Distribution and double generating function of number of patterns in a sequence of Markov dependent multistate trials

Yung-Ming Chang · James C. Fu · Han-Ying Lin

Received: 7 May 2009 / Revised: 28 October 2009 / Published online: 21 May 2010

Abstract In this manuscript, the dual relationship between the probability of number of runs and patterns and the probability of waiting time of runs and patterns in a sequence of multistate trials has been studied via double generating functions and recursive equations. The results, which are established under different assumptions on patterns, underlying sequences and counting schemes, are extensions of Koutras's results (1997, Advances in Combinatorial Methods and Applications to Probability and Statistics, Boston: Birkhäuser). As byproducts, the exact distributions of the longest-run statistics are also derived. Numerical examples are provided for illustrating the theoretical results.

Keywords Simple and compound patterns · Waiting time · Finite Markov chain imbedding · Probability generating function · Double generating function

1 Introduction

The importance and usefulness of runs and patterns arise from their widespread applications in diverse areas of science. Theoretical research on various distributions associated with runs and patterns in a random sequence of multistate trials has been extensively conducted in the literature. Among these distributions, two most com-

Y.-M. Chang (✉)
Department of Mathematics, National Taitung University, Taitung 95002, Taiwan
e-mail: eddchang@nttu.edu.tw

J. C. Fu
Department of Statistics, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada

H.-Y. Lin
Institute of Statistics, National University of Kaohsiung, Kaohsiung 811, Taiwan

monly studied are (i) the distribution of number of runs and patterns (for example, Eryilmaz 2008a; Fu and Koutras 1994; Hirano and Aki 1993; Hirano et al. 1997; Shinde and Kotwal 2006); and (ii) the distribution of waiting time (number of trials) to observe the rth ($r \geq 1$) occurrence of runs and patterns (for example, Aki 1992; Aki and Hirano 1999; Eryilmaz 2008b; Fu and Chang 2002). These two distributions are closely related (see Feller 1968). In fact, the distributions in (i) and (ii) can be derived from each other.

Traditionally, a combinatorial approach was adopted in the study of runs and patterns. Fu and Koutras (1994) introduced a general and efficient method, the finite Markov chain imbedding (FMCI) technique, to study the distributions of several run-related statistics. The basic idea of this approach is to imbed the random variable of interest into a Markov chain. Then the exact distribution of that variable can be expressed in a simple form in terms of the transition probability matrix of its imbedded Markov chain. Since the FMCI technique has been applied successfully to deal with many general problems for runs and patterns, it has become an alternative approach in this field.

In this manuscript, we aim to study the dual relationship between the distribution of number of runs and patterns and the distribution of waiting time of runs and patterns, based on different assumptions on patterns, underlying sequences and counting schemes. We first show that the double generating function of number of patterns can be expressed in terms of the probability generating function (PGF) of waiting time for the first occurrence of patterns, whose transition matrix of imbedded Markov chain has a simpler structure. Following Stanley (1997) techniques, we establish recurrence relations for the exact distributions of number of patterns. As byproducts, the exact distributions of the longest-run statistics are also derived.

In Sect. 2, we introduce basic definitions for runs and patterns and provide some results related to waiting time distributions. The general results are presented in Sect. 3. In Sect. 4, we provide numerical examples to see the performance of our method. Finally, in Sect. 5, we give some concluding remarks.

2 Notations and preliminaries

Let $\{X_i\}$ be a sequence of m-state ($m \geq 2$) random variables defined on the state space $\Gamma = \{c_1, c_2, \ldots, c_m\}$. Unless mentioned otherwise, we assume that $\{X_i\}$ is a sequence of first-order homogeneous Markov-dependent m-state trials.

Definition 1 We say that Λ is a simple pattern if Λ is composed of a specified sequence of k states (k is fixed); i.e., $\Lambda = c_{i_1} \cdots c_{i_k}$, $i_j \in \{1, \ldots, m\}$ for all $j = 1, \ldots, k$.

Definition 2 A proper subpattern of a simple pattern Λ is defined to be a finite sequence having the general form $c_{i_1} \cdots c_{i_j}$, $1 \leq j \leq k - 1$.

Let Λ_1 and Λ_2 be two simple patterns with lengths k_1 and k_2, respectively. Define a segment to be any (contiguous) subset of a simple pattern. For example, let $\Lambda = c_1c_1c_2c_2$ be a simple pattern; then, the subpatterns c_1, c_1c_1 and $c_1c_1c_2$ are segments of Λ. On the other hand, c_2, c_1c_2, c_2c_2, and $c_1c_2c_2$ are segments of Λ, but not subpatterns. We say that Λ_1 and Λ_2 are distinct if neither is a segment of the other. Define the union $\Lambda_1 \cup \Lambda_2$ to be the occurrence of either the pattern Λ_1 or the pattern Λ_2.

Definition 3 We say that Λ is a compound pattern if Λ is composed of a union of l distinct simple patterns; i.e., $\Lambda = \cup_{i=1}^{l} \Lambda_i$.

Next, we introduce some random variables associated with simple and compound patterns. Let $X_n(\Lambda)$ denote the number of a pattern Λ (simple or compound) that occurred in a sequence of n m-state trials. As stated in Sect. 1, the random variable $X_n(\Lambda)$ is closely related to the waiting time of Λ. Therefore, it is important for us to provide the definition of waiting time.

Definition 4 For a given integer r, $r = 1, 2, \ldots$, we define the waiting time $W(r, \Lambda)$ to the rth occurrence of a pattern Λ (simple or compound) to be the minimum number of trials required to observe the rth occurrence of Λ.

For brevity, we denote $W(1, \Lambda)$ as $W(\Lambda)$. From the definitions of $X_n(\Lambda)$ and $W(r, \Lambda)$, we have the following dual relationship (see Feller 1968):

$$X_n(\Lambda) < r \quad \text{if and only if} \quad W(r, \Lambda) > n.$$

Hence the probability of the event $\{X_n(\Lambda) = r\}$ can be computed by

$$\begin{aligned} P(X_n(\Lambda) = r) &= P(X_n(\Lambda) < r + 1) - P(X_n(\Lambda) < r) \\ &= P(W(r + 1, \Lambda) > n) - P(W(r, \Lambda) > n). \end{aligned} \tag{1}$$

In the study of the distributions of $X_n(\Lambda)$ and $W(r, \Lambda)$ $(r > 1)$, the ways of counting patterns should be taken into consideration. In this article, we consider nonoverlapping and overlapping counting schemes.

Given a compound pattern $\Lambda = \cup_{i=1}^{l} \Lambda_i$, it has been shown that the waiting time $W(\Lambda)$ is finite Markov chain imbeddable. Hence, there exists a Markov chain $\{Y_t : t = 0, 1, \ldots\}$ defined on the finite state space

$$\Omega = \{\emptyset\} \cup \Gamma \cup \bigcup_{i=1}^{l} \mathcal{S}(\Lambda_i) \cup \{\Lambda_1, \ldots, \Lambda_l\},$$

where $\emptyset$ is the initial state and $\mathcal{S}(\Lambda_i)$ is the collection of all proper subpatterns of Λ_i, for $i = 1, \ldots, l$. The states in the state space Ω can be renumbered as $\Omega = \{1, \ldots, w, \alpha_1, \ldots, \alpha_l\}$, where $\alpha_1, \ldots, \alpha_l$ denote the absorbing states corresponding to the patterns $\Lambda_1, \ldots, \Lambda_l$, respectively. Hence, the transition probability matrix $\boldsymbol{M}$ of the imbedded Markov chain $\{Y_t\}$ can be obtained in the form

$$\boldsymbol{M} = \left[\begin{array}{c|c} \boldsymbol{N} & \boldsymbol{C} \\ \hline \boldsymbol{0} & \boldsymbol{I} \end{array}\right]. \tag{2}$$

The following results were established by Fu and Chang (2002).

Theorem 1 *Let Λ be a pattern (simple or compound). For a waiting time random variable $W(\Lambda)$, we have*

(i) *the exact distribution of $W(\Lambda)$ is given by*

$$P(W(\Lambda)=n)=\boldsymbol{\xi}\boldsymbol{N}^{n-1}(\boldsymbol{I}-\boldsymbol{N})\mathbf{1}^{\top}, \tag{3}$$

where $\boldsymbol{\xi}=(1,0,\ldots,0)_{1\times w}$ is the initial distribution with $P(Y_0=\emptyset)\equiv 1$ and $\mathbf{1}^{\top}$ is the transpose of the row vector $\mathbf{1}=(1,1,\ldots,1)_{1\times w}$; and

(ii) *the PGF of $W(\Lambda)$ is given by*

$$\varphi_{W(\Lambda)}(s)=1+\left(1-\frac{1}{s}\right)\Phi_{W(\Lambda)}(s), \tag{4}$$

where $\Phi_{W(\Lambda)}(s)=\sum_{i=1}^{w}\phi_i(s)$, and $(\phi_1(s),\ldots,\phi_w(s))$ is the solution of the following simultaneous equations

$$\phi_i(s)=s\boldsymbol{\xi}\boldsymbol{e}_i^{\top}+s(\phi_1(s),\ldots,\phi_w(s))\boldsymbol{N}(i),\ i=1,\ldots,w, \tag{5}$$

where $\boldsymbol{e}_i=(0,\ldots,0,1,0,\ldots,0)$, $i=1,\ldots,w$, are unit vectors, and $\boldsymbol{N}(i)$, $i=1,\ldots,w$, are the column vectors of the matrix $\boldsymbol{N}$.

Chang (2005) has shown that the transition probability matrix associated with the imbedded Markov chain of $W(r,\Lambda)$ $(r>1)$ has the same form as (2). Thus, the results in Theorem 1 are also applicable for $W(r,\Lambda)$. For example, the generating function $\Phi_{W(r,\Lambda)}(s)=\sum_{n=1}^{\infty}s^nP(W(r,\Lambda)\geq n)$ can be obtained in a similar way as Theorem 1(ii) and hence it follows from Equation (4) that

$$\Phi_{W(r,\Lambda)}(s)=\frac{s}{1-s}\left[1-\varphi_{W(r,\Lambda)}(s)\right]. \tag{6}$$

Instead of using Theorem 1, Chang (2005) also suggested an alternative way of finding the PGF $\varphi_{W(r,\Lambda)}(s)$. Let $W(\Lambda_j|\Lambda_1,\ldots,\Lambda_l)$ be the waiting time to the first occurrence of Λ_j, and Λ_j occurs first among all the patterns $\Lambda_1,\ldots,\Lambda_l$, and let $\psi_{W(\Lambda_j|\Lambda_1,\ldots,\Lambda_l)}(s|\boldsymbol{\xi})$ denote its generating function that depends on the initial distribution $\boldsymbol{\xi}$. Under nonoverlapping counting scheme, the PGF $\varphi_{W(r,\Lambda)}(s)$ can be obtained from the following general form:

$$\varphi_{W(r,\Lambda)}(s)=\sum_{i_1,\ldots,i_r\in\{1,2,\ldots,l\}}\prod_{j=1}^{r}\psi_{W_j(\Lambda_{i_j}|\Lambda_1,\ldots,\Lambda_l)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda_{i_{j-1}})\right), \tag{7}$$

where the initial state $\emptyset_{j^\star}$ is the last element $j^\star$ of $\Lambda_{i_{j-1}}$, $j=2,\ldots,r$, $\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda_{i_{j-1}})=(1,0,\ldots,0)$ is the initial distribution, and $\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda_{i_0})=\boldsymbol{\xi}$ by convention. He further indicated that the above formula can be reduced to some well-known cases under different assumptions on patterns, underlying sequences and counting schemes. Write $W(r, \Lambda)$ as

$$W(r, \Lambda) = W_1(\Lambda) + \cdots + W_r(\Lambda),$$

where $W_i(\Lambda),\ i = 1, \ldots, r$, are interwaiting times. We summarize these results as follows:

(i) If $\{X_i\}$ consists of independent and identically distributed (i.i.d.) trials and Λ is a compound pattern with respect to nonoverlapping counting scheme, then $\varphi_{W(r,\Lambda)}(s)$ is given by

$$\varphi_{W(r,\Lambda)}(s) = \left[\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\right]^r, \tag{8}$$

where the initial distribution for the imbedded Markov chain of $W_1(\Lambda)(\equiv W(\Lambda))$ is $\boldsymbol{\xi} = (1, 0, \ldots, 0)$ (with usual initial state $\emptyset$).

(ii) If $\{X_i\}$ consists of Markov-dependent trials and Λ is a simple pattern with respect to nonoverlapping counting scheme, then $\varphi_{W(r,\Lambda)}(s)$ is given by

$$\varphi_{W(r,\Lambda)}(s) = \varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi}) \left[\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)\right]^{r-1}, \tag{9}$$

where the initial state $\emptyset_{j^\star}$ for the imbedded Markov chain of $W_2(\Lambda)$ is the last element $j^\star$ of the pattern Λ and the initial distribution is $\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda) = (1, 0, \ldots, 0)$. In the same case, if overlapping counting scheme is adopted, then the initial distribution for the imbedded Markov chain of $W_2(\Lambda)$ is replaced by $\boldsymbol{\xi}_{\emptyset_{j^\circ}}(\Lambda) = (0, \ldots, 1, \ldots, 0)$, where the position of 1 corresponds to the initial state j° which is the longest subpattern counting backward from the last element of Λ.

To close this section, we point out that Equations (1), (6), (8) and (9) lay the foundation for studying the relationship between the distributions of $X_n(\Lambda)$ and $W(r, \Lambda)$.

3 Distributions of number of patterns

The goal of this section is to study the distributions of $X_n(\Lambda)$. We first show that the double generating function of $X_n(\Lambda)$ can be expressed in terms of the PGF of the waiting time $W(\Lambda)$. Then recurrence relations for the exact distributions of $X_n(\Lambda)$ can be derived by using Stanley (1997) techniques.

3.1 The double generating function of $X_n(\Lambda)$

Let $G(s, t)$ denote the double generating function of $X_n(\Lambda)$; i.e.,

$$G(s, t) = \sum_{n=0}^{\infty} \varphi_{X_n(\Lambda)}(t)s^n = \sum_{n=0}^{\infty}\left(\sum_{r=0}^{\infty} P(X_n(\Lambda) = r)t^r\right)s^n.$$

We establish general results for $G(s,t)$ under different assumptions on patterns, underlying sequences and counting schemes discussed in Sect. 2.

Theorem 2 *Let Λ be a pattern (simple or compound) and $\{X_i\}$ be a sequence of i.i.d. multistate trials. Suppose nonoverlapping counting scheme is used, then the double generating function of $X_n(\Lambda)$ is given by*

$$G_1(s,t) = \frac{1}{1-s}\left[\frac{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})}{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t}\right]. \tag{10}$$

Proof It follows from the definition, Equations (1), (6) and (8) that

$$\begin{aligned}
G_1(s,t) &= \sum_{n=0}^{\infty} \varphi_{X_n(\Lambda)}(t)s^n \\
&= \sum_{r=0}^{\infty}\frac{1}{s}\sum_{n=0}^{\infty}\left[P(W(r+1,\Lambda)\geq n+1) - P(W(r,\Lambda)\geq n+1)\right]s^{n+1}t^r \\
&= \sum_{r=0}^{\infty}\frac{1}{s}\left[\Phi_{W(r+1,\Lambda)}(s) - \Phi_{W(r,\Lambda)}(s)\right]t^r \\
&= \frac{1}{s}\sum_{r=0}^{\infty}\frac{s}{1-s}\left[\varphi_{W(r,\Lambda)}(s) - \varphi_{W(r+1,\Lambda)}(s)\right]t^r \\
&= \frac{1}{1-s}\sum_{r=0}^{\infty}\varphi_{W(r,\Lambda)}(s)t^r - \frac{1}{1-s}\sum_{r=0}^{\infty}\varphi_{W(r+1,\Lambda)}(s)t^r \\
&= \frac{1}{1-s}\sum_{r=0}^{\infty}\left[\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\right]^r t^r - \frac{1}{1-s}\sum_{r=0}^{\infty}\left[\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\right]^{r+1} t^r \\
&= \frac{1}{1-s}\left[\frac{1}{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t} - \frac{\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})}{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t}\right] \\
&= \frac{1}{1-s}\left[\frac{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})}{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t}\right].
\end{aligned}$$

This completes the proof. □

Theorem 3 *Let Λ be a simple pattern and $\{X_i\}$ be a sequence of Markov-dependent multistate trials. Suppose nonoverlapping counting scheme is used, then the double generating function of $X_n(\Lambda)$ is given by*

$$G_2(s,t) = \frac{1}{1-s}\left[1 - \frac{\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})(1-t)}{1-\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t}\right]. \tag{11}$$

Proof The proof is along the lines of the proof of Theorem 2. It follows from Equation (9) that

$$
\begin{aligned}
G_2(s,t) &= \frac{1}{1-s}\sum_{r=0}^{\infty}\varphi_{W(r,\Lambda)}(s)t^r - \frac{1}{1-s}\sum_{r=0}^{\infty}\varphi_{W(r+1,\Lambda)}(s)t^r \\
&= \frac{1}{1-s}\left\{1+\sum_{r=1}^{\infty}\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\left[\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)\right]^{r-1}t^r - \varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\right. \\
&\quad - \varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t \\
&\quad \left. - \sum_{r=2}^{\infty}\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\left[\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)\right]^r t^r\right\} \\
&= \frac{1}{1-s}\left\{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t+\left[\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t\right.\right. \\
&\quad \left.\left. -\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})\left(\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)\right)^2 t^2\right] \Big/ \left(1-\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t\right)\right\} \\
&= \frac{1}{1-s}\left[1-\frac{\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})(1-t)}{1-\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t}\right].
\end{aligned}
$$

This completes the proof. □

When $\{X_i\}$ is a sequence of multistate trials (i.i.d. or Markov-dependent) and Λ is a simple pattern with respect to overlapping counting scheme, the double generating function of $X_n(\Lambda)$ has the same form as Equation (11). The only difference is that the initial distribution for the imbedded Markov chain of $W_2(\Lambda)$ is $\boldsymbol{\xi}_{\emptyset_{j^\circ}}(\Lambda)$ as discussed in Sect. 2.

Remark 1 Koutras (1997) also established the same results as Equations (10) and (11). However, he did not clearly specify the assumptions on patterns, underlying sequences and counting schemes. Particularly, when $\{X_i\}$ is a sequence of Markov-dependent multistate trials and Λ is a compound pattern, there is no simple form for the double generating function of $X_n(\Lambda)$.

From Theorems 2 and 3, it is clear that we can obtain the PGF $\varphi_{X_n(\Lambda)}(t)$ through the double generating function of $X_n(\Lambda)$. For example, by Theorem 2, we have

$$
\varphi_{X_n(\Lambda)}(t) = \frac{1}{n!}D_s^n G_1(s,t)\Big|_{s=0}. \tag{12}
$$

However, when the length of the pattern Λ is large, the analytic form for $G_1(s,t)$ (or $G_2(s,t)$) becomes very complicated. Therefore, differentiating the double generating function $G_1(s,t)$ (or $G_2(s,t)$) n times may be troublesome. In the next section, we will show how to overcome these difficulties and present some new results.

3.2 The exact distributions of $X_n(\Lambda)$

To derive the exact probability distribution for $X_n(\Lambda)$, we start from the case described in Theorem 2. In view of Equation (10), the double generating function $G_1(s,t)$ always has a rational form. Stanley (1997) proposed a method for computing the coefficients of a rational function. Note that $G_1(s,t)$ is a rational function of s, i.e.,

$$G_1(s,t)=\frac{1}{1-s}\left[\frac{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})}{1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t}\right]=\frac{R_1(s)}{Q_1(s,t)}=\sum_{n=0}^{\infty}\varphi_{X_n(\Lambda)}(t)s^n. \quad (13)$$

For brevity, we denote $\varphi_{X_n(\Lambda)}(t)$ as $\varphi_n(t)$. Without loss of generality, we assume that $R_1(s)=\beta_0+\beta_1 s+\cdots+\beta_m s^m$ and $Q_1(s,t)=1+\gamma_1(t)s+\cdots+\gamma_d(t)s^d$ (possibly $m\geq d$). From Equation (13), we have $R_1(s)=Q_1(s,t)\sum_{n=0}^{\infty}\varphi_n(t)s^n$. Then, equating the coefficients of s^n yields the following recurrence relation:

$$\varphi_n(t)=-\sum_{i=1}^{d}\gamma_i(t)\varphi_{n-i}(t)\delta_{n-i}+\beta_n\delta_{m-n}, \quad (14)$$

where

$$\delta_x=\begin{cases}1, & \text{if } x\geq 0,\\ 0, & \text{if } x<0.\end{cases}$$

Note that the denominator of $G_1(s,t)$ is a linear function of t. This fact implies the following lemma:

Lemma 1 $\gamma_i(t)=a_i+c_i t$ *for all* $i=1,\ldots,d$, *where* a_i *and* c_i, $i=1,\ldots,d$, *are constants.*

Proof Since the PGF of $W(\Lambda)$ always has the form

$$\varphi_{W(\Lambda)}(s)=\frac{H(s)}{K(s)},$$

where $H(s)$ and $K(s)$ are polynomials in s. Hence the result follows from the fact that the denominator of $G_1(s,t)$ is $(1-s)(1-\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})t)$. □

Now the exact probability distribution of $X_n(\Lambda)$ can be obtained from Equation (14). Before giving the general results, we notice that given n, the maximum number of occurrence of a simple pattern Λ is determined by the length of Λ. For example, let Λ be a simple pattern with length 3. If $n=20$, then $X_n(\Lambda)$ is at most $\left[\frac{20}{3}\right]=6$ under nonoverlapping counting scheme. For a compound pattern $\Lambda=\cup_{i=1}^{l}\Lambda_i$, we define the minimum length k of Λ to be $k=\min[k_1,\ldots,k_l]$, where k_i denotes the length of the simple pattern Λ_i. We have the following results:

Theorem 4 *Let* Λ *be a compound pattern with minimum length* k *and let* $\{X_i\}$ *be a sequence of i.i.d. multistate trials. Suppose nonoverlapping counting scheme is adopted. Then for a given integer* $r=0,1,\ldots,\left[\frac{n}{k}\right]$, *the exact probability distribution of* $X_n(\Lambda)$ *is given by*

$$P(X_n(\Lambda)=r)=\begin{cases}-\sum_{i=1}^{d}\gamma_i(0)P(X_{n-i}(\Lambda)=0)\delta_{n-i}+\beta_n\delta_{m-n}, & \textit{if}\, r=0,\\ -\sum_{i=1}^{d}\sum_{j=0}^{1}\gamma_i^{(j)}(0)\,P(X_{n-i}(\Lambda)=r-j)\,\delta_{n-i}, & \textit{if}\, r\geq 1,\end{cases} \tag{15}$$

which satisfies the following conditions:

(i) *if* $0\leq n<k$, *then*

$$P(X_n(\Lambda)=r)=\begin{cases}1, & \textit{if}\, r=0,\\ 0, & \textit{if}\, r\geq 1,\end{cases}$$

(ii) *if* $r<0$ *or* $r>\left[\frac{n}{k}\right]$ *or* $n<0$, *then* $P(X_n(\Lambda)=r)=0$.

Proof The proof follows directly from the definition and Equation (14). For $r=0$, we have

$$\begin{aligned}P(X_n(\Lambda)=0)&=\varphi_n(t)\,|_{t=0}\\ &=-\sum_{i=1}^{d}\gamma_i(0)\varphi_{n-i}(0)\delta_{n-i}+\beta_n\delta_{m-n}\\ &=-\sum_{i=1}^{d}\gamma_i(0)P(X_{n-i}(\Lambda)=0)\delta_{n-i}+\beta_n\delta_{m-n}.\end{aligned}$$

Further, for $1\leq r\leq\left[\frac{n}{k}\right]$, we have

$$\begin{aligned}P(X_n(\Lambda)=r)&=\frac{1}{r!}\left.\frac{d^r}{d\,t^r}\varphi_n(t)\right|_{t=0}\\ &=-\sum_{i=1}^{d}\frac{1}{r!}\sum_{j=0}^{r}\binom{r}{j}\gamma_i^{(j)}(0)\,\varphi_{n-i}^{(r-j)}(0)\,\delta_{n-i}\\ &=-\sum_{i=1}^{d}\sum_{j=0}^{r}\frac{1}{j!}\gamma_i^{(j)}(0)P(X_{n-i}(\Lambda)=r-j)\delta_{n-i},\end{aligned}$$

where $\gamma_i^{(j)}(0)=\left.\frac{d^j}{d\,t^j}\gamma_i(t)\right|_{t=0}$ and $\varphi_{n-i}^{(r-j)}(0)=\left.\frac{d^{(r-j)}}{d\,t^{(r-j)}}\varphi_{n-i}(t)\right|_{t=0}$. Finally, note from Lemma 1 that $\gamma_i^{(j)}(t)\equiv 0$ for $j\geq 2$. Hence

$$P(X_n(\Lambda)=r)=-\sum_{i=1}^{d}\sum_{j=0}^{1}\gamma_i^{(j)}(0)P(X_{n-i}(\Lambda)=r-j)\delta_{n-i}.$$

This completes the proof. □

For the case described in Theorem 3, the exact distribution of $X_n(\Lambda)$ can be derived in a similar fashion. Observe that Equation (11) (i.e. $G_2(s,t)$) has a similar form as Equation (10) except that the numerator, denoted by $R_2(s,t)$, involves the parameter t. Thus, $G_2(s,t)$ can be written as a rational form like Equation (13); i.e.,

$$\begin{aligned} G_2(s,t) &= \frac{1}{1-s}\left[1-\frac{\varphi_{W_1(\Lambda)}(s|\boldsymbol{\xi})(1-t)}{1-\varphi_{W_2(\Lambda)}\left(s|\boldsymbol{\xi}_{\emptyset_{j^\star}}(\Lambda)\right)t}\right]=\frac{R_2(s,t)}{Q_2(s,t)} \\ &= \sum_{n=0}^{\infty}\varphi_n(t)s^n. \end{aligned} \tag{16}$$

Again, we assume that $R_2(s,t)=\beta_0(t)+\beta_1(t)s+\cdots+\beta_m(t)s^m$ and $Q_2(s,t)=1+\gamma_1(t)s+\cdots+\gamma_d(t)s^d$. Then we get

$$\varphi_n(t)=-\sum_{i=1}^{d}\gamma_i(t)\varphi_{n-i}(t)\delta_{n-i}+\beta_n(t)\delta_{m-n}, \tag{17}$$

where $\beta_n(t)$ is a linear function of t. We establish the following results:

Theorem 5 *Let Λ be a simple pattern of length k and let $\{X_i\}$ be a sequence of Markov-dependent multistate trials. Suppose nonoverlapping counting scheme is used. Then for a given integer $r=0,1,\ldots,\left[\frac{n}{k}\right]$, the exact probability distribution of $X_n(\Lambda)$ is given by*

$$P(X_n(\Lambda)=r)=\begin{cases} -\sum\limits_{i=1}^{d}\gamma_i(0)P(X_{n-i}(\Lambda)=0)\delta_{n-i}+\beta_n(0)\delta_{m-n}, & \textit{if } r=0,\\ -\sum\limits_{i=1}^{d}\sum\limits_{j=0}^{1}\gamma_i^{(j)}(0)\,P(X_{n-i}(\Lambda)=1-j)\,\delta_{n-i}+\beta_n^{'}(0)\delta_{m-n}, & \textit{if } r=1,\\ -\sum\limits_{i=1}^{d}\sum\limits_{j=0}^{1}\gamma_i^{(j)}(0)\,P(X_{n-i}(\Lambda)=r-j)\,\delta_{n-i}, & \textit{if } r\geq 2, \end{cases} \tag{18}$$

which satisfies the following conditions:

(i) *if $0\leq n<k$, then*

$$P(X_n(\Lambda)=r)=\begin{cases}1, & \textit{if } r=0,\\ 0, & \textit{if } r\geq 1,\end{cases}$$

(ii) *if $r<0$ or $r>\left[\frac{n}{k}\right]$ or $n<0$, then $P(X_n(\Lambda)=r)=0$.*

When Λ is a simple pattern and $\{X_i\}$ is a sequence of i.i.d. or Markov-dependent multistate trials with respect to overlapping counting scheme, the exact distribution of $X_n(\Lambda)$ has the same expression as Equation (18) except that the maximum value of

r is $\left[\frac{n-k}{k-k^o}\right]+1$, where k^o is the length of the longest subpattern counting backward from the last element of Λ.

The major advantage of the recursive formulas derived in this section is their efficiency in computation. Another interesting feature of these formulas is that they can be applied to find the exact distributions of the longest-run statistics. We give brief discussions in the following.

Let $\{X_i\}_{i=1}^n$ be a sequence of m-state trials defined on the state space $\Gamma = \{c_1, c_2, \ldots, c_m\}$. Given a positive integer n and a fixed j, we define $L_n(c_j)$ to be the length of the longest run of the symbol c_j. More precisely,

$$L_n(c_j) = \max_{1\le i\le n-k+1} \{k : X_i = X_{i+1} = \cdots = X_{i+k-1} = c_j\}.$$

Further, we define L_n to be the length of the longest run of one of c_j's; i.e.,

$$L_n = \max_{1\le j\le m} L_n(c_j).$$

The statistics defined above are closely related to $X_n(\Lambda)$. These relations are stated mathematically below:

(i) Given j, let $\Lambda_j = c_j \cdots c_j$ be a c_j run of length k. Then

$$P(L_n(c_j) < k) = P(X_n(\Lambda_j) = 0). \tag{19}$$

(ii) Let $\Lambda = \cup_{j=1}^m \Lambda_j$ be a compound pattern, where Λ_j is a c_j run of length k for $j = 1, \ldots, m$. Then

$$P(L_n < k) = P(X_n(\Lambda) = 0). \tag{20}$$

Based on (19), (20) and previous discussions, the exact distributions for $L_n(c_j)$, $j = 1, \ldots, m$ and L_n can be obtained via Equations (15) and (18).

4 Numerical examples

We have developed computer programs for computing the exact probability $P(X_n(\Lambda) = r)$. It is worth noting that given r and n the exact probabilities $P(X_i(\Lambda) = x)$ for all $i = 0, 1, \ldots, n$ and $x = 0, 1, \ldots, r$ are automatically generated from our computer programs. Therefore, the exact probability $P(r_0 < X_n(\Lambda) < r)$ can be easily obtained. We provide two numerical examples to illustrate our results.

Example 1 Let $\{X_i\}$ be a sequence of i.i.d. four-state trials with possible outcomes A, C, G, and T, respectively. Assume that $P(X_i = \mathrm{A}) = P(X_i = \mathrm{C}) = P(X_i = \mathrm{G}) = P(X_i = \mathrm{T}) = 0.25$ for $i = 1, \ldots, n$. Under nonoverlapping counting scheme, we consider the compound pattern $\Lambda = \mathrm{ACACGTGT} \cup \mathrm{ATTATAAT} \cup \mathrm{CAACGTTG}$. Numerical results for $P(X_n(\Lambda) = r)$ and $P(r_0 < X_n(\Lambda) < r)$ for various values of r_0, r and n are presented in Tables 1 and 2, respectively. The expected value of $X_n(\Lambda)$ can be computed easily; for example, $E[X_n(\Lambda)] = 4.58$ for $n = 100000$.

Table 1 The exact probabilities $P(X_n(\Lambda) = r)$ for the compound pattern $\Lambda =$ ACACGTGT$\cup$ATTATAAT$\cup$CAACGTTG and selected values of r and n

r	n	Exact Prob.	r	n	Exact Prob.
3	10000	0.010068725	20	100000	6.769244373e-08
3	20000	0.051126247	20	200000	7.397372678e-04
3	30000	0.109279727	20	300000	2.538079133e-02
3	40000	0.163960234	50	100000	3.106214787e-34
3	50000	0.202654217	50	200000	3.944099931e-21
3	80000	0.210262429	50	300000	2.663936175e-14
3	100000	0.164380665	80	100000	6.610775615e-69
3	200000	0.013505813	80	200000	1.044034900e-46
3	300000	0.000468036	80	300000	1.419900202e-34

Table 2 The exact probabilities $P(r_0 < X_n(\Lambda) < r)$ for the pattern $\Lambda =$ ACACGTGT $\cup$ ATTATAAT $\cup$ CAACGTTG and selected values of r_0, r and n

r_0	r	n	Exact Prob.	r_0	r	n	Exact Prob.
−1	5	10000	0.999887186	9	21	100000	0.018906703
−1	5	30000	0.986825015	9	21	200000	0.432366606
−1	5	50000	0.917677876	9	21	300000	0.836741046
−1	5	100000	0.517526837	14	101	100000	0.000087841
−1	5	200000	0.049923910	14	101	200000	0.046618242
−1	5	300000	0.002193756	14	101	300000	0.401117175

Example 2 Let $\{X_i\}$ be a sequence of first-order homogeneous Markov chain with possible outcomes A, C, G, and T and transition probability matrix

$$\boldsymbol{A} = \begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \\ 0.2 & 0.3 & 0.2 & 0.3 \\ 0.7 & 0.1 & 0.1 & 0.1 \end{bmatrix}.$$

Given the initial probabilities $P(X_1 = \text{A}) = 0.6$, $P(X_1 = \text{C}) = 0.2$, $P(X_1 = \text{G}) = 0.1$ and $P(X_1 = \text{T}) = 0.1$. Let $\Lambda =$ CTACT be a simple pattern. Under overlapping counting scheme, the longest proper subpattern counting backward from the last element of Λ is "CT" ($\Lambda =$ CTA$\underline{\text{CT}}$). Hence, the maximum possible value for $X_n(\Lambda)$ is $\left[\frac{n-k}{k-k^o}\right] + 1 = \left[\frac{n-5}{5-2}\right] + 1$. With $n = 10000$, the exact distributions of $X_n(\Lambda)$ with respect to overlapping and nonoverlapping counting schemes are shown in Fig. 1. The expected values of $X_n(\Lambda)$ are $E[X_n(\Lambda)] = 563.06$ for overlapping counting and $E[X_n(\Lambda)] = 435.16$ for nonoverlapping counting, respectively.

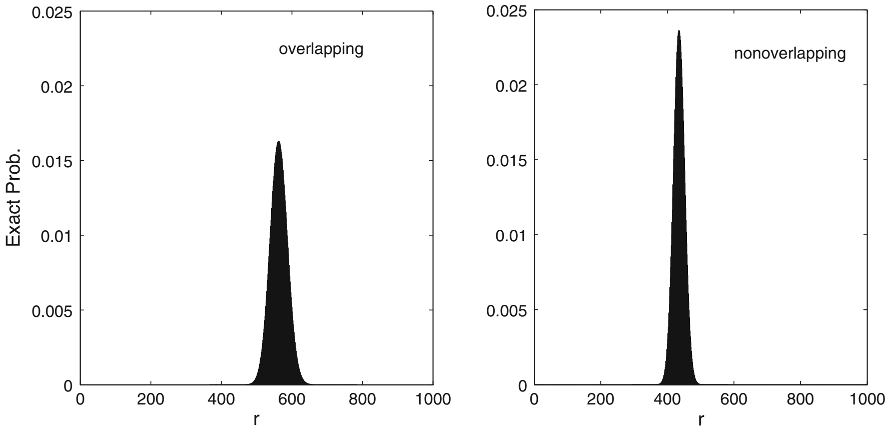

Fig. 1 The exact distributions of $X_{10000}(\Lambda)$ for Λ = CTACT with respect to overlapping and nonoverlapping counting schemes

5 Concluding remarks

In this manuscript, we have derived recursive formulas for the exact probability $P(X_n(\Lambda) = r)$ under different assumptions on the pattern Λ, underlying sequences and counting schemes. It can be seen from our numerical experiments that these formulas perform very well for both large r and n. They are also feasible for symbolic computation for reasonable large pattern and moderate r and n. In addition, our method could be applied to the case when Λ is a simple pattern and $\{X_i\}$ is a sequence of higher-order homogeneous Markov-dependent multistate trials. The key is to specify the initial distribution for the imbedded Markov chain of $W(\Lambda)$. To find the PGF of $W(\Lambda)$ in higher-order Markov chains, we refer to Fu and Lou (2006) for further reference.

Acknowledgments This work was supported in part by the National Science Council of Taiwan under Grant 93-2118-M-390-003 and the Natural Sciences and Engineering Research Council of Canada under the Grant A-9216.

References

Aki, S. (1992). Waiting time problems for a sequence of discrete random variables. *Annals of the Institute of Statistical Mathematics, 44*, 363–378.

Aki, S., Hirano, K. (1999). Sooner and later waiting time problems for runs in Markov dependent bivariate trials. *Annals of the Institute of Statistical Mathematics, 51*, 17–29.

Chang, Y. M. (2005). Distribution of waiting time until the rth occurrence of a compound pattern. *Statistics & Probability Letters, 75*, 29–38.

Eryilmaz, S. (2008a). Distributions of runs in a sequence of exchangeable multi-state trials. *Statistics & Probability Letters, 78*, 1505–1513.

Eryilmaz, S. (2008b). Run statistics defined on the multicolor urn model. *Journal of Applied Probability, 45*, 1007–1023.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. I). New York: Wiley.

Fu, J. C., Chang, Y. M. (2002). On probability generating functions for waiting time distributions of compound patterns in a sequence of multistate trials. *Journal of Applied Probability, 39*, 70–80.

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

Fu, J. C., Lou, W. Y. W. (2006). Waiting time distributions of simple and compound patterns in a sequence of r-th order Markov dependent multi-state trials. *Annals of the Institute of Statistical Mathematics, 58*, 291–310.

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

Hirano, K., Aki, S., Uchida, M. (1997). Distributions of numbers of success-runs until the first consecutive k successes in higher order Markov dependent trials. In N. Balakrishnan (Ed.), *Advances in combinatorial methods and applications to probability and statistics* (pp. 401–410). Boston: Birkhäuser.

Koutras, M. V. (1997). Waiting times and number of appearances of events in a sequence of discrete random variables. In N. Balakrishnan (Ed.), *Advances in combinatorial methods and applications to probability and statistics* (pp. 363–384). Boston: Birkhäuser.

Shinde, R. L., Kotwal, K. S. (2006). On the joint distribution of runs in the sequence of Markov-dependent multi-state trials. *Statistics & Probability Letters, 76*, 1065–1074.

Stanley, R. P. (1997). *Enumerative combinatorics* (Vol. I). Cambridge: Cambridge University Press.