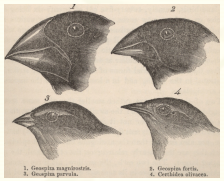
Topic Course: Gene and Protein Evolution



Instructors: Hue Sun Chan, Zhaolei Zhang

Wednesdays 3-5pm

MSB4174

Course outline

1. **Molecular Evolution and Phylogenetics** (02/24)
2. **Use Population Genetics to Detect Positive Selection** (03/02)
3. **Evolution of gene expression and gene regulation**. (03/09)
4. **Evolution of protein structure, interaction, and network**. (03/16)
5. **Synergy between the studies of protein biophysics and protein evolution**. (03/23)
6. **Theory of protein sequence space organization and the dynamics of molecular evolution** (03/30.

Course evaluation

- Attending each lecture on time. (10%)
- Paper presentation. (25%)
- Class participation. (15%)
- Final project – mock grant LOI (50%)

- **Paper Presentation:**
  - 2-3 papers to discuss each week.
  - 10 minutes presentation + 5 minutes discussion

- **Grading criteria**:
  - Understanding of the assigned paper
  - General background knowledge
  - Presentation clarity and skill
  - Ability to answer questions .

Course evaluation

- Attending each lecture on time. (10%)
- Paper presentation. (25%)
- Class participation. (15%)
- Final project – mock grant LOI (50%)

- **Final Project:**
  - Topic: relevant to gene or genome evolution, uses the techniques covered in the course, and has some computational aspect.
  - Has minimal overlap with your own thesis project.
  - Check with the instructor if you are not sure whether the project is appropriate.
  - CIHR style Letter of Intent (LOI) for a 3-year research project .
  - Five pages, single spaced: Abstract, Background & Significance, Experimental Plan, Figures.
  - Budget, References.

Week 1: Molecular Evolution and Phylogenetics

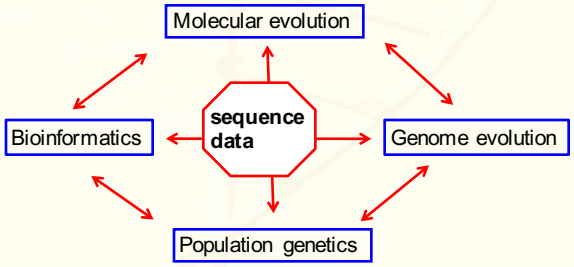
- Introduction and historical background
- Mutations and substitutions
  - Positive, negative, neutral selection, synonymous and nonsynonymous substitutions
- Codon bias
- Neutral theory of evolution
- Phylogenetic trees

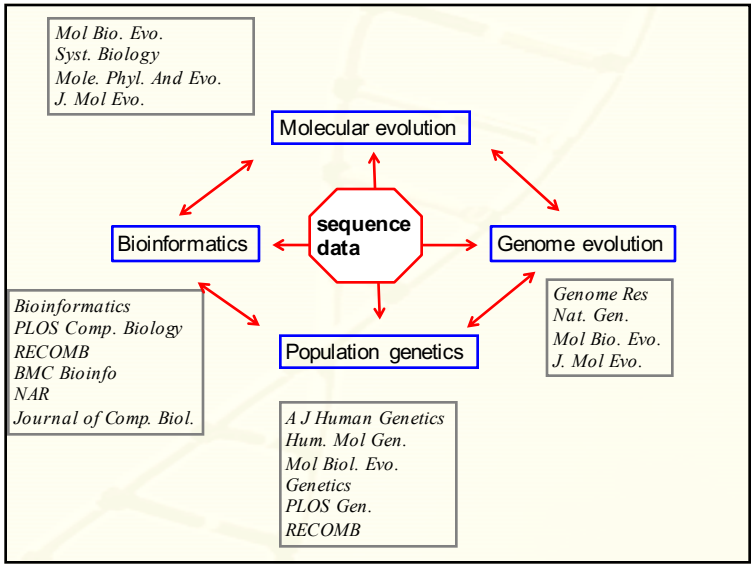
What is Molecular Evolution ?

- 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 , e.g. Xist gene or ribsomes

What is Molecular Evolution ?

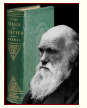
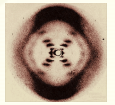
- How and when were a gene and protein created ? How "old" is a gene ? How can we calculate the "age" of a gene ?
- How did the gene evolve to the present form ? What selective forces (if any) influence the evolution of a gene sequence and expression ? Are these changes in sequence adaptive or neutral ?
- 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 function of this gene ?
- How do species evolve? How can evolution of a gene tell us about the evolutionary relationship of species ?





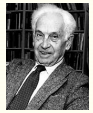
A brief historical perspective

- 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)

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)**"
- "**Nothing in Evolution Makes Sense Except in the Light of …**

Molecular Evolution meets Genome Revolution

2012

Mutations in DNA and protein

- **Synonymous mutations** -> do not change amino acid
- **Nonsynonymous mutations** -> change amino acid

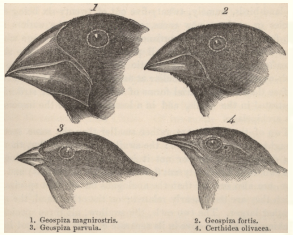
- **Nonsense mutation**: resulting in a pre-mature stop codon
- **Missense mutation**: resulting in a different amino acid
- **Frameshift mutation**: insertion / deletion of 1 or 2 nucleotides
- **Silent** mutation: the same as nonsynonymous mutation

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

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 (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

Ka/Ks - 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
```

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

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

Ka/Ks - 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
```

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

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

Ka/Ks - 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
```

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

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

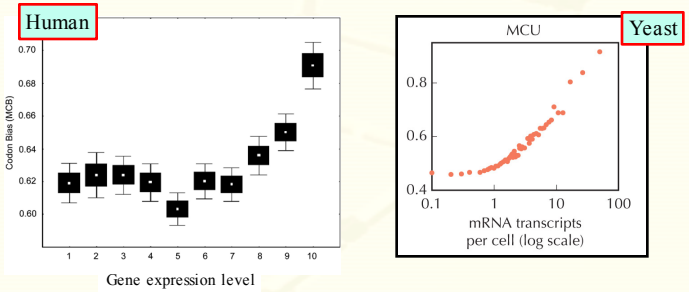
Synonymous substitutions are NOT always neutral

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 TFs, microRNA and RNA binding proteins (RBP)

- More discussions on this at next week.

Highly expressed genes tend to use optimal codons



$$\mathrm{CAI} = \exp\left(1/L \sum_{l=1}^{L} \log\left(w_i(l)\right)\right)$$

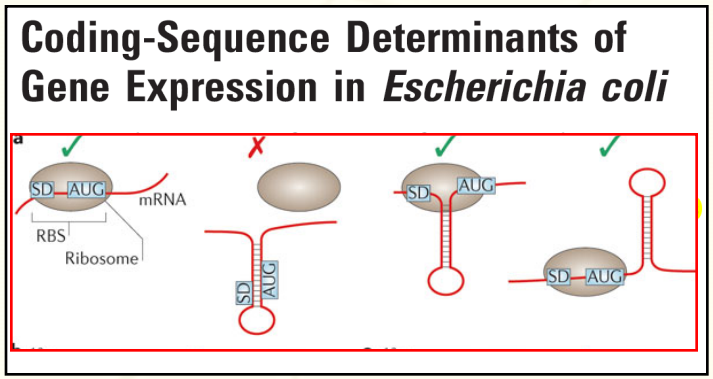
$$w_i = \frac{f_i}{\max(f_j)}$$

$i, j \in$ [synonymous codons for amino acid]

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

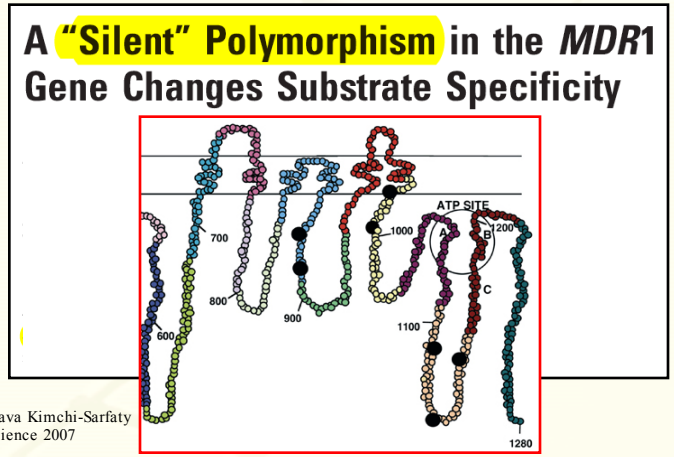
Urrutia and Hurst 2006
Akashi 2004

Synonymous codons influence mRNA secondary structure and gene expression



Kudla .. Plotkn Science 2009

"Rare codons" can influence protein structure



Chava Kimchi-Sarfaty
Science 2007

Methods to detect positive selection

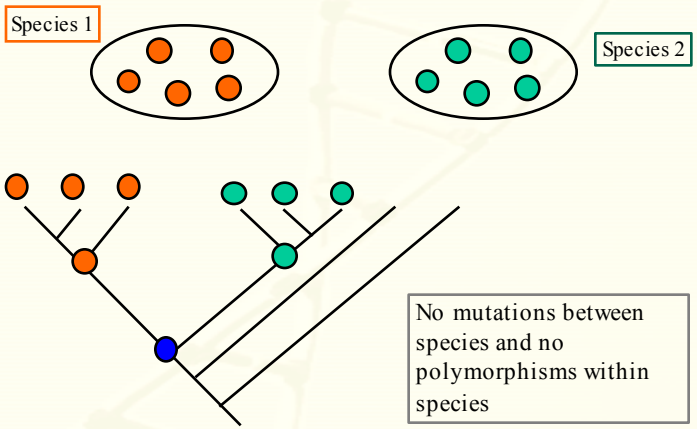
- **Ka / Ks test:** suitable for between species
- **McDonald-Kreitman (MK) test**
  - Compare between species and within species
- **Fixation index** (Fst)
  - Testing difference in allele frequency between populations
- **Linkage disequilibrium (LD)**
  - Look for nonrandom association of alleles at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

All these methods take neutrality as the null hypothesis

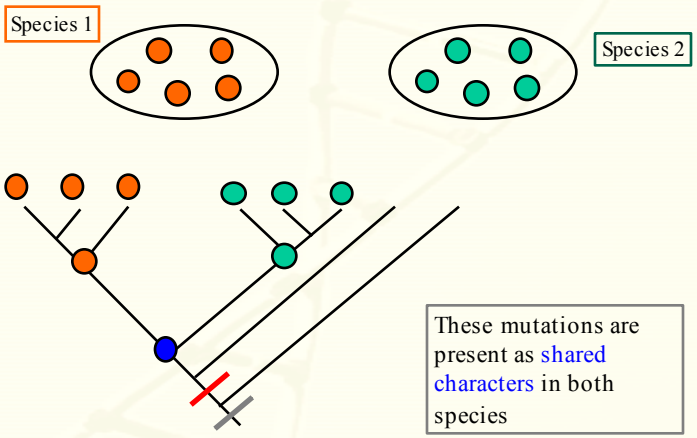
McDonald-Kreitman (MK) test

- McDonald-Kreitman (MK) Test compares divergence between two species with polymorphism within each species.
- Rationale: If a gene evolves neutrally, i.e. the DNA substitutions follow random drift, then the polymorphism within each species should follow similar pattern as divergence between species.
- This predicts similar ratio of synonymous and non-synonymous substitutions between and within species.

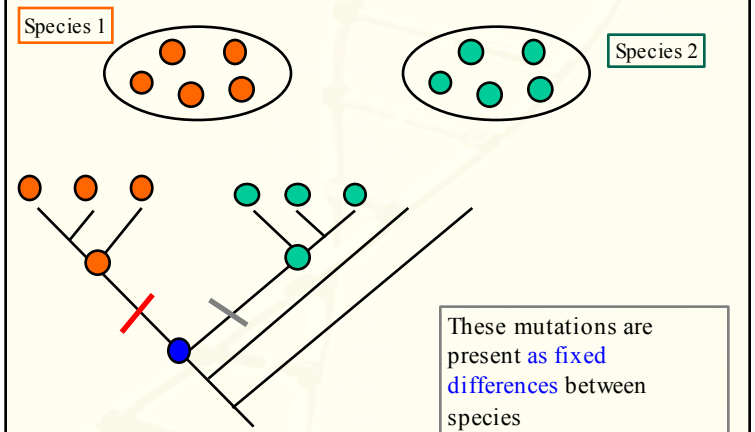
McDonald-Kreitman (MK) test



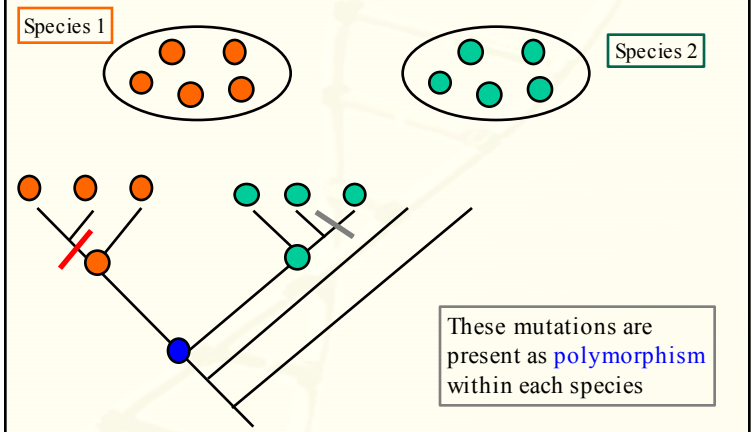
McDonald-Kreitman (MK) test



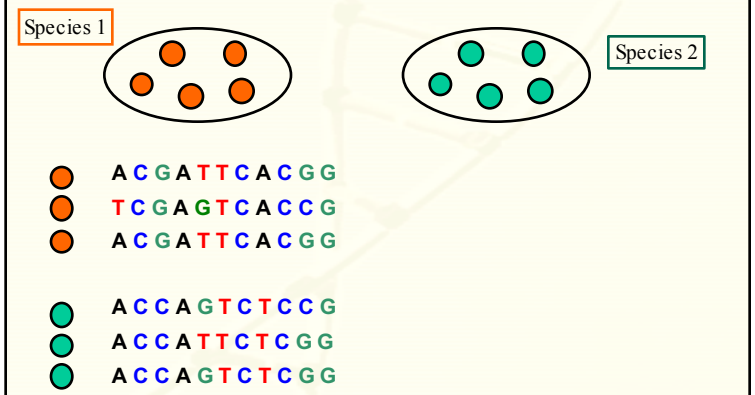
McDonald-Kreitman (MK) test



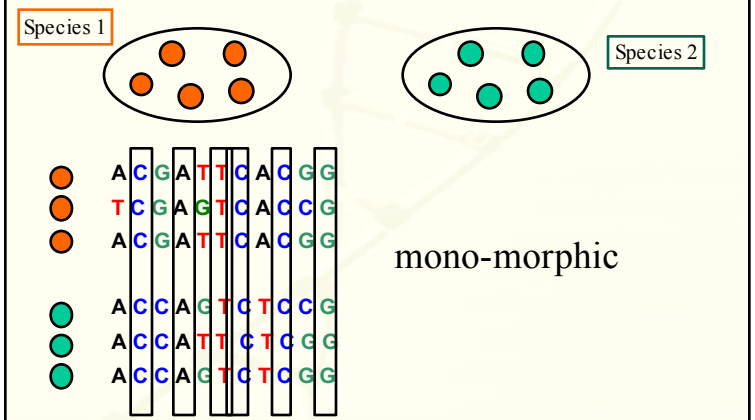
McDonald-Kreitman (MK) test



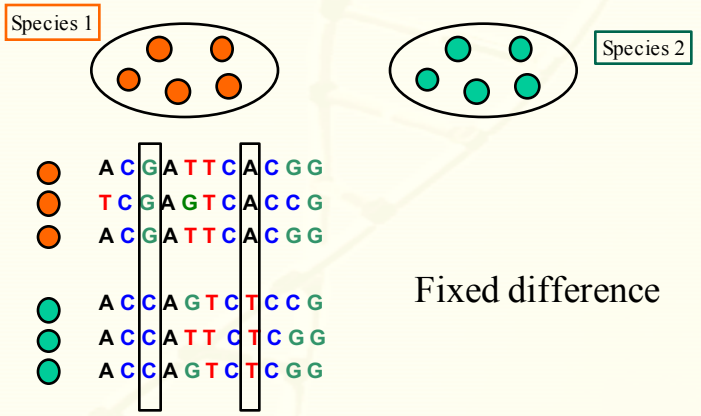
McDonald-Kreitman (MK) test



McDonald-Kreitman (MK) test

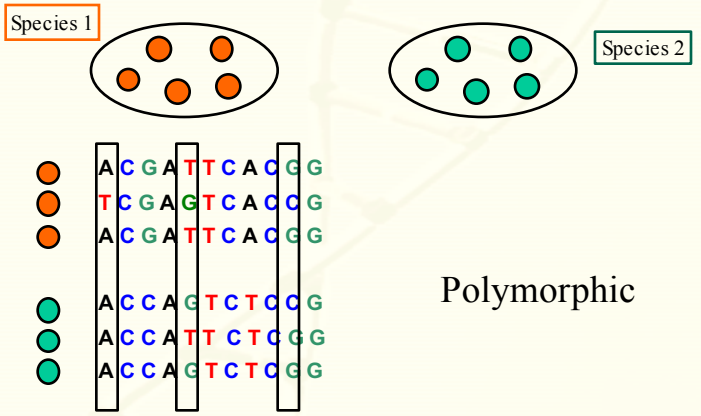


McDonald-Kreitman (MK) test



Fixed difference

McDonald-Kreitman (MK) test



Polymorphic

McDonald-Kreitman (MK) test

	Fixed difference	Polymorphism
Synonymous	W	Y
Non-synonymous	X	Z

Under neutrality: W / X = Y / Z

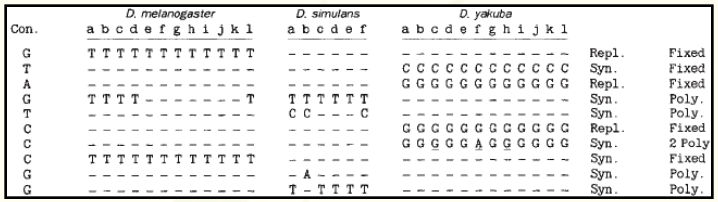
Statistically significant deviation from such null hypothesis can be tested by Chi-square test

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



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.

Smith, Eyre-Walker, Nature, 2002

MK test on real data

This is in contradictory to the neutral theory

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**

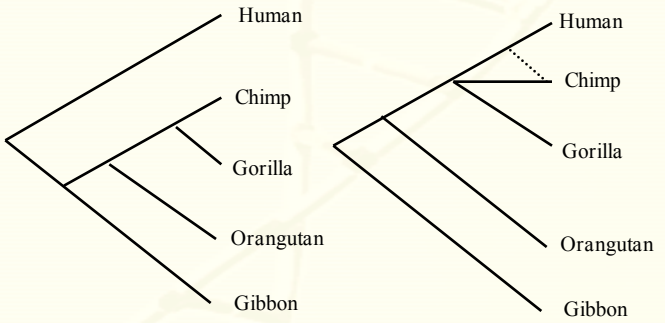
Diet and the evolution of human amylase gene copy number variation

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**

Phylogenetic analysis using DNA sequence

Phylogenetic analysis using DNA sequence



Traditional phylogeny

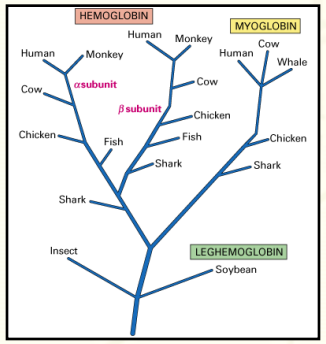
Revised phylogeny using DNA

Two types of questions in Phylogenetic analysis

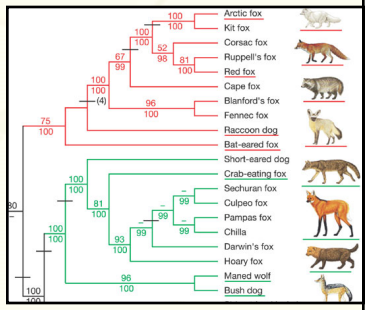
- Phylogenetic inference or "**tree building**":
  - To infer the branching orders and lengths between "taxa" (or genes, populations, species etc).
  - For example, can DNA tell us giant panda is more similar to bear or to dog, and when did they diverge ?
- **Character and rate** analysis:
  - Using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



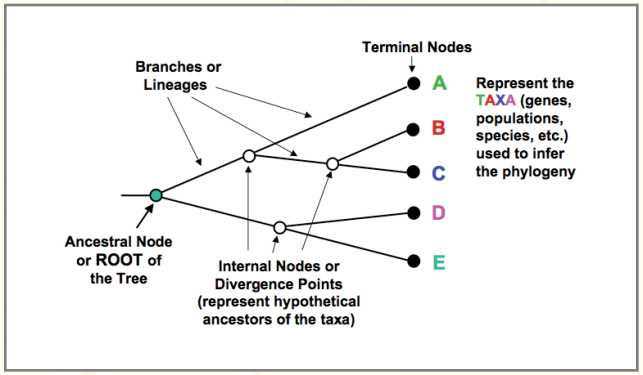
Species Tree



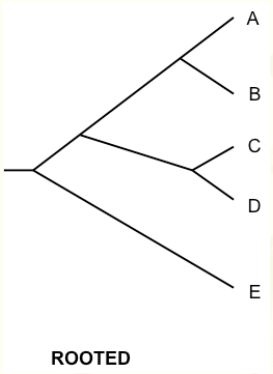
Lindblad-Toh Nature 2005

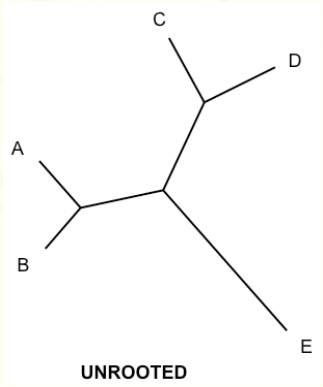
http://www.muhlenberg.edu/depts/biology/courses/bio152

Phylogenetic Tree Terminology

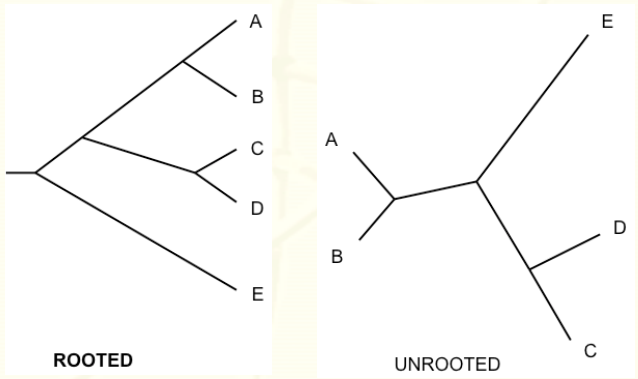


Rooted and unrooted trees

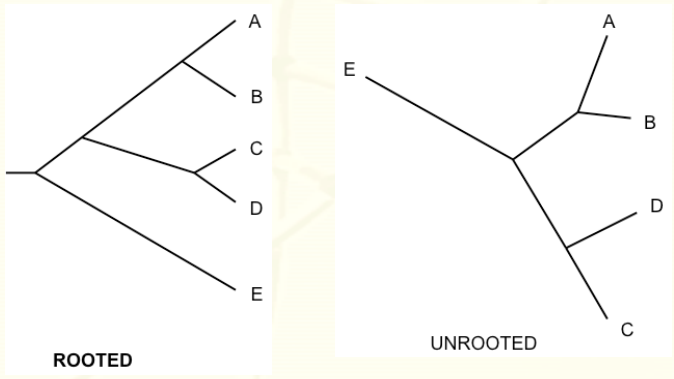


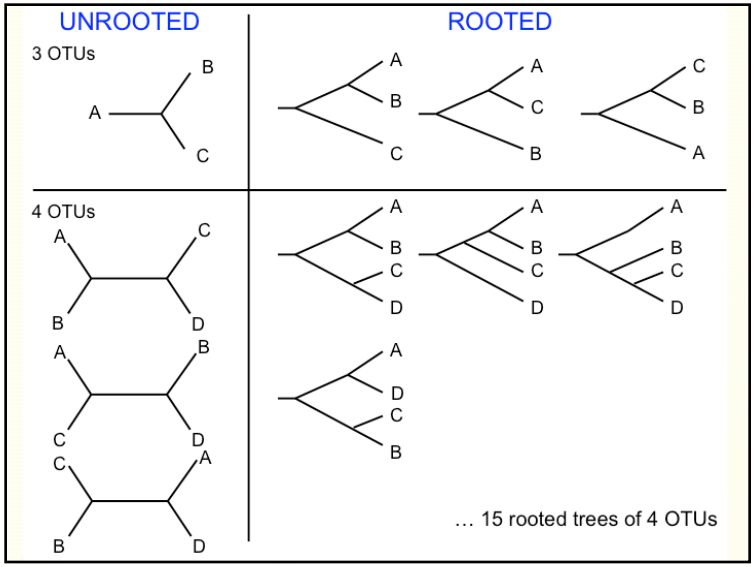


Rooted and unrooted trees

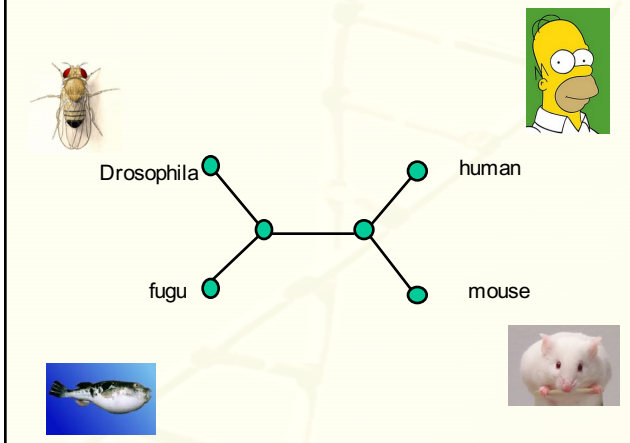


Rooted and unrooted trees





Root a tree using an outgroup



Root a tree using an outgroup

root

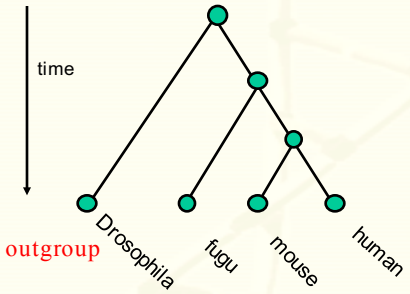
Drosophila
outgroup

human

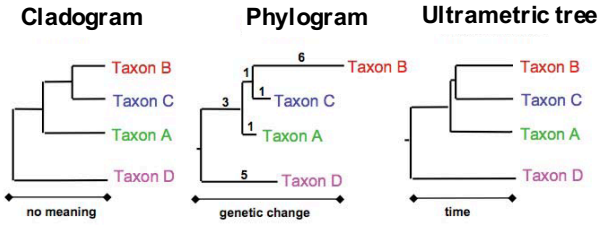
fugu

mouse

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

?

ACTGTTACCGA

ACTGTTACCGA

ACTGTTACCGA

ACTGTTACCGA

ACTGTTACCGA

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**

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

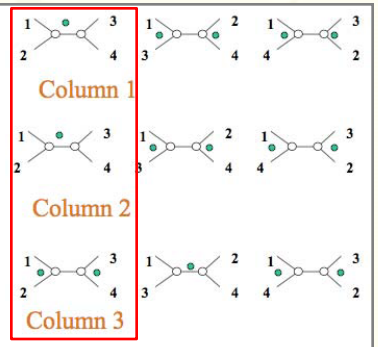
```
1 A A G A G T G C A
2 A G C C G T G C G
3 A G A T A T C C A
4 A G A G A T C C G
```

Three informative columns

- 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

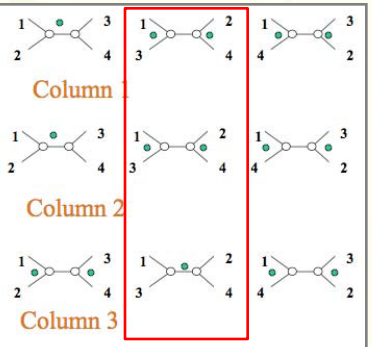


Tree 1: 4 substitutions

- Is a substitution

Maximum Parsimony Example

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

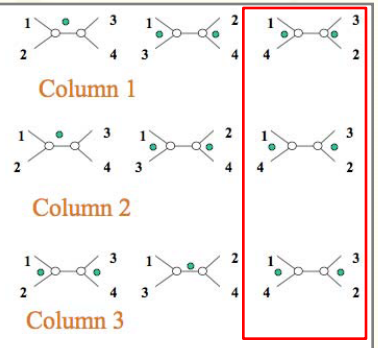


Tree 2: 5 substitutions

- Is a substitution

Maximum Parsimony Example

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



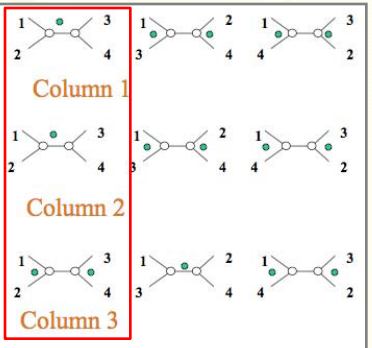
Tree 3: 6 substitutions

- Is a substitution

Maximum Parsimony Example

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

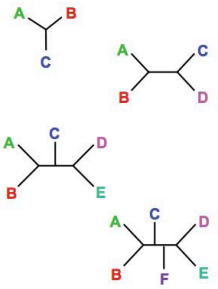
Tree 1: 4
Tree 2: 5
Tree 3: 6



- Is a substitution

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

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

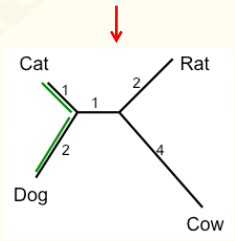
Distance based methods

- 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" ? E.g. 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



What distance metric to use ?

Cat	ATTTGCGGTA
Dog	ATCTGCGATA
Rat	ATTGCCGTTT
Cow	TTCGCTGTTT

?

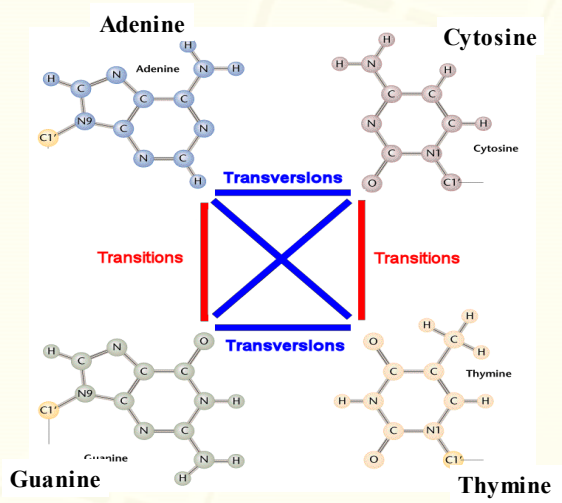
Number of different nucleotides

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

- Multiple substitutions at the same site: the observed differences do not always represent the actual evolutionary events that occurred
- 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.



Jukes & Cantor's one-parameter model

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

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*

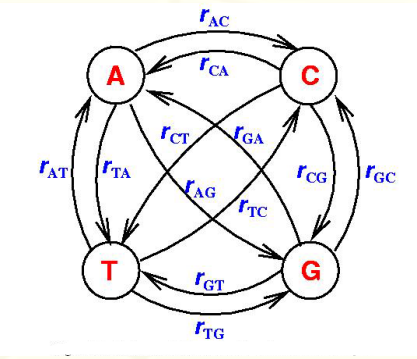
Hasegawa-Kishino-Yano (HKY85) 5-parameter model

	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).

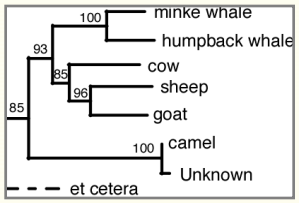
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**																			
S	-1	**4**																		
T	-1	1	**4**																	
P	-3	-1	1	**7**																
A	0	1	0	-1	**4**															
G	-3	0	-2	-2	0	**6**														
N	-3	1	0	-2	-2	0	**6**													
D	-3	0	-1	-1	-2	-1	1	**6**												
E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

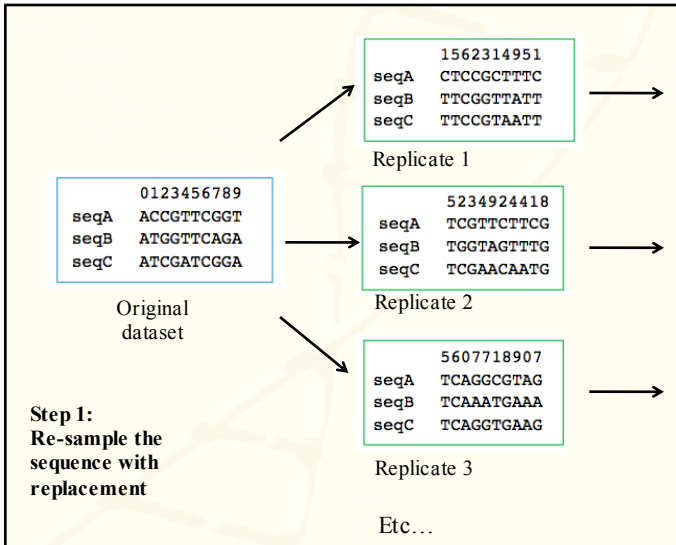
More advanced methods

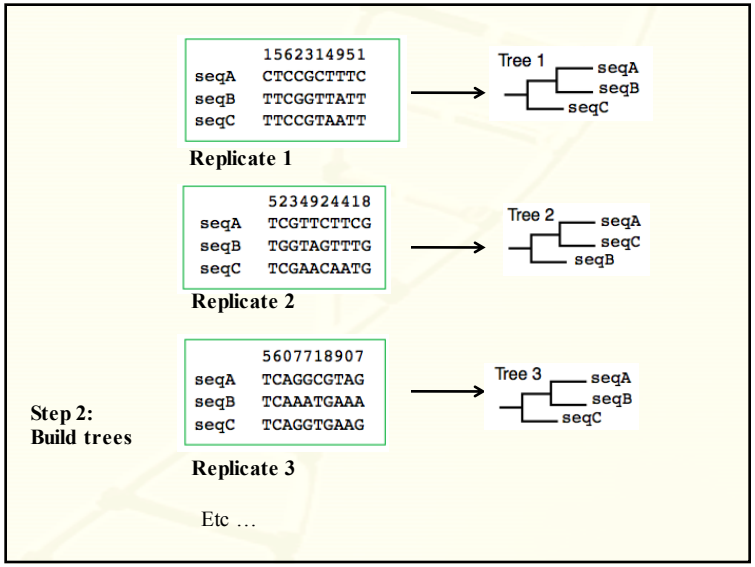
- **Maximum likelihood methods:**
  - ML methods evaluate phylogenetic hypothesis in terms of the **probability** that a proposed model and the parameters gave rise to the observed data. The tree found to have the highest likelihood is considered to be the optimal tree.
- **Bayesian Markov chain Monte Carlo methods**
  - Generate a large population of trees, then take a random walk through the "tree space" until a perfect tree is found.

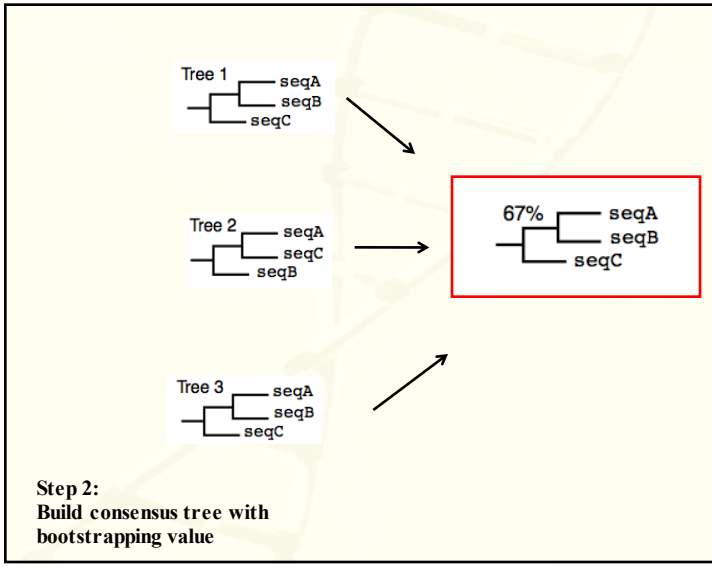
Bootstrapping

- 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.



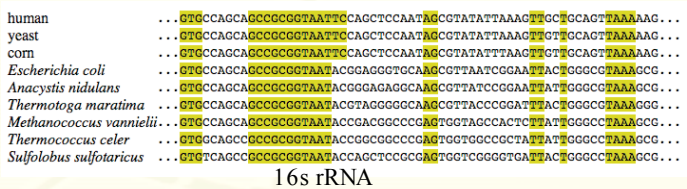






Constructing organism phylogeny from specific genes

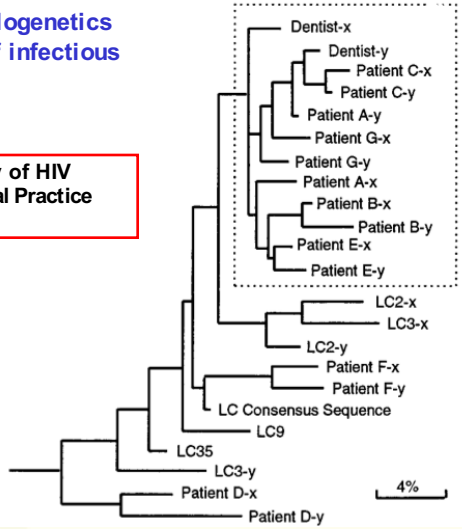
- The gene must be present in all organisms
- The gene cannot be subject to horizontal transfer
- The gene must display an appropriate level of sequence conservation for the divergences of interest, i.e. evolving not too fast and not too slow.
- The gene must be sufficiently large to carry a record of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



16s rRNA

Application of phylogenetics in epidemiology of infectious diseases

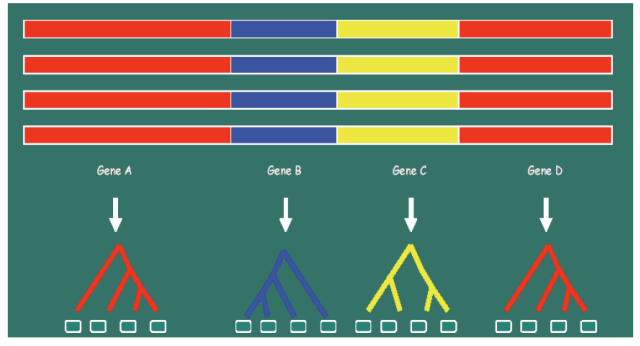
Molecular Epidemiology of HIV Transmission in a Dental Practice
Ou et al Science 1992



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

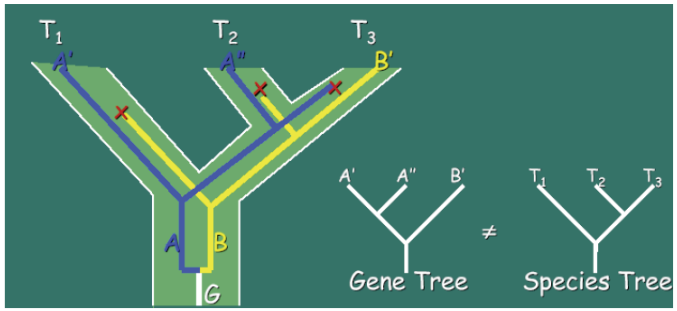
Concatenated Gene Trees



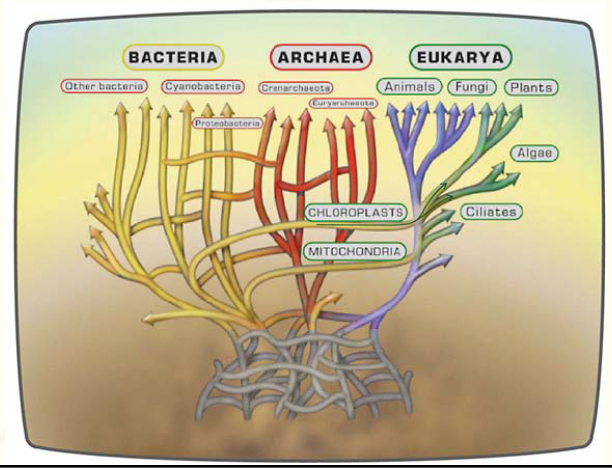
Potential problems: sensitive to ortholog assignment, horizontal gene transfer, sampling errors

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

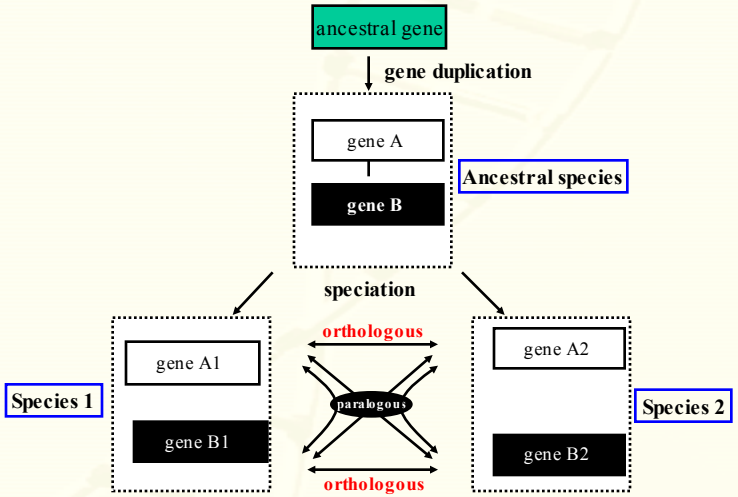
- Gene trees can differ from species tree because of mutation, selection, recombination etc.



Potential issue: Horizontal Gene Transfer



Homologs, orthologs, and paralogs

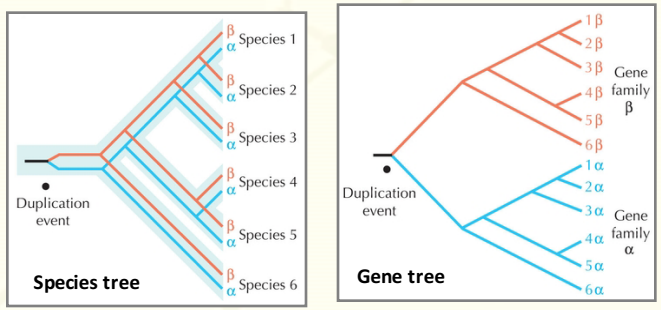


Homologs, orthologs, and paraogs

Homologs: Genes that are descended from a common ancestor.

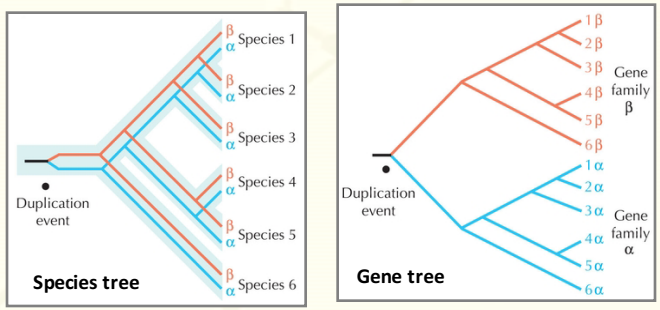
Orthologs: Derived from a single ancestral gene in the last common ancestor of the species, arising due to speciation.

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



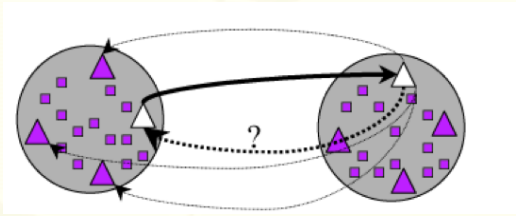
Inparalogs, outparalogs, ohnologs

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



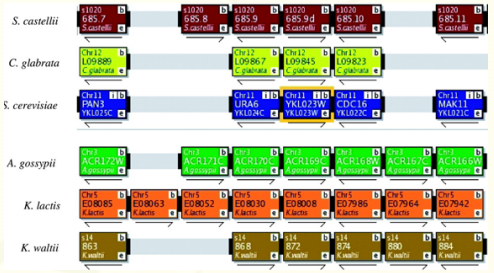
Finding orthologs: Best Bi-directional BLAST hit (BBH)

- 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



Finding orthologs: other methods

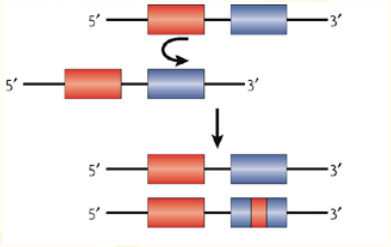
- By phylogenic analysis
- By genomic synteny or gene order, i.e. the orthologs occupy the same genomic region in different species



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



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