Ann Inst Stat Math (2017) 69:497–512
DOI 10.1007/s10463-015-0551-8

The *m*th longest runs of multivariate random sequences

Yong Kong[1]

Received: 14 April 2015 / Revised: 19 November 2015 / Published online: 1 February 2016

Abstract The distributions of the *m*th longest runs of multivariate random sequences are considered. For random sequences made up of k kinds of letters, the lengths of the runs are sorted in two ways to give two definitions of run length ordering. In one definition, the lengths of the runs are sorted separately for each letter type. In the second definition, the lengths of all the runs are sorted together. Exact formulas are developed for the distributions of the *m*th longest runs for both definitions. The derivations are based on a two-step method that is applicable to various other runs-related distributions, such as joint distributions of several letter types and multiple run lengths of a single letter type.

Keywords Generating function · Combinatorial identities · Randomness test · Distribution-free statistical test · Runs length test · Biological sequence analysis

1 Introduction

Run statistics have been used in various disciplines to test the nonrandomness in sequences (Balakrishnan and Koutras 2002; Godbole and Papastavridis 1994; Knuth 1997). For the related topic of scan statistics, see for example (Glaz et al. 2001). The research in this area has been revived recently because of the applications in biological related problems, such as sequence analysis and genetic analysis.

One of the commonly used runs tests is the longest run test. An unusual long consecutive appearance of one type of letter usually indicates the nonrandom nature

✉ Yong Kong
yong.kong@yale.edu

[1] Department of Molecular Biophysics and Biochemistry, W.M. Keck Foundation Biotechnology Resource Laboratory, Yale University, 333 Cedar Street, New Haven, CT 06510, USA

of the process that generates the sequence. These long consecutive appearances (runs), however, are usually obscured by noises or other processes so that they do not reveal themselves to the observers, making the application of the longest run test difficult or impossible. For example, when we consider biological sequences such as DNA sequences, the longest run of one particular letter type might have been broken into several shorter ones due to either biological mutations or errors that occurred in the process of reading out these sequences. For such sequences, it would be difficult to have a single run of statistically significant length. For example, for a binary system of 17 total elements, with 10 elements of the first letter type and 7 elements of the second letter type, the longest run of the first letter type needs a length of 7 to achieve statistically significance with a cutoff of $\alpha = 0.05$: $P(l_0 \geq 7) = 0.049$ [see Eq. (18)]. On the other hand, if we use the second longest run, it requires only $l_1 \geq 4$ to achieve statistically significance for the same significance level: $P(l_1 \geq 4) = 0.041$ [see again Eq. (18)]. One of the goals of this paper is to develop explicit, easily calculated formulas for the mth longest runs of multivariate random sequences, where m is an arbitrary nonnegative integer. As shown in Fig. 1 and Table 1, as m becomes bigger, the distributions become narrower, so it might become easier to tell whether the observed statistic comes from one distribution or the other.

The distributions of the longest runs and other runs-related distributions have been studied by previous researchers for independent trials and Markov-dependent trials (Burr and Cane 1961; Philippou and Makri 1985, 1986; Schilling 1990; Koutras and Papastavridis 1993; Koutras and Alexandrou 1995; Lou 1996; Muselli 1996; Fu et al. 2003; Eryilmaz 2006; Makri et al. 2007). Based on the results of Mood (1940), a distribution is derived which gives probability of at least one run of a given length

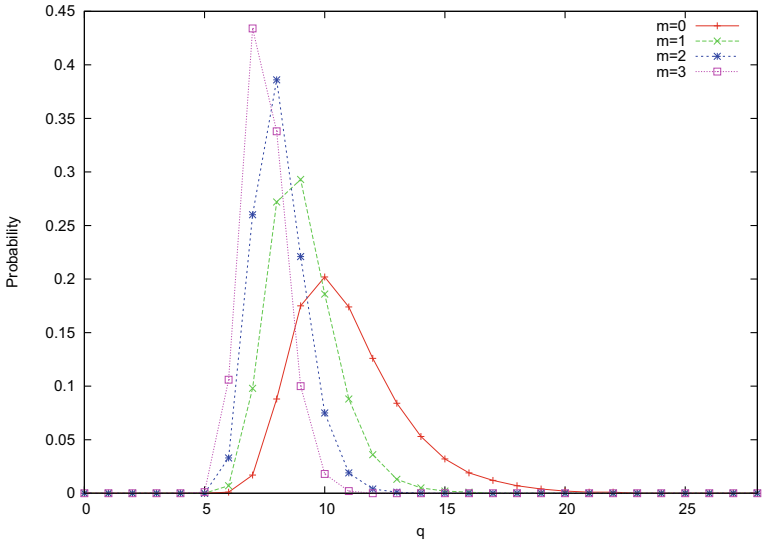

Fig. 1 Probability mass distribution of the mth longest runs of the whole system, for $\mathbf{n} = (n_1, n_2) = (200, 300)$, $m = 0$ to 3. The probability is calculated by $W(\mathbf{n}; q; m)$ in Corollary 5, divided by $\binom{n_1+n_2}{n_1}$

Table 1 Average, the second moment, and variance of the distribution in Fig. 1

m	$E(X)$	$E(X^2)$	σ^2
0	10.997	126.502	5.562
1	9.072	84.309	2.006
2	8.121	67.115	1.165
3	7.494	56.966	0.809

or greater for the special case of binary systems where each kind of object has the same number of elements ($n_1 = n_2$) (Mosteller 1941). The formulas involve double summations. These results are simplified later (Olmstead 1958; Bradley 1968, pp. 255–259), again for binary systems. A recursion-based algorithm is given for the distribution of the longest run of any letter type of multiple object systems (Schuster 1996). Morris et al. (1993) had similar objectives to ours: they obtained exact, explicit formulas for multiple objects containing any minimum collection of specified lengths. Their formulas were obtained by using convoluted combinatorial arguments. Different from their approach, we will use a simple method that can derive various distribution in a unified, almost mechanical way. This is the second major goal of this paper: to introduce a systematic method that can treat various distributions in a unified approach.

Earlier studies of run statistics usually used ingenious *ad hoc* combinatorial methods, which sometimes became very tedious. In Kong (2006) a systematic method to study various run statistics in multiple letter systems was developed. Two of the commonly used run tests, the total number of runs test and the longest run test, were investigated in detail by using the general method. The method was later applied to other commonly used runs tests (Kong 2015a, b, c). In this paper, we extend the method to investigate the distributions of the mth longest runs for multi-letter systems. Two different definitions of the run length order will be studied. For the first definition, the lengths of the runs are sorted separately for each letter type. The formula developed in Kong (2006, Theorem 10) is a special case for this definition, with $m_i = 0$ for each letter type. For the second definition, the lengths of all the runs from all letter types are sorted together. The distributions of both definitions can be considered as special cases of the general two-step method discussed in Sect. 2.

In the general setting the method involves two steps. In the first step, we only need to consider the arrangements of a single letter type that meet the restrictions we impose on that letter type, such as the lengths of the runs or the number of runs, without worrying about the complicated combinations with other letter types. It simplifies the enumeration tasks considerably when only one letter type is considered. The number of such arrangements of a single letter type, say the ith type, with n_i elements and r_i number of runs, is specified in Eq. (4) by $U(n_i, r_i; X_i)$, where X_i is a place-holder for other restrictions in addition of n_i and r_i. In the second step, the quantities $U(n_i, r_i; X_i)$, which are for individual letter types, are then combined together by the function $F(\mathbf{r})$ [as shown in Eq. (1)] to get the distribution of the whole system. The function $F(\mathbf{r})$ gives the number of configurations to arrange in a line r_1 blocks of the first letter type, r_2 blocks of the second letter type, etc., without the blocks of the same letter type touching each other. The explicit expression of $F(\mathbf{r})$ [Eq. (1)] makes it possible to obtain explicit expressions for various kind of run-related distributions. These expressions

can often be simplified by manipulating binomial and multinomial coefficients, which can be done mechanically using the Wilf–Zeilberger method (Petkovšek et al. 1996).

The major results of this article are Theorems 2 and 3. These theorems are for the mth longest runs in systems with arbitrary number of letter types under the two different definitions of the run length ordering.

Both results are obtained by using the simple yet quite general consideration discussed in Theorem 1. It is interesting to note that, when compared to the formulas for the special case of the longest runs ($m = 0$), the only difference is that the general formulas for the arbitrary mth longest runs contain an extra binomial factor $(-1)^m \binom{j-1}{m}$ where j is a summation variable [see Eqs. (13), (17), (18), and (23)].

There are many definitions of runs in the literature. In this article we use the classical definition of Mood (1940), which asserts that consecutive runs of one letter type must be separated by other letter types. This is also the definition we used in the previous work (Kong 2006).

Two kinds of models are usually used when the distributions of runs are studied. If the numbers of elements for each letter type are fixed, the models are known as conditional models. If the elements are not fixed but chosen from a multinomial population, the models are called unconditional. For both of these models, exact finite distributions and asymptotic distributions have been investigated in the past. The results presented in this article are exact distributions conditioned on the compositions of systems under study, i.e., the numbers of each letter type are fixed. These exact distributions are particular useful for relatively short sequences and other situations where asymptotic results cannot be applied. Once the conditional distribution is obtained, it is usually easy to get unconditional distribution by the multinomial theorem.

Throughout the article we reserve the letter k for the number of letter types in the system, and use n_i as the number of elements of the ith letter type. The total number of elements of the system is $n = \sum_{i=1}^{k} n_i$. The letter m (with index if necessary) is used to indicate the run order. For the first definition of the run ordering, $m_i = 0$ is used to index the longest run of the ith letter type, and $m_i = 1$ is the index of the second longest run, etc. For the second definition, since the letters are pooled together when the run lengths are ordered, the subscript on m is no longer needed. In this case, $m = 0$ indicates the longest run of the whole system, and $m = 1$ is the second longest run, etc.

We denote by the bold letters the tuples with k elements, such as $\mathbf{n} = (n_1, n_2, \ldots, n_k)$, $\mathbf{r} = (r_1, r_2, \ldots, r_k)$, $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, and similarly for other symbols. We use $\binom{n}{m}$ for the binomial coefficient (n choose m), and $\left[\begin{smallmatrix} p_1+\cdots+p_k \\ p_1,\ldots,p_k \end{smallmatrix}\right] = \left[\begin{smallmatrix} p \\ p_1,\ldots,p_k \end{smallmatrix}\right] = \left[\begin{smallmatrix} p \\ \mathbf{p} \end{smallmatrix}\right] = p!/(p_1! \ldots p_k!)$ as the multinomial coefficient, with $p = \sum_{i=1}^{k} p_i$. When there is no ambiguity, the k nesting summations will be abbreviated as a single sum for clarity, for example, $\sum_{p_1=1}^{r_1} \cdots \sum_{p_k=1}^{r_k} f(\mathbf{p})$ will be written as $\sum_{p_i=1}^{r_i} f(\mathbf{p})$. The coefficient of x^m of a polynomial $f(x)$ is denoted as $[x^m]f(x)$.

The paper is organized as follows. In Sect. 2, we describe the two-step method outlined above in a general setting. Then in Sects. 3 and 4, the method is applied to obtain the distributions of the mth longest runs, under two different definitions of the run length ordering.

2 A general two-step method for run-related distributions

As discussed in Sect. 1, in the first step we only consider the enumeration of one letter type when its elements are considered alone. The enumeration of one letter type is considerably easier than when all the letter types are considered together. Let assume that for the ith letter type with n_i elements arranged into r_i runs, we impose one or more additional conditions, collectively denoted as X_i. Denote $U(n_i, r_i; X_i)$ as the number of arrangements of n_i elements of the ith letter type in exactly r_i runs with the additional restriction X_i imposed. Various methods can be used to obtain $U(n_i, r_i; X_i)$, with generating function as one of the most powerful and versatile methods [see Eq. (11) for one of such applications].

After obtaining $U(n_i, r_i; X_i)$, we need to put them together to form a k-letter type system. To do this we will use the function $F(\mathbf{r})$, which is the number of ways to arrange in a line r_1 runs of the first letter type, r_2 runs of the second letter type, etc., without two adjacent runs being of the same kind. The explicit expression of function $F(\mathbf{r})$ is given by Kong (2006):

Lemma 1 *The function $F(\mathbf{r})$ is given by*

$$F(\mathbf{r}) = \sum_{\substack{1\le p_i\le r_i\\ 1\le i\le k}} (-1)^{\sum_i (r_i-p_i)} \binom{r_1-1}{p_1-1}\cdots\binom{r_k-1}{p_k-1}\begin{bmatrix} p_1+\cdots+p_k \\ p_1,\ldots,p_k\end{bmatrix}. \tag{1}$$

When $k = 2$, $F(r_1, r_2)$ can be simplified from Eq. (1) to the following trivial expression,

$$F(r_1, r_2) = \binom{2}{r_1 - r_2 + 1} = \begin{cases} 2 & \text{if } r_1 = r_2,\\ 1 & \text{if } |r_1 - r_2| = 1,\\ 0 & \text{otherwise,}\end{cases}$$

which is obvious from the meaning of function $F(\mathbf{r})$.

Proposition 1 *For a system with k letter types, the total number of configurations is given by*

$$R(\mathbf{n}) = \sum_{r_i=1}^{n_i} F(\mathbf{r}) \prod_{i=1}^{k} U(n_i, r_i; X_i), \tag{2}$$

where $U(n_i, r_i; X_i)$ is the number of arrangements of n_i elements of the ith letter type in exactly r_i runs with restrictions X_i imposed.

Often the time we do not want to impose the restrictions on all of the k letter types. For example, we might only be interested in the length of runs of the first letter type, and put no restrictions on the other $k - 1$ letter types. Or we are only interested in the length of runs of the first and the second letter types to obtain their joint distributions. In general, suppose we only impose certain restrictions on some of the k letter types,

which are indexed by $S = \{i_1, i_2, \ldots, \}$. Then the number of configurations of the system $R(\mathbf{n}; S)$ can be written as

$$R(\mathbf{n}; S) = \sum_{r_i} F(\mathbf{r}) \prod_{i \in S} U(n_i, r_i; X_i) \prod_{i \notin S} V(n_i, r_i), \tag{3}$$

where $V(n_i, r_i)$ is the number of arrangements of the n_i elements of the ith letter type in exactly r_i runs without any restrictions.

Theorem 1 *The number of configurations $R(\mathbf{n}; S)$ for a system with k letter types is given by*

$$R(\mathbf{n}; S) = \sum_{r_i, p_i, i \in S} \begin{bmatrix} \sum_{i \in S} p_i + \sum_{i \notin S} n_i \\ p_i, n_i \end{bmatrix} \prod_{i \in S} (-1)^{r_i - p_i} \binom{r_i - 1}{p_i - 1} U(n_i, r_i; X_i), \tag{4}$$

where the set S specifies the subset of letter types on which additional restrictions are imposed.

Proof The expression of $V(n_i, r_i)$ in Eq. (3) for the unrestricted arrangements of exactly r_i runs using n_i elements is given by the well-known formula

$$V(n_i, r_i) = \binom{n_i - 1}{r_i - 1}. \tag{5}$$

A direct interpretation of above expression is to put $r_i - 1$ bars between the $n_i - 1$ spaces formed by the n_i elements to form r_i runs. By using Eqs. (1) and (5) and utilizing the identity

$$\sum_{r=0}^{n} (-1)^r \binom{n}{r} \binom{r}{m} = (-1)^n \delta_{n,m},$$

the sums of r_i in Eq. (3) for $i \notin S$ can be evaluated to $(-1)^{n_i} \delta_{n_i, p_i}$. These δ_{n_i, p_i} in turn filter out the sums of p_i in the explicit expression of $F(\mathbf{r})$ for $i \notin S$ to a single term with $p_i = n_i$, leading to the simplification of $R(\mathbf{n}; S)$ to sums that only involve letter types in S. □

With different assignments of the set S, Theorem 1 can be used to obtain different kinds of distributions, such as joint distributions of two or three letter types, with $S = \{1, 2\}$ and $S = \{1, 2, 3\}$, respectively. Several special cases for this theorem are mentioned here for: (1) $|S| = 0$, (2) $|S| = k$, and (3) $|S| = 1$. If S is empty, then $R(\mathbf{n}; S)$ is simplified to the trivial result $\begin{bmatrix} n \\ n_i \end{bmatrix}$, as it should be:

$$R(\mathbf{n}; S = \varnothing) = \begin{bmatrix} n \\ n_i \end{bmatrix}.$$

If $S = \{1, 2, \ldots, k\}$, i.e., all letter types are subject to restrictions, then

$$R(\mathbf{n}; S = \{1, \ldots, k\}) = \sum_{p_i} (-1)^{p_i} \begin{bmatrix} p \\ p_i \end{bmatrix} \sum_{r_i} \prod_i (-1)^{r_i} \binom{r_i - 1}{p_i - 1} U(n_i, r_i; X_i). \quad (6)$$

If only one letter type has restrictions, say $S = \{1\}$, then

$$\begin{aligned} &R(\mathbf{n}; S = \{1\}) \\ &= \sum_{r_1=1}^{n_1} \sum_{p_1=1}^{r_1} (-1)^{r_1 - p_1} \begin{bmatrix} n - n_1 + p_1 \\ p_1, n_2, \ldots, n_k \end{bmatrix} \binom{r_1 - 1}{p_1 - 1} U(n_1, r_1; X_1) \\ &= \begin{bmatrix} n - n_1 \\ n_2, \ldots, n_k \end{bmatrix} \sum_{r_1=1}^{n_1} \sum_{p_1=1}^{r_1} (-1)^{r_1 - p_1} \binom{n - n_1 + p_1}{p_1} \binom{r_1 - 1}{p_1 - 1} U(n_1, r_1; X_1). \end{aligned}$$

By using the identity

$$\sum_{p=1}^{r} (-1)^p \binom{n + p}{p} \binom{r - 1}{p - 1} = (-1)^r \binom{n + 1}{r},$$

we get

Corollary 1 *For a system with the first letter type restricted while the other letter types are unrestricted, the number of configurations is given by*

$$R(\mathbf{n}; S = \{1\}) = \begin{bmatrix} n - n_1 \\ n_2, \ldots, n_k \end{bmatrix} \sum_{r_1=1}^{n_1} \binom{n - n_1 + 1}{r_1} U(n_1, r_1; X_1). \quad (7)$$

The direct interpretation of Eq. (7) is that the $n - n_1$ elements of the other letter types form $n - n_1 + 1$ intervals in a line (including the two ends). There are $\binom{n-n_1+1}{r_1}$ ways for the elements of the first letter type to choose r_1 out of these $n - n_1 + 1$ intervals to form r_1 runs. The multinomial factor in the front takes care of the number of configurations the elements of the other letter types can form among themselves.

In the following, we will use this two-step method to derive distributions of the mth longest runs under two different definitions.

3 The first definition of the mth longest run: run lengths sorted within each letter type

In this definition, the run lengths are sorted for each letter type separately. For the ith letter, we denote $l_0^{(i)}$ as the length of the longest run of the ith letter type, $l_1^{(i)}$ as the length of the second longest run of the ith letter type, etc. In general, $l_m^{(i)}$ is the length

of the $(m+1)$th longest run of the ith letter type. The lengths of all the runs formed by the ith letter type are ordered as

$$l_0^{(i)} \geq l_1^{(i)} \geq l_2^{(i)} \geq \cdots \geq l_{r_i-1}^{(i)}.$$

In other words, there are at least $m+1$ runs of the ith letter type whose length is longer or equal to $l_m^{(i)}$.

For example, in a $k=4$ system made up of letter types $\{1, 2, 3, 4\}$, if we have the following particular arrangement of the four letter types

$$111 \,|\, 2 \,|\, 111 \,|\, 333 \,|\, 444444 \,|\, 33 \,|\, 11111, \tag{8}$$

then we have $l_0^{(1)} = 5$, $l_1^{(1)} = 3$, $l_2^{(1)} = 3$ for the first letter type 1s, $l_0^{(2)} = 1$ for the second letter type 2s, $l_0^{(3)} = 3$, $l_1^{(3)} = 2$ for the third letter type 3s, and $l_0^{(4)} = 6$ for the fourth letter type 4s. All other $l_m^{(i)} = 0$.

As described in Sect. 2, to use the two-step method to obtain the distribution of the whole system we first focus on a single particular letter type. In the following, if we only deal with one letter type, the index i in $l_m^{(i)}$ will be omitted and we will use l_m for simplicity. Define function $h_m(n, q, r)$ as the number of ways to arrange the elements of a given letter type with n elements in r runs, with the length of $(m+1)$th longest run less than or equal to q, $q \geq 0$, i.e., $l_m \leq q$. In other words, at most m runs can have lengths greater than q. This is a specialization of the generic function $U(n, r; X)$ of Eq. (3), with the parameters m and q jointly act as the restriction parameter X. In the following, we will find an explicit expression for $h_m(n, q, r)$.

By definition it is obvious that

$$h_m(n, q, r) = \sum_{i=0}^{m} \bar{h}_i(n, q, r), \tag{9}$$

where $\bar{h}_i(n, q, r)$ counts for the arrangements which have exactly i runs whose lengths are greater than q. We will first find an explicit expression for $\bar{h}_i(n, q, r)$, then use the above relation to obtain $h_m(n, q, r)$.

Lemma 2 *The number of ways to arrange n elements in r runs with exactly m longest runs of length greater than $q \geq 0$ is given by*

$$\bar{h}_m(n, q, r) = \begin{cases} 0 & q = 0 \text{ and } r \neq m, \\ \binom{n-1}{r-1} & q = 0 \text{ and } r = m, \\ \sum_{j=m}^{\min(r, \lfloor (n-r)/q \rfloor)} (-1)^{m+j} \binom{j}{m} \binom{r}{j} \binom{n-qj-1}{r-1} & \text{otherwise.} \end{cases} \tag{10}$$

Proof To calculate $\bar{h}_m(n,q,r)$, we define generating function $g(x,y,q)$ as

$$g(x,y,q) = (x + \cdots + x^q) + y(x^{q+1} + \cdots) = \frac{x(1-x^q)}{1-x} + y\frac{x^{q+1}}{1-x}. \quad (11)$$

If we expand $g(x,y,q)^r$, then $\bar{h}_m(n,q,r)$ will be the coefficient of $x^n y^m$, since this term counts the number of configurations with exactly m runs whose lengths are greater than q for a total of n elements. We obtain:

$$\begin{aligned}
\bar{h}_m(n,q,r) &= [x^n y^m] g(x,y,q)^r \\
&= [x^n y^m](1-x)^{-r}\left[x(1-x^q) + yx^{q+1}\right]^r \\
&= [x^n](1-x)^{-r}\binom{r}{m}\left[x(1-x^q)\right]^{r-m} x^{m(q+1)} \\
&= [x^n]\binom{r}{m}\sum_l \binom{r+l-1}{r-1}\sum_j (-1)^j \binom{r-m}{j} x^{l+r-m+qj+m(q+1)} \\
&= \binom{r}{m}\sum_j (-1)^j \binom{r-m}{j}\binom{n-(m+j)q-1}{r-1} \\
&= \sum_j (-1)^{j-m}\binom{j}{m}\binom{r}{j}\binom{n-qj-1}{r-1}. \quad (12)
\end{aligned}$$

Equation (12) includes the special case of $q = 0$, which can be checked explicitly. For $q = 0$, we need to put n elements into r runs with exactly m runs whose lengths are greater than zero. Each run, by definition, has a length greater than zero. Hence for $q = 0$, $\bar{h}_m(n,q,r)$ vanishes for all values of m except for $m = r$. This is also reflected in Eq. (11): when $q = 0$, the only term of y in $g(x,y,q)^r$ is y^r. In this case there are $\binom{n-1}{r-1}$ number of ways to arrange n elements into r runs. This can be checked in Eq. (12): the sum has only one nonvanishing term, which is when $j = r = m$, leading to $\binom{n-1}{r-1}$. □

Lemma 3 *The number of ways to arrange n elements in r runs, with the length of $(m+1)$th longest run less than or equal to $q \geq 0$, is given by*

$$h_m(n,q,r) = \begin{cases} 0 & q = 0 \text{ and } r > m, \\ \binom{n-1}{r-1} & q = 0 \text{ and } r \leq m, \\ \sum_{j=0}^{\min(r,\lfloor (n-r)/q \rfloor)} (-1)^{m+j}\binom{j-1}{m}\binom{r}{j}\binom{n-qj-1}{r-1} & \text{otherwise.} \end{cases} \quad (13)$$

Proof From the relation Eq. (9) and expression of $\bar{h}_m(n, q, r)$ in Eq. (10), we have

$$h_m(n,q,r) = \sum_{l=0}^{m} \bar{h}_l(n,q,r) = \sum_j (-1)^j \binom{r}{j} \binom{n-qj-1}{r-1} \sum_{l=0}^{m} (-1)^l \binom{j}{l}$$
$$= (-1)^m \sum_j (-1)^j \binom{r}{j} \binom{n-qj-1}{r-1} \binom{j-1}{m}. \tag{14}$$

By definition, $h_m(n, q, r) = \binom{n-1}{r-1}$ when $m \geq r$, the number of configurations with n elements in r runs. This is reflected in the above expression as $j = 0$ is the only nonvanishing term in the sum when $m \geq r$. As before, some special cases when $q = 0$ should be considered. Apparently when $q = 0$, $h_m(n, q, r) = 0$ if $r > m$. When $q = 0$ and $r \leq m$, Eq. (14) is simplified to $\binom{n-1}{r-1}$. □

With the expression of $h_m(n, q, r)$ in Lemma 3, we can use Eq. (4) in Theorem 1 to get the distribution for the whole system. First, we define two sets of numbers $\mathbf{m} = (m_1, \ldots, m_k)$ and $\mathbf{q} = (q_1, \ldots, q_k)$. Then we denote $N(\mathbf{n}; \mathbf{q}; \mathbf{m})$ as the number of ways to have the $(m_i + 1)$th longest run of the ith letter type equal to or less than $q_i \geq 0$ for all letters: $i = 1, \ldots, k$. In other words, $N(\mathbf{n}; \mathbf{q}; \mathbf{m})$ is the number of ways to arrange the letters so that $\forall i \in \{1, 2, \ldots, k\}$, $l_{m_i}^{(i)} \leq q_i$. The Theorem 10 of Kong (2006) is a special case of $N(\mathbf{n}; \mathbf{q}; \mathbf{m})$ with $\mathbf{m} = (0, \ldots, 0)$, i.e., only the longest run for each letter type is considered there. From Eq. (3) we have

$$N(\mathbf{n}; \mathbf{q}; \mathbf{m}) = \sum_{r_i=1}^{n_i} F(\mathbf{r}) \prod_i h_{m_i}(n_i, q_i, r_i). \tag{15}$$

Equation (15) can be simplified if we use the explicit expression of $F(\mathbf{r})$, as in Eq. (6). If we put $U(n, r, X) = h_m(n, q, r)$ in Eq. (6) and define the last sum in Eq. (6) as

$$H_m(n, q, p) = \sum_r (-1)^r \binom{r-1}{p-1} h_m(n, q, r),$$

then the summation of the running variable r can be carried out and we have

Theorem 2 *The number of ways to arrange the k letter types so that for* $\forall i \in \{1, 2, \ldots, k\}$, $l_{m_i}^{(i)} \leq q_i$ *is given by*

$$N(\mathbf{n}; \mathbf{q}; \mathbf{m}) = \sum_{p_i=1}^{n_i} (-1)^{p_i} \begin{bmatrix} \sum p_i \\ p_i \end{bmatrix} \prod_i H_{m_i}(n_i, q_i, p_i), \tag{16}$$

where

$$H_m(n,q,p) = \begin{cases} (-1)^m \binom{n-1}{p-1}\binom{n-p-1}{m-p} & q=0 \text{ and } n \le m, \\ \binom{n-1}{p-1}\left[(-1)^m\binom{n-p-1}{m-p} - \binom{n-p-1}{n-1}\right] & q=0 \text{ and } n > m, \\ \sum\limits_{j=\lceil (n-p)/(q+1)\rceil}^{\lfloor (n-p)/q\rfloor} (-1)^{n+m+qj+j}\binom{j-1}{m}\binom{n-qj-1}{p-1}\binom{p}{n-qj-j} & \text{otherwise.} \end{cases} \tag{17}$$

The special case of $\mathbf{m} = (0, \ldots, 0)$ has been reported previously (Kong 2006, Theorem 10). Comparing the two expressions we see that the only difference is the extra binomial term $(-1)^m\binom{j-1}{m}$ for the general case of the mth longest runs in Eq. (17).

If we define $L(\mathbf{n}; \mathbf{q}; \mathbf{m})$ as the number of arrangements to have at least one of the k letter types, for example, the ith letter type, to have the length of the (m_i+1)th longest run equal to q_i, i.e., $\exists i \in \{1, 2, \ldots, k\}$, $l_{m_i}^{(i)} = q_i$, then by definition, $L(\mathbf{n}; \mathbf{q}; \mathbf{m}) = N(\mathbf{n}; \mathbf{q}; \mathbf{m}) - N(\mathbf{n}; \mathbf{q} - \mathbf{1}; \mathbf{m})$.

Corollary 2 *The number of arrangements to have at least one of the k letter types to have the length of the $(m_i + 1)$th longest run equal to q_i is given by*

$$L(\mathbf{n}; \mathbf{q}; \mathbf{m}) = N(\mathbf{n}; \mathbf{q}; \mathbf{m}) - N(\mathbf{n}; \mathbf{q} - \mathbf{1}; \mathbf{m}).$$

If we define $W(\mathbf{n}; \mathbf{q}; \mathbf{m})$ as the number of arrangements for all letter types to have the length of the $(m_i + 1)$th longest run equal to q_i, then we have

Corollary 3 *The number of arrangements for all letter types to have the length of the $(m_i + 1)$th longest run equal to q_i is given by*

$$W(\mathbf{n}; \mathbf{m}; \mathbf{q}) = \sum_{p_i=1}^{n_i} (-1)^p \begin{bmatrix} p \\ p_i \end{bmatrix} \prod_{i=1}^{k} \left[H_{m_i}(n_i, q_i, p_i) - H_{m_i}(n_i, q_i - 1, p_i) \right],$$

where $p = \sum_{i=1}^{k} p_i$.

Applying Eq. (13) in Lemma 3 to Eq. (7) of Corollary 1, we can get the number of configurations of at least $m + 1$ runs of the first letter type of length q or greater, regardless of the other letter types:

$$Z(\mathbf{n}; q; m) = \begin{bmatrix} n \\ n_i \end{bmatrix} - \begin{bmatrix} n - n_1 \\ n_2, \ldots, n_k \end{bmatrix} \sum_{r_1=1}^{n_1} \binom{n - n_1 + 1}{r_1} h_m(n_1, q - 1, r_1).$$

The summation of r_1 in the above equation can be carried out, leading to

Corollary 4 *The number of configurations of at least $m + 1$ runs of the first letter type with length q or greater is given by*

$$Z(\mathbf{n}; q; m) = \begin{bmatrix} n-n_1 \\ n_2, \ldots, n_k \end{bmatrix} \sum_{j=1}^{\min(n-n_1+1, \lfloor n/q \rfloor)} (-1)^{m+j+1} \binom{j-1}{m} \binom{n-n_1+1}{j} \binom{n-qj}{n-n_1}. \tag{18}$$

Equation (18) is a generalization of previous results, such as those of Bradley (1968, p. 257). For the mth longest run, again the only difference is the extra binomial term $(-1)^m \binom{j-1}{m}$.

By using Theorem 1 and Lemma 3, the method can easily lead to joint distributions of various kinds. For example, instead of using $S = \{1\}$ to focus only on the lengths of runs of the first letter type, we can use $S = \{1, 2\}$ to obtain joint distributions of both the first and the second letter types. Other possibilities are to introduce more tracking variables in generating function Eq. (11) to track more run lengths within one letter type, instead of only one number q. The details are omitted here.

As for computational complexity, Eq. (15) has 3 nested summations over $n_i, i = 1, \ldots, k$: the inner k summations for $h_m(n, q, r)$ in Eq. (13), the middle k summations for the calculation of $F(\mathbf{r})$, and the outer k summations for variables r_i. Hence the computational complexity for Eq. (15) is $O(\prod_{i=1}^{k} n_i^3)$. Equation (16) of Theorem 2 simplifies the computation to two nested summations over n_i, and the computational complexity is reduced to $O(\prod_{i=1}^{k} n_i^2)$.

4 The second definition the mth longest run: run lengths sorted for all letter types

In Sect. 3, the mth longest runs are ordered within runs formed by individual letter types. In this section, distributions of mth longest runs of the whole system will be developed.

For this definition the lengths of runs are sorted regardless which letter type the run is made up of. The lengths of runs of the whole system are ordered as $l_0 \geq l_1 \geq \cdots \geq l_{r-1}$, where r is the total number of runs of the system. The length of the longest run of the whole system is l_0, with the length of the shortest run labeled as l_{r-1}. In general l_m denotes the length of the $(m+1)$th longest run of the whole system. We define $l_i = 0$ if $i \geq r$. If we use the same example shown previously in (8), then $l_0 = 6$, $l_1 = 5, l_2 = l_3 = l_4 = 3, l_5 = 2, l_6 = 1$, and $l_m = 0$ for $m > 6$.

We define $Q(\mathbf{n}; q; m)$ as the number of ways to arrange the whole system to have the length of the $(m+1)$th longest run less or equal to q, i.e., $l_m \leq q$. The definition of $Q(\mathbf{n}; q; m)$ implies that for all the arrangements counted by $Q(\mathbf{n}; q; m)$, there are at most m runs with lengths greater than q.

As before, $Q(\mathbf{n}; q; m)$ can be expressed by

$$Q(\mathbf{n}; q; m) = \sum_{s=0}^{m} \bar{Q}(\mathbf{n}; q; s),$$

where $\bar{Q}(\mathbf{n}; q; m)$ is the number of arrangements of the whole system where there are exactly m runs with lengths greater than q, regardless of the letter types. The numbers

given by $Q(\mathbf{n}; q; m)$ and $\bar{Q}(\mathbf{n}; q; m)$ are the corresponding quantities on the whole system level of the numbers given by $h_m(n, q, r)$ and $\bar{h}_m(n, q, r)$ discussed in Sect. 3 for a particular given letter type.

To calculate $\bar{Q}(\mathbf{n}; q; m)$, we use the same expression of $\bar{h}_m(n, q, r)$ in Eq. (12), which is the number of ways to arrange n elements of one particular letter type in r runs, with exact m runs longer than q. Again the function $F(\mathbf{r})$ is used to put the whole system together:

$$\bar{Q}(\mathbf{n}; q; s) = \sum_{\substack{m_i=0 \\ \sum m_i = s}}^{s} \sum_{r_i=1}^{n_i} F(\mathbf{r}) \prod_i \bar{h}_{m_i}(n_i, q, r_i).$$

Hence for $Q(\mathbf{n}; q; m)$ we have

$$Q(\mathbf{n}; q; m) = \sum_{s=0}^{m} \sum_{\substack{m_i=0 \\ \sum m_i = s}}^{s} \sum_{r_i=1}^{n_i} F(\mathbf{r}) \prod_i \bar{h}_{m_i}(n_i, q, r_i). \tag{19}$$

Equation (19) can be simplified. First, by using the explicit expression of $F(\mathbf{r})$ of Eq. (1), Eq. (19) can be simplified as

$$Q(\mathbf{n}; q; m) = \sum_{s=0}^{m} \sum_{\substack{m_i=0 \\ \sum m_i = s}}^{s} \sum_{p_i=1}^{n_i} (-1)^{p_i} \begin{bmatrix} \sum p_i \\ p_i \end{bmatrix} \prod_i \bar{H}_{m_i}(n_i, q, p_i), \tag{20}$$

where

$$\bar{H}_m(n, q, p) = \begin{cases} (-1)^m \binom{m-1}{p-1} \binom{n-1}{m-1} & q = 0, \\ \displaystyle\sum_{j=\lceil (n-p)/(q+1) \rceil}^{\lfloor (n-p)/q \rfloor} (-1)^{n+m+qj+j} \binom{j}{m} \binom{n-qj-1}{p-1} \binom{p}{n-qj-j} & \text{otherwise.} \end{cases} \tag{21}$$

The expression of Eq. (20) can be further simplified by getting rid of the selection summation on $\sum m_i = s$ in the second sum. Let us discuss the simplification for $q = 0$ and $q > 0$ separately.

When $q = 0$, if $m \geq n$, we have $Q(\mathbf{n}; 0; m) = \begin{bmatrix} n \\ n_i \end{bmatrix}$. For $q = 0$ and $m < n$, for each summation of m_i in Eq. (20), we can first ignore the selection restriction $\sum m_i = s$, and use a variable t to track m_i later. First look at the sum over one particular m_i:

$$\begin{aligned} \sum_{m_i=p_i}^{n_i} (-1)^{m_i} \binom{m_i - 1}{p_i - 1} \binom{n_i - 1}{m_i - 1} t^{m_i} &= \binom{n_i - 1}{p_i - 1} \sum_{m_i=p_i}^{n_i} (-1)^{m_i} \binom{n_i - p_i}{m_i - p_i} t^{m_i} \\ &= (-1)^{p_i} \binom{n_i - 1}{p_i - 1} (1 - t)^{n_i - p_i} t^{p_i}. \end{aligned}$$

The nested k sums of m_i will then give

$$(-1)^p(1-t)^{n-p}t^p\prod_{i=1}^{k}\binom{n_i-1}{p_i-1},$$

where $n=\sum_k n_i$ and $p=\sum_k p_i$. The selection restriction $\sum m_i = s$ just takes the coefficient of t^s from the above expression:

$$[t^s](-1)^p(1-t)^{n-p}t^p=(-1)^s\binom{n-p}{s-p}.$$

The outmost sum of s can then be carried out:

$$\sum_{s=0}^{m}(-1)^s\binom{n-p}{s-p}=(-1)^m\binom{n-p-1}{m-p}.$$

Putting all together, we have for $q=0$ and $m<n$,

$$Q(\mathbf{n};0;m)=(-1)^m\sum_{p_i}(-1)^p\binom{n-p-1}{m-p}\begin{bmatrix}p\\p_i\end{bmatrix}\prod_i\binom{n_i-1}{p_i-1}. \tag{22}$$

Similarly, for $q>0$ Eq. (20) can be simplified by first doing the sums on each m_i, and then filtering out the term with the selection restriction $\sum_{i=1}^{k} m_i = s$ by taking the coefficient of the t^s term.

In the end, after putting everything together, we obtain

Theorem 3 *The number of configurations of a system with $l_m \le q$ when all lengths of runs are sorted together regardless of letter types is given by, when $q>0$,*

$$\begin{aligned}&Q(\mathbf{n};q;m)\\&=(-1)^{n+m}\sum_{p_i=1}^{n_i}(-1)^p\begin{bmatrix}p\\p_i\end{bmatrix}\\&\quad\times\sum_{j_i=\lceil(n_i-p_i)/(q+1)\rceil}^{\lfloor(n_i-p_i)/q\rfloor}(-1)^{j(q+1)}\binom{j-1}{m}\prod_i\binom{n_i-qj_i-1}{p_i-1}\binom{p_i}{n_i-qj_i-j_i}\end{aligned} \tag{23}$$

with $j=\sum_i j_i$, $n=\sum_k n_i$, and $p=\sum_k p_i$. When $q=0$, if $m\ge n$,

$$Q(\mathbf{n};0;m)=\begin{bmatrix}n\\n_i\end{bmatrix},$$

when $q = 0$ and $m < n$,

$$Q(\mathbf{n}; 0; m) = (-1)^m \sum_{p_i=1}^{n_i} (-1)^p \binom{n-p-1}{m-p} \left[\begin{matrix} p \\ p_i \end{matrix} \right] \prod_i \binom{n_i - 1}{p_i - 1}.$$

If we compare Eq. (23) with Eq. (16), we see that the only difference is in the term $(-1)^{m+(q+1)j} \binom{j-1}{m}$: in Eq. (16) the term is calculated separately for individual letter type as $(-1)^{m_i+(q_i+1)j_i} \binom{j_i-1}{m_i}$, while in Eq. (23) the term is calculated for the whole system using the $j = \sum_i j_i$. From the definitions we see that when $m = 0$, if we set all q_i in Eq. (16) to q, so that $\mathbf{q} = (q, q, \ldots, q)$, $N(\mathbf{n}; \mathbf{q}; \mathbf{0}) = Q(\mathbf{n}; q; 0)$. This can be confirmed by comparing Eqs. (16) and (17) with Eq. (23). For $m > 0$, this will no longer be true.

Corollary 5 *The number of ways to have the length of the $(m + 1)$th longest run as q for the whole system is given by*

$$W(\mathbf{n}; q; m) = Q(\mathbf{n}; q; m) - Q(\mathbf{n}; q - 1; m).$$

As we can see, Eq. (22) is very similar in form to Eq. (28) of Kong (2006), which calculates the number of configurations with the total number of runs as r:

$$T(r; \mathbf{n}) = (-1)^r \sum_{p_i} (-1)^p \binom{n-p}{r-p} \left[\begin{matrix} p \\ p_i \end{matrix} \right] \prod_i \binom{n_i - 1}{p_i - 1}.$$

By the definition of $Q(\mathbf{n}; q; m)$, $Q(\mathbf{n}; 0; m)$ means the number of arrangements to have at most m runs with lengths greater than 0, i.e., with at most m runs. The relation between $Q(\mathbf{n}; 0; m)$ and $T(r; \mathbf{n})$ is obvious:

$$Q(\mathbf{n}; 0; m) = \sum_{r=0}^{m} T(r; \mathbf{n}),$$

which can be checked explicitly.

In Fig. 1 the probability mass distribution of $W(\mathbf{n}; q; m)$ for $\mathbf{n} = (n_1, n_2) = (200, 300)$ (divided by $\binom{n_1+n_2}{n_1}$) is plotted for $m = 0$ to 3. In Table 1, the average, the second moment, and the variance of the same system are listed. The distributions become narrower when m increases.

Acknowledgements This work was supported in part by the Clinical and Translational Science Award UL1 RR024139 from the National Center for Research Resources, National Institutes of Health.

References

Balakrishnan, N., Koutras, M. V. (2002). *Runs and scans with applications*. New York: Wiley.
Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs: Prentice-Hall.

Burr, E. J., Cane, G. (1961). Longest run of consecutive observations having a specified attribute. *Biometrika*, *48*, 461–465.
Eryilmaz, S. (2006). Some results associated with the longest run statistic in a sequence of Markov dependent trials. *Applied Mathematics and Computation*, *175*, 119–130.
Fu, J. C., Wang, L., Lou, W. Y. W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials. *Journal of Applied Probability*, *40*, 346–360.
Glaz, J., Naus, J., Wallenstein, S. (2001). *Scan statistics*. New York: Springer.
Godbole, A. P., Papastavridis, S. G. (Eds.). (1994). *Runs and patterns in probability: selected papers*. Dordrecht: Kluwer Academic Publishers.
Knuth, D. E. (1997). *The art of computer programming*. Seminumerical Algorithms (3rd edn, vol. 2). Boston: Addison-Wesley Longman Publishing Co. Inc.
Kong, Y. (2006). Distribution of runs and longest runs: a new generating function approach. *Journal of the American Statistical Association*, *101*, 1253–1263.
Kong, Y. (2015a). Distributions of runs revisited. *Communications in Statistics Theory and Methods*, *44*, 4663–4678.
Kong, Y. (2015b). Number of appearances of events in random sequences: a new approach to non-overlapping runs. *Communications in Statistics Theory and Methods* (to appear).
Kong, Y. (2015c). Number of appearances of events in random sequences: a new generating function to Type II and Type III runs. *Annals of the Institute of Statistical Mathematics*. doi:10.1007/s10463-015-0549-2.
Koutras, M. V., Alexandrou, V. A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach. *Annals of the Institute of Statistical Mathematics*, *47*(4), 743–766.
Koutras, M. V., Papastavridis, S. G. (1993). On the number of runs and related statistics. *Statistica Sinica*, *3*, 277–294.
Lou, W. Y. W. (1996). On runs and longest run tests: a method of finite Markov chain imbedding. *Journal of the American Statistical Association*, *91*, 1595–1601.
Makri, F. S., Philippou, A. N., Psillakis, Z. M. (2007). Shortest and longest length of success runs in binary sequences. *Journal of Statistical Planning and Inference*, *137*, 2226–2239.
Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics*, *11*, 367–392.
Morris, M., Schachtel, G., Karlin, S. (1993). Exact formulas for multitype run statistics in a random ordering. *SIAM Journal on Discrete Mathematics*, *6*, 70–86.
Mosteller, F. (1941). Note on an application of runs to quality control charts. *Annals of Mathematical Statistics*, *12*, 228–232.
Muselli, M. (1996). Useful inequalities for the longest run distribution. *Statistics & Probability Letters*, *46*, 239–249.
Olmstead, P. S. (1958). Runs determined in a sample by an arbitrary cut. *The Bell System Technical Journal*, *37*, 55–82.
Petkovsěk, M., Wilf, H.S., Zeilberger, D. (1996). $A = B$. Wellesley: A K Peters Ltd.
Philippou, A. N., Makri, F. S. (1985). Longest success runs and Fibonacci-type polynomials. *The Fibonacci Quarterly*, *23*, 338–346.
Philippou, A. N., Makri, F. S. (1986). Successes, runs, and longest runs. *Statistics & Probability Letters*, *4*, 211–215.
Schilling, M. F. (1990). The longest run of heads. *The College Mathematics Journal*, *21*, 196–207.
Schuster, E. F. (1996). The conditional distribution of the longest run in a sample from a multiletter alphabet. *Communications in Statistics Simulation and Computation*, *25*, 215–224.