A Model for Estimating the Relative Increase in the Mutation Rate of Single Nucleotides Close to an Indel Segregating in a Population

We posit that the indels arise at the very low rate $u_{\rm i}$ and are neutral. For both indel and nonindel alleles, let u and $u_{\rm het}$ denote the neutral nucleotide mutation rates in homozygotes and heterozygotes, respectively, and set $f = u_{\rm het}/u$. Let $N_{\rm i}$ and $N_{\rm ni}$ designate the number of neutral mutations close to an indel and nonindel allele, respectively, since the most recent common ancestor (MRCA) of the sample. We seek the expectations $E(N_{\rm i})$ and $E(N_{\rm ni})$.

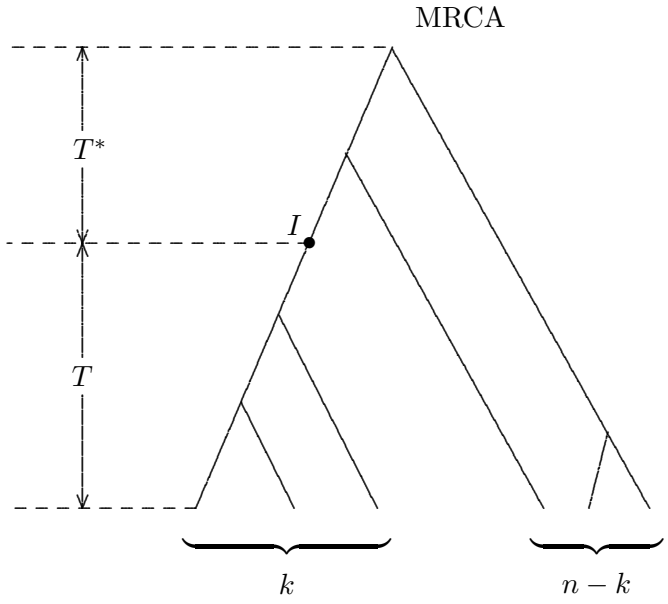

Fig. 1. Neutral genealogy in which k of n genes in a sample contain the neutral indel I. The time back to I is T, and T^* denotes the additional time back to the MRCA of the sample. The number of lineages at I is ν.

In the genealogy in Figure 1, k $(1 \le k \le n-1)$ of n genes in the sample carry the indel I, and the number of lineages at I is ν $(2 \le \nu \le n-k+1)$. We measure all times in units of $2N_{\rm e}$ generations, where $N_{\rm e}$ represents the effective population number. The time back to I is T, and T^*

designates the additional time back to the MRCA. Here and below, we suppress the dependence on k and n.

Let T_{het} and T_{hom} signify the times spent by an indel in heterozygotes and homozygotes, respectively. Then we have

$$T = T_{\text{het}} + T_{\text{hom}}, \tag{1}$$

and our assumptions yield

$$E(N_{\text{i}}) = \tfrac{1}{2}\theta[fE(T_{\text{het}}) + E(T_{\text{hom}}) + E(T^*)] \tag{2}$$

$$= \tfrac{1}{2}\theta[fE(T) + E(T^*) - (f-1)E(T_{\text{hom}})], \tag{3}$$

where $\theta = 4N_{\text{e}}u$.

If at time t in the past, the indel has frequency $X(t)$, then it is in a heterozygote and homozygote with respective conditional probabilities $1-X(t)$ and $X(t)$, whereas the corresponding probabilities for a nonindel are $X(t)$ and $1-X(t)$. The indicator function for homozygosity of an indel at time t is

$$\chi_{\text{hom}}(t) = \begin{cases} 1 & \text{if the indel is in a homozygote at time } t, \\ 0 & \text{if the indel is in a heterozygote at time } t. \end{cases} \tag{4}$$

Therefore,

$$T_{\text{hom}} = \int_0^T \chi_{\text{hom}}(t)dt, \tag{5}$$

which yields the conditional expectation

$$\begin{aligned} E[T_{\text{hom}} \,|\, X(0) = x] &= E\left\{ E\left[\int_0^T \chi_{\text{hom}}(t)dt \;\middle|\; T,\ X(0) = x \right] \right\} \\ &= E\left\{ \int_0^T E[\chi_{\text{hom}}(t)]dt \;\middle|\; X(0) = x \right\} \\ &= E\left\{ \int_0^T X(t)dt \;\middle|\; X(0) = x \right\}. \end{aligned} \tag{6a}$$

Similarly, we derive

$$E[T_{\text{het}} \;\middle|\; X(0) = x] = E\left\{ \int_0^T [1 - X(t)]dt \;\middle|\; X(0) = x \right\}. \tag{6b}$$

We infer from the preceding paragraph that can obtain $E(N_{\rm ni})$ from $E(N_{\rm i})$ by interchanging $E(T_{\rm het})$ and $E(T_{\rm hom})$:

$$E(N_{\rm ni}) = \tfrac{1}{2}\theta[fE(T_{\rm hom}) + E(T_{\rm het}) + E(T^*)] \tag{7}$$

$$= \tfrac{1}{2}\theta[E(T) + E(T^*) + (f-1)E(T_{\rm hom})]. \tag{8}$$

Thus, we must calculate the three mean times in (3) and (8). We assume henceforth that $\theta_{\rm i} = 4N_{\rm e}u_{\rm i} \to 0$ and neglect mutation from the indel to the nonindel.

To compute (6a), we note first that an indel segregating in our sample ($0 < k < n$) has never been fixed in the population. Therefore, we can use Eq. (16) of Maruyama and Kimura (1975) for the mean age of a mutant before fixation, with the following modifcations: we (i) divide by $2N_{\rm e}$ to scale the time; (ii) let the mutation rate approach 0; and (iii) insert a factor ξ into both integrands. This gives

$$E[T_{\rm hom} \,|\, X(0) = x] = 2\int_0^x \xi d\xi + \frac{2x}{1-x}\int_x^1 (1-\xi)d\xi = x. \tag{9}$$

To deduce $E(T_{\rm hom})$ from (9), we require the probability density of $X(0)$ conditional on observing k indels in a sample of n genes. Appealing to Bayes' formula, the population frequency spectrum θ/x (Kimura, 1969, 1971; Ewens, 2004, p. 298), and the sample frequency spectrum $\theta_{\rm i}/k$ (Watterson 1975, p. 266; Ewens, 2004, p. 311), we find the conditional density

$$\phi(x) = \binom{n}{k} x^k (1-x)^{n-k} \left(\frac{\theta_{\rm i}}{x}\right) \Big/ \frac{\theta_{\rm i}}{k}$$

$$= k\binom{n}{k} x^{k-1}(1-x)^{n-k}. \tag{10}$$

Averaging (9) over the beta density (10) leads immediately to

$$a \equiv E(T_{\rm hom}) = \frac{k}{n+1}. \tag{11}$$

From Wiuf and Donnelly (1999, p. 193), we have

$$b \equiv E(T) = \frac{2k}{n-1} - \frac{2}{n} + 2\binom{n-1}{k}^{-1} \sum_{j=2}^{n-k+1} \frac{1}{j}\binom{n-j-1}{k-1}. \tag{12}$$

To evaluate $E(T^*)$, we first condition on ν, the number of lineages at I. We designate the time to the MRCA of j lineages by T_{j1}. Its mean reads (Kingman, 1982)

$$E(T_{\mathrm{j1}}) = 2\left(1 - \frac{1}{j}\right), \tag{13}$$

whence

$$E(T^*) = E[E(T_{\nu 1} \,|\, \nu)] = 2E\left(1 - \frac{1}{\nu}\right). \tag{14}$$

From Eq. (18) of Wiuf and Donnelly (1999) we have ($2 \le j \le n - k + 1$)

$$P(\nu = j) = \binom{n-j}{k-1}\binom{n-1}{k}^{-1}. \tag{15}$$

Substituting (15) into (14) yields

$$c \equiv E(T^*) = 2\binom{n-1}{k}^{-1}\sum_{j=2}^{n-k+1}\left(\frac{j-1}{j}\right)\binom{n-j}{k-1}. \tag{16}$$

Dividing (3) by (8) and recalling (11), (12), and (16), we deduce

$$r \equiv \frac{E(N_{\mathrm{i}})}{E(N_{\mathrm{ni}})} = \frac{fb + c - (f-1)a}{b + c + (f-1)a}. \tag{17}$$

Since we know a, b, and c, and can estimate r as $N_{\mathrm{i}}/N_{\mathrm{ni}}$, we solve (17) for f:

$$f = \frac{a + c - r(b + c - a)}{a(r+1) - b}. \tag{18}$$

Before explaining how to combine data from different values of k, we demonstrate that r increases from 1 to $(b-a)/a$ as f increases from 1 to ∞. First, we rewrite (17) as

$$r = \frac{b-a}{a} - \frac{(b-2a)(b+c)}{a(b+c-a+fa)}. \tag{19}$$

Therefore, it suffices to establish that $b > 2a$.

Next, following the derivation of (9) but now inserting a factor $1-\xi$ instead of ξ into Eq. (16) of Maruyama and Kimura (1975), from (6b) we infer

$$E[T_{\mathrm{het}} \,|\, X(0) = x] = 2\int_0^x (1-\xi)d\xi + \frac{2x}{1-x}\int_x^1 \frac{(1-\xi)^2}{\xi}d\xi \tag{20}$$

$$= -x\left(1 + \frac{2\ln x}{1-x}\right) \tag{21}$$

$$> -x + \frac{2x(1-x)}{1-x} = x. \tag{22}$$

Comparing (22) with (9) informs us that

$$E[T_{\rm het} \,|\, X(0) = x] > E[T_{\rm hom} \,|\, X(0) = x]. \tag{23}$$

From (1) and (23) we obtain

$$E[T \,|\, X(0) = x] > 2E[T_{\rm hom} \,|\, X(0) = x], \tag{24}$$

and now (11), (12), and (24) imply that $b > 2a$.

The preceding result suggests that when we estimate f from (18), we should expect large confidence intervals.

We showed above how to estimate f for each k. Since the true value of the molecular parameter f must be independent of the sample parameter k, we now explain how to estimate f from the entire data set. We display explicitly the dependence on k in (3) and (8):

$$\begin{aligned} E(N_{\rm i}^{(k)}) &= \tfrac{1}{2}\theta[fE(T^{(k)}) + E(T^{*(k)}) - (f-1)E(T_{\rm hom}^{(k)})] \\ &= \tfrac{1}{2}\theta[fb^{(k)} + c^{(k)} - (f-1)a^{(k)}], \end{aligned} \tag{25a}$$

$$E(N_{\rm ni}^{(k)}) = \tfrac{1}{2}\theta[b^{(k)} + c^{(k)} + (f-1)a^{(k)}]. \tag{25b}$$

We invoke (11), (12), and (16), and redefine

$$E(N_{\rm i}) = \sum_{k=1}^{n-1} E(N_{\rm i}^{(k)}), \qquad E(N_{\rm ni}) = \sum_{k=1}^{n-1} E(N_{\rm ni}^{(k)}), \tag{26a}$$

$$a = \sum_{k=1}^{n-1} a^{(k)} = \frac{n(n-1)}{2(n+1)}, \tag{26b}$$

$$b = \sum_{k=1}^{n-1} b^{(k)} = n - 2 + \frac{2}{n} + 2\sum_{j=2}^{n} \frac{1}{j} \sum_{k=1}^{n-j+1} \binom{n-1}{k}^{-1} \binom{n-j-1}{k-1}, \tag{26c}$$

$$c = \sum_{k=1}^{n-1} c^{(k)} = 2\sum_{j=2}^{n} \left(\frac{j-1}{j}\right) \sum_{k=1}^{n-j+1} \binom{n-1}{k}^{-1} \binom{n-j}{k-1}, \tag{26d}$$

Summing (25) over k, we conclude from (26) that (17) and (18) hold with our redefinitions.

Application

We applied the model to yeast data drawn from aligned genome sequences of three strains of *S. cerevisiae* (S288C, RM11, and YJM89) and a closely related species, *S. paradoxus*. We used the outgroup sequence to polarize mutations (both indels and single-nucleotide polymorphisms) segregating in the three strains ($n = 3$). These alignments yielded 1026 instances of indel mutations occurring once ($k = 1$) and 251 instances of indel mutations occurring twice ($k = 2$) in the three *S. cerevisiae* strains (see Table 1). First, we treat the two cases separately; then we combine them.

For $k = 1$, from (11), (12), and (16) we get

$$a^{(1)} = \tfrac{1}{4}, \quad b^{(1)} = \tfrac{5}{6}, \quad c^{(1)} = \tfrac{7}{6}; \tag{27}$$

so (18) simplifies to the estimate

$$f^{(1)} = \frac{17 - 21r}{3r - 7}. \tag{28}$$

As r increases from 1 to 7/3, $f^{(1)}$ increases from 1 to ∞.

For $k = 2$, from (11), (12), and (16) we find

$$a^{(2)} = \tfrac{1}{2}, \quad b^{(2)} = \tfrac{4}{3}, \quad c^{(2)} = 1, \tag{29}$$

whence (18) yields

$$f^{(2)} = \frac{9 - 11r}{3r - 5}. \tag{30}$$

As r increases from 1 to 5/3, $f^{(2)}$ increases from 1 to ∞.

To derive the joint estimate, we use (26), (27), and (29):

$$a = \tfrac{3}{4}, \quad b = \tfrac{13}{6}, \quad c = \tfrac{13}{6}. \tag{31}$$

Now (18) gives

$$f = \frac{35 - 43r}{9r - 17}. \tag{32}$$

As r increases from 1 to $\frac{17}{9}$, f increases from 1 to ∞.

References

Ewens, W. J., 2004 *Mathematical Population Genetics. I. Theoretical Introduction.* Ed. 2. Springer, New York.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

Kimura, M., 1971 Theoretical foundation of population genetics at the molecular level. Theor. Pop. Biol. **2:** 174–208.

Maruyama, T., and M. Kimura, 1975 Moments for sum of an arbitrary function of gene frequency along a stochastic path of gene frequency change. Proc. Natl. Acad. Sci. USA **72:** 1602–1604.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **19A:** 27–42.

Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. Theor. Pop. Biol. **7:** 256–276.

Wiuf, C., and P. Donnelly, 1999 Conditional genealogies and the age of a neutral mutant. Theor. Pop. Biol. **56:** 183–201.