PROBABILITY OF COMMON SUBSEQUENCE LENGTH

BY

JOSEPH G. DEKEN

TECHNICAL REPORT NO. 151
DECEMBER 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 78-07736

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PROBABILISTIC BEHAVIOR OF COMMON SUBSEQUENCE LENGTH

by

JOSEPH G. DEKEN

TECHNICAL REPORT NO. 151
DECEMBER 1979

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT
MCS 78-07736

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also issued as Technical Report No. 139 Under National Science Foundation Grant MCS77-16974, Department of Statistics, Stanford University.
PROBABILISTIC BEHAVIOR OF COMMON SUBSEQUENCE LENGTH

By

Joseph G. Deke

Abstract

Upper and lower bounds are given for the asymptotic ratio of longest common subsequence length to total sequence length, for pairs of sequences of independent letters drawn uniformly from a k-ary alphabet. The method for upper bounds, based on a refinement of overcounting, is presented in detail.
PROBABILISTIC BEHAVIOR OF COMMON SUBSEQUENCE LENGTH

By

Joseph G. Deken

1. Introduction.

The longest common subsequence length of two sequences of letters from a k-ary alphabet is defined to be

\[ \ell (\mathcal{a}, \mathcal{b}) = \max \left\{ \ell (i_1, j_1, i_2, j_2)^p \mid \forall \alpha, \beta = 1, 2 (i_{\alpha, j} > i_{\beta, j-1}, a_{i_{\alpha, j}} = b_{i_{\beta, j}}) \right\} \]

In words, \( \ell (\mathcal{a}, \mathcal{b}) \), for two sequences of length \( n \), is \( n \) minus the minimum number of letters which would have to be deleted (from each sequence, perhaps at different positions) in order for the sequences of letters remaining on \( \mathcal{a} \) and \( \mathcal{b} \) to be identical. The common subsequence length furnishes a natural metric for sequence comparison in a variety of biological and other applications. (Sankoff, Cedergren, Lapalme (1976), Hirschberg (1977)). Natural candidate sequences are amino acids (20 letter alphabet) and nucleotides (4 letter alphabet).

In comparing two sequences by this metric, it is of interest to judge their closeness relative to that of sequences which are in some sense random. There are many possible models for random sequences. One may suppose that all letters appear independently on both sequences, and have equal probability, or that they appear independently but perhaps with different probabilities for different letters. Alternatively there may be a fixed set of letters for each sequence, and the observed sequences obtained by permuting these letters randomly.
For a fairly broad class of models for random sequences (c.f. Deken (1979)), it can be shown that the length of a longest common subsequence, divided by the total sequence length, approaches a constant as the lengths of the random sequences become large. In particular, this fractional length, averaged over all random sequences will approach some constant \( c_k \) which is a function of the random model used and the number \( k \) of letters in the alphabet. The constants \( c_k \) are difficult to determine except in trivial cases, and the only results available for finite alphabets are upper and lower bounds.

Lower bounds for the constants \( c_k \) may be produced by constructing an algorithm which takes two sequences as input and produces a common subsequence as output. If the algorithm is simple enough its probabilistic behavior can be determined, and since the common subsequence produced is by definition no longer than the longest common subsequence, the limit obtained furnishes a lower bound for \( c_k \). Upper bounds may be derived by counting methods, since the probability that the two sequences contain any particular subsequence is usually easily computed. Examples of these upper and lower bound techniques may be found in the papers by Chvatal and Sankoff (1975) and Deken (1979).

Table 1 below gives upper and lower bounds for the constants \( c_k \) derived by these methods, when the sequences are completely independent, with all letters equally likely. The corresponding lower bound algorithms have been described in detail (c.f. Deken (1979)) elsewhere, and will not be restated here. The counting method used to derive the upper bounds is of some independent interest, and so is described below. The bounds in the table are believed to be temporarily the best known.
Table of Bounds for $c_k$

<table>
<thead>
<tr>
<th>Alphabet Size</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.7615</td>
<td>.8602(.8575)</td>
</tr>
<tr>
<td>3</td>
<td>.6153</td>
<td>.7769</td>
</tr>
<tr>
<td>4</td>
<td>.5454</td>
<td>.7181</td>
</tr>
<tr>
<td>5</td>
<td>.5061</td>
<td>.6733</td>
</tr>
<tr>
<td>6</td>
<td>.4716</td>
<td>.6373</td>
</tr>
<tr>
<td>7</td>
<td>.4450</td>
<td>.6074</td>
</tr>
<tr>
<td>8</td>
<td>.4223</td>
<td>.5819</td>
</tr>
<tr>
<td>9</td>
<td>.4032</td>
<td>.5599</td>
</tr>
<tr>
<td>10</td>
<td>.3965</td>
<td>.5406</td>
</tr>
<tr>
<td>11</td>
<td>.3719</td>
<td>.5234</td>
</tr>
<tr>
<td>12</td>
<td>.3589</td>
<td>.5079</td>
</tr>
<tr>
<td>13</td>
<td>.3473</td>
<td>.4942</td>
</tr>
<tr>
<td>14</td>
<td>.3368</td>
<td>.4811</td>
</tr>
<tr>
<td>15</td>
<td>.3273</td>
<td>.4650</td>
</tr>
</tbody>
</table>
2. **Upper Bounds.**

Upper bounds for the constants $c_k$ are derived here by an overcounting method: Suppose we wish to show that $c_k < p$. We compute the number $N(p,n)$ of pairs of sequences of length $n$ containing in common a particular sequence $s$ of length $[pn]$. The total number of pairs of sequences containing a common subsequence of length greater than or equal to $[pn]$ is not greater than $c(p,n) := k^{[pn]} N(p,n)$, and if the ratio $\frac{c(p,n)}{k^{2n}}$ approaches 0 as $n$ increases, then $c_k \leq p$. This method is essentially that used by Chvatal and Sankoff (1975). Our efforts here are focused on reducing the overcounting caused by the fact that in general, one pair of sequences may contain many common subsequences of length $[pn]$.

Consider a sequence $\underline{s} = (s_1, s_2, \ldots, s_{[pn]})$. The probability that a sequence of length $n$ contains $\underline{s}$ may be expressed as the probability that the sum of "waiting times" to collect $[pn]$ digits on an infinite random sequence does not exceed $n \approx \frac{[pn]}{p}$. Since the expected waiting time to collect a digit is $k$, and the values of $p$ of interest will be $> 1/k$, this probability may be bounded by an argument due to Chernoff (1952). Squaring this bound gives an upper bound to the probability that two independent sequences both contain $\underline{s}$.

It is convenient to think of examining the pair of sequences left to right to determine if they contain the subsequence $\underline{s}$. For $i = 1, 2, \ldots, [pn]$, we scan the "top" sequence until the desired digit $s_i$ of $\underline{s}$ appears, then switch to the "bottom" sequence and search for $s_i$ there. A particular pair of sequences is "marked" by the subsequence $\underline{s}$ if this process can be completed without searching more than $n$ digits.
on either sequence. Suppose however that in collecting \( s_i \) we examine a segment \( v \) of \( n_1 > 0 \) digits different from \( s_1 \) on the top sequence and a segment \( v' \) of \( n_2 > 0 \) digits different from \( s_1 \) on the bottom sequence. In this case, there may be a digit \( r \neq s_1 \) contained by both \( v \) and \( v' \) (e.g. in the binary alphabet, there must be a match). In this case, we will not mark the pair of sequences by \( s \), since if both sequences contain \( s \), they also contain the sequence \( s' \) formed by inserting the digit \( r \) before \( s_1 \). In a natural way, we say that the sequence \( s \) is "preceded at \( i \)" by \( s' \). The probability that a sequence is not preceded is the product of the probabilities that it is not preceded at \( i = 1, 2, \ldots, [pn] \), i.e. \( m(k)^{[pn]} \), where \( m(k) \) is the probability (as a function of alphabet size \( k \)) that no match occurs on initial segments of a random sequence before the appearance of a "1" on both sequences. New upper bounds are thus obtained by bounding the probability \( c'(k,p,n) \) that the sum of the collection times for \( s \) does not exceed \( n \) on either the top or bottom sequence and \( s \) is not preceded. Functions \( m(k,p) \) are derived below such that \( c'(k,p,n) \leq m(k,p)^{[pn]} \). This gives the bounds

\[
c_k \leq \inf\{p : \lim_{n \to \infty} k^{[pn]} c'(k,p,n) = 0\} \leq \inf\{p : k m(k,p) < 1\}.
\]

3. Upper Bounds for \( c'(k,p,n) \).

In order to use Chernoff's inequality, we compute the generating function

\[
\varphi_k(t) := \sum_{i=1}^{\infty} p_i(k) t^{i+2},
\]

\[5\]
where \( P_i(k) \) is the probability that \( i \) digits in total occur on top and bottom sequences before the occurrence of "1" on each sequence, and no other digit is common to the two initial segments before the "1"'s. In fact, we have the following formula:

\[
\varphi_k(t) = \sum_{\alpha=0}^{k-1} \binom{k-1}{\alpha} \sum_{\beta=0}^{\alpha} \binom{\alpha}{\beta} (-1)^{\beta} \frac{t^2}{(k-t(\alpha-\beta))(k-t(k-1-\alpha))}.
\]

In probabilistic terms, \( \frac{\varphi_k(e^t)}{\varphi_k(1)} \) is the moment generating function of the sum of the number of digits in the two initial segments, including the "1"'s, conditional on no other matches, so that the desired inequality

\[
c'(k,p,n) \leq m(k,p)[pn], \text{ with } m(k,p) := \inf_{t>0} \{t^{-2/p} \varphi_k(t)\}
\]

Chernoff's inequality and the fact that the maximum of two positive numbers is less than \( n \) if their sum is less than \( 2n \).


To further improve the upper bounds, we consider blocks of digits of length \( b \). A subsequence \( s \) of length \( \lfloor cn \rfloor \) is then considered to be made up of \( \lfloor \frac{cn}{b} \rfloor \) blocks. Precedence rules are formulated so that \( s \) is not preceded if and only if it is not preceded at any block \( i = 1, 2, \ldots, \lfloor \frac{cn}{b} \rfloor \). A sequence \( A \) of length \( n \) is marked by \( s \) if and only if \( s \) is not preceded on \( A \) and the total number of collections (summed over top and bottom sequences and all blocks) for \( s \) does not exceed \( 2n \).

The upper bounds given above were obtained by numerical solution of the equation.
\[ k \ m(k,p) = 1 , \]

and area accurate to four decimal places. The additional upper bound (.8575) for the binary case was obtained by utilizing an extended precedence rule with blocksize \( b = 8 \). Increasing \( b \) causes a geometric increase in the number of terms to be calculated for the moment generating function, and the improvement in the bound seems to be small.
Bibliography


