Conditional Bonferroni-Type Inequalities in Molecular Biology

Jie Chen
University of Massachusetts Boston,Department of Computing Services
100 Morrissey Boulevard
Boston, MA. U.S.A.
jie.chen@umb.edu

Joseph Glaz
University of Connecticut, Department of Statistics
Storrs, CT. U.S.A.
glaz@uconnvm.uconn.edu

Abstract
Scan statistic has been extensively used in molecular biology. In this article, second order Bonferroni-type inequalities based on a conditional scan statistic are derived to test an unusual large negative or positive charge in a sequence of amino acids. A simulation study is presented to evaluated the accuracy of these inequalities. Numerical results presented in this article indicate that these inequalities are tight
Kew Words: Bonferroni-type inequalities, clustering of events, Moving sums, Scan statistic.

1. Introduction

Let $X_1, ..., X_n$ be iid nonnegative integer valued observations. For $2 \leq m < n$ and $1 \leq t \leq n-m+1$, define $Y_t(m) = \sum_{i=t}^{t+m-1} X_i$ to be the number of events that have occurred in a moving window of length m. If $Y_t(m)$ exceeds a preassigned value k, then we will say that k events are clustered in a window of length m or that a *generalized run* of at least k events has occurred in a window of length m. For the case of iid $0-1$ Bernoulli trials with $p = P(X_i = 1)$ being small, one can view the occurrence of a generalized run as a rare coincidence of events, labeled as successes that are occurring in a sequence of n events. This generalizes the notion of a success run of length m that has been studied extensively in the statistical literature (Balakrishnan, Balasubramanian and Viveros 1993, Diaconis and Mosteller 1989, Fu 1986, Fu and Koutras 1994, Godbole 1990, 1991 and 1993, Gordon, Schilling and Waterman 1986, Hirano and Aki 1993, Karlin and Ost 1987, Koutras and Alexandrou 1997, Mott, Kirkwood and Curnow 1990 and Schwager 1983).

For $1 \leq k \leq m < n$ the *unconditional discrete scan statistic* is defined as the largest number of events in any window of size m:

$$S_m = \max_{1 \leq t \leq n-m+1} Y_t(m). \tag{1}$$

S_m is used in testing the null hypothesis that the observations $X_1, ..., X_n$ are identically distributed from a distribution F_0, while under the alternative hypothesis for some $0 \leq t \leq n-m+1$ and $I_t(m) = \{t, ..., t+m-1\}$, X_i, $i \in I_t(m)$ are identically distributed from F_1 and X_i, $i \in \{1, ..., n\} \backslash I_t(m)$ are identically distributed from F_0. It has been show in Glaz and Naus (1991) that the generalized likelihood ratio test reject the above null hypothesis in favor of the alternative hypothesis whenever $S_m \geq k$, where k is determined from a specified significance level for the test. To implement this testing procedure accurate inequalities for $P(S_m \geq k)$ are of great value. Glaz and Naus (1991) derived accurate product-type inequalities for $P(S_m \geq k)$.

In certain applications the total number of events that have occurred is known. In that case we refer to the scan statistics defined in Equation (**??**) as *conditional* scan statistics. For general nonnegative integer valued observations there are no accurate inequalities available. In Section 2 of this article we develop tight Bonferroni-type inequalities for the conditional discrete scan statistics for the case of iid $0-1$ Bernoulli trials, when $n \neq Lm$. An example from an experiment in molecular biology (Karlin, Blaisdell, Mocarski and Brendel 1989) is applied. The special case of $n = Lm$ is illustrated in Chen et.al. (2001). Applications of scan statistic in molecuar biology has been discussed in Waterman (1995). The conditional scan statistic has many other applications include: meteorology (Moye, et. al. 1988), minefield detection (Glaz 1996), molecular biology (Altschul and Erickson 1988, Fu and Curnow 1990, Karlin and Ghandour 1985, Naus and Sheng 1996, Sheng and Naus 1994 and 1996 and Waterman 1995), psychology (Runnels, et. al. 1968), quality control and reliability theory (Balakrishnan, Balasubramanian and Viveros 1993, Chao, Fu and Koutras 1995, Fu and Koutras 1994, Glaz 1983, Greenberg 1970, Saperstein 1972, and Viveros and Balakrishnan 1993, radar detection (Bogush 1972 and Nelson 1978), and sociology (Schwager 1983).

Section 5 presents numerical results to evaluate the performance of the Bonferroni-type inequalities and the Conclusions.

2. Bonferroni-Type Inequality

For the charge alphabet classification of amino acids, there are three possibilities: acidic, neutral and basic, denoted numerically by $-1, 0, 1$, respectively. Microbiologists are interested in determining if a large number of acidic charges $(-1's)$ or a large of number of basic charges $(1's)$ that have occurced in a window of length m is unusual.

Karlin, Blaisdell, Mocarski and Brendel (1989) gave an example of a sequence of $n = 968$ residues of the adenovirus type 2 hexon protein. They have observed 114 acidic charges, 754 neutral charges and 100 basic charges. They inspect windows of length 30. In this example we want to evaluate tight inequalities for observing k or more basic charges in a window of length $m = 30$.

Let $Y_1, \cdots Y_n, n = 968$ be a sequence of $0-1$ Bernoulli trials. Suppose we know that

$a = 100$ successes $(1's)$ have been observed. Consider the moving windows with length $m = 30$. In this case n is not a multiple of m, $n = Lm + \nu$ where $L = 32$ and $\nu = 8$. For $1 \leq i \leq L-1$, define events D_i be

$$D_i = \bigcap_{j=1}^{m+1} \left(Y_{(i-1)m+j} + \cdots + Y_{im+j-1} \leq k-1 \right), \tag{2}$$

and

$$D_L = \bigcap_{j=1}^{\nu+1} \left(Y_{(L-1)m+j} + \cdots + Y_{Lm+j-1} \leq k-1 \right) \tag{3}$$

be the last group of moving windows. We are interested to approximate the tail probability of the conditional scan statistic

$$P(k; m, n, a, \nu) = P\left(\bigcup_{i=1}^{L} D_i^c \right).$$

Employing the second order Bonferroni-type upper inequality in Hunter (1976) and Worsley (1982) we get

$$\begin{aligned} P(k; m, n, a, \nu) &\leq \sum_{i=1}^{L} P\left(D_i^c\right) - \sum_{i=1}^{L-1} P\left(D_i^c \bigcap D_{i+1}^c\right) \\ &= (L-1)P(D_1^c) + P(D_L^c) - (L-2)P(D_1^c \bigcap D_2^c) - P(D_{L-1}^c \bigcap D_l^c) \\ &= 1 + (L-2)q_{2m}(a) - (L-2)q_{3m}(a) - q_{2m+\nu}(a), \end{aligned} \tag{4}$$

where for $1 \leq i \leq L-2$, $q_{2m}(a) = P(D_i), q_{3m}(a) = P\left(D_i \bigcap D_{i+1}\right)$ and for $r = 2, 3$

$$q_{rm}(a) = \sum_{j=0}^{min(rk-r,a)} q(rm|j) \frac{\binom{rm}{j}\binom{n-rm}{a-j}}{\binom{n}{a}}$$

where $q(rm|j) = 1 - P(k; m, rm, j)$ are evaluated using Naus (1974, Theorem 1) and $q_{2m+\nu}(a)$ will be evaluated in next section.

To derive a Bonferroni-type lower inequality for $P(k; m, n, a)$ we employ the approach in Kwerel (1975) to get:

$$P\left(\bigcup_{i=1}^{L} D_i^c \right) \geq \frac{2s_1}{b} - \frac{2s_2}{b(b-1)}, \tag{5}$$

where

$$b = [\frac{2s_2}{s_1} + 2],$$

$$s_1 = \sum_{i=1}^{L} P\left(D_i^c\right) = (L-1)(1 - q_{2m}(a)) + 1 - q_{m+\nu}(a)$$

and

$$s_2 = \sum_{1 \leq i < j \leq L-1} P\left(D_i^c \bigcap D_j^c\right) + \sum_{i=1}^{L-1} P(D_i \bigcap D_L)$$

$$
\begin{aligned}
&= \sum_{i=1}^{L-2} P\left(D_i^c \bigcap D_{i+1}^c\right) + \sum_{1\le i<j-1\le L-2} P\left(D_i^c \bigcap D_j^c\right) + \sum_{i=1}^{L-2} P(D_i^c \bigcap D_L^c) + P(D_{L-1}^c \bigcap D_L^c) \\
&= (L-2)q_{3m}(a) + 0.5(L-2)(L-3)q_{2m,2m}(a) + 0.5(L-1)(L-2)(1-2q_{2m}(a)) \\
&\quad +(L-1)[1-q_{2m}(a)-q_{m+\nu}(a)] + (L-2)q_{2m,m+\nu}(a) + q_{2m+\nu}(a)
\end{aligned}
$$

where

$$
q_{2m,2m}(a) = \sum_{j_1=0}^{\min(2k-2,a)} \sum_{j_2=0}^{\min(2k-2,a-j_1)} q(2m|j_1)q(2m|j_2)\frac{\binom{2m}{j_1}\binom{2m}{j_2}\binom{n-4m}{a-j_1-j_2}}{\binom{n}{a}},
$$

$q_{2m}(a)$, $q_{3m}(a)$ are defined in Equation (**??**), $q_{2m+\nu}(a) = P\left(D_{L-1} \bigcap D_L\right)$ will be evaluated in next section and $q_{m+\nu}(a) = P(D_L)$, and $q_{2m,m+\nu}(a) = P\left(D_1 \bigcap D_L\right)$ have to be evaluated separately. First,

$$
q_{m+\nu}(a) = \sum_{j=0}^{\min(k-1+\min(k-1,\nu),a)} q(m+\nu|j)\frac{\binom{m+\nu}{j}\binom{(L-1)m}{a-j}}{\binom{n}{a}}), \tag{6}
$$

where

$$
q(m+\nu|j) = \sum_{j=0}^{\min(k-1+\min(k-1,\nu),a)} \sum_{i=0}^{\min(m-\nu,k-1,j)} [1-P(k-i;\nu,2\nu,j-i)]\frac{\binom{m-\nu}{i}\binom{2\nu}{j-i}\binom{(L-1)m}{a-j}}{\binom{n}{a}}) \tag{7}
$$

and $P(k;m,n,a)$ can be evaluated using Naus (1974, Theorem 1). The term

$$
q_{2m,m+\nu}(a) = \sum_{j_1=0}^{\min(2k-2,a)} \sum_{j_2=0}^{\min(k-1+\min(k-1,\nu),a-j_1)} q(2m|j_1)q(m+\nu|j_2)\frac{\binom{2m}{j_1}\binom{m+\nu}{j_2}\binom{(L-3)m}{a-j_1-j_2}}{\binom{n}{a}}.
$$

where $q(2m|j_1)$ and $q(m+\nu|j_2)$ can be evaluated using Equations (**??**) and (**??**).

3. Evaluation of $q_{2m+\nu}(a)$

The term

$$
q_{2m+\nu}(a) = \sum_{j=0}^{\min(2k-2+\min(k-1,\nu),a)} q(2m+\nu|j)\frac{\binom{2m+\nu}{j}\binom{(L-2)m}{a-j}}{\binom{n}{a}}, \tag{8}
$$

where $q(2m+\nu|j), j = \sum_{i=1}^{2m+\nu} Y_i$ is the total number of $1's$ in the $2m+\nu$ trails, will be evaluated based on the following algorithm.

Let $Y_1, \cdots, Y_{2m+\nu}$ be a sequence of $0-1$ Bernoulli trials with total number of $1's$ equal to j. We partition the trials $1, \cdots 2m+\nu$ into 9 of disjoint sets biven bwlow:

$$
I_{2i-1} = \{(i-1)m, \cdots, (i-1)m+\nu\}, i = 1,2,3
$$

and

$$
I_{2i} = \{\underbrace{(i-1)m+\nu+1}_{I_{2i_1}}\} \bigcup \{\underbrace{(i-1)m+\nu+2, \cdots, (i-1)m+m-1}_{I_{2i_2}}\} \bigcup \{\underbrace{(i-1)m+m}_{I_{2i_3}}\}, i = 1,2.
$$

Let $\Gamma_1 = I_1 \bigcup I_3 \bigcup I_5$ and $\Gamma_2 = I_2 \bigcup I_4$ be the disjoint index sets. For $t \in \Gamma_i, i = 1, 2$ let $S_t(m) = Y_t + \cdots + Y_{t+m-1}$. For $i = 1, 2$ define

$$M_i = \max_{t \in \Gamma_i} S_t(m)$$

to be the maximum number of the moving windows in the index set Γ_i. Then given $\{n_{i_j}\}$, the number of $1's$ in I_{i_j}, M_1 and M_2 are independent. Therefore we have

$$q(2m+\nu|j) = \sum_{\{n_{i_j}\} \in C} P\left(M_1 < k | \{n_{i_j}\}\right) P\left(M_2 < k | \{n_{i_j}\}\right) P(\{n_{i_j}\}) \tag{9}$$

where the summation extends over the set C of all partitions of j into sets I_{i_j} with integers n_{i_j} satisfying: $n_{2i} = \sum_{j=1}^{3} n_{2i_j}$ for $i = 1, 2$ and $\sum_{i=1}^{5} n_i = j$ and $n_i + n_{i+1} < k$, for $i = 1, 2, 3, 4$ respectively. And n_{i_j} follow an 8 dimensional multivariate hypogeomatric distribution

$$P(\{n_{i_j}\}) = \frac{\binom{\nu}{n_1}\binom{1}{n_{2_1}}\binom{m-\nu-2}{n_{2_2}}\binom{1}{n_{2_3}}\binom{\nu}{n_3}\binom{1}{n_{4_1}}\binom{m-\nu-2}{n_{4_2}}\binom{1}{n_{4_3}}\binom{\nu}{j-n_1-n_2-n_3-n_4}}{\binom{2m+\nu}{j}}.$$

Note that

$$P\left(M_2 \geq k | \{n_{i_j}\}\right) = P\left(k - (n_{2_3} + n_3 + n_{4_1}); m - (\nu+2), 2[m-(\nu+2)], n_{2_2} + n_{4_2}\right), \tag{10}$$

where M_2 is the scan statistic with window length of $m - (\nu + 2)$ conditional on the total number of $1's$ being $n_{2_2} + n_{4_2}$ in $2[m - (\nu + 2)]$ trials such that at least one window contains at least $k - (n_{2_3} + n_3 + n_{4_1})$ $1's$. To derive the conditional distribution of M_1, let $X_{2i-1}(t)$ denote the number of $1's$ in trials $(i-1)m+1 \cdots (i-1)m+t$. Then for $i = 1, 2, 3$

$$X_{2i-1}(t) = \sum_{j=(i-1)m+1}^{(i-1)m+t} Y_j.$$

It follows that

$$\begin{aligned} P\left(M_1 < k | \{n_{i_j}\}\right) &= P\left(\bigcap_{t=0}^{\nu} X_3(t) - X_1(t) + n_1 + n_2 < k, X_5(t) - X_3(t) + n_3 + n_4 < k\right) \\ &= P\left(\bigcap_{t=0}^{\nu} X_1(t) + \alpha_1 > X_3(t) + \alpha_3 > X_5(t) + \alpha_5\right) \\ &= \det |h_{ij}|, \end{aligned}$$

where for $i = 1, 2, 3$

$$\alpha_{2i-1} = .5[l - (2i-1)]k - \sum_{j=i}^{L} n_{2j-1} + \sum_{j=1}^{i-1} n_{2j},$$

$$h_{ij} = \frac{n_{2i-1}!(\nu - n_{2i-1})!}{(n_{2i-1} + \alpha_{2i-1} - \alpha_{2j-1})!(\nu - n_{2i-1} - \alpha_{2i-1} + \alpha_{2j-1})!}$$

and h_{ij} is defined as zero when any of the factorials is less than zero.

Note that for $0 \leq \nu \leq m-1$ Equation (**??**) can be extented to the general case for $L > 3$.

4. A Tighter Bonferroni-Type Inequality

A different method to derive inequalities for $P(k; m, n, a, \nu)$ is similar to the one used in Section 3. Let the events D_i be defined as in Equation (**??**). For $1 \leq i \leq L$, define the event that the number of $1's$ in trials $(i-1)m+1, \cdots, im$ is at least k, denoted by

$$A_i^* = Y_{(i-1)m+1} + \cdots + Y_{im} \geq k \tag{11}$$

and the last event be

$$A_{L+1}^* = Y_{Lm+1} + \cdots Y_{Lm+\nu} \geq k,$$

and $A^* = \bigcup_{i=1}^{L+1} A_i^*$. Let the events $D_i, 1 \leq i \leq L-1$ be defined as in Equation (**??**) and D_L be defined as in Equation (**??**), then similar to Equation (**??**), for $1 \leq i \leq L$ define $G_i = D_i^c \bigcap A^c$ and $G = \bigcup_{i=1}^{L} G_i$. It follows that

$$P(k; m, n, a, \nu) = P(A^*) + P(G). \tag{12}$$

We propose to evaluate $P(A^*)$ exactly using a recursive formula given below and to bound $P(G)$ by a upper inequality of Hunter (1976) and a lower inequality of Kwerel (1975).

Let $h(a, l, m, k, \nu)$ be the number of ways to distribute the a $1's$ among the $n = l \times m + \nu$ trials such that all the $l+1$ groups have less than k $1's$. Abbreviate $h(a, l, m, k, \nu)$ to $h(a, l, \nu)$. Define

$$H(a, l, \nu) = \frac{h(a, l, \nu)}{\binom{n}{a}}.$$

Note that $H(a, l, \nu)$ is the cumulative distribution function of the largest order statistic of $l+1$ dimensional multivariate hypergeometric vector with parameter a and l number of cell with size m and the last cell with size ν evaluated at $k-1$. The following recursion holds:

$$H(a, l+1, \nu) = \sum_{j=0}^{\min(\nu, a, k-1)} \frac{\binom{\nu}{j}\binom{n-\nu}{a-j}}{\binom{n}{a}} H(a-j, l)$$

with the initial conditions $H(a, 1) = 1$ for $a < k$ and $H(a, 1) = 0$ for $a \geq k$. It follows that

$$P(A^*) = 1 - H(a, L+1, \nu).$$

The upper Hunter (1976) inequality for $P(G)$ is given by

$$\begin{aligned} P(G) &\leq \sum_{i=1}^{L} P(G_i) - \sum_{i=1}^{L-1} P\left(G_i \bigcap G_{i+1}\right) \\ &= (L-1)P(G_1) + P(G_L) - (L-2)P(G_1 \bigcap G_2) - P(G_{L-1} \bigcap G_L) \end{aligned}$$

and the second order Bonferroni-type lower inequality of Kwerel (1975) for $P(G)$ is given by

$$P(G) \geq \frac{2s_1^*}{b^*} - \frac{2s_2^*}{b^*(b^*-1)}, \tag{13}$$

where

$$b^* = [\frac{2s_2^*}{s_1^*} + 2],$$

$$s_1^* = \sum_{i=1}^{L} P\left(G_i\right) = (L-1)P(G_1) + P(G_L),$$

and

$$\begin{aligned} s_2^* &= \sum_{j=2}^{L}\sum_{i=1}^{j-1} P\left(G_i \bigcap G_j\right) \\ &= (L-2)P\left(G_1 \bigcap G_2\right) + P\left(G_{L-1} \bigcap G_L\right) \\ &\quad +.5(L-2)(L-3)P\left(G_1 \bigcap G_3\right) + (L-2)P\left(G_1 \bigcap G_L\right). \end{aligned}$$

To evaluate $P\left(G_1\right)$, $P\left(G_1 \bigcap G_2\right)$ and $P\left(G_1 \bigcap G_3\right)$, let $n_i = Y_{(i-1)m+1} + \cdots Y_{im}$ be the number of $1's$ in the windows $[(i-1)m+1, im]$ for $1 \leq i \leq L$. Condition on n_1, n_2, n_3 respectively, to get

$$P\left(G_1\right) = \sum_{n_1=0}^{\min(a,k-1)} \sum_{n_2=\max(0,k-n_1)}^{\min(a-n_1,k-1)} P(D_1^c|n_1,n_2) H(a-n_1-n_2, L-2, \nu) \frac{\binom{m}{n_1}\binom{m}{n_2}\binom{n-2m}{a-n_1-n_2}}{\binom{n}{a}}.$$

$$\begin{aligned} P\left(G_1 \bigcap G_2\right) &= \sum_{n_1=0}^{\min(a,k-1)} \sum_{n_2=\max(0,k-n_1)}^{\min(a-n_1,k-1)} \sum_{n_3=\max(0,k-n_2)}^{\min(a-n_1-n_2,k-1)} P\left(D_1^c \bigcap D_2^c|n_1,n_2,n_3\right) \\ &\quad H(a-n_1-n_2-n_3, L-3, \nu) \frac{\binom{m}{n_1}\binom{m}{n_2}\binom{m}{n_3}\binom{n-3m}{a-n_1-n_2-n_3}}{\binom{n}{a}}, \end{aligned}$$

$$\begin{aligned} P\left(G_1 \bigcap G_3\right) &= \sum_{n_1=0}^{\min(a,k-1)} \sum_{n_2=\max(0,k-n_1)}^{\min(a-n_1,k-1)} \sum_{n_3=0}^{\min(a-n_1-n_2,k-1)} \sum_{n_4=\max(0,k-n_3)}^{\min(a-n_1-n_2-n_3,k-1)} P\left(D_1^c \bigcap D_3^c|n_1,n_2,n_3,n_4\right) \\ &\quad H(a-n_1-n_2-n_3-n_4, l-4, \nu) \frac{\binom{m}{n_1}\binom{m}{n_2}\binom{m}{n_3}\binom{m}{n_4}\binom{n-4m}{a-n_1-n_2-n_3-n_4}}{\binom{n}{a}}, \end{aligned}$$

where

$$P\left(D_1^c \bigcap D_2^c|n_1,n_2,n_3\right) = 1 - P\left(D_1|n_1,n_2\right) - P\left(D_2|n_2,n_3\right) + P\left(D_1 \bigcap D_2|n_1,n_2,n_3\right),$$

$$P\left(D_1^c \bigcap D_3^c|n_1,n_2,n_3,n_4\right) = \left[1 - P\left(D_1|n_1,n_2\right)\right]\left[1 - P\left(D_3|n_3,n_4\right)\right]$$

and $P\left(D_1|n_1,n_2\right)$ and $P\left(D_1 \bigcap D_2|n_1,n_2,n_3\right)$ can be evaluated using Naus(1974, Equations (2.7) and (2.5)) respectively.

The terms $P\left(G_L\right)$, $P\left(G_{L-1} \bigcap G_L\right)$ and $P\left(G_1 \bigcap G_L\right)$ are more complex to evaluated using Huntington and Naus (1975) in the following:

$$P\left(G_L\right) = \sum_{n_L=0}^{\min(a,k-1)} \sum_{n_{L+1}=\max(0,k-n_L)}^{\min(a-n_L,k-1,\nu)} P(D_L^c|n_L,n_{L+1}) H(a-n_L-n_{L+1}, L-1) \frac{\binom{m}{n_L}\binom{\nu}{n_{L+1}}\binom{n-m-\nu}{a-n_L-n_{L+1}}}{\binom{n}{a}}$$

$$P\left(G_{L-1}\bigcap G_L\right) = \sum_{n_{L-1}=0}^{\min(a,k-1)} \sum_{n_L=\max(0,k-n_{L-1})}^{\min(a-n_{L-1},k-1)} \sum_{n_{L+1}=\max(0,k-n_L)}^{\min(a-n_{L-1}-n_L,k-1)} P\left(D_{L-1}^c\bigcap D_L^c|n_{L-1},n_L,n_{L+1}\right)$$
$$H(a-n_{L-1}-n_L-n_{L+1},L-2)\frac{\binom{m}{n_{L-1}}\binom{m}{n_L}\binom{\nu}{n_{L+1}}\binom{n-2m-\nu}{a-n_{L-1}-n_L-n_{L+1}}}{\binom{n}{a}}$$

and

$$P\left(G_1\bigcap G_L\right) = \sum_{n_1=0}^{\min(a,k-1)} \sum_{n_2=\max(0,k-n_1)}^{\min(a-n_1,k-1)} \sum_{n_L=0}^{\min(a-n_1-n_2,k-1)} \sum_{n_{L+1}=\max(0,k-n_L)}^{\min(a-n_1-n_2-n_L,k-1)}$$
$$P\left(D_1^c\bigcap D_L^c|n_1,n_2,n_L,n_{L+1}\right)$$
$$H(a-n_1-n_2-n_L-n_{L+1},L-3)\frac{\binom{m}{n_1}\binom{m}{n_2}\binom{m}{n_L}\binom{\nu}{n_{L+1}}\binom{n-3m-\nu}{a-n_1-n_2-n_L-n_{L+1}}}{\binom{n}{a}}$$

where

$$P(D_L^c|n_L,n_{L+1}) = 1-P(D_L|n_L,n_{L+1}),$$
$$P\left(D_1^c\bigcap D_L^c|n_1,n_2,n_L,n_{L+1}\right) = [1-P(D_1|n_1,n_2)]\,[1-P(D_L|n_L,n_{L+1})]$$

and

$$P\left(D_{L-1}^c\bigcap D_L^c|n_{L-1},n_L,n_{L+1}\right) = 1-P(D_{L-1}|n_{L-1},n_L)-P(D_L|n_L,n_{L+1})$$
$$+P\left(D_{L-1}\bigcap D_L|n_{L-1},n_L,n_{L+1}\right)$$

where $P\left(D_{L-1}\bigcap D_L|n_{L-1},n_L,n_{L+1}\right)$ can be evaluated similarly to Equation (**??**) with slight difference in partition such that

$$P(\{n_{i_j}\}) = \frac{\binom{\nu}{n_1}\binom{1}{n_{2_1}}\binom{m-\nu-2}{n_2-n_{2_1}-n_{2_3}}\binom{1}{n_{2_3}}\binom{\nu}{n_3}\binom{1}{n_{4_1}}\binom{m-\nu-2}{n_4-n_{4_1}-n_{4_3}}\binom{1}{n_{4_3}}}{\binom{m}{n_1+n_2}\binom{m}{n_3+n_4}}.$$

Numerical results for this example are given in Table 9 in Section 5.

5. Numerical Results and Conclusion

In Table 1 we present numerical results for an example in molecular biology outlined in this article. For selected values of k, Bonferroni-type inequalities for $P(k;m,n,a,\nu)$ are given. Bonferroni-type inequalities are evaluated for the probability of observing k or more basic charges in a window of size $m=30$ in a sequence of $n=968$ residues of the adenovirus type 2 hexon protein. From the numerical results one can conclude that observing 11 or more basic residues is unusual for the underlying assumption of modeling the charged residues as iid $-1,0,+1$ trinomial trials, while observing 10 basic residues is not that unusual.

In conclusion, we would like to state that Bonferroni-type inequalities for scan statistics investigated in this article produced tight inequalities that outperformed the inequalities that have been previously derived. The dependence structure captured in the scanning window representation of the scan statistics allowed us to construct relatively simple second order Bonferroni-type inequalities and achieve the desired accuracy.

Table 1. Comparison of Bonferroni-type inequalities for $P(k; m, n, a, \nu)$ ***for*** $m = 30, a = 100$**,** $\nu = 8$ ***and*** $n = 968$

k	$\hat{P}(k; m, n, a, \nu)$	BTLB1	BTLB2	BTUB2	BTUB1
8	.6855	.5696	.6445	.8443	.9899
9	.2925	.2707	.2849	.3061	.3182
10	.0787	.0813	.0817	.0830	.0839
11	.0181	.0184	.0183	.0184	.0185
12	.0032	.0034	.0034	.0034	.0035

NOTE: BTLB1, BTLB2 are Bonferroni-type lower inequalities and BTUB1, BTUB2 are Bonferroni-type upper inequalities for $P(k; , n, m, a, \nu)$, respectively.

REFERENCES

Altschul, S. F. and Erickson, B. W. (1988). Significance levels for biological sequence comparison using non-linear similarity functions, *Bulletin of Mathematical Biology*, **50**, 77-92.

Balakrishnan, N., Balasubramanian, K., and Viveros, R. (1993) On sampling inspection plans based on the theory of runs, *The Mathematical Scientist*, **18**, 113-126.

Bogush, Jr. A. J. (1972). Correlated clutter and resultant properties of binary signals, *IEEE Transactions on Aerospace Electronic Systems*, **9**, 208-213.

Chao, M. T., Fu, J. C. and Koutras, M. V. (1995). Survey of reliability studies of consecutive-k-out-of-n: F & Related Systems *IEEE Transactions on Reliability* Vol. 44, No. 1, 120-127.

Chen,J., Glaz,J., Naus,.J., and Wallenstein, S. (2001). Bonferroni-type Inequalities for Conditional scan Statistics. will be appear in *Statistics & Probability Letters.*

Diaconis,P and Mosteller, F (1989) Methods for studying coincidences, Journal of American Statistical Association, **84**, 853-861.

Fu, J. C. (1986). Reliability of consecutive-k-out-of-n: F system with (k-1)-step Markov dependence, *IEEE Transaction on Reliability*, **35**, 603-602.

Fu, J. and Koutras, M. (1994). Distribution theory of runs: A Markov chain approach, *Journal of American Statistical Association*, **89**, 1050-1058.

Glaz, J. (1983). Moving window detection for discrete data, *IEEE Transaction on Information Theory, IT-29, 457-462.*

Glaz, J. and Naus, J. I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data. *Annals of Applied Probability* **1**, 306-318.

Glaz, J. (1996). Discrete scan statistics with applications to minefields detection. in *Proceedings of Conference SPIE*, Orlando, Florida, SPIE, **2765**, 420-429.

Godbole, A. P. (1990). Specific formulae for some success runs distributions, *Statistics & Probability Letters*, **10**, 119-124.

Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events, *Advances in Applied Probability*, **23**, 851-865.

Godbole A. P. (1993). Approximate reliabilities of m-consecutive-k-out-of-n: failure systems, *Satatstica Sinica*, **3**, 321-327.

Gordon, L., Schilling, M. F., and Waterman, M. S. (1986). An extreme value theory for long head runs, *Probability Theory and Related Fields*, **72**, 279-288.

Greenberg, I. (1970). The first occurrence of n successes in N trails, *Technometrics*, **12**, 627-634.

Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov Ahain, *Statistica Sinica*, **3**, 313-320.

Hunter, D. (1976). An upper bound for the probability of a union. *Journal of Applied Probability.* **13**, 597-603.

Huntington and Naus, J.I. (1975) A simple expression for k^{th} nearest neighbor coincidence probabilities. *Annals of Probability*, **3**, 894-896.

Karlin, S., Blaisdell, B., Mocarski, E., and Brendel, V. (1989). A method to identify distinctive charge configurations in protein sequences With applications to human Herpesvirus polypeptides. *Journal of Molecular Biology.* **205**, 165-177.

Karlin, S., Ghandour, G. (1985). Multiple-alphabet amono acid sequence comparison of the immunoglobulin k-chain constant domain, *Proceeding of the National Academy of Science, USA*, **82**, 8597-8601.

Karlin, S. and Ost, F. (1987). Counts of long aligned word matches among random letter sequences, *Advances in Applied Probability*, **19**, 293-351.

Koutras, M.V. and Alexandrou V.A. (1997) No-parametric randomness test based on success runs of fixed length, *Statistics & Probability Letters*, **32**, 393-404.

Kwerel, S. M. (1975). Most stringent bounds on aggregated probabilities of partially specified dependent probability system. *Journal of the American Statistical Association* **70**, 472-479.

Mott, R. F., Kirkwood, T. B. L. and Curnow, R. N. (1990). An accurate approximation to the distribution of the length of longest matching word between two random DNA sequences, *Bulletin of Mathematical Biology*, **52**, 773-784.

Moye, L. A., Kapadia, A. S. Cech, I. M. and Hardy, R. J. (1988) The theory of runs with applications to drought prediction, *Journal of Hydrology*, **103**, 127-137.

Naus, J. I. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association* **69**, 810-815.

Naus, J. I. and Sheng, K. N. (1996). Screening for unusual matched segments in multiple protein sequences, *Communication in Statististics, Simulation and Computation*, **25**, 937-952.

Nelson, J. B. (1978). Minimal order models for false alarm calculations on sliding windows, *IEEE Transactions on Aerospace and Electronic System*, **15**, 352-363.

Runnels, L. K., Thompson, R. and Runnels, P. (1968). Near perfect runs as a learning criterion, *Journal of the American Statistical Association* , **5**, 362-368.

Saperstein, B. (1972). The generalized birthday roblem, *Journal of the American Statistical Association* , **67**, 425-428.

Schwager, S.J. (1983). Run probabilities in sequences of Markov-dependent trials, *Journal of the American Statistical Association,* **78**, 168-175.

Sheng, K. N. and Naus, J. I. (1994). Pattern matching between two non-alligned random sequences, *Bulletin of Mathematical Biology*, **56**, 1143-1162.

Sheng, K. N. and Naus, J. I. (1996). Matching rectangles in 2-dimensions. *Statistics & Probability Letters*, **26**, 83-90.

Viveros, R., and Balakrishnan, N. (1993).Statistical inference from startup demonstration test data, *Journal of Quality Technology*, **25**, 119-130.

Waterman, M.S. (1995). *Introduction to computational Biology*, Chapman & Hall, London.

Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69**, 297-302.