

4. Evolutionary Genetics

Read: Futuyma: Chap. 22 pp. 625-635 *except* section titled **Phylogenetic Insights**.
Graur & Li: Chap. 3 discusses alignment and correcting for multiple hits in much more detail than we will need; Chap. 4 pp. 99-123 (*except* section titled **Similarity profiles**); pp. 139-142; 146-163.

Note: Nancy Moran will be lecturing on molecular evolution, so we will only touch on some basic theory and phenomena. You'll probably want to read some of the above in more detail for her section of the course.

Speciation and the Transition from Population Genetics to Evolutionary Genetics

This section is concerned with long-term evolution, i.e. evolutionary divergence of species.

Some of the basic factors that determine variation within a species also control divergence of different species: mutation, drift, selection, sex. Difference is time scale.

Diagram below shows gene lineages during speciation in asexual haploid organism (or genome, e.g. animal mitochondria). When speciation begins, e.g. with separation of population into two subpopulations by permanent geographic barrier, the two subpopulations include individuals that are more closely related to individuals from the other clade (polyphyletic, paraphyletic). When speciation is complete, they are reciprocally monophyletic and are distinct clades. Lineage sorting is complete.

This process takes about $4N_e$ generations in a haploid population.

Measuring evolutionary divergence at the sequence level

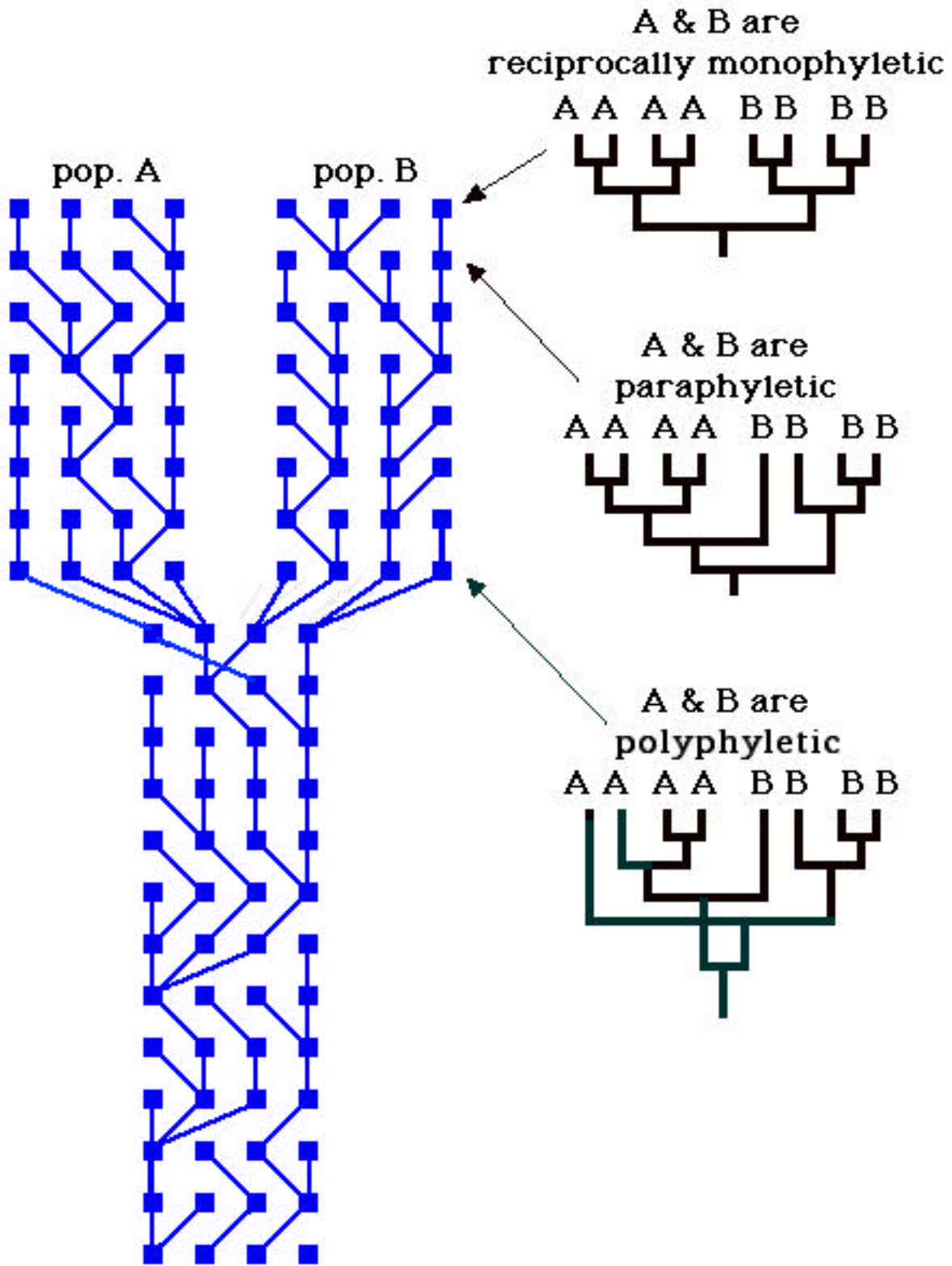
- (1) For each species, sequence one copy of a gene.
- (2) Align the sequences; this is usually done with computer assistance.
- (3) Calculate the sequence difference

sequence difference = d = # sites with mismatches / total sites

sequence identity = $1 - d$ = number of sites without mismatches / total sites

A site is occupied by a base or amino acid in both sequences; gaps are ignored!

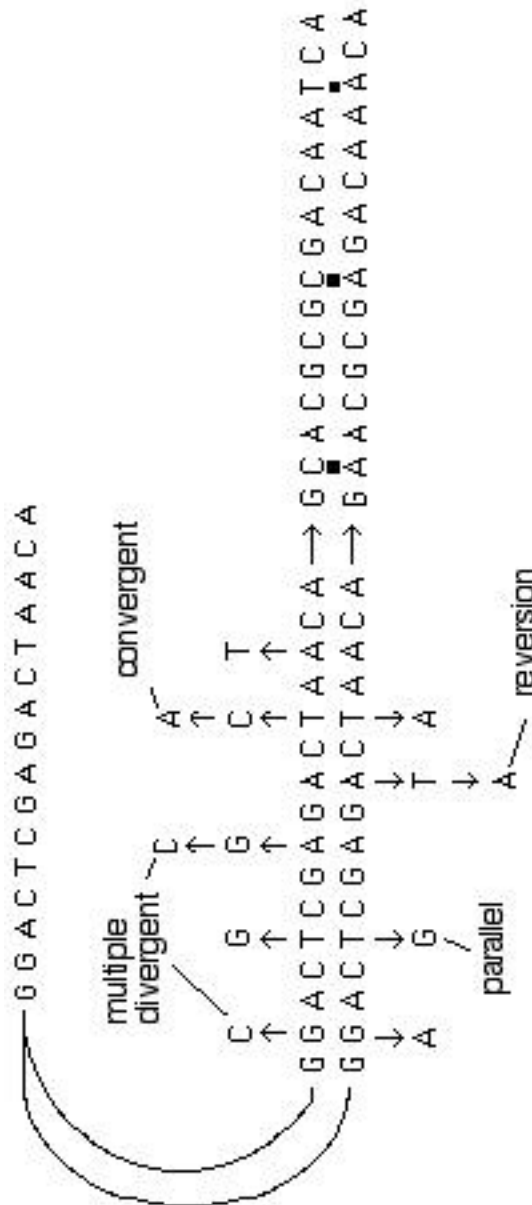
sequence similarity = proportion of sites where the amino acids are identical *or* are believed to be selectively equivalent & interchangeable because they have similar physical properties *or* are, in fact, often substituted



(4) Calculate the sequence divergence = K = distance corrected for multiple hits according to an evolutionary model. The simplest is the Jukes-Cantor model (see Bruce Walsh's lecture notes). For base sequences:

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3}d\right)$$

Often the above calculations are done separately for synonymous and nonsynonymous substitutions, or for 3rd codon position vs. 1st and 2nd codon positions.



Calculating evolutionary rates

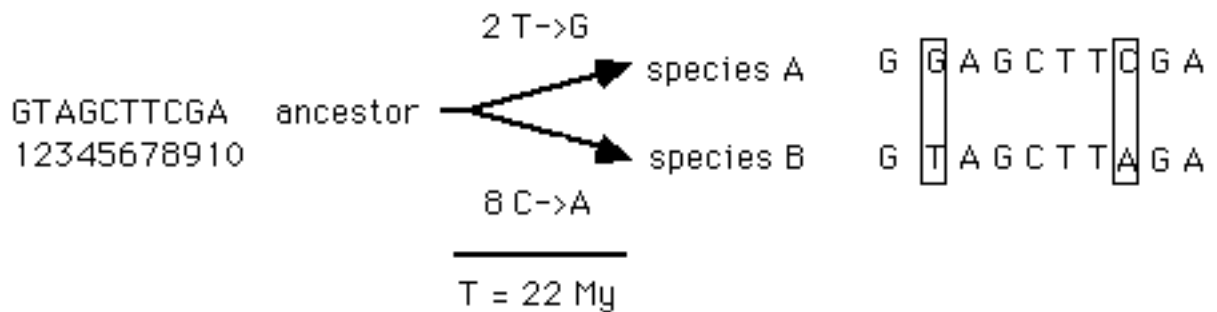
If T is the species divergence time, the rate of evolution = $E = K/2T$

Usually, K is in bp substitutions per site (or per bp), and T is in years, so E is in bp substitutions per site per year.

A note on dimensional analysis: Don't forget that bp substitutions per site per year is the same as bp sub's/site \times years or bp sub's \times site⁻¹ \times years⁻¹, *not* bp sub's/site/year. This is important when you are doing dimensional analysis, i.e. writing out expressions or equations with all dimensions to determine the dimensions of the final result. For example, in this case we have K in bp sub's/site and T in years, so

$$E = \frac{\text{bp subs}}{\text{sites}} \times \frac{1}{\text{years}} = \frac{\text{bp subs}}{\text{sites} \times \text{years}}$$

A diagrammatic summary:



$$d = 2/10 = 0.2 \quad K = - (3/4) \ln [1 - (4/3)d] = 0.23$$

$$E = K/2T = 0.23/2 \times 22 \text{ My} = 5.3 \times 10^{-9} \text{ bp sub's per site per year}$$

The best data on E come from vertebrates, for which we have a good fossil record. But even there the estimates have a very large uncertainty. Plants and invertebrates have a poor fossil record; microorganisms have almost none.

bdelloid rotifers in amber 40 Mya

$$K_s = 0.53 \quad T = 40 \times 10^6 \quad E = K_s/2T = 6.6 \times 10^{-9}$$

Sometimes we can get T from biogeography, or estimate T for a parasite from T for its host.

Representative values and variation (see texts for more examples)

Evolutionary rates of base pair substitution are generally order of 10^{-10} - 10^{-7} , except for RNA viruses which evolve at least three orders of magnitude faster.

Variation in evolutionary rates parallels variation in diversity. (All the following rates are in bp substitutions per site per year unless otherwise stated.)

Evolutionary rates vary among organisms and lineages

| | |
|----------|--|
| rodents | 7.9×10^{-9} |
| primates | $1.3 \times 10^{-9}, 2.2 \times 10^{-9}$ |

Evolutionary rates vary among genomes (following are synonymous substitution rates)

Plants:

| | |
|---------------|----------------------|
| nuclear | 6×10^{-9} |
| chloroplast | 2×10^{-9} |
| mitochondrial | 0.6×10^{-9} |

Primates:

| | |
|---------------|---------------------|
| mitochondrial | 20×10^{-9} |
| nuclear | 5×10^{-9} |

Evolutionary rates vary among genes and other sequences

vertebrates, amino acid sequence divergences:

| | |
|-----------------|----------------------|
| fibrinopeptides | 4×10^{-9} |
| hemoglobins | 1×10^{-9} |
| cytochrome c | 0.2×10^{-9} |

Pseudogenes evolve faster than genes:

mammals:

| | |
|----------------------|-----------------------|
| pseudogenes | 4.85×10^{-9} |
| protein coding genes | 1.81×10^{-9} |
| 18S rRNA genes | 0.04×10^{-9} |
| 28S rRNA genes | 0.50×10^{-9} |

(Note the data for these genes come from different species pairs and are not exactly comparable.)

Evolutionary rates vary among regions of a gene

mammals:

| | |
|-----------------------------|-----------------------|
| exons (total) | 1.81×10^{-9} |
| synonymous substitutions | 4.65×10^{-9} |
| nonsynonymous substitutions | 0.88×10^{-9} |

Explaining evolutionary rates and patterns

Forces are mutation, drift, and directional selection.

Balancing selection usually doesn't last very long after speciation (exception: MHL in humans and chimpanzees).

Key point: only mutations that are fixed will contribute to differences between species.

$$E = 2NuF$$

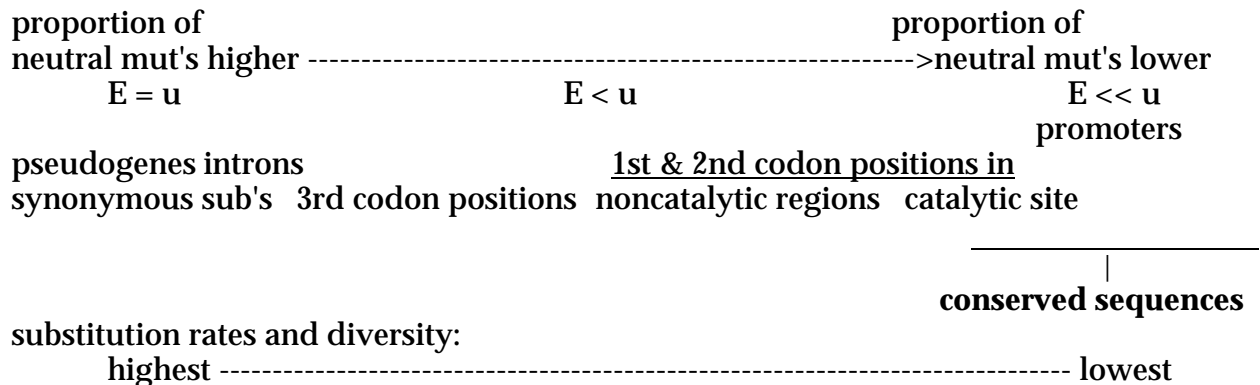
$$\text{fixations/site} = (\text{mutations/site} \times \text{generation})(\text{fixations/mutation})$$

$$\text{neutral mutations in diploids: } F = 1/2N \quad E = u$$

Substitution rate in pseudogenes and synonymous substitutions can be used as an estimate of u . Often easier and more accurate than direct measurement.

Synonymous substitutions in highly expressed genes are subject to weak selection.

Summary of relationship between selection and substitution rates and diversity:



Positive selection for advantageous mutations can be detected in evolutionary studies

In a gene: K_s/K_a ratio

At specific sites:

- Compensating substitutions
- Convergent and parallel evolution

Molecular Clocks

In spite of all the above variations, evolutionary rates can be approximately constant over long periods of time when one looks at a single gene. If one can calculate E for a gene in some lineage, one can estimate the time since divergence of two other lineages by solving $E = K/2T$ for $T = K/2E$. Some people have pointed to cases of extreme divergence (e.g. a few cases of hundred-fold differences in E), or just the high variance in E , and argued that there is no clock. There is; the only question is whether it is good enough for its intended use; the answer will be different in different cases. Also in many cases it is the only game in town.

The really interesting and probably very deep question is: why is the rate of evolution even remotely constant? Why doesn't it vary by a factor of 10^3 or 10^4 ? For E to be constant requires that 4 parameters be constant or vary in a nicely compensating fashion: N , N_e , u , and s .

Sex and Evolution

Hitchhiking and background selection effects don't affect long-term *neutral* substitution rates. But the fixation probability of selected mutations at a site is affected by selected mutations segregating in the background, especially if they are linked to the site (Hill-Robertson effect).

Simplistic example: consider two loci with two alleles: A1 and B1 are advantageous, A2 and B2 are detrimental. In asexual population, starting with genotype A1 B1, mutations occur (usually in different individuals), producing genotypes A2 B1 and A1 B2. Now selection against A2 is confounded by selection for B1, and selection against B2 is confounded by selection for A1.

Result is that detrimental mutations accumulate (Muller's ratchet in soft selection model; meltdown in hard selection model), and advantageous mutations are more likely to be lost.

Selection on background acts like noise interfering with selection at site. Can be viewed as reduced N_e .

Sex counteracts effect by replacing some A1 B1 and A1 B2 with A1 B1 and A2 B2.

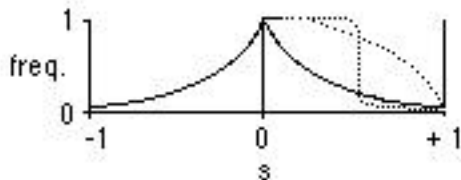
Most general genetic advantage of sex: Natural selection works better with sex.

Prediction: asexual lineages should have high extinction rate and low speciation rate.

A note on mutational load and escaping from meltdown:

In a surviving lineage in an asexual organism, does # compensating substitutions = # detrimental substitutions?

Not necessarily: don't know frequency distribution of s values.



One strongly selected advantageous substitution might compensate for several weakly selected detrimental substitutions.

My intuition: indefinite survival of a lineage requires $u_a s_a = u_d s_d$ where s is the mean selection coefficient, u is the mutation rate, and subscripts a and d refer to advantageous and detrimental mutations.

Another point: Fitness is quantitative multilocus trait. Some forms of epistasis allow selection to act on many detrimental mutations at the same time.