

# Letter to the Editor

## Estimating the Probabilities of Runs of Identical Events Within Biological Sequences

Mark Finkelstein,\* Walter M. Fitch,† Carmine A. Lanciani,‡ and Michael M. Miyamoto‡

\*Department of Mathematics and †Department of Ecology and Evolutionary Biology, University of California–Irvine; and

‡Department of Zoology, University of Florida

Miyamoto and Fitch (1995) recently compared the properties and predictive powers of the covarion hypothesis (Fitch and Ayala 1994) with those of the gamma version of the one-parameter model (Ota and Nei 1994; Yang 1996). Miyamoto and Fitch focused on the degree to which the identities of polymorphic and monomorphic codons for the amino acid sequences of Cu, Zn superoxide dismutase (SOD, EC 1.15.1.1) were the same in mammals and plants, because less overlap (fewer positions being varied in both groups) is predicted between taxa by the covarion process than by the gamma model. Their comparisons relied on both computer simulations of sequence change and spatial analyses of amino acid replacements on the three-dimensional configuration of the protein. The latter analyses of amino acid replacements relative to tertiary structure included an evaluation of a run of eight consecutive varied positions that were polymorphic in plants but not in mammals. These eight sites from the first  $\beta$ -strand and loop at the  $N$ -terminus of the protein constituted the longest run of consecutive polymorphic positions, varied in that group only. Thus, Miyamoto and Fitch (1995) attempted to calculate the probability ( $P$ ) of obtaining a run of eight or more polymorphic sites, restricted to either mammals or plants and anywhere among the varied positions, by their equation (5):

$$P = \sum_{i=8}^j \left(\frac{j}{t}\right)^{i-1} (j - i + 1), \quad (1)$$

where  $t$  is the total number of varied codons that are polymorphic in plants, in mammals, or in both groups ( $t = 74$ ),  $j$  is the number of polymorphic positions unique to plants ( $j = 27$ ) or to mammals ( $j = 24$ ), and  $i$  is the length of all runs from  $i = 8$  to  $j$ . Out of the 74 polymorphic positions, 51 were therefore limited to either mammals or plants only, whereas 23 were varied in both groups.

Miyamoto and Fitch (1995) estimated from equation (1) that the final probability of a run of eight or more polymorphic positions unique to either mammals or plants was no larger than 0.035 for their SOD data. Equation (1) offers a conservative, first-order approximation of  $P$ , as acknowledged by its authors. A limitation of this equation is that it treats runs of different lengths as if they were nonoverlapping.

Key words: run probabilities, biological sequences, covarion hypothesis, superoxide dismutase.

Address for correspondence and reprints: Michael M. Miyamoto, Department of Zoology, University of Florida, Gainesville, Florida 32611. E-mail: miyamoto@zoo.ufl.edu.

*Mol. Biol. Evol.* 15(4):470–472. 1998

© 1998 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

We present a method that gives the exact probability of occurrence of a sequence of at least a given length of “special” sites (e.g., polymorphic sites unique to plants) when the overall length and number of “special” sites are given. The test is a conditional nonparametric test based on observing a run of length  $i$  of one kind in a sequence of length  $t$ . This method permits exact calculations of significance probabilities for statistics where these values were previously obtainable only by simulation or approximation (Stephens 1985; Sawyer 1989). O’Brien and Agin (1994) gave a formula for the probability of observing exactly  $m$  runs each of exact length  $i$ , conditioned on the number of “special” sites and the overall length. From this, it would be possible, at least in principle, to obtain the conditional probability of at least one run of length  $\geq i$ . However, O’Brien and Agin’s (1994) formula is difficult to implement numerically in some instances, as it involves the addition of very large quantities with alternating signs, making the formula numerically unstable. (As an example of such an unstable formula, see Feller [1968, p. 102, formula 2.3].) Lehmann (1975) studied a related statistic: the number of runs, rather than the length of the longest run.

We obtain a different formula, which is both simple and computationally effective. We also present a runs test for parametric models involving Bernoulli trials.

We calculate the probability of a run of varied positions by a conditional runs test. Conditioning on the event that we observe 27 polymorphic positions unique to plants out of 74 positions, what is the probability that a sequence of length 74, with 27 randomly chosen positions filled with P’s (markers for “plants”) and the remaining 47 positions filled with O’s (for “others”), will contain a run of 8 or more P’s? Here, all permutations of a sequence of 27 P’s and 47 O’s are considered equally likely. We first give a theoretical solution to this problem, followed by the numerical results.

Shifting from P’s and O’s to black and white balls, we consider the arrangement of  $j$  black balls and  $t - j$  white balls arranged in a sequence of length  $t$ . We imagine this sequence to be the result of random shuffling of the  $t$  positions and ask for the probability  $P(i, j, t)$  of observing a run of at least  $i$  black balls in the resulting sequence.

To obtain this formula, we define the random variable  $W$  to be the first position where a white ball occurs, and condition on the value of  $W$ :

$$\begin{aligned} P(i, j, t) &= \sum_{k=1}^t P(i, j, t | [W = k])P[W = k] \\ &= \sum_{k=1}^i P(i, j, t | [W = k])P[W = k] \\ &\quad + \sum_{k=i+1}^t P(i, j, t | [W = k])P[W = k]. \quad (2) \end{aligned}$$

Note that if the  $k$ th ball is the first white ball, the first  $k - 1$  balls must have been black, and if  $k \leq i$ , this initial sequence of black balls and one white ball cannot contribute to a run of black balls of length  $i$  or greater. From this, it follows that

$$P(i, j, t | [W = k]) = P(i, j - (k - 1), t - k) \quad \text{if } k \leq i. \quad (3)$$

Further, note that if the first white ball occurs in the  $(i + 1)$ -st position or later, we already have a run of  $i$  or more black balls in a row. Hence,

$$P(i, j, t | [W = k]) = 1 \quad \text{if } k > i. \quad (4)$$

Collecting these observations leads to the recursion formula

$$\begin{aligned} P(i, j, t) &= \sum_{k=1}^i P(i, j - k + 1, t - k)P[W = k] \\ &\quad + \sum_{k=i+1}^t P[W = k] \\ &= \sum_{k=1}^i P(i, j - k + 1, t - k)P[W = k] \\ &\quad + P[W > i]. \end{aligned} \quad (5)$$

The values  $P[W = k]$  are determined by sampling without replacement from an urn with  $j$  black balls and  $t - j$  white balls until the first white ball is observed:

$$P[W = k] = \begin{cases} \frac{j}{t} \frac{j-1}{t-1} \cdots \frac{j-(k-2)}{t-(k-2)} \frac{t-j}{t-(k-1)} & \text{if } k \geq 2 \\ \frac{t-j}{t} & \text{if } k = 1. \end{cases} \quad (6)$$

Using formulas (5) and (6) above, we determined  $P_{\text{plants}}(8, 27, 74) = 0.00707$  and  $P_{\text{mammals}}(8, 24, 74) = 0.00249$ . Thus, the total probability of a run of length eight or more for either plants or mammals in a sequence of 74 positions is bounded above by  $0.00707 + 0.00249$ , which is less than 0.01. This estimate is an upper bound, because it double-counts the simultaneous occurrence of runs of eight or more in both plants and mammals. As a check on these values, we ran 1,000,000 computer-generated random permutations of a sequence of length 74, consisting of 27 P's, 24 M's (for "mammals"), and 23 O's. These simulations showed that 6,983 sequences (or, proportionally, 0.00698) contained at least one run of  $\geq 8$  P's, 2,563 sequences (0.00256) contained at least one run of  $\geq 8$  M's, and 31 sequences (0.00003) contained both. These results agree with the theoretically derived values reported above for plants and mammals. They also indicate that the double-counting due to the simultaneous occurrence of runs of  $\geq 8$  for both plants and mammals is proportionally  $< 0.0001$ .

Stephens (1985, p. 544) reported in his figure 3 that (in our terminology)  $P(15, 23, 30) = 0.0099$ , whereas

our formulas show  $P(15, 23, 30) = 0.0253$ . Stephens also reported that  $P(5, 7, 30) = 0.0047$ , whereas  $P(5, 7, 30) = 0.0035$  according to our equations. These discrepancies come from his assumption of independence of segments, which was used to multiply probabilities in his equation (10), and from the use of the approximation in his equation (10).

There may be situations in which a parametric model is more appropriate. In these cases, a different runs test can be constructed in which the probability  $p$  of occurrence (in any position of a sequence) of an event of a certain kind is fixed (and  $p$  is either known, or unknown and estimated). The sequence of length  $t$  is then a sequence of  $t$  Bernoulli trials with parameter  $p$ . We again seek the probability of observing a run of  $i$  or more "successes" in the  $t$  trials. To determine this probability, we imagine conducting a sequence of coin tosses (each with probability  $p$  of heads) until we observe a sequence of exactly  $i$  heads in a row, at which time we stop. This can be viewed as a Markov chain in which the states are  $\{0, 1, 2, \dots, i\}$  (the state counts the number of heads in a row we have just observed). In this Markov chain, we start in state 0, move from state  $x$  to state  $x + 1$  with probability  $p$ , and move from state  $x$  to state 0 with probability  $(1 - p)$ . State  $i$  is the "stopping state." The probability that this Markov chain terminates in  $t$  steps (trials) or less is exactly the same as the probability that in a sequence of  $t$  Bernoulli trials, we observe a run of  $i$  or more consecutive heads.

To calculate the probability that we terminate our Markov chain in no more than  $t$  trials, we let  $p_n$  be the probability that we stop our chain on exactly the  $n$ th trial. If  $n < i$ , this value is 0, because we cannot have observed  $i$  heads in fewer than  $i$  trials. If  $n = i$ , we must have had exactly  $i$  heads, the probability of which is  $p^i$ . If  $n = i + 1$ , we must have observed tails in the first trial, followed by  $i$  heads, the probability of which is  $(1 - p)p^i$ . In general, for  $n > i + 1$ , for our chain to end on exactly the  $n$ th trial, we must have observed (1) that our chain had not ended by the  $(n - i - 1)$ -st trial, (2) that the  $(n - i)$ -th trial was tails, and (3) that the last  $i$  trials were all heads. Because the Bernoulli trials are independent, a computation with conditional probabilities shows that the probability of the simultaneous occurrence of these three events is the product of their probabilities, namely,

$$p_n = (1 - p_1 - p_2 - \dots - p_{n-i-1})(1 - p)p^i$$

or, noting that  $p_1 = \dots = p_{i-1} = 0$ ,

$$p_n = (1 - p_i - p_{i+1} - \dots - p_{n-i-1})(1 - p)p^i. \quad (7)$$

Thus, each of the  $p_n$ 's is defined recursively in terms of the  $p_x$ 's with  $i \leq x < n - i$ . The desired probability,  $P$ , that the chain terminates in  $t$  or fewer trials, is

$$P = \sum_{n=i}^t p_n. \quad (8)$$

The final conditional estimate of  $< 0.01$  from formulas (5) and (6) for obtaining by chance a sequence with at least one run of eight or more polymorphic sites

unique to either mammals or plants indicates that the observed run of eight varied positions unique to plants at the *N*-terminus of SOD is a significantly unlikely event. Thus, this result is consistent with the claim by Miyamoto and Fitch (1995) that the variable and invariable positions of the two groups are distributed differently across the major regions of the protein. In turn, this result strengthens their conclusion that the covarion hypothesis provides a better explanation of the evolution of SOD than does the gamma version of the one-parameter model.

Molecular biologists and evolutionists often must evaluate the significance of an unusual but short stretch of protein or nucleic acid sequence (e.g., Stephens 1985; Tagle et al. 1988; Sawyer 1989). However, the restricted distributions and short lengths of such sequences can make it difficult to document their statistical significance. The analytical procedures of this study may prove useful in cases like these as they have in further evaluating mammal and plant SODs and their mechanisms of change.

### Acknowledgments

We are grateful to the Associate Editor and the referee of our paper for their helpful suggestions and to H. G. Tucker and J. A. Veeh for their many valuable discussions. We also thank E. Chipouras, J. B. Slowinski, M. R. Tennant, and N. L. White for their useful advice on this research and our respective departments and universities for their financial assistance.

### LITERATURE CITED

- FELLER, W. 1968. An introduction to probability theory and its applications. Vol. 1, 3rd edition. John Wiley & Sons, New York.
- FITCH, W. M., and F. J. AYALA. 1994. The superoxide dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. USA* **91**:6802–6807.
- LEHMANN, E. L. 1975. Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco.
- MIYAMOTO, M. M., and W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**:503–513.
- O'BRIEN, P. C., and M. A. AGIN. 1994. Robust procedures for detecting non-random patterns. Pp. 183–203 in A. P. GODBOLE and S. G. PAPASTAVRIDIS, eds. *Runs and patterns in probability*. Kluwer Academic, Dordrecht, The Netherlands.
- OTA, T., and M. NEI. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* **38**:642–643.
- SAWYER, S. A. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**:526–538.
- STEPHENS, J. C. 1985. Statistical methods of DNA sequencing analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.* **2**:539–556.
- TAGLE, D. A., B. F. KOOP, M. GOODMAN, J. L. SLIGHTOM, D. HESS, and R. T. JONES. 1988. Embryonic  $\epsilon$  and  $\gamma$  globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation, and phylogenetic footprints. *J. Mol. Biol.* **203**:439–455.
- YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *TREE* **11**:367–372.
- STANLEY A. SAWYER, reviewing editor
- Accepted December 2, 1997