Journal of Siberian Federal University. Mathematics & Physics 2017, 10(1), 71–74

УДК 004.412

A Formula for the Mean Length of the Longest Common Subsequence

Sergej V. Znamenskij*
Ailamazyan Program Systems Institute of RAS
Peter the First, 4, Veskovo village,
Pereslavl area, Yaroslavl region, 152021
Russia

Received 10.10.2016, received in revised form 10.11.2016, accepted 20.12.2016

The expected value E of the longest common subsequence of letters in two random words is considered as a function of the $\alpha = |A|$ of alphabet and of words lengths m and n. It is assumed that each letter independently appears at any position with equal probability. A simple expression for $E(\alpha, m, n)$ and its empirical proof are presented for fixed α and $m + n$. High accuracy of the formula in a wide range of values is confirmed by numerical simulations.

Keywords: longest common subsequence, expected value, LCS length, simulation, asymptotic formula.
DOI: 10.17516/1997-1397-2017-10-1-71-74.

Introduction

The random words of lengths m and n in the alphabet α are also reffered as random symbol sequences. We consider the letter appearance in different positions of words as equally probable and independent events. So for those random sequences the expected value of the longest common subsequence length is a function of $E(m, n, \alpha)$, which reflects the similarity of the original words.

Since the behavior of this function is related to a variety of generic algorithms for fuzzy search and differences identification, it attracts the attention of researchers for a three decades [1]. However, both the use of mathematical apparatus as in [2, 3] and numerical modeling (usually with special algorithms) [4, 5, 6] succeeded to clarify situation only in special cases $m = n$ or $\alpha = 2$ (see [7]).

Even for $\alpha = 2$ the asymptotic on $\dfrac{m}{n}$ became clear only resently [8] (just now without detailed proof). Computer calculations E for small m, n, in [9] have identified a similar relation for the $\alpha = 4$.

The work is intended to the detection and empirical proof of this relation with except of huge α and small $m + n$ cases.

* svz@latex.pereslavl.ru

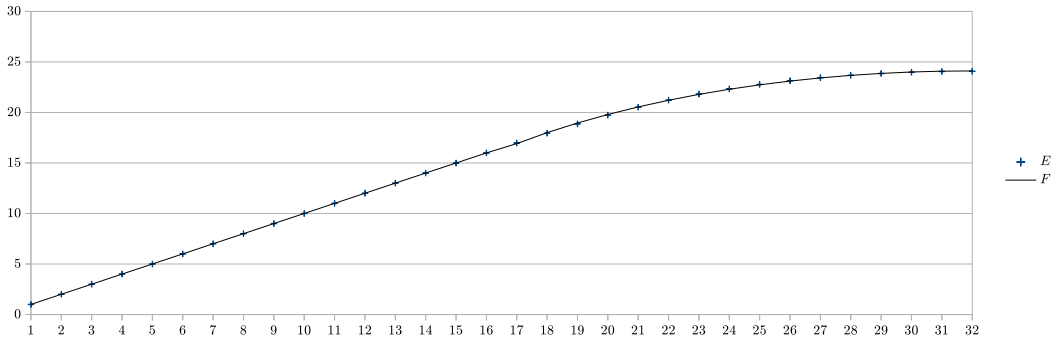

Fig. 1. The case $\alpha = 2$, $n = 64$ with of the least relative accuracy

1. Model

Hypothesis 1. *The functions* $r_x = r_x(\mu, \alpha)$ *and* $r_y = r_y(\mu, \alpha)$ *exists, for which under the notions* $\delta = \dfrac{m-n}{2r_x}$, $r = \sqrt{r_x^2 + r_y^2}$ *the function*

$$F(m,n,\alpha) = \begin{cases} n, & 1 < r\delta \\ \frac{m+n}{2} - r + r_y\sqrt{1-\delta^2}, & -1 < r\delta < 1 \\ m, & r\delta < -1 \end{cases}$$

gives a fine approximation for $E(m,n,\alpha)$ *at least for all* $\alpha < 128$ *and* $50 < m + n < 100000$.

2. Evaluation

Direct evaluation for huge LCS lengths is impossible due to known square complexity of algorithm. Therefore 6 fixed values of $m + n$ and 10 for α were selected and for 6×10 series of 32 triplets (m, n, α) their expected values of LSS lengths were calculated as sample means. A perl XS module with a speed compatible with C compiled code was used. Required number of calculations and processed time were detected in a series of runs attempted to get large enough samples for acceptable accuracy. The full collected sample data is available over email to author.

The r_x and r_y values were calculated which minimizes the mean square error. All the results are presented in the Table 1. We note $\mathcal{I}_{m,n,\alpha}$ the samle set of all calculated lengths of LCS for generated random words, $\overline{E}(m,n,\alpha)$ their means over each $\mathcal{I}_{m,n,\alpha}$ and calculate their experimental standard deviations

$$\sigma_{\overline{E}}(s)) = \max_{n+m=s,\alpha} \sqrt{\frac{1}{|\mathcal{I}_{m,n,\alpha}| - 1} \sum_{i \in \mathcal{I}_{m,n,\alpha}} l_i}, \quad \sigma_F(s) = \sqrt{\frac{1}{31} \sum_{j=1}^{32} (\overline{E}(x_{j,s},\alpha) - F(x_{j,s},\alpha))^2},$$

where $x_{j,s} = \left(\dfrac{js}{64}, s - \dfrac{js}{64}\right)$.

The worst matches from all 6×10 tests are shown on the Fig. 1.

Table 1. The results of empirical proof

α	notation	Total length of both sequences $m+n$					
		64	**256**	**1024**	**8192**	**16384**	**65536**
		Optimal (r_x, r_y) values for fixed $m+n$ and α					
2	r_x, r_y	21.389 25.056	75.265 89.92	286.241 351.613	2231.881 2796.965	4425.896 5528.217	17571.808 21893.814
3		34.289 47.844	117.155 153.426	434.358 551.615	3329.878 4155.315	6606.47 8198.862	26269.533 32525.718
4		36.57 44.483	129.026 150.263	484.121 548.949	3722.583 4135.24	7417.019 8233.956	29509.918 32664.38
5		36.674 38.624	132.592 136.337	502.645 506.569	3881.93 3835.89	7731.749 7628.555	30838.127 30430.938
6		36.407 33.926	133.781 123.444	511.075 465.027	3970.617 3562.51	7904.151 7067.06	31531.221 28201.948
7		36.082 30.304	134.101 112.73	515.004 428.676	4017.635 3310.963	8001.95 6574.387	31922.49 26229.624
8		35.796 27.519	134.119 104.035	517.126 398.526	4043.61 3090.13	8063.285 6155.939	32145.348 24503.713
16		34.539 17.121	132.891 68.706	519.659 270.916	4098.747 2136.885	8181.072 4263.338	32660.916 17022.721
32		33.752 11.037	131.601 46.221	518.343 185.746	4106.893 1481.344	8202.243 2959.956	32763.324 11829.582
128		32.945 4.741	130.073 21.549	515.861 89.437	4104.696 725.75	8202.921 1453.251	32784.888 5820.357
		CPU core time spent ih hours					
	τ	55.1	40.1	13.5	10.7	26.4	399.6
		Total number of CLS calculations					
	$\mathcal{N}$	5658012818	719343652	28547575	489896	100227	103149
		Maximum over (α, n) of LSC length precision					
	$\sigma_{\overline{E}}$	1.577	2.491	4.025	9.408	12.279	23.573
		Precision of F					
	σ_F	0.027	0.039	0.058	0.355	1.043	1.736
		Relative accuracy for F					
	ϵ	0.12%	0.053%	0.027%	0.020%	0.029%	0.011%

Acknowledgments

This work was performed under financial support from the Government, represented by the Ministry of Education and Science of the Russian Federation (Project ID RFMEFI60414X0138).

References

[1] V.Chvátal, D.Sankoff, Longest common subsequences of two random sequences, *J. Appl. Prob.*, **12**(1975), 306–315.

[2] M.A.Kiwi, M.Loebl, J.Matousek, Expected length of the longest common subsequence for large alphabets, *Advances in Mathematics*, **197**(2005), no. 2, 480–498.

[3] G.S.Lueker, Improved bounds on the average length of longest common subsequences, *Journal of the ACM (JACM)*, **56** (2009), no. 3, 17.

[4] R.Bundschuh, High precision simulations of the longest common subsequence problem, *The European Physical Journal B - Condensed Matter and Complex Systems*, **22**(2001), no. 4, 533–541.

[5] J.Boutet de Monvel, Extensive simulations for longest common subsequences *The European Physical Journal B - Condensed Matter and Complex Systems*, **7**(1999), no. 2, 293–308.

[6] R.Baeza-Yates, G.Navarro, R.Gavaldá, R.Schehing, Bounding the expected length of the longest common subsequences and forests, *Theory of Computing Systems*, **32**(1999), no. 4, 435–452.

[7] Kang Ning, Kwok Pui Choi, Systematic assessment of the expected length, variance and distribution of Longest Common Subsequences //arXiv preprint arXiv:1307.2796, 2013.

[8] J.D.Dixon, Longest common subsequences in binary sequences //arXiv preprint arXiv:1307.2796, 2013.

[9] S.V.Znamenskij, A picture of common subsequence length for two random strings over an alphabet of 4 symbols, *Program systems: theory and applications*, **7**(2016), no. 1(28), 201–208.

Формула для средней длины длиннейшей общей подпоследовательности

Сергей В. Знаменский
Институт программных систем РАН
Петра Первого, 4,
Переславльский район, Ярославская обл., 152021
Россия

Математическое ожидание E длиннейшей общей подпоследовательности букв двух случайных слов рассматривается как функция от мощности алфавита $|A|$ и длин m и n этих слов. При этом предполагается, что любая буква независимо и с равной вероятностью оказывается в любой позиции слова. Предъявлено простое выражение для $E(\alpha, m, n)$ при фиксированных α и $m + n$.

Ключевые слова: длиннейшая общая подпоследовательность, математическое ожидание, длина LCS, численное моделирование, асимптотическая формула