Simulations, Computations, and Statistics for Longest Common Subsequences

Qingqing Liu∗‡
Christian Houdré†

May 22, 2017

Abstract

The length of the longest common subsequences (LCSs) is often used as a similarity measurement to compare two (or more) random words. Below we study its statistical behavior in mean and variance using a Monte-Carlo approach from which we then develop a hypothesis testing method for sequences similarity. Finally, theoretical upper bounds are obtained for the Chvátal–Sankoff constant of multiple sequences.

1 Introduction

The study of sequences alignments and comparisons is an important problem in bioinformatics and computer science, where a fundamental issue is to compare two or more sequences and to assess the significance of their similarity or dissimilarity. Within this framework, a general methodology is first to find an optimal alignment of the sequences and then to compute its score. Afterwards, some knowledge of the statistics of the alignment score allows to test hypotheses to tell whether or not the similarity is significant.

To formalize our discussion, let us introduce our framework. Following [10], let \( A \) be a finite alphabet and let \(-\) represent a gap symbol. Let \( \Sigma \) to be the set of non-empty sequences of \( A \), i.e., \( \Sigma = \bigcup_{n \geq 0} A^n \), where \( A^0 = \emptyset \) is the empty string. A sequence \( x \in A^n \) has length \( n \), denoted by \( |x| = n \). Given two sequences \( a = (a_1, \ldots, a_n) \) and \( b = (b_1, \ldots, b_m) \) \( \in \Sigma \), we say that a pair of sequence \( a^\diamond, b^\diamond \in \bigcup_{n \geq 0} (A \cup \{-\})^n \) is an alignment of \( (a, b) \), if the following three conditions are satisfied: (i) \( |a^\diamond| = |b^\diamond| \), (ii) \( a^\diamond_i \neq - \) or \( b^\diamond_i \neq - \) for \( i = 1, \ldots, |a^\diamond| \), i.e., no two gaps are aligned, and (iii) \( a^\diamond|_A = a \) and \( b^\diamond|_A = b \), i.e., the restrictions of \( a^\diamond \) and \( b^\diamond \) to symbols in \( A \) give respectively \( a \) and \( b \).

To measure the similarity of two sequences, assign a score to each alignment and take the score of the best alignment (i.e., with the highest score) as the similarity score of the two sequences. To define an alignment score, we need a score function \( s : A \times A \to \mathbb{R} \), and a gap penalty function \( g : \mathbb{N} \to \mathbb{R} \) which is assumed to be subadditive, i.e.,

\[
\forall k, l : g(k + l) \leq g(k) + g(l).
\]

Given a sequence \( u \in \bigcup_{n \geq 0} (A \cup \{-\})^n \), we say that \( u \) contains a gap of length \( k \) at position \( i \) if \( (u_{i}, \ldots, u_{i+k-1}) \in \bigcup_{n \geq 0} \{-\}^n \), and there is no other subsequence of \( u \) extending \( (u_{i}, \ldots, u_{i+k-1}) \) that is composed uniquely of ‘-’s. Then still following [10], \( \Delta_k(u) \) is defined to be the number of different

---

∗School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160 USA, qqliu@gatech.edu
†School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160 USA, houdre@math.gatech.edu. Research supported in part by the grant #246283 from the Simons Foundation.

Keywords: Longest Common Subsequences, Monte-Carlo Simulation, Hypothesis Testing, Longest Common Increasing Subsequences

MSC 2010: 65C05, 62F03, 60C05, 05A05
gaps of \( u \) having length \( k \), and the score of the alignment is defined as

\[
s(a^*, b^*) = \sum_{1 \leq i \leq |a^*|} \sum_{a^*_i \neq b^*_i} s(a^*_i, b^*_i) - \sum_{1 \leq k \leq |a^*|} \Delta_k(a^*) g(k) - \sum_{1 \leq k \leq |a^*|} \Delta_k(b^*) g(k).
\]

Two types of alignments are commonly used in sequences comparisons, global and local alignment. While a local alignment looks for the segments with best matching scores, the global alignment score corresponds to having as many letter matched as possible in each sequence.

Although the statistics (mean, variance, distribution, etc.) of local alignment scores are well studied [1, 7], there is still much unknown about the statistics of global alignment scores. One of the most analyzed global alignment statistics is the length of the longest common subsequences (LCSs), which is the score of the optimal alignment using the score function

\[
s(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b, \end{cases}
\]

and a zero gap function, i.e., \( g(k) = 0 \) for all \( k \in \mathbb{N} \). Next, given two strings \( a = (a_1, \ldots, a_n) \) and \( b = (b_1, \ldots, b_m) \), a sequence \( c = (c_1, \ldots, c_l) \) is called a common subsequence of \( a \) and \( b \) if there exist indices \( 1 \leq i_1 < i_2 < \cdots < i_l \leq n \) and \( 1 \leq j_1 < j_2 < \cdots < j_l \leq m \) such that \( c_k = a_{i_k} = b_{j_k} \) for \( k = 1, \ldots, l \). Then, the length of the LCS of \( a \) and \( b \) is \( LCS(a, b) = \max\{|c| : c \text{ is a common subsequence of } a \text{ and } b\} \), and we also use LCS to also represent the length of the common subsequences. This definition can be naturally extended to the case of three or more sequences, and when the sequences have same length \( n \), we denote it by \( LC_N \). In the present text, we will only consider the LCSs of sequences of the same length unless otherwise specified.

As far as this paper’s content is concerned, we start by summarizing previous studies on the mean behavior of LCS, some notable LCS algorithms and previous work on Monte-Carlo simulation of LCSs. We then estimate the variance of the length of LCS of two binary random words using Monte-Carlo experiments (Section 3). Based on these results, and on some recent advances on its limiting distribution [17], we build a hypothesis testing method to test whether two sequences are significantly similar or not (Section 4.1) and conduct extensive Monte-Carlo experiments to determine the parameters of the test (Section 4.2 to 4.2). Finally, we extend a classical result of [9] valid for two sequences to an arbitrary finite number of sequences (Section 5) and thus obtain new theoretical upper bounds on the Chvátal and Sankoff constant in that context.
Table 1: Theoretical Bounds for $\gamma_2^2$

|                | lower bound | upper bound |
|----------------|-------------|-------------|
| Chvátal and Sankoff [8, 9] | 0.727273    | 0.86660     |
| Deken [12, 13]                | 0.7615      | 0.8575      |
| Dančík [11]                  | 0.773911    | 0.837623    |
| Lueker [26]                  | 0.788071    | 0.826280    |

2.2 Algorithms for LCSs

Algorithms to find the best alignments (the ones having the maximal score) have also been well studied. Since [27] developed a dynamic programming algorithm for global alignment, many improvements or variants have been developed—[10] for a linear space improvement, [31] for local alignment, [15] for affine gap penalty, [25, 2, 20] for fast heuristic local alignment, and many more. A detailed review of LCS algorithms can be found in [4].

3 Monte-Carlo Simulation for the Variance

The theoretical study of the variance of the length of LCSs is less complete. A general linear upper bound has been obtained in [32]. Lower bounds, also of linear order, have been proved in various biased instances ([23], [18], [19], [24], [14], [3], ···). But the uniform i.i.d. case is still unknown. In [5], it is observed through Monte-Carlo simulation, with $n$ up to 20,000, that the order of the variance of the length of the LCSs of binary random words is at least of order $n^{2\omega'}$, where $\omega' = 0.418 \pm 0.005$. Our simulation shows that when $n$ becomes larger, such deviation also becomes larger and the variance tends to have order $n$.

3.1 Problem Description

Given two sequences $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ having the same length, where $X_i, Y_i \in A$ and where again $A$ is the alphabet, we explore, by Monte-Carlo method, the asymptotic behavior of $\text{Var} \, LC_n$ when $n$ grows large.

To perform Monte-Carlo simulations, we need to select an algorithm to compute the length of the LCSs. The dynamic programming algorithm is classical but not efficient enough. Since our experiments are only for $|A| = 2$ or $|A| = 4$, we choose to use the WMMM algorithm [23], which is according to [4] very efficient in time and memory when $|A|$ is small.

3.2 Experiment Setting

- The alphabet size is 2 ($|A| = 2$);
- For each $n$ we draw 10,000 random sample for Monte-Carlo simulation.

3.3 Experiment Results

3.3.1 $\mathbb{P}(X_1 = 0) = 0.5$, $\mathbb{P}(X_1 = 1) = 0.5$

In this experiment, $n$ ranges from 50,000:50,000:1,000,000. We plot $\text{Var} \, LC_n$ against $n$ under a log-log scale in Figure [1].

We found the following relation between $\text{Var} \, LC_n$ and $n$ using linear regression

$$\text{Var} \, LC_n \approx 0.0297 n^{0.9086}.$$  

1For all the simulations presented in this paper, the experiments were run on the Partnership for an Advanced Computing Environment (PACE).
Figure 1: **Left**: log-log plot of $\text{Var } LC_n$ versus $n$, **Right**: plot of $\text{Var } LC_n/n^{0.9086}$ versus $n$

Figure 2: **Left**: log-log plot of $\text{Var } LC_n$ versus $n$, **Right**: plot of $\text{Var } LC_n/n^{0.9855}$ versus $n$

Figure 3: **Left**: log-log plot of $\text{Var } LC_n$ versus $n$, **Right**: plot of $\text{Var } LC_n/n^{1.0021}$ versus $n$
\subsection{3.3.2} \( \mathbb{P}(X_1 = 0) = 0.1, \mathbb{P}(X_1 = 1) = 0.9 \)

In this experiment, \( n \) ranges from 50,000:50,000:1,000,000. We plot \( \text{Var} \, LC_n \) against \( n \) under a log-log scale in Figure 2.

We found the following relation between \( \text{Var} \, LC_n \) and \( n \) using linear regression
\[
\text{Var} \, LC_n \approx 0.0208n^{0.9855}.
\]

\subsection{3.3.3} \( \mathbb{P}(X_1 = 0) = 0.01, \mathbb{P}(X_1 = 1) = 0.99 \)

In this experiment, \( n \) ranges from 2,050,000:50,000:4,000,000. We plot \( \text{Var} \, LC_n \) against \( n \) under a log-log scale in Figure 3.

We found the following relation between \( \text{Var} \, LC_n \) and \( n \) using linear regression
\[
\text{Var} \, LC_n \approx 0.0042n^{1.0021}.
\]

In all cases, we conjecture that the order of variance of \( LC_n \) is:
\[
\text{Var} \, LC_n \overset{\text{asym}}{\sim} cn,
\]
where \( c \) is a small constant.

\section{4 Hypothesis Testing for the Similarity of two Sequences}

\subsection{4.1 Testing Procedure}

To test the similarity of two sequences, we propose the following hypothesis testing procedure. Assume we have two sequences \( X = (X_1, \ldots, X_n) \) and \( Y = (Y_1, \ldots, Y_n) \), both of length \( n \), and then define the null and alternative hypothesis as
\[
H_0 : X \text{ and } Y \text{ are i.i.d. uniformly generated} \\
H_a : X \text{ and } Y \text{ have high similarity.}
\]

Based on the results of [17], we use the Z-test and the test statistic is
\[
S = \frac{(LC_n)_{\text{obs}} - ELC_n}{\sqrt{\text{Var} \, LC_n}},
\]
where \((LC_n)_{\text{obs}}\) is the observed length of the LCS of the two sequences being tested, while \( ELC_n \) and \( \text{Var} \, LC_n \) are the expectation and variance of the length of the LCSs of two sequences, their values estimated by Monte-Carlo simulation.

The paper [29] proposed a similarity score based on LCS for comparing two sequences without providing a hypothesis testing procedure, where the estimated LCS statistics were computed for \( n \) up to 1000. Below, we develop a hypothesis testing approach and conduct simulations for \( n = 10,000 \) and extensively verified the effectiveness of the testing method on synthetic sequences.

\subsection{4.2 Experimental Verification}

We conducted several experiments to verify the effectiveness of our testing procedure still using the WMMM algorithm. These experiments shares the following assumptions/parameters:

- The alphabet size is 4 (\(|A| = 4\));
- The two sequences \( X \) and \( Y \) have the same length (\(|X| = |Y| = n\));
- The action of inserting a sequence \( Z \) into another sequence \( X \) is controlled by a parameter \( s \). We divide \( Z \) into \( s \) equally long contiguous segments and \( X \) into \( s + 1 \) equally long contiguous segments, and then insert the \( s \) segments from \( Z \) into corresponding positions in the \( s \) gaps of \( X \), as illustrated in Figure 4. We denote this action as \textsc{insert}(\( Z, X, s \)).
With $n = 1,000,000$, we randomly generated 529 pairs of $X$ and $Y$, and compute $\gamma_4^* \approx \frac{\text{LCS}(X, Y)}{n} \approx 0.654$, $c \approx s^2(\text{LCS}(X, Y))/n \approx 0.0075$.

We use $\alpha = 0.05$, $n = 10,000$ in our experiments. For each Monte-Carlo simulation, we draw 10,000 random samples.

Below are the experiment results.

4.2.1 Null Hypothesis

Here $P(S \leq Z_\alpha) = 0.9893$, and the histogram of $(\text{LC}_n)_{\text{obs}} - \gamma_4^* n / \sqrt{cn}$ is in Figure 5.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{histogram.png}
\caption{Histogram of $(\text{LC}_n)_{\text{obs}} - \gamma_4^* n / \sqrt{cn}$}
\end{figure}

4.2.2 Alternative Hypothesis (1)

$H_a$: We randomly generated two uniform i.i.d. sequences $X'$, $Y'$ of length $m$, and insert a sequence $Z$ of length $n - m$ into $X'$ and $Y'$, obtaining $X$ and $Y$. The results for $p = P(S \leq Z_\alpha)$ are in the following table

| $m$   | $n - m$ | $p$     |
|-------|---------|---------|
| 9,000 | 1,000   | 0       |
| 9,300 | 700     | 0.2284  |
| 9,350 | 650     | 0.4286  |
| 9,400 | 600     | 0.6119  |
| 9,500 | 500     | 0.8541  |
| 9,900 | 100     | 0.9884  |
4.2.3 Alternative Hypothesis (2)

$H_a$: We randomly generated two uniform i.i.d. sequences $X'$, $Y'$ of length $m = 5,000$, and inserted a sequence $Z$ of length $n - m = 5,000$ into $X'$ and $Y'$ obtaining $X$ and $Y$. The difference is now that each piece of the sequence $Z$ has been inserted, with probability 0.8 into both $X'$ and $Y'$, with probability 0.1 into $X'$ alone, and with probability 0.1 into $Y'$ alone.

In this case, $\mathbb{P}(S \leq Z_\alpha) = 0$, and the histogram of $(LC_n)_{obs} - \gamma^* n / \sqrt{cn}$ is in Figure 6.

![Figure 6: Histogram of $(LC_n)_{obs} - \gamma^* n / \sqrt{cn}$](image)

4.2.4 Alternative Hypothesis (3)

$H_a$: We randomly generated two uniform i.i.d. sequences $X'$, $Y'$ of length $m = 5,000$, and insert a sequence $Z$ of length $n - m = 5,000$ into $X'$ and $Y'$ obtaining $X$ and $Y$. This time, each piece of the sequence $Z$ was inserted with probability 0.15 into both $X'$ and $Y'$, with probability 0.4 into $X'$ alone, with probability 0.4 into $Y'$ alone, and with probability 0.05 into neither $X'$ nor $Y'$.

![Figure 7: Histogram of $(LC_n)_{obs} - \gamma^* n / \sqrt{cn}$](image)
In this case, \( \mathbb{P}(S \leq Z_n) = 1 \), and the histogram of \( (LC_n)_{obs} - \gamma^*_n n / \sqrt{cn} \) is in Figure 7.

The experiments show that our proposed testing procedure is effective in that the probability \( \mathbb{P}(S \leq Z_n) \) gets closer to zero when the two sequences have higher similarity.

5 Upper Bound on the Expected Length of LCSs for Multiple Sequences

For two sequences and equally likely letters from \( \mathcal{A} = \{0, 1, \ldots, k - 1\} \), upper bounds on \( \gamma^*_k \) are given in [9], a result which can be extended to an arbitrarily finite number of sequences. Below, following [9], we outline the proof of this extension which will provide upper bounds on \( \gamma^*_{k,m} \), where \( m \) now denotes the number of sequences.

Let \( F(n, s, k) \) be the number of sequences of length \( n \) that contains \( s \), where \( s \) is any fixed sequence of length \( \ell \). Then a counting and inductive argument developed in [9] gives:

Lemma 1.

\[
F(n, s, k) = \sum_{j=\ell}^{n} \binom{n}{j} (k-1)^{n-j}.
\]

(5.1)

Since

\[
\binom{n}{j+1} (k-1)^{n-j-1} \leq \binom{n}{j} (k-1)^{n-j}, \text{ for } j \geq n/k,
\]

(5.1) leads to

\[
F(n, s, k) \leq n \binom{n}{\ell} (k-1)^{n-\ell}, \text{ for } \ell \geq n/k
\]

(5.2)

For a fixed \( s \) of length \( \ell \), the number of ordered \( m \)-tuples of length-\( n \) sequences \( (a_1,a_2,\cdots,a_m) \) that all contains \( s \) as a subsequence is \( F^m(n, s, k) \). Then the total number of such \( (m+1) \)-tuples \( (a_1,a_2,\cdots,a_m,s) \) is

\[
G(n,\ell,k) = \sum_{|s|=\ell} F^m(n, s, k),
\]

where the summation is over all the \( k^\ell \) sequences of length \( \ell \).

Now, let \( g(n,\ell,k) \) be the number of \( m \)-tuples \( (a_1,a_2,\cdots,a_m) \) such that \( LC(a_1,a_2,\cdots,a_m) \geq \ell \), then

\[
g(n,\ell,k) \leq G(n,\ell,k).
\]

(5.3)

Next, let \( h^{(m)}_k(\theta) \) be the proportion of all ordered \( (a_1,a_2,\cdots,a_m) \) such that \( LC(a_1,a_2,\cdots,a_m) \geq \ell \).

Lemma 2. Let \( \theta = \ell/n \), then \( h^{(m)}_k \leq (H_k(\theta))^{mn} \), where

\[
H_k(\theta) = \frac{k^{(\theta/m)-1} (k-1)^{1-\theta}}{\theta^m (1-\theta)^{1-\theta}}.
\]

Moreover, \( H_k(\theta) = 1 \) has a unique solution in the interval \([1/k,1]\). Let \( V_k \) be this solution, then \( H_k(\theta) < 1 \), for \( \theta > V_k \).

Proof. By Lemma 1 as well as \([5.2]\) and \([5.3]\),

\[
h^{(m)}_k = \frac{g(n,\ell,k)}{k^{mn}} \leq \frac{G(n,\ell,k)}{k^{mn}} = \sum_{|s|=\ell} \frac{F^m(n, s, k)}{k^{mn}} \leq k^{\ell-mn} \left\{ \binom{n}{\ell} (k-1)^{n-\ell} \right\}^m.
\]
Thus by Stirling’s formula, ,

\[
\lim_{n \to \infty} \left( \frac{h_k^{(n)}}{n}\right)^{1/n} \leq \lim_{n \to \infty} k^{(\ell - mn)/n} \left\{ \frac{n (n - \ell)}{n} \right\}^{m/n} \\
= k^{\theta - m} (k - 1)^{m - m\theta} \lim_{n \to \infty} \left\{ \frac{n (n - \ell)}{n} \right\}^{m/n} \\
= k^{\theta - m} (k - 1)^{m - m\theta} \frac{1}{\theta^m (1 - \theta)^{m - m\theta}} \\
= H_k(\theta)^m.
\]

To prove the second statement of the lemma, note that \(H_k(\theta) > 0\) for all \(\theta \in [1/k, 1)\) and that

\[
\lim_{\theta \to 1} H_k(\theta) = k^{1/m - 1} \lim_{\theta \to 1} \frac{(k - 1)^{1 - \theta}}{\theta^m (1 - \theta)^{1 - \theta}} = k^{-(m - 1)/m} < 1,
\]

while

\[
H_k(1/k) = k^{1/mk} > 1.
\]

But for \(\theta \in [1/k, 1)\),

\[
\frac{dH_k(\theta)}{H_k(\theta)} = \log \frac{(1 - \theta)k^{1/m}}{(k - 1)\theta} = \begin{cases} 
0 & \text{if } \theta > \theta_k \\
< 0 & \text{if } \theta < \theta_k,
\end{cases}
\]

for some \(\theta_k\). Therefore, there exists a unique solution \(V_k \in [1/k, 1)\), and \(H_k(\theta) < 1\) for \(\theta > V_k\). 

Combining the above results leads to:

**Proposition 1.**

\[
\lim_{n \to \infty} \frac{\mathbb{E}LC_n}{n} \leq V_k.
\]

**Proof.** For any \(\epsilon > 0\) satisfying \(V_k + \epsilon < 1\), separate the total \(k^m n\) tuples of \((a_1, a_2, \ldots, a_m)\) into two categories: those with longest common subsequences longer than \((V_k + \epsilon)n\), and those with longest common subsequences with length at most \((V_k + \epsilon)n\). Thus,

\[
\mathbb{E}LC_n \leq (V_k + \epsilon)n \left\{ 1 - h_k^{(n)}(V_k + \epsilon) \right\} + (V_k + \epsilon)n \left\{ h_k^{(n)}(V_k + \epsilon) \right\} \\
\leq (V_k + \epsilon)n + (V_k + \epsilon)n \left\{ h_k^{(n)}(V_k + \epsilon) \right\} \\
\leq (V_k + \epsilon)n + (V_k + \epsilon)n H_k(1/k) = \mathbb{E}LC_n.
\]

Since \(H_k(\theta) < 1\) for \(\theta > V_k\), the last term converges to 0 as \(n \to \infty\). Thus,

\[
\lim_{n \to \infty} \frac{\mathbb{E}LC_n}{n} \leq V_k + \epsilon,
\]

holds for any \(\epsilon\) satisfying \(V_k + \epsilon < 1\). 

Therefore, from the above proposition, \(V_k \in [1/k, 1)\) such that \(H_k(V_k) = 1\) provides an upper bound on \(\gamma_{k,m}^*\). In particular, letting \(k = 2\), i.e., \(\mathcal{A} = \{0, 1\}\), leads to the following table for \(\gamma_{2,m}^*\), where the lower bounds are obtained in [22].

| number of sequences \(m\) | upper bound for \(\gamma_{2,m}^*\) | lower bound for \(\gamma_{2,m}^*\) |
|--------------------------|--------------------------|--------------------------|
| 2                        | 0.866595                 | 0.781281                 |
| 3                        | 0.793926                 | 0.704473                 |
| 4                        | 0.749082                 | 0.661274                 |
| 5                        | 0.719527                 | 0.636022                 |
| 6                        | 0.698053                 | 0.617761                 |
| 7                        | 0.681605                 | 0.602493                 |
| 8                        | 0.668516                 | 0.594016                 |
| 9                        | 0.657797                 | 0.587900                 |
| 10                       | 0.648819                 | 0.570155                 |
The results of [9] have been improved in [13]. The current multi-sequence result can similarly be improved using the approach there. In particular, this gives for three sequences with binary alphabet, the upper bound 0.791, which is slightly better than 0.793026 obtained above. However for four (or more) sequences, even with an alphabet of size 2, this approach becomes rather cumbersome. Simulation results on $E\text{LC}_n$ are also presented, in some multisequence cases, in [28].

6 Acknowledgement

We would like to thank the Partnership for an Advanced Computing Environment (PACE) for providing a high performance infrastructure to simulate and analyze our experimental results, as well as Karim Lounici for his early input and numerous comments on this paper.

References

[1] Stephen F. Altschul and Warren Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

[2] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[3] Saba Amsalu, Christian Houdré, and Heinrich Matzinger. Sparse long blocks and the variance of the longest common subsequences in random words. *arXiv:1204.1009v2 [math-ph]*, September 2016.

[4] Libor Bareš. *Algorithms for Longest Common Subsequence Computation*. Master thesis, Czech Technical University in Prague, 2009.

[5] Jacques Boutet de Monvel. Extensive simulations for longest common subsequences. *The European Physical Journal B - Condensed Matter and Complex Systems*, 7(2):293–308, 1999.

[6] R. Bundschuh. High precision simulations of the longest common subsequence problem. *The European Physical Journal B - Condensed Matter and Complex Systems*, 22(4):533–541, August 2001.

[7] Kun-Mao Chao and Louxin Zhang. *Sequence Comparison*, volume 7 of *Computational Biology*. Springer London, London, 2009.

[8] Vacláv Chvátal and David Sankoff. Longest Common Subsequences of Two Random Sequences. *Journal of Applied Probability*, 12(2):306–315, 1975.

[9] Vacláv Chvátal and David Sankoff. An upper-bound technique for lengths of common subsequences. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, Massachusetts, 1983.

[10] Peter Clote and Rolf Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, 2000.

[11] Vladimír Dancík. *Expected Length of Longest Common Subsequences*. PhD thesis, 1994.

[12] Joseph G. Deken. Some limit results for longest common subsequences. *Discrete Mathematics*, 26(1):17–31, January 1979.

[13] Joseph G. Deken. Probabilistic behavior of longest-common-subsequence length. In David Sankoff and Joseph Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, Massachusetts, 1983.
[14] Ruoting Gong, Christian Houdré, and Jüri Lember. Lower Bounds on the Generalized Central Moments of the Optimal Alignments Score of Random Sequences. *Journal of Theoretical Probability*, pages 1–41, December 2016.

[15] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708, December 1982.

[16] Daniel S. Hirschberg. A Linear Space Algorithm for Computing Maximal Common Subsequences. *Commun. ACM*, 18(6):341–343, June 1975.

[17] Christian Houdré and Ümit Işlak. A Central Limit Theorem for the Length of the Longest Common Subsequences in Random Words. *arXiv:1408.1559v4 [math]*, January 2017.

[18] Christian Houdré and Jinyong Ma. On the Order of the Central Moments of the Length of the Longest Common Subsequences in Random Words. In *High Dimensional Probability VII*, pages 105–136. Birkhäuser, Cham, 2016.

[19] Christian Houdré and Heinrich Matzinger. On the Variance of the Optimal Alignments Score for Binary Random Words and an Asymmetric Scoring Function. *Journal of Statistical Physics*, 164(3):693–734, August 2016.

[20] W. James Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, January 2002.

[21] Marcos Kiwi, Martin Loebl, and Jiří Matoušek. Expected length of the longest common subsequence for large alphabets. *Advances in Mathematics*, 197(2):480–498, November 2005.

[22] Marcos Kiwi and José Soto. On a Speculated Relation Between Chvátal–Sankoff Constants of Several Sequences. *Combinatorics, Probability and Computing*, 18(04):517–532, July 2009.

[23] Jüri Lember and Heinrich Matzinger. Standard deviation of the longest common subsequence. *The Annals of Probability*, 37(3):1192–1235, May 2009.

[24] Jüri Lember, Heinrich Matzinger, Joonas Sova, and Fabio Zucca. Lower bounds for moments of global scores of pairwise Markov chains. *arXiv:1602.05560 [math]*, February 2016.

[25] David J. Lipman and William R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, March 1985.

[26] George S. Lueker. Improved Bounds on the Average Length of Longest Common Subsequences. *J. ACM*, 56(3):17:1–17:38, May 2009.

[27] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[28] Kang Ning and Kwok Pui Choi. Systematic assessment of the expected length, variance and distribution of Longest Common Subsequences. *arXiv:1306.4253 [cs]*, June 2013.

[29] Jens G. Reich, Heinz Drabsch, and Astrid Däumler. On the statistical assessment of similarities in DNA sequences. *Nucleic Acids Research*, 12(13):5529–5543, July 1984.

[30] David Sankoff and Sylvie Mainville. Common subsequences and monotone subsequences. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 363–365. 1983.

[31] Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[32] J. Michael Steele. An Efron-Stein Inequality for Nonsymmetric Statistics. *The Annals of Statistics*, 14(2):753–758, June 1986.
[33] Sun Wu, Udi Manber, Gene Myers, and Webb Miller. An $O(NP)$ sequence comparison algorithm.

*Information Processing Letters*, 35(6):317–323, September 1990.