How Can One Tell In What Direction Evolution is Going?

Are most new mutations that are fixed in a population deleterious or advantageous?

Some biologists feel that most evolutionary change is due to the accidental fixation of weakly deleterious mutations.

This is because most mutations are deleterious, and the chance effects of who mates with whom may cause some good genes to be lost.

Can one estimate the fraction of new mutations that are being fixed that are beneficial, using only DNA from current populations?

Chromosomes can be thought of as long strings of DNA, which in turn can be thought of as long strings of nucleotides A, C, G, T.

A gene is a segment of a chromosome that looks something like

— (Reg)Code. . .(Intron). . .Code. . .(Intron). . .Code(End) —

"Reg" stands for regularity regions. Introns have mostly no effect. In reading frames ("Code") for a gene, *amino acids* are coded by consecutive triples of bases that are called *codons*, for example:

ATG	GCA	GAA	GGC	TTT	AAC	TTC	ATT	GGT	ACC	· · ·
Met	Ala	Glu	Gly	Phe	Asn	Phe	Ile	Gly	Thr	· · ·

Proteins are built from strings of amino acids. There are 20 amino acids and 64 codons. Base changes that change the amino acid are called *replacement*. Changes that do not are called *silent*.

Most silent variation is in the 3$^{\rm rd}$ nucleotide position, and is mostly either one of

(GAT,GAC) = Asp   (GAA,GAG) = Glu   or
(GCT, GCC, GCA, GCG) = Ala

In most cases, any change in either of the first two nucleotides changes is *replacement* (that is, it changes the amino acid). For the four codons for Alanine above, all 3$^{\rm rd}$ position changes are *silent*.

Looking for clues: Suppose that we have m DNA sequences from one species at a particular gene and n DNA sequences from another sequences, for example

Two species of Drosophila
Two species of Arabidopsis (a common weed)
Human beings and chimpanzees

Suppose that we have samples of DNA sequences from two closely-related species:

Species 1: ...T...A......A...C...C...
...T...T......A...G...C...
...T...A......A...C...C...
Species 2: ...G...A......C...T...T...
...G...A......C...T...A...
...G...A......C...T...T...

We then collect counts (*McDonald-Kreitman table*):

	mono. at diff. bases	poly. in either sp.	Sum
Replacement	M_r	P_r	T_r
Silent	M_s	P_s	T_s
(Sum)	T_M	T_P	T

An excess of fixed replacements ($M_r > \frac{T_r T_M}{T}$, so that $\frac{M_r}{T} > \frac{T_r}{T}\frac{T_M}{T}$) suggests favorable mutation. Conversely, a deficit ($M_r < \frac{T_r T_M}{T}$) suggests unfavorable mutation.

DNA changes at n gene loci are likely to be too sparse for individual McDonald-Kreitman tables

	mono. at diff. bases	poly. in either sp.
Replacement	M_{ri}	P_{ri}
Silent	M_{si}	P_{si}

at n loci to be statistically significant, with many small cell values for different i. However, one can show (using the *Mantel-Haenszel strata test* for 2 × 2 tables):

Two Drosophila species (*melanogaster* and *simulans*):
$n = 56$ loci, $Z = 6.35$, $P \approx 2 \times 10^{-10}$

Two Arabidopsis species (*thaliana* and *lyrata*):
$n = 12$ loci, $Z = -4.47$, $P \approx 8 \times 10^{-6}$

Thus, evolution seems to be going uphill for the two Drosophila species, but downhill for the two weeds.

However, can we make this quantitative? What are the selection coefficients involved, per generation? How long does it take?

Random changes due to random choices of mates and who survives happen on a time scale of $N = N_e$ generations, where N_e is the effective population size, so that it is natural to scale time in this way.

It is useful to consider five different kinds of mutations, where s is the amount of selection (relative advantage) per generation:

(i)	$s < 0$, $\lvert sN \rvert \gg 1$	Evolutionary lethal
(ii)	$s < 0$, $\lvert sN \rvert = O(1)$	Weakly deleterious
(iii)	$s = 0$	Neutral
(iv)	$s > 0$, $\lvert sN \rvert = O(1)$	Weakly advantageous
(v)	$s > 0$, $\lvert sN \rvert \gg 1$	Hopeful monsters(?)

Evolutionary lethal mutations can be ignored since they rapidly disappear in time scaled by N generations, and hopeful monsters are never polymorphic in this time scale.

We will restrict ourselves to weakly selected mutations, (ii,iii,iv). This ignores the more interesting "hopeful monsters" (v), but these are probably rare.

PROBABILITY THEORY: Let X_k be the frequency of the mutant base in generation k. Thus $X_0 = 1/N$. Most models for "random mating" are equivalent to

$$\{X_{k+1} \mid X_k = p\} \approx \frac{1}{N}\,\text{Binom}\left(N, \frac{(1+s)p}{1-p+(1+s)p}\right)$$

where s is the selective advantage ($s > 0$) or disadvantage ($s < 0$) per generation. We assume $Ns \approx \gamma$, so that $s \to 0$. Then if $p_1 = (1+s)p/(1+sp)$

$$N\ E(X_{k+1} - p \mid X_k = p) = N(p_1 - p)\ \to\ \gamma p(1-p)$$

$$N\ \text{Var}(X_{k+1} \mid X_k = p) = N\frac{p_1(1-p_1)}{N}\ \to\ p(1-p)$$

$$N\ E(|X_{k+1} - p|^3 \mid X_k = p)\ \le\ C/\sqrt{N}\ \to\ 0$$

This means that, in the time scale $t = k/N$, $X_N(t) = X_{[Nt]} \to X(t)$ for a Markov diffusion process $X(t)$ with infinitesimal generator

$$L_p = \frac{1}{2}p(1-p)\frac{d^2}{dp^2} + \gamma p(1-p)\frac{d}{dp}$$

This means that if

$$\Pr(X(t) \in y + dy \mid X(0) = x) = q(t, x, y)\,dy$$

then $(\partial/\partial t)q(t, x, y) = L_x q(t, x, y)$ for

$$L_x = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} + \gamma x(1-x)\frac{\partial}{\partial x}$$

By purely random forces, both the discrete processes X_k and the diffusion process X_t are eventually trapped at either $p = 0$ (the mutant is lost) or at $p = 1$ (the mutant base becomes fixed in the population). From this it follows that

$$\begin{aligned} \phi(p) &= \Pr(\text{Trapped at } 1 \mid X_0 = p) \\ &= \frac{1 - \exp(-2\gamma p)}{1 - \exp(-2\gamma)} \end{aligned}$$

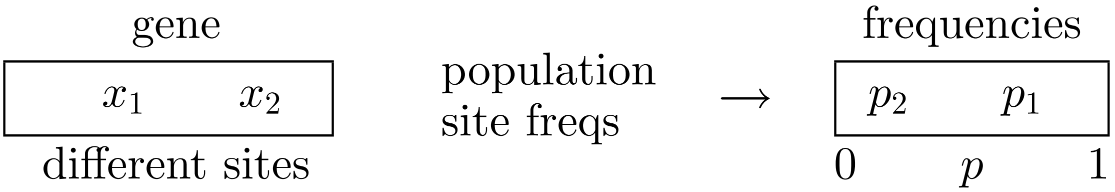
Thus the probability that a particular mutant base is successful should be

$$\phi(1/N) \approx \frac{1}{N}\frac{2\gamma}{1 - \exp(-2\gamma)} \approx \frac{c(\gamma)}{N}$$

Thus most new mutant bases are lost even if $\gamma > 0$ and large, but a number that is proportional to $c(\gamma)/N$ survive to become fixed.

Consider a flux of mutations at rate μ ($= \mu_r$ or $= \mu_s$) at various sites in the gene in various individuals in the population:

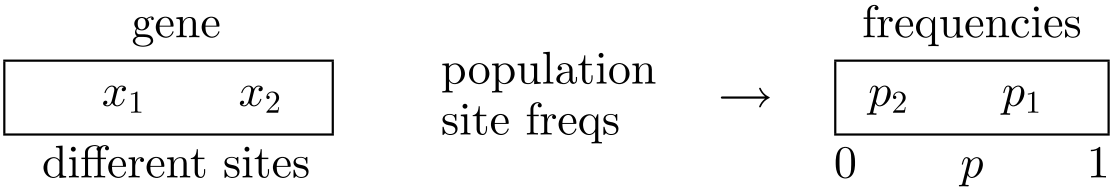
Most new mutant alleles quickly go extinct by drift, but a number survive to have appreciable base frequencies in the population:



Each new mutant (located at some site x_1) eventually satisfies either $p_1 = 0$ (goes extinct) or $p_1 = 1$ (becomes fixed in the population).

Now view the *population frequencies at those sites* for the non-extinct non-fixed mutant bases as a *point process of frequencies* in $(0, 1)$ with an *equilibrium distribution*.

Assume that mutations are so rare that they don't happen twice at the same site, and let $N \to \infty$.


gene

x_1 x_2

different sites

population
site freqs

$\longrightarrow$

frequencies

p_2 p_1

0 p 1

As $N \to \infty$, if we can assume that the processes X_t for different sites at the same locus are independent, the p_i for polymorphic sites at a particular locus form a *Poisson random field*.

This means that the numbers of p_i with $a < p_i < b$ have a Poisson distribution and are independent for nonoverlapping intervals (a, b). The distributions of M_r, P_r for replacement sites and M_s, P_s for silent sites can be determined from this model, and turn out also to be independent Poisson.

The parameters of the limiting model at each locus are:

At silent sites: $\theta_s = N\mu_s =$ silent mutation rate per N generations

At replacement sites: $\theta_r = N\mu_r$ and $\gamma = Ns$: scaled replacement mutation rate and selective advantage or disadvantage per N generations.

The model: We assume

- All new mutations occur at a new site. Silent polymorphisms are neutral. Each new mutant base is subject to constant directional selection, with no epistasis or dominance over sites.

- Sites are unlinked; that is, are statistically independent. (Seems OK by simulation for applications for two related species.)

- At the i^{th} locus, each new replacement mutation has selection coefficient $\gamma \in N(\gamma_i, \sigma_w^2)$ for some γ_i,

- The mean selection coefficient at the i^{th} locus γ_i satisfies $\gamma_i \in N(\mu_\gamma, \sigma_b^2)$

This means that the distribution of γs for new replacement mutations is that of a *random-effects model* in statistics, with between-locus and within-locus variances σ_b^2 and σ_w^2, respectively.

STATISTICS: Under these assumptions, given data $M_{ri}, P_{ri}, M_{si}, P_{si}$, we can write down a *likelihood*

$$L = L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid \mu_\gamma, \sigma_w, \sigma_b, \theta_{si}, \theta_{ri}, \gamma_i, t_{\mathrm{div}})$$

for $1 \le i \le n$, if we have data for n loci. Here t_{div} is the scaled divergence time between the two species. If $n = 56$, there are 172 parameters, of which we are primarily interested in $\mu_\gamma, \sigma_w, \sigma_b, \gamma_i$.

The *maximum likelihood method* says that we should guess those parameter values that maximize L given our data, but we have far too many parameters for numerical maximization methods to work well.

Instead, we will use a *Bayesian technique* called MCMC, for *Markov chain Monte Carlo*. The first step is to assume a "prior distribution" $\pi_0(\theta)$ for $\theta = (\mu_\gamma, \sigma_w, \sigma_b, \theta_{si}, \theta_{ri}, \gamma_i, t_{\mathrm{div}})$ that is a probability distribution in those parameters.

Given our prior distribution $\pi_0(\theta)$ for our parameters θ, the expression $\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si}, \theta)$ is then a joint probability distribution for both our parameters θ as well as our data $M_{ri}, P_{ri}, M_{si}, P_{si}$. We now consider the *conditional* or *posterior distribution*

$$\pi_1(\theta) = C(M,P)\,\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid \theta)$$

where

$$C(M,P) = 1 \Big/ \int \pi_0(z)L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid z)\,dz$$

If we have enough data, this distribution should be concentrated near the true value of θ, and the center of the distribution should not depend on $\pi_0(\theta)$.

Thus we want to find means or median values of various components of $\pi_1(\theta)$. This is a reasonably tractable expression of θ except for the hideously complicated normalizing constant $C(M,P)$, which is here a 172-dimensional integral that does not simplify.

A particular solution to this problem — of finding integrals or median values of a moderately complicated expression $\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si}, \theta)$ times an impossibly complicated normalizing constant — was found by Metropolis *et al.* (1953), and is based on three ideas.

The first idea is to look for a Markov chain Z_n on θ-space (here part of R^{172}) that has $\pi_1(\theta)$ as a stationary measure. If the Markov chain is ergodic, we can estimate the conditional distribution of the parameters θ given our data by considering the sample distribution of a single very long sample path of Z_n.

The second idea is due to Metropolis (1953): Given any Markov-chain transition function $q(\theta_1, \theta_2)$ on θ-space that is symmetric in θ_1 and θ_2, they show how to modify $q(\theta_1, \theta_2)$ in a simple way to form a second Markov-chain transition function $q_1(\theta_1, \theta_2)$ such that $q_1(\theta_1, \theta_2)$ had $\pi_1(\theta)$ as a stationary measure. Hastings (1970) removed the condition of symmetry on $q(\theta_1, \theta_2)$: The resulting slightly more-complicated procedure is called the Metropolis-Hasting algorithm.

The third idea is, for parameters $\theta \in R^{172}$, is break up each step of the Markov chain $Z_n \in R^{172}$ into a sequence of 172 single steps of one-dimensional Markov chains for each component of θ. If each of the one-dimensional Markov chains is ergodic in one dimension, then one can usually show that the resulting 172-dimensional Markov chain is ergodic in R^{172}.

Geman and Geman (1984) introduced the idea, for updating the i^{th} component of θ or Z_n, of sampling from the conditional distribution of that component given the current value of all of the other components as well as the data.

This idea, which is a special case of the Metropolis-Hastings algorithm, carries the colorful name of *Gibbs Sampler*. In practice, many MCMC algorithms are run using Gibbs' sampler steps for some components of $\theta = Z_n$ and Metropolis random-walk steps for other components. In some cases, key components are so highly correlated that this does not work. In that case, groups of components can be updated together.

Our procedure for the $n = 56$ Drosophila loci was, first, to run Z_n for 10,000 steps to *burn in* or "*average out*" the 172 components of Z_n. After the burnin, a single long run of 1,000,000 steps was split into 10 consecutive blocks or subchains of 100,000 steps each. The idea is that if parameter estimates from the 10 subchains are similar, then one has confidence that the process has converged.

The process did converge under the additional condition that $\theta_{ri}/(2\theta_{si}) = q$ was constant across loci. This assumption is that the proportion of replacement sites that are not "evolutionary lethal" was constant across loci. Median and "95% credible intervals" (the middle 95% of the posterior distribution) for the final subchain were

μ_γ :	-5.74	(-20.67, -0.34)
σ_b :	5.41	(3.70, 8.46)
σ_w :	6.20	(2.87, 12.73)
$\sigma_b/(\sigma_b + \sigma_w)$:	0.47	(0.37, 0.61)
$q = \theta_{ri}/2\theta_{si}$:	0.14	(0.09, 0.32)
t_{div} :	2.47	(2.18, 2.80)

Under these conditions, averaging over all 56 loci, the expected proportions of advantageous (non-deleterious) mutations among replacement mutations are

New (nonlethal) mutations	19%
Polymorphic in samples	47%
Fixed differences	93%

This shows that, at least for these two Drosophila species, evolution has proceeded as Darwin would have expected.

An Epilogue: The reason that I considered this "work in progress" is that my co-authors have more recently sent me a new set of 78 Drosophila loci with 3 species and then a "better behaved" dataset with 112 loci. The first did not converge in 1,000,000 steps, even with the assumption on $\theta_{ri}/(2\theta_{si}) = q$, but does converge in $n =$ 20,000,000 iterations without assuming $\theta_{ri}/(2\theta_{si}) = q$.

The reasons why the first datasets are "badly behaved" are illustrated in the next few graphics.

Thank you for coming.