TWO CURRENT PROBLEMS IN MOLECULAR EVOLUTION:

(1) Accounting for the observed pattern of substitutions between species
   • Involves both previous pattern of common ancestry +
     evolutionary forces acting on sequences given this common ancestry.

(2) Accounting for the observed patterns of polymorphism within species

DYNAMICS OF ALLELE FREQUENCY CHANGE: SELECTION IN AN INFINITE
POPULATION

**Directional selection** (one homozygote has the most fit genotype):

Removes variation and generates an allelic substitution.

Simplest case: two alleles, A and a and **additive selection**:

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1+s</td>
<td>1+2s</td>
</tr>
</tbody>
</table>

if we let \( p = \) frequency (a),
\( p \to 1 \) if \( s > 0 \) (allele a is fixed),
\( p \to 0 \) if \( s < 0 \) (allele A is fixed).

**Balancing Selection**:

--- The genotype with the highest fitness is a heterozygote and all other
---- genotypes have lower fitness

--- Maintains variation and prevents allelic substitution.

The simplest form is **overdominant selection**

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1+s</td>
<td>1</td>
</tr>
</tbody>
</table>

Here, if \( s > 0 \), both alleles are maintained by selection (selectively maintained
polymorphism).

**Key points for selection in an infinite population:**

--- Unmeasurably small selection coefficients can cause substitutions (directional
---- selection) or can maintain a polymorphism (balancing selection)

--- Different types of selection are required to account for allelic substitutions and
---- the maintenance of polymorphisms.

--- Thus, effects that cannot be measured in the lab are nonetheless evolutionarily
critical
DYNAMICS OF ALLELE FREQUENCY CHANGE: DRIFT

KEY: Due to random sampling of gametes to form the new generation, allele frequency changes each generation, so that eventually an allele is either fixed or lost.

Example 1: Selfing a heterozygote. After one generation, 1/2 of the offspring are fixed (half for A, the other half for a).

Example 2: A sexual population of size two. Here, we start with two heterozygotes. After one generation, 1/8 of the resulting populations are expected to be fixed.
In general, the Wright-Fisher model of genetic drift predicts:

\[
\text{Prob( } k \text{ copies of A in generation 1 given there are } i \text{ copies in generation 0)} =
\]

\[
\text{Prob( } X(1) = k \mid x(0) = i) = \binom{i}{2N}^k \left(1 - \frac{i}{2N}\right)^{2N-k} \frac{(2N)!}{k! (2N-k)!}
\]  

(1)

where \( k! = k*(k-1)*...*1 \). This is just a binomial random variable with sample size \( 2N \) and success parameter \( p = i/(2N) \).

Let \( U(p) = \) the probability that an allele is fixed given it starts at initial frequency \( p \). For a neutral allele, iterating (1) shows that \( U(p) = p \) (the probability of fixation simply equals its starting frequency).

**Key points:**

---\( \rightarrow \) This repeated sampling results in the fixation of a randomly-chosen allele.

---\( \rightarrow \) Drift, by itself, removes genetic variation.

**Coalescents**

The method of describing the *coalescent* provides a very powerful way for developing sampling theory and statistical tests of neutrality. As the following figure shows, the coalescent is simply the pattern of gene genealogies. Any two alleles can be traced back to a most recent common ancestor, which is called the *coalescent time* for those two alleles. In the figure, the coalescent time for all alleles in this population is 4 generations.
One of the most powerful features associated with the coalescent approach is that if two sequences have a most recent common ancestor $\tau$ generations ago, then the expected number of mutants the two sequences differ by is Poisson distributed with parameter $2\mu\tau$, e.g.,

$$\text{Prob}[\text{sequences differ by } k \text{ mutants}] = \frac{(2\mu\tau)^k e^{-(2\mu\tau)}}{k!} \quad (2)$$

Thus, these sequences differ by an average of $2\mu\tau$ mutants. By knowing the expected distribution of times back to common ancestor, the expected number of mutant differences between sampled alleles can easily be computed.

**Mutation models:**

- **Infinite allele model** (each new mutation is unique)
- **Infinite site model** (each new mutation gives a new segregating site)

\[
\begin{array}{cccccc}
A & A & G & C & C & G & A & \text{allele 1} \\
& \bullet & \bullet & \bullet & \bullet & \\
A & A & C & A & C & A & G & C & \text{allele 2}
\end{array}
\]

The difference between these model is shown by the above sequence data in which there are two alleles (different sequences), but these differ at four segregating sites

- **Step-wise model** (for tandem arrays)

\[
\begin{array}{cccccccc}
AA & AAA & AAAA & AAAA & \text{A}_5 \\
\text{A}_5 & \text{A}_4 & \text{A}_5
\end{array}
\]
The coalescent structure -- depends on $N_e$ and selection

Infinite alleles mutational model

Current alleles: $m, m_2, m_3, m_4$

Infinite sites mutational model

Current alleles: $m_{101}, m_{111}, m_{000}, m_{001}$

Stepwise mutation model

Current alleles: $A_3, A_4, A_5, A_5$
**INTERACTION OF DRIFT AND MUTATION**

**Overview**: Mutation pumps in variation, drift removes it. The interaction of these two evolutionary forces can account for both the maintenance of polymorphisms (standing genetic variability) and the substitution of new alleles.

Let $\mu = \text{the mutation rate}$, and $N$ the idealized population size (more generally, $N$ is replaced by $N_e$, the effective population size).

**Maintaining Polymorphisms**:

A typical pattern of allele frequency change (under drift and mutation) looks like this (over a short time scale)

Here, alleles $a$ and $b$ are initially present, while alleles $c$ and $d$ arise by mutation. At time 1, alleles $a$ and $b$ are segregating; at time 2, $b$ and $c$ are segregating; at time 3 $b$, $c$, and $d$ are all segregating. Note that the average level of heterozygosity remains roughly constant over time, but that the actual alleles (and their frequencies) accounting for this heterozygosity changes over time.

The average expected heterozygosity (at equilibrium) is given by

$$H = \frac{4N\mu}{1 + 4N\mu} \tag{3}$$
which is plotted below.

Note that when $4N\mu >> 1$, the population is highly polymorphic ($H \approx 1$). Since the expected time back to the most recent ancestor for two randomly chosen alleles in the population (the expected coalescent time for the entire population) is $2N$, two randomly-chosen sequences differ (on average) by $2\mu \tau = 4N\mu$ mutants. When $4N\mu >> 1$, two randomly-chosen alleles differ (on average) by at least one mutant and hence are highly polymorphic.

**The Substitution of New Alleles:**

Over longer time scales (plotting only those alleles that go to fixation), we need to distinguish between the **time to fixation** and **the time between appearance of new mutants** destined to become fixed.

1. The expected time between the initial appearance of new mutants destined to become fixed is $\mu^{-1}$ generations.
2. The expected time to fix those mutants that are destined to become fixed is $\approx 4N$ generations.
3. The expected time an allele remains in the population following its introduction as a single copy (most are lost) is $\approx \ln(N)$ generations.
For example, if $\mu = 10^{-7}$ and $N = 10^5$, then the expected time between mutants that are fixed is $10^7$ generations, with each such mutant taking (on average) about 400,000 generations to become fixed. The majority of new mutants ($1 - 1/(2 \times 10^5) = 0.99999$) are lost. Each of these persists in the population for an average of $\ln(10^5) = 11.5$ generations.

**Key:** the interaction of drift and mutation generates both polymorphism and substitution of alleles.

### Interaction of Selection and Drift

**Keys:**

--->
When population size is finite, all alleles are either lost or fixed, and selectively advantageous alleles can become lost and deleterious alleles can become fixed. Thus, one measure to consider is the probability of fixation, $U(p)$.

--->
As population size decreases, drift becomes stronger. Thus, an allele may be effectively neutral in one population and yet strongly selected in another (larger) population.

For additive selection, Kimura found

$$U(p) = \frac{1 - \exp(-4Nes)}{1 - \exp(-4Nes)}$$

(4)

Since a new mutant appears initially as a single copy, $p = 1/(2N)$,

$$U([2N]^{-1}) \approx \frac{[2N]^{-1}}{2N} \quad \text{if } 4Ns << 1 \quad (5a)$$

$$U([2N]^{-1}) \approx 2s \quad \text{if } 4Ns >> 1 \quad (5b)$$

$$U([2N]^{-1}) \approx -2s \exp(4Ns) = 0 \quad \text{if } 4Ns << -1 \quad (5c)$$

**Key points from Equation (5)**

--->
Selection is important if $4Ns >> 1$.

--->
Deleterious alleles can be fixed by chance, but this is very unlikely if $4Ns << -1$.

--->
Even favored alleles are usually lost. From (5b), if $s = 0.05$ (a 5% selective advantage, a huge amount in evolutionary terms) then the probability that a single new mutant becomes fixed is only 10%, no matter how large the population (note 5b is independent of $N$).

**The Expected Substitution Rate $k$**

$$k = E[\# \text{ of new mutants per generation}] \times \text{Prob}[\text{such a mutant is fixed}]$$

hence,

$$k = 2N\mu \ U([2N]^{-1})$$

(6)

For selectively neutral alleles, $U([2N]^{-1}) = [2N]^{-1}$, hence $k = 2N\mu \times [2N]^{-1} = \mu$. 
For a selectively favored allele (from 5a), \( U([2N]^{-1}) = 2s \), hence \( k = 4N\mu_s \).

**Keys:**

--- For neutral alleles, the expected substitution rate is independent of the population size and depends only on the mutation rate. Hence \( k \) is expected to be fairly constant over very different species. This provides one explanation for molecular clocks.

--- For favored alleles, the expected substitution rate increases linearly with both \( s \) and \( N \). Hence, \( k \) is expected to vary widely even among different genes in the same species.

**THE NEUTRAL THEORY OF MOLECULAR EVOLUTION**

Kimura (1968, reviewed in his 1983 book)

The total mutation rate (\( \mu_T \)) = rate of effective neutral mutations (\( \mu_n \)) + rate of deleterious mutants (\( \mu_d \)) + rate of advantageous mutants (\( \mu_a \)), viz

\[
\mu_T = \mu_n + \mu_d + \mu_a.
\]

**Key assumptions:**

--- Most new mutants are either neutral or deleterious, so that \( \mu_T = \mu_n + \mu_d \).

\( \mu_a \) is not zero, but is so small relative to \( \mu_T \) that we can ignore it.

--- Therefore, essentially all substitutions and polymorphisms are due to effectively neutral mutations, so that \( k = \mu_n \) and \( H = 4N\mu_n / (1+4N\mu_n) \).

--- Thus, the neutral theory assumes that **strong selection** is operating, by constantly removing deleterious mutations.

**Support:**

--- Predicts a molecular clock as \( k = \mu \) under the neutral model. Molecular clocks are commonly seen in sequence data.

--- Predicts that regions of the genome under no selection should evolve the fastest (as these regions have the fewest selective constraints). In accordance with theory, 3rd base positions evolve faster than 1st and 2nd base positions, introns evolve even faster still, and pseudogenes evolve the fastest.

**Problems:**

--- Clocks should run on **generations**, but many protein-sequence clocks run on **years**.

--- The clock is more variable than predicted from the neutral theory.

--- Polymorphism data is not well explained.

--- Too many rare alleles.

--- The range of observed heterozygosity is too narrow.
USING DNA DATA TO DETECT SIGNALS OF BALANCING AND DIRECTIONAL SELECTION

The key idea here follows from considerations of the coalescent: the "deeper" the coalescent (longer time back to the most recent common ancestor), the more variation is expected to arise from mutation. Following synonymous 3rd-base substitutions (which are assumed, at least to first-order, to be effectively neutral), we can look for signatures of past (and present) selection at linked sites by examining the standing level of variability. The figure below shows the expected structures of three coalescents: A: a neutral allele, B: an allele being substituted in the population by selection (this substitution occurs fairly rapidly relative to the expected time to fixation under drift), and C: two alleles under balancing selection. While some neutral divergence is expected to arise within both classes of alleles, the expected time back to the common ancestor of both can be quite long. Some MHC alleles show shared polymorphisms within species over millions of years.

Likewise, sites tightly linked to these selected alleles will show similar coalescents. In regions tightly linked to a site that has recently been fixed due to directional selection, a selective sweep occurs, removing most variation (because all alleles have a very recent common ancestor). Signals that directional selection may be common are seen in regions of Drosophila with reduced recombination, such as near the chromosome telomeres. In such regions, the amount of 3rd-base polymorphism is rather sharply reduced relative to similar polymorphisms at chromosome regions with higher recombination.

CLOCKS AND THE EXPECTED AMOUNT OF SEQUENCE DIVERGENCE

A problem with estimated the expected number of substitutions between alleles from two different species is the problem of correcting for multiple substitutions.

Note that by comparing the two current sequences, we would only observe one difference (since the ancestral sequence is not known), when in fact two substitutions have occurred.
The Jukes-Cantor model:

The expected amount of divergence at the nucleotide level (assuming neutral alleles) has been examined theoretically under a number of mutational models. The simplest is the Juke-Cantor model, which assumes that each nucleotide has an equal chance of mutating to each other nucleotide (i.e., assumes that the transition and transversion rates are the same). Here, the expected divergence

\[ d = \frac{3}{4} \left[ 1 - \exp \left( -\frac{8}{3} \mu t \right) \right] \]  

which is plotted in the figure below. If we know the mutation rate \( \mu \), then we can estimate the expected divergence time by

\[ \hat{t} = \mu^{-1} \left( -\frac{3}{8} \ln \left( 1 - \frac{4}{3} \hat{d} \right) \right) \]

This figure shows the importance of picking the molecule with the right clock speed is critical if we are trying to estimate the expected divergence time. If the amount of divergence is too small, or too large, it is very difficult to estimate \( t \) (due to sampling effects). For example, if \( t > 1.25 \mu \) (for the uppermost curve, this corresponds to \( t / \mu^* = 0.5 \)), then \( d > 0.72 \). Due to sampling effects, this amount of divergence is very difficult to distinguish from complete divergence (\( d = 0.75 \)) and hence this amount of divergence is consistent with any time from \( t = 1.25\mu \) to \( t = \infty \). The best power occurs when \( d = 0.1 \) to \( 0.6 \) (or \( t = 0.05/\mu \) to \( 0.6/\mu \)).
REFERENCES:


----------

Bruce Walsh
Ecology and Evolutionary Biology
University of Arizona
jbwalsh@ccit.arizona.edu
http://nitro.biosci.arizona.edu