always consistent

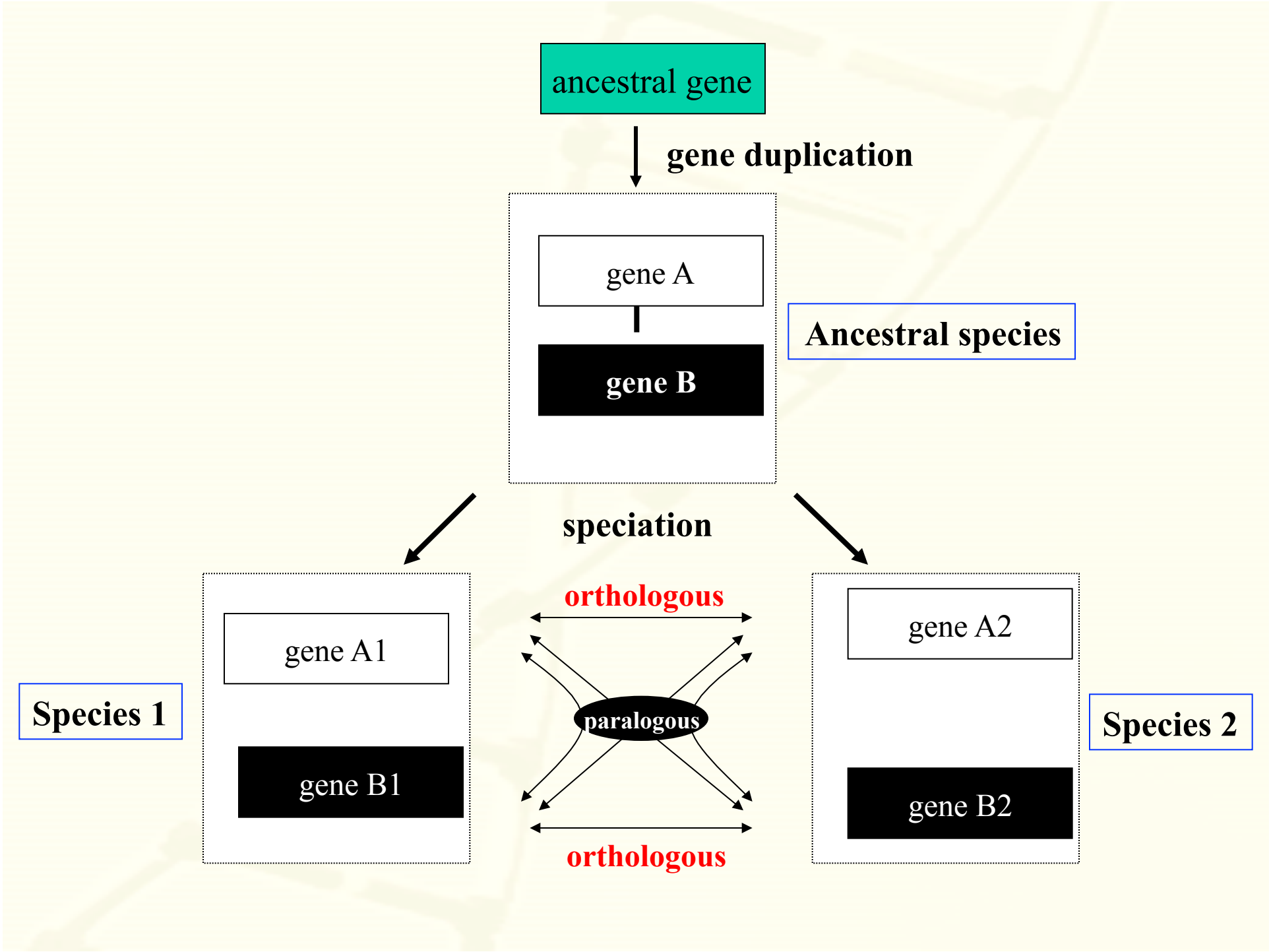
- Gene trees can differ from species tree because of mutation, selection, recombination etc.



Potential issue: Horizontal Gene Transfer



Homologs, orthologs, and paralogs

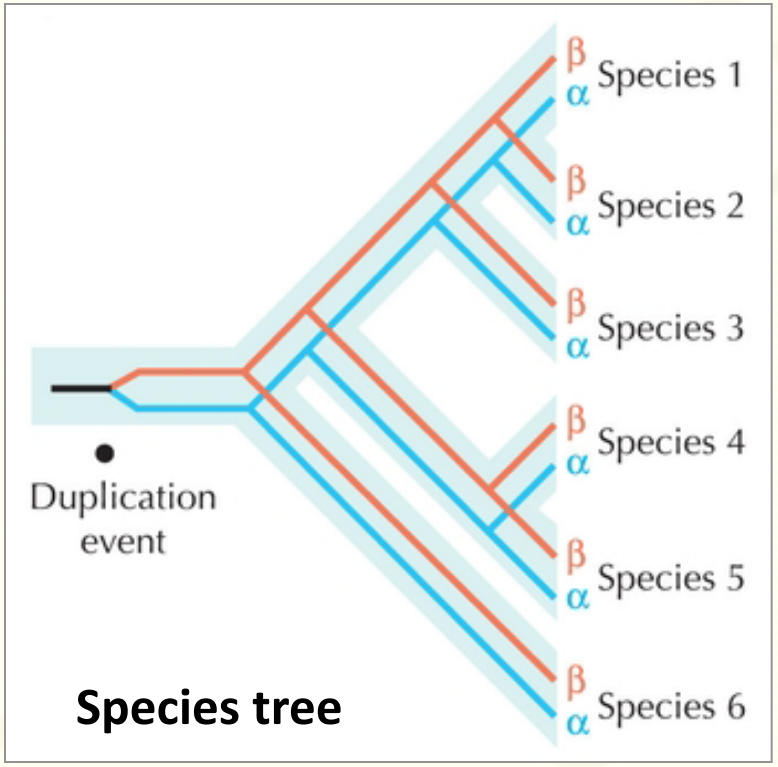

ancestral gene
gene duplication
gene A
gene B
Ancestral species
speciation
orthologous
paralogous
orthologous
gene A1
gene B1
Species 1
gene A2
gene B2
Species 2

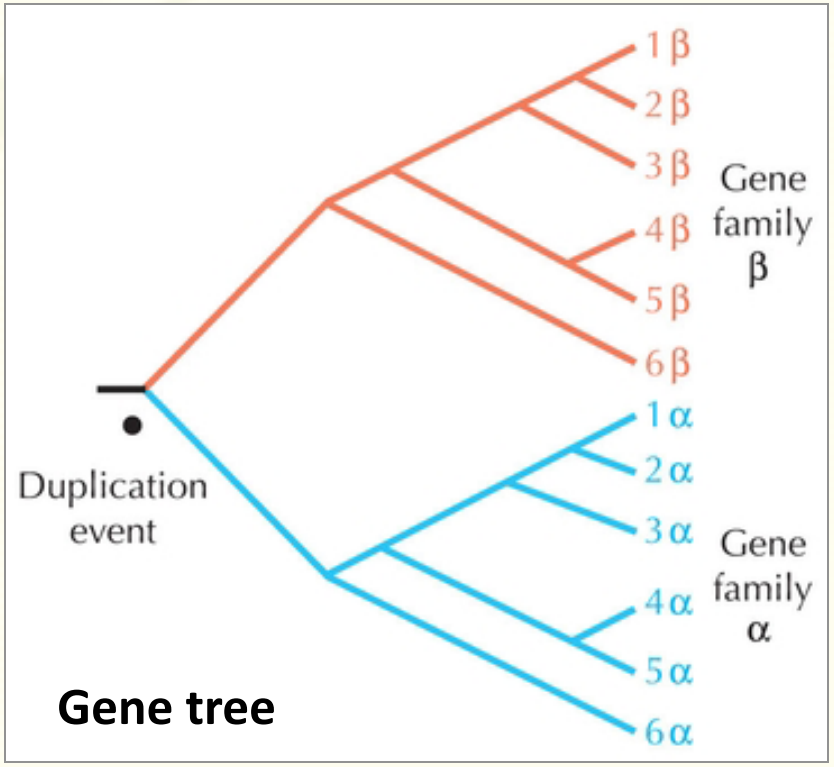
Homologs, orthologs, and paraogs

Homologs: Genes that are descended from a common ancestor.

Orthologs: Derived from a single ancestral gene in the last common ancestor of the species, arising due to speciation.

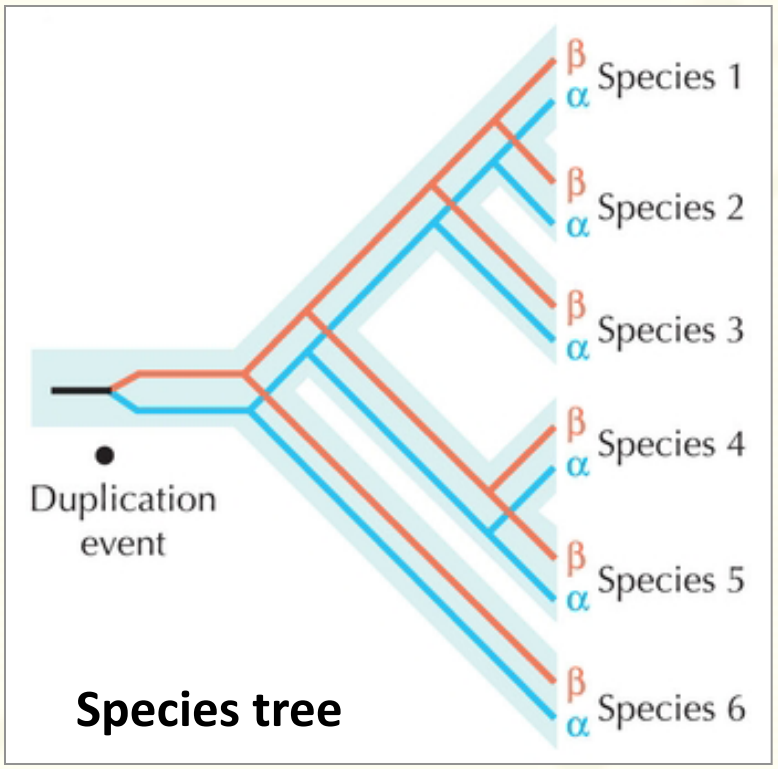
Paralogs: Homologous sequences that are separated by gene duplication within the ancestral species.

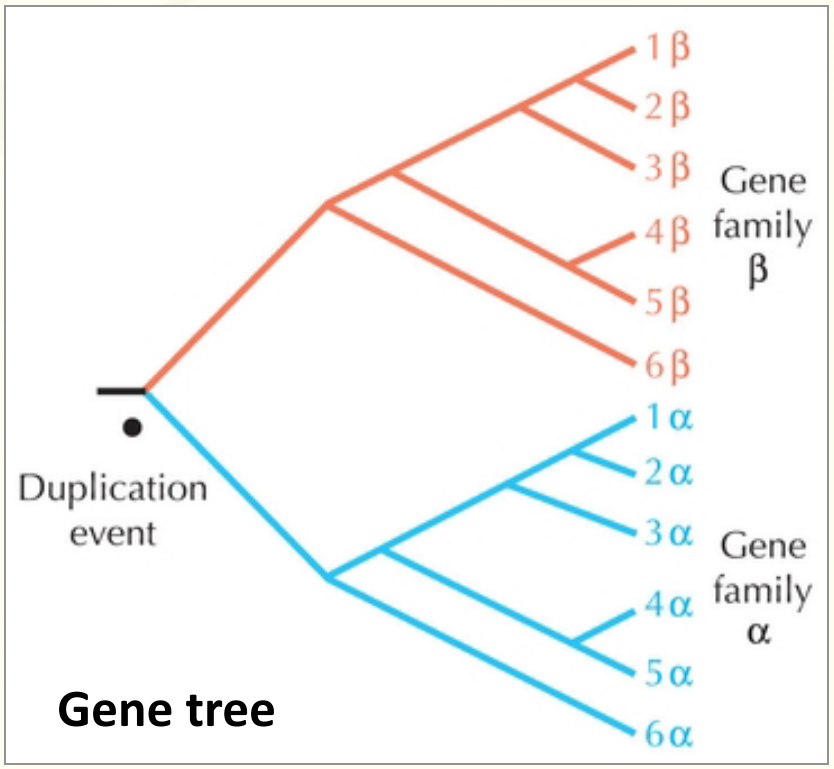




Inparalogs, outparalogs, ohnologs

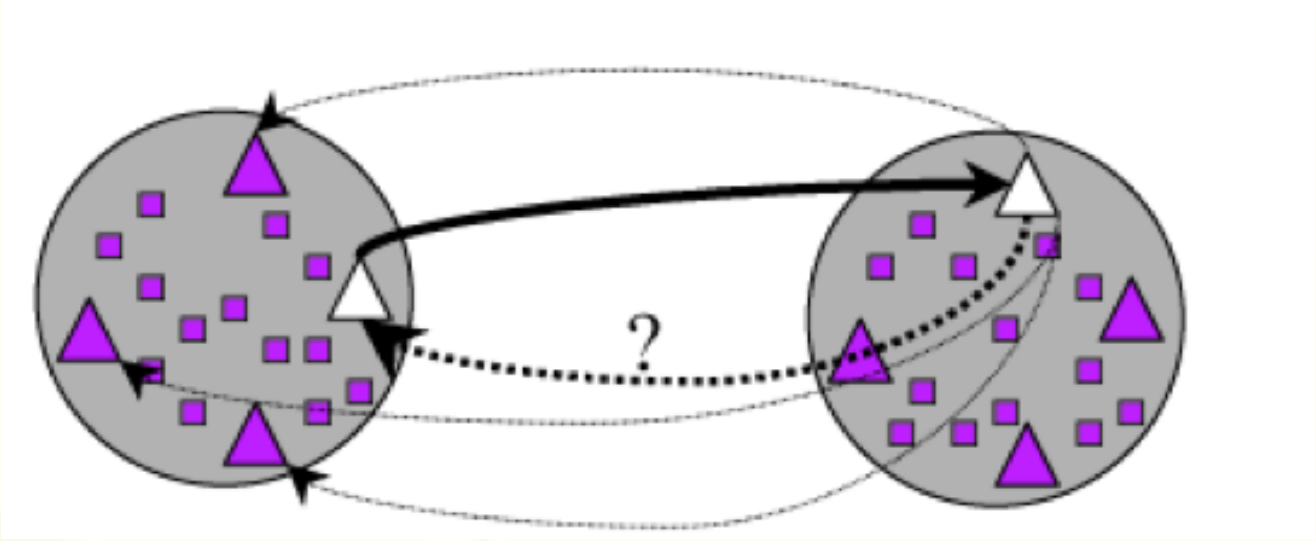
- Inparalogs (symparalogs): within species paralogs
- Outparalogs (alloparalogs): between species paralogs
- Ohnologs: paralogs resulted from whole genome duplication





Finding orthologs:
Best Bi-directional BLAST hit (BBH)

- BLAST gene A in genome 1 against genome 2: gene B is best hit
- BLAST gene B against genome 1: if gene A is best hit A and B are orthologous
- Similar but more rigorous methods: Inparanoid, OrthoMCL



Genome 1 **Genome 2**

Finding orthologs: other methods

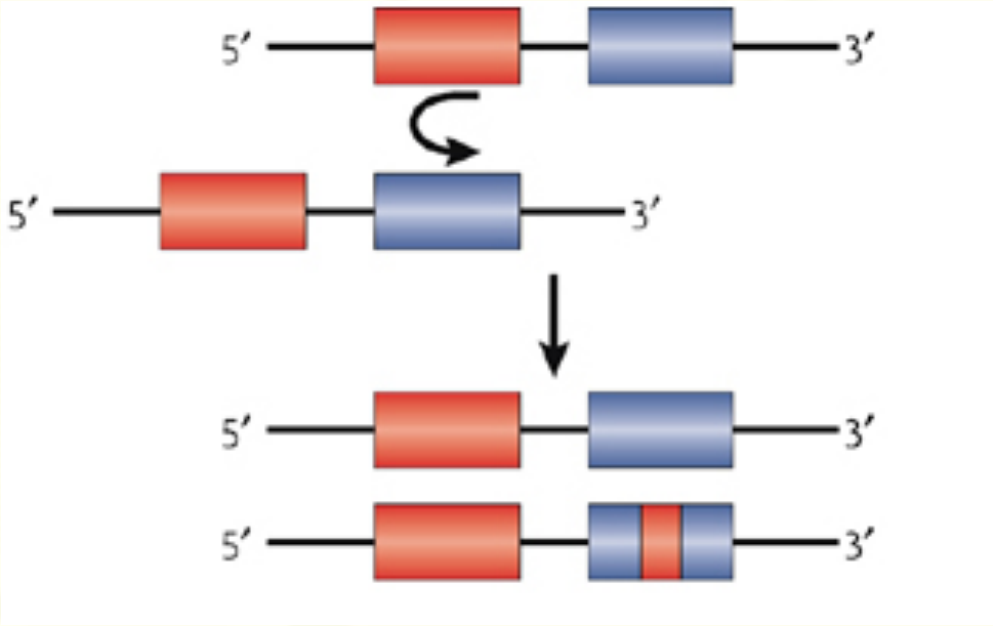
- By phylogenic analysis

- By genomic synteny or gene order, i.e. the orthologs occupy the same genomic region in different species

Yeast Gene Order Browser, Wolfe Lab, http://wolfe.gen.tcd.ie/ygob/

Gene conversion can confuse ortholog assignment

- Gene Conversion: The transfer of DNA sequences between two homologous genes, most often by unequal crossing over during meiosis



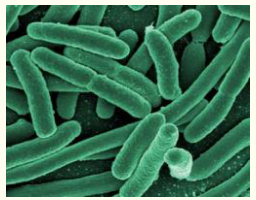
Experimental Evolution

- **Experimental evolution**: testing evolutionary theory using microorganism grown in designed and controlled conditions conditions in the laboratory.
- This allows direct study of the forces shaping the evolution of genes and genomes including mutation, recombination, selection, genetic drift, and gene flow.
- It also allows to control the mutation rate, population size, environmental structure, strength of selection, the opportunity for genetic exchange…
- The genomic sequence, gene expression level, fitness and phenotypes can be quickly measured by high-throughput genomics technique such as next-gen sequencing.

Animal domestication (e.g. dogs, cattle) can be considered as experimental evolution too.

E. coli long-term evolution experiment

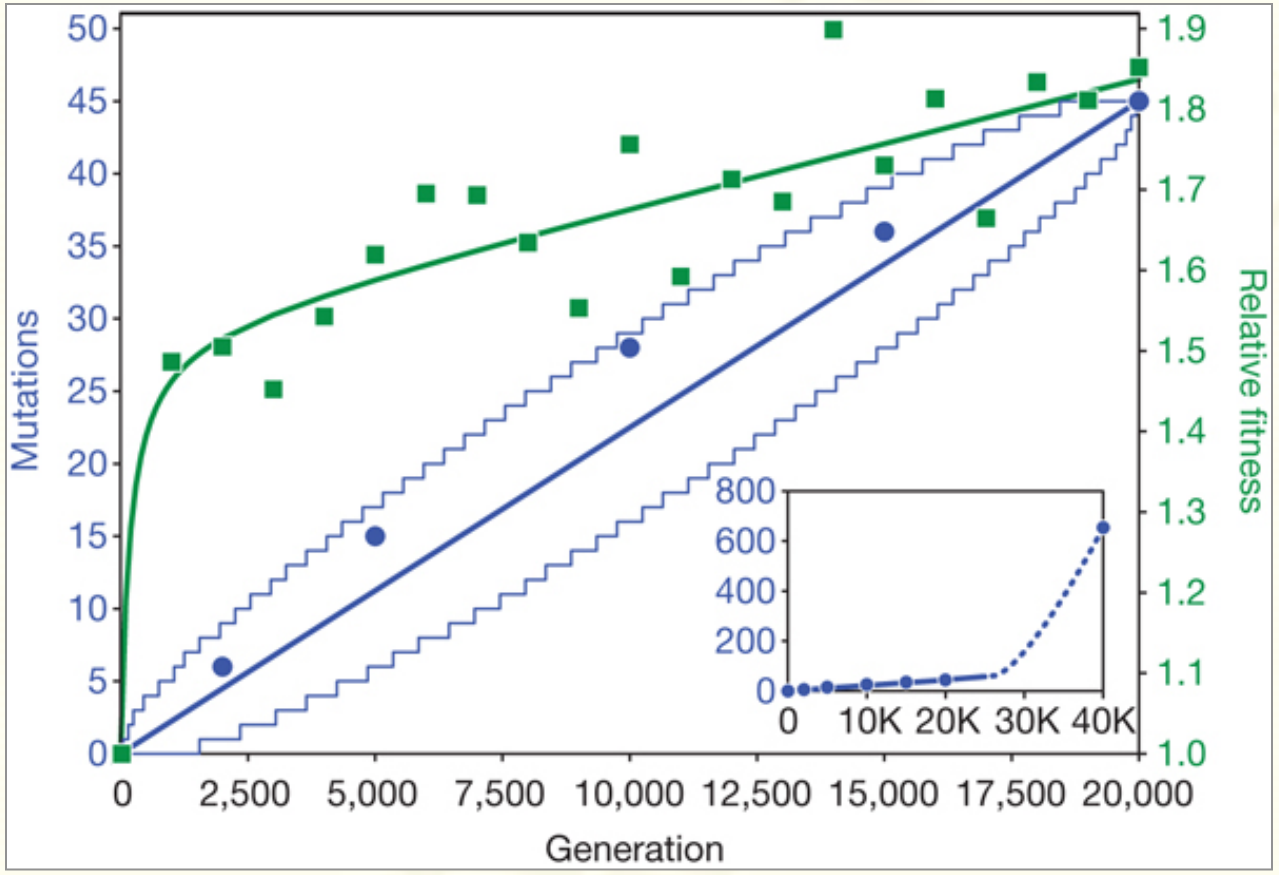
- Richard Lenski at Michigan State
- 24/02/1988: initial 12 nearly identical asexual strains are grown in minimum media
- Every day,1% of each population from each flask is transferred to a flask of fresh growth medium and let grow.
- Every 75 days (500 generations), representative samples of each population are frozen for future studies.
- Until Feb 2010, 50,000 generations



Genome evolution and adaptation in a long-term experiment with *Escherichia coli*

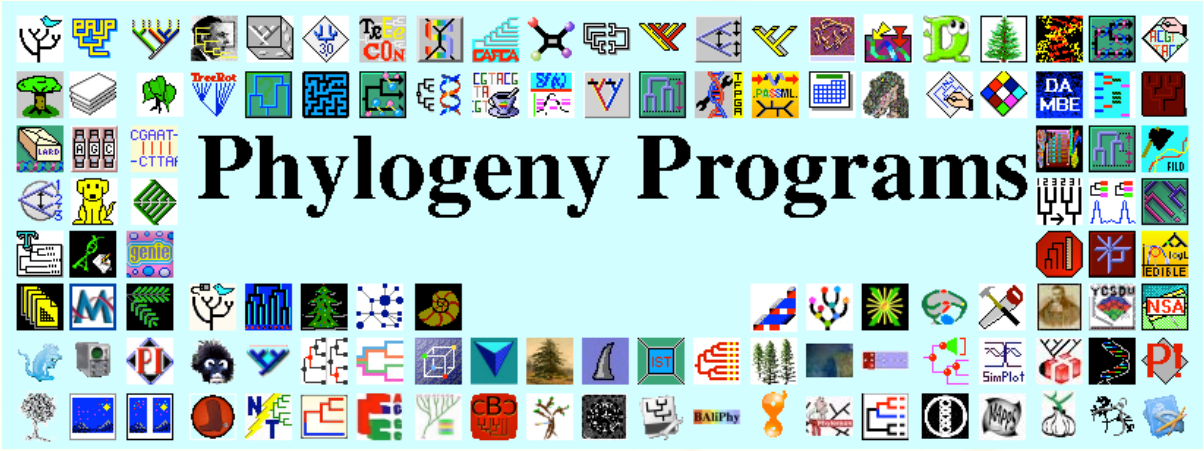
Jeffrey E. Barrick[1]*, Dong Su Yu[2,3]*, Sung Ho Yoon[2], Haeyoung Jeong[2], Tae Kwang Oh[2,4], Dominique Schneider[5], Richard E. Lenski[1] & Jihyun F. Kim[2,6]

Barrick et al Nature 2008



- Mutations accumulated at a near-constant rate even as fitness gains decelerated over the first 20,000 generations.
- Almost all mutations are beneficial mutations.
- After 20,000 generations, mutations are mostly neutral.

Molecular Evolution Software



366 phylogeny software on Joe Felsenstein's website
http://evolution.genetics.washington.edu/phylip/software.html

PHYLIP (PHYLogeny Inference Package)

PAML: Phylogenetic Analysis by Maximum Likelihood (Ziheng Yang)

MEGA: Molecular Evolutionary Genetics Analysis

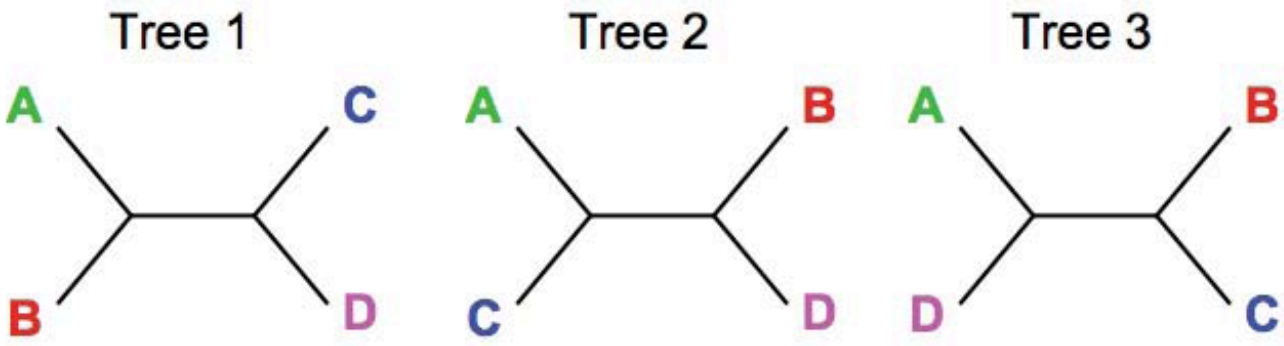
MEGA
Molecular Evolutionary Genetics Analysis

End of lecture

Questions ?

Unrooted Trees

There are three possible unrooted trees for four taxa (A, B, C, D)



Phylogenetic tree building (or inference) methods are aimed at discovering which of the possible unrooted trees is "correct". We would like this to be the "true" biological tree — that is, one that accurately represents the evolutionary history of the taxa. However, we must settle for discovering the *computationally correct* or *optimal* tree for the phylogenetic method of choice.

Jukes & Cantor 1969

	A	C	G	T
A	X	α	α	α
C	α	X	α	α
G	α	α	X	α
T	α	α	α	X

1 parameter
equiprobable changes

Kimura 1980

	A	C	G	T
A	X	α	$\kappa.\alpha$	α
C	α	X	α	$\kappa.\alpha$
G	$\kappa.\alpha$	α	X	α
T	α	$\kappa.\alpha$	α	X

2 parameters
transition rate ≠
transversion rate

Tamura 1992

	A	C	G	T
A	X	$\alpha\frac{1-\theta}{2}$	$\kappa\alpha\frac{1-\theta}{2}$	$\alpha\frac{1-\theta}{2}$
C	$\alpha\frac{\theta}{2}$	X	$\alpha\frac{\theta}{2}$	$\kappa\alpha\frac{\theta}{2}$
G	$\kappa\alpha\frac{\theta}{2}$	$\alpha\frac{\theta}{2}$	X	$\alpha\frac{\theta}{2}$
T	$\alpha\frac{1-\theta}{2}$	$\kappa\alpha\frac{1-\theta}{2}$	$\alpha\frac{1-\theta}{2}$	X

3 parameters
stationary GC% = $\theta \neq 50\%$