Longest Common Subsequence Problem for Sequences of Independent Blocks

Bielefeld University
Stochastics and Real World Models
International Graduate College

FELIPE TORRES

International Graduate College "Stochastics and Real World Models"
Beijing — Bielefeld

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 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 a LCS, let us consider the DNA-alphabet $\Sigma = \{A, G, C, T\}$. Let us consider two sequences $x = ACGTAGCA$ and $y = ACCGTATA$. If we compare them letter by letter the great similarity does not become obvious:

$$\begin{array}{c|c|c|c|c|c|c|c|c}
x & A & C & G & T & A & G & T & A \\
\hline
y & A & C & C & G & T & A & C & A
\end{array} \tag{1}$$

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 non-corresponding letter pairs. It is better to use gaps and to allow 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:

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
x & A & C & & G & T & A & G & & T & A \\
\hline
y & A & C & C & G & T & A & & C & & A
\end{array} \tag{2}$$

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 $z = ACGTAA$, which consists of the pair of matched letters:

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|c}
x & A & C & & G & T & A & G & & T & A \\
\hline
y & A & C & C & G & T & A & & C & & A \\
\hline
z & A & C & & G & T & A & & & & A
\end{array} \tag{3}$$

In the alignment 3, the sequence $z = ACGTAA$ 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 $X = X_1X_2 \dots X_n$ and $Y = Y_1Y_2 \dots Y_n$ 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, Chvàtal-Sankoff [1] proved that the limit $\gamma := \lim_{n\to\infty} \mathrm{E}[L_n]/n$ exists. The constant γ depends on the distribution of X_1 and Y_1. However, until the date the exact 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 $\gamma \approx 0.81$. Neither it is known in general if $\mathrm{VAR}[L_n]$ is of linear order in n. Steele in [3] proved that there exists $C > 0$ constant such that $\mathrm{VAR}[L_n] \leq Cn$. 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 $\mathrm{E}[L_n]/n$ converges at a rate of order $\sqrt{\log n/n}$ by using first passage percolation methods.

Waterman in [2] conjectured that the variance grows linearly namely $\mathrm{VAR}[L_n] = \Theta(n)$, which is still an open conjecture for many kind of distributions of X and Y (including the uniform distribution for X_1 and Y_1), though in recent years Matzinger and collaborators had proved the conjecture to be true for some especial low-entropy cases ([6], [7], [8], [9]).

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: fix an integer $l > 0$ and take $B_{X1}, B_{X2}, \dots$ and $B_{Y1}, B_{Y2}, \dots$ two sequences of i.i.d. variables uniformly distributed in $\{l-1, l, l+1\}$ as follows:

$$P(B_{Xi} = l-1) = P(B_{Xi} = l) = P(B_{Xi} = l+1) = 1/3$$

$$P(B_{Yj} = l-1) = P(B_{Yj} = l) = P(B_{Yj} = l+1) = 1/3$$

We call the runs of 0's and 1's blocks. Let $X^\infty = X_1X_2X_3\dots$ be the binary sequence so that the i-th block has length B_{Xi}, taking the first symbol at random. Similarly let $Y^\infty = Y_1Y_2Y_3\dots$ be the binary sequence so that the i-th block has length B_{Yi}, taking the first symbol ar random too. Let X denote the sequence obtained by only taking the first n bits of X^∞: $X = X_1X_2X_3\dots X_n$ and similarly$Y = Y_1Y_2Y_3\dots Y_n$. Let L_n denote the length of the LCS of X and Y, namely $L_n := |\mathrm{LCS}(X, Y)|$. In this context we have:

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

$$\mathrm{VAR}[L_n] = \Theta(n)$$

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 approches 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 $l-1$ up 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 $l+1$ up. Then we change the length of both these blocks to l. The resulting new sequence is denoted by $\tilde{X}$. Let $\tilde{L}_n$ denote the length of the LCS after our modification of X. Hence $\tilde{L}_n := |\mathrm{LCS}(\tilde{X}, Y)|$. The next theorem proves that if ou block length changing operation has typically a biased effect on the LCS than the order of the fluctuation of L_n is $\sqrt{n}$, 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:

$$\mathrm{P}\Big(\ \mathrm{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon \ \Big) \geq 1 - \exp(-n^\alpha). \tag{4}$$

Then,

$$\mathrm{VAR}[L_n] = \Theta(n)$$

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_{ij} designate the proportion of aligned block pairs which take a block in X having length i and a block in Y having length j. Let $F^n(q)$ denote the event that any optimal alignment of X and Y leaves out at most a proportion $q \in [0, 1]$ of bocks 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:

$$\min \left(\frac{(p_{l-1,l} + p_{l-1,l+1})(1-9q)}{p_{l-1,l-1} + p_{l-1,l} + p_{l-1,l+1}} - \frac{p_{l+1,l+1}(1-3q)}{p_{l+1,l-1} + p_{l+1,l} + p_{l+1,l+1}} - 3q \right)$$

under the conditions:

$$q \in [0, q_0]$$

$$\sum_j p_{l-1,j} \geq ((1/3) - q_0)/2$$

$$\sum_j p_{l+1,j} \geq ((1/3) - q_0)/2$$

$$\sum_{i,j \in I} p_{ij} = 1, p_{ij} \geq 0, \forall i, j \in I$$

$$-2\left(q\ln(q) + (1-q)\ln(1-q)\right) + (1-4q)(\ln(1/9) + H(p)) \geq 0$$

where we define

$$H(p) := \sum_{i,j \in \{l-1,l,l+1\}} p_{ij} \ln(1/p_{ij})$$

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

$$\mathrm{P}\Big(\ \mathrm{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon \ \Big) \geq 1 - e^{-n^\beta} - \mathrm{P}(F^{nc}(q_0))$$

where $\beta > 0$ is a constant not depending on n and $\mathrm{P}(F^{nc}(q_0))$ 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 Theorem 1 one can use $l_0 = 5$.
2. Theorem 2 is the most technical part of the thesis which involved the use of certain kind of functional inequalities, control of renewal processes, asymptotic expansions , large deviation techniques and random walk estimates.

References

[1] V. Chvatal and D. Sankoff. *Longest common subsequences of two random sequences.* J. Appl. Probability, 12 : 306–315, 1975.

[2] M. S. Waterman. *Estimating statistical significance of sequence alignments.* Phil. Trans. R. Soc. Lond. B, 344:383-390, 1994.

[3] M. J. Steele. *An Efron- Stein inequality for non-symmetric statistics.* Annals of Statistics, 14:75–758, 1986.

[4] R. Arratia and M. S. Waterman. *A phase transition for the score in matching random sequences allowing deletions.* Ann. Appl.

[5] Kenneth S. Alexander. *The rate of convergence of the mean length of the longest common subsequence.* Ann. Appl. Probab., 4(4):1074–1082, 1994.

[6] J. Lember, H. Matzinger, and C. Durringer. Deviation from mean in sequence comparison with a periodic sequence. Alea, Volume 3:1–29, 2007.

[7] F. Bonetto and H. Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. Latin American Journal of Probability and Mathematics, Volume 2:195–216, 2006.

[8] C.Houdre, J. Lember, H. Matzinger. *On the longest common increasing binary subsequence.* C.R. Acad. Sci. Paris, Ser. I 343:589–594, 2006.

[9] J. Lember, H. Matzinger. *Stardard Deviation of the Longest Common Subsequence.* Ann. Probab. Volume 37, Number 3: 1192-1235, 2009.

[10] H. Matzinger, F. Torres. *Fluctuations of the longest common subsequence for sequences of independent blocks.* Submitted, August 2009.