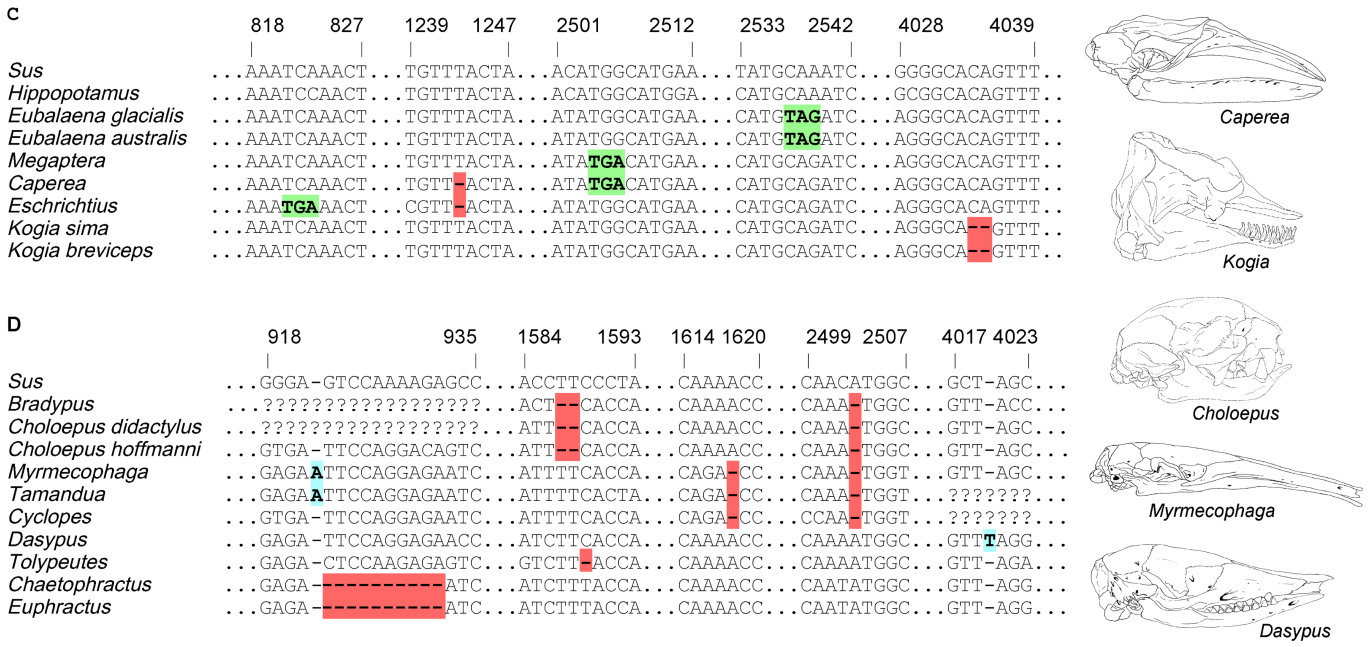
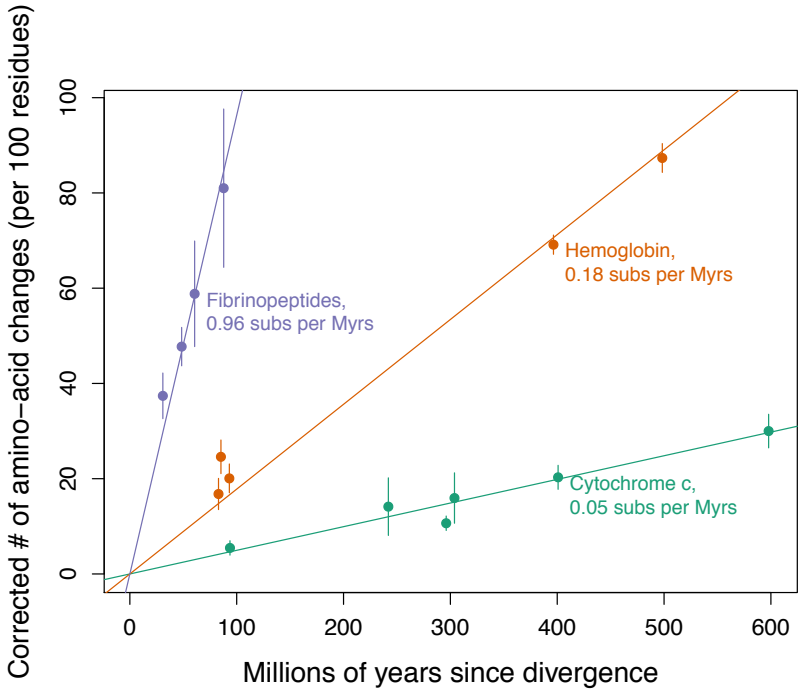
Coop, Chapter 5: Intro-5.1.3

The Population Genetics of Divergence and Molecular Substitution

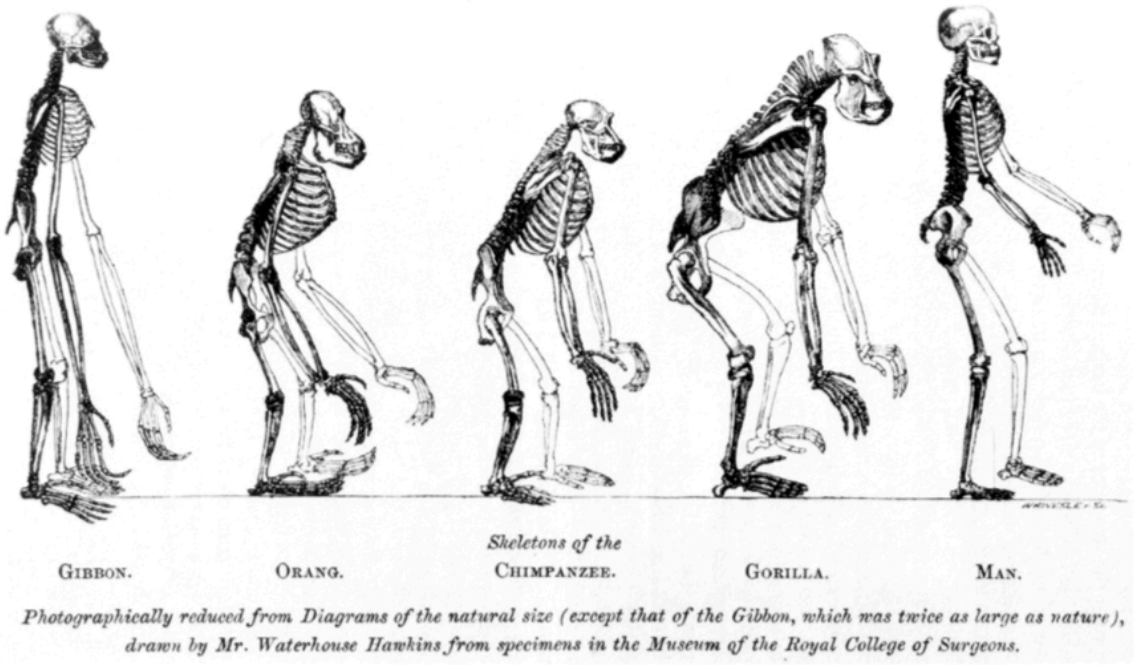
The neutral substitution process





Introduction

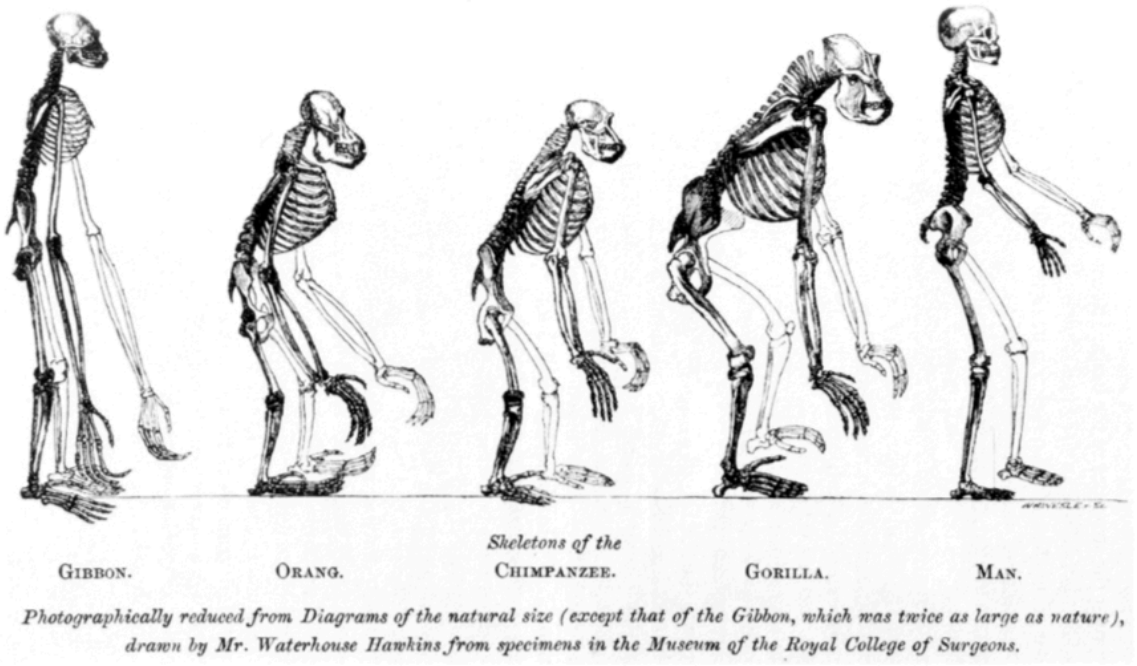
- Humans and chimpanzees, our most closely related species of the Great Apes, differ consistently (have substitutions) at only 1% of loci across the genome
- These changes have arisen through mutation over the last 7 million years
- Unique patterns of substitution can be observed across the Great Ape species
- Each of these mutations arose as a single mutation within a population, rose in frequency, and fixed



Human	accacagcatttgttagttactgccaagaagcctgtatctgtagggtaaaatcctcgctgaagtgggttg
Chimp	.....g.....c.....
Gorilla	.....cc.....
Orangutan	.....c.....c.....c.....
Gibbon	.....c.....---.....
Crab-eating macaque	g.....gg...c.....c..t.t.....

Introduction

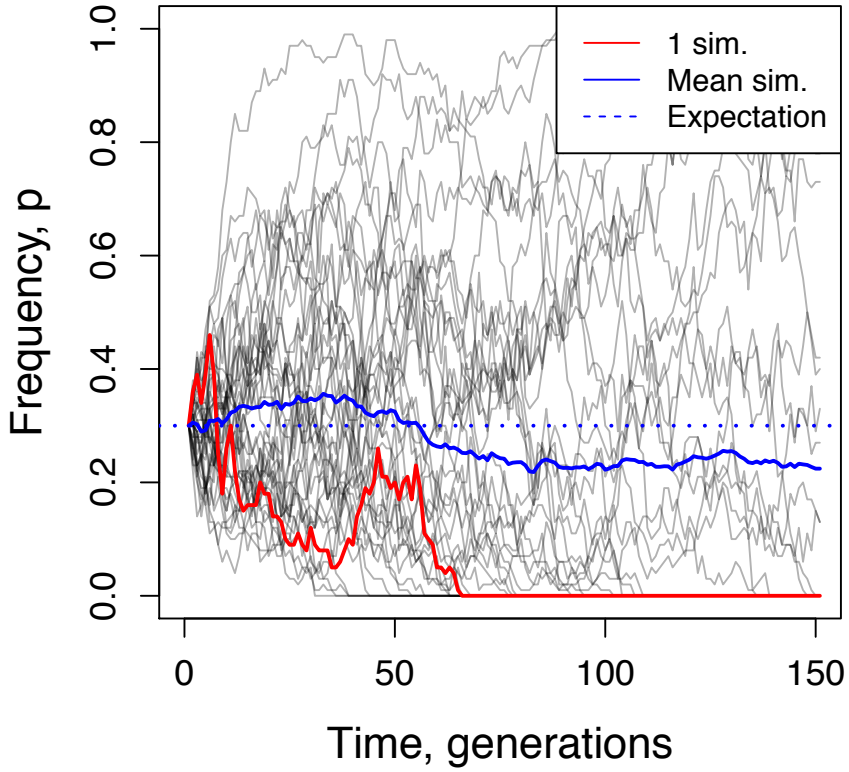
- What Evolutionary forces drove this process?
- Many of these changes were due to adaptive substitutions, new mutations that conferred greater fitness
- Many more, however occurred in non-coding DNA and are likely neutral
- One major goal in Molecular Evolution is to find the loci that are under constraint or that underlie adaptation on a particular lineage
- Expectations under neutrality are a good place to start in working toward this goal




```
Human                accacagcatttgttagttactgccaagaagcctgtatctgtagggtaaaatcctcgctgaagtgggttg
Chimp                ....................................g..........c...................
Gorilla              ................................................cc..................
Orangutan            .........c.........c..............................c..................
Gibbon               ..................c..............................---.................
Crab-eating macaque  g............gg...c..............................c..t.t.............
```

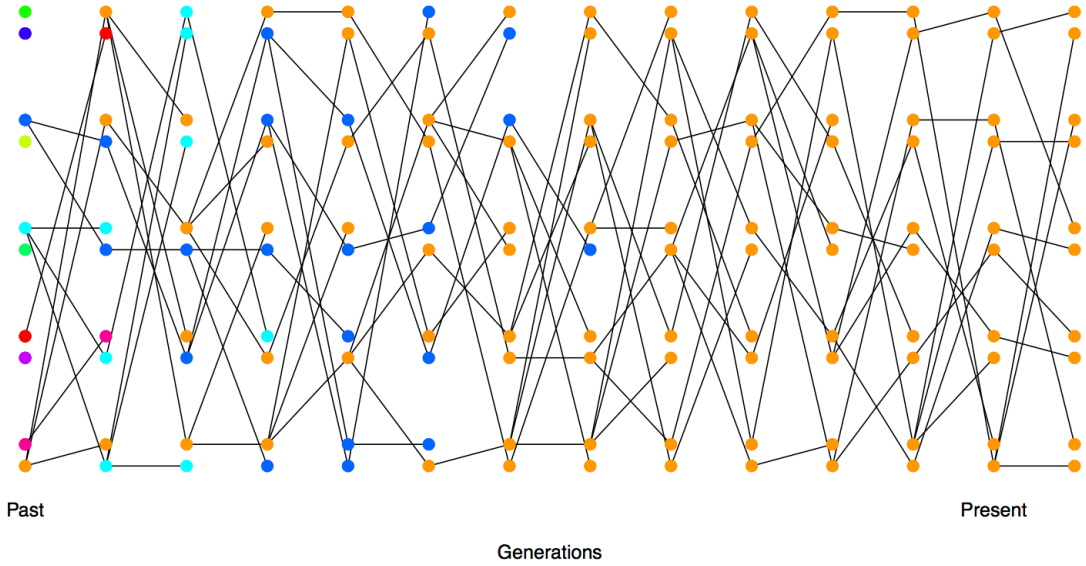
5.1 The Neutral Substitution Process

- So how can a neutral mutation fix and create a substitution?
- Most new mutations will be lost through drift, with the probability of fixation initially being $\frac{1}{2N}$
- While this may seem like a vanishingly small probability, there are enough new mutations over the many, many generations in the history of life, that we see a large amount of neutral substitutions accumulating



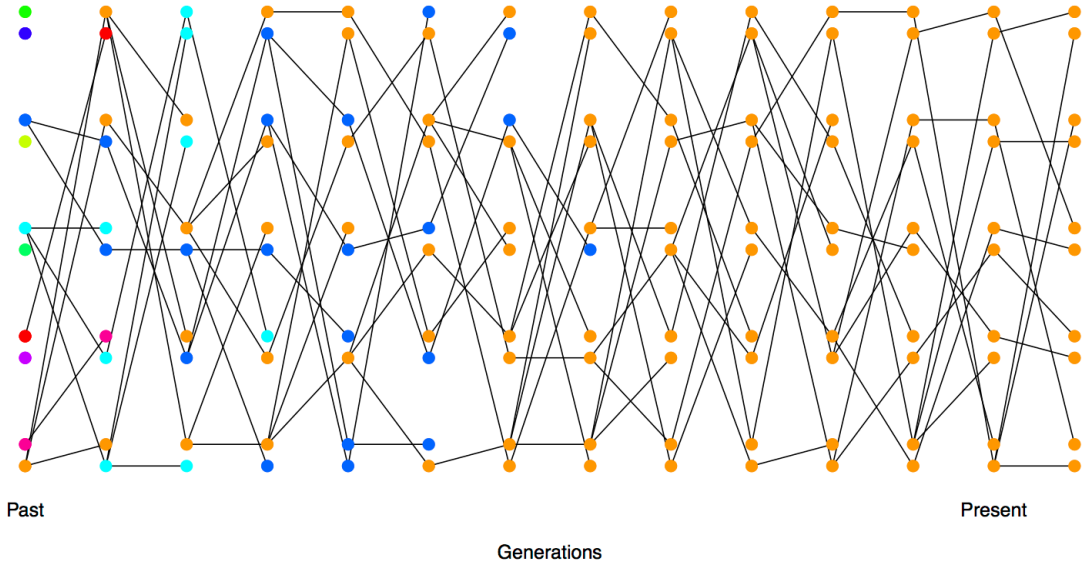
5.1.1 Probability of the eventual fixation of a neutral allele

- When a substitution occurs, all of the individuals in subsequent generations can trace their ancestry back to that single allele
- For example, by the 9th generation in the figure to the right, an orange substitution has occurred and all subsequent individuals trace their history back to this allele
- The probability of this allele's fixation in the first generation was $\frac{1}{2N} = \frac{1}{10} = 0.1$



5.1.1 Probability of the eventual fixation of a neutral allele

- More generally, we can say that the probability of any allele fixing is it's number of copies divided by the total number of alleles in the population: $\frac{i}{2N}$
- In other words, its probability of fixation is just its frequency (p) in the population
- We've seen previously, that it takes $\approx 4N$ generations for a large sample of alleles to coalesce to the T_{MRCA}
- Similarly, it will take $\approx 4N$ generations for a new mutation to fix as a substitution



5.1.2 Rate of substitution of neutral alleles

- We can think of substitution as fixation within an isolated population and therefore the rate of substitution as the rate of new mutations in this population reaching fixation
- We will assume that new mutations are either highly deleterious (C) or neutral ($1 - C$)
- A total of $2N\mu(1 - C)$ neutral mutations will enter our population each generation
- The probability that a mutation fixes is $\frac{1}{2N}$, so the rate at which neutral mutations arise that will ultimately fix in the population is:

$$2N\mu(1 - C)\frac{1}{2N} = \mu(1 - C) \qquad (5.1)$$

5.1.2 Rate of substitution of neutral alleles

- Now, to consider the number of substitutions that may distinguish two isolated populations or species, we can consider they have been separated for T generations
- Taking this into account, we can predict the number of accumulated substitutions is:

$$2\mu(1-C)T \qquad (5.2)$$

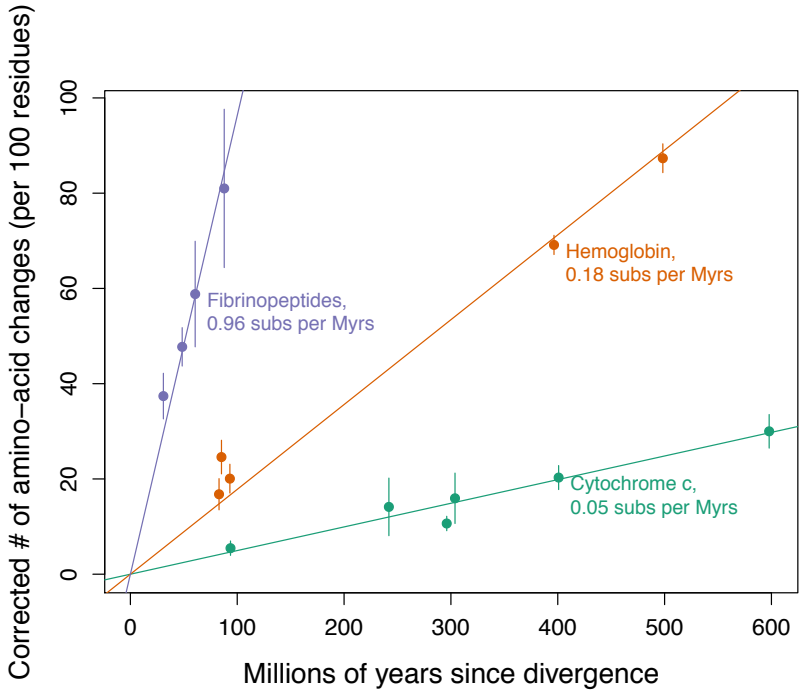
- Conveniently, the population size has cancelled out of the neutral substitution rate

5.1.3 Implications for the Molecular Clock

- Equation 5.2 tells us that molecular evolution in a genomic region is governed by constraint (C)
- Empirical evidence supports this; variation accumulates at a rate from fastest to slowest:

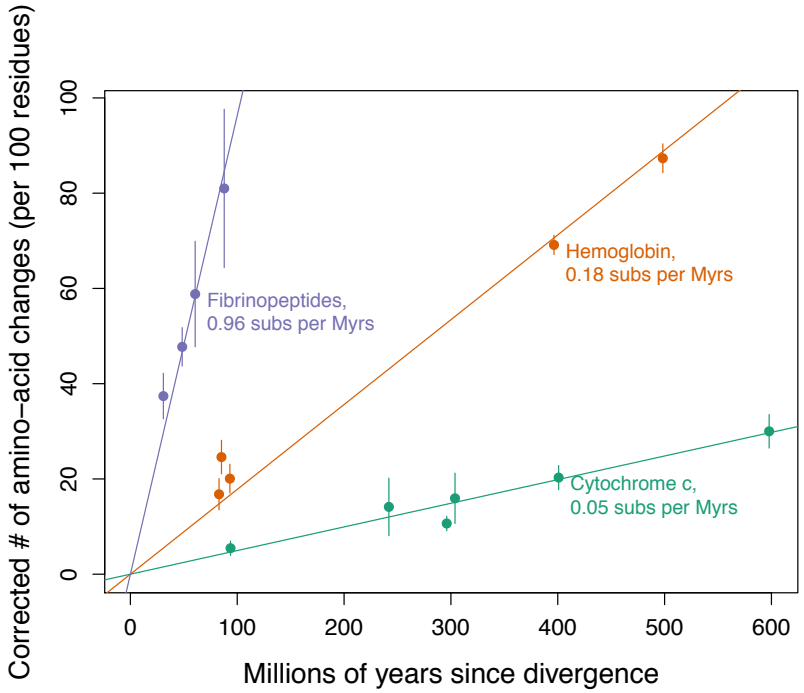
noncoding DNA > synonymous > non-synonymous in less vital proteins > non-synonymous in vital proteins

- For example, Fibrinopeptides are less constrained than Cytochrome c
- This means that we can look at constraint across the genome, even in non-coding regions, to gauge the functional importance of a region



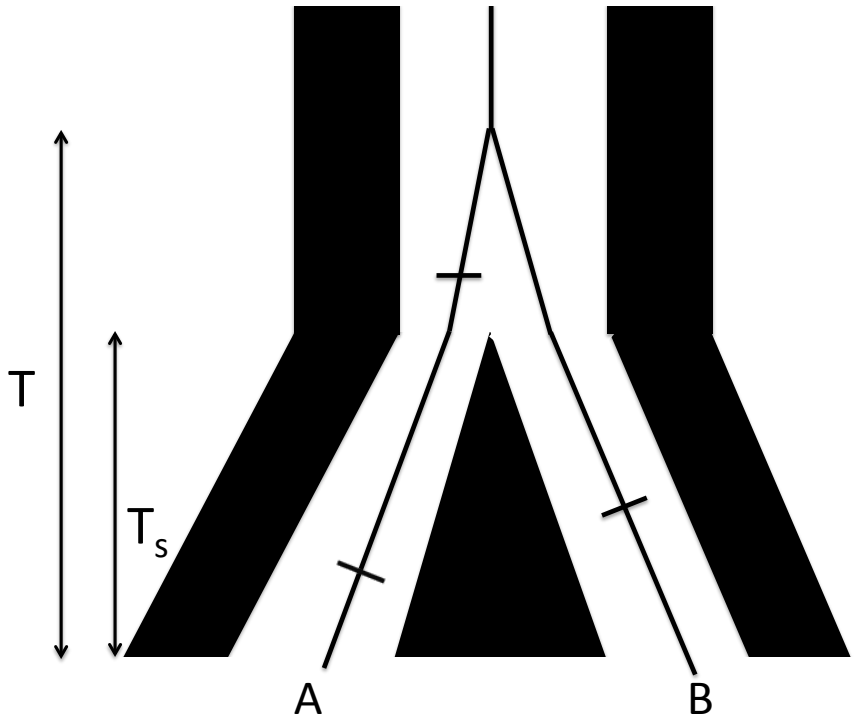
5.1.3 Implications for the Molecular Clock

- It is also exciting that Equation 5.2 is consistent with Zuckerkandl and Pauling's (1965) hypothesis of a constant molecular clock
- There is a linear relationship between increase of non-synonymous substitutions and the T_{MRCA} based on the fossil record
- This assumes, however, that the mutation rate in generations is consistent across species, which has been shown to not be the case



5.1.3 Implications for the Molecular Clock

- Typically, we would think of time T as the number of generations since populations/species split
- However, sorting of ancestral variation can contribute to substitutions
- For example, in the figure to the right, lineages A and B coalesce further in the past than the split time T_S
- The top mutation was polymorphic in the ancestral species but contributes to divergence in A and B
- We can relate these times, assuming the ancestral populations size was N_A using:

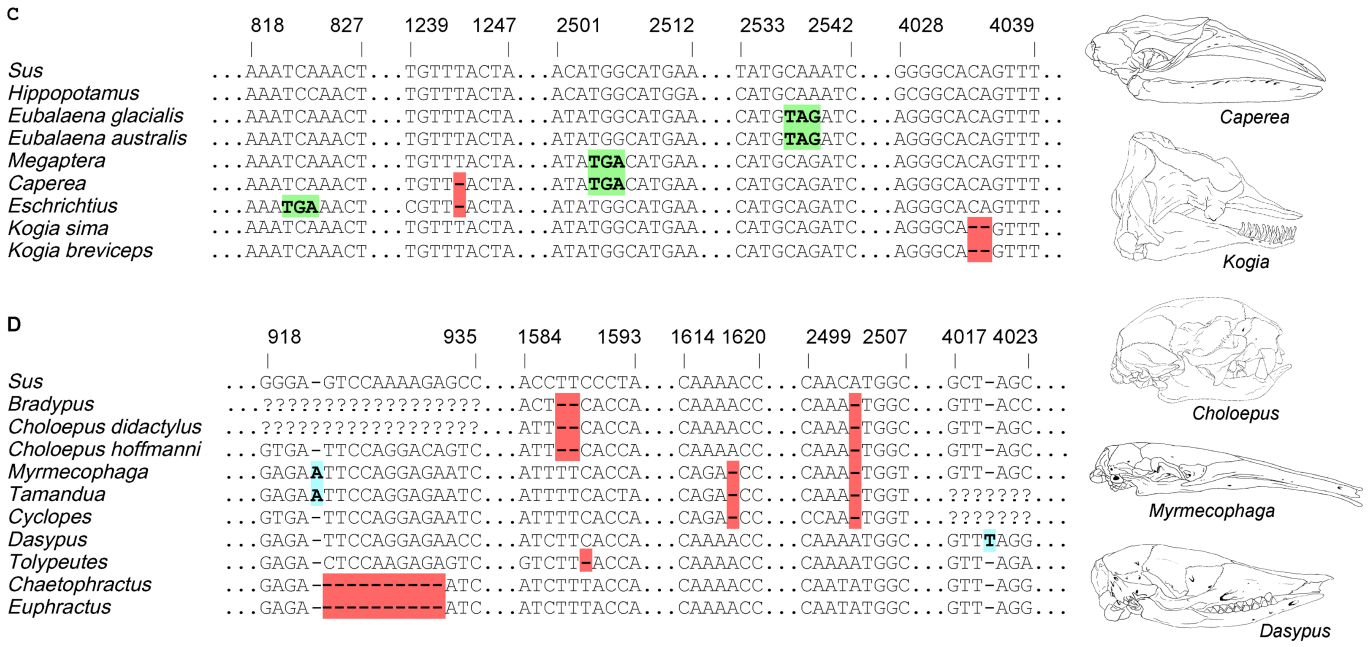


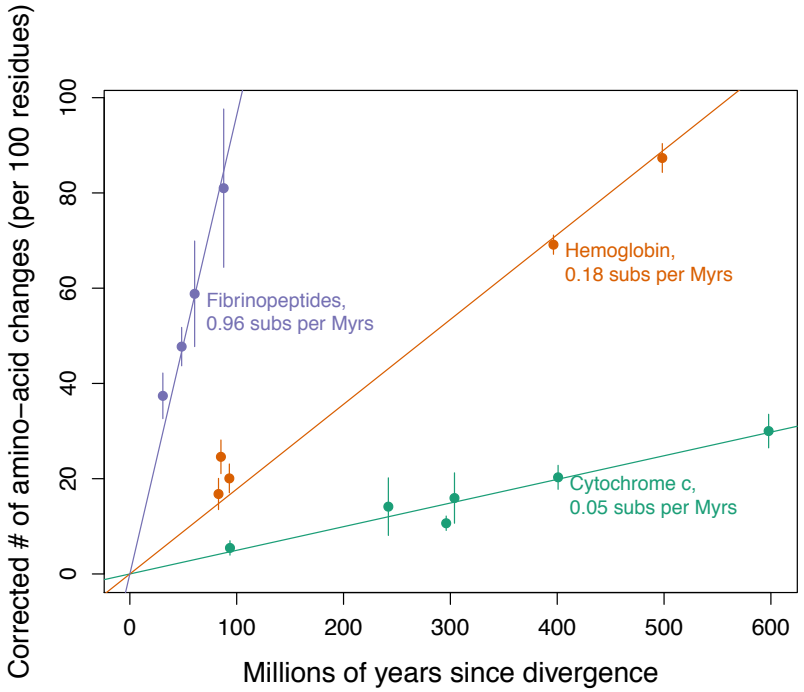
$$T = T_s + 2N_A. \qquad (5.3)$$

Coop, Chapter 5: 5.2-5.2.1

The Population Genetics of Divergence and Molecular Substitution

Tests of Molecular Evolution





5.2.1 Rate of synonymous and non-synonymous substitutions

- A common strategy in Molecular Evolution is to compare rates of substitution at various classes of genomic sites (*e.g.*, the ratio of non-synonymous and synonymous substitutions in a gene)

- To summarize these, for example, we can count the number of observed non-synonymous substitutions in a gene and divide this by the total number of sites where a non-synonymous mutation is possible (referred to as d_N); d_S is calculated similarly

- For most protein-coding genes we see $d_N/d_S < 1$, meaning fewer non-synonymous substitutions are expected proportionately due to constraint

5.2.1 Rate of synonymous and non-synonymous substitutions

- In order to estimate the level of constraint on non-synonymous sites, let's assume that all synonymous sites are neutral and accumulate as $d_s = 2T\mu$
- A fraction C of non-synonymous mutations are too deleterious to contribute to polymorphism
- After T generations, for neutral non-synonymous we would expect:

$$d_N = 2T(1 - C)\mu \qquad (5.4)$$

- Which, after dividing by d_S, gives us:

$$d_N/d_S = (1 - C) \qquad (5.5)$$

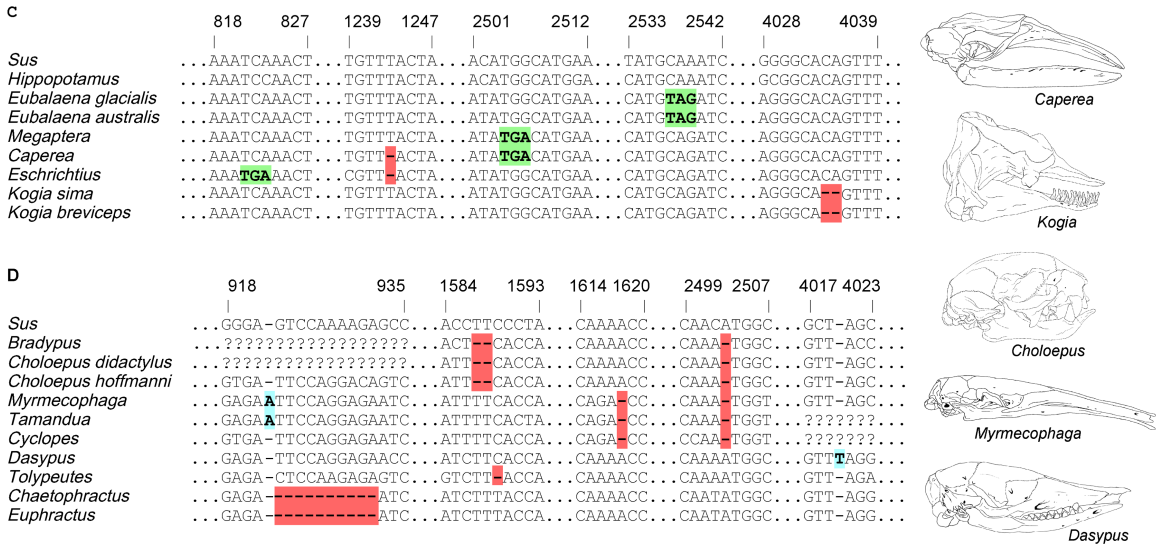
- And $C = 1 - d_N/d_S$

5.2.1 Rate of synonymous and non-synonymous substitutions

- We can test whether our gene is under constraint by determining whether d_N/d_S is significantly less than 1
- We can also assess this in the context of a phylogeny, to see if particular branches are under very strong constraint
- At times, the function a gene serves is no longer important and it is no longer constrained
- Without constraint, non-synonymous mutations are just as free to evolve as synonymous mutations and d_N/d_S = 1

5.2.1 Rate of synonymous and non-synonymous substitutions

- The Enamlin protein is an excellent example where the underlying gene is no longer constrained in some lineages
- This protein forms the enamel of teeth and certain mammals like sloths, whales, and aardvarks no longer need hard teeth because of dietary changes
- They have lost their hard teeth enamel and the Enamlin gene is no longer under constraint
- pseudogenizing substitutions have arisen



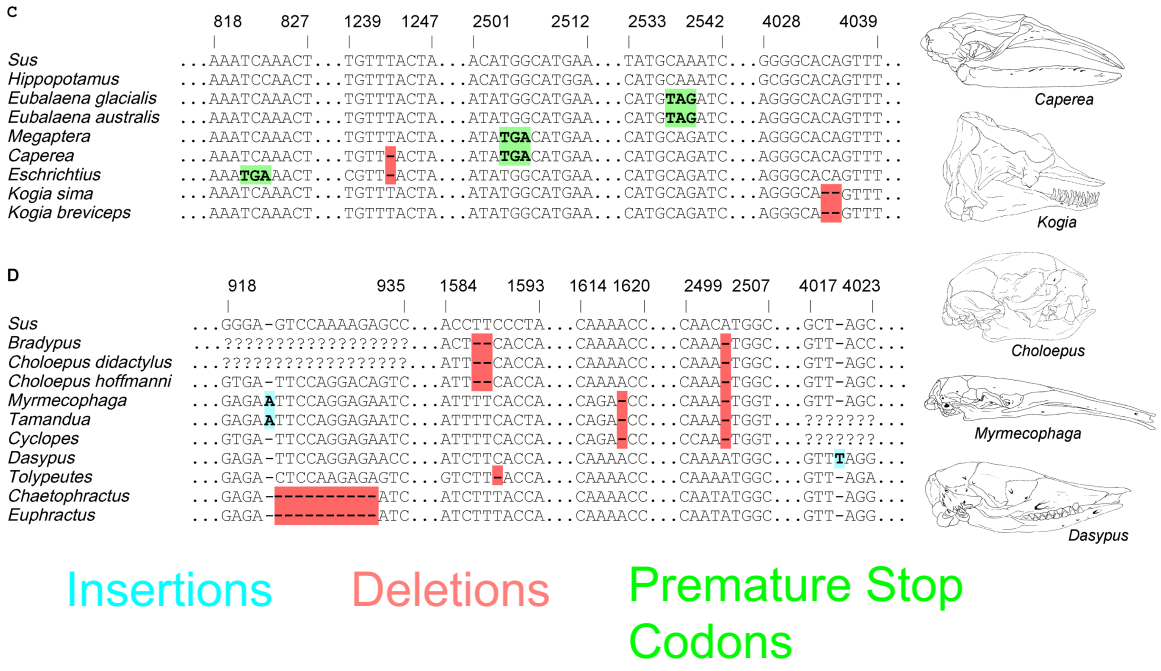
Insertions Deletions Premature Stop Codons

5.2.1 Rate of synonymous and non-synonymous substitutions

- Meredith and colleagues (2009), who carried out the Enamlin study, found that species with a functional gene and hard teeth enamel had a d_N/d_S = 0.51, significantly less than 1 and consistent with constraint

- Species with an Enamlin pseudogene had a d_N/d_S = 1.02

- Certain species appeared to be transitioning from a functional to a non-functional Enamlin gene with d_N/d_S = 0.83-0.98



5.2.1 Rate of synonymous and non-synonymous substitutions

- Not all mutations are either deleterious or neutral. Some fraction will be beneficial
- Let's assume that a fraction B of non-synonymous mutations are beneficial and arise each generation at a rate of $2NuB$
- Particularly when they are rare, beneficial mutations can be lost from the population
- With the probability of fixation of a beneficial mutation represented as f_B, we can rewrite the prevalence of non-synonymous substitutions between two populations as:

$$dN = 2T(1 - C - B)\mu + 2T \times (2N\mu B) \times f_B \qquad (5.6)$$

- and continuing to assume all synonymous mutations are neutral:

$$d_N/d_S = (1 - C - B) + 2NBf_B \qquad (5.7)$$

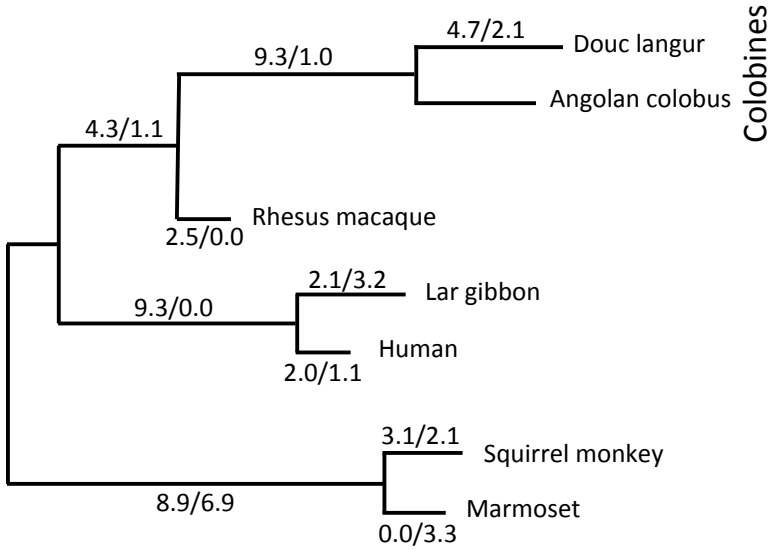
5.2.1 Rate of synonymous and non-synonymous substitutions

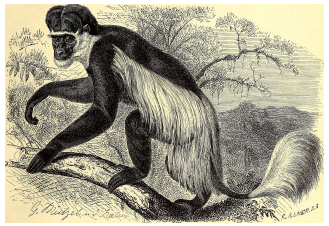
$$d_N/d_S = (1 - C - B) + 2NBf_B \qquad (5.7)$$

- When a particular gene has accumulated a number of beneficial substitutions, d_N/d_S > 1, and the inference is that this gene is rapidly evolving
- We can identify rapidly evolving genes across the complete phylogeny or in particular branches
- This is a very conservative test as the bulk of non-synonymous sites may still be neutral or deleterious even when beneficial substitutions do occur

5.2.1 Rate of synonymous and non-synonymous substitutions

- A classic example of adaptive evolution and d_N/d_S > 1 is the lysozyme protein in primates
- This protein is important for the breakdown of bacterial walls and shows rapid evolution on the lineages leading to apes and the Colobine monkeys
- Colobines eat leaves and digest them through bacterial fermentation and then breakdown bacteria to extract the energy using lysozymes
- Colobines have convergently evolved this ability with cows and Hoatzins (a leaf-eating bird)





5.2.1 Rate of synonymous and non-synonymous substitutions

- As mentioned above, d_N/d_S is a very conservative test for identifying adaptive evolution
- We can improve sensitivity by correcting for the amount of constraint we see on non-synonymous sites at a particular locus in polymorphism (within-population) data
- If we see very little non-synonymous variation and more synonymous variation, even in polymorphism data, the gene is likely under constraint
- McDonald and Kreitman devised a test based on this (often referred to as the “MK” test)

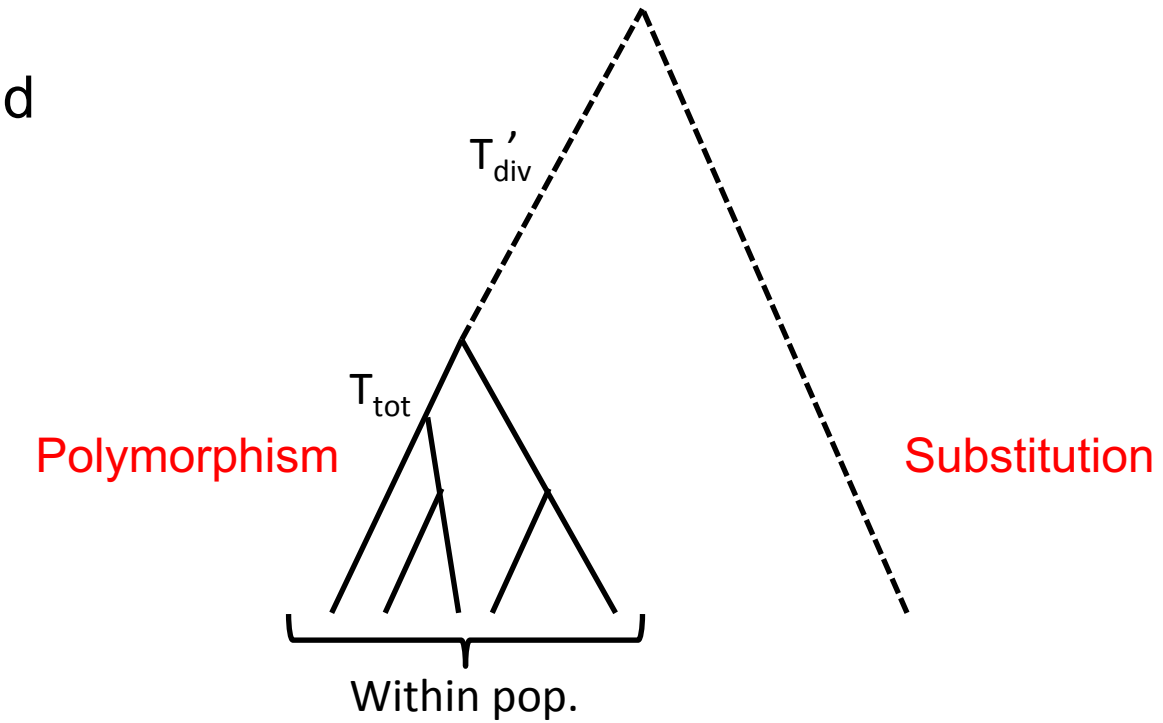
5.2.1 Rate of synonymous and non-synonymous substitutions

- The MK test partitions polymorphism and substitution data into non-synonymous and synonymous sites:

	Poly.	Fixed
Non-Syn.	P_N	D_N
Syn.	P_S	D_S
Ratio	P_N/P_S	D_N/D_S

- Given the phylogeny to the right, our expectations for polymorphism and substitutions become:

	Poly.	Fixed
Non-Syn.	$\mu L_N(1-C)T_{tot}$	$\mu L_N(1-C)T'_{div}$
Syn.	$\mu L_S T_{tot}$	$\mu L_S T'_{div}$
Ratio	$L_N(1-C)/(L_S)$	$L_N(1-C)/(L_S)$



5.2.1 Rate of synonymous and non-synonymous substitutions

- Under a strict neutral model, we expect the ratio of non-synonymous to synonymous variants to be the same for both polymorphism and substitution data
- We can determine if these are equivalent with a 2x2 test of our table
- If the ratio of $\frac{N}{S}$ is higher for substitution than polymorphism data, we have evidence that, after accounting for constraint, non-synonymous substitutions are accumulating more rapidly than might be expected

5.2.1 Rate of synonymous and non-synonymous substitutions

- As an application of the MK test, Frentiu and colleagues (2007) looked at evolution of opsin genes, responsible for color vision, in two butterfly species

- Adaptive substitutions relative to polymorphism constraint was assessed in an MK test:

	Poly.	Fixed
Non-Syn.	2	12
Syn.	9	4
Ratio	$^2/_9$	$^3/_1$

- There is a strong excess of non-synonymous substitutions