Evolutionary population genetics: a brief introduction

Pradipta Ray,
BIOL 6385 / BMEN 6389,
The University of Texas at Dallas
(some material based on content by PR in Eric Xing’s 10-810 Carnegie Mellon class)
accidental death (drift)

Some errors during DNA production/copying become incorporated into gametes (mutation)

Plants that are better adapted survive, grow more quickly, and produce more young than others (natural selection)

Loss of pollen through chance events (genetic drift)

Loss of some seed through chance events (genetic drift)
Forces shaping observations

• Finite sample size: we can't sample the whole population

• Sample bias: is our sample representative?
Fixing

• Fate of an allele in long run: evolutionary race among alleles
  – either dies out: “fixed” at 0 (known as loss)
  – intermediate situation: “fixed” at intermediate value, determined by equilibrium distribution of stochastic process (known as equilibrium)
  – wipes out all other alleles (becomes monoallelic): “fixed” at 1 (known as fixing)

Even at equilibrium, alleles could still be fixed or lost under drift
Mutation: a cursory look
Mutation models: how many alleles

- Bi – allelic model: Two alleles exist – a mutation will change the allele from type A to type B (or a deleterious new allele which will vanish)
- Multi – allelic model: Many alleles exist – effect of mutation may be difficult to predict without explicit model
- Infinite – allelic model: Every mutation creates a new allele (convenient, not true)
Mutation models: how many sites

Infinite sites model:

• Every mutation happens at a new locus

• Expected no of substitutions / site $<< 1$

• Plausible assumption for low mutation rate, short evolutionary history studies
Mutation models: parameters

- Typical assumption: all mutations equally likely to occur (does not mean all mutations equally likely to survive)

\[ N \text{ individuals} \rightarrow 2N^2 \text{ alleles} \rightarrow 2 \mu N^2 \text{ mutations in every generation} \]
Selection : a cursory look
Selection models : fitness

• **Natural selection** : Some changes are more important for survival or lineage propagation based on environmental and other factors : **fitness fn** selects some traits over others

[Image: ultraviolet_light.png]

http://evolution.berkeley.edu/evosite/
Selection models: phenotypic selection

• Impossible to enumerate (aleph – 2 kinds!)

• Difficult to parameterize (curve fitting)

• Tricky to estimate (how many samples are enough depends on complexity of curve)

Wikipedia

BIOL 6385, Computational Biology
Selection models: genomic selection

• At the phenotypic level: various kinds of selectional forces are at play
• At the genomic level: these translate to 3 basic kinds:
  – Positive selection: Advantageous changes are accelerated
  – Negative / purifying selection: Deleterious changes are discarded
  – Neutral selection: Selection plays no part
Facts about selection

• Any preference for one kind of change over another (simple eg: transition vs transversion)
• Operates at every granularity: nucleotide, codon, protein, etc
• Operates at both allelic and genotypic / haplotypic level
• Operates at every resolution: population, subspecies, species : one of the driving forces of speciation
A myth about selection

• Selection causes all but the fittest allele to vanish = myth (but it potentially could!)
  — why? We’ll discuss this shortly

• Think HKY 85 model:

\[ Q = \begin{pmatrix} * & k \pi_C & \pi_A & \pi_G \\ k \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & k \pi_C \\ \pi_T & \pi_C & k \pi_A & * \end{pmatrix} \]

\( k \to \text{some form of selection} \)

Equilibrium distr. \( \rightarrow \pi_A \pi_C \pi_G \pi_T \)

\( \text{not 100000} \)
Selectional models

• Genotype based model
  – each genotype $a_i$ has a different selectional coefficient $s_i$

• Can we incorporate our sampling bias into selectional models?
How selection works
Modelling selection

- Fitness = expected no of offspring in the next generation
  - for the neutral model with fixed population, all genotypes have fitness = 1
- Various models of fitness
  - Dominant disease
  - Recessive disease
  - Heterozygous advantage
  - Directed selection

\[
\begin{array}{ccc}
\text{AA} & \text{Aa} & \text{aa} \\
1-s & 1 & 1 \\
1 & 1 & 1-s \\
1-s & 1 & 1-s \\
1 & 1-hs & 1-s \\
\end{array}
\]
Drift : a cursory look
Drift

• What is drift
  – random, directionless fluctuations in allele frequency from generation to generation

• It is the act of randomly sampling a finite no of alleles from the previous generation
  – we don’t observe the whole population: can we incorporate fluctuations / errors due to our finite sample size into drift models?
Parameterizing drift

• Drift is a result of finite sampling
  – sample size should be our parameter

• What is our sample size ?
  – N individuals, 2N alleles
  – Population size is fixed : sample size = 2N
  – We will see later how to handle situations where
    population size is not fixed
Selection and drift

**Role of Selection:** More prominent in larger pop.

- Directional pressure: \( E(x_{t+1} - x_t) > 0 \) or \( E(x_{t+1} - x_t) < 0 \)
  - based on advantageous or deleterious allele

- Under only selection and drift:

**Diagram:**
- In some stretches, drift overshadows selection
- Concerted upwards pressure
- Concerted downwards pressure
Population size

• Smaller population
  – greater chance of losing allele by drift alone: may undo evolutionary optimization by removing advantageous allele before selection can play a role

• Larger population
  – lesser role of drift, greater role of selection in variation reduction
  – greater no of mutations if rate is fixed

Simulation under only drift
Population bottleneck

• Loss of alleles

• Alleles driven to irrecoverable frequencies: absorbed to 0

• Founder / bottleneck effects

Wikipedia
In a world with only variation reduction

• Sooner or later, all but one allele will go extinct (which one may not be predictable: no stationary distribution)

• If the population size is infinite, then drift will have no role in the long term prospects of an allele [only the fittest allele will survive]
Static models

• Analyzing the allele frequencies at a particular time snapshot (one single generation’s allele frequencies are modelled)

• Modelling allele diversity: how much diversity can we expect in a population?
  – should it depend on the mutation rate?
  – should it depend on the population size?
Chinese restaurant process: modelling a diverse population

- Chinese restaurant process: infinite alleles: static or dynamic model?

\[ \Pr(B_n = B) = \frac{\prod_{b \in B} (|b| - 1)!}{n!} \]

byu.edu

OLD TBL
\[ \frac{161 - \alpha}{n + \theta} \]

NEW TBL
\[ \frac{\theta + 1B|\alpha}{n + \theta} \]

MORE GENERIC MODEL

byu.edu
Expected number of tables

• Relating allelic diversity with population size

\[
\frac{\Gamma(\theta + n + \alpha)\Gamma(\theta + 1)}{\alpha\Gamma(\theta + n)\Gamma(\theta + \alpha)} \quad - \quad \frac{\theta}{\alpha}.
\]

• Looks at a particular snapshot in size, when population = n
  – Whether static or dynamic depends on your question: are you modelling evolution of no of alleles? Evolution of population in terms of allele frequency?

UT Dallas
The University of Texas at Dallas

BIOL 6385, Computational Biology
Dynamic model: multiple snapshots in time

• How do populations change at each table as more members join, and as a function of the number of tables?

• Will the largest table will stay the largest after a doubling of the population?
Standard dynamic pop genetic models

• Allele frequency models (absolute or relative)
Variations on the model

- 2 near-“absorbing states”: 0 and 1
  - not truly absorbing if mutation can occur

- Continuous valued
  - relative frequencies of alleles modelled

- Discrete valued
  - Granularity of relative frequencies determined by population

- Discrete time: generations
  - overlapping generations: eg human
  - non overlapping generations: eg annual plants

- Continuous time: brownian motion, diffusion process
Standard pop genetic models

• Models may not necessarily be Markovian

• Imagine a situation where an individual’s fecundity is bounded
  – the next generation’s allele frequencies will also depend on the history of the population, not just the current allele frequencies
Standard pop genetic models

- What should Markovian models look like?

\[ AF(0) \rightarrow AF(1) \rightarrow AF(2) \rightarrow \ldots \]

Discrete: time unit = generations

- Bayesian networks are a good way to visualize the dependencies
Standard pop genetic models

• What is observed?

\[ AF(0) \rightarrow AF(1) \rightarrow AF(2) \rightarrow \ldots \]

\[ PH(0) \rightarrow PF(1) \rightarrow PH(2) \rightarrow \ldots \]

– in the old days, phenotype …
Standard pop genetic models

• What is observed?

\[ \text{AF(0)} \rightarrow \text{AF(1)} \rightarrow \text{AF(2)} \rightarrow \ldots \]

\[ \text{Pf(0)} \rightarrow \text{Pf(1)} \rightarrow \text{Pf(2)} \rightarrow \ldots \]

– these days, we observe both ...
Standard pop genetic models

- Evolution is really operating at the level of genotype / haplotype
Standard pop genetic models

- Evolution is really operating at the level of genotype / haplotype
- A2G models: census of alleles, but we need genotype
Drift + selection + mutation

Drift + selection + mutation + migration

Drift + selection + mutation + recombination

Beanbag model

AT EQUILIBRIUM w/ no D, u, m or s

EQUILIBRIUM DISTRIBUTION EXISTS FOR SUFFICIENTLY LARGE POPULATION:

MIGRATION = PERTURBATION
RECOMBINATION INCREASES VARIATION IN MULTIALLELIC MODEL

Drift + selection + mutation

+ multiallele haplotype

+ recombination

Drift + mutation

Drift + selection

Drift

Drift + mutation

Drift + selection + mutation

---

NO EQUILIBRIUM DISTRIBUTION EXISTS FOR SUFFICIENTLY LARGE POPULATION:
Classic idealized pop models

• Wright – Fisher model:
  – Mendelian inheritance
  – no sexual selection (allele frequencies don’t differ in sexes)
  – no overlapping generations (in the most complicated deviation from model, we use continuous time)
  – sex ratio = 1
  – effective population = actual population
  – fixed population
  – no selection
  – finite population (discrete valued stoch process)

• Other popular models may trade off such oversimplification for less tractable inference
State space diagram under the Wright - Fisher model

- More connectivity than a random walk
- Bi allelic state space for finite population

Blue transitions allowed if mutation is modeled
Classic idealized pop models

• Other models: Moran model
  – overlapping generations
  – one or two individuals (depending on evolutionary model) selected to reproduce by sampling
  – new individual created and added to pool
  – one individual is killed to retain constant population
State space diagram for the Moran model

Random walk!

Blue edges allowed only if mutation is modeled
Urn models of pop evolution

• George Polya

• Notion of sampling a non-homogeneous population

• Variations on a theme
  – Without replacement / with replacement / with replacement and duplication / new urn
Analogous action on urns

- Overlapping / non overlapping generations
Analogous action on urns

• Mutation
  – new color : infinite alleles model
  – under no such assumption, the duplication merely causes the ball to change color
Analogous action on urns

- Drift: Random fluctuations in frequency from generation to generation: it is the act of the random sampling
  - how is death implicitly modelled in W F models?
Analogous action on urns

• Selection: Sampling is disproportionate to no of balls of each color
Analogous action on urns

- Population effects: the effects of a finite sized jar

balls are indivisible during sampling
Which model should I choose?

- Depends on your need
  - modelling Drosophila with overlapping generations (Moran model)
  - modelling populations where non-overlapping generations are a good approximation (W–F model: we will be studying this from now on)
  - more complicated situations (diffusion model: we will study the basics of the diffusion process at the end of the lecture)
Bean bag models & deviations
Null model of population genetics

• Simplest dynamic model of allele frequencies
• Assumption: allele (and thus genotype) frequencies are in equilibrium
  – no forces at work: drift, selection, mutation, etc

• Beanbag model: alleles only move around like beans in a bean bag from gen to gen
  – even without mutation or selection, can really happen only in an infinite (v large) population where drift doesn’t affect allele frequency
Genotype frequencies

• Why model genotype frequencies?
  – remember, traits determined by genotypes (selection acts on genotypes)

• Hardy Weinberg equilibrium

\[ p + q = 1 \]
Why?

• Alleles segregate independently

• Even if you start with genotype frequencies not at HWE, in one generation genotype frequencies will be at HWE

• Allele frequencies are the same for both sexes (no sexual selection): could allele frequencies be different in the two genders at HWE?
Key properties

• Allele frequency is not affected by alleles segregating into different genotypes

• For pure dominance models, allele frequencies can be estimated from phenotype frequencies
Sex linked alleles

• p & q for heterogametic sex
• \(p^2, 2pq\) and \(q^2\) for homogametic sex

\[
\text{Census: } x_1 \rightarrow A, \quad x_2 \rightarrow a
\]

freq. of sex-linked allele.

At equilibrium, what are the diff. genotypes and their frequencies?
HWE : multiple alleles

Consider: \((p_1V_1 + p_2V_2 + \cdots + p_nV_n)(p_1V_1 + p_2V_2 + \cdots + p_nV_n)\)

→ what is the coefficient of \(V_i V_j\)?

\[ P \{A_i \text{ observed } x_i \text{ times, } A_j \text{ observed } x_j \text{ times}\} \]
\[ = \binom{x_i+x_j}{x_i} \cdot \frac{(x_i+x_j)!}{x_i!x_j!} \]
\[ = \binom{x_i+x_j}{x_i} \cdot \frac{(x_i+x_j)!}{x_i!x_j!} \]

\[ \therefore i \neq j \Rightarrow p(A_iA_j) = 2p_i p_j, \quad i = j \Rightarrow p(A_iA_i) = p_i^2 \]

for higher ploidy, the coeff. of \(\{A_i \text{ observed } x_i \text{ times for } i = 1, 2, \ldots, m\}\)

COUNTING \(\binom{x_1!x_2! \cdots x_m!}{x_1!x_2! \cdots x_m!}\)
Ewens sampling formula: distributions over partitions

• Finite sample size

\[ \Pr(a_1, \ldots, a_n; \theta) = \frac{n!}{\theta(\theta + 1) \cdots (\theta + n - 1)} \prod_{j=1}^{\nu} \frac{\theta^{a_j}}{j^{a_j} a_j!}, \]

• Role of theta

No selection
Space of allele frequencies

The simplex for 3 alleles

\[ p + q + r = 1 \]
Corresponding genotype space

- Higher dimensional space:
  - choose 2 from k options, with repeats

- Under HWE:
  - constrained space
Test for HWE: categorical tests

• Measuring deviations in the simplex

\[
\begin{array}{ccc}
AA & aa & Aa \\
n_{AA} & n_{aa} & n_{Aa} \\
\end{array}
\]

Observed vs Estimated

- $\hat{p}^2 \cdot n$ for $AA$
- $(1-\hat{p})^2 \cdot n$ for $aa$
- $2\hat{p}(1-\hat{p}) \cdot n$ for $Aa$

$\hat{p} = \frac{n_{AA} + n_{Aa}}{n}$

• Perform a categorical test: are the 2 rows drawn from the same distribution?
  - eg chi square (degrees of freedom = no of genotypes – no of alleles)
So, we arent in HWE, ...

• what next?

• A categorical test only tells us whether the population is under HWE, doesn’t tell us the likelihood of observing a non-HWE equilibrium

• Without explicitly modelling the different kinds of forces, we may put a prior over the space of genotype frequencies, based on sampling or prior knowledge
Dirichlet distributions: non-HWE equilibrium models

- For HWE violations, we want to move away from the \((p^2, 2pq, q^2)\) parameterization
- “Pushing” the point on the simplex to a region of the simplex: Dirichlet distrn
\[ P(x_1 \mid x) = \int P(x \mid \theta) P(\theta \mid x) d\theta \]
\[ = \int \prod_{j=1}^{m} \theta_j^{n_j(x)} \left( \frac{1}{c(\alpha)} \prod_{j=1}^{m} \theta_j^{\alpha_j-1} \right) d\theta \]
\[ = \frac{1}{c(\alpha)} \int \prod_{j=1}^{m} \theta_j^{n_j(x)+\alpha_j-1} d\theta \]
\[ = \frac{c(N(x)+\alpha)}{C(N(x)+\alpha)} \]

\[ C(\beta) = \int \prod_{j=1}^{m} \theta_j^{\beta_j-1} d\theta = \prod_{j=1}^{m} \frac{\Gamma(\beta_j)}{\Gamma(\beta_j)} \]
\[ \beta_s = \sum_{j=1}^{m} \beta_j \]

\[ P(X_{n+1} = k \mid x_1: n, \alpha) = \frac{\Gamma(k+\alpha_k)}{\Gamma(n+\alpha_s)} \]
Moving on from the bean-bag model

• We want to model the evolutionary dynamics of the allele frequencies
  – population may not be in equilibrium: we may want to characterize the trajectory such populations take towards their long run configurations
  – even if it is in equilibrium, we may want to find out the nature of forces (mutation, drift, selection) acting on it
Simplest violation of the bean bag model

- Allele populations remain in equilibrium if sampling can be done faithfully at every generation

- Finite populations
  - finite sampling comes up with distribution with errors wrt original population distribution
Drift only models
Drift only model: long run

- No stationary distribution
  - What is the probability that drift will fix one allele and not the other: probability of fixation?
  - Proportional to the relative frequencies
  - Remember gambler’s ruin and the random walk

We will discuss this again when studying coalescents.
Drift only model: heterozygosity

- With no mutation, **identity by descent**
- $H_t = \Pr$ of picking two different alleles in the population at time $t$
- For bi allelic model, $H_0 = 2x_0 \left( 1 - x_0 \right)$

$$E_{x_1, x_0} (H_1) = \sum_{x_1} 2x_1 (1-x_1) P \left( x_1 = x_1, x_0 = x_0 \right)$$

$$= H_0 \left( 1 - \frac{1}{2n} \right)$$

$$E (H_t) = H_0 \left( 1 - \frac{t}{2n} \right)$$

*Allele diversity retained only for $N \to \infty$*
Variance in the sampling process

\[ \Delta x = x_t - x_{t-1} \]

\[ E(\Delta x) = 0 \iff \text{TRUE RANDOM, NOT DIRECTIONAL} \]

\[ V(\Delta x) \uparrow \text{as } N \downarrow \iff \text{EFFECT OF DRIFT INCREASES WITH SMALLER POPULATION} \]

\[ V(\Delta x) = \frac{x(1-x)}{2N} \]
$2N^2 \text{ pop.} \rightarrow \text{ allele pop: } \begin{align*} K_t, & \quad 2N-K_t \\ \text{rel. freq:} & \quad x, \quad 1-x \end{align*}$

$k_{t+1} \sim \text{Bin} (2N, x)$

$V(\Delta x) = V(x_{t+1} - x_t)$

$= V(x_{t+1} - \text{const.}) = V(x_{t+1})$

$= V \left( \frac{k_{t+1}}{2N} \right)$

$= \frac{1}{(2N)^2} \cdot V(k_{t+1})$

$= \frac{1}{(2N)^2} \cdot 2N \cdot x \cdot (1-x)$

$= \frac{x (1-x)}{2N}$

Var. of binomial
Modelling more complex directional change

• Under drift, expected value of change in allele frequency in one generation = 0

• However, empirically, we know that allele frequencies show directed change: selection
Selection only models
Selection only models

• N is assumed to be large: drift has little effect
• Usually variation decreasing force

• But, what happens if the heterozygous allele has maximum fitness?
  – variation increasing force: if variation is thought of as degree of heterozygosity of the population

• How to model changing allele frequencies?
Selection only models

• Important notion: mean fitness of population

\[ \begin{align*}
\text{fitness} & \quad \text{Orig. freq.} \\
\text{Expected no. of offspring freq of gene type} & \quad \text{next gen.}
\end{align*} \]

\[ \begin{align*}
\omega_\text{AA} & = x^2 \omega_\text{AA} \\
2x(1-x) & = 2x(1-x) \frac{\omega_\text{AA}}{\bar{\omega}} \\
(1-x)^2 & = (1-x)^2 \frac{\omega_\text{aa}}{\bar{\omega}}
\end{align*} \]
Change in allele frequencies

\[ x' \leftarrow \text{allele freq. in next gen.} \]

\[ x' = \left[ 2x^2 \cdot \frac{w_{AA}}{\bar{w}} + 2x(1-x) \cdot \frac{w_{AB}}{\bar{w}} \right] \cdot \frac{1}{2} \]

\[ = \frac{x}{2} \left( x w_{AA} + (1-x) w_{AB} \right) \]

\[ \triangle x = x' - x = \frac{x}{\bar{w}} \left( \bar{w}_{A} - \bar{w} \right) = \frac{x(1-x)(\bar{w}_A - \bar{w})}{\bar{w}} \]

\[ \bar{w} \]

No. of alleles = 2
No. of genotypes
Adding to the mix

• So far, we have modelled how existing alleles compete with each other over generations

• But how do these different alleles get created?

• Modelling the primary driving force of polymorphism: mutations
Mutation only models
Frequency of mutation

• Mutation rate: no of de novo mutations as a fraction of the total population

\[ \text{Mutation rate} = \mu \]

\[ \text{fraction of population with mutation after 1 gen} = \mu \]

\[ \text{after n gen} = 1 - (1-\mu)^2 \]

• Does this really happen?

No, selection & drift drive most mutations to 0 frequency (loss)
Mutation only models

• May not be very interesting to look at long run probabilities: no drift or selection to balance allele frequencies at equilibrium: no drift or selection to drive them to loss or fixing

• Instead, more realistic models will try to model mutation in a setting where alleles are lost due to drift or selection
Drift – mutation models
Fixation prob of neutral mutation

• Initial prob of novel mutation = \( \frac{1}{2N} \) (no two mutations are same under infinite alleles model)

• Remember drift only models: the probability of fixing this mutation would be its starting relative frequency = \( \frac{1}{2N} \)

• Under infinite sites model, fixation rate of any new mutation (from a generation):

\[
2N\mu \times \frac{1}{2N} = \mu
\] (INDEP. OF POPULATIONS !)
Mutation with drift: neutral model

For driftonly model, \( G_t = 1 - H_t \)

\[
G_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) G_t
\]

With a mutation rate \( \mu \)

\[
G_{t+1} = (1 - \mu)^2 \left[ \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) G_t \right]
\]

At equilibrium, \( G_{t+1} = G_t \) for \( \mu \ll N \)

\[
G_{eq} = \frac{1}{1 + 4N\mu}, \quad H_{eq} = \frac{4Nm}{1 + 4Nm}
\]
Problems!

• Can we estimate N (population) if we know mutation rate and heterozygosity
  – heterozygosity : sample population
  – mutation rate = substitution rate (why? later...)

• We get wrong answers for N using well established data sets for humans (we get N = 6000) and Drosophila (we get N = 200,000)
  – why? Real populations may not be following W–F model
Effective population size

- $N_e$ may be much less than $N$
- How to estimate $N_e$
  - Variance in no of offspring
  - Fluctuating population
  - Gender skew

$$N_e = \frac{N}{\sigma^2}$$

$$N_e = \frac{1}{\frac{1}{t} \sum \frac{1}{N_i}}$$

$$N_e = \frac{4NmNf}{Nm + Nf}$$

SIMPLE WAY TO CHECK: SIMULATE UNDER WRIGHT FISHER
What is $N_e$ anyway?

- It is the population of an ideal W-F model that would approximate the population dynamics of the current population under study.
Gender imbalance and effective population

- For organisms with matriarchal or patriarchal clans, the approximation should be different.

\[ N_e = \frac{4N_m N_f}{N_m + N_f} \]

Remember, \( N_m + N_f = N \)
N versus \( N_e \)

- Estimated \( N_e \)
- Census \( N \)

\( N = N_e \)
Mutation – selection models
Continuous allele frequencies

• From now on, we will consider that allele frequencies can be modelled as continuous.

• We can now take derivatives wrt the allele frequency!
Rate of change of allele frequencies

- As $N$ is assumed to be large, allele frequencies can be modelled as continuous. Derivatives wrt $x$ can be taken:

$$E_s [\Delta x] = \frac{x(1-x)}{2 \bar{w}} \frac{dq_\bar{w}}{dx}$$

Max selected force for intermediate allele freq.

$\bar{w} \to$ mean fitness

allele frequency increases if allele increases pop. fitness
Rate of change of allele frequencies

\[
\frac{d\bar{\omega}}{dx} = \frac{d}{dx} \left[ x^2 \omega_{AA} + 2x(1-x) \omega_{AB} + (1-x)^2 \omega_{BB} \right]
\]

\[
= 2x \omega_{AA} + 2(1-x) \omega_{AB} - 2x \omega_{AB}
- 2(1-x) \omega_{BB}
\]

\[
2(\bar{\omega}_A - \bar{\omega}_B) = 2(x \omega_{AA} + (1-x) \omega_{AB}
- x \omega_{AB} - (1-x) \omega_{BB})
\]

\[
= \frac{d\bar{\omega}}{dx}
\]
Selection – mutation balanced model

**Genotype**
- AA
- Aa
- aa

**Fitness**
- Before selection
- After selection (un normalized)

**Recessive Disease Model (Multiple Sclerosis)**

- How "s" can be empirically calculated in F2
Selection – mutation balanced model

• Balancing of allele frequency by mutation and selection
  – Why is drift not considered here?

\[
E_s(\Delta q) = \frac{q(1-q)}{2W} \frac{d\bar{w}}{dq} \sim -sq^2(1-q) \sim sq^2
\]

\[
E_m(\Delta q) = \mu(1-q) \sim \mu
\]

\text{Equil: } E_s(\Delta q) = -E_m(\Delta q)

\[
q_{\text{Equil}} = \sqrt{\frac{\mu}{s}}
\]
Working this out for other models

• Can be worked out for other sets of selection coefficients
  – eg, another model

\[
\begin{array}{c|c|c}
AA & Aa & aa \\
1-s & 1 & 1 \\
\end{array}
\]

\[ q_{EQUIL} = \frac{\mu}{s} \]
Drift – mutation – selection models
Continuous-valued, continuous-time stochastic processes

- Continuous valued, continuous time processes:
  - discontinuous in time/jump: sample paths discont

\[ X_t \]

- many notions of continuous in time: Sample – continuous: all sample paths \textit{almost surely} continuous (eg diffusion process)
Diffusion process

\[ P_r \left( X_t = \alpha \right) = \sum_b P_r \left( X_{t-k} = b \right) P_r \left( X_t = \alpha \mid X_{t-k} = b \right) \]

For us:

\[ P_r \left( X_t = i \right) = \sum_{\Delta i} P_r \left( X_{t-1} = i + \Delta i \right) P_r \left( X_t = i \mid X_{t-1} = i + \Delta i \right) \]

(Forward Kolmogorov Eqn)

For continuous-valued, continuous-time process:

\[ \phi \left( x ; t \right) \rightarrow \text{density fn for observing allele frequency } x \text{ at time } t \]

\[ g \left( x - \varepsilon, \varepsilon, \delta t \right) \rightarrow \text{prob. of going from } x - \varepsilon \text{ to } (x - \varepsilon) + \varepsilon \text{ in time } \delta t \]

\[ \phi \left( x ; t + \delta t \right) = \int \phi \left( x - \varepsilon ; t \right) g \left( x - \varepsilon, \varepsilon, \delta t \right) d\varepsilon \]
Diffusion process

\[ \phi(x; t+\Delta t) = \int \left[ \phi(x) g(x) - \varepsilon \frac{\partial}{\partial x} \left( \phi(x) g(x) \right) + \frac{\varepsilon^2}{2} \frac{\partial^2}{\partial x^2} \phi(x) g(x) \right] \, dx \]

\[ \phi(x-\varepsilon; t) \quad g(x-\varepsilon, \varepsilon, \Delta t) \]

\[ = \phi(x) \int g(x) \, dx - \frac{\partial}{\partial x} \left[ \phi(x) \int g(x) \, dx \right] \]

\[ + \frac{1}{2} \frac{\partial^2}{\partial x^2} \left[ \phi(x) \left( \int g(x) \, dx \right)^2 \right] \]

Expected change to \( x \) in time \( \Delta t \)

\[ \phi(x; t+\Delta t) - \phi(x; t) = -\frac{\partial}{\partial x} \left[ \phi(x, t) \frac{1}{\Delta t} \int g(x-\varepsilon, \varepsilon, \Delta t) \, d\varepsilon \right] \]

\[ + \frac{1}{2} \frac{\partial^2}{\partial x^2} \left[ \phi(x, t) \frac{1}{\Delta t} \int g(x-\varepsilon, \varepsilon, \Delta t) \, d\varepsilon \right] \]

Expected change to \( x^2 \) in time \( \Delta t \)

\[ \Delta t \]

\[ \frac{\partial}{\partial x} \left[ \phi(x, t) \frac{1}{\Delta t} \int g(x-\varepsilon, \varepsilon, \Delta t) \, d\varepsilon \right] \]

\[ + \frac{1}{2} \frac{\partial^2}{\partial x^2} \left[ \phi(x, t) \frac{1}{\Delta t} \int g(x-\varepsilon, \varepsilon, \Delta t) \, d\varepsilon \right] \]
Diffusion process

\[
M(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \int \mathcal{E} g(x-\mathcal{E},t,\delta t) d\mathcal{E}
\]

\[
V(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \int \mathcal{E}^2 g(x-\mathcal{E},t,\delta t) d\mathcal{E}
\]

[\text{E: for small } \mathcal{E}(y), V(y) \sim \mathbb{E}(y^2)]

\[
\frac{\partial \phi(x;t)}{\partial t} = -\frac{\partial}{\partial x} \left\{ \phi(x;t) M(x,t) \right\} + \frac{1}{2} \frac{\partial^2}{\partial x^2} \left\{ \phi(x;t) V(x,t) \right\}
\]

\[
\phi(x;t) = \text{Const} \exp \left( 2 \int \frac{M(x,t)}{V(x,t)} dt \right)
\]
Modelling the diffusion process

• We are now dealing with densities, not probabilities

• So far, preference for one kind of change over another was exclusively modelled through selection
  – now for each kind of mutation (\( A \rightarrow B \)), we have a mutation rate (may not be agnostic to nature of change)
Modelling the diffusion process

• Model mutation rates

• Model selection coefficients

• Model the functions of mean and variance of the rate of change of the alleles

\[ M(x, t) & \ U(x, t) \]

– additional parameters may be needed (eg. variance contributed by selection?)
Equilibrium frequencies: adaptive mutation

\[ M_{\delta x} = s \alpha (1-\alpha) + \left[ (1-\alpha) \mu - \alpha 2s \right] \]

\[ V_{\delta x} = \frac{\alpha (1-\alpha)}{2Ne} \]

\[ \phi(\alpha) = \frac{C}{V_{\delta x}} e^{\frac{\int \frac{M_{\delta x}}{V_{\delta x}} \, dx}{2}} \]

At equilibrium freq is \( \phi(\alpha) \)

\[ = C e^{\frac{4Ne \delta \alpha}{4Ne \mu - 1} (1-\alpha)} \]
Fixation prob: adaptive mutation

- Diffusion process

\[ \text{Prob of fixation} = u(x) \]
\[ u'(x) = \frac{du(x)}{dx} \]

**Backward Kolmogorov Eqn:**

\[ u(x) = \int u(x + \delta x) q(\delta x, x, \delta t) \delta x \]
\[ u(x + \delta x) \approx u(x) + \delta x u'(x) + \frac{\delta x^2}{2} u''(x) \]

Substituting:

\[ u(x) \approx \int u(x) q(\delta x, x, \delta t) + \int \delta x u'(x) q(\delta x, x, \delta t) \]
\[ + \int \frac{\delta x^2}{2} u''(x) q(\delta x, x, \delta t) \]

\[ A \rightleftharpoons a \]

Under infinite allele model, what are the mutation rates?
Fixation prob of adaptive mutation

\[ M_{\delta x} = \int \delta x \, g(\delta x, x, \delta t) \, \text{d}x \leftarrow \text{MEAN} \]

\[ V_{\delta x} = \int (\delta x)^2 \, g(\delta x, x, \delta t) \, \text{d}x \leftarrow \text{VARIANCE} \]

\[ u(x) = u(x) + u'(x) \, M_{\delta x} + u''(x) \, V_{\delta x} \]

\[ u'(x) \, M_{\delta x} + u''(x) \, V_{\delta x} = 0 \]

\[ s x (1-x) \, u'(x) = -\frac{x(1-x)}{4Ne} \, u''(x) \]

\[ 4Ne \, s \, u'(x) = -u''(x) \]

\[ u(x) = C \cdot e^{-4Ne s x} + D \]

Boundary: \( u(0) = 0, \ u(1) = 1 \)

\[ u(x) = \frac{e^{-4Ne s x} - 1}{e^{-4Ne s} - 1} \]
Rate of evolution

\[ u \left( \frac{1}{2N} \right) \approx \frac{2sNe/N}{1 - e^{-qNe}} \approx \frac{2sNe}{N} \]

= Chance of fixation of new mutation

• Rate of evolution = Rate of observed mutations
  = Rate of mutation X rate of fixation

\[ \approx (2\mu N) \times 2sNe/N \]

= \[ 4\mu sNe \]

RESULT RELATING DRIFT, SELECTION & MUTATION!
Paradigms of selection

• 3 regimes: based on value of $N_e X s$

\[ \sim \ | \ \rightarrow \text{NEUTRAL SELECTION} \]

\[ \gg \ | \ \rightarrow \text{POSITIVE SELECTION} \]

\[ \ll \ | \ \rightarrow \text{NEGATIVE SELECTION} \]

$N_e$ → effective population

$s$ → selectional coefficient,

expected no. of offspring
Neutral Theory

• Motoo Kimura (1968)

• Very large fraction of fixed mutations (both within and between species) are the result of truly random processes (drift) and not of directed selection
  – previously, it was thought natural selection main driver of fixed mutations
  – do not confuse neutral theory with neutral model (which is any model of evolution under no / neutral selection)
Neutral Theory

• Functional sites: Most mutations deleterious and immediately removed by negative selection.

\[ \text{Substitution Rate} = \mu_{\text{neutral}} \]

• -ve correlation between functional significance and substitution rate: more functionally significant \( \rightarrow \) more types of mutations likely to be deleterious / more types of mutations less likely to be fixed \( \rightarrow \) lower neutral mutation rate & lower substitution rate.

\[ \text{Allele freq} = f(\mu_{\text{neutral}}, N_e) \]
Explanation of molecular clock

• Neutral mutation rate is expected to be constant across species and lineage

• Completely random mutations would accrue linearly over time
  – Branch length = Expected no of substitutions = Substitution rate X time = Neutral mutation rate X time = constant X time
Evidence

• For neutral theory: Functionally important sites show lower substitution rates wrt functionally unimportant sites
• For neutral theory: molecular clock
• Against neutral theory: Only accounts for strongly deleterious and neutral mutations. Evidence exists of weakly deleterious mutations.
Selectionist – neutralist debate

- Ohta: Nearly – neutral theory
  - strongly deleterious alleles get wiped out
  - weakly deleterious alleles get fixed under mutation – selection balance

\[ N_e s \approx 1 \text{ or } N_e s < 1 \]

**Allele freq.** = \( f(\mu, s, N_e) \)

- INCREASE IN CR.
- INCREASE IN CR.: ADVANTAGEOUS
- INCREASE IN CR.: NEUTRAL
- DELETING NEUTRAL

**slightly higher** than \( \mu_{\text{NEUTRAL}} \)
Identifying neutrality

• Biggest challenge in using neutral theory: which changes are neutral?

• Question to address: which phenotypes are affected on which natural selection can act?
A complicated situation

• What about mutations in transcription factor binding site?
  – If the change increases binding affinity?
  – If the change decreases binding affinity?
  – If it causes no (negligible) change?

• Difficult to say due to compensatory binding sites nearby: difficult to quantify from binding alone: expression levels of genes need to be observed: still may not be enough
A less complicated situation: codons

- Simple situation: coding region: silent mutations (which do not change the coded amino acid) are termed neutral.

- Other changes are deemed non-neutral.

- For a MSA, no. and nature of mutations need to be figured out on the tree relating the sequences (or averaged over trees).
**Codon table**

- **Synonymous & non synonymous mutations**

<table>
<thead>
<tr>
<th>First Position</th>
<th>Second Position</th>
<th>Third Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>phe</td>
<td>cys</td>
</tr>
<tr>
<td>UUC</td>
<td>leu</td>
<td>c</td>
</tr>
<tr>
<td>UUA</td>
<td>leu</td>
<td>A</td>
</tr>
<tr>
<td>UUG</td>
<td>leu</td>
<td>G</td>
</tr>
<tr>
<td>CUU</td>
<td>leu</td>
<td>U</td>
</tr>
<tr>
<td>CUC</td>
<td>leu</td>
<td>C</td>
</tr>
<tr>
<td>CUA</td>
<td>leu</td>
<td>A</td>
</tr>
<tr>
<td>CUG</td>
<td>leu</td>
<td>G</td>
</tr>
<tr>
<td>AUA</td>
<td>ile</td>
<td>U</td>
</tr>
<tr>
<td>AUG</td>
<td>ile</td>
<td>C</td>
</tr>
<tr>
<td>AUA</td>
<td>ile</td>
<td>A</td>
</tr>
<tr>
<td>AUG</td>
<td>ile</td>
<td>G</td>
</tr>
<tr>
<td>GUU</td>
<td>val</td>
<td>U</td>
</tr>
<tr>
<td>GUC</td>
<td>val</td>
<td>C</td>
</tr>
<tr>
<td>GUA</td>
<td>val</td>
<td>A</td>
</tr>
<tr>
<td>GUG</td>
<td>val</td>
<td>G</td>
</tr>
</tbody>
</table>
Detecting selection in codons

- Goal is to identify regions in genes where rate of amino acid change (rate of non-synonymous mutation) is greater or lesser than the rate of neutral (synonymous) mutation.
dN / dS

- dN = no of non synonymous changes, dS = no of synonymous changes
- Ratio : > , = , < 1 : positive, neutral or negative selection

- How to put probabilities on such hard constraints ?
  - distribution of (dN – dS) for gold standard sets of neutral and non neutral sites
An example

Nonsynonymous

Arg  Gln  Val
AGA  CAA  GTA

A → G Mutation

CAG  CGA  GTA
Arg  Arg  Val

Synonymous

Arg  Gln  Val
AGA  CAA  GTA

AGA  CAG  GTA
Arg  Gln  Val
Counting $dN$ & $dS$

- Another example:

\[
\begin{align*}
\text{CAA} & \quad \text{CAA} \quad \text{GTA} \quad \text{CAA} \quad \text{AGA} \\
\text{CAG} & \downarrow \quad \text{CGA} \quad \text{GTA} \quad \text{CAG} \quad \text{AGA} \\
\text{S} & \quad \text{N} \quad \text{S}
\end{align*}
\]

\[
\frac{dN}{dS} = \frac{1}{2} < 1
\]
Counting dN & dS over a tree

- **MSA**

\[
\frac{dN}{dS} = 1, \text{ so neutral}
\]

\[
dN = 1
\]

\[
dS = 1
\]
COUNTING $dn$ & $ds$

CA A   CAA   CAA
CGA   CAG   CGA

For pairwise, it is simple:

CA A   CAA   CAA
CGA   CAG   CGA

$dn = 2$, $ds = 1$. We are done
For a multiple Seq Alignment, it is not so simple.

\[ \begin{align*}
  x_1 & : \text{CAA} \\
  x_2 & : \text{CGA} \\
  x_3 & : \text{CAG} \\
  x_4 & : \text{CAA}
\end{align*} \]

\{ what \ is \ dn, \ what \ is \ ds? \}

To count mutations, we want to first fix the level at which evolution is operating. [e.g. nucleotide, codon]

We can use nucleotide level
Next we need a model of evolution [ML or Parsimony, e.g.] for simplicity, we assume topology is known.

\[ \begin{align*}
\rightarrow  & \quad x_1 \quad \rightarrow \quad x_3 \\
\rightarrow  & \quad x_2 \quad \rightarrow \quad x_4 \\
\rightarrow  & \quad CAA \quad \rightarrow \quad CAG \quad \rightarrow \quad CAA \\
\rightarrow  & \quad CGA
\end{align*} \]
Now, we need to calculate the ancestral states for each position of alignment.

There are 16 ways (4 x 4) to assign the ancestors. For parsimony, we choose the assignment(s) with min. # of mutations.
These assignments for least no. of mutations may not be unique. For e.g., there may be 3 ways to assign ancestors for position 2, and 2 ways to assign ancestors for position 3.

Now, to put back the nucleotides:

Under independence model of each site:

\[
\begin{align*}
\text{CA} & \rightarrow \text{CAG} = 1 \hspace{1cm} \text{we were lucky, the assignments were unambiguous. Now its easy to count.} \\
\text{CA} & \leftrightarrow \text{GA} = 1 \hspace{1cm} \\
\text{CA} & \leftrightarrow \text{CA} = 1 \hspace{1cm}
\end{align*}
\]
There are 2 ways in which this situation gets more complicated.

- **Ambiguous Assignments**
  - **Pos 1**
    - C  C  C  C
  - **Pos 2**
    - C  C  C  A  G
  - **Pos 3**
    - A  T  T  T  A  A  T  T  A

Texas A&M University

BIOL 6385, Computational Biology
Now, instead of a single \((1 \times 1 = 1)\) potential ancestral assignment, you have \((1 \times 3 \times 2 = 6)\) six potential ancestral assignments. [We assumed independence of sites]

We will get six \(d_{n/ds}\) ratios - which one is right?
B) There is another problem how about branches with multi.
sites changed?

CAT  ____________  CGA
CAT  <->  CGT  <->  CGA
CAT  <->  CAAG  <->  CGA

Ordering of changes may change \(\text{dn/ds ratio}\). So, even for each tree, we have multiple \(\text{dn/ds ratios}\).
This is when you wish you had used maximum likelihood models.

In general,

Problem A: Solution: Use likelihood framework to pick the most likely ancestors (joint or marginal?), or sum over likelihood of trees

Problem B: ML trajectories are difficult to calculate. Easier to do Viterbi analysis of coupon chain of CTMP which models codon evolution
Word of warning

• Remember, a mode of selection over a set of sites does not guarantee that the same mode of selection will operate on a subset of the sites!
McDonald – Kreitman test

- Synthesis: species and population genetic models: test for ancient selectional forces

- Between species and within species dN & dS compared by categorical tests

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>$D_s$</td>
<td>$P_s$</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>$D_n$</td>
<td>$P_n$</td>
</tr>
</tbody>
</table>
Notion behind the MK test

• Deleterious mutations may persist in populations for a few generations due to drift, very unlikely to become fixed.
  – contribute to polymorphism, but not divergence.

• Advantageous / adaptive mutations become fixed in populations pretty fast: contribute little to polymorphism, appear as fixed differences between species.

• Compare no of fixed to polymorphic differences for synonymous and nonsynonymous mutations deviations from the neutral theory can be detected.
So, what can we do with these tools?

• Given initial allele frequency, and selectional coefficients and mutation rates
  – predict probability of fixation and/or equilibrium frequencies

• Given allele frequencies in equilibrium
  – estimate heterozygosity and other notions of genetic variability and estimate effective population size, mutation rates, selection coefficients

• Given alleles and model of change for a set of loci, predict the nature and degree of selection
More complications

- Genetic hitchhiking
- Modelling multiple loci
- Models of recombination
  - linkage between loci

- Polymorphism as a function of recombination rates
What if ...

• We are less interested in the evolutionary parameters
• More interested in the genealogy?
Coalescent theory

- Purely historical, not predictive
- Retrospective, may be generative
- Genealogical tree to MRCA
- (Bad) analogy in phylogenetics: tree reconstruction

John Kingman
Coalescent theory

• Visualization of the coalescent:
  http://www.ucl.ac.uk/tcga/presentations/TCGA_Augss/TCGA_MW_Seminar4.ppt

• Deriving the coalescent:
  http://bio.classes.ucsc.edu/bio107/Class%20pdfs/W05_lecture14.pdf
A few uses: genetic fingerprinting

• Pick a set of loci s.t. no of allelic configurations (genotypic or haplotypic) approaches the no of individuals in the population

  – Not enough selection, and sufficiently high rate of mutation that it is conserved across individuals (effective population is same for all alleles)
A few uses: reconstructing ancestry

• Paternal and maternal lineages: avoid confounding recombination
  – paternal: Y chromosome
  – maternal: mtDNA (mitochondrial eve)

• Distinguishing divergence from gene flow

• Admixture components: relative contributions of founder populations
That’s all, folks!

More reading (on the website)

Comparing different methods:
- Phylogenetic vs pop genetic
- Historic vs predictive
- ML vs Bayesian

etc ...

Which one to use?

Larson, The Far Side
Summary

- Population genetics: Toolkit for understanding a more fine-grained evolutionary picture, merges evolutionary theory with quantitative genetics (population genomics: whole genome view)

- Evolutionary process: cooking pot, alleles: ingredients, drift, mutation, selection, recombination, population structure and migration, stochasticity: recipe

- Changes in allele frequencies: outcome of the process!

- Often, the goal is to observe the outcome and make evidence-driven guesses about missing pieces of the recipe
  - GENEALOGY ESTIMATION AND INFERENCES: identifying evolutionary relationships between individuals and using such relationships for inference: estimating allele genealogy, coalescents, pedigree based inference
  - POPULATION GENETICS: evolutionary forces: mutation rates, selectional model, recombination rate, demography: migratory model, population size
  - ASSOCIATION STUDIES (CLASSICAL GENETICS): genotype – phenotype relationships: phenotype-associated loci, epistasis model, quantitative trait models
(Some) things we didn’t cover

• Gene tree – species tree reconciliations

• Violating W-F models in additional ways: Inbreeding, migration, ancestry & demographic models

• Modelling multi locus dynamics: recombination

• Quantitative genetics
Acknowledgements

• Eric Xing
• Dannie Durand
• Gil McVean