J. Appl. Prob. 22, 228–234 (1985)
Printed in Israel

RENEWAL THEORY FOR SEVERAL PATTERNS

STEPHEN BREEN
MICHAEL S. WATERMAN* AND
NING ZHANG,* *University of Southern California*

Abstract

Discrete renewal theory is generalized to study the occurrence of a collection of patterns in random sequences, where a renewal is defined to be the occurrence of one of the patterns in the collection which does not overlap an earlier renewal. The action of restriction enzymes on DNA sequences provided motivation for this work. Related results of Guibas and Odlyzko are discussed.

DISCRETE RENEWAL THEORY; GENERATING FUNCTIONS; DNA SEQUENCES

The results presented in this paper were motivated by a study of certain patterns, known as restriction sites, in DNA sequences. DNA sequences can be thought of as finite words over a four-letter alphabet {A, T, G, C}, and restriction sites are (sequence-specific) locations of positions in double-stranded DNA where it is cut by a protein known as a restriction enzyme. See Watson (1977). For example, the restriction enzyme Hpa II cuts at occurrences of CCGG while Hha I cuts at GCGC. Such enzymes, under proper conditions, cut the DNA at all non-overlapping occurrences of the restriction sites. The enzymes are applied alone (a single digest), in batches of two (a double digest), and so on.

The goal was, then, to derive the distribution of fragment lengths when a specific set of enzymes was used to digest DNA. We assumed the DNA was an independent, identically distributed sequence of letters where the letter probabilities were p_A, p_T, p_G, p_C. A direct application of Theorem 3 and the corollary to DNA sequences appears in Waterman (1983). To simplify the current discussion we use the canonical sequence of heads and tails (H and T), and generalize the renewal theory of Feller (1968).

The earliest work we can locate which considers such problems is Leslie (1967) who treats the recurrent event $E(k, g)$ which occurs at the end of a group of k

Received 29 September 1983; revision received 13 February 1984.

Postal address for all authors: Department of Mathematics, University of Southern California, DRB 306, University Park, Los Angeles, CA 90089–1113, USA.

* Work supported by a grant from the System Development Foundation.

H's no two of which are separated by more than g T's, provided $E(k, g)$ does not occur at any of the preceding $(k-1)$ H's of the group. He does note the general problem, but does not solve it.

More recently the papers of Guibas and Odlyzko (1980), (1981) study string overlaps and generating functions. They define polynomials similar to our matching polynomials and generating functions related to but distinct from ours. Their basic concern is the first occurrence of any of the patterns, and the details of proof also differ in an essential manner. Their applications concern a coin-tossing game of Penney (1969) and an important algorithm from computer science, the Boyer–Moore (1977) string-matching algorithm. We shall close this paper with some remarks relating the results of Guibas and Odlyzko to ours.

To begin the discussion we define the matching polynomial $AB(z)$ between two patterns (finite strings) A and B. This polynomial is obtained from considering whether or not overlapping portions of the patterns match, where we proceed as follows. Let $A = \text{HTHTH}$ and $B = \text{HHTH}$.

		Match bit	Power of z	P (unmatched portion of A)
$A =$	HTHTH			
$B =$	HHTH	0	z^0	$1 = 1$
	HHTH	0	z^1	$P(\text{H}) = p$
	HHTH	1	z^2	$P(\text{TH}) = qp$
	HHTH	0	z^3	$P(\text{HTH}) = qp^2$
	HHTH	1	z^4	$P(\text{THTH}) = q^2p^2$

and

$$AB(z) = z^2pq + z^4p^2q^2.$$

Therefore if $A = \alpha_1\alpha_2\cdots\alpha_a$ and $B = \beta_1\beta_2\cdots\beta_b$, we formally define

$$AB(z) = \sum_{k=1}^{a} z^{a-k}\varepsilon_k P(\alpha_{k+1}\alpha_{k+2}\cdots\alpha_a)$$

where $\varepsilon_k = 1$ if $b \geqq k$ and $\alpha_k = \beta_b$, $\alpha_{k-1} = \beta_{b-1}, \cdots, \alpha_1 = \beta_{b-k+1}$ and $\varepsilon_k = 0$ otherwise. A convenient way to write this sum is

$$AB(z) = \sum_{k \in AB} z^{a-k}P(\alpha_{k+1}\alpha_{k+2}\cdots\alpha_a).$$

Take a collection of patterns $\{A, B, \cdots, Q\}$ such that no pattern in the collection is a substring of another. Guibas and Odlyzko refer to this as a *reduced* set of patterns. In addition, we asume that each pattern I in our set of patterns has $P(I) \neq 0$. For a given sequence $x_1x_2\cdots x_n$, we proceed from left to right marking all occurrences of members of $\{A, B, \cdots, Q\}$ which do not overlap other marked occurrences of the collection.

For example let our collection equal {TTH, THT, HH} and let

$$x = \text{TTTTHTHHHTHTTHHTT}.$$

Each time a pattern is marked, a renewal is said to occur. The renewals are named by the specific pattern occurring so that an HH renewal is said to occur at position 8 in the above sequence. Define

$$u_A(n) = P(\text{pattern } A \text{ renewal at } n),$$

$$u_A(0) = 1,$$

and

$$U_A(z) = \sum_{n=0}^{\infty} u_A(n) z^n.$$

Renewals defined in this fashion do, for each pattern, constitute a discrete renewal process according to the classical definition of Feller. We are studying the related renewal processes corresponding to $A, B, \cdots, Q$.

Now assume pattern A occurs, ending at position n ($n \geqq 1$). Then either there is an A renewal at n, or some pattern $A, B, \cdots, Q$ has had an earlier renewal that overlaps this occurrence of A. This yields the equation

$$(1)\quad P(A) = \sum_{k \in \mathcal{A}_A} u_A(n-k) P(x_{a-k+1} \cdots x_a) + \cdots + \sum_{k \in \mathcal{A}_Q} u_Q(n-k) P(x_{a-k+1} \cdots x_a).$$

Multiplying both sides by z^n ($n \geqq |A| =$ length of A),

$$P(A) z^n = \sum_{k \in \mathcal{A}_A} u_A(n-k) z^{n-k} z^k P(x_{a-k+1} \cdots x_a) + \cdots + \sum_{k \in \mathcal{A}_Q} u_Q(n-k) z^{n-k} z^k P(x_{a-k+1} \cdots x_a).$$

Summing on n, and writing an equation for each pattern,

$$\frac{z^{|A|} P(A)}{1-z} = (U_A(z) - 1) \cdot AA(z) + \cdots + (U_Q(z) - 1) \cdot AQ(z)$$

$$\cdots$$

$$\frac{z^{|Q|} P(Q)}{1-z} = (U_A(z) - 1) \cdot QA(z) + \cdots + (U_Q(z) - 1) \cdot QQ(z).$$

This system is written in matrix form by

$$(2)\quad \begin{pmatrix} AA(z) \cdots AQ(z) \\ \cdots \\ QA(z) \cdots QQ(z) \end{pmatrix} \begin{pmatrix} U_A(z) - 1 \\ \cdots \\ U_Q(z) - 1 \end{pmatrix} = \frac{1}{1-z} \begin{pmatrix} z^{|A|} P(A) \\ \cdots \\ z^{|Q|} P(Q) \end{pmatrix}.$$

If the matrix of pattern polynomials is denoted by $\Delta(z)$, note that $\Delta(0) = I$. Therefore $\det(\Delta(0)) = 1$ and $\det \Delta(z) \neq 0$ in a neighborhood of $z = 0$ so that $\Delta^{-1}(z)$ exists in this neighborhood. The system (2) then has solution

$$\begin{pmatrix} U_A(z) \\ \cdots \\ U_Q(z) \end{pmatrix} = \begin{pmatrix} 1 \\ \cdots \\ 1 \end{pmatrix} + \frac{1}{1-z}\Delta^{-1}(z)\begin{pmatrix} z^{|A|}P(A) \\ \cdots \\ z^{|Q|}P(Q) \end{pmatrix}. \tag{3}$$

Since $0 \leqq u_A(n) \leqq 1$, all $U_A(z), \cdots, U_Q(z)$ are analytic for $|z|<1$. The right-hand side of (3) is the meromorphic continuation of these analytic functions to the complex plane, having only a finite number of poles due to $1/(1-z)$ and the zeros of $\det \Delta(z)$. Note also that it follows from (3) that $\Delta^{-1}(z)K(z)$ exists for $|z|<1$, where $K(z)$ is the column matrix on the far right of (3).

The generating functions $F_A(z), \cdots, F_Q(z)$ are defined by

$$F_I(z) = \sum_{n=1}^{\infty} f_n z^n$$

where

$$f_n = P(\text{the first } I \text{ renewal at the } n\text{th trial}).$$

As noted by Feller

$$F_A(z) = 1 - \frac{1}{U_A(z)}, \cdots, F_Q(z) = 1 - \frac{1}{U_Q(z)}.$$

Then

$$F'_A(z) = \frac{U'_A(z)}{U_A^2(z)}, \cdots, F'_Q(z) = \frac{U'_Q(z)}{U_Q^2(z)},$$

where defined. If μ_A denotes the mean recurrence time of A renewals, etc., it is known to follow from the definition of $F(z)$ that $\mu_A = F'_A(1), \cdots, \mu_Q = F'_Q(1)$.

To find $F'_A(1)$ we evaluate $\lim_{z\to 1} F'_A(z)$. Since $U_A(z)$ is meromorphic, we have

$$U_A(z) = \sum_{n=-m}^{\infty} a_n (z-1)^n$$

in a neighborhood of $z=1$. First, $m \geqq 1$. If not, then $U_A(z)$ is analytic at $z=1$ and $\lim_{z\to 1}\Delta^{-1}(z)K(z) = 0$ follows from (3). However, since $K(z) = \Delta(z)\Delta^{-1}(z)K(z)$, this would yield $K(1)=0$, which is clearly a contradiction.

$$\frac{U'_A(z)}{U_A^2(z)} = \frac{\sum_{n=-m}^{\infty} n a_n (z-1)^{n-1}}{\left[\sum_{n=-m}^{\infty} a_n (z-1)^n\right]^2} = \frac{-m a_{-m}(z-1)^{-m-1} + \cdots}{a_{-m}^2 (z-1)^{-2m} + \cdots},$$

so that

$$F'_A(1) = \lim_{z\to 1} \frac{U'_A(z)}{U_A^2(z)} = \begin{cases} -(a_{-m})^{-1}, & \text{if } m = 1 \\ 0, & \text{if } m > 1. \end{cases}$$

Since $\mu_A \geqq |A| > 0$, $m = 1$ and

$$-\begin{pmatrix} a_{-1}^{A} \\ \cdots \\ a_{-1}^{O} \end{pmatrix} = \lim_{z \to 1} (1-z) \begin{pmatrix} U_A(z) \\ \cdots \\ U_O(z) \end{pmatrix} = \Delta^{-1}(1) \begin{pmatrix} P(A) \\ \cdots \\ P(Q) \end{pmatrix}. \tag{4}$$

We can summarize our results as follows.

Theorem 1. For any reduced set of patterns, $\{A, B, \cdots, Q\}$, the generating functions $U_A(z), \cdots, U_O(z)$ satisfy

$$\begin{pmatrix} AA(z) \cdots AQ(z) \\ \cdots \\ QA(z) \cdots QQ(z) \end{pmatrix} \begin{pmatrix} U_A(z) - 1 \\ \cdots \\ U_O(z) - 1 \end{pmatrix} = \frac{1}{1-z} \begin{pmatrix} z^{|A|} P(A) \\ \cdots \\ z^{|O|} P(Q) \end{pmatrix}.$$

Theorem 2. If $\Delta(z)$ equals the matching polynomial matrix in Theorem 1, $\Delta^{-1}(z)K(z)$ exists for $|z| < 1$ and for $z = 1$.

Theorem 3.

$$\begin{pmatrix} 1/\mu_A \\ \cdots \\ 1/\mu_O \end{pmatrix} = \Delta^{-1}(1) \begin{pmatrix} P(A) \\ \cdots \\ P(Q) \end{pmatrix}.$$

The equation of Theorem 3 follows from the calculation of (4). Since $u_I(n) \to 1/\mu_I$ as $n \to \infty$, the system of equations

$$\Delta(1) \begin{pmatrix} 1/\mu_A \\ \cdots \\ 1/\mu_O \end{pmatrix} = \begin{pmatrix} P(A) \\ \cdots \\ P(Q) \end{pmatrix}$$

follows easily from (1). Theorem 2 assures us the system has a solution.

If we set

$$U(z) = \sum_{n=0}^{\infty} u(n) z^n,$$

where $u(n) = P(\text{some renewal at } n\text{th trial})$, $u(0) = 1$, then

$$U(z) = \sum_{I=A}^{O} (U_I(z) - 1) + 1.$$

If μ is the mean recurrence time between renewals, then we have the following corollary.

Corollary.

$$\mu = \frac{1}{1/\mu_A + \cdots + 1/\mu_O}.$$

For an illustrative example we use the above patterns $A = \mathrm{TTH}$, $B = \mathrm{THT}$, and $C = \mathrm{HH}$. $|A| = |B| = 3$, $|C| = 2$, $P(\mathrm{H}) = p = 1 - P(\mathrm{T}) = 1 - q$.

$$\begin{pmatrix} AA(z) & AB(z) & AC(z) \\ BA(z) & BB(z) & BC(z) \\ CA(z) & CB(z) & CC(z) \end{pmatrix} = \begin{pmatrix} 1 & z^2qp & 0 \\ zq & 1+z^2pq & 0 \\ zp & 0 & 1+zp \end{pmatrix}$$

so that the above system becomes

$$\begin{pmatrix} 1 & z^2qp & 0 \\ zq & 1+z^2pq & 0 \\ zp & 0 & 1+zp \end{pmatrix} \begin{pmatrix} U_A(z)-1 \\ U_B(z)-1 \\ U_C(z)-1 \end{pmatrix} = \frac{1}{1-z} \begin{pmatrix} z^3q^2p \\ z^3q^2p \\ z^2p^2 \end{pmatrix}.$$

Solving this system, we obtain

$$\begin{pmatrix} U_A(z) \\ U_B(z) \\ U_C(z) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + (1+zp)^{-1}(1+z^2pq-z^3q^2p)^{-1} \times (1-z)^{-1} \begin{pmatrix} (1+zp)z^3q^2p \\ (1+zp)(1-zq)z^3q^2p \\ z^2p^2(1+(qp-q^2)z^2-z^3q^2p) \end{pmatrix}$$

and from (4) (left-hand equality)

$$\begin{pmatrix} 1/\mu_A \\ 1/\mu_B \\ 1/\mu_C \end{pmatrix} = \frac{1}{(1+p)(1+p^2q)} \begin{pmatrix} (1+p)q^2p \\ (1+p)q^2p^2 \\ p^2(1-q^2+p^2q) \end{pmatrix}.$$

For $p = q = 1/2$,

$$\begin{pmatrix} 1/\mu_A \\ 1/\mu_B \\ 1/\mu_C \end{pmatrix} = \begin{pmatrix} 1/9 \\ 1/18 \\ 7/54 \end{pmatrix},$$

and $\mu_A = 9$, $\mu_B = 18$, $\mu_C = 7.7\ldots$ (These values should be compared to the naive estimates of $1/P(A) = 8$, $1/P(B) = 8$, and $1/P(C) = 4$.) Finally

$$\mu = \frac{1}{1/\mu_A + 1/\mu_B + 1/\mu_C} = \frac{27}{8} = 3.375,$$

which should be compared to the naive estimate of $1/(P(A)+P(B)+P(C)) = 2$.

Finally we relate the generating functions of Guibas and Odlyzko (1981) to those of this paper. Define

$$G(z) = \sum_{n=0}^{\infty} g(n) z^{-n}$$

$$G_A(z) = \sum_{n=0}^{\infty} g_A(n) z^{-n}$$

where

$$g(n) = P(\text{none of } A, \cdots, Q \text{ occur in the first } n \text{ trials}),$$

$$g(0) = 1,$$

$$g_A(n) = P(A \text{ occurs at } n\text{th trial and none of } A, \cdots, Q \text{ occurs earlier}),$$

$$g_A(0) = 0.$$

Then

$$F(z) = G_A(z^{-1}) + G_B(z^{-1}) + \cdots + G_Q(z^{-1}),$$

while in our set-up we obtain F via $U_A, \cdots, U_Q$ through

$$F(z) = 1 - \{(U_A(z) - 1) + (U_B(z) - 1) + \cdots + (U_Q(A) - 1) + 1\}^{-1}.$$

Also

$$\begin{aligned} G(z^{-1}) = \sum_{n=0}^{\infty} g(n) z^n &= \sum_{n=0}^{\infty} \left(\sum_{k=n+1}^{\infty} f_k \right) z^k \\ &= \sum_{k=0}^{\infty} f_k \sum_{n=0}^{k-1} z^n = \sum_{k=0}^{\infty} f_k \left(\frac{1 - z^k}{1 - z} \right) \\ &= \frac{1 - F(z)}{1 - z}, \end{aligned}$$

or

$$F(z) = 1 - (1 - z) G(z^{-1}).$$

Although Guibas and Odlyzko derive properties of first passage of any pattern and our interest is in the distribution of number of trials between a pattern's occurrence, the generating functions $G(z)$ and $F(z)$ can be obtained from each other.

References

BOYER, R. S. AND MOORE, J. S. (1977) A fast string searching algorithm. *Comm. ACM* **20**, 762–772.

FELLER, W. (1968) *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edn. Wiley, New York.

GUIBAS, L. J. AND ODLYZKO, A. M. (1980) Long repetitive patterns in random sequences. *Z. Wahrscheinlichkeitsth.* **53**, 241–262.

GUIBAS, L. J. AND ODLYZKO, A. M. (1981) String overlaps, pattern matching, and nontransitive games. *J. Combinatorial Theory* A **30**, 183–208.

LESLIE, R. T. (1967) Recurrent composite events. *J. Appl. Prob.* **4**, 34–61.

PENNEY, W. (1969) Problem: penney-ante. *J. Recreational Math.* **2**, 241.

WATERMAN, M. S. (1983) Frequencies of restriction sites. *Nucleic Acids Res.* **11**, 8951–8956.

WATSON, J. D. (1977) *Molecular Biology of the Gene*, 3rd edn. W. A. Benjamin, Menlo Park, CA.