1. Definitions, known Results and Conjectures

Let $X$ and $Y$ be two finite strings over a finite alphabet $Σ$. A common subsequence of $X$ and $Y$ is a subsequence which is a subsequence of both $X$ as well as of $Y$. A Longest Common Subsequence (LCS) of $X$ and $Y$ is a common subsequence of $X$ and $Y$ of maximal length. In order to get familiar with the definition of an LCS, let us consider the DNA-alphabet $Σ = \{A, C, G, T\}$. Let us consider two sequences $x = AGCTACT$ and $y = ACCTGAATA$. If we compare them letter by letter the great similarity does not become obvious:

$$\begin{array}{cccccccc}
A & G & C & T & A & C & T & A \\
\hline
A & C & C & T & G & A & T & A
\end{array}$$

(1)

The reason is that some letters "got lost" so that they are present only in one of the two sequences. When we align without leaving any gaps for those lost letters, we mostly align corresponding letter pairs. It is better to use gaps and to allow an aligning a letter with a gap. Then the letters which are present only in one of the two sequences get aligned with gaps. Another different alignment is provided by:

$$\begin{array}{cccccccc}
A & G & C & T & A & C & T & A \\
\hline
A & G & C & T & A & C & T & A
\end{array}$$

(2)

We now see a much better coincidence between the two sequences. We displayed in 1 and 2 two possible alignments between $x$ and $y$, 1 without gaps and 2 with gaps. So, when we speak about an alignment, we automatically assume that it only aligns same-letter-pairs or letters with gaps. Each such an alignment defines a common subsequence. An alignment aligning a maximum number of letter pairs is called an optimal alignment. The Common Subsequence defined by an optimal alignment is hence a LCS. For example, the alignment 2 defines the common subsequence $z = AGCTAA$, which consists of the pair of matched letters:

$$\begin{array}{cccccccc}
A & G & C & T & A & C & T & A \\
\hline
A & G & C & T & A & C & T & A
\end{array}$$

(3)

In the alignment 3, the sequence $z = AGCTAA$ is a common subsequence of $X$ and $Y$ with maximal length, therefore a LCS of $x$ and $y$.

1.1 Probabilistic Model

Assume now that $X = X_1X_2...X_n$ and $Y = Y_1...Y_n$ are two i.i.d. strings independent of each other over the same finite alphabet $Σ$. Let $L_n$ denote the length of the LCS of $X$ and $Y$. Using a sub-additivity argument, Chvátal-Sankoff [1] proved that the limit $γ = \lim_{n\to\infty} E[L_n]$ exists. The constant $γ$ depends on the distribution of $X$ and $Y$. However, the value of $γ$ is not known in even such simple cases as when one has two equally likely symbols though there are many simulation showing that $γ ≈ 0.51$. Neither is it known in general if $V[A,L_n]$ is of linear order in $n$. Steele and Talagrand [3] proved that there exists $C > 0$ constant such that $V[A,L_n] ≤ Cn$. Arratia and Waterman [4] derived a law of large deviation for $L_n$ for fluctuations on scales larger than $\sqrt{n}$. The LCS-problem can be formulated also as a last passage percolation problem with correlated weights, moreover Alexander [5] proved that $E[L_n] = o(n)$ converges at a rate of order $\log n/n$ by using first passage percolation methods.

Waterman in [2] conjectured that the variance grows linearly namely $V[A,L_n] = O(n)$, which is still an open conjecture for many kind of distributions of $X$ and $Y$ (including the uniform distribution for $X_i$ and $Y_j$), though in recent years Manthey and collaborators had proved the conjecture to be true for some especial low-entry cases [6].

2. Sequences of Independent Blocks

The aim of this PhD project was to prove, for the first time in the literature, that Waterman’s conjecture is true in a non-low entropy model [10]. The model for $X$ and $Y$ is the following:

- $X$ and $Y$ are two independent sequences of i.i.d. variables uniformly distributed in $[1,1,1,1]$
- $P(B_{X_1} = 1-1) = P(B_{X_2} = 1) = P(B_{X_3} = 1-1) = 1/3$
- $P(B_{X_1} = 1-1) = P(B_{X_2} = 1) = P(B_{X_3} = 1-1) = 1/3$

We call the runs of 0’s and 1’s blocks. Let $X^y = \{X_i, X_j, X_k\}$ be the binary sequence so that the $i$-th block has length $B_{X_1}$, taking the first symbol at random. Similarly let $Y^y = \{Y_i, Y_j, Y_k\}$ be the binary sequence so that the $i$-th block has length $B_{X_1}$, taking the first symbol at random too. Let $X$ denote the sequence obtained by only taking the first $n$ bits of $X^\infty$: $X = X_1X_2X_3...X_n$ and similarly $Y = Y_1Y_2Y_3...Y_n$. Let $L_n$ denote the length of the LCS of $X$ and $Y$, namely $L_n = \{L_n(X, Y)\}$. In this context, we have:

**Theorem 1.** There exists $l_0 > 0$ so that for all $l \geq l_0$, we have that

$$\text{VAR}[L_n] = O(n)$$

for every $n$ large enough.

We show that the above theorem is equivalent to proving that “a certain random modification has a biased effect on $L_n$.” This is a technique with similar approaches in other papers (for instance see [7], [9]). We choose at random in $X$ a block of length $l+1$ and at random one block of length $l+1$. This means that all the blocks in $X$ of length $l+1$ have the same probability to be chosen and then we pick one of those blocks of length $l+1$ and also that all the blocks in $X$ of length $l+1$ have the same probability to be chosen and we pick one of those blocks of length $l+1$ and then we change the length of both these blocks to $l$. The resulting new sequence is denoted by $X$. Let $L_n$ denote the length of the LCS after our modification of $X$. Hence $L_n = \{L_n(X, Y)\}$. The next theorem proves that if ou block change operation has a bias effect on the LCS then the order of the fluctuation of $L_n$ is $n^{\alpha}$, namely:

**Theorem 2.** Assume that there exists $\epsilon > 0$ and $\alpha > 0$ not depending on $n$ such that for all $n$ large enough we have:

$$\text{Pr}[|E[L_n] - L_n(X, Y)| ≥ \epsilon n] \geq 1 - \epsilon n^{-\alpha}.$$  (4)

Then,

$$\text{VAR}[L_n] = O(n)$$

for every $n$ large enough.

Turns out that condition 4 can be verified by considering an optimization problem for the proportion of aligned block pairs and the proportion of left out blocks in the optimal alignment. Let $p_1$ designate the proportion of aligned block pairs which take a block in $X$ having length and a block in $Y$ having length $y$. Let $F(n)$ denote the event that any optimal alignment of $X$ and $Y$ leaves at least a proportion $\delta$ of blocks in $X$ and leaves out the same proportion $q$ of blocks in $Y$.

**Theorem 3.** Assume that there exists $q_0$ in $[0, (1/3)]$ such that the following minimizing problem:

$$\min \sum_{i,j} p_{i,j} \geq \left(1/3 - q_0\right)/2$$

under the conditions:

$$\sum_{i,j} p_{i,j} = 1, p_{i,j} ≥ 0, \forall i, j \in \{0, 1\}$$

has a strictly positive solution. Let this minimum be equal to $2c > 0$. Then we have that:

$$\text{Pr}\left[|E[L_n] - L_n(X, Y)| ≥ \epsilon n\right] \geq 1 - e^{-n^2} - F(n^{2c})$$

where $\delta > 0$ is a constant not depending on $n$ and $F(n^{2c})$ is exponentially small.

It is important to emphasize two aspect which are not explicitly mentioned above:

1. the optimization problem in Theorem 3 was solved analytically and numerically though is a hard optimization problem, from where in Theo-rem 1 one can use $n_0 = 5$.
2. Theorem 2 is the most technical part of the the- sis which involved the use of certain kind of functional inequalities, control of renewal pro- cesses, asymptotic expansions, large deviation techniques and random walk estimates.

References