Letter to the Editor

Clustered Mutations Have No Effect on the Overdispersed Molecular Clock: A Response to Huai and Woodruff

David J. Cutler

Center for Population Biology, University of California, Davis, California 95616

IN a recent paper, Huai and Woodruff (1997) make the mistake of equating substitution with mutation, and arrive, quite erroneously, at the conclusion that clustered mutations can have a significant effect on the pattern of molecular evolution. I will show that clustered mutations have no effect at all on the index of dispersion. A large index of dispersion remains a central unexplained observation in molecular evolution.

The neutral theory predicts that the number of mutations that arise in a population in t generations, which ultimately become fixed in the population, will be Poisson distributed with mean ut , where u is the per sequence, per generation mutation rate. Therefore, the variance in the number of substitutions, S_t , will equal the mean at any neutral locus. The ratio of the variance in the number of substitutions to the mean number is often called the index of dispersion, $R(t)$,

$$R(t) = \frac{\text{Var}[S_t]}{E[S_t]},$$

which equals one under the neutral theory. Numerous studies have attempted to estimate $R(t)$ for protein encoding loci [see Ohta (1995) for the most comprehensive study to date], and it has been consistently shown that $R(t)$ for replacement substitutions is significantly larger than one (Gillespie 1991).

Huai and Woodruff make two important observations. First, in a series of experiments, Woodruff *et al.* (1996) show that premeiotic mutations lead to clusters of identical mutations in many offspring of a single individual. Next, they note that the theory that predicts $R(t) = 1$ assumes that all mutations arrive in the population at frequency $1/2N$. These two observations cause them to rightly conclude that our predictions for $R(t)$ should be reanalyzed assuming that the starting frequency for new mutations can be larger than $1/2N$ and that the frequency can be variable.

The simplest neutral model that allows easy calculation of $R(t)$ assumes infinite sites and constant population size N . By the neutral assumption, the probability

that any mutation will eventually fix in the population is simply the initial frequency of that mutant. Mutations come in two types, meiotic, which enter the population at frequency $1/2N$, and clusters, which enter the population at frequency P , where $0 < P < 1$ is a random variable. Let u be the mutation rate to unique mutants. By the infinite sites assumption, all meiotic mutants are unique, as are all clusters, but each cluster counts as a single mutational event. We will consider two different levels of recombination. Sites will either be assumed to recombine freely (Kimura 1969), or no recombination will be permitted (Watterson 1975).

Let M_t be the number of unique mutations in a period of t generations. M_t is Poisson distributed with mean and variance $2Nut$. Label each of these M_t mutations with a unique number between 1 and M_t . Associate with mutation j , $1 \leq j \leq M_t$, a random variable X_j . Let X_j equal 1 if mutation j ultimately fixes in the population, and 0 otherwise. Thus, X_j is the indicator that mutation j ultimately fixes in the population. Even though X_j depends on the random variable P , it is, nonetheless, a simple Bernoulli random variable with moments

$$E[X_j] = E[P], \quad (1)$$

$$\text{Var}[X_j] = E[P](1 - E[P]), \quad (2)$$

$$\text{Cov}[X_i, X_j] = 0, \quad \forall_i \neq j. \quad (3)$$

Equation 3 requires some comment. The validity of (3) is clear for free recombination. For the no recombination case we need an argument analogous to the one presented in Birky and Walsh (1988). Let $Y_i(t)$ be the frequency of mutation i t generations after it originated. Let $Y_j(t)$ be the same for j . Let $f_i(y|t) dy$ be the $\text{Prob}[Y_i(t) = y]$, and let $f_j(y|t)$ be the analogous density for $Y_j(t)$. $E[Y_i(0)] = E[Y_j(0)] = E[Y_i(t)] = E[Y_j(t)] = E[P]$. Without loss of generality, assume mutation j is older than mutation i , and assume it arose T generations before i . Consider $E[X_i X_j]$

$$E[X_i X_j] = \text{Prob}[\text{Both } i \text{ and } j \text{ fix}].$$

For both i and j to fix, i must arise on a chromosome containing mutation j , and i must fix. The probability that i fixes is $E[Y_i(0)] = E[P]$, and the probability that

i arises on a chromosome containing j is simply j 's frequency in the generation i arises, so

$$\begin{aligned} E[X_i X_j] &= \text{Prob}[i \text{ fixes}] \times \text{Prob}[j \text{ fixes} \mid i \text{ fixes}] \\ &= \text{Prob}[i \text{ fixes}] \times \text{Prob}[i \text{ arose on a chromosome containing } j] \\ &= E[Y_i(0)] \int y f_j(y|T) dy \\ &= E[Y_i(0)] E[Y_j(T)] \\ &= E[P]^2. \end{aligned}$$

So

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i] E[X_j] = 0.$$

Let S_t be the number of substitutions that originate in the population during these t generations. Clearly,

$$S_t = \sum_{j=1}^{M_t} X_j.$$

Since M_t has been constructed to be independent of the X_j 's, and the X_j 's are mutually independent (Equation 3), we have,

$$E[S_t] = E[M_t] E[X_j], \quad (4)$$

$$\text{Var}[S_t] = E[M_t] \text{Var}[X_j] + \text{Var}[M_t] (E[X_j])^2, \quad (5)$$

(see Feller 1968, p. 301). Noting that $E[M_t] = \text{Var}[M_t]$,

$$R(t) = \frac{\text{Var}[S_t]}{E[S_t]} = E[X_j] + \frac{\text{Var}[X_j]}{E[X_j]} = 1, \quad (6)$$

which holds when there are no clusters ($P = 1/2N$), as well as when there are clusters (P a random variable). Thus, we see clustered mutations can not effect the index of dispersion in any way, and that the index of dispersion is always exactly one.

Huai and Woodruff's overestimation of $R(t)$ stems from two errors. First, they model clustered mutations as if they cause distinct mutations to be copied into several offspring of a single individual, rather than a single unique mutation to be copied into those offspring. As a result, they allow u to be a random variable, rather than P . By failing to make the subtle distinction between u and P , Huai and Woodruff allow the same premeiotic mutation to fix more than once. This error is not substantial, however, and only leads one to conclude

that $R(t) \approx 1 + O(1/2N)$. As population size increases, the distinction becomes unimportant.

The other error is far more serious. Huai and Woodruff derive the variance to mean ratio of the number of mutations, not the variance to mean ratio of the number of substitutions. As we can see from Equation 5, increases in the mutational variance will propagate into increases in the substitutional variance by a factor of $(E[X_j])^2 \approx (1/2N)^2$, whereas all other terms are of order $1/2N$. Thus, by considering the variance to mean ratio of mutations, rather than substitutions, Huai and Woodruff overestimate the increase in $R(t)$ by a factor $2N$. Taking their two errors together they found $R(t) \approx 1 + 2N \times O(1/2N) \approx 1 + O(1)$, which they concluded was a significant effect. The analysis here shows that $R(t)$ is exactly one. This analysis uses the infinite sites, free-recombination or no-recombination model of the gene. The equivalent analysis for an arbitrary level of recombination (Hudson 1983) is more difficult, but should lead to the same conclusion that $R(t) = 1$.

I would like to thank John Gillespie and Hiroshi Akashi for many suggestions. This work was supported by fellowships from The Center for Population Biology, and The Institute for Theoretical Dynamics at UC Davis.

LITERATURE CITED

- Birky, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**: 6414-6418.
- Feller, W., 1968 An Introduction to Probability Theory and Its Applications. John Wiley & Sons, Inc., New York.
- Gillespie, J. H., 1991 The Causes of Molecular Evolution. Oxford University Press, New York.
- Huai, H., and R. C. Woodruff, 1997 Clusters of identical new mutations can account for the 'overdispersed' molecular clock. *Genetics* **147**: 339-348.
- Hudson, R. R., 1983 Properties of a neutral allele model with intra-genic recombination. *Theor. Pop. Biol.* **23**: 183-201.
- Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**: 893-903.
- Ohta, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* **40**: 56-63.
- Watterson, G. A., 1975 On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**: 256-276.
- Woodruff, R. C., H. Huai and J. N. Thompson, Jr., 1996 Clusters of identical new mutations in the evolutionary landscape. *Genetica* **98**: 149-160.

Communicating editor: N. Takahata