*The Annals of Applied Probability*
1991, Vol. 1, No. 2, 306–318

# TIGHT BOUNDS AND APPROXIMATIONS FOR SCAN STATISTIC PROBABILITIES FOR DISCRETE DATA

By Joseph Glaz[1] and Joseph I. Naus

*The University of Connecticut and Rutgers–The State University of New Jersey*

Let $X_1, X_2, \dots$ be a sequence of independently and identically distributed integer-valued random variables. Let $Y_{t-m+1,t}$ for $t = m$, $m+1, \dots$ denote a moving sum of $m$ consecutive $X_i$'s. Let $N_{m,T} = \max_{m \le t \le T}\{Y_{t-m+1,t}\}$ and let $\tau_{k,m}$ be the waiting time until the moving sum of $X_i$'s in a scanning window of $m$ trials is as large as $k$. We derive tight bounds for the equivalent probabilities $P(\tau_{k,m} > T) = P(N_{m,T} < k)$. We apply the bounds for two problems in molecular biology: the distribution of the length of the longest almost-matching subsequence in aligned amino acid sequences and the distribution of the largest net charge within any $m$ consecutive positions in a charged alphabet string.

**1. Introduction.** Scientists in a variety of fields seek the distribution of the maximum value of a moving average or sum. This paper derives tight bounds and accurate approximations for this distribution for the case of a moving sum of independently and identically distributed integer-valued random variables. The bounds are computed and evaluated for two problems in molecular biology.

Let $X_1, X_2, \dots$ be a sequence of i.i.d. integer-valued random variables. Let $Y_{r,t} = \sum_{i=r}^{t} X_i$ for $t \ge r \ge 1$, and be 0 otherwise. $Y_{1,m}, Y_{2,m+1}, Y_{3,m+2}, \dots$ represents a moving sum of the $X$'s within a window of $m$ trials. Section 2 provides bounds and approximations for the distribution of $\max_{m \le t \le T}\{Y_{t-m+1,t}\}$.

In studying amino acid sequences, molecular biologists look at various classification schemes: a chemical alphabet (eight letters), a functional alphabet (four letters), a charge alphabet (three letters, $-1, 0, +1$) and others [Karlin and Ghandour (1985)]. These scientists compare sequences corresponding to different species looking for long aligned subsequences that match on most of the positions. They seek to determine what is an unusually long match. Erdös and Révész (1975), Arratia, Gordon and Waterman (1986) and Karlin and Ost (1988) have provided asymptotic results. These serve as rough approximations for the null distribution of the length of the longest almost-matching subsequence. In this application, $X_i$ takes the value 1 if the amino acid sequences match in position $i$, and 0 otherwise. Section 3 specializes the bounds and approximations of Section 2 for this problem.

Received November 1989; revised October 1990.

[1] Research supported in part by the University of Connecticut Research Foundation.

*AMS* 1980 *subject classifications.* 60F10, 60F99.

*Key words and phrases.* Longest matching subsequences, scan statistics, clustering probabilities.

Molecular biologists also seek to determine what is unusual blocking within individual amino acid sequences. For the charge alphabet they seek to know what is an unusually large positive charge within any subsequence of given length [see Karlin, Blaisdell, Mocarski and Brendel (1989)]. For this application, $X_i$ takes the values $-1, 0, 1$. Section 4 deals with this problem, providing bounds and approximations.

**2. Bounds and approximations.** Let $X_1, X_2, \dots$ be i.i.d. discrete random variables with $P(X_i = j) = P_j$ for $j = 0, 1, 2, \dots, c$; $P(X_i = j) = 0$ elsewhere, where $\Sigma_{j=0}^{c} P_j = 1$. Let $Y_{r,t} = \Sigma_{i=r}^{t} X_i$ for $t \geq r \geq 1$, $Y_{r,t} = 0$ otherwise, let

$$N_{R,T} = \max_{R \leq t \leq T} \{Y_{t-m+1,t}\};\ \tau_{k,m} = \inf\{t \geq 1, \text{ s.t. } Y_{\max(1,t-m+1),t} \geq k\}.$$

$\tau_{k,m}$ is the waiting time until the sum of the $X_i$'s in a scanning interval of length $m$ is as large as $k$. We seek to find the distribution of $\tau_{k,m}$. Let

$$G_{k,m}(T) = P(\tau_{k,m} > T) = P(N_{m,T} < k), \qquad f_{k,m}(t) = P(\tau_{k,m} = t).$$

Abbreviate $\tau_{k,m}$, $G_{k,m}(T)$ and $f_{k,m}(t)$ to $\tau$, $G(T)$ and $f(t)$ when $k$ and $m$ are understood. We derive the following bounds for $G(T)$.

THEOREM 1. *For $c < k$; $i, j, n$ integers $\geq 1$, let*

$$A_{1,n} = f((n+1)m), \qquad A_{j,n} = A_{1,n}(1 - A_{j-1,n})^{-nm+1}$$

*and*

$$B_{1,n} = f(nm)/G((n+1)m - 1), \qquad B_{j,n} = f(nm)(1 + B_{j-1,n})^{nm}.$$

*For $T \geq im$,*

$$G(T) \leq G(im)(1 - A_{j,n})^{T-im} \quad \text{for } T \geq (n+1)m \tag{2.1}$$

*and*

$$G(T) \geq G(im)/(1 + B_{j,n})^{T-im} \quad \text{for } T \geq nm. \tag{2.2}$$

The proof of Theorem 1 follows (2.2f).

REMARKS. For $j = 1$, the bounds (2.1) and (2.2) converge as $G(nm) \to 1$ and $f(nm) \to 0$. If we let the total number of trials become large, the bounds will be tight if $(1 - f((n+1)m))^{T-im}$ is close to $1/(1 + f(nm)/G((n+1)m - 1))^{T-im}$. Theorem 1 can be modified to handle the case where some of the values that the $X$'s take are greater than $k$, are unbounded or are negative.

To apply the bounds (2.1) and (2.2) in specific cases, we need to evaluate $f(nm)$ and $G(im)$ and $G((n+1)m - 1)$ for some appropriate integer values of $i$ and $n$. The larger $i$ and $n$, the better the bounds appear, though the harder it is to evaluate $G$ and $f$. The simplest cases of the bounds are given in inequalities (2.1a)–(2.2f).

UPPER BOUNDS.

$$G(T) \le G(2m)(1 - f(2m))^{T-2m} \quad \text{for } T \ge 2m, \tag{2.1a}$$

$$G(T) \le G(3m)(1 - f(2m))^{T-3m} \quad \text{for } T \ge 3m. \tag{2.1b}$$

The bounds (2.1a) and (2.1b) can be improved by replacing $f(2m)$ by $A_{j,1}$, where we compute $A_{j,1}$ by setting $A_{1,1} = f(2m)$, and iterating $A_{j,1} = f(2m)(1 - A_{j-1,1})^{-m+1}$.

LOWER BOUNDS.

$$G(T) \ge G(2m)/(1 + (f(m)/G(2m-1)))^{T-2m} \quad \text{for } T \ge 2m, \tag{2.2a}$$

$$G(T) \ge G(2m)/(1 + (f(2m)/G(3m-1)))^{T-2m} \quad \text{for } T \ge 2m, \tag{2.2b}$$

$$G(T) \ge G(3m)/(1 + (f(2m)/G(3m-1)))^{T-3m} \quad \text{for } T \ge 3m. \tag{2.2c}$$

Note that since $G(3m) \le G(3m-1)$, (2.2c) implies

$$G(T) \ge G(3m)/(1 + (f(2m)/G(3m)))^{T-3m} \quad \text{for } T \ge 3m. \tag{2.2d}$$

We will prove below that $G(2m)G(2m-1) \le G(3m-1)$ and thus (2.2b) implies (2.2e), and (2.2c) implies (2.2f).

$$G(T) \ge G(2m)/(1 + (f(2m)/G(2m)G(2m-1)))^{T-2m} \quad \text{for } T \ge 2m, \tag{2.2e}$$

$$G(T) \ge [G(2m)]^2/(1 + (f(2m)/G(2m)G(2m-1)))^{T-3m} \quad \text{for } T \ge 3m. \tag{2.2f}$$

To prove Theorem 1, we first prove three preliminary lemmas. We use the known result [see Esary, Proschan and Walkup (1967)]:

LEMMA 1. *If $X_1, X_2, \ldots, X_n$ are independent random variables, then for any two n-variable real-valued functions f and g that are both coordinatewise monotone increasing (or both decreasing)*

$$E[f(X_1, \ldots, X_n)g(X_1, \ldots, X_n)] \ge E[f(X_1, \ldots, X_n)] \times E[g(X_1, \ldots, X_n)]. \tag{2.3}$$

LEMMA 2. *For $t \ge nm$, $c < k$ and $n = 1, 2, \ldots,$*

$$f(t) \le G(t - nm) f(nm). \tag{2.4}$$

PROOF. For $t \geq (n+1)m$,

$$f(t) \equiv P(\tau = t) = P\left(\{N_{m,t-1} \leq k-1\} \cap \left(\bigcup_{i=1}^{c} D_i\right)\right) = \sum_{i=1}^{c} f_i(t), \tag{2.5}$$

where

$$D_i = (Y_{t-m+1,t-1} = k-i) \cap (X_t \geq i) \tag{2.6}$$

and

$$f_i(t) = P(\{N_{m,t-1} \leq k-1\} \cap D_i).$$

Write

$$\begin{aligned} f_i(t) &= P(\{N_{m,t-nm} \leq k-1\} \cap \{N_{t-nm+1,t-1} \leq k-1\} \cap D_i) \\ &\leq P(\{N_{m,t-nm} \leq k-1\} \cap \{N_{t-(n-1)m,t-1} \leq k-1\} \cap D_i) \\ &= G(t-nm) f_i(nm), \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} f_i(nm) = P\big(&\{N_{m,nm-1} \leq k-1\} \\ &\cap \{Y_{(n-1)m+1,nm-1} = k-i\} \cap \{X_{nm} \geq i\}\big). \end{aligned} \tag{2.8}$$

Substituting inequality (2.7) into (2.5) yields $f(t) \leq G(t-nm)\Sigma_{i=1}^{c} f_i(nm) = G(t-nm)f(nm)$, proving Lemma 2 for $t \geq (n+1)m$. To verify that inequality (2.4) holds for $nm \leq t \leq (n+1)m - 1$, define $G(0) = 1$, $G(t) = P(X_1 + \cdots + X_t \leq k-1)$ for $1 \leq t \leq m$; replace $N_{m,t-nm}$ by $Y_{1,t-nm}$ and drop $\{N_{t-(n-1)m,t-1} \leq k-1\}$ if $n = 1$ in (2.7). □

LEMMA 3. *For* $t \geq (n+1)m$, $c < k$, $n = 1, 2, \ldots,$

$$f(t) \geq f((n+1)m)G(t-nm). \tag{2.9}$$

PROOF. Define the events

$$E_1 = \{N_{m,t-nm} \leq k-1\}$$

and

$$E_{2,i} = \{N_{t-nm,t-1} \leq k-1\} \cap \{Y_{t-m+1,t-1} = k-i\}.$$

Write

$$f_i(t) = P(E_1 \cap E_{2,i})P(X_t \geq i), \tag{2.10}$$

where $E_1$ depends on $X_1, \ldots, X_{t-nm}$ and $E_{2,i}$ depends on $X_{t-(n+1)m+1}, \ldots, X_{t-1}$.

Let $S = (X_{t-nm+1}, \ldots, X_{t-1})$. Condition on the set of values in $S$. Let $I(E_1)$ and $I(E_{2,i})$ be indicator functions, $I(E) = 1$ if $E$ occurs, 0 otherwise. Conditional on $S$, $I(E_1)$ and $I(E_{2,i})$ are both coordinatewise decreasing functions of the $X$'s not in $S$. Apply Lemma 1 to find

$$P(E_1 \cap E_{2,i}|S) \geq P(E_1|S)P(E_{2,i}|S) = P(E_1)P(E_{2,i}|S). \tag{2.11}$$

Average both sides of inequality (2.11) over the distribution of $X$'s in $S$ to find

$$\begin{aligned} P(E_1 \cap E_{2,i}) &\geq P(E_1)P(E_{2,i}) = G(t-nm)P(E_{2,i}) \\ &= G(t-nm) f_i((n+1)m)/P(X_t \geq i). \end{aligned} \tag{2.12}$$

Substituting into (2.10) and summing over $i = 1, \dots, c$ completes the proof of Lemma 3. □

Proof of Theorem 1. Use Lemma 2 to find the lower bounds (2.2) and Lemma 3 to find the upper bounds (2.1) of Theorem 1. From Lemma 3, for $t \geq (n+1)m$,

$$\begin{aligned} f(t) \equiv G(t-1) - G(t) &\geq f((n+1)m)G(t-nm) \\ &\geq f((n+1)m)G(t-1), \end{aligned} \tag{2.13}$$

since $G(t)$ is decreasing in $t$. From (2.13) and letting $f((n+1)m) \equiv A_{1,n}$,

$$G(t-1)/(1-A_{1,n})^{t-1} \geq G(t)/(1-A_{1,n})^{t} \tag{2.14}$$

so that $G(t)/(1-A_{1,n})^t$ is decreasing in $t$. Thus for $T \geq im$ and $T \geq (n+1)m$,

$$G(T) \leq G(im)(1-A_{1,n})^{T-im}. \tag{2.15}$$

This is the upper bound (2.1) for $j = 1$.

Further, from the fact that $G(t)/(1-A_{1,n})^t$ is decreasing for $t \geq (n+1)m$,

$$G(t-nm) \geq G(t-1)(1-A_{1,n})^{t-nm}/(1-A_{1,n})^{t-1}. \tag{2.16}$$

Substituting this into inequality (2.13) gives

$$(1-A_{2,n})G(t-1) \geq G(t), \tag{2.17}$$

where $A_{2,n} = A_{1,n}(1-A_{1,n})^{-nm+1}$. Iterating in this fashion gives upper bounds (2.1) for $j = 1, 2, \dots$ .

For the lower bounds (2.2), use Lemma 2 to find for $t \geq nm$,

$$f(t) = G(t-1) - G(t) \leq G(t-nm) f(nm). \tag{2.18}$$

From Lemma 1, we note that for $t \geq m$,

$$\begin{aligned} G(t+u) &= P(\{N_{m,t} \leq k-1\} \cap \{N_{t+1,t+u} \leq k-1\}) \\ &\geq G(t)G(u+m-1), \end{aligned} \tag{2.19}$$

since $I(N_{m,t} \leq k-1)$ and $I(N_{t+1,t+u} \leq k-1)$ are both decreasing functions of the $X$'s. In particular, (2.19) implies that for $t \geq nm$,

$$G(t-nm) \leq G(t)/G((n+1)m-1). \tag{2.20}$$

Substituting (2.20) into (2.18) gives

$$G(t-1) \leq (1+B_{1,n})G(t), \tag{2.21}$$

where $B_{1,n} = f(nm)/G((n+1)m-1)$. Continuing as before yields the lower bounds (2.2). This completes the proof of Theorem 1. □

Our proof of Theorem 1 generalizes the approach of Janson (1984), who derived bounds for a specific continuous-time process, the Poisson process. Note that (2.19) implies that $G(3m-1) \geq G(2m)G(2m-1)$ and thus that inequalities (2.2b) and (2.2c) respectively imply (2.2e) and (2.2f). For i.i.d. discrete random variates, there exist computationally feasible procedures to evaluate $G(t)$ for $t \leq 2m$. These general procedures are derived in Saperstein (1976), and the details of adaptation are available in Glaz and Naus (1989). Thus, in general, one can compute $G(2m)$, $G(2m-1)$, $f(2m) = G(2m-1) - G(2m)$ and $f(m)$, and evaluate the bounds (2.1a), (2.2a), (2.2e) and (2.2f). In the cases studied, (2.2e) and (2.2f) are superior to (2.2a). In certain applications, as illustrated in Section 3, one has a general formula for $G(3m)$, and can use the better bounds (2.1b) and (2.2d). In certain cases the bounds (2.2e) and (2.2f) can be improved by replacing $B_{1,2}$ in (2.2) with an iterated value of $B_{j,2}$, given in Theorem 1, and then replacing the terms $G(3m)$ and $G(3m-1)$ with their lower bounds $[G(2m)]^2$ and $G(2m)G(2m-1)$, respectively.

Recently Hoover (1988) has derived a sequence of decreasing (increasing) Bonferroni-type upper (lower) bounds for the probability of a finite union (intersection) of events. Using these results gives the following lower bound for $G(T)$.

Theorem 2. *For $m < t < T$,*

$$G(T) \geq (T-t+1)G(t) - (T-t)G(t-1); \tag{2.22}$$

*and if $T \geq 2m$,*

$$G(T) \geq G(2m) - (T-2m)f(2m). \tag{2.23}$$

Proof. From Hoover (1988), Theorem 1, it follows that for a sequence of stationary events $E_1, \dots, E_N$,

$$\begin{aligned} P\left(\bigcup_{i=1}^{N} E_i\right) &\leq NP(E_1) - (N-1)P(E_1 \cap E_2) \\ &\quad - \sum_{j=2}^{n-1} (N-j)P\left\{E_1 \cap \left(\bigcap_{i=1}^{j-1} E_{i+1}^c\right) \cap E_{j+1}\right\}, \end{aligned} \tag{2.24}$$

where $1 < n < N$, and $\sum_{j=k}^{l} a_j \equiv 0$ for $l < k$. In our case, (2.24) leads to the inequality

$$\begin{aligned} G(T) &\geq 1 - (T-m+1)P(E_1^c) - (T-m)P(E_1^c \cap E_2^c) \\ &\quad - \sum_{j=2}^{n-1} (T-m+1-j)P\left\{E_1^c \cap \left(\bigcap_{i=1}^{j-1} E_{i+1}\right) \cap E_{j+1}^c\right\}, \end{aligned} \tag{2.25}$$

where $E_j = \sum_{i=j}^{j+m-1} x_i < k$. Let $t = n + m - 1$, then for $m < t < T$,

$$G(T) \geq 1 - (T - m + 1)P(E_1^c) - (T - m)P(E_1^c \cap E_2^c) \tag{2.26}$$
$$- \sum_{j=2}^{t-m} (T - m + 1 - j)P\left\{E_1^c \cap \left(\bigcap_{i=1}^{j-1} E_{i+1}\right) \cap E_{j+1}^c\right\}.$$

For $m < t < T$, (2.26) reduces to (2.22). It is easy to verify that for $T \geq 2m$, $G(T) \geq G(2m) - (T - 2m)f(2m)$. This concludes the proof of Theorem 2. □

REMARKS. For $t = m$ the Hoover bound reduces to the Bonferroni bound and for $t = m + 1$, it reduces to the Hunter (1976) bound. Note that for large $T(T > 2m + G(2m)/f(2m))$, the right-hand side of inequality (2.23) is negative; this is a deficiency of the lower bound (2.23) relative to (2.2). Section 3 gives a comparison of the Hoover bounds with (2.2). In all our computations the lower bound (2.2) was superior to the Hoover bound (2.23). We illustrate the comparison in Table 1.

For applications where one can evaluate $G(2m)$ and $G(3m)$, Naus (1982) showed that the approximation

$$G(T) \doteq G(2m)[G(3m)/G(2m)]^{T/m-2} \quad \text{for } T \geq 3m, \tag{2.27}$$

is highly accurate. In cases where one can only find $G(t)$ for $t \leq t_1 < 3m$, Saperstein (1976), Samuel-Cahn (1983) and Glaz and Johnson (1984) suggest approximations of the form

$$G(T) \doteq G(t_1 - 1)[G(t_1)/G(t_1 - 1)]^{T-t_1+1} \tag{2.28}$$

for $T \geq t_1$. For an appropriate value of $t_1$, approximation (2.28) falls between the lower and upper bounds in Theorems 1 and 2.

In certain applications the researcher is particularly interested in $E(\tau)$, the expected waiting time until the sum of the $X$'s in the scanning window of $m$ consecutive trials is at least equal to $k$. For example, in quality control one looks at the waiting time until there is a run or an almost perfect run of observations above the mean [Mosteller (1941)]. In acceptance sampling, plans are based on the waiting time until there are $k$ unacceptable lots in $n$ consecutive batches [Anscombe, Godwin and Plackett (1947)]. Compute

$$E(\tau) = \sum_{t=0}^{\infty} G(t) = \sum_{t=0}^{2m-1} G(t) + \sum_{t=2m}^{\infty} G(t). \tag{2.29}$$

Substituting the right-hand side of the bounds in (2.1) and (2.2) for $i = 2$ into the last sum on the right-hand side of (2.29) gives

$$E(\tau) \leq \sum_{t=0}^{2m-1} G(t) + \left[G(2m)/A_{j,1}\right] \tag{2.30}$$

and

$$(2.31)\qquad E(\tau) \geq \sum_{t=0}^{2m-1} G(t) + \left[G(2m)(1 + B_{j,2})/B_{j,2}\right]$$

for any $j$, where $A_{j,1}$ and $B_{j,2}$ are defined in Theorem 1.

**3. The longest matching subsequence allowing mismatches.** Let $(Y_1, Y_2, \dots, Y_T)$ and $(Z_1, Z_2, \dots, Z_T)$ be two amino acid sequences of an $r$-letter alphabet. The $Y$ and $Z$ sequences are said to match in position $i$ iff $Y_i = Z_i$. Let $X_i = 1$ if $Y_i = Z_i$, and $X_i = 0$ otherwise. Then $X_1, X_2, \dots$ is a sequence of 0's and 1's, where a long run of 1's represents a long match between the $Y$ and $Z$ sequences. The simplest null model is where a match in position $i$ is independent of matches in other positions, and $P(Y_i = Z_i) = P(X_i = 1) = p$ for $i = 1, 2, \dots, T$. Note that this case does not require all letters in the alphabet to have the same probability of occurrence, nor do the probabilities of a given letter have to be the same for the two sequences.

The distribution of the length of the longest run of 1's in a Bernoulli sequence is well known, and both exact and asymptotic results are available. Molecular biologists not only seek to gauge the significance of long perfect matches between amino acid sequences, but also that of almost perfect matches. Allowing up to $c$ mismatches means there can be up to $c$ 0's among the 1's (in the scanning interval). Let $V_c$ denote the length of the longest subsequence of 0's and 1's containing $c$ or fewer 0's. The event $V_c \geq m$ is equivalent to the event that there exists a subsequence of length $m$ that contains at least $m - c$ 1's. Let $k = m - c$. The probability of the occurrence of this generalized run in $T$ trials is $1 - G(T)$. For this case, $G(T)$ is given exactly [Naus (1974)], asymptotically [Gordon, Schilling and Waterman (1986)] and by a highly accurate approximation [Naus (1982)].

The microbiologist's intuitive focusing on large near matches makes sense in a situation where they seek to distinguish between the following hypotheses.

Null hypothesis. $P(X_i = 1) = p$ for $i = 1, \dots, T$.

Alternative hypothesis. There exists some trial $t$ such that for $i = t, t + 1, \dots, t + m - 1$, $P(X_i = 1) = p^*$; for $i = 1, 2, \dots, t - 1$ and for $i = t + m, \dots, T$, $P(X_i = 1) = p$; and $p^* > p$.

The generalized likelihood ratio test rejects the above null hypothesis in favor of the alternative hypothesis whenever $N_{m,T}$, the maximum number of 1's in any $m$ consecutive trials, is large. This follows because the likelihood ratio can be written as proportional to

$$(p^*(1 - p)/p(1 - p^*))^{Y_{t-m+1,t}}$$

and the generalized likelihood ratio rejects whenever $\max_{m \leq t \leq T}\{(\text{l.r.})\} > C$.

Since $p^* > p$ this is equivalent to rejecting whenever $\max_{m \le t \le T}\{Y_{t-m+1,t}\} \ge C^*$.

Recall from Section 2 that $G(T)$ gives the distribution of $N_{m,T}$, as well as yielding that of $\tau$. To get the bounds (2.1) and (2.2) for $G(T)$, we need at least $G(2m)$ and $f(2m)$. The following theorem gives simple expressions for the present application in terms of binomial probabilities.

THEOREM 3. *Let* $X_1, X_2, \dots$ *be* i.i.d. *random variates with* $P(X_i = 1) = p$; $P(X_i = 0) = 1 - p = q$. *Let* $b_k = b(k; m, p) = \binom{m}{k} p^k q^{m-k}$; $b_{k-1} = b(k-1; m-1, p)$ *and* $F_b(r; s, p) = \sum_{i=0}^{r} b(i; s, p)$. *For* $k > 2$,

$$G_{k,m}(2m) = F_b^2(k-1; m, p) - (k-1) b_k F_b(k-2; m, p) + mpb_k F_b(k-3; m-1, p) \tag{3.1}$$

*and*

$$f(2m) = (p/k) b_{k-1}[kqb_{k-1} + (k - mp) F_b(k-2; m-1, p)]. \tag{3.2}$$

PROOF. Equation (3.1) is equation (4.2) in Naus (1982). [Equations (4.3)–(4.7) in that paper give $G(3m) \equiv Q_3'$.] To prove (3.2), write

$$f(2m) \equiv P\{N_{m,2m-1} < k \mid X_m = 0,\ X_{2m} = 1,\ Y_{m+1,2m-1} = k-1\} \times qpb_{k-1}. \tag{3.3}$$

To evaluate the conditional probability on the right-hand side of (3.3), condition further on $Y_{1,m-1} = i$. Apply Corollary 1, equation (3.7) in Naus (1974) letting $(k, m, n_1, n_2)$ there be $(k, m-1, i, k-1)$ here (the key idea being that when $x_m = 0$, we can find the distribution of $N_{m,2m-1}$ by looking at the corresponding problem of scanning $2m - 2$ trials with a scanning window of $m - 1$ trials). We then average over the binomial distribution of $Y_{1,m-1}$ to find (3.2). Note that

$$f(2m) = G_{k,m}(2m-1) - G_{k,m}(2m) \tag{3.4}$$

and conditioning on $X_m$,

$$G_{k,m}(2m-1) = qG_{k,m-1}(2m-2) + pG_{k-1,m-1}(2m-2); \tag{3.5}$$

one can use (3.1) and (3.5) to evaluate $f(2m)$ through (3.4). Use of (3.2) is simpler. □

EXAMPLE 1. $k = 8$, $m = 10$, $p = 0.5$. From (3.1), $G_{8,10}(20) = 0.802806854$. From (3.2), $f_{8,10}(20) = 0.01323509$. From Naus (1982), (4.3)–(4.7), $G_{8,10}(30) = 0.682788968$.

Now suppose we seek to estimate (and bound) the probabilities $G_{8,10}(T)$ for $T > 30$. From Theorem 1, $A_{1,1} = f(20) = 0.01323509$; $A_{j,1} = A_{1,1}(1 - A_{j-1,1})^{-9}$, which after 10 iterations stabilizes to $A_{10,1} = A_{100,1} = 0.0151898841$. One can use any $A_{i,1}$ but the largest value gives the best bound. We use the approximation based on $G(3m)$. From the bounds (2.1b), iterated version, (2.2d) and approximations (2.27) and (2.28), we find the

TABLE 1
*Bounds and approximations for $G(T)$ for $k = 8$, $m = 10$, $p = 0.5$*

| $T$ | Lower bound | | Approximation | | Upper bound (2.1b) |
|---|---|---|---|---|---|
| | **Hoover (2.23)** | **(2.2d)** | **(2.28)** | **(2.27)** | |
| 40 | 0.538 | 0.564 | 0.580 | 0.581 | 0.586 |
| 50 | 0.406 | 0.465 | 0.492 | 0.494 | 0.503 |
| 60 | 0.273 | 0.384 | 0.418 | 0.420 | 0.431 |
| 70 | 0.141 | 0.317 | 0.355 | 0.357 | 0.370 |
| 80 | 0.009 | 0.261 | 0.301 | 0.304 | 0.318 |

values in Table 1. Table 1 also compares the lower bound (2.2d) with the Hoover bound (2.23).

**4. The largest possible charge in a scanning interval.** For the charge alphabet classification of amino acids, there are three letters (acidic, neutral and basic) with three possible charges $X_i = -1, 0$ or $+1$. Let $P(X_i = -1) = P_1$, $P(X_i = 0) = P_2$ and $P(X_i = +1) = P_3 = 1 - P_1 - P_2$. The variable $Y_{t-m+1,t}$ represents the combined net charge in the $m$ trials ending at trial $t$. $N_{m,T}$ is the largest net charge within any $m$ consecutive trials, anywhere in a sequence of $T$ trials (letters). Microbiologists are interested in determining whether large net charges within a window of length $m$ in a given sequence are unusual, and seek the null distribution of $N_{m,T}$. See, for example, Karlin, Blaisdell, Mocarski and Brendel (1989), page 167, and Brendel and Karlin (1989).

Here the $X_i$'s can take negative values. Let $X_i^* = X_i + 1$, and let $N_{m,T}^*$ be the corresponding quantity in the sequence of $X^*$'s. Since $N_{m,T}^* = N_{m,T} + m$,

$$G_{k,m}^*(T) = P(N_{m,T}^* < k) = P(N_{m,T} < k - m) = G_{k-m,m}(T). \tag{4.1}$$

Thus we can apply Theorem 1 for $c = 2$. To apply Theorem 1 requires that we evaluate $f(2m)$ and $G(2m)$. The combinatorial argument used to evaluate $f(\ )$ and $G(\ )$ for the case in Section 3 breaks down because the counting method [based on either the reflection principle or the Karlin–McGregor theorem, see Naus (1974)] does not permit jumping over states. We use an alternative computational procedure due to Saperstein (1976). For special cases we can use a direct combinatorial procedure to evaluate $G(2m)$, $G(3m)$ and $f(2m)$. We first illustrate this direct procedure.

EXAMPLE 2. $m = 3$, $P_1 = P_2 = P_3$. To find $G_{k,3}(6)$, note that each of six trials can take any of three possible equally likely values. There are $3^6 = 729$ equally likely sampling points, and a simple counting program will find for each $k$ how many of the points lead to $N_{3,6} < k$. The ratio of this number of points to 729 gives $G(2m) = G_{k,3}(6)$. A similar direct counting approach finds $G(2m - 1)$ (there are $3^5 = 243$ arrangements to check) and $G(3m)$ (there are $3^9 = 19{,}683$ arrangements to check). Note that to find $G(T)$ by this procedure

TABLE 2
$G_{k,m}(T)$ *for* $m = 3$; $T = 2m - 1, 2m, 3m - 1, 3m, 4m$;
$P_1 = P_2 = P_3$ for charged alphabet

| $k$ | $3^5G(2m-1)$ | $3^6G(2m)$ | $3^8G(3m-1)$ | $3^9G(3m)$ | $3^{12}G(4m)$ |
|---|---|---|---|---|---|
| 3 | 222 | 648 | 5,524 | 16,128 | 401,392 |
| 2 | 172 | 466 | 3,444 | 9,358 | 187,826 |
| 1 | 99 | 233 | 1,316 | 3,124 | 41,842 |
| 0 | 40 | 76 | 287 | 556 | 4,057 |
| −1 | 9 | 13 | 28 | 41 | 129 |
| −2 | 1 | 1 | 1 | 1 | 1 |

for $T$ large is not practical as there are $3^T$ arrangements to check. Even for $G(4m) = G(12)$ there are 531,441 arrangements to check. However, to bound and approximate $G(T)$ for large $T$ only requires $G(2m-1)$ and $G(2m)$, and for a better approximation and bounds, $G(3m)$ and $G(3m-1)$.

Table 2 gives the entire distribution of these quantities [and also $G(4m)$ for checking purposes] for this example.For example, for $k = 2$, $m = 3$,

$$G(2m) = G_{2,3}(6) = 466/3^6 = 0.639231824,$$

$$G(2m-1) = 0.70781893, \qquad G(3m) = 0.475435655.$$

$A_{1,1} \equiv f(2m) = G(2m-1) - G(2m) = 0.0685871061$; iterating, $A_{10,1} = A_{20,1} = 0.0812555472$. $B_{1,2} < B_{2,2}$ so use $B_{1,2} = 0.130662$. Use bounds (2.1b) and (2.2c) to find $0.329 \le G_{2,3}(12) \le 0.369$. From Table 2, the exact value is $G_{2,3}(12) = 187{,}826/3^{12} = 0.3534$. From approximation (2.27), $G_{2,3}(12) \doteq G(2m)[G(3m)/G(2m)]^2 = 0.3536$. Table 3 compares approximations (2.27) and (2.28) with simulations (based on 100,000 trials), and with bounds (2.1b) iterated, and (2.2c) for larger $T$. Approximations (2.28a) and (2.28b) are approximation (2.28) with $t_1 = 2m, 3m$.

Glaz and Naus (1989) extend Table 2 for the cases $m = 4, 5$. The same approach can be used for the cases $m = 6, 7$. For larger $m$, or unequal $P_i$, more efficient computational procedures are needed, and these are provided by a method of Saperstein (1976), who derives a recursion formula for evaluating $G_{k,m}(i)$ for $m + 1 \le i \le 2m$.

TABLE 3
$G_{2,3}(T)$ *for charged alphabet,* $P_1 = P_2 = P_3$

| | Lower | Approximation | | | | Upper bound |
|---|---|---|---|---|---|---|
| $T$ | bound (2.2c) | (2.28a) | (2.28b) | (2.27) | Simulation | (2.1b) |
| 18 | 0.157 | 0.188 | 0.195 | 0.196 | 0.193 | 0.222 |
| 21 | 0.109 | 0.139 | 0.145 | 0.145 | 0.143 | 0.172 |
| 24 | 0.075 | 0.102 | 0.108 | 0.108 | 0.108 | 0.133 |
| 27 | 0.052 | 0.075 | 0.080 | 0.080 | 0.081 | 0.103 |
| 30 | 0.036 | 0.055 | 0.059 | 0.060 | 0.061 | 0.080 |

TABLE 4
*Probability of a positive net charge of k or more*
$m = 30, T = 968, P_1 = 114/968, P_2 = 754/968, P_3 = 1 - P_1 - P_2$

| $k$ | **Lower bound (2.1a)** | **Approximation (2.28)** | **Simulation (10,000 trials)** | **Upper bound (2.2e)** |
|---|---|---|---|---|
| 8 | 0.1777 | 0.1779 | 0.1771 | 0.1787 |
| 9 | 0.0584 | 0.0584 | 0.0578 | 0.0585 |
| 10 | 0.0161 | 0.0161 | 0.0160 | 0.0161 |
| 11 | 0.003828 | 0.003828 | 0.0039 | 0.003828 |

Saperstein's iterative procedure yields $G(2m - 1)$, $G(2m)$ and thereby $f(2m)$ and thus enables us to bound $G(T)$ via (2.1) and (2.2). We follow this approach to use (2.2e) and the iterated form of (2.1a) to derive lower and upper bounds for the null tail probabilities $P(N_{m,T} \geq k) = 1 - G_{k,m}(T)$. Glaz and Naus (1989) table bounds for selected values of $T$, $m$ and $k$.

Karlin, Blaisdell, Mocarski and Brendel (1989), page 167, give an example of a sequence of $T = 968$ $X$'s (residues of the adenovirus type 2 hexon protein), and fit a model of independent $X_i$'s, with $P(X_i = -1) = 114/968$, $P(X_i = 0) = 754/968$ and $P(X_i = +1) = 100/968$. They take a window length of $m = 30$. They estimate that a *net* positive charge of at least 9.7 anywhere within a window length of 30 is required for significance at the 0.01 level. Table 4 gives the iterated bounds (2.1a) and (2.2e), and approximation (2.28). For this example, the bounds and approximation give the probability of a net positive charge of 10 or more as 0.0161. Note that the bounds in Table 4 are very tight for small (and even moderate) values of $1 - G(T)$.

**Acknowledgment.** The authors would like to thank Samuel Karlin for introducing them to the charged alphabet problem and many other molecular biology problems.

## REFERENCES

ANSCOMBE, F. J., GODWIN, H. J. and PLACKETT, R. L. (1947). Methods of deferred sentencing in testing. *J. Roy. Statist. Soc. Ser. B* **7** 198–217.

ARRATIA, R., GORDON, L. and WATERMAN, M. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971–993.

BRENDEL, V. and KARLIN, S. (1989). Association of charge clusters with functional domains of cellular transcription factors. *Proc. Nat. Acad. Sci. U.S.A.* **86** 5698–5702.

ERDÖS, P. and RÉVÉSZ, P. (1975). On the length of the longest head-run. *Topics in Information Theory. Colloq. Math. Soc. J. Bolyai* **16** 219–228. Keszthely, Hungary.

ESARY, J. D., PROSCHAN, F. and WALKUP, D. (1967). Association of random variables with applications. *Ann. Math. Statist.* **38** 1466–1474.

GLAZ, J. and JOHNSON, B. McK. (1984). Probability inequalities for multivariate distributions with dependence structures. *J. Amer. Statist. Assoc.* **79** 436–441.

GLAZ, J. and NAUS, J. (1989). Tight bounds and approximations for scan statistic probabilities for discrete data. Technical Report, Dept. Statistics, Rutgers Univ., New Brunswick, N.J.

GORDON, L., SCHILLING, M. F. and WATERMAN, M. S. (1986). An extreme value theory for long head runs. *Probab. Theory Related Fields* **72** 279–288.

HOOVER, D. R. (1988). Component complement addition upper bounds—an improved inclusion–exclusion method. Technical Report, Dept. Statistics, Univ. South Carolina.

HUNTER, D. (1976). An upper bound for the probability of a union. *J. Appl. Probab.* **13** 597–603.

JANSON, S. (1984). Bounds on the distributions of extremal values of a scanning process. *Stochastic Process. Appl.* **18** 313–328.

KARLIN, S., BLAISDELL, B., MOCARSKI, E. and BRENDEL, V. (1989). A method to identify distinctive charge configurations in protein sequences, with application to human Herpesvirus polypeptides. *J. Molecular Biol.* **205** 165–177.

KARLIN, S. and GHANDOUR, G. (1985). Multiple-alphabet amino acid sequence comparison of the immunoglobulin $k$-chain constant domain. *Proc. Nat. Acad. Sci. U.S.A.* **82** 8597–8601.

KARLIN, S. and OST, F. (1988). Maximal length of common words among random letter sequences. *Ann. Probab.* **16** 535–563.

MOSTELLER, F. (1941). Note on an application of runs to quality control charts. *Ann. Math. Statist.* **12** 228–232.

NAUS, J. (1974). Probabilities for a generalized birthday problem. *J. Amer. Statist. Assoc.* **69** 810–815.

NAUS, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77** 177–183.

SAMUEL-CAHN, E. (1983). Simple approximations to the expected waiting time for a cluster of any given size for point processes. *Adv. in Appl. Probab.* **15** 21–38.

SAPERSTEIN, B. (1976). The analysis of attribute moving averages: MIL-STD-105D reduced inspection plans. Paper presented at Sixth Conf. Stochastic Processes and Applications, Tel Aviv.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CONNECTICUT
STORRS, CONNECTICUT 06269-3120

DEPARTMENT OF STATISTICS
HILL CENTER FOR MATHEMATICAL SCIENCE
BUSCH CAMPUS
RUTGERS–THE STATE UNIVERSITY OF NEW JERSEY
NEW BRUNSWICK, NEW JERSEY 08903