A new method of approximating the probability of matching common words in multiple random sequences

George HAIMAN and Cristian PREDA

Laboratoire de Mathématiques Paul Painlevé, UMR 8524 CNRS
Université Lille 1, Bât M2, Cité Scientifique,
F-59655 Villeneuve d'Ascq Cedex, France.

E-mail : george.haiman@upmc.fr, cristian.preda@polytech-lille.fr

The problem

- Σ – an alphabet of finite size σ. E.g. $\Sigma = \{A, T, C, G\}$.
- R independent and identicaly distributed sequences of letters over Σ of lenght T

  Model : A sequence is formed by i.i.d. letters drawn from Σ upon a probability distribution $p = (p_1, \ldots, p_\sigma)$.

We are interested in

$\mathbb{P}_R(m, T) = \mathbb{P}(\exists$ a word of lenght m common to the R sequences$)$

Interesting situations : observe a word of lenght n common to the R sequences with $n \geq m$ and $\mathbb{P}_R(m, T) \leq 0.05$.

- Karlin and Ost (Ann. Prob., 1988),

$$\mathbb{P}_R(m,T) \simeq 1 - e^{-(1-\lambda)T^R\lambda^m},$$

with $\lambda = \sum_{k=1}^{\sigma} (p_k)^R$.

- Naus and Sheng (Bull. Math. Biol., 1997),

$$\begin{aligned} N_R(m,T) = &\ \text{number of } \textit{distinct} \text{ words of length } m \\ &\ \text{common to the } R \text{ sequences} \\ \mathbb{P}_R(m,T) = &\ 1 - \mathbb{P}(N_R(m,T)=0) \simeq 1 - e^{-\hat{\mathbb{E}}(N_R(m,T))}, \end{aligned}$$

Approximation of $\mathbb{P}_R(m,T)$ **by** $\mathbb{E}(N_R(m,T))$

If p is the uniform distribution then we show that

$$\mathbb{E}(N_R) - B_R \ \leq \ \mathbb{P}_R(m,T) \ \leq \ \mathbb{E}(N_R)$$

with

$$B_R \leq \begin{cases} 0.4\mathbb{E}(N_R) & \text{if } R = 2 \\ 0.07\mathbb{E}(N_R) & \text{if } R = 3 \\ 0.05\mathbb{E}(N_R) & \text{if } R \geq 4 \end{cases}$$

Accurate approximation for small values of $\mathbb{E}(N_R)$.

Exact computation of $\mathbb{E}(N_R(m,T))$

For each word $\omega \in \Sigma^m$ of length m, let

$$\pi_T(\omega) = \mathbb{P}(\omega \text{ appears in a sequence of length T of i.i.d. letters})$$

Then,

$$\mathbb{E}(N_R(m,T)) = \sum_{\omega \in \Sigma^m} (\pi_T(\omega))^R.$$

Computation of $\pi_T(\omega)$

$$\pi_T(\omega) = \mathbb{P}(\omega \textit{ appears in } L_1, \ldots, L_T\}.$$

Robin and Daudin (1999), Rahman and Rivals (2000), Nuel (2008).

$\pi_T(\omega)$ depends on p and the *self-overlapping* structure of ω.

If p is the uniform distribution then $\pi_T(\omega)$ depends only on the *self-overlapping* structure of ω.

Self-overlapping structure: auto-correlation, period and ancestors

$m = 8, \omega = AABAABAA$.

AABAABAA	
AABAABAA	ok (1)
AABAABAA	no (0)
AABAABAA	no (0)
AABAABAA	ok (1)
AABAABAA	no (0)
AABAABAA	no (0)
AABAABAA	ok (1)
AABAABAA	ok (1)

Autocorrelation of ω : $c = 10010011$

Period : $q = 3$ (AAB). Rest : $r = 2$ (AA)

$m = 8,$

$\omega = AABAABAA :$

- Autocorrelation c=10010011.

- Period $q = 3$, Rest : $r = 2$.

(tail-overlapping)

$\omega = ABBBBBBB :$

- Autocorrelation c=10000000

- Period : $q = 8$; No rest.

(non overlapping).

$\omega = ABBABBAB :$

- Autocorrelation c=10010010.

- Period $q = 3$, Rest : $r = 2$.

(not tail-overlapping)

If $\sigma = 4$ and $m = 15$, there are $4^{15} = 1073741824$ distinct words to which corresponds a list of 57 autocorrelation vectors !

Guibas and Odlyzko (1981) : the number of autocorrelations does not depend on the size of Σ!

c	$\|c\|$	c	$\|c\|$
100000000000000	738478848	100000001000100	576
100000000000001	247205292	100000001000101	108
100000000000010	50071296	100000001000111	36
100000000000011	15710604	100000001001001	240
⋮	⋮	⋮	⋮
100000001000001	14928	101010101010101	12
100000001000010	3072	111111111111111	4
100000001000011	960		

- If p is the uniform distribution then two words having the same autocorrelation have the same probability to appears in a given random sequence.

  Then,
$$\mathbb{E}(N_R(m,T)) = \sum_{c \in \mathcal{C}} (\pi_T(c))^R \times |c|.$$

  $m = 15$, $|\mathcal{C}| = 57$.
- If p is not the uniform distribution, one needs to compute $\pi_T(\omega)$ for each ω of Σ^m.

Let consider that $T = (K+1)m$. For $k = 1, \dots, Km+1$ let

$$\nu_k = \begin{cases} 1 & \text{if } \left(L_k, \dots, L_{k+m-1}\right) = \omega \\ 0 & \text{otherwise.} \end{cases}$$

and define $X_n = X_n(\omega)$, $n = 1, \dots, K$ as

$$X_n = \max\left\{\nu_k \mid (n-1)m + 1 \leq k \leq nm\right\}$$

$\{X_n\}_n$ forms a 1-dependent stationary sequence of r.v.'s and

$$\pi_T(\omega) = 1 - \mathbb{P}(X_1 = 0, \ldots, X_K = 0).$$

Put

$\alpha_n = \alpha_n(\omega) := \mathbb{P}(X_1 = 0, \ldots, X_n = 0)$ and

$\beta_n = \beta_n(\omega) := \mathbb{P}(X_1 = 1, \ldots, X_n = 1), \ \ n \geq 1.$

Then, (see Haiman (1981)),

$$\alpha_1 = 1 - \beta_1, \ \ \alpha_2 = \alpha_1 - \beta_1 + \beta_2$$

and for $k \geq 3$,

$$\alpha_k = \alpha_{k-1} - \alpha_{k-2}\beta_1 + \alpha_{k-3}\beta_2 + \ldots + (-1)^{k+1}\beta_{k-1} + (-1)^{k+2}\beta_k.$$

Idea : Compute β_k, $k = 1, \ldots, K$, thus deducing

$$\pi_T(\omega) = 1 - \alpha_K.$$

Sequence : $\underbrace{L_1 L_2 \ldots L_m}\underbrace{L_{m+1} \ldots L_{2m}} L_{2m+1} \ldots L_{(K+1)m}$

For $s, l \in 1, ..., m$, let

$$\pi(s) = \mathbb{P}(\omega \text{ starts the first time in } L_s \text{ in } L_1 L_2 \ldots L_m)$$

and

$$\begin{aligned}\pi(s,l) = \ & \mathbb{P}(\omega \text{ starts the first time in } L_s \text{ in } L_1 L_2 \ldots L_m \\ & \text{and the first time in } L_{m+l} \text{ in } \ L_{m+1} \ldots L_{2m})\end{aligned}$$

Proposition. *Let*

$$\mu(s,l) = \mu_\omega(s,l) = \frac{\pi(s,l)}{\pi(s)}, \quad s,l = 1,\dots,m.$$

Then, for $n \geq 2$*,*

$$\beta_n = \sum_{s=1}^{m}\sum_{l=1}^{m} \pi(s)\mu^{n-1}(s,l).$$

where μ^k *is the* k*-th power of matrix* μ*.*

If ω is non overlapping then $\beta_n = (\mathbb{P}(\omega))^n C_{n+m-1}^n$.

An application related to the longest success run in Bernoulli trials.

Let $\Sigma = \{0, 1\}$, 0="failure", 1="success", and consider the word of length m, $\omega = (1, \ldots, 1)$. Let $\mathbb{P}(L_i = 1) = p$, $\mathbb{P}(L_i = 0) = 1 - p = q$, $0 < p < 1$.

Then

$$\pi(s)_\omega = \begin{cases} p^m & s = 1, \\ qp^m, & 1 < s \leq m, \end{cases}$$

and

$$\mu_\omega = \begin{pmatrix} p^m & qp^m & qp^m & \cdots & \cdots & qp^m \\ p^{m-1} & 0 & qp^m & \cdots & \cdots & qp^m \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ p^2 & 0 & \cdots & \cdots & 0 & qp^m \\ p & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

Let V(T) denote the length of the longest success run in T Bernoulli $\mathcal{B}(1,p)$ trials. The following exact formula of Bateman (1948) is of practical use only for a small T/m (see also Haiman (2007)) :

$$\mathbb{P}(V(T) \geq m) = \sum_{j=1}^{[T/m]} (-1)^{j+1} \left[p + (N - jm + 1)\frac{q}{j} \right] C_{N-jm}^{j-1} p^{jm} q^{j-1}.$$

If $T = (K+1)m$, $K \in \mathbb{N}$, we have

$$\mathbb{P}(V(T) \geq m) = 1 - \mathbb{P}\{\omega \text{ does not appear in } L_1, \ldots, L_T\} = 1 - \alpha_K$$

Some exact values for $\mathbb{P}(V(T) \geq m)$. $p = \mathbb{P}(succes)$.

m	T	p	$\mathbb{P}(V(T) \geq m)$
20	113620	0.5	0.050159
20	6620020	0.5	0.950156
20	660	2/3	0.050680
20	37420	2/3	0.950449

m	T	p	$\mathbb{P}(V(T) \geq m)$
40	116×10^9	0.5	0.050131
40	68×10^{11}	0.5	0.950953
40	1920040	2/3	0.050758
40	110800040	2/3	0.950518

Example of computation of $\mathbb{E}(N_R(m,T))$.

Let $\Sigma = \{A, B, C, D\}$

Exact values for $\mathbb{E}(N_R(m,T))$, $m = 10$.

$\mathbb{P}(A)$	$\mathbb{P}(B)$	$\mathbb{P}(C)$	$\mathbb{P}(D)$	R	T	$\mathbb{E}(N_R)$
0.25	0.25	0.25	0.25	2	240	0.050438
				3	3810	0.049635
0.23	0.27	0.23	0.27	2	240	0.062572
				3	3810	0.082672
0.20	0.30	0.20	0.30	2	240	0.085676
				3	3810	0.122567
0.10	0.40	0.10	0.40	2	240	1.548562
				3	3810	3.287721

Exact values for $\mathbb{E}(N_R(m,T))$, $m = 15$.

$\mathbb{P}(A)$	$\mathbb{P}(B)$	$\mathbb{P}(C)$	$\mathbb{P}(D)$	R	T	$\mathbb{E}(N_R)$
0.25	0.25	0.25	0.25	2	7350	0.050106
				3	375000	0.045710
0.23	0.27	0.23	0.27	2	7350	0.056843
				3	375000	0.071045
0.20	0.30	0.20	0.30	2	7350	0.067546
				3	375000	0.182311
0.10	0.40	0.10	0.40	2	7350	0.0943522
				3	375000	4.543662

Approximation of $\mathbb{P}_R(m,T)$ by $\mathbb{E}(N_R(m,T))$

Let suppose that the letters $L_1, \ldots, L_T$, $T = (K+1)m$ are drawn from an alphabet Σ of four letters and are uniformly distributed. For each $\omega \in \Sigma^m$ let

$$Z_\omega = \begin{cases} 1 & \text{if } \omega \text{ appears in all R sequences,} \\ 0 & \text{otherwise.} \end{cases}$$

We then have

$$N_R = N_R(m,T) = \sum_{\omega \in \Sigma^m} Z_\omega.$$

and

$$\mathbb{P}_R(m,T) = \mathbb{P}(N_R \geq 1) = \mathbb{P}(\sum_{\omega \in \Sigma^m} Z_\omega \geq 1).$$

$$\mathbb{P}_R(m,T) \leq \mathbb{E}(N_R) = \sum_{\omega \in W} \mathbb{P}(Z_\omega = 1).$$

We determine a $B_R = B_R(m,T) \geq 0$ such that

$$\mathbb{E}(N_R) - B_R \leq \mathbb{P}_R(m,T), \quad R \geq 2,$$

with B_R small with respect to $\mathbb{E}(N_R)$.

Let ω and ω' be two distinct words of length m and let

$$\begin{aligned}\zeta_T(\omega,\omega') &= \{\omega \text{ and } \omega' \text{ appear in } (L_1, \ldots L_T)\} \\ &= \{\exists t, t' \in \{1, \ldots, T-m+1\} \mid (L_t, \ldots, L_{t+m-1}) = \omega \\ &\quad and\ (L_{t'}, \ldots, L_{t'+m-1}) = \omega'\}\end{aligned}$$

Let

$N_R^c = \#\{$distinct couples of distinct words common to the R sequence

We then have :

$$\mathbb{E}(N_R^c) = \frac{1}{2} \times \sum_{\substack{(\omega,\omega') \in (\Sigma^m)^2 \\ \omega \neq \omega'}} \left[\mathbb{P}(\zeta_T(\omega,\omega'))\right]^R$$

and

$$\mathbb{E}(N_R^c) = \mathbb{P}(N_R = 2) + 3\mathbb{P}(N_R = 3) + \sum_{k=4}^{\infty} C_k^2 \mathbb{P}(N_R = k).$$

$$\mathbb{E}(N_R)-\mathbb{E}(N_R^c) = \mathbb{P}(N_R = 1)+\mathbb{P}(N_R = 2)+\sum_{k=4}^{\infty}(k-C_k^2)\mathbb{P}(N_R = k),$$

and since $k - C_k^2 < 0$ for $k \geq 4$, we have

$$\mathbb{E}(N_R) - \mathbb{E}(N_R^c) \leq \mathbb{P}(N_R \geq 1)$$

We show that

$$\mathbb{E}(N_R^c) \leq \begin{cases} 0.4\mathbb{E}(N_R) & \text{if } R = 2 \\ 0.07\mathbb{E}(N_R) & \text{if } R = 3 \\ 0.05\mathbb{E}(N_R) & \text{if } R \geq 4 \end{cases}$$

Numerical examples

m	R	$T = (K+1)m$	B_R	$\mathbb{E}(N_R) - B_R$	$\mathbb{E}(N_R)$	K. and O. app.
10	2	110	0.00329	0.00623	0.00953	0.00861
	2	240	0.01802	0.03241	0.05043	0.04036
	3	2240	0.00076	0.00928	0.01005	0.00953
	3	3810	0.00474	0.04489	0.04963	0.04606
15	2	3300	0.0032	0.00675	0.01005	0.00757
	2	7350	0.01760	0.03250	0.05106	0.03703
	3	226515	0.00072	0.00935	0.01007	0.00940
	3	375000	0.00388	0.04182	0.04570	0.04197

References

1. Guibas L. J., Odlyzko A. M. (1981) *Periods in Strings*, Journal of Combinatorial Theory, Series A, 30, 19–42.
2. Haiman G. (1999) *First passage time for some stationary processes*, Stochastic Processes and their Applications, 80, 231–248.
3. Haiman G. (2007) *Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of* 1*-dependent stationary sequences*, Journal of Statistical Planning and Inference, 137:3, 821–828.
4. Karlin S., Ost F. (1988) *Maximal length of common words among random sequences*, Ann. Prob.,16, p. 535–563.
5. Naus J.I., Sheng K-N (1997) *Matching among multiple random sequences*, Bulletin of Mathematical Biology, vol. 59, No. 3, p. 483–495.
6. Nuel G. (2008) *Pattern Markov chains : optimal Markov chain embedding through deterministic finite automata*. Journal of Applied Probability, 45, 226–243.
7. Robin S., Daudin J.J. (1999) *Exact distribution of word occurrences in a random sequence of random letters*, J. Appl. Prob., 36, 179–193.
8. Rahmann S., Rivals E. (2000) *Exact and efficient computation of the expected number of missing and common words in random texts*. In D. Sankoff and R. Giancarlo (eds.), Proceedings of the 11th Symposium in Combinatorial Pattern Matching (CPM2000), 375–387, Berlin, Springer-Verlag.