Question:

Is molecular evolution driven more by selection,

or is it driven more by mutation and drift?

Natural selection seems dominant mechanism ➔ phenotypic evolution
(e.g. morphology, physiology, behavior, life history …)

Hypothesis 1: Evolution of molecules (DNA, protein sequences) should also be driven by natural selection

Hypothesis 2: Random processes (genetic drift …) are more important than natural selection in molecular evolution

Haldane's paradox:

- Early population geneticists believed that most polymorphisms are maintained by *balancing selection* (e.g. selection against homozygotes, or alternatively, frequency-dependent selection)
- Balancing selection implies a "genetic load" for the population, because homozygotes are less fit than heterozygotes. (Genetic load = allele diversity that a population carries along, possibly even reducing fitness if not all individuals have the same allele)
- When protein electrophoresis became available, it was found that a very large number of genes are actually polymorphic. This appeared to imply an unacceptably high genetic load for human and other populations. It was difficult to explain why selection would favor polymorphisms (and a high genetic load) at the expense of fitness.

Natural selection seems dominant mechanism ➔ phenotypic evolution
(e.g. morphology, physiology, behavior, life history …)

Hypothesis 1: Evolution of molecules (DNA, protein sequences) should also be driven by natural selection

Hypothesis 2: Random processes (genetic drift …) are more important than natural selection in molecular evolution

Relevant Evidence:

Observation 1. Genome size is extremely variable among species

Observation 2. Molecular evolution is "decoupled" from morphological evolution

Observation 3. Many genes have uniform evolutionary rates over evolutionary time

Obs. 1: Genome Size Varies Among Species

Hypothesis (1950s-60s) - DNA content should be correlated with phenotypic and developmental complexity

Total DNA = C value
= DNA in haploid (1N) genome - in kilobases (kb)

N allows all comparisons:

- haploid, diploid, polyploid spp.
- sexual, asexual spp.
- variable ploidy spp.
  N vs. 2N (gametophyte vs. sporophyte; parasites)

C-value paradox

- Amount of DNA present in the genome seems unrelated to the complexity of the organism.

Species	C value (kb)
Navicola pelliculosa (diatom)	35,000
Drosophila melanogaster (fruitfly)	180,000
Paramecium aurelia (ciliate)	190,000
Gallus domesticus (chicken)	1,200,000
Erysiphe cichoracearum (fungus)	1,500,000
Cyprinus carpio (carp)	1,700,000
Lampreta planeri (lamprey)	1,900,000
Boa constrictor (snake)	2,100,000
Parascaris equorum (roundworm)	2,500,000
Carcarias obscurus (shark)	2,700,000
Rattus norvegicus (rat)	2,900,000
Xenopus laevis (toad)	3,100,000
***Homo sapiens* (human)**	**3,400,000**
Nicotiana tabaccum (tobacco)	3,800,000
Paramecium caudatum (ciliate)	8,600,000
Schistocerca gregaria (locust)	9,300,000
Allium cepa (onion)	18,000,000
Coscinodiscus asteromphalus (diatom)	25,000,000
Lilium formosanum (lily)	36,000,000
Amphiuma means (newt)	84,000,000
Pinus resinosa (pine)	68,000,000
Protopterus aethiopicus (lungfish)	140,000,000
Ophioglossum petiolatum (fern)	160,000,000
Amoeba proteus (amoeba)	290,000,000
Amoeba dubia (amoeba)	670,000,000

Compiled by Li and Graur (1991) from Cavalier-Smith (1985), Sparrow et al. (1972), and other references. The C value for humans is highlighted for reference.

Obs. 1: Genome Size Varies Among Species

Genus	(Common name)	C value (kb)
Navicola	(diatom)	35,000
Drosophila	(fruitfly)	180,000
Gallus	(chicken)	1,200,000
Cyprinus	(carp)	1,700,000
Boa	(snake)	2,100,000
Rattus	(rat)	2,900,000
Homo	(human)	3,400,000
Schistocerca	(locust)	9,300,000
Allium	(onion)	18,000,000
Lilium	(lily)	36,000,000
Ophioglossum	(fern)	160,000,000
Amoeba	(amoeba)	290,000,000

"Complex" vertebrates low
"Simple" plants 30X "complex" vertebrates
"Primitive" insects exceed "advanced" vertebrate span
"Primitive" fern 6X "advanced" angiosperms
"Simple" unicellular species span entire range

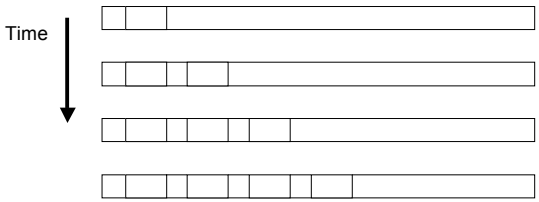
➔ "C value paradox"

Repetitive DNA

Short non-coding DNA sequences that increase in number, usually due to their ability to "jump around" the genome (e.g. transposons)

Also called "junk" DNA

Repetitive DNA



Example: Up to 10% of the human genome consists of 500,000 or more copies of *ALU* sequences, which have no function other than to assist their own replication

Alu elements are approximately 300 bp in length and derive their name from a single recognition site for the restriction enzyme *Alu*I

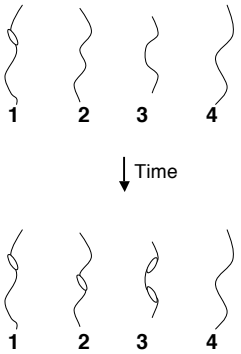
Alu is an example of a so-called "jumping gene" -- a transposable DNA sequence that "reproduces" by copying itself and inserting into new chromosome locations.

Each *Alu* element has an internal promoter for RNA polymerase III needed to independently initiate transcription of itself. The inserted *Alu* is transcribed into messenger RNA by the cellular RNA polymerase. Then, the mRNA is converted to a double-stranded DNA molecule by reverse transcriptase. Finally, the DNA copy of *Alu* is integrated into a new chromosomal locus at the site of a single- or double-stranded break. As this process repeats, the genome accumulates more *Alu* elements

Transposable genetic elements

- Genetic sequences that are able to replicate and insert into any position in the genome.
- Cost of carrying these elements seems to be quite small, although some may be mutagenic
- Rarely, if ever, do they confer an individual benefit
- Help to explain C-value paradox

Transposons



Transposons

Transposons in *Drosophila*:

- Make up 10% of genome
- Between 500 to 1,000 transposons in each individual
- Transposon movement is a major source of deleterious mutations, if they insert themselves into gene coding sequence

Obs. 2 - Decoupled Molecular & Morphological Evolution

Four possible outcomes comparing phenotypic (e.g. morphology) and genetic similarities between two species:

Case	Phenotypic Similarity	Genetic Similarity	Expected
1	Low	Low	if DNA correlated complexity COMMON
2	High	High	if DNA correlated complexity COMMON
3	High	Low	if DNA evolves faster than phenotype COMMON
4	Low	High	if DNA evolves slower than phenotype COMMON

Case 3: Morphologically conservative spp. (frogs, salamanders, corals…)

Case 4: Great apes (man, chimpanzee) **morphologically distinct Genetically similar**

- **Chromosomes** - chimp has **1 pair** more)
- **Protein (allozyme)** - differ **<1.06%**
- **Amino acid (AA) sequence** - differ **0.39%**
- **Gene sequence** - **13,454** homologous genes
  - **29% identical** - **71%** average **2 AA, 3 silent** changes

Obs. 2 - Decoupled Molecular & Morphological Evolution

Four possible outcomes comparing phenotypic (e.g. morphology) and genetic similarities between two species:

Case	Phenotypic Similarity	Genetic Similarity	Expected
1	Low	Low	if DNA correlated complexity
2	High	High	if DNA correlated complexity
3	High	Low	if DNA evolves faster than phenotype
4	Low	High	if DNA evolves slower than phenotype

Property	Frogs	Placental mammals
Number of species	3050	4600
Number of orders	1	16-20
Age of group	150 myr	75 myr
Rate of morphological evolution	slow	fast
Rate of molecular evolution	standard	standard

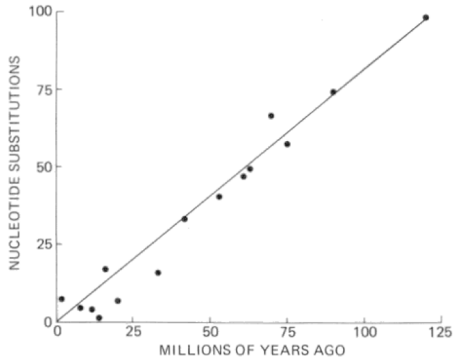
Compare frogs with placental mammals. Frogs have been around almost twice as long, yet their morphological diversity is much less than that of placental mammals, despite a fairly standard rate of molecular evolution in both.

Is the 1% - 5% difference between human and chimp genome sequences enough to account for the magnitude of phenotypic differences between them?

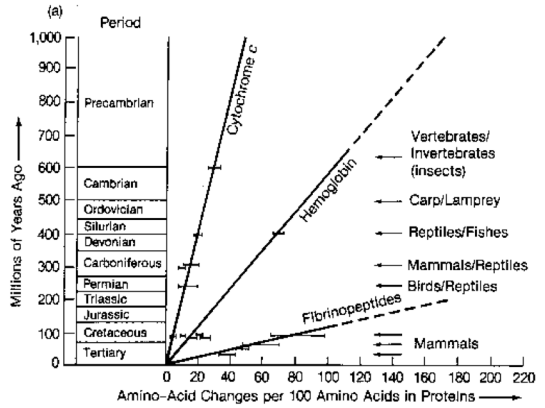
Obs. 3 - Molecular Evolutionary Rates Constant = "Molecular Clock"

Hypothesis: Each gene has characteristic, roughly constant, mutation rate - can be used to estimate time since lineages diverged

"Clocks" for different proteins "tick" at different rates, indicating that protein evolution occurs independently among genes



Inferred pairwise nucleotide substitutions among 17 mammal species from seven gene products, plotted against date of divergence as estimated from the fossil record. The line is drawn from the origin through the oldest point (marsupial / placental divergence at 125 MYBP). The strong linear relationship suggests that molecular differences between pairs of species are proportional to the time of their separation, rather than the degree of organismal difference.



However, different genes "tick" at different rates… probably due to differences in the strength of purifying selection

Obs. 3 - Molecular Evolutionary Rates Constant = "Molecular Clock"

Hypothesis: Each gene has characteristic, roughly constant, mutation rate - can be used to estimate time since lineages diverged

"Clocks" for different proteins "tick" at different rates, indicating that protein evolution occurs independently among genes

1960s - 1970s - The Great Neutralist-Selectionist Controversy

1. All evolution consists of adaptations driven by natural selection

vs.

2. Natural selection does not explain molecular evolution & "clocks"

Arguments against molecular adaptation by natural selection:

- **Phenotypic adaptation often very rapid; clocks slow**
- **Selection rates change drastically (with fitness differentials) and decline as approach fixation**
- **Independent evolution needs a separate selective pressure for each protein gene**

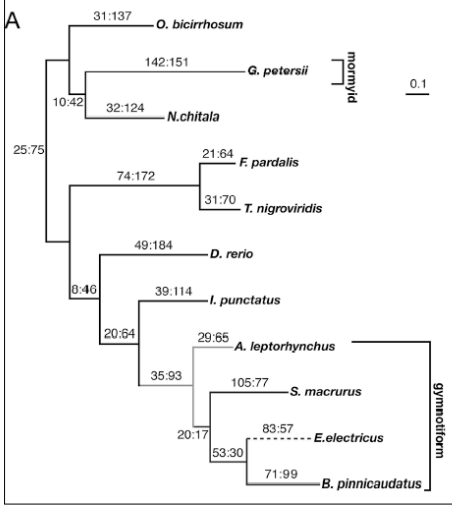
Moto Kimura's *neutral theory* of molecular evolution provides an explanation

- Claim: The large majority of observed molecular polymorphisms reflect neutral changes, and not outcomes of selection. Likewise, most substitutions observed between homologous genes are selectively neutral.

Moto Kimura's *neutral theory* of molecular evolution provides an explanation

- Claim: The large majority of observed molecular polymorphisms reflect neutral changes, and not outcomes of selection. Likewise, most substitutions observed between homologous genes are selectively neutral.

- Implications: Gene (protein) families evolve through neutral mutations and purifying selection. Drift becomes more important than selection, for molecular evolution. Most genes (proteins) have not been improved during the period of metazoan evolution.



Selection undoubtedly occurs, but as the final "layer" over a constant, ongoing foundation of neutral mutation and drift

Maximum Likelihood Models

Models that assume all nucleotides occur at equal frequencies (25%)

1. The Jukes-Cantor (JC) model
    1. All substitutions are equally likely.
    2. All nucleotides occur at the same frequency (25%).
    3. One parameter: the rate of subsitution (alpha).
2. Kimura two parameter (K2P) model
    1. Transitions and transversions happen at different rates.
    2. All nucleotides occur at the same frequency.
    3. Two parameters: transition rate (alpha) and transversion rate (beta).

Models that allow the four nucleotides to be present in different frequencies

1. Felsenstein (F84) & Hasegawa-Kishono-Yano (HKY85) models
    1. Two closely related models -- they use different calculations to model essentially the same thing
    2. Transitions and transversions occur at different rates
    3. Nucleotides occur at different frequencies
2. General time reversible (GTR) model
    1. Assumes a symmetric substitution matrix (and thus is time reversible)
    2. In other words, A changes into T with the same rate that T changes into A.
    3. Each pair of nucleotide substitutions has a different rate
    4. Nucleotides can occur at different frequencies

Conclusions and New Hypotheses

1. Most molecular evolution at the level of the whole genome must be non-adaptive (i.e. not responding to natural selection)

2. Proportion of DNA coding for functional genes is very small (<<5%)

3. Decoupling suggests different evolutionary processes must act at molecular and phenotypic levels

4. Observations suggest natural selection is a minor evolutionary process at the molecular level, and that random processes (e.g. drift) are much more important

4. "Clock-like" evolution of proteins inconsistent with way natural selection acts

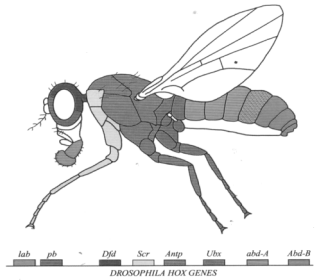
5. Random silent mutations and random drift of "neutral" alleles drive most molecular evolution

Why might selection be weaker when it comes to molecular evolution?

- Not all DNA variation results in protein variation
- Not all protein variation results in phenotype variation
- Not all phenotype variation results in changes in fitness

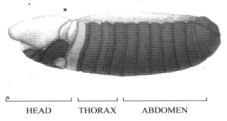
The evolution of biodiversity is probably due more to the evolution of *regulatory* relationships and gene *interactions*, rather than strict and simple evolution of gene sequences.

Morphological diversity seems to be more the result of regulatory differences during development, rather than the result of raw gene sequence differences

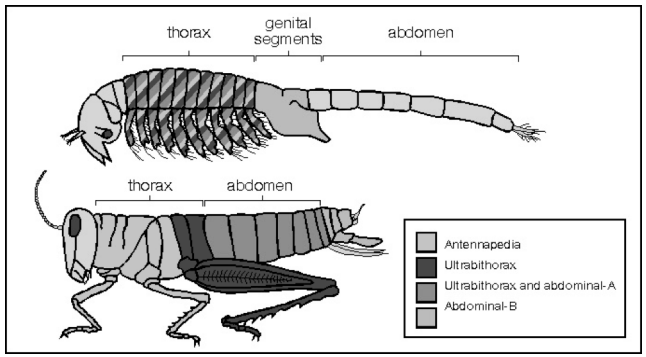


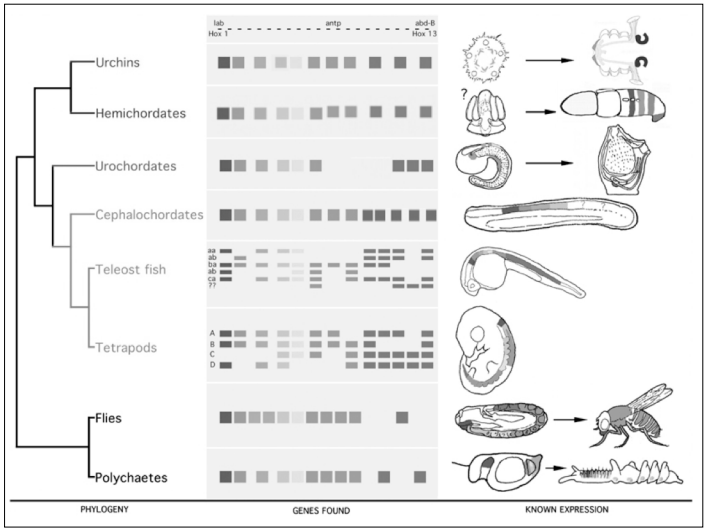
Adult

Embryo



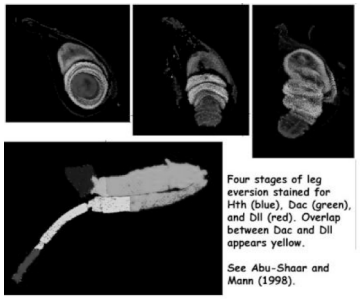
Expression patterns of *Hox* genes, and their relationships with downstream targets, are what generated body plan diversity





Evolving regulatory relationships can result in old genes having new functions (example: developmental gene *distalless* (*dll*)

dll in appendage development



Four stages of leg eversion stained for Hth (blue), Dac (green), and Dll (red). Overlap between Dac and Dll appears yellow.

See Abu-Shaar and Mann (1998).

dll in wing eyespot development