J. Appl. Prob. **12,** 306–315 (1975)
Printed in Israel

LONGEST COMMON SUBSEQUENCES OF TWO RANDOM SEQUENCES

VACLÁV CHVÁTAL AND
DAVID SANKOFF, *Université de Montréal*

Summary

Given two random k-ary sequences of length n, what is $f(n,k)$, the expected length of their longest common subsequence? This problem arises in the study of molecular evolution. We calculate $f(n,k)$ for all k, where $n \leqq 5$, and $f(n,2)$ where $n \leqq 10$. We study the limiting behaviour of $n^{-1}f(n,k)$ and derive upper and lower bounds on these limits for all k. Finally we estimate by Monte-Carlo methods $f(100,k)$, $f(1000,2)$ and $f(5000,2)$.

RANDOM SEQUENCES; COMMON SUBSEQUENCES; MATCHES

1. Introduction

In the study of the evolution of long molecules such as proteins or nucleic acids, it is common practice to try to construct a large set of correspondences, or matches, between two such molecules. Mathematically, this is just the problem of finding a longest common subsequence of two given finite sequences. A quadratic algorithm for doing this is available (Sankoff (1972)). It is often difficult to judge whether this set of correspondences is significantly large, i.e., contains more correspondences than one would expect in the case of two random molecules of the same length and subunit composition. Tests of significance are unavailable except on a Monte-Carlo basis (Sankoff and Cedergren (1973)), since nothing is known about the distribution of the length of the longest common subsequence. As a first step in the study of this distribution, this note investigates its mean value.

We introduce the following notation.

Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)$, $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$ be two sequences. A common subsequence, or $(\boldsymbol{a}, \boldsymbol{b})$-*match* is a set $M = \{(i_k, j_k): 1 \leqq k \leqq m\}$ with $1 \leqq i_1 < i_2 < \cdots < i_m \leqq n$, $1 \leqq j_1 < j_2 < \cdots < j_m \leqq n$ and $a_i = b_j$ for each $(i,j) \in M$. The size of a largest $(\boldsymbol{a}, \boldsymbol{b})$-match will be denoted by $v(\boldsymbol{a}, \boldsymbol{b})$. By a *k-ary sequence* we mean one whose terms come from $\{1, 2, \cdots, k\}$. We shall study the function $f(n, k)$ defined as the mean value of $v(\boldsymbol{a}, \boldsymbol{b})$ over all the k^{2n} ordered pairs $(\boldsymbol{a}, \boldsymbol{b})$ of *k-ary* sequences of length n.

Received in revised form 13 June 1974.

2. Exact formulae for $f(n,k)$ with small n

Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)$ and $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$ be two k-ary sequences. The pair $(\boldsymbol{a}, \boldsymbol{b})$ will be called *normal* if, setting $a_{n+j} = b_j$ for all j, we have $a_1 = 1$ and

$$a_j \leqq \max\{a_1, a_2, \cdots, a_{j-1}\} + 1 \qquad (2 \leqq j \leqq 2n).$$

Let $N(n,v,t)$ denote the number of normal pairs $(\boldsymbol{a}, \boldsymbol{b})$ with $v(\boldsymbol{a}, \boldsymbol{b}) = v$ and $\max\{a_1, a_2, \cdots, a_{2n}\} = t$. Clearly, the number of pairs $(\boldsymbol{c}, \boldsymbol{d})$ where $\boldsymbol{c}, \boldsymbol{d}$ are k-ary sequences of length n with $v(\boldsymbol{c}, \boldsymbol{d}) = v$ is equal to

$$\sum_{t=1}^{2n} N(n,v,t) \cdot (k)_t$$

where $(k)_t$ is the falling factorial $k(k-1)\cdots(k-t+1)$. Hence

$$\begin{aligned} f(n,k) &= \frac{1}{k^{2n}} \sum_{v=0}^{n} v \sum_{t=1}^{2n} N(n,v,t) \cdot (k)_t \\ &= \frac{1}{k^{2n}} \sum_{v=0}^{n} v \sum_{t=1}^{2n} N(n,v,t) \sum_{j=1}^{t} s(t,j) k^j \\ &= \sum_{j=1}^{2n} \sum_{t=j}^{2n} s(t,j) \sum_{v=0}^{n} v N(n,v,t) k^{j-2n} \end{aligned}$$

where $s(t,j)$ are the Stirling numbers of the first kind (Riordan (1958)). Note that $N(n,v,2n) = 0$ unless $v = 0$ and so

$$f(n,k) = \sum_{j=1}^{2n-1} \sum_{t=1}^{2n-1} s(t,j) \sum_{v=0}^{n} v N(n,v,t) k^{j-2n}.$$

Also

$$N(n,v,2n-1) = \begin{cases} n^2 & \text{if } v = 1 \\ 0 & \text{if } v > 1 \end{cases}$$

and so the coefficient of $f(n,k)$ at k^{-1} is

$$s(2n-1,2n-1) \sum_{v=1}^{n} v N(n,v,2n-1) = n^2.$$

We have evaluated $N(n,v,t)$ for $1 \leqq n \leqq 5$ and arrived at the following formulae.

$$\begin{aligned} f(1,k) &= k^{-1}, \\ f(2,k) &= 4k^{-1} - 5k^{-2} + 3k^{-3}, \\ f(3,k) &= 9k^{-1} - 27k^{-2} + 60k^{-3} - 71k^{-4} + 32k^{-5}, \\ f(4,k) &= 16k^{-1} - 84k^{-2} + 380k^{-3} - 1146k^{-4} + 2085k^{-5} - 2018k^{-6} + 771k^{-7}, \\ f(5,k) &= 25k^{-1} - 200k^{-2} + 1500k^{-3} - 8200k^{-4} + 30640k^{-5} - 75096k^{-6} \\ &\quad + 113748k^{-7} - 94790k^{-8} + 32378k^{-9}. \end{aligned}$$

The values of these functions for $1 \leqq k \leqq 15$ are given in the table below.

TABLE 1

	$f(1,k)$	$f(2,k)$	$f(3,k)$	$f(4,k)$	$f(5,k)$
$k = 1$	1.000000	2.000000	3.000000	4.000000	5.000000
2	.500000	1.125000	1.812500	2.523438	3.246094
3	.333333	.888889	1.477366	2.090535	2.718742
4	.250000	.734375	1.253906	1.801453	2.363899
5	.200000	.624000	1.096640	1.594317	2.108546
6	.166667	.541667	.977109	1.435968	1.912269
7	.142857	.478134	.881954	1.309838	1.754954
8	.125000	.427734	.803955	1.206201	1.625155
9	.111111	.386831	.738692	1.119008	1.515694
10	.100000	.353000	.683220	1.044309	1.421763
11	.090909	.324568	.635470	.979404	1.340005
12	.083333	.300347	.593927	.922366	1.267999
13	.076923	.279472	.557455	.871776	1.203953
14	.071429	.261297	.525179	.826554	1.146514
15	.066667	.245333	.496417	.785862	1.094633

Moreover, we have evaluated $f(n,2)$ for all $n = 1, 2, \cdots, 10$; the results are given in Table 2 in proportion to n.

TABLE 2

n	$f(n,2)/n$
1	0.500000
2	0.562500
3	0.604167
4	0.630859
5	0.649219
6	0.663330
7	0.674491
8	0.683640
9	0.691303
10	0.697844

3. Limiting behaviour of $f(n,k)$

Klarner and Rivest (personal communication) have observed that $f(n,k)$ is superadditive with respect to n, that is, $f(n_1+n_2,k) \geqq f(n_1,k)+f(n_2,k)$. Thus, by Fekete's theorem (Fekete (1923)),

$$\lim_{n\to\infty} n^{-1}f(n,k) = \sup_n n^{-1}f(n,k). \tag{1}$$

We shall denote the common value of (1) by c_k. Klarner and Rivest asked whether $c_2 = 1$; we shall show that this is not the case.

Lemma 1. Let $g(c,n,k)$ denote the number of pairs $(\boldsymbol{a},\boldsymbol{b})$ of k-ary sequences of length n with $v(\boldsymbol{a},\boldsymbol{b}) \geqq cn$. If

$$(ck^{\frac{1}{2}})^c(1-c)^{1-c} \geqq 1$$

then $g(c,n,k) = o(k^{2n})$.

Proof. Let $G(c,n,k)$ denote the number of ordered triples $(\boldsymbol{a},\boldsymbol{b},M)$ where $\boldsymbol{a},\boldsymbol{b}$ are k-ary sequences of length n and $M = \{(i_k,j_k): 1 \leqq k \leqq m\}$ is an $(\boldsymbol{a},\boldsymbol{b})$-match with $m = [\![cn]\!]$, so that $m \leqq v(\boldsymbol{a},\boldsymbol{b})$. There are exactly $\binom{n}{m}^2$ ways of selecting i_k's and j_k's with $1 \leqq i_1 < i_2 < \cdots < i_m \leqq n$ and $1 \leqq j_1 < j_2 < \cdots < j_m \leqq n$; once this is done, there are exactly k^{2n-m} appropriate choices of $(\boldsymbol{a},\boldsymbol{b})$. Hence

$$g(c,n,k) \leqq G(c,n,k) = \binom{n}{m}^2 k^{2n-m}$$

since all pairs $(\boldsymbol{a},\boldsymbol{b})$ counted in $g(c,n,k)$ must have at least one $(\boldsymbol{a},\boldsymbol{b})$-match of size m. By Stirling's formula, we have

$$\binom{n}{m}^2 \sim \frac{n}{2\pi m(n-m)} \cdot \left[\frac{n^n}{m^m(n-m)^{n-m}}\right]^2.$$

Now,

$$\frac{n}{2\pi m(n-m)} \sim \frac{1}{2\pi c(1-c)n} = \frac{O(1)}{n}$$

and

$$\frac{n^n}{m^m(n-m)^{n-m}} = \left[\frac{1}{c^c(1-c)^{1-c}}\right]^n \left[\frac{cn}{m}\right]^{cn} \left[\frac{n-cn}{n-m}\right]^{n-cn} \left[\frac{m}{n-m}\right]^{cn-m}$$

$$= (c^c(1-c)^{1-c})^{-n} \cdot O(1).$$

Hence

$$G(c,n,k) \sim k^{2n-m} \cdot \frac{O(1)}{n} \cdot (c^c(1-c)^{1-c})^{-2n}$$

$$= O(1) \cdot \frac{k^{2n}}{n}((ck^{\frac{1}{2}})^c(1-c)^{1-c})^{-2n} = o(k^{2n})$$

which completes the proof.

Now, for each integer k and for each x with $k \geqq 2$, $0 < x < 1$, set

$$h_k(x) = k^{x/2} x^x (1-x)^{1-x}.$$

Note that $\lim_{x \to 0} h_k(x) = 1$, $\lim_{x \to 1} h_k(x) = k^{\frac{1}{2}}$ and

$$\frac{d}{dx} h_k(x) = h_k(x) \cdot \log \left[\frac{x k^{\frac{1}{2}}}{1-x} \right].$$

Hence, for each k, there is a unique solution of

$$0 < x < 1, \quad h_k(x) = 1.$$

Denote this solution by y_k. Values of y_k with $2 \leqq k \leqq 15$ are shown in the following table, to six-decimal accuracy.

TABLE 3

k	y_k
2	0.905118
3	0.829982
4	0.772908
5	0.727666
6	0.690556
7	0.659318
8	0.632493
9	0.609090
10	0.588410
11	0.569942
12	0.553304
13	0.538199
14	0.524397
15	0.511713

Theorem 1. If $k \geqq 2$ then $c_k \leqq y_k$.

Proof. By Lemma 1, we have $g(y_k, n, k) = o(k^{2n})$ and so

$$\frac{f(n,k)}{n} \leqq \frac{1}{nk^{2n}} (g(y_k, n, k)n + (k^{2n} - g(y_k, n, k)) y_k n)$$

$$= y_k + o(1).$$

Note that $\lim_{k \to \infty} y_k = 0$ and so $\lim_{k \to \infty} c_k = 0$.

4. Lower bounds on c_k

For each pair $(\boldsymbol{a}, \boldsymbol{b})$ of k-ary sequences of length n, we shall construct a certain $(\boldsymbol{a}, \boldsymbol{b})$-match M of size $v'(\boldsymbol{a}, \boldsymbol{b})$ and show that $f'(n,k)$, the average of $v'(\boldsymbol{a}, \boldsymbol{b})$ over all k^{2n} ordered pairs $(\boldsymbol{a}, \boldsymbol{b})$, satisfies

$$\lim_{n\to\infty} n^{-1} f'(n,k) = 2k^2/(k^3 + 2k - 1). \tag{2}$$

The construction of M is described below. The main idea is to begin by looking for the 'first' matching pair (a_i, b_j) where $i = 1$ or $j = 1$. For example, suppose we examine the pairs $(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_1, b_3)$ and finally find the first matching pair, namely (a_3, b_1). Then we include (a_3, b_1) in M and proceed to look for the 'first' matching pair in the sequences $a_4, a_5, \cdots, a_n$ and $b_3, b_4, \cdots, b_n$. We continue until one or both sequences are exhausted.

Step 0. Let $\alpha_i = a_i$, $\beta_i = b_i$ and $S(i) = T(i) = i$ for all $i = 1, 2, \cdots, n$. Let FLAG $= 1$ and $M = \varnothing$.

Step 1. If FLAG $= 1$, check successively

$$(\alpha_1, \beta_1), (\alpha_1, \beta_2), (\alpha_2, \beta_1), \cdots, (\alpha_1, \beta_d), (\alpha_d, \beta_1), \cdots$$

until α or β is exhausted or until we find a pair with $\alpha_i = \beta_j$. If FLAG $= -1$, check the pairs in the order

$$(\alpha_1, \beta_1), (\alpha_2, \beta_1), (\alpha_1, \beta_2), \cdots, (\alpha_d, \beta_1), (\alpha_1, \beta_d), \cdots.$$

In the case of exhaustion, stop; otherwise add the pair $(S(i), T(j))$ to M.

Step 2. Note that $i = 1$ or $j = 1$ or both.
If $i \leqq 2$ and $j \leqq 2$, set

$$i' = i + 1,\ j' = j + 1.$$

If $i = 1$ and $j \geqq 3$, set

$$i' = \begin{cases} j-1 & (\text{FLAG} = 1) \\ j & (\text{FLAG} = -1) \end{cases},\ j' = j + 1.$$

If $i \geqq 3$ and $j = 1$, set

$$i' = i + 1,\ j' = \begin{cases} i & (\text{FLAG} = 1) \\ i-1 & (\text{FLAG} = -1) \end{cases}.$$

Step 3. Let $p = S(i') - 1$, $q = T(j') - 1$ and redefine

$$S(i) = p + i, \quad \alpha_i = a_{S(i)}$$
$$T(j) = q + j, \quad \beta_j = b_{T(j)}$$

for all i,j with $1 \leqq i \leqq n-p$, $1 \leqq j \leqq n-q$. Reverse the sign of FLAG and go to Step 1.

Lemma 2. For infinite sequences $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$, we have

$$E(i'+j'-2) = \frac{k^3+2k-1}{k^2}$$

where i', j' are defined as in the preceding algorithm and $E(\cdot)$ denotes mathematical expectation.

Proof. Consider the sequence of pairs in case FLAG = 1, that is,

$$(\alpha_1,\beta_1),(\alpha_1,\beta_2),(\alpha_2,\beta_1),\cdots,(\alpha_1,\beta_d),(\alpha_d,\beta_1),\cdots.$$

The event that any of these pairs contains equal terms has probability $1/k$ and this is also the conditional probability given any or all the preceding pairs. Hence the probability that the rth pair will be the first equal one is $(k-1)^{r-1}/k^r$. Now,

$$i'+j'-2 = \begin{cases} 2 & \text{if } r = 1, \\ 3 & \text{if } r = 2, \\ r & \text{if } r \geqq 3. \end{cases}$$

Therefore

$$E(i'+j'-2) = 2\cdot\frac{1}{k} + 3\cdot\frac{k-1}{k^2} + \sum_{r=3}^{\infty} r\frac{(k-1)^r}{k^r} = \frac{k^3+2k-1}{k^2}.$$

The same can be shown for case FLAG = -1.

Theorem 2. For all k, we have $c_k \geqq \dfrac{2k^2}{k^3+2k-1}$.

Proof. Obviously, it will suffice to prove (2). Let $X_1, X_2, \cdots$ be successive values of $i'+j'-2$ found by the algorithm when applied to the infinite sequences $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$. It is clear that the X_i's are independent, identically distributed random variables (indeed, in each cycle, equality or inequality of pairs is independent of all previous cycles). Let

$$x_k = \frac{2k^2}{k^3+2k-1}.$$

The symmetry ensured by the alternation of sign of FLAG ensures that after $w = 2u$ cycles of the algorithm, the total number p (resp. q) of the a^*_i's (resp. b^*_j's) that have been used up satisfies

$$E(p) = E(q) = \tfrac{1}{2}wE(i'+j'-2) = w/x_k.$$

Furthermore,

$$\Pr\left[\left|\frac{p}{w} - \frac{1}{x_k}\right| > \varepsilon\right] = \Pr\left[\left|\frac{q}{w} - \frac{1}{x_k}\right| > \varepsilon\right] = o(1)$$

by the law of large numbers. Now a pair $(\boldsymbol{a}, \boldsymbol{b})$ of random sequences of length n can be considered as being the first n terms of $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$. If the algorithm (applied to $\boldsymbol{a}, \boldsymbol{b}$) halts during the $(w+1)$th cycle then the first w cycles are the same as the first w cycles of the algorithm applied to $\boldsymbol{a}^*$ and $\boldsymbol{b}^*$. Now, after $\| nx_k \|$ cycles of the algorithm applied to $\boldsymbol{a}^*, \boldsymbol{b}^*$, we have

$$\Pr(p > n(1+\varepsilon) \text{ or } p < n(1+\varepsilon)) = \Pr\left[\left| \frac{p}{[\![nx_k]\!]} - \frac{1}{x_k} \right| > \varepsilon \right]$$

$$= o(1)$$

and so

$$\Pr(n(1-\varepsilon) \leqq p \leqq n(1+\varepsilon) \text{ and } n(1-\varepsilon) \leqq q \leqq n(1+\varepsilon)) = 1-o(1).$$

Hence with probability $1-o(1)$, at least $[\![nx_k]\!] - n\varepsilon$ and at most $[\![nx_k]\!] + n\varepsilon$ cycles of the algorithm (applied to $\boldsymbol{a}^*, \boldsymbol{b}^*$) operate within $\boldsymbol{a}$ and $\boldsymbol{b}$ since $n\varepsilon$ successive terms in a sequence can give rise to at most $n\varepsilon$ cycles of the algorithm. Equivalently,

$$\Pr(| v'(\boldsymbol{a}, \boldsymbol{b}) - [\![nx_k]\!] | \leqq n\varepsilon) = 1-o(1)$$

and so $\lim_{n\to\infty} n^{-1} f'(n,k) = x_k$.

Values of x_k with $2 \leqq k \leqq 15$ are given in the table below.

TABLE 4

k	x_k
2	0.727273
3	0.562500
4	0.450704
5	0.373134
6	0.317181
7	0.275281
8	0.242884
9	0.217158
10	0.196271
11	0.178994
12	0.164477
13	0.152115
14	0.141465
15	0.132197

5. Monte-Carlo estimates for $f(100, k)$ and c_2

To obtain further information about c_k, we carried out two series of Monte-Carlo simulations. First, for $n = 100$ and for each $k = 2, \cdots, 15$, we generated

100 pairs $(\boldsymbol{a}, \boldsymbol{b})$ of random k-ary sequences and calculated $v(\boldsymbol{a}, \boldsymbol{b})$ in each case. We denote by $m_{k,n}$ the average value of $n^{-1}v(\boldsymbol{a}, \boldsymbol{b})$ in a given sample. For large n, this quantity may be considered an estimate of c_k. Values of $m_{k,100}$ are tabulated in Table 5, and may be compared with the upper and lower bounds in Tables 2 and 4. Table 5 also contains $s_{k,100}$, where

$$s_{k,n}^2 = \sum_{\substack{(\boldsymbol{a},\boldsymbol{b}) \\ \text{in} \\ \text{sample}}} (n^{-1}v(\boldsymbol{a}, \boldsymbol{b}) - m_{k,n})^2/(\text{sample size} - 1)$$

is an unbiased estimator of the variance of $n^{-1}v(\boldsymbol{a}, \boldsymbol{b})$.

TABLE 5

k	$m_{k,100}$	$s_{k,100}$
2	0.7814	0.0243
3	0.6855	0.0210
4	0.6242	0.0176
5	0.5778	0.0211
6	0.5332	0.0208
7	0.5065	0.0214
8	0.4812	0.0219
9	0.4593	0.0211
10	0.4423	0.0208
11	0.4268	0.0200
12	0.4126	0.0193
13	0.4003	0.0212
14	0.3827	0.0212
15	0.3712	0.0198

To estimate c_2 more closely, a second series of simulations was carried out for $k = 2$ and $n = 10$, 100, 1000, and 5000. Table 6 lists $m_{2,n}$ and $s_{2,n}$, as well as the size of the sample used to make these estimates.

TABLE 6

n	$m_{k,n}$	$s_{k,n}$	sample size
10	0.6991	0.1079	1000
100	0.7806	0.0238	100
1000	0.80529	0.00468	100
5000	0.8082	0.0015	6

On the basis of these simulations, it seems fair to conjecture that $c_2 > 4/5$ and that the variance of $v(\boldsymbol{a}, \boldsymbol{b})$ is $o(n^{2/3})$.

Note added in proof

The bounds y_k in Table 3 can be improved by a refinement of the argument in Lemma 1. This depends on the observation that, given a k-ary sequence $\boldsymbol{a}$ of length m, the number of k-ary sequences of length $n \geqq m$ which contain $\boldsymbol{a}$ as a subsequence is a function only of m, n and k, and is independent of the structure of $\boldsymbol{a}$. The new upper bound z_k is the unique solution in $(1/k, 1)$ of

$$k^{1-x/2}(k-1)^{x-1}x^x(1-x)^{1-x} = 1.$$

For example, $z_2 = 0.866595$.

References

[1] FEKETE, M. (1923) Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Math. Z.* **17**, 228–249.

[2] RIORDAN, J. (1958) *An Introduction to Combinatorial Analysis.* John Wiley, New York.

[3] SANKOFF, D. (1972) Matching sequences under deletion/insertion constraints. *Proc. Nat. Acad. Sci. U.S.A.* **69**, 4–6.

[4] SANKOFF, D. AND CEDERGREN, R. J. (1973) A test for nucleotide sequence homology. *J. Mol. Biol.* **77**, 159–164.