LETTER TO THE EDITOR

Dear Editor,

A note on 'The statistical analysis of direct repeats in nucleic acid sequences'

Shukla and Srivastava (1985) developed a probability model to account for the possible 'dependence' among repeated segments of the sequences. The probability of the repeats was evaluated by assuming the random arrangement of the nucleic bases within the sequences. The authors provided formulae for the mean and variance of tandem repeats, $T_k(n)$ (originally written as T_k on page 22 of their paper), which are as follows (keeping the same notation):

$$(1) \qquad \boldsymbol{E}(T_k(n)) = N_k P_k(n)$$

and

$$(2) \qquad V(T_k(n)) = N_k P_k(n) Q_k(n) + 2 \sum_{j=1}^{\min(n+k-1, N_k)} (N_k - j)[P_k(n+j) - (P_k(n))^2].$$

The equation (2) was derived from

$$(3) \qquad V(T_k(n)) = N_k P_k(n) Q_k(n) + 2 \sum_{j=1}^{\min(n+k-1, N_k)} (N_k - j) \operatorname{Cov}(d_{i,i+k}, d_{i+j,i+j+k}).$$

When $k = 1$, (2) is obtained from (3) by noting that $\operatorname{Cov}(d_{i,i+k}, d_{i+j,i+j+k}) = [P_k(n+j) - (P_k(n))^2]$. However, the following numerical example shows that (2) cannot be generalized to all values of k. Assuming $N = 20$, $k = 4$, $n = 1$ and equiprobability of nucleotide bases, the variance of $T_4(1)$ equals -1.3125 if (2) is used.

The variance of $T_k(n)$ can be derived by considering

$$(4) \qquad \begin{aligned} \operatorname{Cov}(d_{i,i+k}, d_{i+j,i+j+k}) &= \boldsymbol{E}(d_{i,i+k}, d_{i+j,i+j+k}) - (\boldsymbol{E}(d_{i,i+k})\boldsymbol{E}(d_{i+j,i+j+k})) \\ &= \boldsymbol{P}(d_{i,i+k} = 1 \wedge d_{i+j,i+j+k} = 1) - P_k(n)^2. \end{aligned}$$

$\boldsymbol{P}(d_{i,i+k} = 1 \wedge d_{i+j,i+j+k} = 1)$ (henceforth denoted by $P_d(j, k)$) can be obtained by considering different combinations of k, n and j.

1. *$k < n$ and $j \leqq n$.* In this situation, n-tuples, Y_i and Y_{i+j}, form an $(n+j)$-tuple Y'_i, and also n-tuples, Y_{i+k} and Y_{i+j+k}, form another $(n+j)$-tuple Y'_{i+k}. Thus, $P_d(j, k)$ equals $\boldsymbol{P}(d_{i,i+k}(n+j) = 1) = P_k(n+j)$.

$$P_d(j, k) = \prod_{l=1}^{k} \left(\sum_{i=1}^{4} P_i^{[(n+j+k-l)/k]+1} \right) = P_k(n+j).$$

https://doi.org/10.2307/3214911 Published online by Cambridge University Press

2. $k < n$ *and* $n < j < n + k$.

$$P_d(j,k) = \prod_{l=k_1}^{k_2}\left(\sum_{i=1}^{4} P_i^{[(n+j+k-l)/k]+1}\right)\left(\prod_{l=n+k-j+1}^{k}\left(\sum_{i=1}^{4} P_i^{[(n+k-l)/k]+1}\right)\right)^2 \quad \text{if } k_1 \leqq k_2$$

$$= \prod_{l=1}^{k_2}\left(\sum_{i=1}^{4} P_i^{[(n+j+k-l)/k]+1}\right)\prod_{l=k_1}^{k}\left(\sum_{i=1}^{4} P_i^{[n+j+k-l)/k]+1}\right)$$

$$\times\left(\prod_{l=n+k-j+1}^{k}\left(\sum_{i=1}^{4} P_i^{[(n+k-l)/k]+1}\right)\right)^2 \quad \text{if } k_1 > k_2$$

where

$$\begin{aligned} k_1 &= j+1-k[(j+1)/k] && \text{if } j+1 > k[(j+1)/k] \\ &= k && \text{if } j+1 = k[(j+1)/k] \\ k_2 &= n+k-k[(n+k)/k] && \text{if } n+k > k[(n+k)/k] \\ &= k && \text{if } n+k = k[(n+k)/k]. \end{aligned}$$

3. $k < n$ *and* $j \geqq n + k$. In this situation, two pairs of n-tuples are mutually separated. So

$$P_d(j,k) = \left(\prod_{l=1}^{k}\left(\sum_{i=1}^{4} P_i^{[(n+k-l)/k]+1}\right)\right)^2 = (P_k(n))^2.$$

4. $k \geqq n$ *and* $j \leqq n$. This circumstance is similar to Case 1 and the probability is

$$P_d(j,k) = \prod_{l=1}^{k}\left(\sum_{i=1}^{4} P_i^{[(n+j+k-l)/k]+1}\right) = P_k(n+j).$$

5. $k \geqq n$, $n < j \leqq k < n + j$. In this case, the overlapping sub-segment of Y_{i+j} and Y_{i+k} form an $(n+j-k)$-tuple. If $d_{i,i+k} = 1$ and $d_{i+j,i+j+k} = 1$ are satisfied, we can find two more sub-segments within Y_i and Y_{i+j+k}, respectively, which are identical to the above-mentioned $(n+j-k)$-tuple. The probability of obtaining the three identical sub-segments is $(\Sigma P_i^3)^{n+j-k}$. Considering the remaining four sub-segments (or $(k-j)$-tuples) on Y_i, Y_{i+j}, Y_{i+k} and Y_{i+j+k}, we find

$$P_d(j,k) = \left(\sum_{i=1}^{4} P_i^3\right)^{n+j-k}\left(\sum_{i=1}^{4} P_i^2\right)^{2(k-j)}.$$

6. $k \geqq n$, $n < j < n + j \leqq k$. In this situation, all four n-tuples are mutually separated. Thus

$$P_d(j,k) = \left(\prod_{l=1}^{k}\left(\sum_{i=1}^{4} P_i^{[(n+k-l)/k]+1}\right)\right)^2 = (P_k(n))^2.$$

7. $k \geqq n$, $k < j < n + j$. Interchanging the positions of j and k in the inequality, $n \leqq k < j \leqq n + j$, and following the discussion in Case 5, we find

$$P_d(j,k) = \left(\sum_{i=1}^{4} P_i^3\right)^{n+k-j}\left(\sum_{i=1}^{4} P_i^2\right)^{2(j-k)}.$$

https://doi.org/10.2307/3214911 Published online by Cambridge University Press

Employing these formulae, the variance of $T_k(n)$ for the previously mentioned numerical example is found to be 3.0.

Reference

SHUKLA, R. AND SRIVASTAVA, R. C. (1985) The statistical analysis of direct repeats in nucleic acid sequences. *J. Appl. Prob.* **22**, 15–24.

Yours sincerely,

JIANLIANG ZHANG
RAKESH SHUKLA

Institute of Environmental Health,
Division of Biostatistics,
University of Cincinnati Medical Center,
Wherry Hall (#183)
Cincinnati, OH 45267,
USA