Molecular evolution

We have talked about A and a, B and b, etc.

These are alleles that conferred fitnesses, but what does that mean exactly?

They are DNA sequences that differ in some way from one another, presumably making different proteins or something

The field of molecular evolution studies these alleles at the molecular level
What exactly is a new mutation?

**Original allele:** ATACAGAT
**Mutant allele:** ATAACAGAT

This mutation in the DNA may:

- Change the protein sequence
  - change protein structure
  - change protein biochemical function
  - have no effect on structure or function

- Change the protein expression
  - protein expressed in new tissue
  - protein expressed at new time
<table>
<thead>
<tr>
<th>Era</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre 1960's</td>
<td>It was thought that there is probably little genetic variation</td>
</tr>
<tr>
<td>1960's</td>
<td>Allozyme (different forms of the same enzyme or protein) studies revealed high levels of variation, many allozymes segregating in populations</td>
</tr>
<tr>
<td>1980's+</td>
<td>DNA sequencing really begins and variation is all over the place, allozyme alleles are revealed as major polymorphisms and many minor polymorphisms were discovered.</td>
</tr>
</tbody>
</table>
Organism 1 ATACCAGATTTAGCAGTTAGCACCAGTGATGAC
Organism 2 ........................................G
Organism 3 ........................................C........G
Organism 4 ...........................................
Organism 5 ........................................G
Organism 6 ........................................A........G
Organism 7 ........................................A
Organism 8 ........................................A
Organism 9 ........................................A........G....T
Organism 10 ........................................A

Major polymorphism (9, 23) and minor polymorphism (18, 28)

Heterozygosity vs polymorphism

Linkage equilibrium (between 9 & 23)
Both selection and drift can cause linkage disequilibrium. Disequilibrium will decline with distance from selected allele and we can use this to guess at selected alleles.
<table>
<thead>
<tr>
<th>Organism</th>
<th>DNA Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATACCAGATTAGCAGTTAGCACCAGTGAGGATGAC</td>
</tr>
<tr>
<td>2</td>
<td>....................................</td>
</tr>
<tr>
<td>3</td>
<td>........................................C...G.........T</td>
</tr>
</tbody>
</table>
| 4        | ........................................G..........
| 5        | ........................................T |
| 6        | ........................................A..........
| 7        | ........................................A..........
| 8        | ........................................A..........
| 9        | ........................................A..........
| 10       | ........................................C...A..........

What does selection for the C mutation do?
What does selection for the C mutation do?
What does selection for the C mutation do?
What does selection for the C mutation do?
Organism 1  ATACC\textcolor{red}{C}G\textcolor{red}{A}ATAGCAGTTAGCACGAGTGGGATGAC
Organism 2 ........................................
Organism 3 ........................................T...T
Organism 4 ........................................
Organism 5 ........................................T
Organism 6 ........................................T...
Organism 7 ........................................
Organism 8 ........................................
Organism 9 ........................................
Organism 10 ........................................

Regions of very low polymorphism indicate \textit{selective sweeps}

\textbf{C} was selected
\textbf{A} and \textbf{G} "hitchhiked"
Neutral mutation

One allele always takes over the population, goes to fixation

If all alleles are equal, the chance for the “A” allele to be the one that fixes depends on it’s frequency

\[ \pi(p) = p \]

New mutation in single individual: \[ \pi(1/2N) = 1/2N \]

How long does this take?

\[ t(1/2N) = 4N \]
Advantageous mutation

If all alleles are not equal the chance for the “a” allele to be the one that fixes depends on the fitnesses

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

New mutation in single individual:

\[ \pi \left( \frac{1}{2N} \right) \sim 2s \]  
(Prob. is twice adv. in het.)

How long does this take?

\[ t_{fix} = ? \]
Advantageous mutation

How long does this take? \( t_{\text{fix}} = ? \)

\[
\tilde{t}_1(p) = \frac{1}{p} \int \psi(\xi)u(\xi)(1-u(\xi))d\xi + \frac{1-u(p)}{u(p)} \int_0^p \psi(\xi)u^2(\xi)d\xi
\]

Where \( u \) and \( \psi \) are given by the following:

\[
u(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx}
\]

\[
\psi(x) = \frac{1}{\int_0^1 G(x)dx} G(x)
\]

\[
G(x) = \exp\left\{ -\int_0^x \frac{2M}{V\delta_\xi} \right\}
\]

![Diagram showing genetic cycles](image)
Advantageous mutation

If all alleles are not equal the chance for the “a” allele to be the one that fixes depends on the fitnesses

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

New mutation in single individual: \( \pi(1/2N) \sim 2s \)

How long does this take? \( t_{\text{fix}} \sim (2/s)[ln(2Ns)+1] \)
Two forces change allele freq.

(1) $\Delta p$ due to random changes (genetic drift):
$$\Delta p = \frac{pq}{2N} \quad \text{(less for large } N\text{)}$$

(2) $\Delta p$ due to selection:
$$\Delta p = \frac{pq}{W_{\text{mean}}}\left[p(w_{11}-w_{12})+q(w_{12}-w_{22})\right]$$
$$\Delta p = \frac{pq}{W_{\text{mean}}}\left[p(s)+q(s)\right]$$
$$\Delta p = \frac{pq}{W_{\text{mean}}}\left[s(p+q)\right]$$
$$\Delta p = \frac{pq}{W_{\text{mean}}}$$
and since $W_{\text{mean}} \approx 1$ for small $s$ values
$$\Delta p \approx spq$$

When is: $\Delta p_{\text{sel}} > \Delta p_{\text{drift}}$?
$$spq > \frac{pq}{2N} ?$$
$$s > \frac{1}{2N} \quad \text{(selection works better in larger pop)}$$
Two forces influence fixation probability, \( \pi \)

\[
\begin{align*}
w_{11} &= 1 + 2s \\
w_{12} &= 1 + s \\
w_{22} &= 1
\end{align*}
\]

(1) If \( s = 0 \), allele "A" is neutral (genetic drift):

\[
\pi = \frac{1}{2N} \quad \text{(less for large N)}
\]

(2) If \( s = 0 \), allele "A" is advantageous or deleterious (selection):

\[
\pi = \frac{1 - e^{-2Ns}}{1 - e^{-s}} \quad \text{which reduces to } \pi = 2s
\]

When is:

\[
\pi_{\text{sel}} > \pi_{\text{drift}}? \\
2s > \frac{1}{2N} ? \\
s > \frac{1}{4N} \quad \text{(selection works better in larger pop)}
\]
Neutral
\[ p(\frac{1}{2N}) = \frac{1}{2N} \]

Advantageous
\[ p(\frac{1}{2N}) = 2s \]

Rate of substitution, assuming a mutation rate of \( \mu \):

\[ 2N\mu(\frac{1}{2N}) = \mu \]
\[ 2N\mu2s = 4N\mu s \]

Assuming that advantageous mutations are rare the overall rate of substitution should be determined by the rate of neutral evolution.

Neutral theory of molecular evolution (Motoo Kimura):
Most substitutions are neutral so the rate of substitution = \( \mu \)

Is this true?
Constant rate of substitution?

If the rate of substitution were constant - and - we could date some splits accurately with fossils

This would allow us to create a molecular clock
Molecular clock

If the rates are constant and we have dates for at least one split
1. Build a tree
2. Compute the number of changes on each branch
3. Date one split using fossils or geology
4. Infer the age of other splits
We would expect populations with shorter generation times to diverge faster.

We don’t see that.

Why not?
We would expect populations with shorter generation times to diverge faster.

We don’t see that.

Why not?

Nearly neutral theory (Tomoko Ohta and Motoo Kimura)
Most mutations are *nearly* neutral, slightly deleterious. Since selection is more effective in larger populations, these new mutations fix at a slower rate in larger populations. This slows per-generation substitution rate of larger populations, the ones with shorter generation times.
Neutral theory (Motoo Kimura)
- Most fixed mutations neutral
- Substitution rate is therefore $\mu$ for all species

Nearly neutral theory (Tomoko Ohta and Motoo Kimura)
- Most fixed mutations are nearly neutral, slightly deleterious.
- The balancing of stronger selection against these mutations in larger populations and the shorter generation times of these species results in the constant rate we observe.

Selectionist theory (Gillespie)
- Almost nothing is truly neutral and many fixations are beneficial
- Molecular clocks are untrustworthy

How to decide?

This is the first topic in Bio 412/512, also discussed in Bio 472/572
What sorts of DNA or protein changes occur?

"A" and "a"
"B" and "b"

Organism 1

Organism 2

ATACCAGATTAGCAGTTAGCCCCAGTGGATGAC

...............A...............T...............  

Transversion  Transition

Purines: A,G
Pyrimidines: C,T
Silent/synonymous:

CAA CGT CCG ACA AGT        CAG CGT CCG ACA AGT
val ala gly cys ser        val ala gly cys ser

No change in protein, almost always neutral

Replacement/nonsynonymous:

CAA CGT CCG ACA AGT        TAA CGT CCG ACA AGT
val ala gly cys ser        ile ala gly cys ser

Changes protein: often deleterious, sometimes beneficial
Nonsense:

\[
\text{CAA CGT CCG ACA AGT} \quad \text{CAA CGT CCG ACT\text{T} AGT}
\]

\[
\text{val ala gly cys ser} \quad \text{val ala gly stop}
\]

Results in a truncated protein, almost always deleterious

**Insertion/deletion, indel:**

\[
\text{CAA CGT CCG ACA AGT} \quad \text{CAA CGT CCG TAC AAG T}
\]

\[
\text{val ala gly cys ser} \quad \text{val ala gly met phe}
\]

Results in a frameshift, almost always deleterious
Microsatellites “mutate” or change by recombination which is **much** more likely than a brand new mutation.

This is why microsatellites are appropriate for closely related species or individuals.

They are NOT appropriate for divergent species or groups (totally scrambled).

http://www.dakotacom.net/~clamunyon
**Mutation**: change in a single individual

**Polymorphism**: when a mutation is observed at a certain frequency

**Substitution**: change in entire population due to fixation of a mutation

What influences rates of substitution?

1. Microsatellite repeat number change very fast

2. Nucleotide substitutions

   **Transitions** are faster than **transversions** because the mutation rate is higher

---

![Diagram showing transitions and transversions between nucleotides A, G, C, T.](image-url)
Nucleotide changes that affect amino acids will be fixed (substitute) slower than ones that don’t

- “junk” DNA changes faster than coding regions
- introns change faster than exons
- 3rd position in codons will change faster than 2nd or 1st
- Silent/Synonymous (aa. not changed) substitutions will be faster than Replacement/nonsynonymous (aa. changed)
Amino acid substitutions
These will be more rare than nucleotide substitutions
Some of these may be more or less likely than others

Conservative substitution
(little change in charge or size)

Radical substitution
(more change in charge or size)

TAT → AAT
Isoleucine (ile) → Leucine (leu)
hydrophobic → hydrophobic

TAT → TCT
Isoleucine (ile) → Arginine (arg)
hydrophobic → hydrophilic
Functional importance of genes
Some genes are more equal than others

Observations have shown that the fastest evolving genes are ones with many copies doing very similar tasks

Olfactory receptor genes
Major histocompatibility complex (MHC) genes

Redundancy may make some genes less deleterious when changed (or change may be favored to increase diversity)

Slowest changing genes are fundamentally important ones like histones (DNA wraps around the histones)
Rates of molecular evolution

(1) **Microsatellite** repeat number

(2) Nucleotide substitution rate influenced by mutation rate:
    transitions > transversions

(3) Nucleotide substitution rate influenced by fitness effect:
    Silent/synonymous > replacement/nonsynonymous
    nonsynonymous > Nonsense and frameshift

(4) Amino acid substitutions influenced by fitness effect:
    Conservative substitution > Radical substitution

(5) Functional importance of gene
    Redundant genes > unique or important gene

Which one you should use depends on the time scale you are studying.