On Pattern Frequency Occurrences in a Markovian Sequence

Mireille Régnier* and Wojciech Szpankowski†
INRIA, Rocquencourt, 78153 Le Chesnay Cedex, France
Department Computer Science, Purdue University, W. Lafayette, IN 47907, U.S.A.

Consider a *given* pattern and a random text generated by a Markovian source of any order. We study the frequency of pattern occurrences in a random text when overlapping copies of the pattern are counted separately (cf. [1, 2, 3]). More precisely: let us consider two strings, a *given* pattern string $H = h_1h_2 \dots h_m$ and a random text string $T = t_1t_2 \dots t_n$ of respective lengths equal to m and n over an alphabet $\mathcal{S}$ of size V. We shall write $P(H_i^j) = \Pr\{T_{i+k}^{j+k} = H_i^j\}$ for the probability of the substring $H_i^j = h_i \dots h_j$ occurrence in the random text T_{i+k}^{j+k} between symbols $i+k$ and $j+k$ for any k. We define also the H-conditional generating function of a language $\mathcal{L}$ (i.e., a set of words satisfying certain properties) as $L_{\mathrm{H}}(z) = \sum_{w\in\mathcal{L}} P(w|w_{-m} = h_1 \cdots w_{-1} = h_m)z^{|w|}$, where w_{-i} stands for a symbol preceding the first character of w at distance i. Finally, we define the so called *autocorrelation polynomial* $A(z)$ as: *Given a string H, let $A_{\mathrm{H}}(z) = \sum_{k\epsilon HH} P(H_{k+1}^m|H_1^k)z^{m-k}$ where HH is the set of positions of H for which a prefix of H is equal to a suffix of H, e.g., $k \in HH$ means that the last k symbols of H are equal to the first k symbols of H.*

In the sequel, we denote by $O_n(\mathrm{H})$ a random variable representing the number of occurrences of H in a random text T of size n. We introduce the generating function of the language $\mathcal{T}_r$ of words that contains exactly r occurrences of H: $T^{(r)}(z) = \sum_{n\ge 0}\Pr\{O_n(\mathrm{H}) = r\}z^n$ for $|z| \le 1$. We also define a bivariate generating function as follows: $T(z,u) = \sum_{r=1}^{\infty} T^{(r)}(z)u^r = \sum_{r=1}^{\infty}\sum_{n=0}^{\infty}\Pr\{O_n(\mathrm{H}) = r\}z^n u^r$. Our main results for the Markovian model are summarized in the following two theorems proved in [4] (see [5] for extensions of these findings to approximate pattern occurrences):

Theorem 1 *Let H be a given pattern of size m, and T be a random text of length n generated according to a stationary Markov chain (of any order) over a V-ary alphabet $\mathcal{S}$.*
The generating functions $T^{(r)}(z)$ and $T(z,u)$ can be computed as follows:

$$\begin{aligned} T^{(r)}(z) &= R(z)M_{\mathrm{H}}^{r-1}(z)U_{\mathrm{H}}(z)\ , \\ T(z,u) &= R(z)\frac{u}{1-uM_{\mathrm{H}}(z)}U_{\mathrm{H}}(z)\ , \end{aligned}$$

where, after defining

$$D_H(z) = (1-z)(A_{\mathrm{H}}(z) + (P_H(H) - P(H))z^m) + z^m P(H)\ ,$$

we have

$$\begin{aligned} M_{\mathrm{H}}(z) &= 1 + \frac{z-1}{D_H(z)}\ , \\ U_{\mathrm{H}}(z) &= \frac{1-M_{\mathrm{H}}(z)}{1-z} = \frac{1}{D_H(z)}\ , \end{aligned}$$

and $R(z) = z^m P(H)U_{\mathrm{H}}(z)$. In the above, $P(H) = P(w = H)$ and $P_H(H) = P(w = H|w_{-1}^{-m} = H)$.

The above theorem is a key to the next asymptotic results. These results are derived using analytical tools.

Theorem 2 *Let the hypotheses of Theorem 1 be fulfilled, and in addition $nP(H) \to \infty$. The following results hold.*

(i) Moments. *For all $n \ge m$*

$$\begin{aligned} EO_n(H) &= P(H)(n-m+1)\ , \\ \mathrm{Var}\ O_n(H) &= nP(H)c_1 + P(H)c_2\ , \end{aligned}$$

where c_1 and c_2 are explicitly computable constants (cf. [4]).

(ii) Case $r = O(1)$. *Let ρ_H be the smallest root of $D_H(z) = 0$ outside the unit circle $|z| < 1$, and let $\rho > \rho_H$. Then:*

$$\Pr\{O_n(\mathrm{H}) = r\} \sim \sum_{j=1}^{r+1}(-1)^j a_j \binom{n}{j-1}\rho_{\mathrm{H}}^{-(n+j)}$$

where a_j are explicitly computable constants (cf. [4]).

(iii) Case $r = EO_n + x\sqrt{\mathrm{Var}\ O_n}$. *Let $x = O(1)$. Then:*

$$\Pr\{O_n(H) = r\} = \frac{1}{\sqrt{2\pi c_1 n}}e^{-\frac{1}{2}x^2}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right)\ ,$$

(iv) Case $r = (1+\delta)EO_n$. *Let $a = 1+\delta$ and $\delta \ne 0$. Define $\rho(t)$ to be the root of $1 - e^t M_{\mathrm{H}}(e^\rho) = 0$, and ω_a to be the root of $\rho'(\omega_a) = a$ Then:*

$$\Pr\{O_n(H) = r\} = \frac{1}{\omega_a\sqrt{2\pi c_1 n}}e^{-((n-m+1)I(a)}\left(1 + O\left(\frac{1}{n}\right)\right)$$

where $I(a) = a\omega_a - \rho(\omega_a)$.

References

[1] L. Guibas and A. W. Odlyzko, String Overlaps, Pattern Matching, and Nontransitive Games, *J. Combin.Theory Ser. A*, 30, 183-208, 1981.

[2] S. R. Li, A Martingale Approach to the Study of Occurrences of Sequence Patterns in Repeated Experiments, *Ann. Probab.*, 8, 1171-1176, 1980.

[3] B. Prum, F. Rodolphe, and E. Turckheim, Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequence, *J.R. Stat. Soc. B*, 57, 205-220, 1995.

[4] M. Regnier and W. Szpankowski, A Last Word on Pattern Frequency Occurrence in a Markovian Sequence?, Purdue University, CSD-TR-96-043, 1996.

[5] M. Regnier and W. Szpankowski, On the Approximate Pattern Occurrence in a Text, Purdue University, CSD-TR-96-083, 1996.

*Supported by the ESPRIT LTR Project No. 20244 (ALCOM IT), and NATO Collaborative Grant CRG.950060.
†Supported by NSF Grants NCR-9206315 and NCR-9415491, and NATO Collaborative Grant CRG.950060.