On the variance an uniqueness of longest common subsequence

J. Lember

University of Tartu, Estonia

with

H. Matzinger

and

A. Vollmer

University of Bielefeld, Germany

Longest common subsequence

$\mathcal{X}$ – alphabet (finite set)
$X, Y \in \mathcal{X}^n$ two sequences of length n

$$X = X_1, \ldots, X_n, \quad Y := Y_1, \ldots, Y_n.$$

Common subsequence of X and Y is any subsequence of X that is also contained in Y.

Formally: $X_{i_1}, \ldots, X_{i_k}$ is a subsequence of X, if $i_1 < i_2 < \cdots < i_k$; $X_{i_1}, \ldots, X_{i_k}$ is a common subsequence of X and Y, if it is a the same time a subsequence of Y, i.e. there exists $j_1 < \cdots < j_k$ such that

$$X_{i_1} = Y_{j_1}, \quad X_{i_2} = Y_{j_2} \quad , \ldots, \quad X_{i_k} = Y_{j_k}.$$

X =ATAGCGT, Y =CAACATG. A common subsequence of X and Y is AACG, because it is a subsequence of X

Longest common subsequence

$\mathcal{X}$ – alphabet (finite set)
$X, Y \in \mathcal{X}^n$ two sequences of length n

$$X = X_1, \ldots, X_n, \quad Y := Y_1, \ldots, Y_n.$$

Common subsequence of X and Y is any subsequence of X that is also contained in Y.

Formally: $X_{i_1}, \ldots, X_{i_k}$ is a subsequence of X, if $i_1 < i_2 < \cdots < i_k$; $X_{i_1}, \ldots, X_{i_k}$ is a common subsequence of X and Y, if it is a the same time a subsequence of Y, i.e. there exists $j_1 < \cdots < j_k$ such that

$$X_{i_1} = Y_{j_1}, \quad X_{i_2} = Y_{j_2} \quad , \ldots, \quad X_{i_k} = Y_{j_k}.$$

$X =$ATAGCGT, $Y =$CAACATG. A common subsequence of X and Y is AACG, because it is a subsequence of X and Y.

Longest common subsequence

$\mathcal{X}$ – alphabet (finite set)
$X, Y \in \mathcal{X}^n$ two sequences of length n

$$X = X_1, \ldots, X_n, \quad Y := Y_1, \ldots, Y_n.$$

Common subsequence of X and Y is any subsequence of X that is also contained in Y.

Formally: $X_{i_1}, \ldots, X_{i_k}$ is a subsequence of X, if $i_1 < i_2 < \cdots < i_k$; $X_{i_1}, \ldots, X_{i_k}$ is a common subsequence of X and Y, if it is a the same time a subsequence of Y, i.e. there exists $j_1 < \cdots < j_k$ such that

$$X_{i_1} = Y_{j_1}, \quad X_{i_2} = Y_{j_2} \quad , \ldots, \quad X_{i_k} = Y_{j_k}.$$

$X =$ATAGCGT, $Y =$CAACATG. A common subsequence of X and Y is AACG, because it is a subsequence of X and Y.

The longest common subsequence of X and Y is any common subsequence of X and Y that is of maximal length.

X =ATAGCGT, Y =CAACATG AT – a common subsequence;
X =ATAGCGT, Y =CAACATG AACG – LCS.
LCS is often not unique:
X =ATAGCGT, Y =CAACATG AACT – LCS.

The length of LCS, denoted by L_n is used to measure the "relatedness" or "closeness" of X and Y. The bigger L_n (relative to n), the more closed X and Y presumably are.

In our example, $n = 7$, $L_7 = 4$. Thus

$$\frac{L_n}{n} = \frac{4}{7}.$$

Is it big enough?

Applications

Computational molecular biology:
- comparing DNA sequences, $\mathcal{X} = \{A, T, G, C\}$
- comparing protein alignments, $|\mathcal{X}| = 20$ amino acids.

Linguistics: $\mathcal{X}$ - (usual) alphabet.

The easiest case to study: $\mathcal{X} = \{0, 1\}$.

Common subsequence can be represented by an alignment with gaps (alignment with insertions and deletions (indels)). Like

A	T	A	G	C	G	T
C	A	A	C	A	T	G

	A	T	A	G	C			G	T
C	A		A		C	A	T	G	

Another example:

f	a	n	t	h	a	s	t	i	c
f	n	t	a	s	t	i	q	u	e

f	a	n	t	h	a	s	t	i	c			
f		n	t		a	s	t	i		q	u	e

Hamming score would be 1, whilst $L_{10} = 7$, LCS is $fntasti$.

A more general sequence comparison scheme: every alignment (with gaps and mismatch) has a score that can be calculated pair-wise, where: 1) matching with gap costs $-\delta$, 2) mismatching costs $-\mu$, 3) match rewords 1.
One seeks for optimal alignment achieving the highest score.

LCS – a special case of optimal alignment with $\delta = 0$, $\mu > 0$.

Back to LCS.
To distinguish the related sequences from unrelated ones (using LCS), it is important to know about the LCS for unrelated sequences.

Stochastic model: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are the first elements of ergodic processes. Unrelated – the processes are independent.

Then L_n – random variable. (Asymptotic) properties of L_n? .

LCS - superadditive:

$$\underbrace{\begin{vmatrix} X_1 & \cdots & X_n & X_{n+1} & \cdots & X_{n+m} \\ Y_1 & \cdots & Y_n & Y_{n+1} & \cdots & Y_{n+m} \end{vmatrix}}_{L(1 \ldots n+m)} \geq \underbrace{\begin{vmatrix} X_1 & \cdots & X_n \\ Y_1 & \cdots & Y_n \end{vmatrix}}_{L(1 \ldots n)} + \underbrace{\begin{vmatrix} X_{n+1} & \cdots & X_{n+m} \\ Y_{n+1} & \cdots & Y_{n+m} \end{vmatrix}}_{L(n+1 \ldots n+m)}$$

Kingman's subadditve ergodic thm: $\exists$ constant γ such that

$$\frac{L_n}{n} \to \gamma \quad \text{a.s and in } L_1.$$

Independent Bernoulli random variables

Consider the easiest case: $X = X_1, \ldots, X_n$ and $Y = Y_1, \ldots, Y_n$ are i.i.d. Bernoulli with parameter θ and independent of each other.

Mean

γ (Chvatal-Sankof constant) is unknown even for this case. If $\theta = 0.5$ then $\gamma \approx 0.81$. If $\theta \neq 0.5$ then γ is (presumably) bigger.

Variance

Chvatal-Sankof conjecture (1975): if $\theta = 0.5$, then $\mathrm{Var}(L_n) = o(n^{\frac{2}{3}})$.

Steele (1986): $\mathrm{Var}(L_n) \leq \mathbf{P}(X_1 \neq Y_1)n$.

Waterman's conjecture (1994): Steele's bound's cannot be improved: $\mathrm{Var}(L_n) \asymp n$ i.e. $\exists C > c > 0 : cn \leq \mathrm{Var}(L_n) \leq Cn$.

Independent Bernoulli random variables: variance study

Aim: To prove Waterman's conjecture (iid Bernoulli case).

Conditioning on Y: Suppose the outcome of Y is fixed. So, $X_1, \dots, X_n$ is i.i.d Bernoulli, $Y_1, \dots, Y_n$ is non-random.

1) Y is unicolor. Then $L_n = \sum X_i$, $\quad$ Var$(L_n) = \theta(1-\theta)n$: linear.

X :	0	1	1	1	0	1	0	0	1	0
Y :	1	1	1	1	1	1	1	1	1	1

2) $Y_1, \dots Y_{\frac{n}{2}}$ one color, $Y_{\frac{n}{2}+1}, \dots Y_n$ another − variance linear.

X :	0	1	0	1	0	1	0	1	1	0	1	0	1	1	0	1	0	1	0	1
Y :	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0

3) the first $\frac{1}{4}$-th of Y one color, the second another color, and so on. Variance linear

Y : –

In 1) – 3), the variance was driven by the long unicolor blocks proportional to n. What if Y is non-random but as "mixed" as possible?

4) Y is periodic.

X :	0	0	0	0	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0	0
Y :	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Thm (Matzinger, Durringer, L.) If Y is non-random and periodic and X iid Bernoulli with $\theta = 0.5$, then $\mathrm{Var}(L_n) \asymp n$.

Relax the assumption that Y is non-random. Let X and Y both be iid Bernoulli with very low entropy, so θ is very small (for both).

X :	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
Y :	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0

The optimal alignment typically will align mostly 0's:

X :		0		1	1	0	0		0	0	0	0	0	0	1	0	1	0	0	0	0	0	0		0
Y :	0	0	0		1	0	0	1	0	0	0	0	0	0		0	1	0	0	0				1	0

Thm (Matzinger, L.) Let X and Y be independent iid Bernoulli with parameter θ. If θ is small enough, then $\mathrm{Var}(L_n) \asymp n$.

Modeling the relatedness

X and Y have a a common ancestor process: $Z_1, Z_2, \ldots$ that is $\mathcal{X}$-valued iid process.

The ancestor process will mutate independently: there are iid random mappings $f_1, f_2, \ldots \mathcal{X} \to \mathcal{X}$ giving the mutated process $f_1(Z_1), f_2(Z_2), \ldots$.

Finally, some elements of will be deleted via deletion process $D_1, D_2, \ldots$ that is iid Bernoulli. The elements corresponding to 1 remain. These elements are X.

The sequence Y is modeled from the same ancestor process via independent mutations and deletions.

$Z:$	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	common ancestor
$f(Z):$	$f_1(Z_1)$	$f_2(Z_2)$	$f_3(Z_3)$	$f_4(Z_4)$	$f_5(Z_5)$	$f_6(Z_6)$	X mutations
$D^x:$	0	1	1	0	0	1	X deletions
$X:$		X_1	X_2			X_3	
$h(Z):$	$h_1(Z_1)$	$h_2(Z_2)$	$h_3(Z_3)$	$h_4(Z_4)$	$h_5(Z_5)$	$h_6(Z_6)$	Y mutations
$D^y:$	1	1	1	0	1	0	Y deletions
$Y:$	Y_1	Y_2	Y_3		Y_4		

Both processes are still iid but they are not independent. The process

$$(X_1, Y_1), (X_2, Y_2), \ldots$$

is still ergodic, so by superadditivity

$$\frac{L_n}{n} \to \gamma_R \quad \text{a.s and in } L_1.$$

We call X_i and Y_j related if they have the same ancestor. In the Example, X_2 and Y_3 are related.

Aim: To distinguish the related case from unrelated one. One way is to look at L_n, it is worth of looking at all optimal alignments.

2D representation of an alignment

Let X =ATACCGT, Y =CAACATG.
There are 2 LCS: AACG and AACT.

To AACG corresponds 2 alignments:

	A	T	A	C	C			G	T
C	A		A		C	A	T	G	

	A	T	A	C	C			G	T
C	A		A	C		A	T	G	

Which can be represented by the following plots:

G						$*$	
T							
A							
C					$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

G						$*$	
T							
A							
C				$*$			
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

The alignments corresponding to AACT are

G							
T							$*$
A							
C					$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

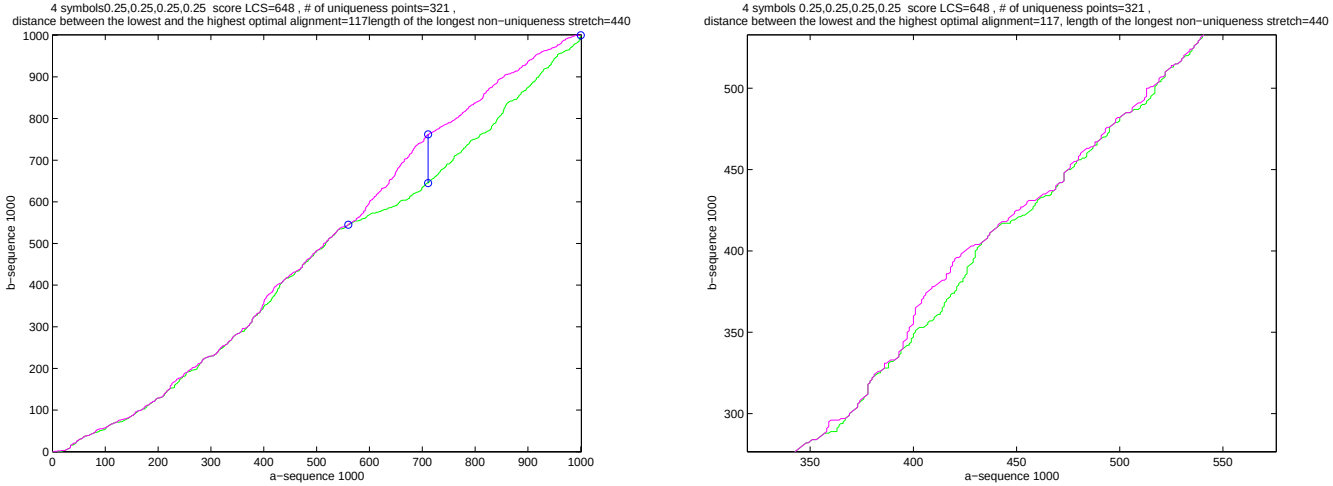
G							
T							$*$
A							
C				$*$			
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

Putting them all in one plot, we can see the uniqueness part as well as the lowest and highest alignment:

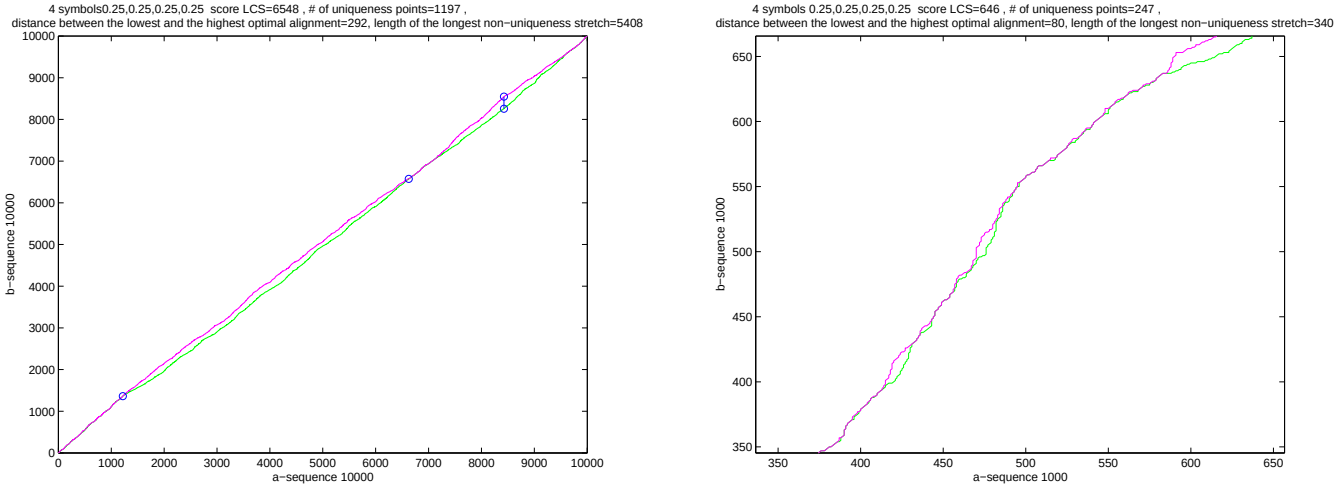
G						$*$	
T							$*$
A							
C				$*$	$*$		
A			$*$				
A	$*$						
C							
	A	T	A	C	C	G	T

It is easy to see that the lowest and highest alignment always exists; they are possible to find by dynamic programming.

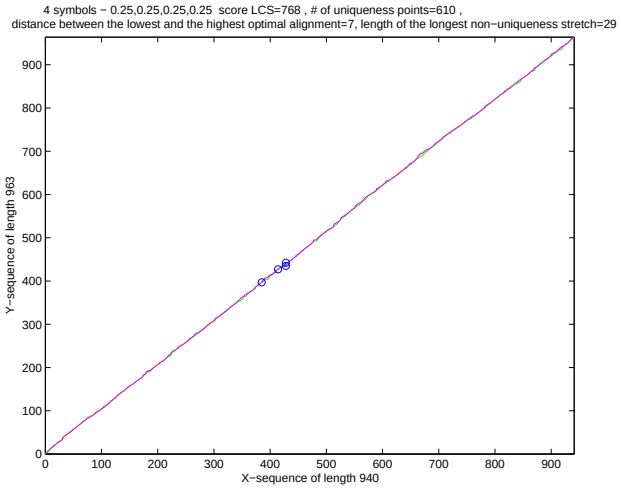
To be more illustrative (for big n), we join the dots by a line to get an alignment graph.

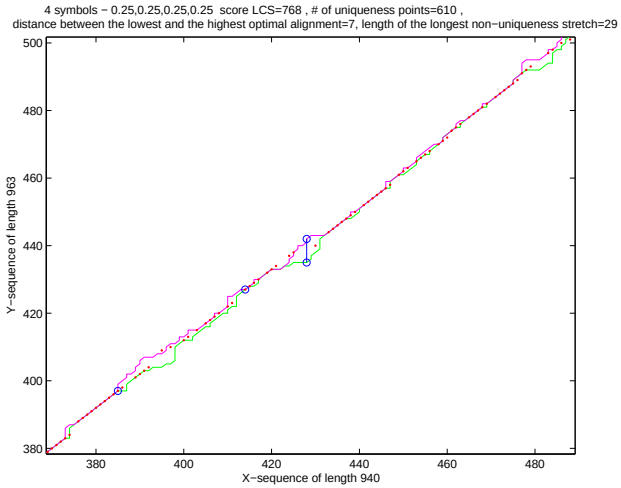


X and Y are independent, $n = 1000$. Right: zoomsection.

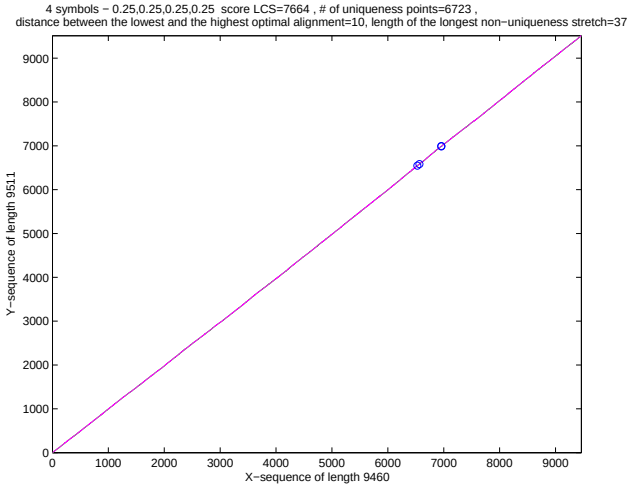


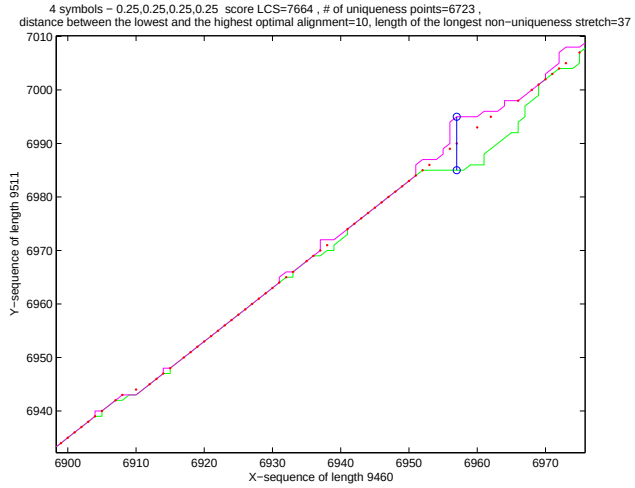
X and Y are independent, $n = 10000$. Right: zoomsection.





X and Y are related, $n = 1000$. Right: zoomsection (red dots: related pairs, equal to ancestor.)





X and Y are related, $n = 10000$. Right: zoomsection.

From the pictures, one can clearly see the difference between the related and unrelated case. How to measure it? Some first ideas:

– Maximal vertical (horizontal) distance

– The length of the maximal non-uniqueness stretch

– Maximal Hausdorff's distance

...

Thm (Matzinger, L.) Assume X and Y are related. Let V_n be the maximal vertical distance between the highest and lowest alignment. Under some assumptions

$$P\Big(V_n > 2C \ln n\Big) \leq Dn^{-1},$$

where C and D are constants.