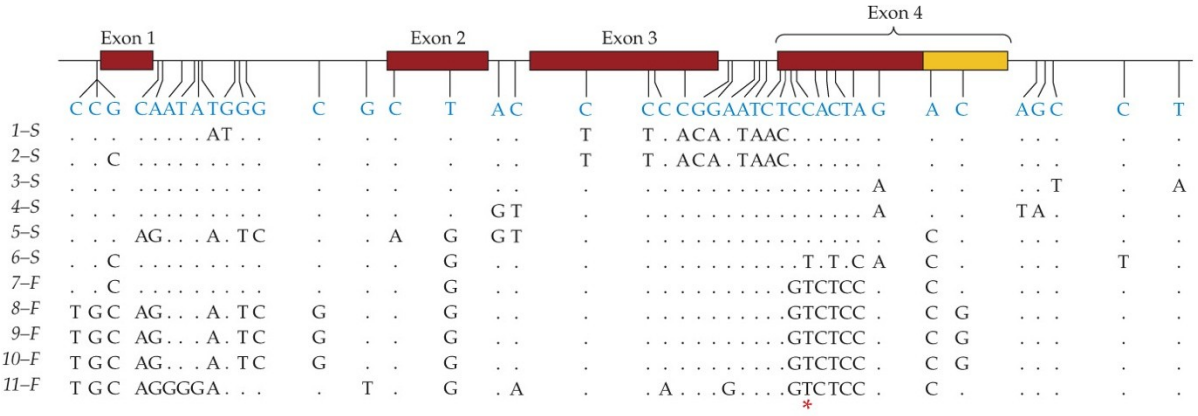
Molecular evolution



MOLECULAR AND GENOME EVOLUTION 1e, Figure 2.3

The Figure shows sequences of the ADH gene in 11 inbred Drosophila lines. Assume that the mutation rate is the same along the region. The first six are fast (S) haplotypes, the last five slow (S) haplotypes due to the variant marked by a red star which changes protein mobility and is the only non-synonymous variant

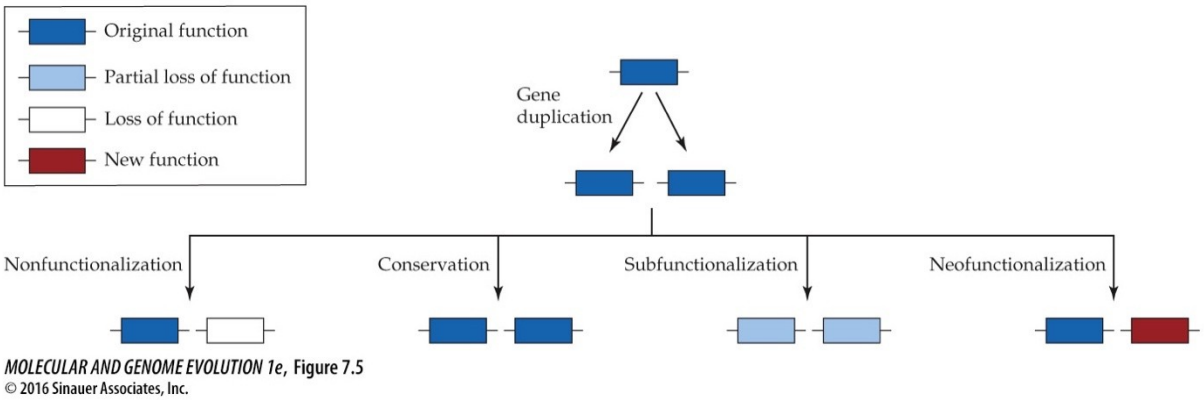
1. There is evidence of purifying selection against coding changes
2. The sites are all in linkage equilibrium
3. There is evidence of recombination between the fast and the slow haplotype
4. Variants outside exons are always without functional consequences

Estimators of mutation rate

The alignment of five sequences above have four segregating sites

1. Wattersons estimator of theta is 4
2. Tajimas estimator of theta is 2.2
3. Tajimas D is greater than zero
4. Population expansion is expected to give a negative Tajimas D

Gene duplication

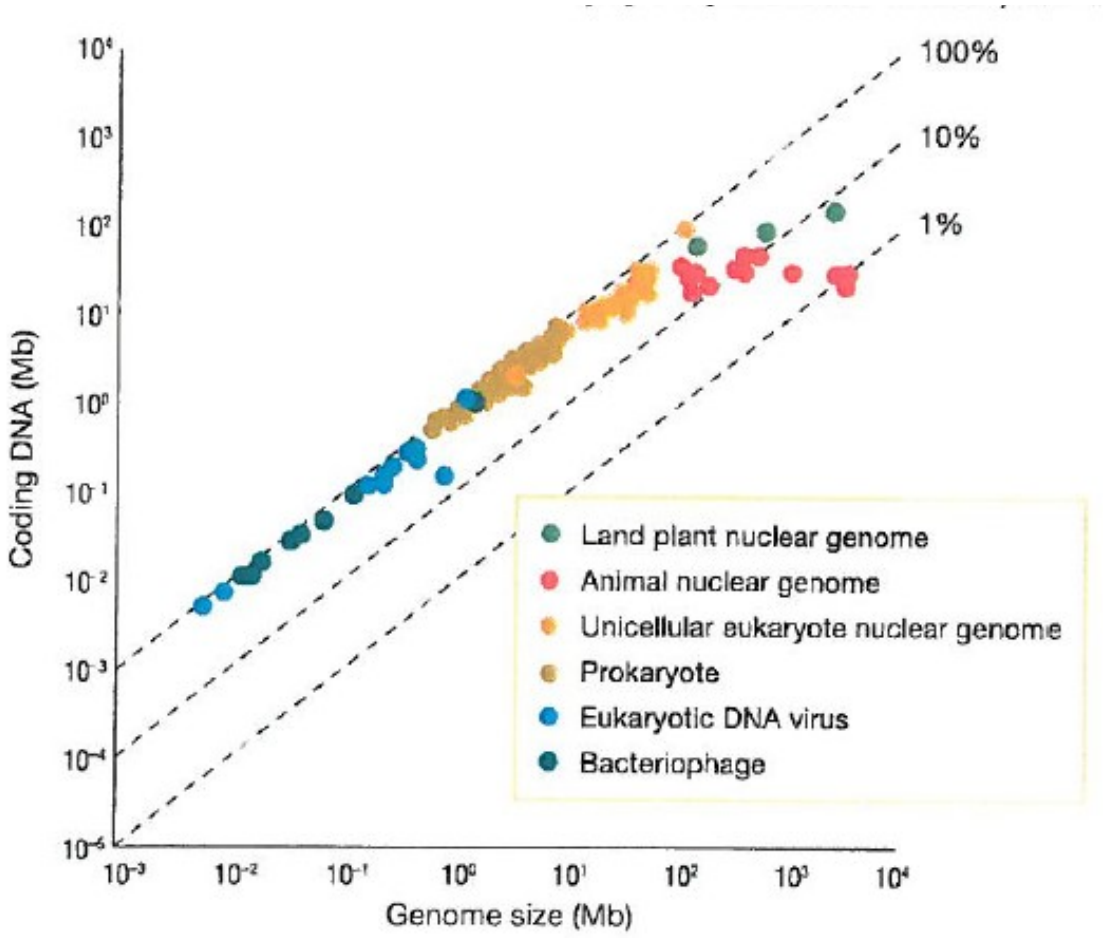


MOLECULAR AND GENOME EVOLUTION 1e, Figure 7.5

The figure shows possible fates of gene copies after a gene duplication event

1. Nonfunctionalisation is most likely to happen
2. Subfunctionalisation is likely to make it deleterious to lose one of the two copies
3. Gene duplication has been important for colour vision evolution
4. Gene duplication is unlikely to happen in vertebrates

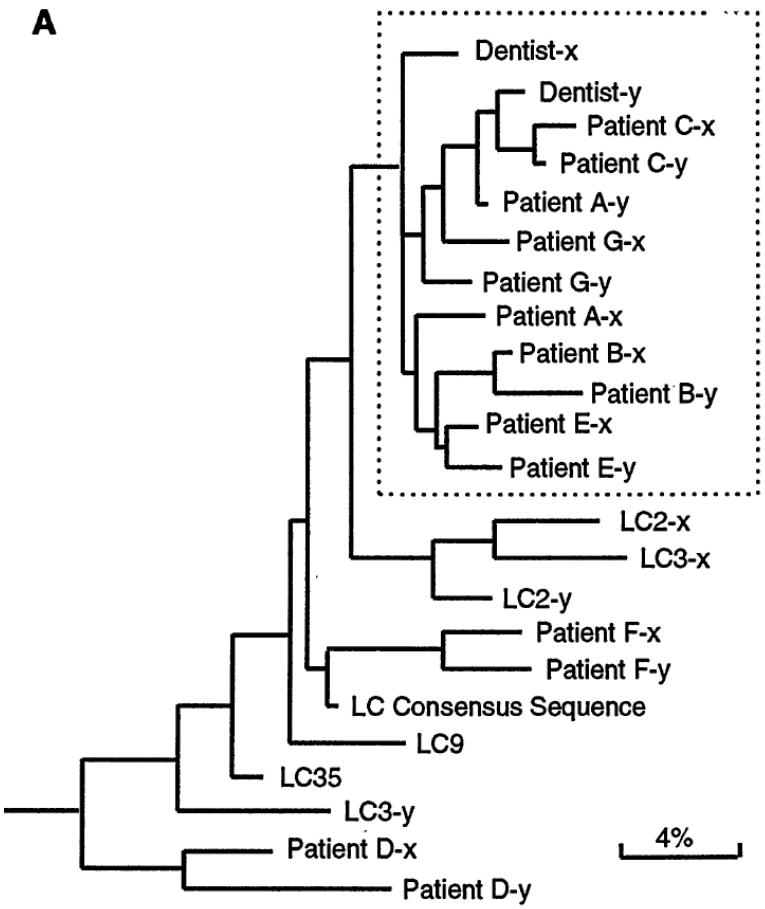
Genome evolution



The Figure shows the amount of coding DNA as a function of Genome size.

1. The population size is typically smaller in species with larger genomes
2. Animal genomes consists mostly of non-coding DNA
3. Single cell organisms have the main part of their genome coding
4. All parts of the DNA sequence in the human genome has some biological function

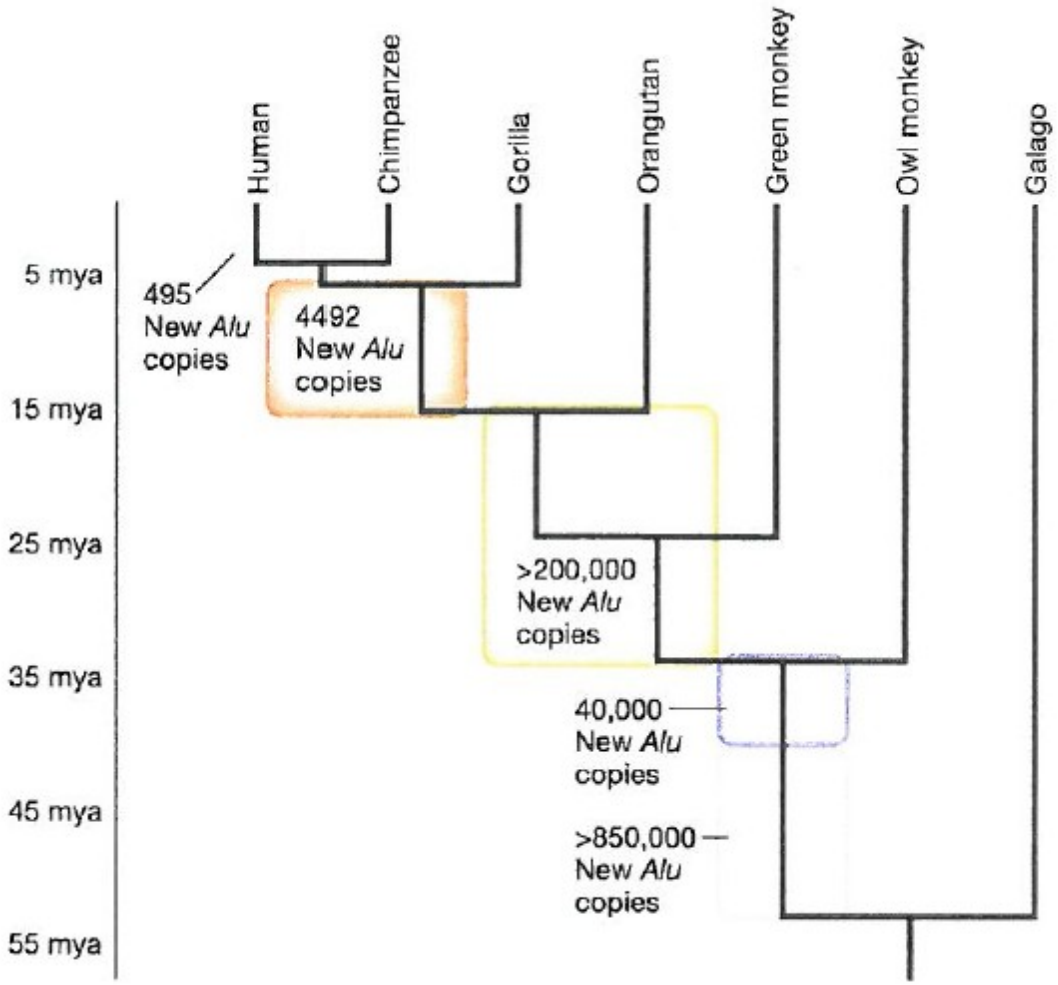
Phylogenetics



The Figure shows a phylogenetic trees of HIV nucleotide sequences used in a criminal case against a dentist (from paper by Ou et al). Assume that branches are well supported by bootstrap

1. The sequences from each patient are monophyletic
2. The nucleotide distance between LC9 and Dentist-x is larger than between LC9 and Patient F-x
3. The difference between the two dentist sequences are more different than between the two sequences from patient E
4. The tree is consistent with five out of seven patients being infected by the Dentist

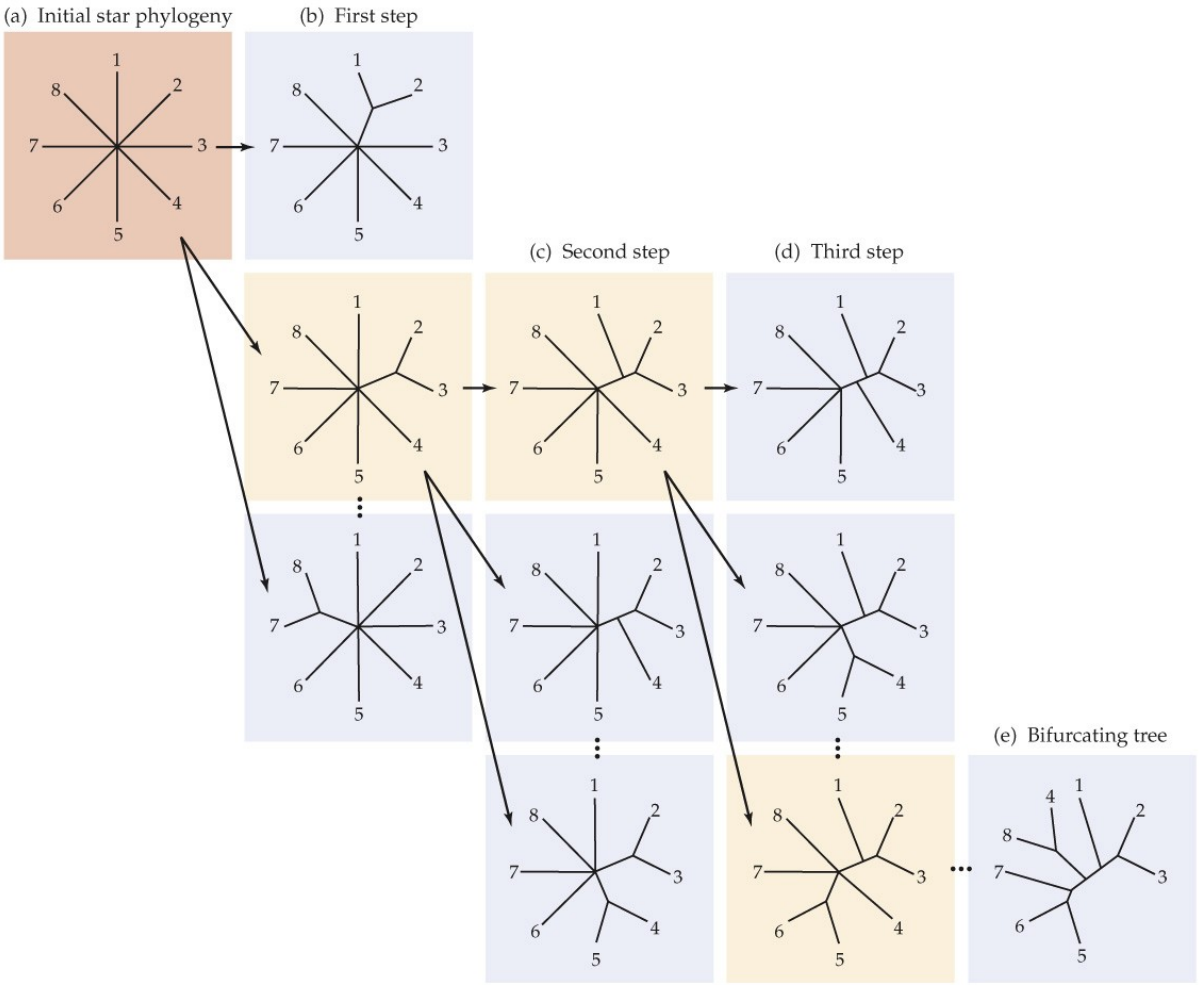
Genome evolution



ALU elements. The figure shows the number of new Alu elements acquired on the part of the primate phylogeny leading to humans

1. The number of new Alu elements acquired by the genome is approximately constant per year over the last 50 million years of human evolution
2. The Alu elements may promote exon shuffling in the human genome
3. The Alu element itself codes for an important biological function in humans
4. Most Alu elements are kept because they give the species an evolutionary advantage

Neighbor-joining

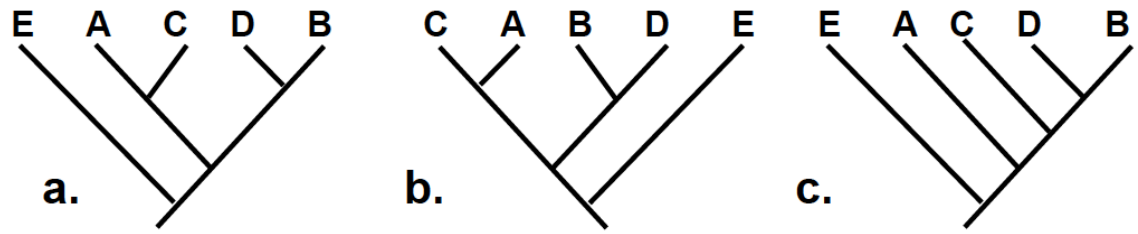


MOLECULAR AND GENOME EVOLUTION 1e, Figure 5.20

The Figure from the Graur book illustrates the principle of neighbor joining

1. The Neighbor joining algorithm first joins the most similar pair of sequences
2. The Neighbor joining algorithm inserts one new internal branch for each iteration
3. The Neighbor joining algorithm guarantees to find the shortest tree
4. Neighbor joining picks the pair of sequences that if joined will decrease the length of the tree most

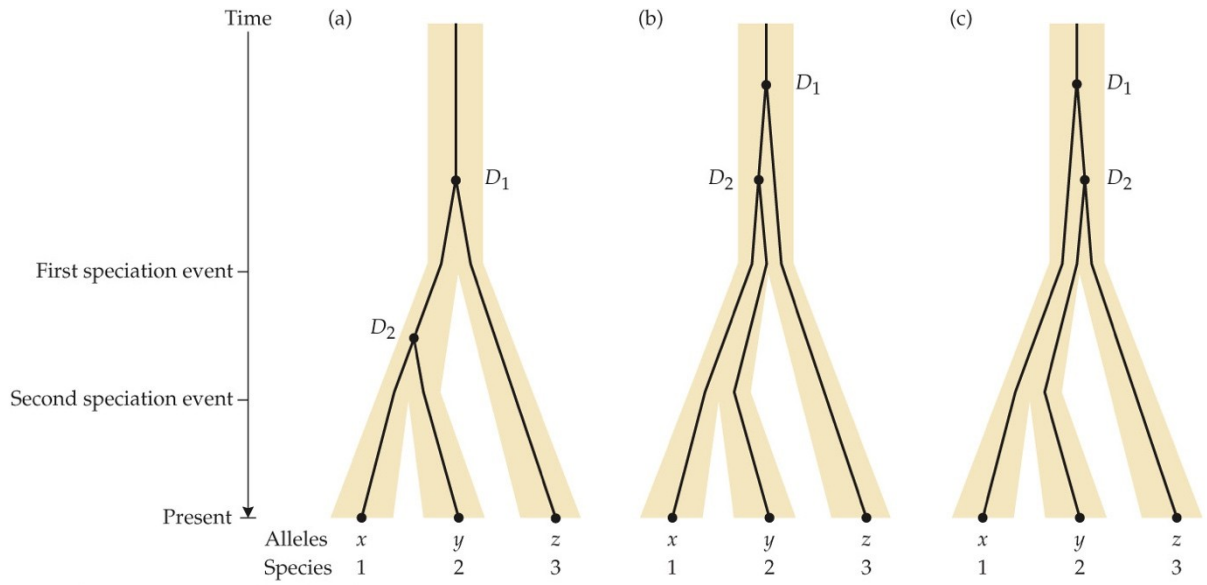
General tree understanding.



The figure above shows three rooted trees of 5 species

1. Tree a and b have the same branching pattern
2. In tree c, the common ancestor of A and B is also a common ancestor or C and B
3. In tree a, C and D are closer than A and B
4. There are 105 possible rooted trees for 5 species

Incomplete lineage sorting



MOLECULAR AND GENOME EVOLUTION 1e, Figure 5.11

The Figure shows the different possible gene trees for three species 1, 2 and 3 with species tree ((1,2),3).

1. Subfigure (b) is an example of the gene tree differing from the species tree
2. The expected time to the most recent common ancestor for *x* and *y* is the same in Subfigure (b) and (c)
3. When the population size becomes larger the probability of observing Subfigure (c) becomes larger
4. When the time between first and second speciation event becomes smaller the probability of observing Subfigure (c) becomes larger

Coalescence theory

You compare two randomly mating populations of diploid individuals: one of size 5000 and one of size 15000.

1. The effective population size is always the census number of individuals in the population
2. The expected time it takes for two sequences to find their most recent common ancestor is proportional to the size of the total population.
3. Two sequences sampled from the small population are expected to find a common ancestor in a third of the time it takes two sequences sampled from the large population.
4. The expected time it takes for five sequences to find their most recent common ancestor in the large population is the same as the expected time it takes for fifteen sequences to find their most recent common ancestor in the small population.

Coalescence theory

Assuming the infinite sites model, a randomly mating population of 10,000 diploid individuals, and that mutation is introducing new mutants into each sequence at rate 10e-5 per generation.

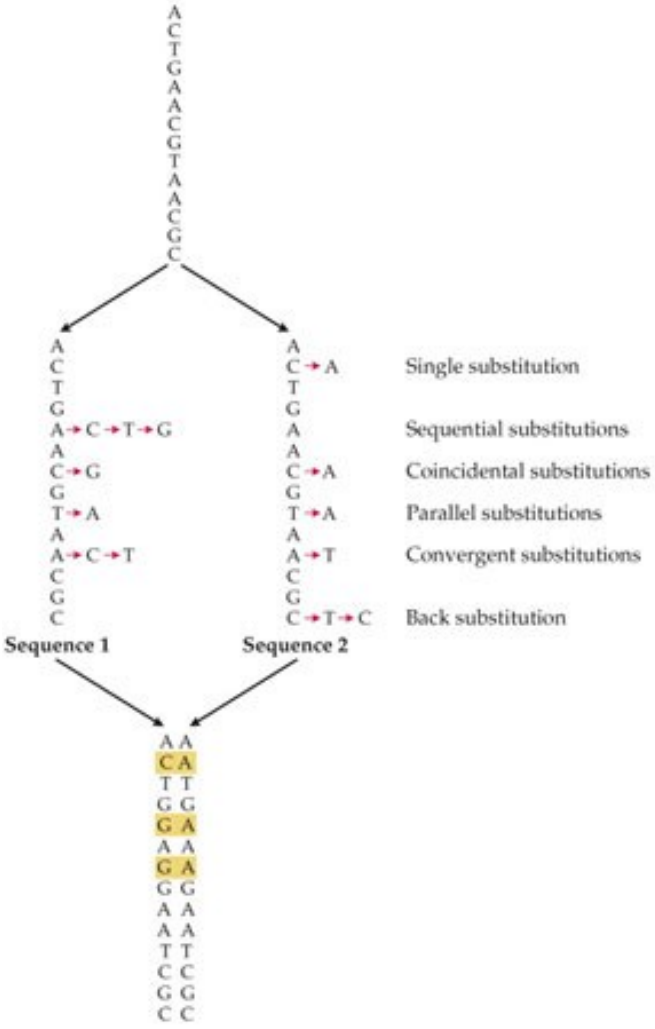
1. The divergence between two sequences sampled from the population reflects both mutation rate and population size.
2. The expected number of pairwise differences between your sampled sequences is directly proportional to the total branch length of the tree connecting the sequences.
3. The expected branch length separating two of the sequences sampled from the population is 40,000 generations.
4. If 10 genes are samples, there is more than 50% chance that the MRCA of the sample is the same as the MRCA of the whole population

Gene trees and species trees

You have built a phylogeny of ten species from an alignment of very long sequences (~1Mb) from each species.

1. The inner nodes in your tree represent the average coalescence time over the alignment.
2. A bootstrap value of 100% guarantees that there can be no part of the alignment with a different gene tree.
3. Larger ancestral populations result in larger divergence times of species.
4. Short internal branches are more likely associated with phylogenetic incongruence than long internal branches.

Molecular evolution



MOLECULAR AND GENOME EVOLUTION 1e, Figure 3.6

The Figure shows what really happened to two sequences since their common ancestor. Only three differences are found today but more than three events occurred.

1. The P distance between the sequences is 3/10
2. The real evolutionary distance between the sequences is 13/14
3. More transitions than transversions happened in the history of the sequences
4. The sequences have evolved according the infinite sites mutation model

Alignment

(e) Matrix fill completed

		1 G	2 A	3 A	4 T	5 T	6 C	7 A	8 G	9 T
	0	0	0	0	0	0	0	0	0	0
1 G	0	5	1	1	1	1	1	1	5	1
2 G	0	5	2	–2	–2	–2	–2	–2	6	2
3 A	0	1	10	7	3	3	3	3	3	3
4 T	0	1	6	7	12	8	8	8	8	8
5 C	0	1	6	3	8	9	13	9	9	9
6 G	0	5	6	3	8	5	9	10	14	10
7 A	0	1	10	11	8	5	9	14	10	11

MOLECULAR AND GENOME EVOLUTION 1e, Figure 3.14 (Part 5)
© 2016 Sinauer Associates, Inc.

(f) Traceback (without terminal gaps)

		1 G	2 A	3 A	4 T	5 T	6 C	7 A	8 G	9 T
	0	0	0	0	0	0	0	0	0	0
1 G	0	5								
2 G	0		2							
3 A	0			7	3					
4 T	0				12	8				
5 C	0						13	9		
6 G	0								14	
7 A	0									11

GAATTCAGT
GGA–TC–GA

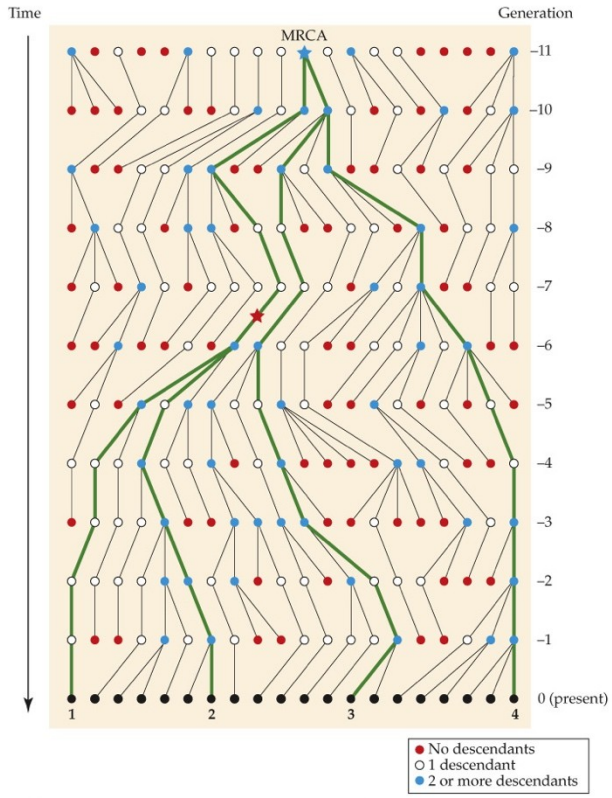
GAATTCAGT
GGAT–C–GA

MOLECULAR AND GENOME EVOLUTION 1e, Figure 3.14 (Part 6)
© 2016 Sinauer Associates, Inc.

The figures show the last two steps in a global alignment procedure from Graur

1. A long insertion is more expensive than a short insertion in this example
2. The algorithm is guaranteed to give the alignment with the highest score?
3. If I added a T to both sequences and reperformed the alignment, the total score would still be 11
4. The algorithm only works if sequences are orthologous

Coalescence



MOLECULAR AND GENOME EVOLUTION 1e, Figure 2.11

The figure shows a coalescence process for a sample of four genes in total population of 2N=20 genes.

1. The time to the most recent common ancestor in this example is longer than expected for this population size.
2. Sequence 1 and 2 are expected to be more similar than sequence 3 and 4.
3. The time to the most recent common ancestor of all four sequences is expected to be less than twice the time to the most recent common ancestor for two sequences
4. The Figure illustrates the case of a growing population

Hitch-hiking

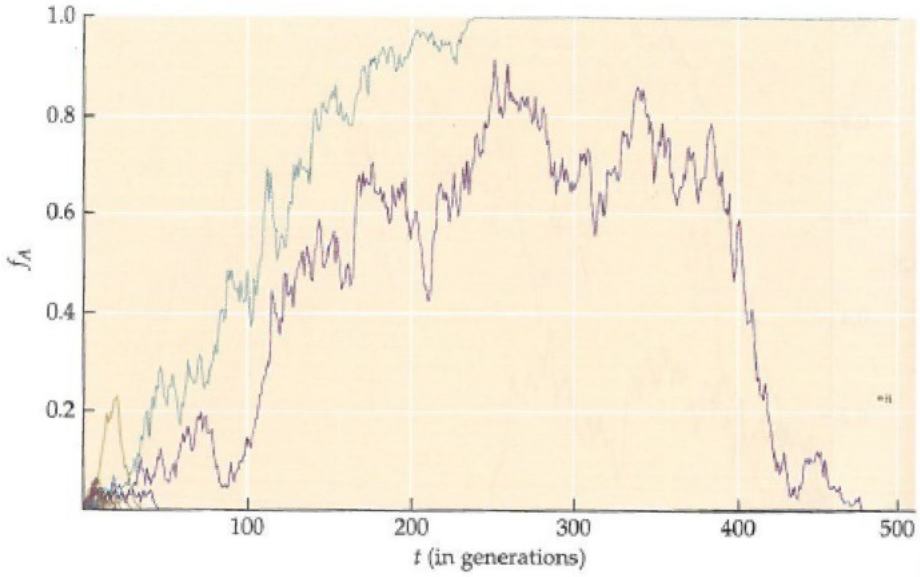
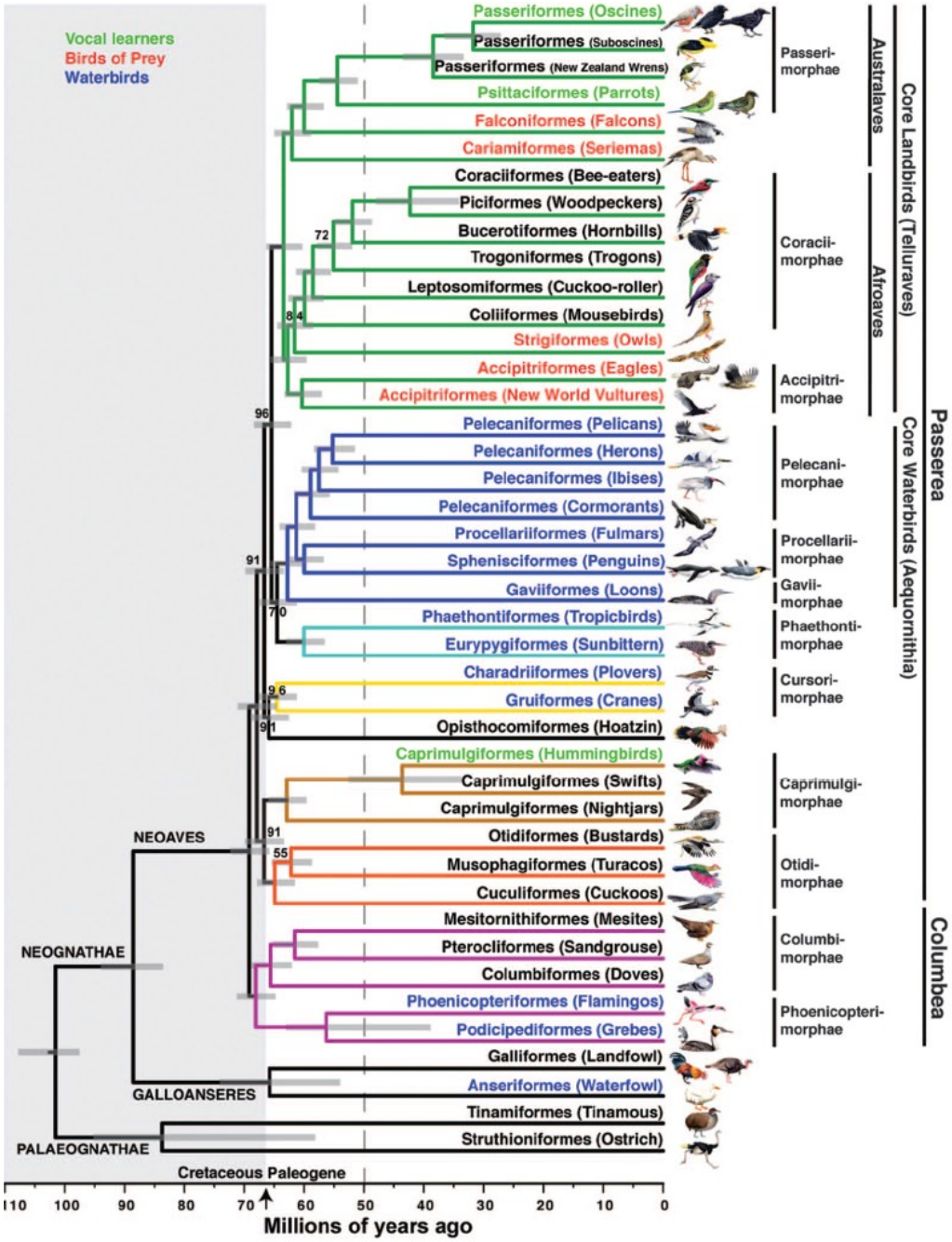


Figure 8.7 Fifty replicate trajectories of a slightly deleterious allele with an additive effect. $s = -0.005$, $N = 100$.

The Figure from Nielsen and Slatkin book shows simulations of a slightly deleterious allele, Ns=-0.5

1. We expect that less than 5% of new mutations will be fixed
2. The probability of fixation decreases with increasing frequency of the new mutation
3. For Ns=-0.5, selection is much more important than genetic drift for the fate of a new mutation
4. Segregation of deleterious alleles is expected under the neutral theory

Phylogeny



The illustrations show a phylogeny of bird species created with a maximum likelihood approach that also assumes a molecular clock. Bootstrap values are shown when above 50

1. The speciation rate of birds seems to have increased within the last 20 million years
2. Results are consistent with a burst of speciation occurring soon after the extinction of dinosaurs 65 million years ago
3. According to the phylogeny, birds of prey are monophyletic
4. Vocal learning appears to have evolved only once

Selection

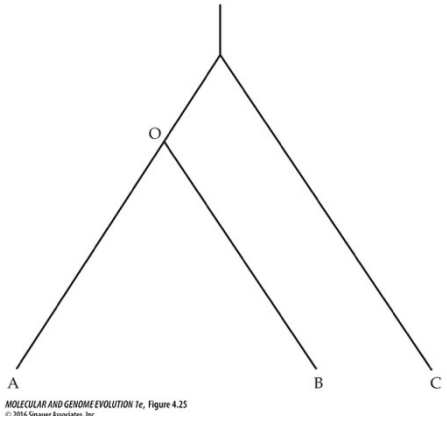
1. Overdominance in fitness (heterozygote advantage) is likely to lead to a stable polymorphism
2. Selection is much stronger than genetic drift when $Ns<<1$
3. Fixation of a strongly advantageous allele with s=0.01 is faster in a large population (N=100,000) than in a small population (N=10,000)
4. The probability that a new mutation with s=2% is lost is about 96%

Selective sweeps

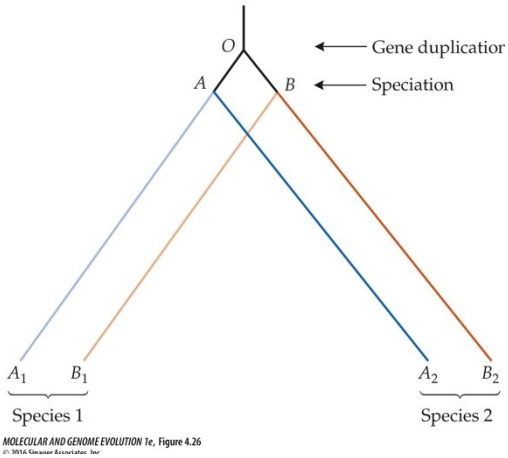
1. A selective sweep removes variation in a larger region when the recombination rate is high
2. A selective sweep removes variation in a larger region in a large population than in a small population
3. A selective sweep will reduce diversity within a species more than the divergence between species
4. A partial sweep leads to some of the haplotypes being nearly of completely identical around the site under selection
5.

Molecular clock

Consider the following Figures from the Graur book



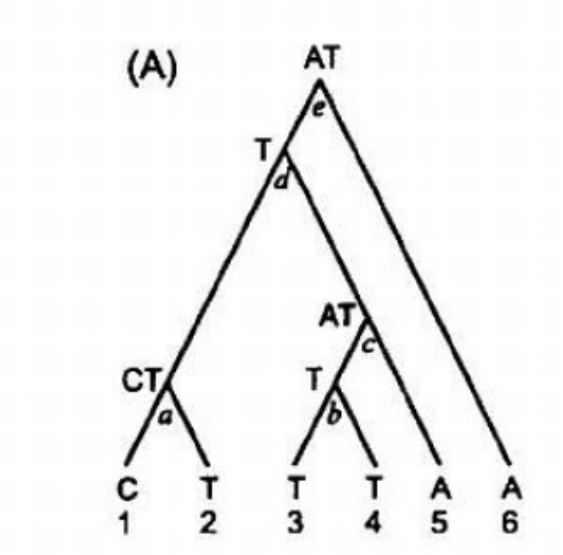
MOLECULAR AND GENOME EVOLUTION 1e, Figure 4.25



MOLECULAR AND GENOME EVOLUTION 1e, Figure 4.26

1. The relative rates test using the sequences A, B and C in the illustration to the left tests whether A and C have different distances to the root.
2. The gene duplication test (illustration right) tests against the hypothesis that two genes A, and B, each evolved at the same rate in two different species
3. Neutral evolution and a constant mutation rate are sufficient for a molecular clock to exist
4. A molecular clock exists between humans and plants

Parsimony



The Figure illustrates how to infer the minimum number of substitutions for one site on a given tree using the parsimony principle.

1. The tree shown is the most parsimonious tree for this site
2. Node *c* should be a T and not an A under parsimony
3. Node *a* should be a C and not a T under parsimony
4. The parsimony number of changes on the tree is 4

Substitution models

Table 3.2 Models of nucleotide substitution.

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_C
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	ag_T	bg_C	cg_G
T	αg_A	-	αg_C	αg_G	ag_A	-	dg_C	eg_G
C	αg_A	αg_T	-	αg_G	bg_A	dg_T	-	fg_G
G	αg_A	αg_T	αg_C	-	cg_A	eg_T	fg_C	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

The Figure above shows some of the most commonly used substitution models for DNA sequences

1. The equilibrium frequencies of the four nucleotides are the same under Jukes-Cantor and Kimura's model.
2. The HKY model is nested in the Equal-input model
3. The difference in the number of free parameters in the HKY and Kimura model is four
4. Simpler models have less variance

A likelihood ratio test can be used when

1. Testing whether a tree is a maximum parsimony tree
2. Testing between two nested substitution models
3. Testing whether bootstrap values are significant
4. Testing whether a set of sequences evolve according to a molecular clock

Substitution models

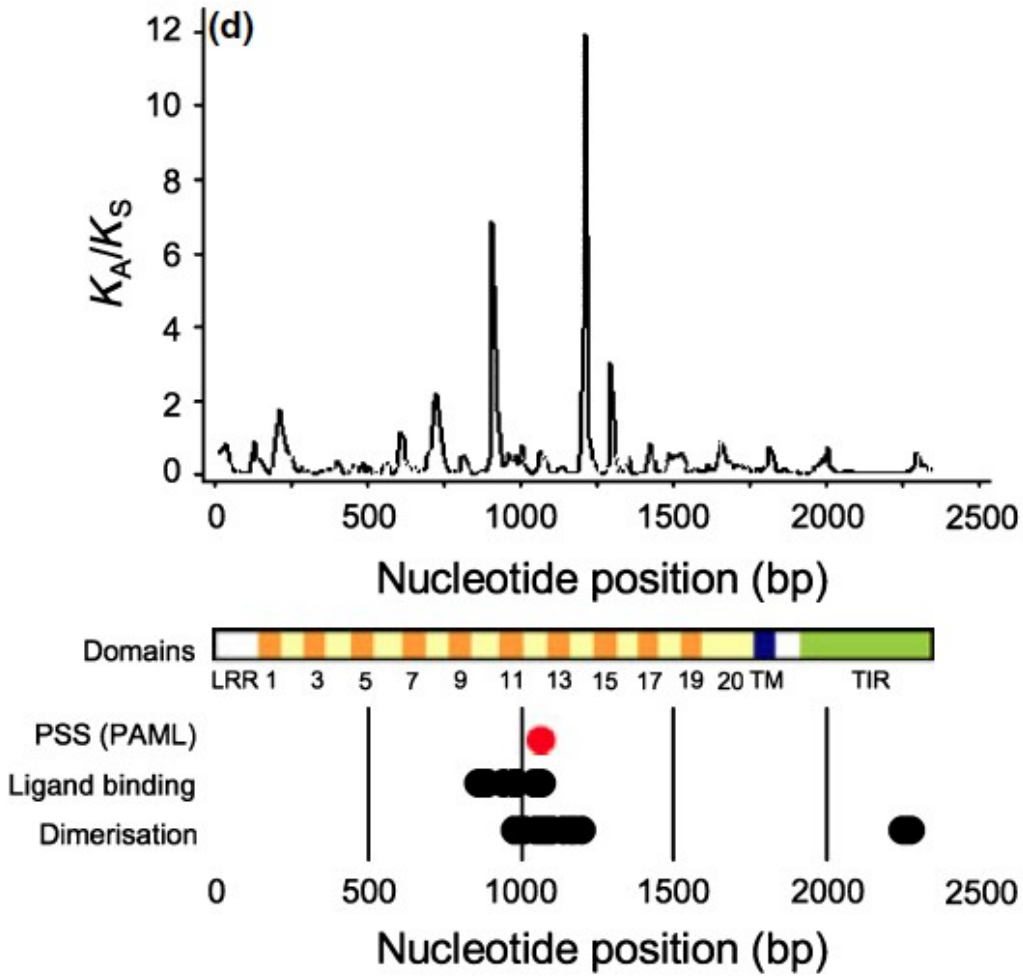
Table. Maximum Likelihood fits of 24 different nucleotide substitution models

Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)
TN93+G	15	50404.479	50262.125	-25116.060	n/a	0.05	24.68	0.310	0.249	0.310	0.132
TN93+G+I	16	50415.948	50264.105	-25116.050	0.00	0.05	24.68	0.310	0.249	0.310	0.132
HKY+G	14	50418.129	50285.265	-25128.631	n/a	0.05	24.81	0.310	0.249	0.310	0.132
HKY+G+I	15	50420.437	50278.083	-25124.039	0.44	0.05	26.78	0.310	0.249	0.310	0.132
GTR+G	18	50438.296	50267.473	-25115.733	n/a	0.05	24.68	0.310	0.249	0.310	0.132
GTR+G+I	19	50449.747	50269.434	-25115.713	0.00	0.05	24.68	0.310	0.249	0.310	0.132
TN93	14	50456.076	50323.212	-25147.604	n/a	n/a	23.13	0.310	0.249	0.310	0.132
TN93+I	15	50467.533	50325.179	-25147.587	0.00	n/a	23.13	0.310	0.249	0.310	0.132
HKY	13	50471.857	50348.483	-25161.240	n/a	n/a	23.13	0.310	0.249	0.310	0.132
HKY+I	14	50483.321	50350.457	-25161.227	0.00	n/a	23.13	0.310	0.249	0.310	0.132
GTR	17	50489.801	50328.468	-25147.231	n/a	n/a	23.14	0.310	0.249	0.310	0.132
GTR+I	18	50501.255	50330.432	-25147.213	0.00	n/a	23.14	0.310	0.249	0.310	0.132
T92+G	12	51723.229	51609.346	-25792.671	n/a	0.05	24.48	0.279	0.279	0.221	0.221
T92+G+I	13	51734.043	51610.669	-25792.333	0.03	0.05	24.53	0.279	0.279	0.221	0.221
T92	11	51774.609	51670.216	-25824.107	n/a	n/a	23.11	0.279	0.279	0.221	0.221
T92+I	12	51786.073	51672.190	-25824.093	0.00	n/a	23.11	0.279	0.279	0.221	0.221
K2+G	11	51935.868	51831.474	-25904.736	n/a	0.05	24.42	0.250	0.250	0.250	0.250
K2+G+I	12	51940.505	51826.621	-25901.309	0.40	0.05	25.76	0.250	0.250	0.250	0.250
K2	10	51986.222	51891.319	-25935.658	n/a	n/a	23.10	0.250	0.250	0.250	0.250
K2+I	11	51997.688	51893.294	-25935.646	0.00	n/a	23.10	0.250	0.250	0.250	0.250
JC+G	10	52885.220	52790.317	-26385.157	n/a	0.06	0.50	0.250	0.250	0.250	0.250
JC+G+I	11	52896.711	52792.317	-26385.157	0.00	0.06	0.50	0.250	0.250	0.250	0.250
JC	9	52916.559	52831.146	-26406.572	n/a	n/a	0.50	0.250	0.250	0.250	0.250
JC+I	10	52927.957	52833.053	-26406.526	0.00	n/a	0.50	0.250	0.250	0.250	0.250

The Figure shows the result of testing substitution models for a dataset in MEGA. Models are ranked according to BIC

1. The best model is the most complex one in this case
2. There is evidence for different nucleotide frequencies for the investigated gene
3. There is evidence for a gamma distributed heterogeneity in substitution rates on investigated gene
4. The model with the lowest likelihood score is preferred in this example

Evolution of a coding gene



Estimation of synonymous (Ks) and non-synonymous (Ka) substitution rates have been made in an innate immunity gene evolving in rodents as shown in the Figure above. We can conclude that

1. The gene is unlikely to be functional
2. Some regions of the gene appear to be evolving under purifying selection
3. Positive selection appears to have been targeting part of the region involving ligand binding
4. Over time the amino acid sequences is expected to evolve fastest in regions where K_A/K_S are largest if mutation rates are constant

The molecular clock

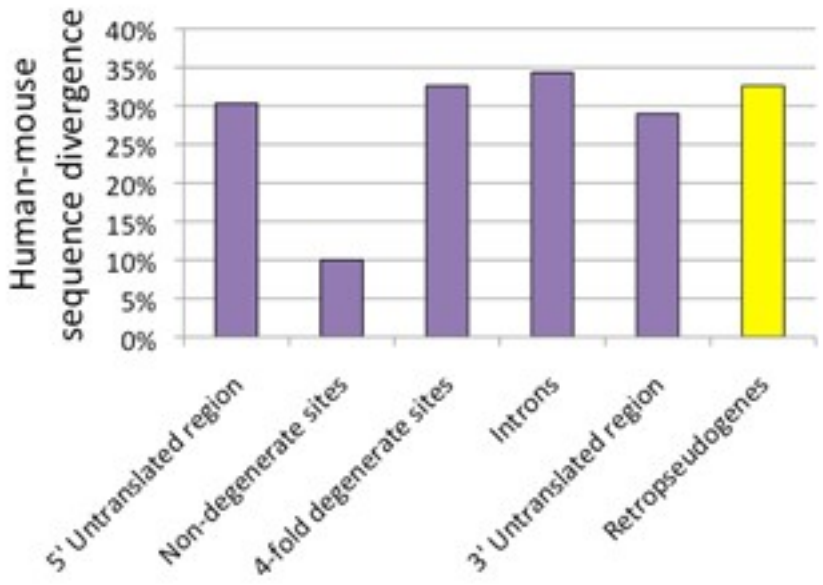
1. Both protein sequences and nucleotide sequences can follow a molecular clock
2. If the assumption of a molecular clock is true, then it can be used to date the times of divergence for a set of species
3. If mutation occurs at rate *u* and all mutation is neutral, then the substitution rate is 4*u* and independent of N
4. Increasing the effective population size N speeds up the molecular clock

Gene trees and species trees

Consider any three species.

1. Incomplete lineage sorting can occur even if species diverged many million years ago
2. Recombination eliminates incomplete lineage sorting.
3. The amount of incomplete lineage sorting will be larger if the time between speciation events is small.
4. The proportion of incomplete lineage sorting will be smaller if the ancestral population sizes are large.

Purifying selection



Complete genomic data for humans and mice allows the comparison of genetic sequences between the two species to calculate sequence divergence in different regions of genes and in retro-pseudogenes. (Data from: Makalowski, W. *et al*. (1998) and Zheng, D. *et. al*. (2007).).
We assume when interpreting the data above that all mutations arising only fall in two categories: neutral mutations and deleterious ones.

1. Non-degenerate sites are functional and therefore evolve more slowly than pseudogenes
2. Codon usage bias should affect 4-fold degenerate sites more than 3' untranslated sites
3. Introns are less constrained than non-degenerate sites.
4. The data presented above proves that the mouse genome evolved under stronger purifying selection than the human genome.

Selection

1. More conserved proteins typically have lower d_N/d_S ratios
2. A d_N/d_S ratio of 1 is evidence of strong purifying selection
3. An average $d_N/d_S < 1$ for a given gene implies that no codons in the gene can be under positive selection
4. The McDonald-Kreitman test suggests positive selection only if $d_N > d_S$