The creation-mutation-selection model: mutation rates and effective population sizes

Gordon Irlam
gordoni@gordoni.com

Los Altos, California, United States

July 16, 2023

[***Pre-publication draft – feedback solicited***]

Abstract

The creation-selection-mutation model makes predictions regarding the fitness of asexual and sexual populations in an environment that incorporates both positive and negative selection. The model predicts the optimal spontaneous mutation rate for a sexual population as one in which the fitness losses associated with positive and negative selection are equal. The model depends upon three mutation related rates: the rate of adaptive mutational opportunities, the rate of negative mutational site creation, and the spontaneous mutation rate. These three mutation related rates are estimated based on a comparison of substitution rates at nonsynonymous and synonymous sites in the genomes of related eukaryotic species. For eukaryotes, the rate of adaptive mutation opportunities is found to typically be in the range 10^{-3} to 10^{-2} population wide adaptive mutational opportunity sites per sexual generation. Negative sites are typically created at the rate 10^{-1} to 10^{1} sites per haploid genome per sexual generation. And the spontaneous mutation rate is typically in the range 10^{-9} to 10^{-8} spontaneous mutations per creation-mutation-selection model site per sexual generation. Effective population sizes are also computed based on the assumption of optimal mutation rates. That effective population sizes appear reasonable, adds some evidence to the claim that evolution tunes the mutation rate towards a near optimal value.

Keywords: adaptive mutation rate, deleterious mutation rate, spontaneous mutation rate, optimal mutation rate.

Introduction

The creation-selection-mutation model is a mathematical model for the fitness of asexual and sexual populations in the presence of positive and negative selection[1]. The model depends on three mutation related rates: the rate of population wide adaptive mutational opportunity site creation, the per organism rate of negative mutational opportunity site creation, and the spontaneous mutation rate at these sites. Excluding neutral sites, these three rates are denoted Γ_p^*, Γ_n^*, and μ_{ss}, respectively, and are expressed as rates per asexual or sexual generation. For finite genome sizes, there are really only two independent mutation related rates; Γ_n^* is determined by the length of the genome under the control of negative selection and the spontaneous mutation rate.

By comparing substitution rates at nonsynonymous and synonymous sites in the genomes of relatively recently diverged sexual species with a known divergence time and a known time between sexual generations, the three mutation related rates will be estimated. The motivation being with these rates in hand it should then be possible to compute the advantage of sex.

In the creation-selection-mutation model, the optimal spontaneous mutation rate for sexual populations is one for which the fitness losses from positive and negative selection are equal. Too low a mutation rate, and adaptive sites will take too long to fix. Too high a mutation rate, and negative selection will exert a heavy toll. Effective population sizes are computed based on the assumption of optimal mutation rates. For animals, these effective population sizes are found to be eminently reasonable, adding support to the hypothesis that evolution tunes the mutation rate for maximal fitness.

Results

The aim of this study is to determine the approximate range of the key mutational parameters, not their precise values. For particular species others may have come up with more precise estimates of some of the parameter values, but this is not known to have been previously done with a consistent framework that spans species and parameter values. To simplify things it is assumed spontaneous mutations occur at random across the full length of the genome.

In the creation-selection-mutation model a site is a binary entity that is first created, and subsequently satisfied. This is different from coding sites on the genome, where a site is a base pair that can take on any one of four possible values. Which type of site is being discussed should be obvious from the context.

Method of parameter estimation

Consider two related species. Let A and S denote the number of nonsynonymous and synonymous sites that are either in common between two orthologous genes, or alternatively the coding portions of the genomes, of the two species. Let D_n and D_s be the number of nonsynonymous and synonymous substitutions occurring between the two genes or genomes. Let $K_a = \frac{D_n}{A}$ (also known as d_N) and $K_s = \frac{D_s}{S}$ (also known as d_S).

The spontaneous mutation rate

Let g be the generation time, and T be the time since the two species diverged. Let μ_{by} be the DNA mutation rate for the species per base pair per unit of time. Then the mutation rate per base pair per sexual generation, μ_{bs}, is given by,

$$\mu_{bs} = \mu_{by} g \tag{1}$$

μ_{by} can be estimated from the genome wide synonymous mutation rate,

$$\mu_{by} = \frac{K_s}{2T} \tag{2}$$

Since there is assumed to be only one correct mutation for a given adaptive mutational opportunity, with the other two mutations leaving the site on average no better and no worse off, the rate of satisfying mutations per sexual generation, μ_{ss}, is,

$$\begin{aligned} \mu_{ss} &= \frac{1}{3}\mu_{bs} \\ &= \frac{1}{3}\frac{K_s g}{2T} \end{aligned} \tag{3}$$

Positive selection

Let α be the fraction of nonsynonymous substitutions that are positively selected; as opposed to being neutral nonsynonymous substitutions. In the long run in our model, the rate of site creation is equal to the rate of substitution. Consequently, the per generation rate of site creation is,

$$\Gamma_p^* = \alpha \frac{A' K_a g}{2T} \tag{4}$$

where A' is the value of A adjusted to take into account nonsynonymous sites in the genome that weren't analyzed.

Estimating alpha

It is possible to estimate α, if, in addition to estimates of the number of substitutions, data on polymorphisms are available[2, 3]. Table 1 shows a sampling of estimates for α. As can be seen there is considerable variability regarding the value of α. For animals, α has an approximate mean value of 0.6 and a standard deviation of 0.2.

Species	Estimate of α	Year	Source
Homo sapiens and old world monkeys	0.35	2001	[4]
Homo sapiens and *Pan troglodytes*	0.10 - 0.13	2007	[5]
Homo sapiens and *Pan troglodytes*	0.10 - 0.20	2008	[6]
Homo sapiens and macaques	0 or 0.31 - 0.40	2009	[7]
Mus musculus and *Mus famulus* or *Rattus*	0.57	2010	[8]
Drosophila simulans and *Drosophila yakuba*	0.45	2002	[2]
Drosophila melanogaster and *Drosophila simulans*	0.29	2007	[9]
Drosophila americana and *Drosophila ezoana*	0.57	2007	[10]
Drosophila miranda and *Drosophila pseudoobscura*	0.44 - 0.61	2008	[11]
Drosophila melanogaster and *Drosophila simulans*	0.52	2009	[7]
44 animal species pairs	0.58 ± 0.20	2016	[12][Supplement 1]
9 out of 10 plant species	0.0	2010	[13]

Table 1: Some estimates of α the fraction of positively selected nonsynonymous substitutions.

Negative selection

Random deleterious mutations to the genome can be corrected either by purifying selection or back mutation. Back mutation is likely to be rare, so we focus on purifying selection. We model deleterious mutations as occurring at a rate Γ_n^* per organism lineage.

Let a_n^* be the fraction of nonsynonymous mutations that are truly deleterious. Truly deleterious mutations are both deleterious and non-neutral. Assuming mutations are distributed randomly across sites, this is the same as the fraction of the nonsynonymous coding sites that are being maintained by negative selection. That is, mutations of the nonsynonymous sites are neither neutral nor beneficial. Let n_a be the fraction of nonsynonymous sites in the fraction of the genome under the control of negative selection. n_a will be less than 1 if non-coding regions are under the control of negative selection. The per generation rate at which deleterious mutations are occurring is given by,

$$\begin{aligned}\Gamma_n^* &= \frac{a_n^* A' \mu_{by} g}{n_a} \\ &= \frac{a_n^* A' K_s g}{2 n_a T} \text{ by equation 2}\end{aligned} \tag{5}$$

Estimating the nonsynonymous deleterious mutation fraction

To compute Γ_n, we need to come up with an estimate for a_n^*, the fraction of nonsynonymous mutations that are truly deleterious.

Let a_p^* be the fraction of nonsynonymous mutations that are truly beneficial. Let N be the population size. Let $\mathrm{E}_{mut}[s_p^*]$ be the mean selection coefficient for new true positive mutations. For a sexual population in the creation-mutation-selection model, the fixation probability of new mu-

tations is approximately equal to the selection coefficient[1]. Consequently the rate of fixation of beneficial mutations per unit time for a particular site is $Na_p^*\mu_{by}\,\mathrm{E}_{mut}[s_p^*]$. The beneficial fixation rate for a particular site is also $\frac{\alpha K_a}{2T}$. Consequently,

$$\begin{aligned} a_p^* &= \frac{\alpha K_a}{2T\mu_{by}N\,\mathrm{E}_{mut}[s_p^*]} \\ &= \frac{\alpha K_a}{K_s N\,\mathrm{E}_{mut}[s_p^*]} \text{ by equation 2} \end{aligned} \tag{6}$$

α and $\frac{K_a}{K_s}$ are both less than 1. Some theoretical and experimental work suggests the distribution of fitness effects of new beneficial mutations is exponential with small fitness effect mutations being more common than large effect mutations[14, 15], while other experimental work rejects this hypothesis and suggests fitness effects might follow a normal distribution[16, 17]. For the exponential distribution a mean fitness effect of 0.087 has been reported[15]. This was for asexual bacteria. For sexual eukaryotes with their larger genomes, the mean fitness effect might be quite a bit smaller, but even then, given the impact of N in equation 6, a_p^* is likely to be very close to zero. For a normal distribution the mean fitness effect is likely to be larger, making a_p^* even closer to zero.

$$a_p^* \approx 0$$

The fraction of nonsynonymous substitutions that are neutral is $1-\alpha$, and neutral substitutions occur at the rate $\frac{K_a}{2T}$, giving a per site nonsynonymous neutral substitution rate, k,

$$k = (1-\alpha)\frac{K_a}{2T}$$

Let the fraction of nonsynonymous mutations that are neutral be a_0. According to the neutral theory, the neutral mutation rate is equal to the neutral substitution rate[18],

$$\begin{aligned} a_0\mu_{by} &= k \\ a_0 &= (1-\alpha)\frac{K_a}{K_s} \text{ by equation 2} \end{aligned}$$

Since mutations are either true positives, true negatives, or neutral,

$$\begin{aligned} a_p^* + a_n^* + a_0 &= 1 \\ a_n^* &\approx 1-(1-\alpha)\frac{K_a}{K_s} \end{aligned} \tag{7}$$

Over an entire genome $\frac{K_a}{K_s}$ will almost certainly be less than 1. It follows then that a_n^* will be greater than α.

Species	Lesser genes	Genes aligned	A	K_a	K_s	$\frac{K_a}{K_s}$	A'
Homo sapiens and *Pan troglodytes*	19,932	17,678	2.2×10^7	0.0034	0.014	0.25	2.7×10^7
Mus musculus and *Rattus norvegicus*	22,517	18,162	2.2×10^7	0.031	0.19	0.16	2.7×10^7
Gallus gallus and *Phasianus colchicus*	16,248	14,312	1.9×10^7	0.017	0.11	0.16	2.1×10^7
Xenopus laevis and *Xenopus tropicalis*	21,885	17,864	2.3×10^7	0.049	0.29	0.17	2.8×10^7
Oryzias latipes and *Nothobranchius furzeri*	22,145	17,031	2.2×10^7	0.11	0.98	0.11	2.9×10^7
Drosophila simulans and *Drosophila yakuba*	14,217	12,854	1.6×10^7	0.035	0.29	0.12	1.8×10^7
Plasmodium vivax and *Plasmodium gonderi*	5,389	3,343	3.1×10^6	0.15	4.17	0.035	8.3×10^6
Arabidopsis thaliana and *Camelina sativa*	27,271	21,557	2.0×10^7	0.055	0.27	0.20	2.5×10^7
Elaeis guineensis and *Cocos nucifera*	26,295	16,665	1.6×10^7	0.033	0.11	0.30	2.4×10^7
Populus trichocarpa and *Hevea brasiliensis*	31,543	15,327	1.5×10^7	0.13	0.74	0.17	3.3×10^7

Table 2: Genome wide estimates of coding sites and substitution rates. Lesser genes is the lesser number of genes of the two species. A is the numbers of nonsynonymous sites. K_a and K_s are the number of substitutions per nonsynonymous and synonymous site respectively. A' is the number of nonsynonymous sites adjusted for genes and segments of genes that weren't analyzed.

Parameter estimation

Comparison of genomes

To estimate Γ_p^*, Γ_n^*, and μ_{ss}, a number of relatively recently diverged species pairs were chosen. Species were selected based on the availability of sequenced genomes, availability of estimates of divergence times, and availability of organism generation times. The pairs need to have diverged relatively recently so that the average generation time is meaningful. It is assumed that the number of within species polymorphisms is small in comparison to the number of between species substitutions, so that all differences between genomes can be considered to represent substitutions. Many of the selected species are model organisms. Model organisms are often proposed on the basis of their short generation times. This might introduce a slight bias leading to underestimates of typical values for Γ_p^*, Γ_n^*, and μ_{ss}.

For each species pair orthologous genes were identified using protein-protein BLAST to determine reciprocal best hits. Protein sequences of orthologous genes were then aligned using the Needleman-Wunsch algorithm. Aligned protein sequences were mapped back to aligned nucleotide sequences. Genes containing tandem repeat regions were excluded. Estimates of A, S, K_a, and K_s for each gene pair were made. Genome wide estimates of A and S were computed as the sums of individual gene pairs, and genome wide estimates of K_a and K_s computed as A and S weighted averages of the values of the individual gene pairs. Estimates of Γ_p^*, Γ_n^*, and μ_{ss} are likely underestimates on account of regions of the genome with the greatest variability not aligning and being excluded from the analysis. This becomes increasingly significant for longer divergence times. See Materials and methods for a more detailed description of the methodology. The results are shown in Table 2.

The K_a and K_s values for humans and chimps of 0.0034 and 0.014 compare reasonably well to previously reported values of 0.0029 and 0.013 respectively[19][Supplement S23, site weighted K_a and K_s values divided by $2T$].

To account for unanalyzed genes the value of A was multiplied by the smaller of the coding sequence sizes for the two species divided by the analyzed coding sequence size giving A'. Use of the smaller

coding sequence size is important because while most analyzed species are diploid, *Xenopus laevis* is tetraploid[20], and *Camelina sativa* is hexaploid[21].

Estimating the nonsynonymous fraction in the fraction of the genome under the control of negative selection

To compute Γ_n^* we need to estimate n_a, the fraction of nonsynonymous sites in the fraction of the genome under the control of negative selection.

Let L be the length of the genome in base pairs, and l_n be the fraction of L under the control of negative selection. In theory n_a can be computed from,

$$\begin{aligned} a_n^* A' &= n_a l_n L \\ n_a &= \frac{a_n^* A'}{l_n L} \end{aligned} \tag{8}$$

For humans and chimps $\frac{K_a}{K_s} = 0.25$, and from Table 1, $\alpha \approx 0.15$ leading by equation 7 to $a_n^* = 0.79$. For humans $l_n = 0.054$[22], $L = 3.1{\times}10^9$, and $A' = 2.7{\times}10^7$. By equation 8 this results in a value for n_a of 0.13. For many other species l_n is unknown. Not knowing any better, we assume that the same value of n_a applies for most other species.

Plasmodium spp. have very small compact genomes, $L = 3.0{\times}10^7$. $\frac{K_a}{K_s} = 0.035$, so assuming $\alpha = 0.58$, gives $a_n* = 0.99$. $A' = 8.3{\times}10^6$, implies $n_a l_n = 0.27$, which if $n_a = 0.13$ implying $l_n > 1$, which is impossible. We naively assume $n_a = l_n = 0.52$ for *Plasmodium* spp.

Calculation of the mutation related rates

Using equation 7 it is possible to compute estimates for a_n^*. Estimates of α are based on Table 1, except that for plants we assume an α of 0.1 rather than the somewhat implausible value of 0.0. An α of 0.1 appears within the 95% confidence interval of 8 of the 10 plants reported by Gossmann et al.[13]. Using equations 4, 5, and 3, it is then possible to compute estimates for Γ_p^*, Γ_n^*, and μ_{ss}. The results are presented in Table 3.

For the considered species pairs excluding plants, Γ_p^* is in the range 0.0012 to 0.026, or roughly 10^{-3} to 10^{-2}. Γ_n^* is in the range 0.16 to 4.2, or roughly 10^{-1} to 10^1. And μ_{ss} ranges from $4.3{\times}10^{-10}$ to $1.3{\times}10^{-8}$, or roughly 10^{-9} to 10^{-8}.

The estimate $\Gamma_p^* = 0.0012$ for *Drosophila simulans* and *Drosophila yakuba* can be compared to a published estimate of a rate of adaptive substitution of 0.0022 for the same species based on an analysis of 35 genes[2].

Table 4 shows $\frac{\Gamma_p^*}{g}$, the rate of adaptive site creation per unit time, μ_{bs}, the rate of spontaneous mutation per base pair sexual generation, and μ_{by}, the rate of spontaneous mutation per base pair per unit time.

Species	T	g	α	n_a	a_n^*	Γ_p^*	Γ_n^*	μ_{ss}
Homo sapiens and *Pan troglodytes*	6.7×10^6	25	0.15	0.13	0.79	0.026	4.2	8.5×10^{-9}
Mus musculus and *Rattus norvegicus*	20.9×10^6	0.5	0.57	0.13	0.93	0.0057	0.45	7.7×10^{-10}
Gallus gallus and *Phasianus colchicus*	34.1×10^6	1	0.58	0.13	0.93	0.0030	0.24	5.3×10^{-10}
Xenopus laevis and *Xenopus tropicalis*	64.0×10^6	3	0.58	0.13	0.93	0.019	1.4	2.3×10^{-9}
Oryzias latipes and *Nothobranchius furzeri*	93.0×10^6	1.5	0.58	0.13	0.95	0.015	1.7	2.6×10^{-9}
Drosophila simulans and *Drosophila yakuba*	11.4×10^6	0.1	0.45	0.13	0.93	0.0012	0.16	4.3×10^{-10}
Plasmodium vivax and *Plasmodium gonderi*	9.5×10^6	0.18	0.58	0.52	0.99	0.0068	0.62	1.3×10^{-8}
Arabidopsis thaliana and *Camelina sativa*	9.4×10^6	0.2	0.10	0.13	0.82	0.0015	0.46	9.7×10^{-10}
Elaeis guineensis and *Cocos nucifera*	43.0×10^6	50	0.10	0.13	0.73	0.046	8.7	2.2×10^{-8}
Populus trichocarpa and *Hevea brasiliensis*	80.0×10^6	25	0.10	0.13	0.85	0.066	25	3.9×10^{-8}

Table 3: Estimates of a_n^* and mutation related rates. T is the time since the species diverged in years. g is the estimated generation time in years. α is the proportion of substitutions that are adaptive. n_a is the fraction of nonsynonymous sites in the fraction of the genome under negative selection. a_n^* is the proportion of nonsynonymous coding sites that are maintained by negative selection. Γ_p^* is the rate of true positive site creation per generation. Γ_n^* is the rate of true negative site creation per haploid genome per generation. μ_{ss} is the rate of mutation per creation-mutation-selection model site per generation.

Species	$\frac{\Gamma_p^*}{g}$	μ_{bs}	μ_{by}	N_e
Homo sapiens and *Pan troglodytes*	0.0010	2.5×10^{-8}	1.0×10^{-9}	7.4×10^5
Mus musculus and *Rattus norvegicus*	0.0115	2.3×10^{-9}	4.6×10^{-9}	1.7×10^7
Gallus gallus and *Phasianus colchicus*	0.0030	1.6×10^{-9}	1.6×10^{-9}	2.4×10^7
Xenopus laevis and *Xenopus tropicalis*	0.0063	6.9×10^{-9}	2.3×10^{-9}	5.9×10^6
Oryzias latipes and *Nothobranchius furzeri*	0.0101	7.9×10^{-9}	5.2×10^{-9}	3.4×10^6
Drosophila simulans and *Drosophila yakuba*	0.0122	1.3×10^{-9}	1.3×10^{-8}	1.7×10^7
Plasmodium vivax and *Plasmodium gonderi*	0.0375	4.0×10^{-8}	2.2×10^{-7}	8.2×10^5
Arabidopsis thaliana and *Camelina sativa*	0.0074	2.9×10^{-9}	1.5×10^{-8}	3.3×10^6
Elaeis guineensis and *Cocos nucifera*	0.0009	6.5×10^{-8}	1.3×10^{-9}	2.5×10^5
Populus trichocarpa and *Hevea brasiliensis*	0.0026	1.2×10^{-7}	4.6×10^{-9}	6.8×10^4

Table 4: Estimates of additional parameters. $\frac{\Gamma_p^*}{g}$ is the rate of true positive site creation per year. μ_{bs} is the rate of mutation per base pair per sexual generation. μ_{by} is the rate of mutation per base pair per year. N_e is the effective population size assuming the optimal mutation rate.

Species	μ_{bs}	Source
Homo sapiens	2.0×10^{-8}	[27][Table 4, $C\mu_b$]
Homo sapiens	2.5×10^{-8}	[28]
Mus musculus	3.8×10^{-9}	[29]
Mus musculus	5.7×10^{-9}	[30][mean value]
Mus musculus	1.1×10^{-8}	[27][Table 4, $C\mu_b$]
Drosophila melanogaster	2.8×10^{-9}	[31]
Drosophila melanogaster	3.5×10^{-9}	[32]
Drosophila melanogaster	8.5×10^{-9}	[27][Table 4, $C\mu_b$]

Table 5: Estimates of the spontaneous mutation rate, μ_{bs}, by various authors. μ_{bs} is the rate of mutation per base pair per sexual generation.

The rate of adaptive site creation roughly ranges from one every thirty years to one every thousand years.

Plasmodium has a very high spontaneous mutation rate per sexual generation. The average mutation rate is reported elsewhere to be 1.7×10^{-9} per generation[23]. However, the *Plasmodium* life cycle involves at least 200 generations per year[24] but only one sexual generation every 65 days[25]. This would imply a spontaneous mutation rate per sexual generation of at least 6.1×10^{-8} and a spontaneous mutation rate per year of 3.4×10^{-7}. This compares to the estimates made here of 4.0×10^{-8} and 2.2×10^{-7} respectively. This high rate of spontaneous mutation may go some way to explaining why *Plasmodium* is able to rapidly evolve drug resistance[26].

μ_{bs} roughly ranges from 10^{-9} to 10^{-7}. Some of the reported mutation rates can be compared to estimates reported by others as shown in Table 5. While similar in magnitude, a discrepancy of more than a factor of two exists between the estimate made here and the lowest value reported by others for *Drosophila* spp.

Excluding *Plasmodium*, μ_{by} roughly ranges from 10^{-9} to 10^{-8}.

Effective population sizes assuming optimal mutation rates

Even though deleterious mutations appear to occur per genome at roughly one hundred times the rate of adaptive mutational opportunities, deleterious mutations have a more fleeting existence because alleles to satisfy them already exist in the population, and so they each individually exert a smaller fitness cost.

Assuming evolution drives spontaneous mutation rates towards the value that produces the maximum population fitness, the fitness losses coming from positive selection will exactly equal those coming from negative selection[1]. Then the optimal mutation rate is given by[1],

$$\mu_{bs} \approx \frac{3\Gamma_p^*}{N\Gamma_n^*} \tag{9}$$

This equation can be trivially rearranged to give N_e, the effective population size implied by μ_{bs}, Γ_p^*, and Γ_n^*, assuming the optimal per site mutation rate.

$$N_e \approx \frac{3\Gamma_p^*}{\Gamma_n^* \mu_{bs}}$$

Using equations 2, 1, 4, and 5.

$$N_e \approx \frac{6\alpha n_a K_a T}{a_n^* K_s^2 g} \tag{10}$$

The resulting estimates of N_e, computed using 10 are shown in Table 4.

If the hypothesis that spontaneous mutation rates are evolutionarily tuned towards producing a fitness level that maximizes the ability to adapt by positive selection while minimizing the cost of negative selection was false we could get any sorts of random values out of equation 10. We don't. Most of the values appear eminently reasonable. Thus the hypothesis appears to be a reasonably good hypothesis.

The effective population for *Plasmodium* spp. is 8.2×10^5, which may on first glance appear smaller than expected, and much smaller than the actual population, but it must be remembered that *Plasmodium* undergoes a severe population bottleneck every time it is transmitted from host to host. Another way of looking at things is a small effective population size and a small size of the portion of the genome under the control of negative selection explains why *Plasmodium* has such a high rate of spontaneous mutation. These two terms appear as the denominator in the formula for the optimal mutation rate[1],

$$\mu_{bs} \approx \sqrt{\frac{3\Gamma_p^*}{N l_n L}}$$

The effective population size for *Elaeis guineensis* (African oil palm) and *Cocos nucifera* (coconut palm), and *Populus trichocarpa* (black cottonwood tree) and *Hevea brasiliensis* (rubber tree) are small. This may be because the fixed physical location of plants might limit the extent of random mating. Or perhaps something unexplained is going on with plants; they also reportedly have low values for α.

Discussion

Genomic analysis suggests that for sexual species Γ_p^* is typically in the range 10^{-3} to 10^{-2} population wide adaptive mutational opportunity sites per sexual generation. Γ_n^* is typically in the range 10^{-1} to 10^{1} negative sites per haploid genome per sexual generation. And μ_{ss} is typically in the range 10^{-9} to 10^{-8} spontaneous mutations per creation-mutation-selection model site per sexual generation. This is the area of the parameter space of interest when seeking to assess the advantage of sex.

For animals effective population sizes computed under the assumption of optimal mutation rates appear eminently reasonable, suggesting that evolution tunes the spontaneous mutation rate to produce optimal fitness. Whether this also applies to plants isn't clear.

Materials and methods

See Supplement 1 for the bioinformatics analysis code and results of the gene by gene analysis of each species pair[33].

The coding sequence (CDS) of genes from sequenced species genomes were obtained from NCBI. Genomes were filtered to remove any non-nuclear genes, documented pseudogenes, or duplicate protein isoforms. At this stage multiple distinct protein isoforms of each gene were potentially present. This is important. When seeking to assess the divergence of closely related genomes, filtering out shorter isoforms of each gene might have resulted in the longest isoform of orthologous genes in each species having few exons directly in common but still being related. This would have led to undercounting of orthologous sites and misestimating substitution rates.

Reciprocal best hits were computed using protein-protein BLAST[34] using the task blastp-fast, the BLOSUM90 matrix, soft masking of low complexity regions, a minimum expect value of 10^{-6}, and reporting the single best high-scoring segment pair. For each gene of the first species having multiple isoforms with reciprocal best hits, the reciprocal best hit of the isoform with the highest number of matching residues was then selected to represent an orthologous gene pair.

Protein sequences were aligned to each other using the Needleman-Wunsch[35] algorithm implemented by Biopython PairwiseAligner[36]. Scores of match 5, mismatch -8, gap open -50, end gap open -25, gap extend -2, and end gap extend -1, were found to do a good job of conservatively predicting the alignments for most orthologous gene pairs examined. Nucleotide sequences were then aligned based on the protein sequences. Stop residues '*', unknown nucleotides 'X', and leading gaps '-' were not aligned, but the remainder of the protein sequence was aligned.

Mutation rates may vary over the genome. For each orthologous gene pair PAML's CODEML[37] was used to count the number of synonymous and nonsynonymous sites and estimate the number of substitutions that have occurred. Rarely CODEML failed to deliver a result within its search bounds, producing an estimate of 50.0 for the separation time. Such results were excluded from the analysis.

Genes containing tandem repeats may bias the analysis due to tandem repeat copy number polymorphism and random gene conversion within such genes. So Tandem Repeat Finder[38] was used on the nucleotide sequence to exclude such genes from the analysis. Default recommended parameter values were used, match 2, mismatch 7, delta (indels) 7, PM 80, PI 10, minscore 50, except maxperiod was increased from 500 to 2,000 to ensure identifying the human FLG gene which has repeats of period 972.

Estimates of K_a and K_s are highly variable for small numbers of residues. Only genes having an aligned protein length of 20 residues or more with a Tandem Repeat Finder score of 100 or less were selected for inclusion in the resulting analysis. A score of more than 100 corresponds to more than 16 residues making up any discovered total tandem repeat region length.

Species	Accession
Homo sapiens	GCF_000001405.39
Pan troglodytes	GCF_002880755.1
Mus musculus	GCF_000001635.27
Rattus norvegicus	GCF_015227675.2
Gallus gallus	GCF_000002315.5
Phasianus colchicus	GCF_004143745.1
Xenopus tropicalis	GCF_000004195.4
Xenopus laevis	GCF_017654675.1
Oryzias latipes	GCF_002234675.1
Nothobranchius furzeri	GCF_001465895.1
Drosophila simulans	GCF_016746395.2
Drosophila yakuba	GCF_016746365.2
Plasmodium vivax	GCF_000002415.2
Plasmodium gonderi	GCF_002157705.1
Arabidopsis thaliana	GCF_000001735.4
Camelina sativa	GCF_000633955.1
Elaeis guineensis	GCF_000442705.1
Cocos nucifera	GCA_008124465.1
Populus trichocarpa	GCF_000002775.4
Hevea brasiliensis	GCF_001654055.1

Table 6: Accession numbers of genomes used in this study.

Genome wide estimates of A and S were produced by summing the individual gene values. K_a and K_s for the genome were computed as the A and S weighted averages of the individual gene values.

Divergence times were estimated using TimeTree[39].

Noting that laboratory lifespans are typically longer than wild lifespans, generation times were estimated as follows. *Homo sapiens* and *Pan troglodytes*, 25 years[40]. *Mus musculus* and *Rattus norvegicus*, 6 months based on reproductive life spans of 7 to 8 months and 12 to 15 months respectively. *Gallus gallus* and *Phasianus colchicus*, 1 year for *Gallus*[41]. *Xenopus laevis* and *Xenopus tropicalis*, 3 years based on sexual maturity of 12 to 18 months and 4 to 6 months and laboratory lifespans of 10 to 15 years and 10 years respectively[42, 43, 44]. *Oryzias latipes* and *Nothobranchius furzeri*, 1.5 years, based on 50% laboratory mortality after 3 years and exactly 1 year respectively[45, 46]. *Drosophila simulans* and *Drosophila yakuba*, 0.1 years[2]. *Plasmodium vivax* and *Plasmodium gonderi* 65 days, based on *Plasmodium falciparum* and *Plasmodium reichenowi*[25]. *Arabidopsis thaliana* and *Camelina sativa* 0.2 year, based on a life cycle of as little as 6 weeks and a crop season of 85-100 days respectively. *Elaeis guineensis* and *Cocos nucifera* 50 years, based on economic lifespan of 30 years and lifespan of 100 years or more and a lifespan of 80 to 100 years respectively. *Populus trichocarpa* and *Hevea brasiliensis* 25 years, based on being suitable for timber production after 25 years and an economic life of 25 years respectively[47]. It should be noted that *P. trichocarpa* and *H. brasiliensis* are both Malpighailes, but so are annuals of the genus *Viola*. If the ancestors of these trees were short lived, the generation time would be much smaller.

Accession numbers of genomes used in this study are shown in Table 6.

Conflict of interest disclosure

The author declares they have no financial conflicts of interest in relation to the content of this manuscript.

Supplements

Supplement 1 - Mutation rate analysis software and results.
https://doi.org/10.5281/zenodo.8080182

References

[1] Gordon Irlam. The creation-mutation-selection model: the model and mathematical analysis. *BioRxiv*, 2023. doi:10.1101/2023.06.26.546616.

[2] Nick GC Smith and Adam Eyre-Walker. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):1022–1024, 2002.

[3] Ana Filipa Moutinho, Thomas Bataillon, and Julien Y Dutheil. Variation of the adaptive substitution rate between species and within genomes. *Evolutionary Ecology*, 34(3):315–338, 2020.

[4] Justin C Fay, Gerald J Wyckoff, and Chung-I Wu. Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234, 2001.

[5] Jun Gojobori, Hua Tang, Joshua M Akey, and Chung-I Wu. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings of the National Academy of Sciences*, 104(10):3907–3912, 2007.

[6] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, 4(5):e1000083, 2008.

[7] Adam Eyre-Walker and Peter D Keightley. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, 26(9):2097–2108, 2009.

[8] Daniel L Halligan, Fiona Oliver, Adam Eyre-Walker, Bettina Harr, and Peter D Keightley. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics*, 6(1):e1000825, 2010.

[9] Joshua A Shapiro, Wei Huang, Chenhui Zhang, Melissa J Hubisz, Jian Lu, David A Turissini, Shu Fang, Hurng-Yi Wang, Richard R Hudson, Rasmus Nielsen, et al. Adaptive genic evolution in the Drosophila genomes. *Proceedings of the National Academy of Sciences*, 104(7):2271–2276, 2007.

[10] Xulio Maside and Brian Charlesworth. Patterns of molecular variation and evolution in Drosophila americana and its relatives. *Genetics*, 176(4):2293–2305, 2007.

[11] Doris Bachtrog. Similar rates of protein adaptation in Drosophila miranda and D. melanogaster, two species with different current effective population sizes. *BMC Evolutionary Biology*, 8(1):1–14, 2008.

[12] Nicolas Galtier. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS Genetics*, 12(1):e1005774, 2016.

[13] Toni I Gossmann, Bao-Hua Song, Aaron J Windsor, Thomas Mitchell-Olds, Christopher J Dixon, Maxim V Kapralov, Dmitry A Filatov, and Adam Eyre-Walker. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, 27(8):1822–1832, 2010.

[14] H Allen Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526, 2003.

[15] Rees Kassen and Thomas Bataillon. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics*, 38(4):484–488, 2006.

[16] Darin R Rokyta, Craig J Beisel, Paul Joyce, Martin T Ferris, Christina L Burch, and Holly A Wichman. Beneficial fitness effects are not exponential for two viruses. *Journal of Molecular Evolution*, 67(4):368, 2008.

[17] Michael J McDonald, Tim F Cooper, Hubertus JE Beaumont, and Paul B Rainey. The distribution of fitness effects of new beneficial mutations in Pseudomonas fluorescens. *Biology Letters*, 7(1):98–100, 2011.

[18] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[19] The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[20] Adam M Session, Yoshinobu Uno, Taejoon Kwon, Jarrod A Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, Akira Hikosaka, Atsushi Suzuki, Mariko Kondo, et al. Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, 538(7625):336–343, 2016.

[21] Sateesh Kagale, Chushin Koh, John Nixon, Venkatesh Bollina, Wayne E Clarke, Reetu Tuteja, Charles Spillane, Stephen J Robinson, Matthew G Links, Carling Clarke, et al. The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, 5(1):3706, 2014.

[22] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[23] Selina ER Bopp, Micah J Manary, A Taylor Bright, Geoffrey L Johnston, Neekesh V Dharia, Fabio L Luna, Susan McCormack, David Plouffe, Case W McNamara, John R Walker, et al. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genetics*, 9(2):e1003293, 2013.

[24] Dorothy E Loy, Weimin Liu, Yingying Li, Gerald H Learn, Lindsey J Plenderleith, Sesh A Sundararaman, Paul M Sharp, and Beatrice H Hahn. Out of Africa: origins and evolution of the human malaria parasites Plasmodium falciparum and Plasmodium vivax. *International Journal for Parasitology*, 47(2-3):87–97, 2017.

[25] Gavin G Rutledge, Ulrike Böhme, Mandy Sanders, Adam J Reid, James A Cotton, Oumou Maiga-Ascofare, Abdoulaye A Djimdé, Tobias O Apinjoh, Lucas Amenga-Etego, Magnus Manske, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature*, 542(7639):101–104, 2017.

[26] Toshihiro Mita and Kazuyuki Tanabe. Evolution of Plasmodium falciparum drug resistance: implications for the development and containment of artemisinin resistance. *Japanese Journal of Infectious Diseases*, 65(6):465–475, 2012.

[27] John W Drake. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences*, 870(1):100–107, 1999.

[28] Michael W Nachman and Susan L Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[29] Adeolu B Adewoye, Sarah J Lindsay, Yuri E Dubrova, and Matthew E Hurles. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nature Communications*, 6(1):6684, 2015.

[30] Brandon Milholland, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*, 8(1):1–8, 2017.

[31] Peter D Keightley, Rob W Ness, Daniel L Halligan, and Penelope R Haddrill. Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics*, 196(1):313–320, 2014.

[32] Peter D Keightley, Urmi Trivedi, Marian Thomson, Fiona Oliver, Sujai Kumar, and Mark L Blaxter. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, 19(7):1195–1201, 2009.

[33] Gordon Irlam. The creation-mutation-selection model: mutation rates and effective population sizes: Supplement 1, 2023. `doi:10.5281/zenodo.8080182`.

[34] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[35] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[36] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[37] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[38] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27(2):573–580, 1999.

[39] Sudhir Kumar, Glen Stecher, Michael Suleski, and S Blair Hedges. Timetree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819, 2017.

[40] Adam Eyre-Walker and Peter D Keightley. High genomic deleterious mutation rates in hominids. *Nature*, 397(6717):344–347, 1999.

[41] Mahendra Mariadassou, Marie Suez, Sanbadam Sathyakumar, Alain Vignal, Mariangela Arca, Pierre Nicolas, Thomas Faraut, Diane Esquerré, Masahide Nishibori, Agathe Vieaud, et al. Unraveling the history of the genus Gallus through whole genome sequencing. *Molecular Phylogenetics and Evolution*, 158:107044, 2021.

[42] Souqi Liao, Wenyan Dong, Hui Zhao, Ruijin Huang, Xufeng Qi, and Dongqing Cai. Cardiac regeneration in Xenopus tropicalis and Xenopus laevis: discrepancies and problems. *Cell & Bioscience*, 8(1):1–3, 2018.

[43] Barney Reed. Good practice guidance for the housing and care of Xenopus iaevis. *Animal Technology and Welfare*, 8(3):137, 2009.

[44] Thomas Naert, Dionysia Dimitrakopoulou, Dieter Tulkens, Suzan Demuynck, Marjolein Carron, Rivka Noelanders, Liza Eeckhout, Gert Van Isterdael, Dieter Deforce, Christian Vanhove, et al. RBL1 (p107) functions as tumor suppressor in glioblastoma and small-cell pancreatic neuroendocrine carcinoma in Xenopus tropicalis. *Oncogene*, 39(13):2692–2706, 2020.

[45] Nobuo Egami. Further notes on the life span of the teleost, Oryzias latipes. *Aging in Cold Blooded Animals*, page 118, 1974.

[46] Eva Terzibasi Tozzini, Alexander Dorn, Enoch Ng'oma, Matej Polačik, Radim Blažek, Kathrin Reichwald, Andreas Petzold, Brian Watters, Martin Reichard, and Alessandro Cellerino. Parallel evolution of senescence in annual fishes in response to extrinsic mortality. *BMC Evolutionary Biology*, 13(1):1–12, 2013.

[47] Yi Peng Teoh, Mashitah Mat Don, and Salmiah Ujang. Assessment of the properties, utilization, and preservation of rubberwood (hevea brasiliensis): a case study in malaysia. *Journal of Wood Science*, 57(4):255–266, 2011.