Molecular evolution: traditional tests of neutrality

Guest lecture by Dr. Daniel Neafsey
Research Scientist, Broad Institute

Mutation+Selection=Evolution
Relative importance of each for maintaining variation in population?

Early Criticism of Darwin
Blending inheritance, ‘gemmales’

Fleeming Jenkin (1867):
\[ \text{Var}[X(t+1)] = \frac{1}{2} \text{Var}[X(t)] \]

Mendelian Inheritance
published 1865-66, rediscovered 1900

Law of Segregation:
- allelic variation
- offspring receive 1 allele from each parent
- dominance/recessivity
- parental alleles ‘segregate’ to form gametes

Law of Independent Assortment

Simple case: no selection

The Hardy-Weinberg Law (1908)

Requires:
- infinite population size
- random mating
- non-overlapping generations
- no selection, mutation, or migration

The Hardy-Weinberg Law

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency at time 0:</td>
<td>( u_0 )</td>
<td>( v_0 )</td>
<td>( w_0 )</td>
</tr>
</tbody>
</table>

\[ u_0 + v_0 + w_0 = 1 \]

frequency of \( A \) (\( p_0 \)) = \( u_0 + v_0/2 \)

frequency of \( a \) (\( q_0 \)) = \( w_0 + v_0/2 \)

\[ p_0 + q_0 = 1 \]
The Hardy-Weinberg Law

Genotype: AA Aa aa
Frequency at time 0: $u_0 \ v_0 \ w_0$

<table>
<thead>
<tr>
<th>Mating Pair</th>
<th>Frequency Offspring</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA x AA</td>
<td>$u_0^2$ 1 0 0</td>
</tr>
<tr>
<td>AA x Aa $u_0v_0$</td>
<td>$\frac{1}{2}$ $\frac{1}{2}$ 0</td>
</tr>
<tr>
<td>Aa x AA $u_0v_0$</td>
<td>$\frac{1}{2}$ $\frac{1}{2}$ 0</td>
</tr>
<tr>
<td>Aa x Aa</td>
<td>$v_0^2$ $\frac{1}{4}$ $\frac{1}{2}$ $\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Frequency of AA in next generation:

\[ u_1 = u_0^2 + u_0v_0 + \frac{1}{4}v_0^2 \]
\[ = (u_0 + \frac{v_0}{2})^2 \]
\[ = p_0^2 \]

The Hardy-Weinberg Law

If assumptions met:

- allele frequencies don’t change
- after a single generation of random mating, allele frequencies are:
  \[ u = p^2 \]
  \[ v = 2pq \]
  \[ w = q^2 \]
- entire system characterized by one parameter ($p$)

Deviation from expectations indicates failure of 1 or more assumptions—selection?

HW application: Sickle cell anemia

<table>
<thead>
<tr>
<th>Observed Counts</th>
<th>Expected Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS 834</td>
<td></td>
</tr>
<tr>
<td>Ss 161</td>
<td>2pq *1000 = 129</td>
</tr>
<tr>
<td>ss 5</td>
<td></td>
</tr>
</tbody>
</table>

$p = \sqrt{0.834} = 0.91$
$q = \sqrt{0.005} = 0.071$

Approach: Detect selection through comparison to neutral expectation

Kimura: neutral theory
Ewens: sampling formula
Coalescence

Neutral Theory History

- Motoo Kimura (1924-1994)
- 1968: a large proportion of genetic change is not driven by selection
- Adapted diffusion approximations to genetics
- Deal with finite pops

Genetic Drift

- no drift
  - infinite pop
- drift
  - finite pop
Neutral allele diffusion

Neutral allele diffusion built on foundations of diffusion theory and extended the idea of ‘identity by descent’ (ibd). It is sample-based and shifted focus to inferential methods, introducing the ‘infinite alleles’ model.

Ewens sampling formula (1972)

- infinite number of states into which an allele can mutate, therefore each mutation assumed unique
- $2N\mu$ new alleles introduced each generation, derived from existing alleles
- initial allele frequency = 1/(2N)
- every allele eventually lost

Ewens Sampling formula

Probability that a sample of $n$ gene copies contains $k$ alleles and that there are $a_1, a_2, ..., a_n$ alleles represented 1, 2, ..., $n$ times in the sample:

$$ P(a_1, a_2, ..., a_n) = \frac{n!\Theta^k}{\Theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j}a_j!} $$

where

$$ \Theta_{(n)} = \Theta(\Theta + 1)...(\Theta + n - 1) $$

and $a_j$ is the number of alleles found in $j$ copies

Expected Site Frequencies

Under diffusion, probability of an allele whose frequency is between $x$ and $x+\delta x$ is:

$$ f(x)\partial x = \Theta x^{-1} (1 - x)^{\Theta - 1} \partial x $$

where

$$ \Theta = 4Nu $$

$N = $ population size

$\mu = $ mutation rate

Infinite alleles model

Infinite alleles model is the model that assumes an infinite number of states into which an allele can mutate, deriving from existing alleles in each generation. The initial allele frequency is 1/(2N), and every allele eventually is lost.

Ewens Sampling formula

The Ewens Sampling formula is a probability distribution that describes the number of different alleles in a sample. It is based on the assumption of an infinite number of alleles and a constant mutation rate.

Expected Site Frequencies

The expected site frequencies are derived from the Ewens Sampling formula and represent the probability of observing a specific number of different alleles in a sample.
The Coalescent

Alternate, ‘backwards’ approach to generating expected allele frequency distributions

\[ i = 3 \text{ or } 1 \]
\[ i = 2 \]
\[ i = 1 \]

infer tree structure (genealogy), because tree structure dictates pattern of polymorphism in data

The Coalescent

How far back in time did a sample share a common ancestor?

\[ T_{\text{pop}} \approx 4N \text{ generations} \]

Coalescent inference

\[ P(\text{pattern}) = \sum P(\text{appropriate mutations}|G)P(G) \]

summary statistics obviate need to actually sum over all genealogies

Sample of size 2:

\[ P(\text{coal}) = \frac{1}{2N} \]

\[ f(t_2) = \frac{1}{2N} e^{-\frac{t_2}{2N}} \quad \text{for } t_2 \]

\[ P(k) = \left( \frac{\Theta}{\Theta + 1} \right)^k \left( \frac{1}{\Theta + 1} \right) \]

Probability of \( k \) mutation events before two sequences coalesce

Turning neutral models into selection tests

Three polymorphism summary statistics:

\[ S \quad \text{no. of segregating sites in sample} \]
\[ \pi \quad \text{avg. no. of pairwise differences} \]
\[ \eta_i \quad \text{no. of sites that divide the sample into } i \text{ and } n-i \text{ sequences} \]
Frequency-based selection tests

Tajima (1989) proposed:

\[ D = \frac{\pi - S / a_i}{\sqrt{\text{Var}(\pi - S / a_i)}} \]

where \( a_i = \sum_{i=1}^{n} \frac{1}{i} \)

Fu and Li (1993) proposed:

\[ D^* = \frac{S / a_i - \frac{n-1}{n} \eta_i}{\sqrt{S / a_i - \frac{n-1}{n} \eta_i}} \]
\[ F^* = \frac{\pi - \frac{n-1}{n} \eta_i}{\sqrt{\pi - \frac{n-1}{n} \eta_i}} \]

Neutral Expectation
(no selection, no structure, constant population size)

\[ E(\eta_i) \approx 0 \]

Positive Selection (Sweep)

Positive D, D*, F*

Balancing Selection

Positive D, D*, F*

Population Structure/Subdivision

Positive D, D*, F*
Population Expansion

\[
E(\eta_i)
\]

Negative D, D*, F*

Polymorphism vs. Divergence

Divergence between species should reflect variation within species

HKA Test

A Test of Neutral Molecular Evolution Based on Nucleotide Data
Richard K. Holmes, Martin R. Frommer, and Monika Halková
Journal of Molecular Evolution, 2006

Polymorphism vs. Divergence

Species A

Species B

Poly vs. Divergence

Species A

Species B

Divergence between species should reflect variation within species

Polymorphism/divergence with a twist: site classes

Synonymous changes: don't affect amino acid

UCU \Rightarrow \text{UCC} \Rightarrow \text{Serine}

Nonsynonymous (replacement) changes: new amino acid

UCU \Rightarrow \text{UUC} \Rightarrow \text{Phenylalanine}

MK Test

Adaptive protein evolution at the Adh locus in Drosophila
John H. McDonald & Martin Kreitman

Polymorphism/divergence with a twist: site classes

Synonymous changes: don't affect amino acid

UCU \Rightarrow \text{UCC} \Rightarrow \text{Serine}

Nonsynonymous (replacement) changes: new amino acid

UCU \Rightarrow \text{UUC} \Rightarrow \text{Phenylalanine}

Rate-based selection metric:

\[d_{N}/d_{S}\]

\[d_{S} = \text{no. nonsynonymous changes/ no. nonsynonymous sites}\]

\[d_{D} = \text{no. synonymous changes/ no. synonymous sites}\]

Counting codon 'sites' example: CAT

Histidine is encoded by only one other codon: CAC

CAT

P(T\Rightarrow C) = \text{fractional syn sites}

P(C\Rightarrow G or A) = \text{fractional nonsyn sites}

MK Test requires only 1 locus, but polymorphism data from 2 species.

Adh exhibits an excessive proportion of replacement fixed differences.
Rate-based selection metric: $d_N/d_S$

- $d_N/d_S < 1$: purifying selection
- $d_N/d_S = 1$: neutral expectation
- $d_N/d_S > 1$: positive selection

Codon Bias and Translation
Codon bias: the unequal usage of synonymous codons
- Thought to reflect selection for optimal translational efficiency and/or translational accuracy.

Distribution of Codon Bias Estimates for 6,453 Cryptococcus Genes

Preferred vs. Unpreferred Substitutions
Synonymous codon families were divided into 2 classes:

- "Preferred Codon" - disproportionately used in genes w/ high codon bias
- "Unpreferred Codon" - disproportionately used in genes w/ low codon bias

Comparing the Rates of Preferred and Unpreferred Synonymous Substitutions

$K_p = (# Preferred Substitutions) / (# Preferred Sites)$
$K_u = (# Unpreferred Substitutions) / (# Unrelated Sites)$

- $K_p/K_u < 1$: Excess Unpreferred Substitution
- $K_p/K_u = 1$: Neutral Expectation
- $K_p/K_u > 1$: Excess Preferred Substitution
**Selection Tests Summary**

- Allelic frequency spectrum tests (Tajima's D)
- Polymorphism/divergence tests (HKA, MK)
- Rate-based metric: $d_N/d_S$

**Correlates with $d_N/d_S$ (or just $d_N$)**

- expression level (-)
- dispensability (+)
- protein abundance (-)
- codon bias (-)
- gene length (+)
- number of protein-protein interactions (-)
- centrality in interaction network (-)